SOCIAL MEDIA MINING FOR VETERINARY EPIDEMIOLOGICAL SURVEILLANCE

Submitted by

Amir Samuel Munaf

for the degree of

Doctor of Philosophy

at the Division of Computing Science and Mathematics

The University of Stirling

September 2023

STATEMENT OF DECLARATION

I declare that this thesis is an original report of my research, has been written by me and has not been submitted for any previous degree. The experimental work is almost entirely my own work; the collaborative contributions have been indicated clearly and acknowledged. Due references have been provided on all supporting literatures and resources.

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification.

Signed………………………………. ……………………………………………

ACKNOWLEDGMENTS

I would like to thank my family, friends and supervisors for their continued support and guidance throughout this PhD.

"You are now about to witness the strength of street knowledge"

Table of Contents

Glossary

ACF: Autocorrelation Function

ADF: Augmented Dicky-Fuller

AH: Animal health

AI: Artificial intelligence

AIC: Akaike Information Criteria

ANN: Artificial neural network

APHA: Animal and Plant Health Agency

API: Application Programming Interface

ARMA: Auto-regressive Moving Average

ARIMA: Auto-regressive Integrated Moving Average

ARIMAX: Auto-regressive Integrated Moving Average with Exogenous variables

AVIN: Avian Influenza

AUC: Area Under the Curve

BGC: Bagging Classifier

BIC: Bayesian Information Criterion

CDC: Centre for Disease Control

COVID-19: Coronavirus disease 2019

CSV: Comma-Separated Values

CV: Computer vision

CSS: Cascading Style Sheets

DEFRA: Department for Environment, Food and Rural Affairs

DS: Data science

DT: Decision Tree

ES: Epidemiological surveillance

GFT: Google Flu Trends

GIS: Geographic Information System

HBM: Health Belief Model

HPAI: Highly Pathogenic Avian Influenza

HTML: Hypertext Markup Language

HQIC: Hannan–Quinn information criterion

ILI: Influenza-like Illness

IoT: Internet of things

JSON: JavaScript Object Notation

KDE: Kernal density estimation

KNN: K-nearest neighbour

LDA: Latent Dirichlet Allocation

LR: Logistic Regression

ML: Machine Learning

MNB: Multinomial Naïve Bayes

NLP: Natural language processing

NLTK: Natural Language Toolkit

NWA: Network Analysis

PACF: Partial Autocorrelation Function

PCA: Principal components analysis

PCC: Pearsons' Correlation Coefficient

PHE: Public Health England

PHS: Public Health Scotland

ROC: Receiver Operating Characteristic

RF: Random Forrest

RCUK: Research councils UK

ReSST: Real time syndromic surveillance team

RF: Random Forrest

SARIMA: Seasonal Auto-regressive Integrated Moving Average

SAVSNET: Small Animal Veterinary Surveillance Network

SMOTE: Synthetic Minority Oversampling Technique

SNA: Social network analysis

SQL: Structured Query Language

SRUC: Scotland's Rural College

SS: Syndromic surveillance

SVC: Support Vector Classifier

SVD: Singular Value Decomposition

TF-IDF: Term frequency-inverse document frequency

TOE: Technology Organization Environment

TAM: Technology Acceptance Model

TPB: Theory of Planned Behaviour

WHO: World Health Organisation

XML: Extensible Markup Language

List of Figures

List of Tables

Publications

Title	Journal	Link
Text mining of veterinary forums for epidemiological surveillance supplementation	**Springer nature** *Social network analysis and mining*	https://doi.org/10.1007/s13278-023-01131-7
Social media network analysis of smallholder livestock farming communities	**Heliyon** *Life Sciences*	https://doi.org/10.1016/j.heliyon.2023.e23265
Spatio-temporal evaluation of social media as a tool for livestock disease surveillance	**One Health** *Public Health*	https://doi.org/10.1016/j.onehlt.2023.100657
Text classification of UK smallholding communities through Twitter	**BMC Veterinary research** *Zoonotic diseases collection*	https://doi.org/10.21203/rs.3.rs-2670842/v1

Abstract

Extensive records are kept in the UK regarding large-scale farms, which include information on farm sizes, locations, disease outbreaks, and the movement of animals. This data enables a nuanced understanding of the disease risks associated with commercial farms. Unfortunately, there is a lack of documented data on small-scale farms, making it difficult to evaluate the risks linked with them, despite literature inferring that they play a crucial part in epidemiological surveillance.

The primary aim of this project was to evaluate the viability of using social media data as an instrument of passive surveillance for both identifying smallholding communities and early disease detection. This includes assessing the availability and quality of sufficient data, in addition to deriving meaningful inferences about the animal health population within the United Kingdom. Through the use of numerous data science techniques, such as text classification, topic modelling, social network analysis, and spatio-temporal analysis, it was possible to gain insights into the demographics, concerns, and interactions of these communities.

Offering a new perspective on disease surveillance and control for policymakers, veterinarians, and agricultural experts, social media platforms have great potential to supplement traditional surveillance, as indicated by the findings. While the research faced limitations, such as the rapidly evolving nature of social media and the specific focus on English-language platforms only, it still added valuable insights to the growing body of knowledge. With the ever-increasing integration of digital and physical domains in today's world, this research points towards new opportunities for interdisciplinary research in data science and livestock farming.

Main contributions from this work:

- **Digital Surveillance Mechanism:** Formulated an innovative methodology for monitoring and analysing smallholder discussions, concerns and actions on the internet in niche fora.
- **Predictive Modelling:** Machine learning models have been introduced that can classify smallholding users based on their profile descriptions, providing a valuable tool for rapid identification.
- **Disease Outbreak Analysis:** Leveraged spatio-temporal analysis to link online discussions with real-world events, providing a potential early warning system for disease outbreaks.
- **Network Analysis:** Unveiled the complex social dynamics of the smallholder community, pinpointing crucial nodes and pathways of information diffusion.

1/. Introduction

In the UK, we have significant amounts of data on large-scale pig and poultry farmers, including information on farm sizes, disease outbreaks and transportation of animals[1]. These commercial holdings are obligated to provide this information to the Department for Environment, Food and Rural affairs (DEFRA) and risk financial sanctions for non-compliance to animal health regulations[2].

We therefore have access to data which can contribute to an understanding of the risks, in terms of disease outbreaks, associated with these commercial farms. However, data on small-scale farmer/smallholders is significantly lacking in both quality and quantity[1]. This creates a large disparity between the actual number of these premises and those which are directly recorded in official registries and databases. Those individuals who own fewer than fifty birds are not legally required to register their flock with the Government, hence it results in a substantial proportion of livestock being omitted from surveillance databases and consequently complicating the control of notifiable diseases like Highly Pathogenic Avian Influenza (HPAI)[3].

Recent outbreaks of HPAI in the UK demonstrate the risks posed to such producers, as well as the potential risks that they pose to larger scale commercial producers[4]. Effective animal disease surveillance should account for such smallholders, who are known to use social media (e.g., Facebook and Twitter), advice fora such as Quora, and search engines like Google to learn more about animal husbandry and health[1].

Establishing underlying information about the farm population at risk is critical for the effective development and application of disease surveillance and control. Pig and poultry premises represent two useful, contrasting examples: the existence and locations of small-scale pig holders can be inferred from some recorded data, but small poultry holders are particularly problematic, as there is no requirement for recording of any kind in the UK[5].

This project attempts to bridge the gap in the smallholder database by using modern, non-traditional sources of data to firstly recognise the types and locations of smallholding premises. In addition, comparing the results to what is known about these premises from more conventional sources to establish the efficacy of the findings. In continuation, the types of information that such smallholders search for was also determined; with respect to their animals and the concerns they have. These livestock-related concerns consisted of clinical signs of disease, other health-related issues pertaining to animal husbandry, biosecurity queries and the use of medicines in these animals.

Scotland's Rural College (SRUC) Veterinary Services collects information on reports of suspected cases of disease[6]. Among poultry, this information is collected particularly from small or backyard poultry operations. Search terms trends were compared with the occurrence of known animal health events (e.g., HPAI outbreaks) and with veterinary surveillance data to establish the utility of social media and other non-traditional data sources in the identification of such trends and events.

Finally, these new data sources were scanned and analysed automatically using the appropriate data science tools for finding relevant social media content. Automating the process will allow an on-going service to run, continuously updating key measurements and reflecting the current state of smallholder animal health.

1.1 Aims and Objectives

- To use non-traditional data sources (e.g., social media and search engine searches) to gain information about small holder pig and poultry farmers, where they are, how they interact, and what their animal health related concerns are.

- To better inform our knowledge of the demographics of small holders for purposes of animal disease surveillance and control.

- To characterise the networks of small-scale pig and poultry holders as identified in this data.

- To compare any trends identified in social media and search engine searches to known animal health related events, to ascertain the utility of these data sources for identifying such events in the future.

- To continuously and automatically monitor those trends to produce further data and insights into small holder animal health by developing a set of tools that can be used by the industry to drive further pig/poultry health improvements.

1.2 Research questions

The following are the research questions proposed in this PhD and are all centred around gathering passive intelligence of smallholding communities:

1. ***Is it possible to differentiate between a smallholder and non-smallholder user based on their online profile and content?***
2. ***What are the differences between the social media extracted data and the current knowledge database?***
3. ***How do these communities exchange information and what is the information being discussed?***
4. ***Can social media sources be used as an early predictor of outbreaks?***
5. ***Is it possible to extract locations of smallholders and/or outbreak hotspots?***

To answer these research questions, the combination of all 3 disciplines in figure 1 were incorporated, presenting a multi-layered approach to tackling each objective:

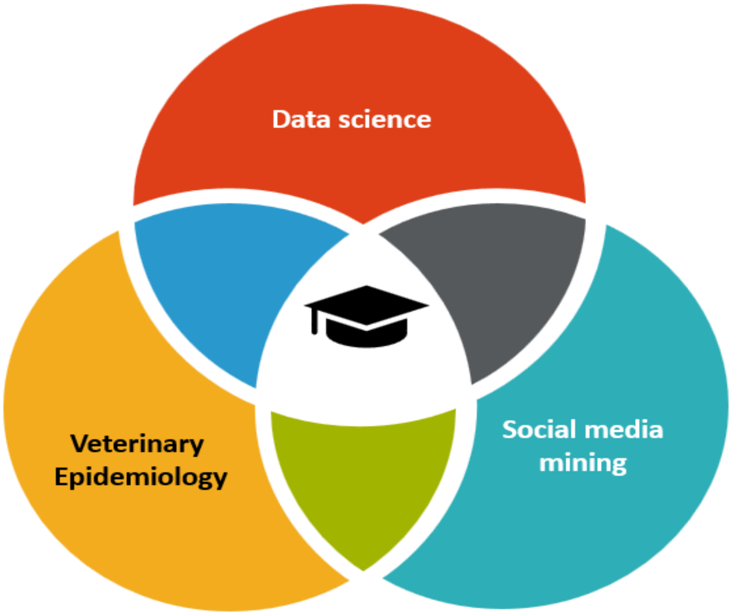


FIGURE 1: INTERSECTION OF DISCIPLINES

1.3 Rationale of Study Focus

The decision to research pig and poultry smallholders is backed by reasons like economics, epidemiology, and socio-cultural factors. This section clarifies why this specific group was chosen, highlighting their important but often overlooked role in livestock epidemiology.

1.3.1 Epidemiological Importance

The agricultural economy relies on pig and poultry farming, which plays a vital role in ensuring global food security[7]. Small-scale farmers are vital for development in regions, as they support livelihoods and enhance the resilience of food systems. This research aims to shed light on their practices, challenges, and potentials, providing a holistic understanding of agricultural productivity and sustainability. Pig and poultry smallholders have been overlooked in academic research and policy-making discussions, despite their importance. The overall agricultural landscape suffers from gaps in our understanding of the operations, risks, and contributions of these farmers because of a lack of focus[8]. By focusing on this group, the study fills a research gap and provides valuable insights for customising interventions and support for smallholders.

1.3.2 Representative Cohort for Surveillance Studies

Whilst studies have been conducted on the individual species, sparse literature exists to examine overlaps and distinctions between each cohort. Due to their size, pigs are easily traceable through extensive pig networks around the UK[9]. Conversely, because of the procurement of poultry within backyard operations, the intelligence surrounding poultry keepers is severely hindered. By examining two dichotomous cohorts of smallholdings in regards to the current database, we can build a knowledge base which tests the overlap between existing, verified information in the pig/poultry cohorts through lab reports and governmental records, with the newly tested surveillance mechanism of passive listening through social media[10].

Biosecurity challenges in small-scale pig and poultry farming are unique. Disease outbreaks are more common in smaller farms as a result of lax preventive measures. Therefore, surveillance is essential in these settings to protect animal health, ensure food security, and prevent zoonotic diseases that can impact humans[11].

1.3.3 Implications for Broader Small-Scale Communities

While this research centres on pig and poultry smallholders, the methodologies and findings apply and are scalable to the wider smallholder community. The surveillance tools and analytical techniques developed can be adapted to other smallholder groups, which is pivotal for creating comprehensive animal health surveillance strategies.

Expanding this research to include all members of the livestock community is justified by the diverse range of farming practices and species found within the sector. By broadening the focus, a more comprehensive understanding of the landscape can be achieved, ensuring that surveillance systems are strong, inclusive, and adaptable to the unique characteristics of various smallholding operations. However, the question whether a project such as this can realistically be extrapolated to fit the entire community based on social media activity remains to be verified. This will be addressed in more detail in the discussion chapter.

1.3.4 Rationale Summary

Overall, analysing contrasting cohorts of pig and poultry smallholders enriches the research by offering a comparative perspective. It highlights the need for comprehensive and diverse studies that encompass different agricultural practices and their varying levels of information. This decision supports the study's goals of deepening our comprehension of these farmers, enhancing disease monitoring, and advising policy. It acknowledges the significance of small-scale agriculture in the wider economic and food security context and aims to amplify the voices and experiences of those typically overlooked in academic research.

1.3 Research Impact

The results of this work will have an impact on disease surveillance in the UK smallholding and broader livestock industries, particularly in relation to what can be achieved with supplementary social media data. It demonstrates the benefits of data science approaches to this type of system and could therefore lead to projects in different production sectors.

1.4 Project Structure

The structure of this thesis is illustrated below in figure 2:

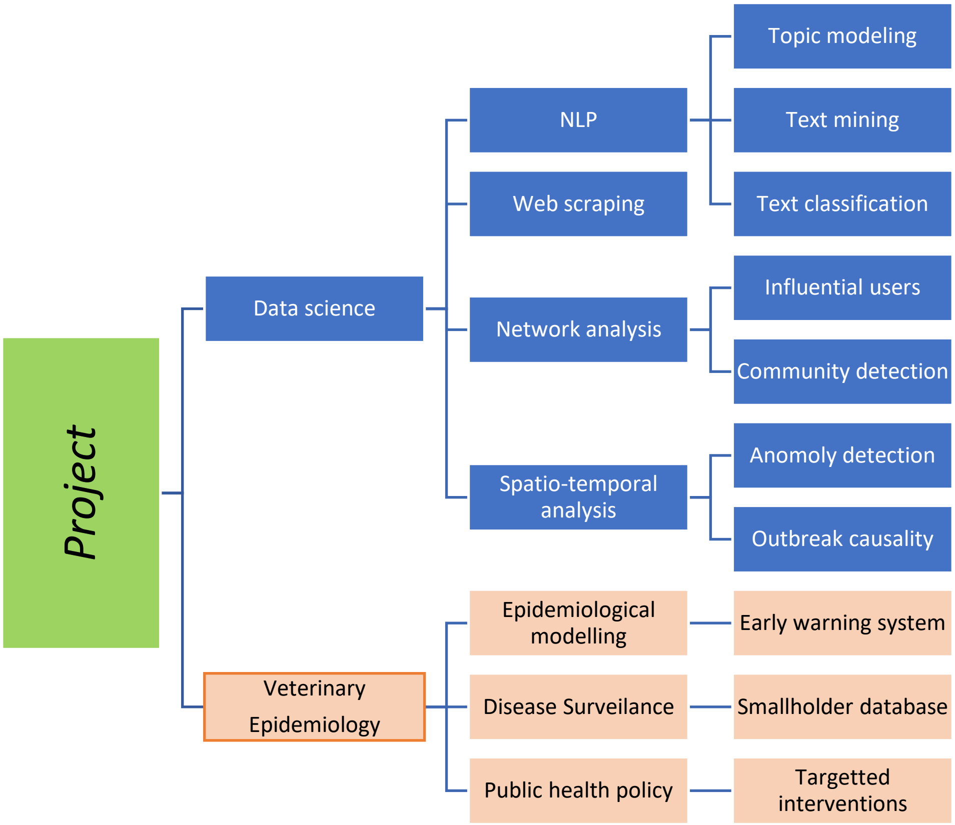


FIGURE 2: INTERDISCIPLINARY APPROACH OF PHD

1.5 Thesis structure

Table 1 below summarises the content of this report, with each main chapter possessing a summary box to serve as a prelude:

Section	Description
Introduction & Background	Introduces the current state of disease surveillance systems in the UK, and presents the aims, objectives, and structure of this project.
Literature review	Presents a synthesis of the current and previous literature in the fields of public health surveillance, epidemiological surveillance, social media and data science.
Research methodology	Describes the data extraction procedure, analytical methodologies, justifications and ethical considerations.
Experimental Chapter 1	Presents results from user classification.
Experimental Chapter 2	Presents results from network analysis.
Experimental Chapter 3	Presents results from forum scraping.
Experimental Chapter 4	Presents results from Spatio-temporal analysis.
Discussion	Offers a synthesis of the results from all four experimental chapters to provide recommendations around the information required for surveillance supplementation. Discusses the research impact of the work from the perspectives of academia, and how to implement this on a national scale alongside current surveillance methods.
Conclusion	Concludes this report with remarks around the essential findings and contributions of this work.

TABLE 1: THESIS STRUCTURE

1.6 Background

Over the past two decades, we have witnessed substantial paradigm shifts in epidemiological disease surveillance due to the rapid expansion of technological enterprise, in the form of the internet and social media[12]. The public health domain has been able to reap the benefits of increased quantities of health research derived from the internet boom, by subsequently creating progressively innovative epidemiological models[13]. We have witnessed this in contemporary research such as tracking COVID-19 mandated vaccination sentiments and opinions, and geolocation mapping of behaviour adherence during lockdown[14].

Moreover, the internet has been utilised to assist in the determination of disease outbreaks, monitor the proliferation of infectious disease, and to appraise outbreaks in the population[15]. This has proven to be very effective in the monitoring of breakouts of sexually transmitted diseases, tuberculosis and hepatitis in the USA[15]. The methods in which communication tools are used in traditional public health surveillance have also been strengthened with the aid of internet surveys and social media polls, proving to be more time and cost-efficient than the conventional approaches of telephone/mail/face-to-face surveys[16].

"Social media" pertains to various sources of internet engagement platforms such as Twitter, Facebook, Instagram, Quora and other specific forums for public engagement[12]. These platforms can all be utilised to facilitate the collection of passive data for analysis through web scraping and Application Programming Interfaces (API) calling.

The rise of different avenues of social media in the past decade has allowed public health researchers to enhance their epidemiological surveillance (ES) databases and has resulted in researchers obtaining the ability to mine previously unavailable information[17]. This includes current trending topics, information dispersion, user engagement, social demographics, community characteristics and influential users/communities, all from these social media posts[13]. A multitude of social media platforms are available free of charge to the public and offer users the ability to communicate and network with members on a global scale, thus the transfer of news and information is expeditated. This has opened up tremendous opportunities in both the epidemiological and public health fields, wherein the application of data mining algorithms can be applied to aid and track potential disease outbreaks within both the human and animal domain. The willingness to incorporate such dynamic big data has positively augmented the field of ES and outbreak modelling, as readily available public-data can be harvested, with potential insights being drawn and included alongside the standard routinely collected data. Relevant examples of this include google trends data for influenza-symptom detection[18], Ebola hotspot recognition[19], time series COVID-19 pandemic monitoring [20] and Avian Influenza risk surveillance [21].

Traditional methods of data utilisation for the purposes of ES have revolved around standard routine data sources[22]. Data collection falls within two categories; primary and secondary, with the former pertaining to first-hand data collection for a specific study, and the latter relating to already-available data which can be implemented for a study[23]. Primary data (e.g. interviews) is usually time-demanding and arduous in nature; hence researchers have begun possessing an inclination to also incorporate secondary data (e.g. health/veterinary records) for efficiency and optimisation in study completion[23].

Social media data falls under the rubric of secondary data, however over the past few years has raised questions around its efficacy due to quality and accuracy issues[17]. This highlights the dire need for greater advancements within this field of health informatics, or Infodemiology; a phrase coined by eHealth researcher Dr Gunther Eysenbach in 2004[24]. Eysenbach also devised the term Infoveillance; the adoption of internet search and social media data to assist with epidemiological surveillance, with the ability of refining parameters to particular regions to identify hotspots.

One of the key research benefits associated with social media datasets is the available characteristics of online communities and how these users interact amongst each other[25]. This provides access of data into otherwise private groups who previously only used their own communities to share ideas, advice and information. We can also observe clusters of communities being formed from user engagement and can deploy social network analysis (SNA) to comprehend how this information is spread[26]. Those users within clusters often have demographic metadata associated with their account, therefore can be modelled at a population level through techniques such as clustering. With this knowledge, we can delve into a branch of mathematics known as network analysis, whereby the relationships between individual users and their corresponding interactions can be quantified via a network diagram[27].

Obtaining the ability to quantify the diffusion of social media information at a population level, enables public health researchers and epidemiologists to determine if the information they relay is firstly reaching the right groups (those at risk), and secondly, how these users utilise this information (e.g. share with others)[28]. Moreover, influential users can be detected within communities, based on the number of shared posts, likes and followers one has, and the interaction amongst these smaller communities may consist of potentially useful insights[29].

The COVID-19 outbreak in late 2019 highlights the dire need for dynamic disease surveillance and the usage of internet data to aid in this process[30]. A Canadian based Artificial intelligence firm named BlueDot, were amongst the first researchers to identify the possible outbreak of Covid-19 utilising social media and Google trends data; thus, reinforcing the potential benefit to both the human and veterinary epidemiological domain[31]. Results from their analysis described that social media has the ability to make people aware of symptoms and subsequent control measures, whilst simultaneously encouraging correct public health behaviour (i.e. face masks, social distancing).

Particularly, the COVID-19 pandemic, triggered by a novel coronavirus (SARS-CoV-2), has produced an unparalleled sheer quantity of online searches, through Google searches and social media outlets[20]. This might indicate the occurrence of genuine infections amongst those who actively searched, nevertheless it could be a by-product of the influx of news generated on social media; hence, emphasising the need for validation with actual cases when examining correlations between search trends and "ground truth"[32]. We have witnessed numerous recent studies incorporating Infoveillance techniques to this pandemic to track various topics such as public opinion, compliance, attitudes towards authority and agreement with scientific consensus[20].

Only recently have public health researchers implemented data science techniques in the mode of forecasting disease progression and identifying risk of infection[33]. Text mining algorithms can be employed over patient health records, and computer vision algorithms deployed on medical images to predict signs and symptoms of disease[34]. The success of this has opened up enormous opportunities to transfer such methods into the veterinary field; with larger commercial farms benefiting from "future farming systems", wherein image recognition is used on both livestock and crops for optimisation, early-detection of disease and preventative procedures[35]. The team at SRUC's Precision Livestock Farming department is dedicated to helping farmers enhance efficiency and decrease the carbon impact of their systems through data-driven solutions and decision support platforms[36].

The following examples in figure 3 are some of the solutions offered and reinforces the potential of incorporating data science tools within the livestock health:

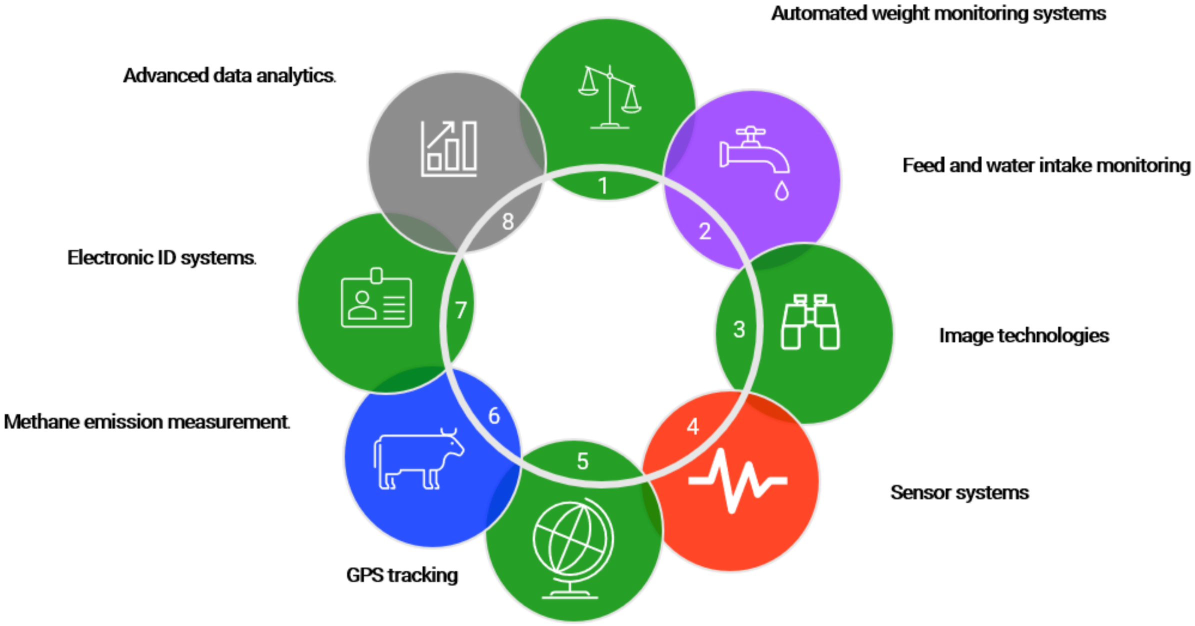


FIGURE 3: DATA SCIENCE TOOLS IN LIVESTOCK

The cost of specialised equipment and state of the art analytics largely omits the smallholder communities from partaking in such endeavours, hence the opportunity to apply some of these data science methods via the medium of the internet is available and has great potential to be valuable[30]. Additionally, these communities tend to be more closed off and private in their affairs, adhering to a lifestyle of self-sustenance. Therefore direct surveillance may not be the most realistic or effective tool for this population, but rather a passive form of data gathering through social media surveillance may produce the best results in understanding them.

From the perspective of the pig industry, the demand for robust veterinary epidemiology surveillance methods via the utilisation of innovative predictive analytics has been strengthened by the various swine flu (H1N1) outbreaks since 2009[37]. This culminated in a global economic cost of over half a trillion dollars as a consequence of deceased or infected pigs, in addition to the disproportionate amount of human deaths witnessed in Africa and Southeast Asia[38]. Whilst knowledge pertaining to pig owners is more widely known to the government than poultry, biosecurity issues within poultry livestock movements and disposal is still vital, and this information needs to be relayed to the right cohorts of smallholders[1].

1.6.1 Overview of UK smallholdings

Within the UK, a smallholding is generally classified as an area of land adjoining living premises, with the size of it in the region of fifty acres and is commonly recognised as a dwelling for raising/breeding livestock organically on open land[1]. In relation to owning poultry livestock, APHA doesn't impose mandatory government requirements for smallholders to register their flock if they own less than fifty animals, ranging from any combination of the poultry species (i.e. chicken, duck, geese, turkey etc). The registration of small flocks of poultry to a governmental agency is voluntary; hence surveillance of these smallholders is very difficult which leads to problems with biosecurity. Moreover, the relative lack of engagement with local authorities compounds this issue and adds to the complexities faced by authorities in public health surveillance[3].

The popularity of a smallholding lifestyle is ever-increasing within the UK and has proven to be extremely common throughout many third world agriculturally-dependant countries[39]. There are approximations surrounding the quantity of smallholder farms globally, with figures over half a billion individual farms, providing sustenance for nearly two billion citizens[40]. It has been implied that the increase in health awareness of society has played a major role in this increase in demand, as households strive for better food quality, animal welfare and sustainability; all of which can be achieved through this lifestyle. Smallholding farming in the UK is a type of agriculture that involves the cultivation of small plots of land, by individual farmers or families[41]. The main aim of this type of farming is to produce food for personal consumption and to generate a small income from the sale of surplus produce. Furthermore, it is becoming increasingly popular in the UK as people look for ways to live more sustainably, grow their own food and reduce their dependence on the commercial food system.

This mode of farming requires a lot of hard work and dedication, as farmers need to be able to manage their land, livestock and crops effectively to ensure the success of their venture. Often involving the undertaking of tasks such as ploughing, planting, harvesting, weeding, fencing and maintaining animal housing. Many individuals also choose to sell their produce at local markets or through farm-to-table initiatives, which helps to connect them with their local communities and provide a steady income stream[42]. In terms of crops, they typically grow a range of vegetables, fruit, herbs, and flowers and many also keep livestock such as chickens, pigs, sheep, and cattle simultaneously. Some farmers also produce honey and other value-added products such as jams, pickles, and cider. In addition to food production, many also choose to incorporate other activities into their business, such as agritourism, where visitors can come to the farm to experience life on a smallholding and purchase produce[43].

One of the main benefits accrued through this lifestyle is that it can provide a more sustainable, self-sufficient way of life, where farmers have control over the quality and origin of the food they consume. It can also provide a sense of community and connection to the land, as well as a source of income. However, it can also be challenging, as individuals must be able to manage their resources effectively and overcome a range of obstacles, such as weather conditions, pests and diseases, and the need for access to markets[41].

In terms of government support, there are several initiatives in place to help smallholding farmers in the UK, including grants for land improvement, support for agritourism, and training programmes[44]. However, there are also challenges in regard to the access to markets, as many struggle to compete with larger commercial farms in terms of scale and efficiency. This can make it difficult to sell their produce at a fair price and generate a stable income.

Social media is one method they can utilise to connect with customers, promote their products, and share their experiences and knowledge with others[45]. Platforms like Facebook, Instagram, and Twitter can provide them with a wider audience, helping them to reach new customers and build their brand. By sharing photos and updates about their farm, farmers can give customers a behind-the-scenes look at what goes into producing the food they sell, and engaging relationships can be built with their customers through engagement, thus increasing brand loyalty and helping to ensure a steady demand for their products. In addition to connecting with customers, social media can be employed to network with other farmers and industry professionals. Furthermore, through the process of sharing best practices and lessons learned, farmers can help one another to improve their operations and overcome the common challenge mentioned previously[41].

The pig industry in Scotland alone consists of a population well over 300,000[46], with 10% of these being reared for breeding purposes[1]. Furthermore, this sector accounts for approximately 3% of the Scottish Agricultural Output, with an economic output of £85 Million[47]. The threat of disease outbreaks stemming from smaller-scale pig premises is often overlooked. A holistic approach of disease modelling which factors in the importance of sound biosecurity measures for backyard pig keepers is essential to effectively quell outbreaks.

Within the UK, almost 70% of pig producers fall into the category of smallholders, owning less than ten pigs[48]. Given this statistic, pig networks and subsequent biosecurity is severely ambiguous as the sparse, unaccounted for nature of these holdings is often under-reported. Regulating bodies consider small-scale produces to be high risk in relation to disease spread, as the owners may be unaware of the latest biosecurity measures needed, in addition to less extensive regulation as compared to larger commercial farms[1]. The gap in literature stems from the lack of research regarding the extent of knowledge the smallholders receive and if they choose to adopt such measures for public health purposes.

Linkages have been made between these backyard keepers and the role which they play in avian associated outbreaks, both endemic (Salmonella) and exotic (Highly pathogenic avian influenza) alike[4]. In recent years, there has been a resurgence of interest in small-scale poultry farming as a way to produce food locally and sustainably[49].
The migratory nature of wild birds vastly increases the probability of disease within poultry and fowl livestock, thus sound biosecurity advice is regularly disseminated by governmental agencies such as The Animal and Plant Health Agency (APHA) to quell the spread of disease[50]. They play a significant role in the spread and control of disease outbreaks in the livestock industry[1] and can contribute to the spread of diseases by failing to follow proper biosecurity measures, such as not properly cleaning and disinfecting equipment and facilities, or by not properly controlling the movement of animals in and out of their farms[5]. Furthermore, they can also play a critical role in controlling disease outbreaks by reporting suspected cases of illness to the relevant authorities, following recommended quarantine and treatment procedures, and by participating in disease control programs. Preventative measures can be adopted to reduce the risk of disease, such as providing proper nutrition, housing, and ventilation for their animals, as well as implementing good hygiene practices[51].

As the requirement to register small flocks with governmental agencies is unnecessary, understanding the communities and networks of such a demographic becomes difficult from a public health surveillance perspective[52]. Researchers have adopted primary data collection methods through questionnaires and surveys, delivered in-person and via post[6]. Questionnaires were designed to capture both demographic and biosecurity related information from the target population, with distinct segments covering animal husbandry, location, movement, health and enterprise.

In the case of large-scale disease outbreaks, the cooperation of smallholders is crucial in controlling the spread of the disease and protecting public health[53]. They must be informed and educated about the dangers of disease and the importance of proper biosecurity measures, and they must be provided with the resources and support they need to implement these measures effectively[54]. By implementing proper biosecurity measures and participating in disease control programs, smallholders can help to reduce the risk of disease and protect public health. Despite these challenges, this lifestyle continues to grow in popularity, and many believe it has an important role to play in the future of UK agriculture[42].

1.6.2 The multifaceted landscape of Smallholders

The UK's agricultural community is diverse, and smallholders represent an essential segment, which includes various distinct types[55]. The multifaceted nature of this population contributes to the UK's agricultural fabric, ensuring food security, supporting the rural economy, promoting biodiversity, and preserving cultural heritage. Without a nuanced understanding of their distinctive characteristics and roles, policy and support development may fall short. Figure 4 provides an overview of the diversity of this cohort:

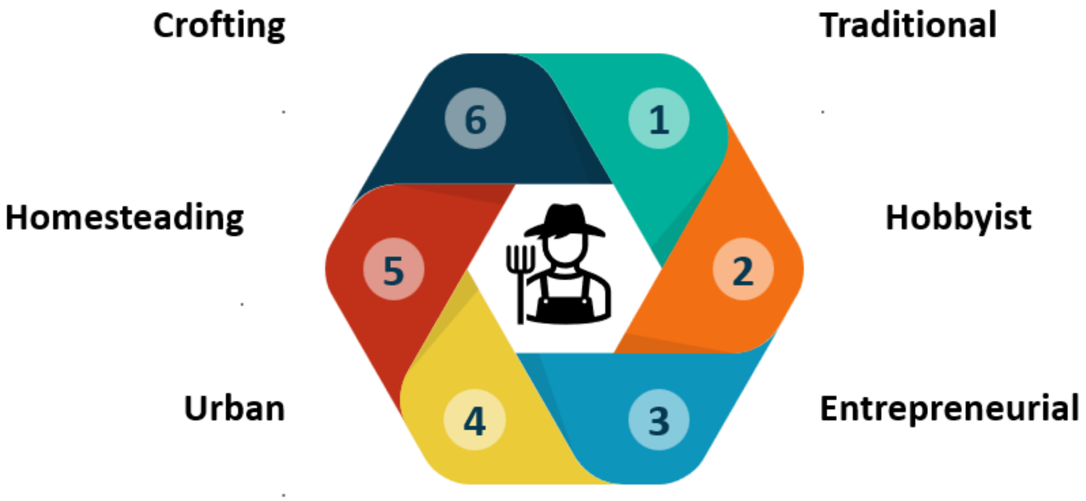


FIGURE 4: TYPES OF SMALLHOLDINGS

1/. Traditional smallholders' preservation of historical methods and local breeds testifies to the heritage of farming[56]. Generally, they operate on a small scale, prioritising quality and tradition over mass production through mixed farming. Their commitment to sustainable agriculture is evident in their emphasis on soil health, biodiversity, and community well-being. In several regions, traditional smallholders act as guardians of agricultural history, keeping valuable skills and knowledge alive for future generations.

2/. Farming for personal fulfilment and enjoyment is the primary motivation of hobbyists, rather than making a profit[57]. Focusing on specific hobbies like gardening, raising poultry, or beekeeping. Even though financial gain isn't the primary objective, they still play a part in local food systems and sometimes create a stronger bond with nature and community. Their dedication to agriculture improves local diversity and creates educational and engaging opportunities.

3/. Specialised or niche markets are targeted by entrepreneurial or commercial smallholders operating on a smaller scale, with the aim of profitability. They might concentrate on high-value crops, artisanal goods, or selling directly to customers through farmers' markets or community-supported agriculture programs. Although they are small, commercial smallholders are inclined towards innovative practices, big data technologies, and marketing strategies[58]. They may be more inclined to use social media as a marketing tool, utilising platforms such as Instagram and TikTok to showcase their inventory and target the younger demographics of society.

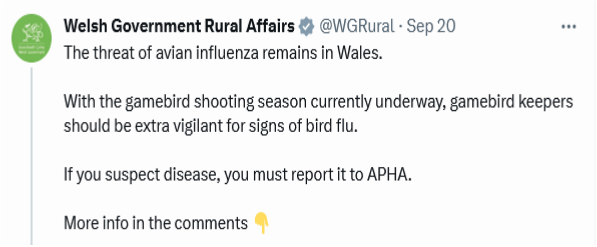
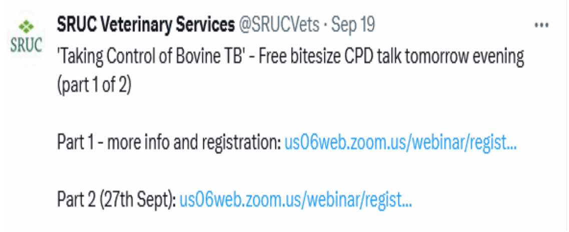
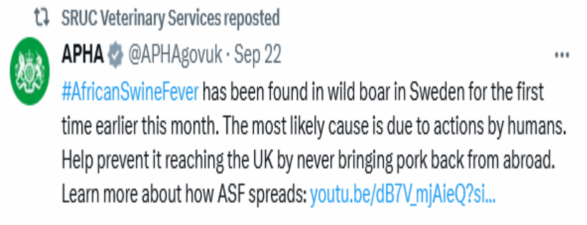
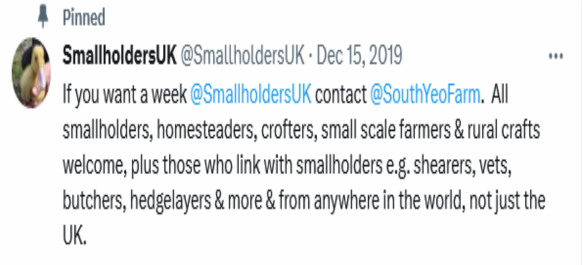
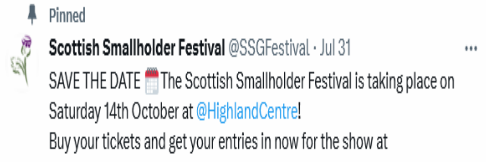
4/. City spaces like rooftops, balconies, and community gardens are used by urban smallholders for growing crops[59]. The main focus is often on cultivating vegetables, herbs, and small-scale livestock to provide fresh and local produce in city environments. Food security, environmental sustainability, and community engagement are positively impacted by these individuals. By following their practices, urban residents are encouraged to connect with food sources and become more aware of sustainable living and local food systems.

5/. Small-scale farmers who practice homesteading seek self-sufficiency through farming, crafting, and household management[60]. They usually cultivate crops, raise livestock, and create items for their personal use, frequently using traditional techniques like preserving food, carpentry, and stitching. Homesteading is more than just growing crops; it is a philosophy that emphasises self-sufficiency, sustainability, and a connection to nature. The approach fosters a holistic way of living by integrating food production, family, and community. Online communities, such as Reddit's homesteading subreddit, https://www.reddit.com/r/homestead/, provides opportunities for these individuals to communicate with like-minded individuals.

6/. Crofters are unique farmers that can mainly be found in the Scottish Highlands and Islands[61]. It has roots in community and familial traditions, involves subsistence farming on small parcels of land known as crofts. These individuals engage in an array of activities such as raising livestock, planting crops, and fishing or performing other trades to increase their income[62]. The connection to communal grazing lands is a defining characteristic of crofting, which encourages crofters to work together and share responsibilities. It is governed by specific legislation and local regulations, is an important tool for preserving cultural heritage, promoting environmental stewardship, and supporting rural communities. They are custodians of a rich and complex socio-cultural landscape that plays a vital role in Scotland's rural identity.

1.6.3 Avenues of information available to smallholders

Information and advice about livestock are often sought in several ways, including face-to-face networking with other smallholders, attending exhibitions (e.g. The annual Scottish Smallholder Festival), referring to veterinary guidance and online resources. The table below summarises the various options available:

Information source	Description	Evidence
Veterinarians and Livestock Advisors	Veterinarians and livestock consultants can offer specialist information and guidance on animal health and well-being. These experts can be contacted for particular queries or continuous assistance.	 Welsh Government Rural Affairs @WGRural · Sep 20 The threat of avian influenza remains in Wales. With the gamebird shooting season currently underway, gamebird keepers should be extra vigilant for signs of bird flu. If you suspect disease, you must report it to APHA. More info in the comments 👇
Agricultural Colleges and Extension Services	Training and guidance on best practises for livestock management, such as feeding, housing, and disease control, may be obtained through agricultural colleges (i.e. SRUC) and extension agencies (SAC consulting).	 SRUC Veterinary Services @SRUCVets · Sep 19 'Taking Control of Bovine TB' - Free bitesize CPD talk tomorrow evening (part 1 of 2) Part 1 - more info and registration: us06web.zoom.us/webinar/regist... Part 2 (27th Sept): us06web.zoom.us/webinar/regist...
Online Resources	Social media, government websites, forums, and discussion groups are among the many internet tools accessible. These resources may provide information about animal husbandry, nutrition, and health.	 SRUC Veterinary Services reposted APHA @APHAgovuk · Sep 22 #AfricanSwineFever has been found in wild boar in Sweden for the first time earlier this month. The most likely cause is due to actions by humans. Help prevent it reaching the UK by never bringing pork back from abroad. Learn more about how ASF spreads: youtu.be/dB7V_mjAieQ?si...
Networking with other Smallholders	Networking may be an excellent source of knowledge and assistance. Many smallholders join local groups or clubs to discuss their experiences, ask questions, and get assistance from more experienced farmers. As this is usually done in a face-to-face setting, more traditional smallholders may prefer this route.	 Pinned SmallholdersUK @SmallholdersUK · Dec 15, 2019 If you want a week @SmallholdersUK contact @SouthYeoFarm. All smallholders, homesteaders, crofters, small scale farmers & rural crafts welcome, plus those who link with smallholders e.g. shearers, vets, butchers, hedgelayers & more & from anywhere in the world, not just the UK.
Livestock Shows and Exhibitions	Fairs and exhibits may be another avenue to learn about different breeds, observe animals in action, and chat with industry professionals.	 Pinned Scottish Smallholder Festival @SSGFestival · Jul 31 SAVE THE DATE 📅 The Scottish Smallholder Festival is taking place on Saturday 14th October at @HighlandCentre! Buy your tickets and get your entries in now for the show at

Books and Guides	There are several publications and guidelines on livestock management available, covering a wide variety of issues such as feeding, housing, and disease control. These sites can give detailed information and can be accessed as required. Local libraries are a valuable resource to provide such information, particularly in rural areas.	COUNTRY Smallholding BRITAIN'S BIGGEST-SELLING SMALLHOLDING MAGAZINE INSIDE Your FREE Poultry magazine £3.99 MARCH 2018 LOVE YOUR SOIL The key to a successful veg patch! BREEDING SUCCESS Pregnancy and birth YOUR GOOD LIVES Smallholder profiles

TABLE 2: SMALLHOLDER INFORMATION SOURCES

In conclusion, they have access to a range of resources for information and advice on livestock, including veterinarians and livestock advisors, agricultural colleges (i.e. SRUC), online resources (i.e. social media), networking with other similar individuals, livestock shows, exhibitions (i.e. Scottish Smallholder Festival), books and guides. By utilising these resources, they can gain the necessary knowledge and support they need to manage their livestock effectively[63].

1.6.4 Current surveillance systems in the UK

Livestock surveillance systems play a crucial role at the local, national, and international levels in detecting and preventing biological and chemical risks such as animal diseases, zoonotic diseases or syndromes, toxins, or contaminants to secure public health and food safety[16]. These systems aim to assist public health needs, including the creation of early warning systems for new, exotic, and resurging diseases, efficient disease control, and monitoring of disease patterns both spatially and temporally. Surveillance data serves as a foundation for international trade regulations and is crucial for creating contingency plans to protect human and animal health and rural economies from the consequences of widespread disease outbreaks, and for mitigating the impact of animal disease and climate change on the environment[23].

Animal health monitoring in the UK is primarily performed by farmers submitting specimens to any of the eight regional Disease Surveillance Centres for diagnosis and post-mortem examination[2]. This is supported by disease reporting at slaughterhouses, including required reporting of notifiable illnesses, zoonoses, and passive monitoring of wildlife diseases, as well as active surveillance for particular pathogens or diseases, such as Trichinella spiralis and Bovine Viral Diarrhoea[64]. Furthermore, industry-led programs, such as pig assurance schemes, offer disease information to farmers, but this information is not usually combined with other surveillance systems. As a result, the monitoring system involves multiple separate and uncoordinated systems, implemented by different entities[65].

The current state of animal health surveillance has come under heavy criticism recently. It is funded both by the UK Government and through fees charged to farmers for diagnostic services[16]. A thorough review conducted in 2011 found that the existing veterinary surveillance system would require extra resources, which were deemed unlikely to be obtained, to continue operating in its current form. The review also identified the potential to enhance efficiency in disease surveillance. Since the review, stakeholders have been involved to increase transparency and accountability and to ensure the veterinary surveillance system's goals align with the needs of its users[66]. The process is visualised below in figure 5:

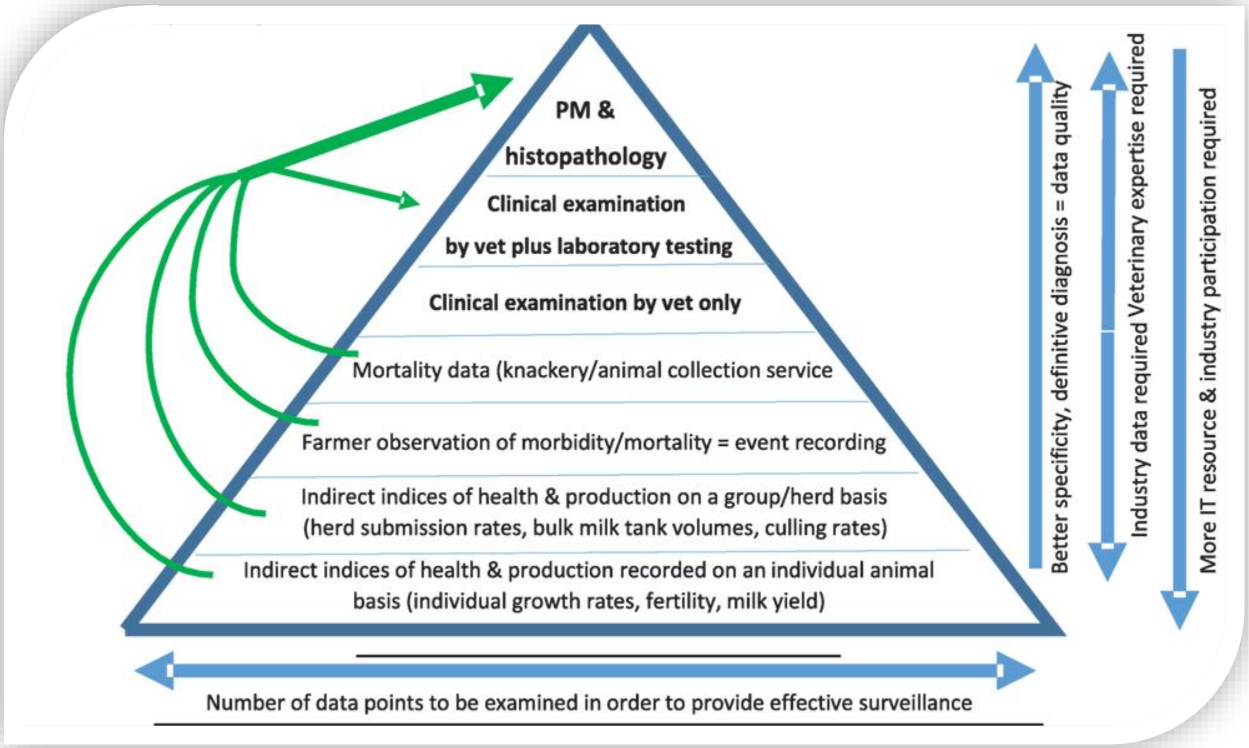

PM &
histopathology
Clinical examination
by vet plus laboratory testing
Clinical examination by vet only
Mortality data (knackery/animal collection service
Farmer observation of morbidity/mortality = event recording
Indirect indices of health & production on a group/herd basis
(herd submission rates, bulk milk tank volumes, culling rates)
Indirect indices of health & production recorded on an individual animal
basis (individual growth rates, fertility, milk yield)
Better specificity, definitive diagnosis = data quality
Industry data required Veterinary expertise required
More IT resource & industry participation required
Number of data points to be examined in order to provide effective surveillance

FIGURE 5: SURVEILLANCE PYRAMID

Due to the economic issues surrounding the allocation of both human and financial resources to bolster surveillance initiatives, the government is faced with tough decisions regarding how to prioritise potential threats. Existing and contemporary data collection, analytical methodologies and insight derivation structures underlie these outcomes[67]. To establish disease-free status, sources of evidence from both economics and epidemiology should be consulted, and existing surveillance methods should be evaluated. Both quantitative and qualitative epidemiological approaches are frequently influenced by the investigators' perspectives of the system at risk and the belief that past events will predict the future[16]. Figure 6 below illustrates the multi-faceted structure of public health surveillance, with the interconnectedness of various data nodes constructing a holistic model of intelligence. Highlighted in red is the newest addition to this model and is the focus of this thesis, with the results of this project strengthening the efficacy of such data sources in future applications.

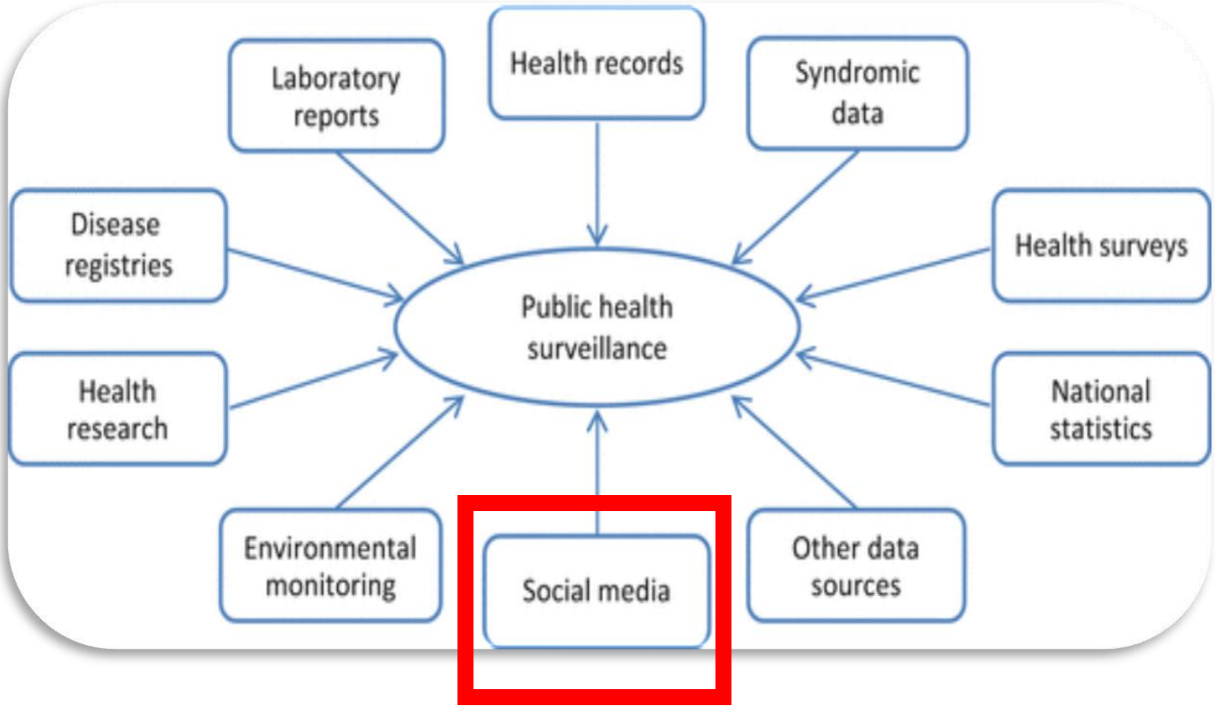


FIGURE 6: DATA SOURCE SPIDER DIAGRAM

1.6.5 Shifting paradigms of surveillance

The dawn of the 21st century has seen continued augmentations in this field, with the employment of Machine Learning (ML) and health informatics techniques for the analyses of large data collected through automated systems[68]. This evolution within the agricultural industry has also impacted the physical tools utilised in data and image capture (e.g. drones, sensors), thus the application of big data techniques is necessary to make sense of such voluminous and dynamic data (which included numbers, text and images).

The different envisioned futures in disease surveillance involve varying levels of uncertainty in data collection and analysis. As the current trend continues, farmers may adopt precision agriculture technology, but this could lead to fewer submissions from farmers and a decrease in veterinary resources for generating conventional surveillance data such as clinical samples[69]. As real-time animal, plant, and environmental health data are generated and collected through new technologies, the insights gained from scenarios such as state-led, export-led, and industry-led surveillance become more relevant. In a future data economy, control over data sharing, including secure transfer, management, storage, and accessibility, will be as crucial as technological advancements in smart systems to ensure standardised and integrated data that can be transformed into useful knowledge. The free trade of both animal health data and animal products may play a key role in driving economic growth in the future[16].

On the other hand, if technology-driven alternatives to traditional data collection are not widely adopted, other more extreme scenarios become more likely (such as individual-led surveillance). Both high and low volume, as well as high- and low-quality data collection, have the potential to create risks and widen inequalities between those who have access to data and those who do not. This, combined with a lack of strategy for making variable quality information publicly available or with political shifts towards increased state control over data and services, could lead to a loss of trust in expert opinions and a decrease in stakeholder belief in the value of investing in a scientific evidence base for policy making[23].

To address the uncertainty of the impact of novel and emerging diseases, participants in the "Industry-led surveillance future" suggested the introduction of legislation for mandatory reporting of "health risk states." A Health Risk States Scheme is currently used in human health in the UK to flag potential threats to public health at an early stage, even if the causative agent is not known[70]. This system could benefit from co-localisation and resource sharing between veterinary and human health laboratories. The proposed strategy would ensure that early warning signs of potential concern are shared, addressing the possible conflicts between private companies and veterinarians over reporting. There is currently no comparable system in animal health in the UK or in international animal disease reporting, where notifications are based on suspicion or confirmation of specific pathogens. Early detection systems, such as the Programme for Monitoring Emerging Diseases, encourage voluntary reporting of similar non-specific information[71].

1.6.6 Social media as a surveillance tool

The monitoring of diseases has changed considerably over the years and is anticipated to do so going forward. This evolution is the result of technological developments, particularly easier internet access and more powerful computers, which allowed for the development of digital disease surveillance systems[72]. The term "digital disease surveillance" refers to the creation or application of systems for forecasting the incidence or prevalence of diseases in the present or the future using internet-based data[73]. By providing real-time data and trends for health outcomes, the use of social media and internet-based search engines has also created new opportunities for extending disease surveillance. This information is frequently added to outpatient, hospital, and lab-based surveillance systems. Traditional methods rely on people seeking medical attention, which might result in an miscalculation of the overall disease burden, as witnessed with the overestimation of Google Flu trends models[18].

On the other hand, certain forms of digital surveillance rely only on digital data without necessarily completing established systems for keeping an eye on people's health. For instance, it has been demonstrated that isolated Salmonella enterica outbreaks correspond with searches for terms like "diarrhoea" and "food poisoning" in Google[74]. Although the relationship with conventional health surveillance is not entirely evident, this type of data may nonetheless offer useful health information. Examples include gathering information on public opinion towards immunisation and health practises like tracking diet and quitting smoking[75]. Due to easier access to the internet and technological improvements, there is now more public health information being monitored electronically.

By gathering real-time data and trends, social media and internet-based search engines have created new potential for surveillance as digital data can be added to conventional public health monitoring systems, such as those that rely on hospital and laboratory-based data, improving the representativeness of the overall illness burden[15]. The use of search query data, online restaurant reservation logs, and social media are just a few examples of digital surveillance systems that are distinct from traditional surveillance systems. With tools like Google Flu Trends, Influenzanet, and FluNearYou, digital public health monitoring is most commonly employed in the field of influenza surveillance[76].

These systems have promise, but there are also uncertainties around accuracy, privacy, and collaborations between tech firms and the surveillance state. People may actively or passively choose to disclose their health-related information on social media[77]. As active participation (such as tweeting about symptoms) is likely to be a more accurate picture of actual sickness than passive engagement, it can have an impact on the accuracy and specificity of surveillance data collected via digital sources (e.g. visiting a Wikipedia page on influenza). Because some people are more likely to actively post their health-related information on social media than others, the demographic that these tactics can reach may also differ and the potential for sampling bias could occur.

Numerous studies have shown that in addition to collecting and analysing data, social media has also been used as a platform for conducting experiments and observational studies[78]. For example, researchers might create fake profiles on Facebook and manipulate the content that is shared in order to study how people respond to different types of information. Furthermore, they might also observe real people's interactions on social media platforms to study the dynamics of online communities and the ways in which people communicate and share information.

While social media has many potential benefits as a research tool, it is important to note that there are also limitations and challenges associated with this approach. Researchers have purported that one of the main challenges is the issue of data accuracy and reliability, as people may not always be truthful or accurate when posting online[78]. Additionally, these platforms can be biased in terms of who is using them and what information is being shared, which can lead to incorrect or misleading conclusions if not considered.

Despite these limitations, it has proven to be a valuable resource for investigators looking to understand more about people, their thoughts, and behaviours. As it continues to evolve and become an increasingly important part of our lives, it is likely that it will continue to play a crucial role in shaping our understanding of the world around us.

Although the use of machine learning algorithms can aid in making sense of the enormous amount of data from digital sources, issues with data quality and privacy concerns still need to be resolved. The use of machine learning in digital surveillance systems is a tool for processing massive amounts of data; however, the precision and efficacy of the predictions depend on a number of variables, including the quality and specificity of the data, the choice of algorithms and their parameters, and the capability to efficiently pre-process the data. Despite machine learning's triumphs in other areas, greater prediction in the context of digital monitoring is not a given[79].

Table 3 presents a synthesis of social media sources that often used for epidemiological research:

Social media source	Description
Twitter	Tracking the transmission of infectious illnesses and monitoring public health incidents since it delivers real-time information from individuals and organisations.
Facebook	Examining public opinion and attitude on health concerns, as well as gathering information on illness spread through online networks.
Instagram	Collecting visual data on health concerns such as disease transmission in different regions, in addition to tracking public opinion and attitudes.
Reddit	Users may publish and discuss health-related subjects. It can be used to obtain disease-spread information as well as monitor public opinion and attitude.
Forums/Blogs	Blogs may be employed to gather data and opinions from those with specialised expertise and experience in the subject, as well as to give in-depth information and analysis on health concerns.

TABLE 3: SOCIAL MEDIA PLATFORMS IN EPIDEMIOLOGICAL RESEARCH

It is important to note that while these sources can provide valuable information for epidemiological research, it is crucial to validate and verify the information before using it for research purposes. They aren't always reliable, and it is vital to consider the source of the information and its credibility prior to commencing analysis.

1.6.7 Improving smallholder surveillance

Improved monitoring on smallholders can help detect illnesses and outbreaks more effectively, and efficiently apply control measures[80]. Some techniques for improving surveillance are listed below in figure 7:

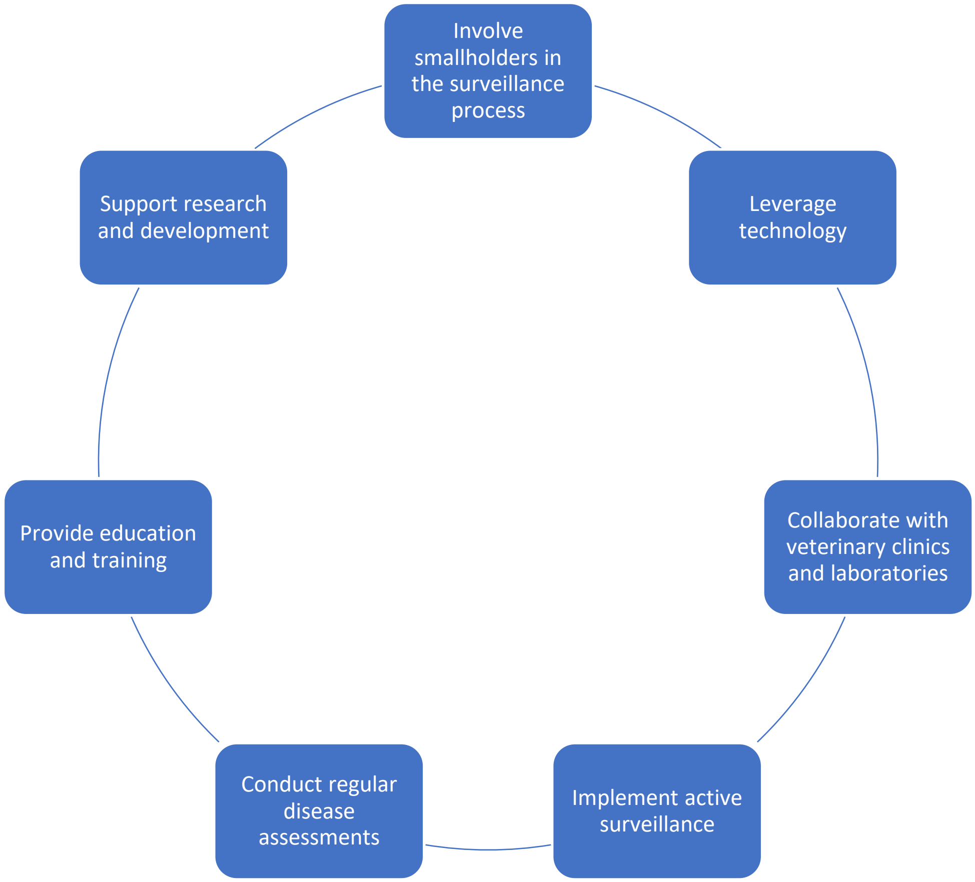

Involve smallholders in the surveillance process
Leverage technology
Collaborate with veterinary clinics and laboratories
Implement active surveillance
Conduct regular disease assessments
Provide education and training
Support research and development

FIGURE 7: SMALLHOLDER SURVEILLANCE IMPROVEMENT

Firstly, encouraging individuals to report symptoms and incidences of disease in their animals can assist to enhance the accuracy and timeliness of surveillance information. They can be educated to spot symptoms and report events via computerised reporting systems or social media platforms.

Secondly, utilising technology, such as GPS tracking and remote sensing, can give significant information on livestock movements and distribution, as well as environmental variables that may contribute to disease transmission. Furthermore, collaboration with veterinary clinics and laboratories can aid in improving the accuracy and completeness of surveillance data. These organisations can provide electronic reporting systems, diagnostic tools, and other resources to aid with surveillance activities.

Moreover, active surveillance, such as sentinel farms and environmental monitoring, can also aid in the detection of illnesses and epidemics by providing early warning of disease episodes and assist in the prioritisation of management actions.

In continuation, frequent disease evaluations can be of benefit through identifying areas of concern and assessing the efficacy of control strategies. Assessments of this kind can be based on information gathered from a variety of sources, such as social media, electronic reporting systems, GPS tracking, and environmental monitoring. Smallholders' capacity to participate to monitoring activities and protect their animals from illnesses can be improved by providing education and training on disease detection and reporting, as well as biosecurity measures.

Finally, funding R&D to improve surveillance technology and methodologies can significantly increase the usefulness of surveillance initiatives. This might involve the creation of new diagnostic equipment, the advancement of data analysis techniques, and the investigation of novel disease surveillance methodologies, as explored in this project.

By taking a multi-faceted approach and leveraging technology, education, and collaboration, it is possible to improve surveillance on smallholders and to enhance the ability to detect and control diseases and outbreaks in animals.

1.6.8 Advancing Surveillance Using Digital Epidemiology

Disease surveillance is crucial in veterinary epidemiology to detect, control, and mitigate outbreaks, safeguarding public health and food security. Conventional veterinary surveillance methods, which rely on periodic reporting and physical data collection, often miss real-time or emerging data[81]. This is especially the case in settings such as those of small-scale farmers, who are often not well-represented in official records. In order to fill these gaps, this research project makes use of non-traditional data sources such as social media and online searches to establish a surveillance system that is more dynamic and immediate. The goal is to use online behaviour and data patterns to identify early signs of disease outbreaks, essential for quick response strategies.

1.6.8.1 Limitations of conventional surveillance systems

Traditional surveillance frameworks form the basis of veterinary epidemiology. These methods have built-in delays in data collection, transmission, and processing, which can impede the prompt detection and response to disease outbreaks. The foot-and-mouth disease outbreak in the UK in 2001 demonstrated the substantial delays in disease detection resulting from depending on physical inspections and farmer reports[82]. Lack of immediate reporting and detection of signs led to the rapid spread of the outbreak.

Due to limited infrastructure, small-scale farmers are often excluded from surveillance networks[80]. Disease reporting and diagnosis can be delayed in regions with many smallholder farms, like rural areas of Scotland and Wales, due to limited access to veterinary services and the sparse distribution of farms. Sample collection and verification are made more challenging by the logistical difficulties of reaching remote locations such as the Shetlands.

Additionally, the accuracy and effectiveness of traditional methods are further complicated by under-reporting caused by factors such as stigma, regulatory or economic repercussions[83]. Reporting suspected disease cases can be a concern for farmers due to potential culling, economic losses, and regulatory sanctions. under-reporting was a significant problem during avian influenza outbreaks, as farmers hesitated to report poultry symptoms out of fear of business disruptions and disease stigma.

These challenges demonstrate the importance of implementing proactive, real-time surveillance approaches. By employing digital tools alongside traditional methods, the UK's veterinary surveillance system can enhance its ability to detect and respond promptly.

1.6.8.2 Integration of Digital Epidemiology

The integration of digital epidemiology findings with traditional surveillance data offers a more comprehensive approach to disease monitoring. By employing machine learning and natural language processing, this study examines vast datasets to accurately identify potential outbreaks through the detection of disease-related keywords and symptom descriptions, drawing from examples in literature examining the detection of outbreaks of bovine tuberculosis and avian influenza within the UK in the past[84], [85].

Methods like topic modelling aid in identifying discussion patterns regarding particular illnesses. During an avian influenza outbreak, topic modelling of social media posts uncovered growing worries about poultry symptoms[86]. This facilitated faster identification of infected areas and potential carriers, empowering authorities to effectively guide farmers on biosecurity measures. Another powerful tool is sentiment analysis, which assesses the urgency and severity of discussions about particular disease symptoms. An example of this was found during the COVID-19 pandemic, whereby researchers studied public sentiment on Twitter in relation to the incidence of foot-and-disease coinciding with the pandemic[87]. The findings illustrated that increases in foot-and-mouth disease incidence were closely tied to higher levels of public anxiety, highlighting the need for effective government communication strategies to mitigate such worries.

Furthermore, in the context of location analysis, mapping disease mentions using geospatial analysis is a valuable tool in identifying potential hotspots. Spatial mapping in the UK connected regions with suspected bovine tuberculosis to specific areas using discussions on "badger sightings" and "cattle skin lesions"[84]. The analysis provided valuable information on how the virus spreads and what factors increase the risk, resulting in a more focused response.

1.6.7.3 Barriers to implementation

Nevertheless, there are challenges associated with the use of digital data. Noise in social media data, like irrelevant or misleading information, needs to be filtered meticulously and aided with domain expertise to avoid false alarms[88]. Sometimes routine culling measures were mistaken for outbreaks, causing unnecessary panic among farmers and the public. Hence, the development of reliable filtering algorithms is crucial to differentiate accurate disease signals from misinformation.

Moreover, dealing with linguistic diversity is yet another major obstacle, considering the various languages and dialects spoken in different areas. Regional dialects and varying terminology for symptoms and diseases in the UK, like "Newcastle disease" or "fowl pest," can make the analysis more complex. NLP models need to be adaptable to identify these distinctions and correctly interpret the information.

When handling sensitive digital information, it is crucial to prioritise data privacy and ethical considerations[89]. Anonymisation and transparency should be the focus of data collection practices to ensure trust and compliance with data privacy regulations. The conversations around social media data scraping and public acceptability are still highly contentious and evolving and will be mentioned in greater detail in the methodology and the discussion chapters.

1.6.9 Health belief model

The Health Belief Model (HBM) is a psychological framework that uses an individual's beliefs and perceptions to predict health behaviours and may be used in the context of social media to study how people interact and respond to health-related information, campaigns, or interventions provided on these platforms[90].

Perceived vulnerability, perceived severity, perceived advantages of preventative activities, and perceived obstacles are the four primary categories of beliefs in the HBM framework[91]. The conviction in one's ability to effectively adopt a behaviour, known as self-efficacy, was included later, and has been demonstrated to increase the model's applicability[92]. Recent research has found that obstacles and advantages are the biggest determinants of health behaviour, with these two factors having a higher influence when concentrating on preventative activities than acute diseases/sickness[93].

Individuals may be exposed to information on social media addressing the prevalence of a specific health concern, which may impact their opinion of their own chances of developing that illness, a concept known as perceived susceptibility. Additionally, health campaigns or information published on social media can influence an individual’s impression of the gravity or repercussions of a health condition, a concept known as perceived severity, urging them to take preventive measures.

Social media may also be used to communicate the perceived benefits of adopting a certain behaviour or taking action to prevent or manage a health issue. Furthermore, by giving knowledge, tools, or support to overcome perceived barriers, social media can assist in addressing perceived barriers, or an individual's view of the barriers they may experience while adopting a given health activity.

The HBM has previously been useful in characterising a wide variety of preventative behaviours for illnesses and behaviours that are well documented, enhance the likelihood of early disease identification, and for which the ramifications of any behaviour modifications are typically well recognised[93]. Yet, in most cases, the model was used and evaluated in reasonably established health contexts, allowing individuals to comprehend and assess risks in order to make educated decisions about their personal health behaviour[94].

Notably, every behaviour is culturally influenced, especially when it impacts others, and individuals deliberately alter their answers to match with others' expectations. Because the objective of the activity is critical, this cultural interpretation of behaviour is directly significant for the HBM. Prior studies mostly focused on preventative behaviours linked to non-communicable diseases or disorders, which are often individual-centred behaviours that differ significantly from those connected to pandemics, in which each individual's actions have a cascading influence on others[95].

Integrating the HBM into smallholder social media surveillance can aid in assessing their health-related beliefs, behaviours, and attitudes, and might help to build focused treatments and communication methods that meet their individual needs and concerns. Text mining techniques, namely sentiment analysis, in addition to network analysis, can be adopted to analyse identified patterns related to perceived susceptibility, severity, benefits and barriers amongst the extracted data. After the patterns and trends have been established, this data can be exploited to create tailored health communication campaigns and interventions that address the unique concerns, attitudes, and barriers of smallholders.

1.6.10 Data science in public health and epidemiology

In the last 15 years, data science (DS) techniques applied to both public health and epidemiology have advanced significantly as a result of the boom in internet usage and big data availability[96]. What began as a small, diversified group of academics from several professions analysing the growing volume of internet data for epidemiological reasons has now developed into a new subject all on its own, creating a substantial overlap between the fields of data science, predictive statistics, epidemiology, public health and data visualisation.

DS has the potential to fundamentally alter how scientific research is conducted, and it will be up to public health scholars and practitioners to mould this viewpoint in order to enhance population health. The DS pipeline includes a number of DS-related activities, such as formulating hypotheses, planning research, gathering, storing, manipulating, and processing data, creating and using suitable analysis techniques and distributing findings. This description includes well-known components from many fields, both inside and outside of public health, as well as the explicit objective of preserving and enhancing health.

In addition to adjusting to the emergence of DS, public health has the potential to be a driving force in its continued development[97]. The goal of public health is to understand the consequences of illness, damage, and disability, provide remedies for health inequalities, establish causal links, and offer proof of the efficacy of interventions. Researchers have the critical thinking abilities to judge whether study designs are appropriate for examining scientific hypotheses and carefully examine the measuring and sampling techniques used to collect data[23]. Additionally, to promote improvements in population health, research is fundamentally interdisciplinary and collaborative and draws on a variety of quantitative and qualitative skills.

Public health is in a good position to influence the direction of DS in this aspect because of its longstanding emphasis on ethics[98]. In the foreseeable future, the amount and complexity of data will continue to increase from a variety of sources, such as genomic data from bio-samples, exposure to environmental pollutants, continuous lifestyle behaviour monitoring by wearable technology, detailed health care records from electronic systems, and digital footprints like social media activity, search queries, and online and offline transactions[97].

By identifying the health issues that need to be resolved and assessing the effectiveness of interventions created to address those issues, researchers and practitioners can help to improve population health and well-being. Furthermore it may play a significant role in ensuring that algorithms are used in ways that support, rather than undermine, fairness in health outcomes by including ethical considerations into the development and application of data science methodologies.

How successful ML solutions are depends much on the type and characteristics of the input data as well as the performance of the learning algorithms[99]. Building ML data-driven systems for classification, regression, clustering, and reinforcement learning successfully requires the use of association rule learning and reinforcement learning approaches. Several aspects must be considered when using ML models to develop effective applications with high model accuracy, including data size, feature selection criteria, train/test split ratio, hyperparameter tuning, data class imbalance, algorithm training, and class imbalance[100].

A few of the complicated issues that may be resolved with the use of AI in animal health (AH) include host-pathogen interactions, precision-based animal and human medicine, and quantitative and predictive epidemiology[101]. Benefits of AI include better-targeted medicines, quicker and more accurate risk assessments, and improved representation of complex biological systems. Additionally, it improves case identification and disease diagnosis. The special problems in AH can also act as a catalyst for AI research because of the particular data, constraints, and analytical aims[102].

Three major issues that also affect AH can be solved AI: Understanding a situation and its development, such as the spread of an epidemic, is the first step[102]. The second step is perception, which in AH entails recognising patterns (such as recurring sequences of observations), shapes (such as proteins), and signals (such as increased mortality compared to a baseline) across various scales. The third step is computer-aided decision-making, or, more realistically, human decision support (e.g., expert systems, diagnostic assistance, resource allocation)[101].

Understanding, embracing, and applying AI and ML are essential veterinary researchers and experts to improve the standard and efficiency of our animal health and service. This rapidly expanding field of study and knowledge shouldn't be viewed as a threat to traditional methods. However, it is imperative that caution must be applied when integrating cutting-edge technology into the current animal health system. Awareness of the drawbacks and dangers of using AI and ML, particularly how biased datasets can lead to possibly incorrect findings and how "black box" algorithms make decisions that are difficult to explain. But given the unparalleled rate at which data is being gathered, AI and ML will offer crucial insight into a wide range of prospective tools for health improvement and supplementation.

It is also important to address certain significant drawbacks and issues with using AI models to improve decision-making in terms of disease preparation[103]. First, each and every viral illness reveals distinct natural traits (e.g. transmission route, infectivity, incubation period). Second, during the early stages of an infectious illness outbreak, the understanding of the disease may be restricted. Third, data pertaining to diseases may be stored in various forms, necessitating intensive information extraction efforts. The tuning of Artificial Intelligence (AI) algorithms to conditions related to certain diseases may be necessary. In other words, not all illness situations may be amenable to a one-size-fits-all strategy[103].

1.6.11 Social physics applied to online networks

The behavioural patterns and dynamics of social systems are analysed and understood using methodologies and tools borrowed from the Physics field[104]. Coined in the 19th century, the term has recently gained new life thanks to researchers such as Alex Pentland[105]. The approach relies heavily on data, utilising big data, data mining, machine learning, and network analysis.

By applying this method, it is possible to understand interaction patterns amongst smallholders through the mapping and analysis of the networks that individuals form through interactions and information exchange. Insights into information flow, especially in relation to animal health, can be gained through an understanding of network structures. Additionally, predictive models can incorporate social interactions and behavioural patterns by analysing relevant datasets. The models can play a critical role in forecasting disease outbreaks and how smallholder farmers receive information about them.

To understand the changes in networks and activities, temporal analysis is necessary[106]. Continuous monitoring of trends and networks can provide insights into how smallholder farmers' roles in disease outbreaks change. Moreover, this analysis aids in grasping how farmers' concerns and interactions develop, furthering the aim of continuous monitoring for the procurement of more data and insights into animal health.

1.6.12 Representativeness of Social Media Data

Over the years, the fast-paced development of social media has introduced both possibilities and obstacles in surveillance. Understanding the representativeness of collected data is crucial for governments to make effective investments in technology in this changing environment. Although this issue is important, most literature only scratches the surface.

Significant changes in social media platforms have directly influenced data collection methods. The emergence of TikTok has introduced a new era in content discovery through its fast video-sharing and algorithmic recommendation system[107]. With its emphasis on user engagement and short videos, rather than follower count, TikTok set itself apart from traditional platforms and fundamentally altered the way trends and information circulate.

In an effort to compete with TikTok, Instagram and YouTube have shifted its attention from photo-sharing to short-form video content with the introduction of Instagram Reels[108]. As a result of this shift, content creators have been prompted to embrace fresh formats and modify their strategies, leading to changes in the types of data that are collected and analysed.

The recent rebranding of Twitter as X has led to structural modifications that have had an impact on user behaviour[109]. By incorporating newer features like X Spaces (audio rooms) and paid subscriptions, the goal is to enhance engagement and generate fresh data points. The nature of public conversations on the platform has been impacted by the changes in moderation policies. This shift from textual platforms such as Facebook and Twitter to predominantly image/video-based platforms also have implications on surveillance.

Maintaining the representativeness of social media data continues to be a significant obstacle[110]. User preferences can impact the collected data as platforms cater to diverse demographic groups. Petutschnig et al (2021) found that younger people tend to be particularly drawn to TikTok, whereas Instagram has a wider appeal across different age groups[110]. Politically active individuals are often drawn to X because of its real-time information sharing capabilities.

Moreover, the online presence is shaped by socioeconomic factors, which also impact the representativeness of data. Due to the distinct user demographics on each platform, biases can arise, causing opinions and behaviours that may not accurately reflect broader societal trends. Therefore, reinforcing the need for a holistic surveillance method that incorporates all of these platforms, in addition to a combination of textual, image and audio data.

1.7 Parallels with existing research

Limited research has been conducted on smallholder livestock farming in the UK, leaving this sector largely uncharted. In response to this gap, the current research seeks to employ and customise techniques that are typically used in other extensively researched domains, as highlighted by figure 8.

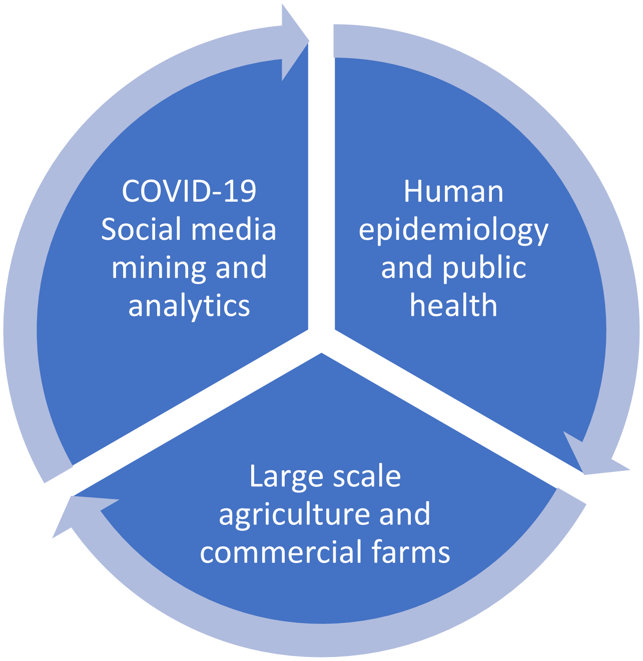

COVID-19 Social media mining and analytics
Human epidemiology and public health
Large scale agriculture and commercial farms

FIGURE 8: PARALLELS WITH EXISTING RESEARCH

Initially, epidemiological studies have provided a valuable framework by mainly focusing on comprehending patterns, distributions, and determinants of health in populations. The purpose is to understand smallholder livestock systems in the UK more profoundly by drawing parallels between these methodologies and the dynamics within the wider scope of veterinary epidemiology.

Another source of research stems from the vast research available on commercial farming. Despite the difference in scale, the methods found within surveillance and monitoring of these larger communities can be transferred over to smaller-scale cohorts, whilst simultaneously accounting for the difference in demographics and farming objectives.

Lastly, social media platforms have emerged as a valuable data source due to the recent increase in digital research, particularly during the COVID-19 pandemic. By using the techniques employed to analyse public sentiment and discourse during the pandemic, this study seeks to gain insight into the perspectives and concerns of the smallholder livestock community in the UK using similar methods. A large proportion of COVID-19 passive listening research revolved around online discussions, particularly through Twitter, therefore providing a foundation to apply these same methods to this project.

By drawing from the existing body of work and making the necessary adjustments so that it is relevant to the population, this project aims to take a step further in filling the interdisciplinary gap within data science, animal health and social media mining.

2/. Literature review

This chapter provides an overview of the primary research areas that are relevant to this thesis. The discussion is structured into four main categories: Data science, surveillance, social media and smallholdings:

1. The first part covers the application of data science tools in both public health and epidemiological research. Literature covers the three main analytical elements used in this thesis; NLP, social network analysis and time series/anomaly detection.
2. The second portion focuses on both public health and epidemiological surveillance within the UK and presents a thorough review of the numerous methods of surveillance, including physical and digital monitoring.
3. The third section dives into the usage of social media as a research tool, presenting an overview of the various social media sites, their capabilities, and the kinds of data that may be acquired from them to supplement research.
4. The final portion of the chapter investigates smallholdings within the UK, in addition to their significance to local economies and food production. The conversation also looks at the influence of social media and other communication methods as a means of dissemination and sharing information amongst these communities.

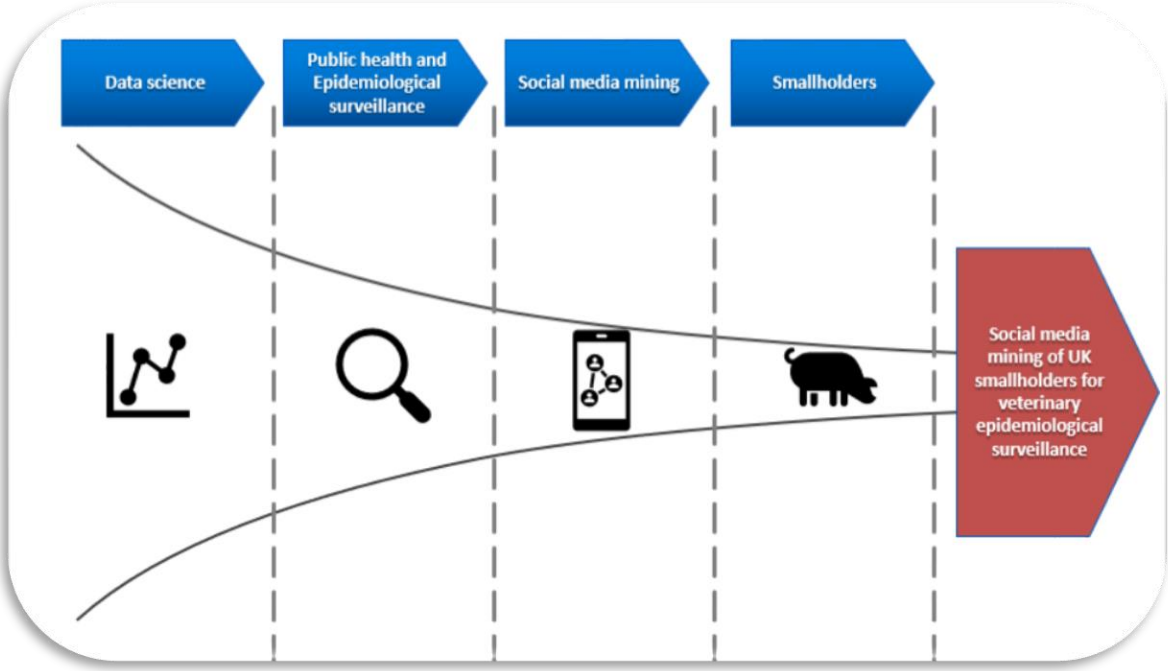

Data science
Public health and Epidemiological surveillance
Social media mining
Smallholders
Social media mining of UK smallholders for veterinary epidemiological surveillance

FIGURE 9: LITERATURE REVIEW FUNNEL PLOT

The funnel plot in figure 9 visually depicts the order of this chapter and the quantity of literature available within this interdisciplinary project, with the width of the funnel illustrating the size of the body of work that currently exists. The final section in red is the gap in existing literature, which this project aims to fill by combining all these four fields together. Through this PhD project, the chasm in current research attempts to be filled by examining how data science, epidemiological surveillance, social media mining, and smallholdings are interconnected within the UK. Despite the isolated aspects of these areas, insufficient research has been performed to explore their intersection.

2.1 Data science in public health and epidemiology

Predicting outbreaks of Cholera and Dengue fever have been extensively studied by South American and African researchers due to its high incidence within these continents[111]. Researchers from Brazil used Machine Learning (ML) algorithms to predict dengue fever outbreaks in the cities of São Paulo and São Luís do Maranhão, by analysing climate data, socioeconomic information, and past dengue cases to make their predictions[112], [113]. Their results showed that the Random Forrest (RF) algorithm yielded the highest overall accuracy and was capable of predicting Dengue outbreaks in advance. By using a combination of datasets, the authors were able to create a larger pool of data to work from and assess the important variables through Principal Components Analysis (PCA). Similar studies have been conducted to explore the accuracy of RF models in forecasting Cholera in coastal India, through the amalgamation of satellite and various environmental data[114]. In continuation, Deep learning was applied by Wang et al (2020) to estimate the prevalence of Hand-Foot-and-Mouth Disease brought on by EV-A71 in Beijing, China, between 2011 and 2018[115]. The authors gathered information on a range of meteorological variables, including temperature, humidity, wind speed, and air pressure, in addition to other pertinent data, such the number of cases recorded and the monthly vaccination rate. These studies illustrate the advances in research on large scale datasets, often disparate, and how DS can be used as an effective tool in disease outbreak prediction.

DS has also been crucial in improving vaccine delivery during outbreaks, allowing the models to optimise delivery strategies by analysing patterns in patient behaviour, logistics and healthcare resource utilisation[116]. The COVID-19 pandemic was a prime example of the importance surrounding robust vaccine delivery systems, as those with the highest risks of infections needed to be seen to first. ML models possess the ability of performing risk stratification on population health data in order to be used for logistic analysis. A different approach was used by Lincoln et al (2022) to examine if vaccine hesitancy in a population could be predicted through Logistic regression[117]. Shapley additive explanations and permutation feature importance were employed to determine the most important features and highlighted that conspiracy beliefs and political affiliations were the highest-ranking features in determining vaccine uptake. Such information can be used effectively by public health agencies to alter and adapt their intervention methods to target previously hesitant demographics[118], reinforcing that DS plays a crucial role in public health guideline formulation and dissemination.

Health disparities relate to inequalities in health outcomes and access to health care across various groups depending on criteria including race, ethnicity, poverty, education, and geographic location. DS has played a significant role in tracking these differences. Populations that are most susceptible to health inequalities can be identified using predictive analytics, and treatments can be tailored to these populations. To determine which people are most at risk for a certain health outcome, such as diabetes or heart disease, ML algorithms may be trained on enormous databases of health outcomes and demographic data[119]. An increasing number of people are worried that the widespread use of ML algorithms in healthcare would further widen existing health inequities[120]. For disadvantaged populations like women or members of racial or ethnic minorities, it is widely known that ML algorithms employed for image-based medical diagnosis, risk prediction, and guiding triage choices underperform. There is another side to the story, too, as academics working in the fields of epidemiology and health economics increasingly use ML algorithms to spot and even remedy health inequities[121].

Recent studies have looked at comparing a ML approach to widely used early warning and severity score systems, to understand the association between racial bias and mortality in hospital admissions[122]. The authors stated in their findings that the algorithm looked at in this study, despite its flaws, offers promise as one of many essential steps in reducing racial disparities in health care. Grote and Keeling argued the legal use of affirmative ML algorithms in public health overfit for socioeconomically and racially underrepresented sub-populations and outperform them compared to sub-populations that have historically had advantages[120].

Another aspect of DS being beneficial in public health is through improving patient outcomes by identifying risk factors and developing targeted interventions[123]. Predicting diabetes through risk stratification models[124], identifying patients at highest risk of heart disease based on lifestyle factors and biometric measurements[125] and Natural language Processing (NLP) of health forums to detect individuals expressing symptoms of depression[126] all reinforce the effectiveness of DS and internet mining techniques being applied to a vast array of public health issues.

Evaluation of the efficacy of public health initiatives relies heavily on DS. Data scientists may analyse the effectiveness of a certain intervention and quantify its impact by gathering and analysing data on healthcare utilisation, health outcomes, and health habits. Making decisions regarding future public health initiatives based on this information will ensure that resources are being spent efficiently to promote population health[127]. Researchers have attempted to incorporate ML tools within smoking cessation apps to assist those who wish to quit[128]. ML models have the ability to offer personalised interventions and health plans based on the users data input history, proving to be more effective at helping quit smoking than traditional generalised approaches often used by healthcare systems[128]. Examples of other addictive behaviour interventions using ML methods have been applied to gambling, alcohol and drug use[129]. These findings imply that supervised learning, in particular, is being utilised more often in addiction psychiatry to guide medical decisions.

Future research should concentrate on examining the integration of big data for infectious diseases and knowledge discovery, as well as modelling methodologies for reliable infectious disease surveillance utilising different data sources[100]. It is intended that attempts to use well-established theories and methodology from computer science, information technology, disease modelling, and disease epidemiology would be made across disciplines in order to build a big data analytic approach for disease modelling.

2.1.1 Data science in veterinary epidemiology

In order to help farmers better support their decision-making and demonstrate the efficiency and welfare of their animals, it may be possible to analyse large integrated datasets[130]. There is very little use of ML models in veterinary care as compared to human medicine, and not enough research has been focused on using ML to address veterinary data issues[131].

Smith et al (2020) sought to determine the type of pig holding facilities in the UK based on movement data[132]. The authors used both analytical and ML approaches to classify the pig holdings into different categories such as indoor breeding and finishing, outdoor breeding and finishing, and outdoor grazing. Results indicated that ML algorithms, specifically RF, provided better results compared to other analytical approaches and were able to accurately predict the type of pig holding facilities based on the movement data. This study provide valuable insights into the efficiency and accuracy of different analytical and ML approaches for classifying pig holdings in the UK.[132]

Taneja et al (2020) applied a mixture of clustering and ML models to detect lameness in dairy cattle, with the assistance of Internet-of-things (IoT) software[133]. For sectors like the dairy industry, IoT, fog computing, cloud computing, and data-driven methodologies together offer a huge opportunity to raise production by gaining actionable insights to improve agricultural operations, hence increasing output and efficiency[133]. A similar approach was taken to forecast Bovine Mastitis in dairy cows, through the comparison of eight ML algorithms; Linear Discriminant analysis, Generalized Linear Model, Naïve Bayes, Classification and regression trees, K-Nearest Neighbour, Support vector machine, RF and Neural networks[134]. The findings point to ML algorithms as a potential tool to help farmers make better decisions, whilst enhancing monitoring techniques and assisting farmers in anticipating which cows would have a high somatic cell count on the upcoming test day. The agriculture sector has actively begun and moved towards smart, tech-enabled solutions to enhance farming operations, boost production, and increase output. The idea of "smart dairy farming" is no longer just a theoretical construct; it has really begun to take shape as several areas, including machine learning, have discovered useful applications in this area[135].

Machado et al applied a RF classifier to understand which variables were pertinent in predicting Bovine Viral Diarrheal virus[136]. A target system and its process, which are often complicated and nonlinear, may be understood better through the application of RF, which enables a novel method of modelling and information extraction from observational data. To the authors' knowledge, there are just a few ML-based research in veterinary epidemiology, and the majority of them continue to disregard the significance of correct and meticulous tuning of model parameters[136]. In a cross-sectional study, for instance, RF was employed to evaluate the risk variables that may have contributed to the spread of pH1N1 from people to pigs in Cameroon, Central Africa[137]. In continuation, Artificial neural network (ANN), RF, and other ML methods were employed by Bradley and Rajendran (2021) to boost adoption rates at animal shelters. This entailed estimating how long each animal will remain in shelters by considering important factors such the animal's kind (dog, cat, etc.), age, gender, breed, size, and shelter location[138]. Furthermore, decision trees have been applied in the veterinary, therapeutic, and environmental areas to build straightforward and understandable ML models[139].

AI and ML might help with disease detection, animal welfare, population medicine, education, and industry elements of the veterinary sector[140]. ANN, case-based, rule-based, and fuzzy logic systems were used to detect a variety of fish ailments over 20 years ago, and the authors determined that the veterinary field was less constrained than the human health field. This should be viewed as a chance to create really novel diagnostic algorithms that demonstrate ideas that might eventually be applied to human and other veterinary medical situations.

2.1.2. Text mining in animal health

Textual information may be obtained from a variety of sources, including scholarly papers, electronic health records, social media, and other web-based tools[37]. Within animal health, the objective is to gather insights that may be utilised to enhance health and well-being, such as forecasting therapy efficacy and generating novel remedies[141].

The healthcare field has witnessed large quantities of research using text mining in electronic patient records to uncover any insights within free-text, even as early as 2001[142]. Parallel studies have subsequently been conducted in animal health, with veterinary reports becoming the focus of NLP techniques to unearth clinical signs. Rodriguez et al (2021) conducted an extensive longitudinal study of over 100,000 tumours affecting companion animals between 2018 and 2019 from Electronic Pathology Records.[143] The results of this study have been adopted into veterinary research and aided both surveillance and clinical decision making. Similarly, Jones-Diette et al (2019) piloted the use of content analysis and text mining for information synthesis and analysis, based on veterinary electric patient records[141]. Both of these studies highlighted the under-utilisation of free-text corpus' in the veterinary field, and the text mining of this unstructured text can aid early disease surveillance through the extraction of clinical signs.

Contiero et al (2019) applied text mining algorithms to examine the interest in scientific literature surrounding pig welfare[37]. They presented a comprehensive analysis of the current scientific literature regarding the experience of pain in pig production and used text mining techniques to examine the published research to identify trends, patterns, and gaps in the knowledge of pig pain. To investigate pain-relieving remedies, 47 years of literature was retrieved, and text mining was performed on the abstracts of these studies. Because this phrase was commonly related with 'acute stress,' text mining identified vocalisations as the primary pain markers in pigs. Building from this analysis, Costa et al (2019) extended this methodology by incorporating stakeholder perceptions regarding animal welfare, collected through an internet survey[144].

Another component of NLP is topic modelling, which aims to find hidden subjects in unstructured text and has the capacity to group comparable documents of text together based on word/term frequencies[145]. Zuliani et al (2021) applied topic modelling on a mountain livestock systematic review to understand the trends and topics discussed over the past forty years[146]. The ten highlighted subjects were deemed to be a suitable foundation for additional and more extensive study (e.g., systematic review) of specific research or geographical regions. Furthermore, the authors proclaimed that in order to identify change drivers and comprehend existing and future issues in mountain livestock production, a genuinely holistic and multidisciplinary research strategy is required.

Sentiment analysis is a branch of NLP in which the goal is to automatically classify the sentiment represented in free text[147], through the classification of emotions and perceptions into positive and negative categories. Largely used in public health research to determine public attitudes regarding new health guidelines, and the volume of literature in this topic has significantly risen since COVID-19[148]. Myszewski et al. (2022) conducted a comprehensive examination of both veterinary and human medicine literature from Twitter in order to understand bias in unfavourable attitudes across the two fields[149]. Their findings revealed that veterinary medicine had a larger bias towards positive results reporting than human medicine.

2.1.3 Social Network analysis

Tools for social network analysis (SNA) are used to examine the interactions between individuals inside groups[150]. They are helpful for analysing the interrelationships and social structures of people or organisations, as well as their working routines. Social media is scanned using social influence network analysis to find important individuals, groups, or trends[151].

SNA has been applied rather rigorously to the public health and epidemiology fields, covering both developed and non-developed nations alike, with the primary goal of quantifying inherent structures within disease transmission, communication networks between communities and animal interaction modelling[152]. Moreover, in public health, for example, SNA can be used to understand the spread of infectious diseases by identifying the relationships between infected individuals and the factors that influence the spread of disease. In epidemiology, SNA can be used to understand the relationships between individuals and their risk factors, as well as to identify the key individuals or groups that play a central role in disease transmission[152].

The traditional approach centred around the spread of infectious disease within populations, with the AIDS epidemic of the 1980's playing a critical part in being the catalyst for increased research into the application of SNA's into epidemiology[153]. The developments in this field spilled over into the veterinary domain, with zoonotic disease SNA's becoming increasingly in demand as a result of outbreaks such as African Swine Fever in 1997[154].

As a consequence of the incorporation of internet data to aid surveillance, researchers harnessed the opportunity to apply SNA theory to the interaction networks of social media users; the same way infectious diseases are modelled[154]. The exchange of information on the internet unravels hidden communities, with influential users of these communities generating the largest engagement with their information dispersal and adopting the role as central nodes within these networks. Furthermore, social media data provides the ability to also visualise text networks, often referred to as semantic network analysis[155]. The automated retrieval of text data though text mining algorithms and the subsequent application of SNA's has proven to be a largely successful endeavour even over 30 years ago[156].

Related work surrounding social media network analysis in the animal health domain is very limited, primarily due to SNA being applied for the purposes of disease outbreaks[157], [158]. Empirical gaps in knowledge need to be addressed to create a multitude of network visualisation options in the veterinary field which utilise social media networks to aid in surveillance.

The work by Albizua et al (2020) explored how social networks influence farming practises and promote agricultural sustainability[159]. This research examined how social networks may be used to help farmers embrace more sustainable agricultural practises, such as utilising more environmentally friendly methods, decreasing waste, and increasing yields. The results indicated that social networks could help distribute information about sustainable agricultural techniques and shape farmers' attitudes and behaviour. Their results emphasised the significance of considering the structure of these networks, such as the types of relationships between farmers and the role of important farmers in impacting the attitudes and actions of others. The authors concluded that social networks have the potential to play a significant role in promoting sustainable agriculture, but that additional research is required to better understand how they may be utilised to help farmers adopt more sustainable practises.

Research has suggested that one of SNA's advantages is its capacity for handling massive volumes of data and offering a comprehensive perspective of intricate social networks[152]. Additionally, it may be used to find patterns and connections that may not be immediately obvious from other types of data analysis. In contrast, it does have certain restrictions, for instance, SNA assumes that connections between people or things are fixed, which isn't always the case as dynamic systems in social media network are most likely prevalent. In continuation, it also makes the supposition that connections are reciprocal, i.e., that if one person is connected to another, the opposite is also true. As certain connections are unbalanced, this may not always be the case, as evident within Twitter networks in particular[160].

Animal health professionals can employ SNA to develop disease prevention plans, including specialised vaccination campaigns or quarantine restrictions, and to evaluate the success of such efforts[152]. Dealing with illnesses that are difficult to identify or prevent, such as those that are carried by wildlife or via indirect contact, can be very helpful when using this strategy. Overall, social network analysis is a valuable weapon in the toolbox of researchers and practitioners in the field of animal health because it offers a potent tool for comprehending and limiting the spread of illnesses in animal populations.

2.1.4 Community detection

Communities are important for understanding and modelling complex systems because of the assumption that they have dense connections inside a group and sparser connections to the rest of the network[161]. The number, diversity, and complexity of contemporary networks are expanding tremendously[162]. Newer and distinct communication network types, such multi-agent, IoT, ad-hoc, wireless sensors, cloud-based, co-citation, and social networks, are developing as a result of network modifications and growth in internet-based data applications.

The addition of social media augments this field as both conversational and friend networks can be modelled through network analysis[163]. Social media networks differ in terms of the entities and the kinds of relationships they represent, but despite this, they represent a substantial source of intelligence since they store the online behaviours and contributions of huge numbers of users.

In addition to gaining insights into the social phenomena and processes that exist in our society, studying such networks allows one to extract knowledge that is useful for a variety of information management and retrieval tasks, such as online content navigation and recommendation[164]. Data mining techniques, however, have significant difficulties when analysing these networks since they virtually always have enormous sizes and a highly dynamic nature[165].

Several approaches for detecting communities have been developed, namely the Louvain algorithm[166], Label Propagation[167] and Infomap[168]. These techniques relied only on the network's connection topology. They may not be able to reliably identify community members due to the sparse and noisy network structure of social networks such as Twitter[169]. The Louvain method is one of the most often used algorithms for detecting community, although it contains a flaw that might result in poorly linked communities.

Other studies have determined communities only based on user material[161], [170]; however the results are insufficient since communities are only dependent on subject. Some studies addressed this issue by including user content information into the link network in community detection and combining it to increase the quality of community algorithms[169].

A vast range of mechanisms connected to one another through a wide variety of interactions and linkages make up the ecosystem of social media apps. Social media networks, which include online items as their vertices and the relationships/interactions between them as their edges, offer a pleasant depiction of social media data[164]. Vertices can represent a variety of players, including individuals, content (e.g. Blog posts, images, and videos), and even metadata (e.g. subject categories and tags). Additionally, depending on the network development process, the edges of social media networks might be basic, weighted, directed, or multiway (i.e. linking more than two entities)[163].

The relevance of community discovery in social networks has been widely investigated, with earlier work from Clauset et al. analysing the structure of very large social networks in order to uncover purchasing habits in 2004[171]. However, these methods focus solely on the link and graphical body of social networks and ignore user interactions and interests, as well as the influence of users on social networks on the Internet. In order to address this issue, researchers have proposed alternative methods based on the Bayesian eventuality model[172].

2.1.4.1 Twitter

Twitter's social network has recently grown at a remarkable pace with users exploiting social networks to express their thoughts, feelings, and ideas on numerous subjects[169]. They frequently form groups with other people who have similar beliefs or events, and share information with other users through extensive communication networks. Quantifying these networks can be accomplished by examining the trends of tweets, retweets, mentions, responses, and other types of user activity[25].

Bello et al (2016) analysed online user communities on Twitter to determine the attitudes of these groups concerning vaccinations, in the midst of preventable disease outbreaks such as Measles, polio and Influenza[173]. This study initially examined WHO data of vaccination coverage rates before employing community detection techniques to identify communities. To locate the communities, they used a hierarchical clustering approach based on network links and user interests, with the implementation of numerous algorithms including Infomap and Label propagation. The results from this study resulted in further studies being conducted in public health perceptions surrounding vaccination hesitancy communities[174], as witnessed in the numerous contemporary studies during COVID-19[175], [176], [177]. These methods are beneficial for describing community opinion in public health applications.

Building from previous studies, in an attempt to improve community discovery, Lam et al (2018) employed semantic analysis and the following/follower connections to determine edge weight, in addition to incorporating topic modelling to analyse discourse[178]. Kanavos et al (2018) suggested a novel method for identifying significant communities in social networks based on emotional behaviour and user profile[179]. This technique relied on the content of each user post to determine user emotion based on the user's impact metric on a certain topic and this method gave a unique impact measure to each user, which was then utilised to determine the most influential communities.

Amati et al (2021) demonstrated how to use the social network identification approach to find certain subjects in Italian tweets[180]. Communities with varying interests and content were identified using a flow-based community identification approach in online discussion. The authors also improved a community discovery algorithm by combining knowledge graphs. This approach combines user context information with additional qualities via hierarchical concept maps to detect nodes in the community and community that matches the context of the community.

Benabdelkrim et al (2021) used a generated graph via an edge attributed multilayer network of Twitter membership lists to find local groups and shared interests of people[181]. The approach entails embedding layer textual information in a Euclidean space and using it to group together nodes belonging to semantically near layers in the same community. To that purpose, they created a pattern mining method for extracting communities from numerical data. Furthermore, by combining symbolic and numerical approaches, this method was especially well-suited for identifying communities in multilayer networks[182].

Finally, Akachar et al (2019) introduced a novel technique to detect communities that take into account both embedding users and message content[183]. This strategy depicted embedding people by retweet relationships among users while message content was based on semantic similarity of the users.

2.1.4.2 Facebook

The Facebook network is a large, complex, and dynamic system that poses a number of challenges for community detection. One of the main challenges is the sparsity of the network, as most users on Facebook have a relatively small number of connections[184]. Another challenge is the presence of noise in the network, as many interactions on Facebook are non-meaningful or irrelevant. Furthermore, the lack of ground truth data makes it difficult to verify the accuracy of the detected communities. Lastly, the network is highly dynamic, with users constantly changing their behaviour and interests, which can make it difficult to accurately detect communities over time[184].

In addition to the algorithmic approaches, there have been several studies that have attempted to characterise the structure of Facebook communities and understand the factors that drive community formation. Ferrara was amongst the first researchers to conduct such a study back in 2012, setting the foundation for further studies to be undertaken[185]. For example, literature has indicated that communities on Facebook tend to be highly homophilic, with members having similar demographic characteristics, such as age, gender, and location[186]. Other studies have found that communities on Facebook tend to be formed around specific topics or interests, such as politics, sports, or entertainment[187].

2.1.4.3 Forums

Forums are a popular venue for detecting communities since they provide a plethora of information about user interactions and the subjects being discussed[188]. Researchers and companies may obtain insights into the structure and behaviour of online communities and understand the variables that drive community creation by examining this data[189].

Wu (2014) suggested a method for evaluating WeChat social network interactions and creating social network diagrams that emphasise user closeness[190]. WeChat users were grouped depending on the security of their accounts. While mob phenomena have been well-studied in sociology and public health previously, Albaldi & Weir (2018) noticed that user categorisation based on social influence was particularly significant in social network forensic investigations[191]. They identified prospective flash mob organisers by categorising discussion forum users based on their online activity.

Similarly, research conducted by Kumar et al. (2018) investigated the nature of disagreements between communities on Reddit[192]. Their findings highlighted that such disputes are often launched by a small number of groups: fewer than 1% of communities initiate 74% of conflicts. Conflicts are typically sparked by highly engaged members of the community, but they are carried out by fewer active people. Furthermore, they discovered that disagreements are characterised by the establishment of echo chambers, in which users predominantly communicate with other members of their own group. Conflicts have negative long-term repercussions and lower overall activity of users in the targeted areas[193].

2.1.4.4 Influential nodes

Influential nodes in community discovery algorithms can have an important influence in deciding the eventual partitioning of the network into communities. For example, if an important node is allocated to a certain community, it may attract other nodes in the network to that community, influencing the overall general structure. As a result, the function of significant nodes in the community discovery process must be carefully considered[194]. Tulu et al (2018) stated that like all networks, certain nodes (users) tend to rise to the top of the information flow hierarchy and become influential users within these communities[195]. They yield the ability to control the narrative due to their large following and attention garnered, with their posts receiving a greater number of engagements compared to others. Engagement metrics in Twitter for example are based on likes, comments, retweets and followers.

Recent political scandals involving Cambridge Analytica's nefarious activity in influencing voters highlights how susceptible networks are to manipulation[196]. Sentiments towards purchasing decisions, voting and scientific guidelines can all be swayed by the influential members of the community, as reinforced by recent literature using Facebook data[197], [198].

The importance of identifying and predicting these individuals becomes a challenge for researchers. Moreover, attempting to quantify their influence adds a level of complexity as social media networks tend to be dynamic nature; perpetually changing sizes and users being promoted/demoted in "influential" capacities[199]. Alotaibi & Rhouma (2022) stated that social media consumers disclose their beliefs, knowledge and ideas on a diverse range of topics, which can often be contentious in nature and thus lead to opinion-fuelled debates[200]. The extent to which the user has influence as a result of their online post depends heavily on not only their relative level of celebrity amongst their peers, but the very nature of the topic itself, i.e. COVID-19 mandatory vaccinations, lockdowns etc[118].

Wang & Zhao (2020) explored the effects of a topic/subject on its referenced and shared behaviour by others henceforward[201]. This is a great method for initial exploratory analysis however omits the underlying structural information of the papers in the network, as the topic or theme of the paper is not the only viable predictor of influential nodes. Therefore the need to adopt a dualistic approach which takes into consideration both the structural location of users in a network, and the topics/opinions shared by them is integral in order to build a more robust predictive model[29]. A structural model extracts the centrality of the characteristics of nodes, whilst a topic-model identifies the prominent topics disseminated within the networks.

Using ML methodology to identify prominent nodes in a network offers various advantages over previous methods[202]. The ML model can better predict which nodes are genuinely significant in the network by integrating the semantic components of the nodes[203]. The semantic part of the model relates to the findings of a LDA model applied to the data. The node characteristics are retrieved and used in the machine learning algorithm throughout this process, allowing it to evaluate not just the network topology but also the underlying material or subjects linked with each node.

Twitter is a social network which is comprised of both content and context, with the latter referring to the structure or formation of the social network itself[204]. Dhokar et al (2021) incorporated topic-based approaches by aggregating the content of a user post with its context[205]. The authors continue by stating that influence is defined as the individual's social impact and the impact this has on the remainder of the community in which they operate.

A difficult issue which arises in larger and more complex social networks is the ability to robustly forecast a collection of influential users (nodes) within the network. Past literature usually only focusses on the relationship between the nodes, whilst omitting the local update effects on the diffusion process[202].

2.1.5 User classification

In social media, user classification/categorisation is the process of identifying and categorising individuals based on their activity, interests, demographics, and other criteria[27]. A survey of recent literature suggests that ML, particularly deep learning techniques such as neural networks and convolutional neural networks, is the most widely utilised strategy[206]. Researchers have also investigated how to extract significant characteristics from user-generated material and social network structures using SNA and NLP techniques.

Stemming from preliminary research in Twitter content categorisation, i.e. classifying a tweet as sports, entertainment, politics etc based on the textual inputs, user categorisation takes a similar approach but with the user demographics[207]. Manually labelling a training dataset with two or more coders prior to any text classification is required.

Within public health, studies have employed such approaches in categorising illnesses amongst the population[208]. This included differentiating between dementia and Alzheimer's based on tweet content and matching keywords to domain-specific dictionaries. User classification has also been adopted to distinguish the type of entity that the profile or/and tweet originates from. Alsudias & Rayson (2020) explored methods to categorise users who disseminated content pertaining to Infectious diseases into "Academic", "Media", "Government", "Health professionals" and "public"[206]. Their findings reinforce that multiclass classification studies such as this require a more substantial training dataset, as a balanced distribution for each class is preferable[206]. Similar work has not been replicated in the veterinary domain, however similar approaches using with electronic health veterinary records have been applied for the automatic labelling of disease based on clinical signs[209].

In continuation, Kim et al (2017) attempted to classify Twitter users that discussed E-cigarettes into 5 distinct categories, each pertaining to varying degrees of E-cigarette usage[210]. Their text classification model also included metadata features, such as profile colour and background, and resulted in prediction accuracies in exceeding 90%. Similar methods have been conducted which considers multiple features from Twitter, not just the tweet content itself. Gilani et al (2017) performed a binary classification on Twitter accounts to separate bots from human user, and achieved comparatively high predictions through their holistic model[211]. Daouadi et al (2019) also combined both statistical and content-based approaches to classify bots from human users, reinforcing the need for a hybrid approach to user categorisation, particularly with social media data[212].

While the accuracy of categorisation algorithms has considerably increased in recent years, a number of issues remain. Dealing with noisy and skewed data, overcoming privacy problems, and resolving ethical difficulties associated to the use of personal information for classification purposes are among these[206]. As social media platforms develop and evolve, more advanced user categorisation systems that can adapt to changing user behaviours and preferences are required.

2.1.6 Time series and anomaly detection

Anomaly detection or outlier detection is a method applied to time series analysis to identify abnormalities within the data points, relative to the expected average[213]. The cause of such anomalies could be caused by innate variability or inconsistencies within the data itself. Outliers permeate throughout social media data due to its volume and instantaneous nature, and thus distinguishing the relevant data from the noise is a crucial element behind any robust early disease detection tool[214].

The public health field has produced numerous early-detection tools through the application of social media data, and this has seen a considerable increase since the advent of COVID-19. Influenza-based detection instruments from Twitter data have been popular modes of corroborating validated outbreak data, with the keyword-based discovery algorithms from a spatial and temporal perspective. The effectiveness of such tools has been reinforced in earlier works by Paul et al (2014), Culotta (2010), Achrekar (2011) and Sharpe (2017)[215], [216], [217], [218].

The increase in urgency surrounding zoonotic diseases (i.e., avian influenza, swine flu) has merged the fields of animal and human health, illuminating the need for adequate intelligence systems for early detection warnings. Building on from the work of Robertson and Yee (2016)[21], Yousefinaghani et al (2019) assessed the efficacy of a time series model, combined with outlier detection, to create an early warning system of Avian Influenza based on Twitter activity in North America[86]. The authors employed web scraping methods to extract relevant tweets from the Twitter stream that contained terms linked to Avian Influenza. Moreover, the data was then evaluated to detect regional and temporal trends that might indicate the start of an outbreak. The study discovered that Twitter data may be used to detect early stages of an outbreak and pinpoint its geographic location and that integrating Twitter data with other sources of data, such as government reports and news articles, can enhance outbreak detection accuracy.

Chen et al (2019) applied the same methodology, but with Chinese internet search query data in an attempt to detect the H7N9 strain of Avian Influenza in humans[219]. Confirmed laboratory cases were mapped against Weibo (a popular social media platform in China) posting data between 2013 and 2017, to explore spatial and temporal trends of the strain. They discovered that H7N9-related search queries were substantially linked with the actual number of cases, implying that internet search data may be utilised as a tool for early epidemic identification. In addition, the researchers also discovered many phrases related with H7N9 outbreaks, such as "pneumonia," "fever," and "bird flu," which might be used to improve the accuracy of early warning systems. The study reinforces the potential of internet search data as a supplement to traditional disease monitoring and epidemic detection approaches in China, where traditional surveillance methods may have limitations. However, they emphasised the need for more studies to address concerns such as data quality, representativeness, and privacy.

Regarding the integration of such methods into syndromic surveillance systems, anomaly detection algorithms have been implemented in veterinary laboratory data by combining Shewhart control charts and Holt-Winters exponential smoothing, by setting thresholds for the observed frequencies. Research performed by Dorea et al[209] depicted a visual representation of an early-warning system through which the algorithm has learned from a training dataset, and any new data point which exceeded the detection limit, is flagged as an outlier. The implementation of such detection algorithms for social media data is the next step for syndromic surveillance.

2.1.7 Peak detection using social media

Building on from anomaly detection, social media peak detection methods have been suggested to expose events and happenings from social geo-tagged data[220]. There are usually two primary stages involved in these processes: (i) extracting space-time attributes from social data and modelling them as time series, and (ii) detecting peaks in time series to identify deviations from a user's usual behaviour[221].

Automated peak detection methods have been developed in the animal health field to identify behavioural events. An automated peak detection method was employed in a study to detect feeding attempts in Risso's dolphins and blue whales[222]. The norm-jerk time series was used to accomplish this, which records the acceleration's rate of change. The study found that their method for detecting peaks was effective in automatically identifying animal behaviours from multi-sensor tag data with high precision for behaviours that are appropriately represented by the data time series.

The potential of peak detection methods via social media has been demonstrated in revealing events and happenings from social geo-tagged data across various domains. Public health, epidemiology, and animal health have employed these methods to enhance our understanding of events and behaviours[223].

2.2 Surveillance

2.2.1 Surveillance in public health and epidemiology

Syndromic surveillance (SS) is a public health term denoting a range of approaches to identify both individual and population-based health markers which are evident, prior to the commencing of a diagnosis[224]. This refers to indicators of disease symptoms, clinical signs and behaviours preceding the official lab test verification of an infectious disease. Furthermore, it is an epidemiological method adopted globally by governmental agencies such as the World Health Organization (WHO) to recognise premature clusters of disease or symptoms promptly to quell the possibility of an outbreak[225] .

Literature shows that traditional methods of devising such models include the aggregation of population health data produced prior to a confirmed diagnosis, which allows early interventions to be enforced by the public health agencies, such as the Ebola epidemic of 2013[19]. The authors purport that developing robust mechanisms for the early detection of disease has been a global priority for over half a century, with SS evolving in-line with internet advancements to create novel methods of early detection[226].

The H1N1 influenza outbreak from 2009 and the Ebola epidemic from 2013 are contemporary instances highlighting that modelling outbreaks is not always accurate or reliable[227]. However, the discovery and examination of irregular health-related instances successfully permit the accurate classification of infectious diseases outbreaks. Dissimilar to previous ES systems, which entailed often arduous and time-consuming efforts to acquire data, SS enhanced this field by incorporating real-time data via the automation of protocols and reports. The dynamic nature of these systems makes them invaluable for disease outbreak discovery, observation, and tracking, as purported by Chen et al (2020)[228].

The increased popularity of these methods gained momentum through incidences of bio-terrorist activities in the early 21st century; for example, the anthrax assaults of 2001, and epidemics of rising communicable diseases, such as SARS[229]. This underpinned the requirement to identify examples symptomatic of a potential emergence of human pathogens rapidly and before disease progression. Applying the methods available by health bioinformatics, dynamic monitoring practices were cultivated to apply readily accessible pre-diagnosis data, such as pharmacy sale data pertaining to prescriptions, non-attendances to work/school, presenting complaints and triage comments in hospital, or lab test requests[68], [209].

The deficiency pertaining to the validity of pre-diagnosis data was a key component in the shift towards surveillance which examined clusters of diseases (i.e., syndromes), which despite being less detailed and particular than confirmed lab data, proved to be significantly quicker[230]. The supposition of this type of surveillance system isn't centred around its ability to generalise disease burdens on a population level but rather that it is perceptive to fluctuations in the intensities of disease at a population level, thus reinforcing early detection efficacy[230].

Within the UK, Public Health England (PHE) have devised a Real-time Syndromic Surveillance Team (ReSST), which aggregates in-depth general practice (GP) and hospital emergency patient data, in addition to triage information to enable the early detection of disease spread on both a regional and population level[231]. This was an augmentation to the previous NHS direct syndromic surveillance system which played an integral role in quelling the Influenza epidemic of 2009. This system was further enhanced in 2011 with the introduction of NHS 111 syndromic surveillance system, which allowed clinicians to analyse the chief complaint from each call to detect possible symptoms and proved to be successful during the winter of 2013-14. More recently, both PHE and Public Health Scotland (PHS) have utilised Covid-19 related syndrome surveillance to suppress regional outbreaks, with early detections being the forefront for country-wide lockdowns[232].

2.2.2 Surveillance in veterinary epidemiology

The rapid advancements of disease and infection regulation procedures have had a substantial positive impact on the veterinary epidemiology field; from traditional in-person inspections and timely report submissions to automated data collection and advanced epidemiological intelligence tools for analysis and policy making[233]. Historically, the animal health field has always lagged behind the human health field in regard to early detection tools, methodologies and implementation, however in the aftermath of recent major zoonotic diseases (e.g. Avian influenza H5N8, African Swine Fever), governing bodies have sought to bridge this chasm as the interconnectedness of the two domains is strikingly apparent[234]. There still remain large disparities in surveillance capabilities globally, as the 1st world is able to benefit from advanced technologies, unlike the 3rd world which lacks the resources; which proves to be a problem due to the inception of some zoonotic diseases stemming from 3rd world countries, hence the risk of pandemics is still relatively high as early detection is not as robust in these regions[235].

Veterinary data stems from both wild and domesticated livestock alike and offer potential for tremendous insights into the spread of zoonotic diseases, with a successful European example being utilising serum gathered from domestic pigeons, for the premature detection of West Nile Virus in Greece[236]. Similar to the public health field, there still remains a large chasm between developed and developing nations in relations to the viability and accuracy of animal surveillance systems, and global efforts of a standardised reporting system still remain at a distance[237].

In this novel environment, offering viable and ample methods for methodical data storage and analysis plays a key role in attaining the objectives of veterinary disease surveillance[238]. Despite the notion of aggregating disparate data sources for improved intelligence not being new[239], the field of veterinary health informatics has advanced to include new modalities of surveillance via geographic information systems (GIS), predictive analytics for outbreak extrapolation, spatio-temporal analysis and visualisations.

Researchers from The University of Liverpool devised a tool for small animal surveillance referred to as The Small Animal Veterinary Surveillance Network (SAVSNET), which observes infection in small animals during the first vet visit, utilising data from lab results and dynamic compilations of consultation records from vet practices[240]. Such systems have proven to be successful on a small scale in the UK, however an extension to poultry or pigs has not yet been implemented, possibly due to the difficulties in data collection.

Literature within the animal health surveillance domain pertains mostly to zoonotic diseases and is heavily focused on companion animals primarily due to their immediacy with humans. Earlier research has explored the correlations between both wild and domestic animals; Leblondlet et al (2007) examined the susceptibility of wild birds and horses to West Nile Virus, through the use of surveillance methods[241]. The convergence between animal and human health was heavily researched in North America, with the introduction of the North Dakota Electronic Animal Health Surveillance system in 2007[242]. The incorporation of veterinary data alongside patient data to identify analogous symptoms proved to be a feasible study by Shafferet et al (2007) and Maciejewski et al (2007) articulated a set of guidelines to combine hospital emergency data with veterinary data[243].

The success of these tools has been debated more recently, with little information being known on the viability of amalgamating the two fields. Currently in the UK, systems to link these databases together are not in place, however APHA provide regular surveillance reports on the welfare of a wide range of animals; with both pigs and poultry possessing dedicated geographic dashboards for in-depth visualisations[244].

2.2.3 Infoveillance and Infodemiology

The success of SS opened up avenues for researchers to examine new data sources to include into their analysis to detect outbreaks and spread of infections, via the medium of the internet[24]. This spawned two new branches of health informatics entitled Infodemiology, and Infoveillance: the former pertaining to the use of internet data for epidemiological supplementation and the latter referring to the use of internet data for surveillance[24]. Johnson et al (2004) were amongst the first researchers to adopt such a mechanism in the early 21st century, whereby they examined Google and web searches to divulge information regarding the dispersion and extent of influenza in the USA[15]. The success of this work was the foundation for further studies which would build the credibility of these newly developed surveillance systems, and both terms fall under the categorisation of SS. The automated and incessant flow of free-text internet data proves to be invaluable to researchers due to being time-efficient and monetarily cheap; two traits which were constraints in traditional surveillance. The precise issues with Googe Flu Trends (GFT) have been well-explained and debated, but the bigger issue was that the underlying data was privately owned, making it impossible to replicate and do independent research on the technique[18].

Research indicates that the adoption of this additional data source has proved successful in some applications yet contains methodological challenges and data validity scepticism[245]. There has been an increasing understanding of the usefulness of social media analysis in public health and disease monitoring in recent years. The growing availability of internet data, as well as the need for more efficient and effective means of tracking disease outbreaks, have contributed to this trend[15]. A rising body of evidence supports the value of Infodemiology as many studies have shown that social media data may be used to track the transmission of illnesses such as dengue fever, foodborne illness, and influenza-like illness (ILI)[246].

Literature has illustrated that one of the major issues, however, is the requirement to authenticate the data and assure its correctness. This is especially crucial in the context of disease outbreaks, where precise data is required to enable effective decision-making and response[247]. Notwithstanding the limitations, there is widespread agreement that this novel technique of disease surveillance has the potential to complement and improve established methods. The data, for example, may be utilised to offer early warning signs of disease outbreaks and to assist public health professionals in responding more rapidly and effectively[74].

Facebook and Twitter are examples of avenues of data extraction to aid in both public health and veterinary surveillance[227]. There are estimates of over two billion global Facebook and four hundred million active Twitter users; all of which provides an ever-increasing flow of public opinion and first-hand news[248]. In particular, social media has been especially effective in curating a plethora of passively collected data ranging from sociodemographic information, lifestyle behaviours and even attitudes/sentiments towards products and ideas. Figure 10 illustrates the implementation of Infoveillance techniques over the past thirty years, stemming from email scraping to the advanced social media scraping of today:

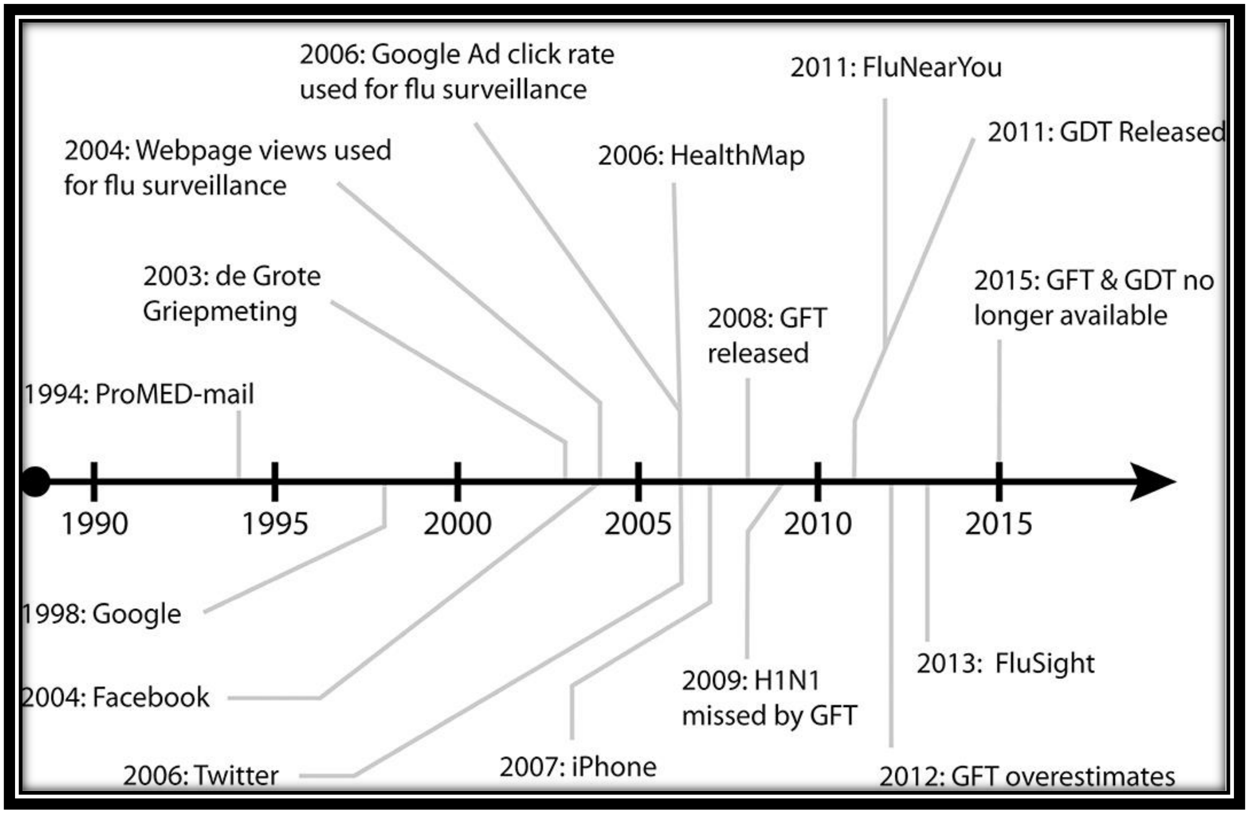


FIGURE 10: INFOVEILLANCE TIMELINE

The Ebola epidemic in Western Africa in 2014 is a pertinent example of how rapidly public health related information can be transmitted via social media; Allowing mathematical models to be derived to ascertain how this information disseminated to the online population[19]. The dynamic nature of social media news means that the public are able to access the information significantly faster than other traditional avenues, such as the news on TV, however at the expense of validity. Similar to a communicable disease, social media follows an analogous route of transmission through one-to-one communication, resulting in online messages going "viral". Obtaining the ability to quantify the diffusion of social media information at a population level, enables public health researchers and epidemiologists to determine if the information they relay is firstly reaching the right groups (those at risk), and secondly, how these users utilise this information (e.g. retweet to others). Moreover, influential users can be detected within communities, based on the number of retweets, likes and followers one has, and the interaction amongst these smaller communities may consist of potentially useful patterns[25].

Health informatics researchers have been discovering methods to incorporate social media and other internet data to deliver time/cost effective predictions of disease occurrence. Tracking influenza incidence and prevalence through Twitter hashtag searches has been hugely successful in both the USA and Europe, whereas the insights drawn from Google trends have proven to be less accurate in comparison with the USA sentinel influenza-like illness surveillance system[76].

Researchers also reinforced the additional benefit of social media as a communication tool to increase global perception and understanding of outbreaks[33]. Furthermore, Infoveillance techniques have been adopted to model the variations in epidemic awareness, in addition to public attitudes/sentiments towards health measures enforced by governments. Recent studies have used Infoveillance to ascertain variations in global perceptions to Coronavirus 2 SARS-CoV-2 (COVID-19) related governmental lockdown enforcements[148]. An epidemiological study of this magnitude would not be possible in such a short timeframe without the employment of social media data.

Critics of such approaches have vocalised concerns regarding the ethics surrounding the acquisition and analysis of social media data to inform public health policy[249]. Academic guidelines surrounding its usage for research still remain ambiguous as the notion of both informed consent and anonymity remain in doubt, thus researchers have attempted to abide by ethical frameworks through the mining of only publicly available data as opposed to data from private groups.

Literature within the animal and veterinary field is very scarce, with the efficacy of Infoveillance still unapplied and invalidated within this domain. However, augmentations in current SS methods to include internet search-based data has started to gain popularity amongst UK and European researchers.

To summarise, Infodemiology is a fast-developing science with the potential to significantly contribute to public health and disease surveillance. Yet, considerable effort needs to be done in order to fully realise its potential and solve future epidemiological problems.

2.2.4 Infoveillance and Infodemiology – Post COVID-19

The pandemic in early 2020 resulted in a tremendous increase in online activity pertaining to information searching and opinion sharing of COVID-19[250]. Infodemiology of COVID-19 refers to the study of how information about the virus and the pandemic was distributed, spread, and perceived in society. During the pandemic, the internet and social media became primary sources of information about the virus for many people and the sheer amount of information and misinformation about the disease can be overwhelming, leading to confusion and anxiety[251].

The rapid spread of information also led to the spread of misinformation and conspiracy theories, which can be harmful to public health efforts. For example, misinformation about the virus's origin, transmission, and treatments can lead to distrust in health authorities and decisions that harm public health[251]. Inaccurate information can also lead to incorrect personal choices and behaviours, such as not wearing masks or avoiding vaccines.

In order to combat misinformation and promote accurate information, health authorities and organisations have used various strategies during this time, including social media campaigns, targeted messaging, and partnerships with tech companies[251]. Additionally, various fact-checking organisations have emerged to independently verify information and to counter misinformation[252].

The function of the media in information dissemination has also been discussed in the discipline of Infodemiology. While the media has been critical in keeping the public informed about the epidemic, they have also been chastised or sensationalising events, disseminating disinformation, and failing to offer enough context and analysis[252].

Contemporary examples of studies adopting social media for COVID-19 investigation are summarised below:

- Monitoring disinformation regarding COVID-19 and its treatments[253].
- Assessing public perceptions regarding vaccinations through the tracking of changes in public opinions towards immunisations using web search data[254].
- Understanding vaccine hesitation through conducting a survey study to determine the variables that lead to vaccination hesitancy in the general population[255].
- Tracking the spread of novel variations using genomic sequencing data to trace the distribution of new variants and their influence on public health outcomes[256].
- Monitoring the impact of preventative measures by using mobile app data to track the effectiveness of mask use on disease transmission[257].
- Observing the effect of social distancing measures through the utilisation of mobile location[258].

- Measuring the impact of school closures via web search data to analyse changes in public attitudes and behaviours around school closures[259].

These studies show how Infoveillance may be used to monitor the epidemic and analyse the impact of public health initiatives. Infoveillance has assisted in informing public health decision-making and improving our response to the epidemic by giving real-time data. The findings of these research also underscore the significance of increasing digital and media literacy, as well as giving accurate and accessible information to the public during public health crises. It also highlights the importance of continuing efforts to monitor, verify, and correct misinformation regarding other public health concerns.

2.2.5 Spatial surveillance

Digital surveillance tools are now a valuable data source in public health and epidemiology, for mapping the spread of infectious diseases in near real-time, especially since COVID-19. Bhatia et al. [119] provided a detailed overview of digital surveillance, covering types, methods, and challenges. Digital sources such as online media reports, social media posts, mobility data, and contact tracing apps were discussed. They put forward a structure for integrating digital surveillance data with mathematical models to estimate disease transmission potential and spatial distribution. The authors highlight the importance of enhancing the quality, validity, and availability of digital surveillance data for public health research and decision-making. Despite this, they acknowledge the ethical and legal implications that digital surveillance may pose regarding data privacy, security, and ownership, which aligns with previous research with comparable findings[260], [261].

The agricultural and livestock domain has seen fast transformations in spatial surveillance[262]. This shift represents a significant step towards sustainable and efficient agricultural practices, thanks to technological advancements. Few studies have explored the intricacies and benefits of spatial surveillance for small-scale agriculture, emphasising its capacity to enhance crop production, encourage sustainable farming techniques, and aid decision-making[263]. However, smallholding communities may resist the mass adoption of such technologies due to lack of resources, knowledge, and concerns over ethics and privacy, as corroborated by the works of Bhatia et al[264].

2.2.6 Societal Acceptance of Social Media Surveillance

Before delving into the benefits and methods of social media surveillance, and its potential as a supplementary tool, questions need to be addressed around the likelihood of such a technology being adopted within wider society, let alone a more closed off cohort such as smallholders. Since the Cambridge Analytica scandal, we have witnessed significant public aversions to both social media scraping and government-based surveillance methods[265]. This aversion not only acts as a significant barrier to this project but raises questions around the long-term feasibility of such passive surveillance methods.

Westerlund et al conducted a thorough study exploring citizens' views on digital identity, state intelligence activities, and biodata security[266]. The trust and acceptance of government use of personal data were critically examined in light of these factors. The research underscores the increasing alarm among news media and human rights organizations regarding the emergence of a surveillance state that relies on mass surveillance and a lack of trust in citizens. Following the aftermath of the COVID-19 pandemic, public anxiety and worry has increased, given the rapid digitalisation and the connection between state and corporate surveillance. The results of the study highlight the ethical and social dilemmas presented by digital surveillance in today's world, where the use of big data significantly influences public discussions and policymaking.

2.2.6.1 Theoretical Context

Two essential theories, the Technology Acceptance Model (TAM) and the Theory of Planned Behaviour (TPB), have been crucial in establishing a theoretical framework for technology adoption. According to Davis (1989), the TAM suggests that adoption is determined by perceived usefulness and ease of use[267]. This model has shown exceptional insight in the fields of healthcare, epidemiology, and public health. Research within animal health has explored the perceptions of health providers towards telehealth technology, demonstrating its ability to promote remote connections and interdisciplinary collaboration in palliative care[268].

Furthermore, within healthcare settings, the model has revealed how technology adoption can enhance patient outcomes, data management, collaboration, skill development, and reduce organisational expenses. AlQudah et al (2021) identified perceived usefulness and ease of use as influential factors in public health, impacting the acceptance of eHealth technologies and providing a basis for improving healthcare delivery and technology adoption[269].

Similarly, attitudes, subjective norms, and perceived behavioural control are emphasised in Ajzen's (1991) TPB[270]. The TPB has been applied in various fields. For example, it successfully predicted pet owners' food choices in the field of animal health[271]. By enhancing its predictive power, the model has also shown its usefulness in predicting and interpreting health-related behaviours, with perceived behavioural control playing a significant role[272].

However, these models might overlook the complexities that are inherent to smallholder farmers. Decision-making in this context goes beyond rationality and is influenced by generational knowledge, community cohesion, and a strong bond with the land. When studying surveillance acceptance, socio-cultural factors must be considered, highlighting the need for a more comprehensive model that can accommodate these unique factors, in addition to the constructs proposed by TAM and TPB.

2.2.6.2 Global acceptance within the small-scale community

International studies have indicated that small-scale farmers generally approach surveillance technologies with a mix of caution and optimism[273], [274]. The main reason for this optimism stems from the potential advantages that these technologies provide, especially in terms of improving biosecurity and achieving collective benefits. These benefits greatly enhance the productivity and sustainability of small-scale farming, thereby contributing to food security and the development of rural areas. In continuation, additional studies within Southeast Asia have highlighted the importance of surveillance technologies in protecting community agricultural interests but have paralleled the narrative of the previous authors in regards to scepticism around long-term adoption[275]. The synthesis of this literature emphasises jhow despite the research being conducted at different time points and continents, the results are largely similar.

Nevertheless, there is an undertone of uneasiness stemming from worries about data protection and government monitoring. Literature shows that small-scale farmers hesitate to share their data out of concerns about misuse and exploitation[276]. Their concern is that too much government oversight could negatively affect their livelihoods by using their data in harmful ways. The caution indicates the necessity for better communication and stronger privacy measures to address these concerns without compromising surveillance integrity[277].

It is crucial to acknowledge that smallholder communities are not uniform, as highlighted in figure 4 . There is substantial diversity in their literacy levels, access to the internet/technology, and integration with markets. Across various smallholder communities, the acceptance of surveillance technologies differs significantly due to their heterogeneity, thus making any assumptions about technology acceptance across these groups would be inaccurate, as each community has distinct requirements, obstacles, and perspectives.

2.2.6.3 Willingness to Share Data

Trust in institutions is crucial for smallholders when deciding whether or not to share data[278]. The perception of trust in institutions encourages smallholders to share their information, as stated by research conducted Borthakur and Chhatpar (2021)[279]. The authors concluded that building trust involves transparent practices, open communication, and showcasing the benefits of data sharing in agriculture. By showcasing how data contributed by smallholders has contributed to the development of more effective disease management tools and expanded market opportunities, institutions can foster trust and ultimately enhance productivity and profitability.

The decision to share data depends heavily on the perceived direct benefits, with prominent examples of such notions being found within healthcare[280]. Farmers are more inclined to participate in data sharing programs if they believe they will gain direct benefits like better agricultural techniques, promotions, financial incentives, or increased market opportunities[281]. Highlighting the practical gains of data sharing in agriculture, including better disease management tools and wider market access, is essential for building trust and increasing productivity and profitability. This is in line with what other studies have found regarding the significance of trust in institutions for data sharing[282], [283].

2.2.6.4 Barriers to Acceptance

In the existing body of literature, several barriers to societal acceptance of surveillance are consistently identified and grouped into four key areas: privacy concerns, potential data misuse, transparency issues, and historical distrust of authorities[284]. These barriers are not separate but instead interconnected, often strengthening each other.

Firstly, concerns about privacy pose a major obstacle as people fear their personal data being accessed and abused. These concerns frequently result in hesitation or outright rejection of data-sharing initiatives[89]. Consequently, such worries are worsened by uncertainties of data misuse, where gathered information could be abused for identity theft, scams, or illicit activities.

The acceptance of surveillance is hindered by additional transparency problems. Mistrust and resistance are widespread due to a lack of clear, accessible information about data collection, use, and safeguarding. This scepticism is based on a history of distrust in authorities due to past incidents of surveillance overreach and abuse of power. Thus, historical contexts contribute to uncertainty and the reinforcement of public resistance towards surveillance practices.

Finally, surveillance overreach and data breaches, as reported by the media, intensify these concerns and resistance. Existing literature reviews suggest that farming communities may have a deeper scepticism towards data-sharing and surveillance than what is commonly recognised, especially small-scale individuals as a major incentive of adopting this lifestyle is to be “off-grid”[278].

2.2.6.5 Participatory approaches to improve acceptance

Roop et al. (2023) argue that including smallholder input in the development of surveillance systems can be advantageous, as suggested by their preliminary work conducted pre and post COVID-19[285]. However, there are concerns about the scalability and authenticity of these participatory approaches, as some critics point out[286]. Despite advocating for transparent practices, questions linger about the effectiveness of these strategies in diverse smallholder contexts, given the wide disparities in literacy and understanding of governance policies[287].

2.3 Social media as a research tool

Social media has become a ubiquitous part of modern society, with billions of people worldwide using platforms like Facebook, Twitter, Instagram, and LinkedIn to connect, share information, and express their opinions. As a result, it has also become a valuable tool for researchers looking to understand people's thoughts, behaviours, and attitudes[288].

One of the primary ways that it has been used as a research tool is by collecting and analysing large amounts of data from public posts, comments, and other online content. Researchers have used ML algorithms and NLP techniques to extract insights from this data, such as identifying patterns, themes, and trends in the topics being discussed[17]. This has provided valuable information about what people are thinking and feeling, what they are interested in, and how they are behaving.

Another way it has been used as a research tool is by conducting surveys and polls through platforms like Facebook and Twitter[288]. Researchers have targeted large numbers of users immediately and easily by disseminating questionnaires, and then be analysing this information to gain insights into specific topics or populations.

Table 4 below provides a glance at the use of social media in various domains in contemporary studies:

Title	Topic	Authors	Year	Social media source
"Digital and virtual spaces as sites of extension and advisory services research: social media, gaming, and digitally integrated and augmented advice”	Agriculture	Klerkx[289]	2021	Twitter/Facebook/Instagram
"Drivers and challenges of precision agriculture: a social media perspective"	Agriculture	Ofori & El-Gayar[290]	2021	Twitter
"Social Media Use for Health Purposes: Systematic Review”	Healthcare	Chen & Wang[291]	2021	Facebook/LinkedIn/Twitter/Instagram/Pinterest WeChat/Weibo
"Prevalence of Health Misinformation on Social Media: Systematic Review"	Healthcare	Suarez-LLedo & Alvarez-Galvez [88]	2021	Twitter/Facebook
"Using Social Media Marketing to Create Brand Awareness, Brand Image, and Brand Loyalty on Tourism Sector in Indonesia"	Marketing	Rimadias et al[292]	2021	TikTok
"Understanding the effect of social media marketing activities: The mediation of social identification, perceived value, and satisfaction”	Marketing	Chen & Lin[293]	2019	Twitter/Facebook

"A practical guide to analysing online support forums"	Psychology	Smedley & Coulson[294]	2018	Forums
"Dreaddit: A Reddit Dataset for Stress Analysis in Social Media"	Psychology	Turcan & McKeown[295]	2019	Reddit
"Identifying customer knowledge on social media through data analytics"	Data Science	He et al[296]	2018	Twitter
"A novel big data analytics framework for smart cities"	Data Science	Osman[297]	2019	Twitter
"Leveraging media and health communication strategies to overcome the COVID-19 infodemic"	Public Health	Mheidly & Fares [298]	2020	Twitter
"A New Application of Social Impact in Social Media for Overcoming Fake News in Health"	Public Health	Pulido et al[299]	2020	Reddit/Facebook/Twitter
"Integrating digital and field surveillance as complementary efforts to manage epidemic diseases of livestock: African swine fever as a case study"	Veterinary Epidemiology	Tizzani et al[300]	2021	Twitter
"The Assessment of twitter's potential for outbreak Detection: Avian Influenza Case Study"	Veterinary Epidemiology	Yousefghani, et al.[86]	2019	Twitter

TABLE 4: SUMMARY OF SOCIAL MEDIA AS A TOOL IN VARIOUS DOMAINS

These instances highlight the development and significance of social media as a research instrument in a range of fields and subjects, from marketing and psychology to public health and epidemiology.

2.4 Smallholdings and backyard farming

Backyard farming and smallholdings play a crucial role in agriculture, promoting biodiversity, food security, and sustainable development[39]. The cultivation of a small area of land for both commercial and subsistence production is a defining feature of these agricultural practices. Socio-economic, environmental, and cultural factors have influenced their evolution, resulting in diverse practices, motivations, and outcomes.

Bradley et al. (2021) conducted a study that examined the motivations of smallholders, highlighting the correlation between economic aspirations, lifestyle preferences, and a love of agriculture[301]. Likewise, similar studies indicate that many people participate in backyard farming due to ecological worries, the longing to be self-sufficient, and a preference for organic and locally grown produce[302].

Within Eastern Europe, backyard pigs were proven to be a major factor in the spread of African and Classical swine fever [303]. The lack of biosecurity tracing ability within smallholder pig farmers frequently presents a risk to the larger commercial farms, as pig movements and welfare practices are not as thoroughly documented with backyard keepers hence the cross-contamination of disease is highly possible [11].

Backyard producers possess a greater proclivity to vary from larger-scale commercial producers in the enforcement of biosecurity, which could be a consequence of deficiencies in awareness and overall information regarding proper techniques to manage their livestock[11]. Furthermore, an added dimension of risk is generated when there is a combination of varying livestock within the same holding, resulting in elevated levels of potential disease spread from inter-animal interactions[304].

It is difficult to apply the knowledge we currently have from 3rd world countries, wherein this type of subsistence farming is common, and attempt to employ this information to the British landscape[305]. The deficiency of information results in challenging evaluations of backyard livestock systems in relation to disease modelling and early intervention thus requiring a substantial effort to fill this chasm.

2.5 Conclusion

This evaluation of the literature provides a detailed synthesis of the multidisciplinary domains linked to this project, as well as an overview of how they interconnect to meet the proposed research questions. The structure was created to provide a comprehensive knowledge of the convergence of veterinary epidemiology, social media mining and data science, as well as to emphasise the gaps in the research that I aim to fill in the following chapters.

3/. Methodology

Based on the gaps in knowledge addressed in the previous chapter, this section presents an overview, justification and critique of the procedures and tactics employed to collect data from social media, as well as the research methodology used to evaluate and interpret the data. Consisting of web scraping, API calling, data storage and pre-processing in the initial stage, followed by text mining, social network analysis, user classification and time series analysis. Furthermore, ethical considerations and privacy concepts that influenced the data gathering procedure, such as informed permission, confidentiality and privacy are covered.

3.1 Introduction

Literature from the previous section indicates that with the digitalised world, smallholding communities may adopt social media as their main source of information pertaining to their livestock needs, advice and regulation, especially since the COVID-19 pandemic shifted the majority of communication networks to online. The open-source, publicly available nature of social media platforms provides a vessel for information exchange, and this catalogue of data can be extracted and analysed. Presented in this section is the methodology behind the data extraction and analysis of this project. Multiple sources of social media were used to target as wide of an audience as possible, as some smallholders prefer larger platforms such as Twitter, whereas others have a penchant for more niche platforms, namely specific livestock forums. A possible reason behind the two preferences could relate to wanting larger community reach, through the former, and the latter could pertain to wanting to maintain a smaller, close-knit network of like-minded individuals.

3.2 Study population

The project's study population are the small-scale farming communities in the UK, especially those engaged in raising pigs and poultry. The analysis had to take a more holistic approach with the study population due to sample size limitations, which were particularly apparent in the user classification and social network analysis chapters. In comparison, the forum scraping segment was specifically focused on the pig and poultry subsidiaries due to the abundance of data available in those forums for extraction and analysis.

The funnel plot in figure 11 illustrates the specific parameters of the study population, starting with the entire farming/agricultural population and narrowing down to smallholders, pig/poultry keepers, who are based in the UK and have a social media presence. These stringent parameters affected the sample sizes of the study; however measures were put in place to curtail such events. These included expanding the time scale of the data extraction, employing a greater number of search terms including abbreviations and colloquialisms, and using multiple data sources i.e. Twitter and forums.

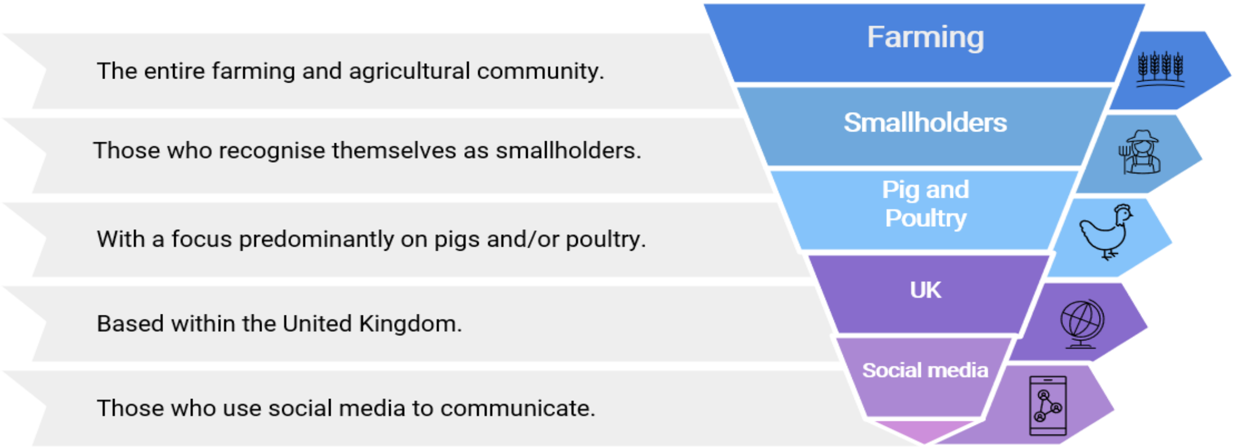

The entire farming and agricultural community.
Those who recognise themselves as smallholders.
With a focus predominantly on pigs and/or poultry.
Based within the United Kingdom.
Those who use social media to communicate.
Farming
Smallholders
Pig and Poultry
UK
Social media

FIGURE 11: STUDY POPULATION FUNNEL PLOT

3.2.1 Selection Criteria

Word-matching search parameters were implemented for the Twitter data scraping, with users being removed if the location field was empty or contained a location outside of the UK. For the user categorisation of smallholder profiles and social network analysis, all followers from the specific account were extracted, and once again the location filter was used. The forum scraping analysis employed the same location filter and selected every user within the pig and poultry forums. Finally, in the case of the spatio-temporal analysis, Avian flu related word matching was employed, a time filter between 01/01/2021 and 31/12/2022 was utilised and locations outside of the UK were filtered out.

3.2.2 Sampling

The data for this project was collected from two key online platforms, Twitter and a smallholder forum (www.accidentalsmallholder.net), using a purposive non-random sampling strategy. The chosen method was considered suitable because of the research questions put forward, which called for targeting particular groups of social media users engaged in discussions related to veterinary epidemiology[306]. The selection of the sample was based on specific criteria that aligned with the research objectives. Both manual and automated methods were used to extract relevant posts over a specific timeframe, adhering to ethical standards like data protection, anonymity, and consent, to carry out the data collection process. Although purposive non-random sampling can have limitations such as selection bias and limited generalisability, this approach captured rich, context-specific data and yielded invaluable insights into the research topic[307].

3.2.3 Justification of Twitter selection

Among the various data sources considered, Twitter was ultimately chosen as the primary one. One of the main reasons behind its popularity is because of a diverse user base, which includes smallholders, agricultural organisations, veterinarians, and other stakeholders in the farming community. With such diversity, the dataset becomes enriched and varied, capturing the essence of smallholding in the UK.

By employing Twitter's real-time nature, the smallholder community's most current and topical discussions can be captured, which is invaluable. Real-time data is crucial for the following chapters that focus on spatio-temporal analysis and disease outbreak detection, where capturing temporality is crucial.

Twitter's public API is capable of extracting historical tweets, as well as metadata such as the time of posting, the geographical location of the user, and the network of retweets and mentions, enabling the collection of significant amounts of data. With this added information, a more comprehensive understanding of online behaviour can be achieved.

Extracting useful data from Twitter discussions is a difficult task because they are often noisy and filled with off-topic comments. Despite the noise in the data, the advanced text mining, topic modelling, and social network analysis methodologies used in this research can extract valuable insights.

3.2.4 Justification of forum selection

The primary data source was selected based on the fact that the forum is considered to be the foremost online community for smallholdings in the UK. The website provides a wide variety of topics to explore, including animal husbandry, biosecurity measures, entrepreneurship, and livestock management. The forum had over 57,000 active participants at the time of this research, all exchanging ideas and information. Smallholder-centric information can be easily extracted from the platform, thanks to its engaged community and vast repository. On forums, you'll find more meaningful discussions about livestock-related topics, unlike Twitter where discussions can often be noisy and off-topic.

3.3 Research design

The research design encompasses the overall approach for a specific aspect of a study[308]. It involves four primary stages, namely, the selection of a research paradigm, methodology, strategy, and technique, which together form the framework for data collection and analysis[309].

3.3.1 Research approach

As this PhD deals with a combination of qualitative (i.e. thematic analysis, content analysis through text mining) and quantitative (e.g. statistics of tweet frequencies during avian influenza outbreaks) techniques, a mixed-methods approach was deemed to be the most apt.

A mixed research technique combines qualitative and quantitative research methodologies to create a more thorough comprehension of a study topic[310]. This includes gathering and evaluating numerical data via surveys, experiments, or other quantitative research methods, as well as qualitative data via interviews and focus groups. The findings from both data sets are then combined by the researcher to give a more thorough understanding of the study topic.

Social media data is naturally complicated and multidimensional, with networks creating massive volumes of data in a variety of formats, including text, photos, videos, and user-generated material. A mixed methods approach helps to capture this complexity by combining several data sources and analytic approaches and has the potential to assist with unearthing insights that might otherwise be overlooked if only one approach were used[311]. For example, quantitative analysis may assist in uncovering patterns and trends in social media data, but qualitative analysis can help offer context and interpretation.

3.3.2 Research method

Secondary data research is gathering and analysing data that has already been obtained by other researchers or organisations[312]. This form of study is frequently used to support primary research, which is gathering fresh data through surveys, trials, or other means.

It can be a valuable tool for research especially when conducting primary research is difficult or expensive. A variety of sources are included, namely social media, academic journals, publications, government data, corporate data, online database repositories and reports containing data related to the study issue are examples of published literature.

Social media falls under the rubric of secondary data[312]. Social media networks create massive volumes of data, such as text, images, and videos, which may be examined to learn about user behaviour, opinions, and trends. Yet, it is critical to examine the limits of such data, namely sample representativeness, bias, and data quality. While gathering and analysing social media data, researchers must also follow ethical requirements, such as gaining informed consent and respecting users' privacy.

3.4 Data collection

This project implemented two distinct methods to collect relevant data: web scraping of a livestock forum and activating the Twitter API to scrape historical tweets related to smallholdings, avian influenza and friend networks.

3.4.1 Web scraping

Online scraping, also known as web crawling, is the process of autonomously extracting data from websites using software[313]. This approach is very useful in current sectors like as Business Intelligence since it allows us to extract structured data from Hypertext Markup Language (HTML) text when it is not offered in machine-readable forms such as JavaScript Object Notation (JSON) or Extensible Markup Language (XML). Compared to manual data entry, using a web scraping program yields more thorough, accurate, and consistent data. As such, web scraping is an essential tool in modern fields, requiring multiple technologies such as spidering and pattern matching for proper implementation.

This was performed using a cascading style sheets (CSS) extractor tool in the Mozilla Firefox browser to retrieve forum data from https://www.accidentalsmallholder.net/, using a combination of R and Python. A more detailed explanation of the process is described in chapter 6.

Justification

The justification for opting to scrape such a forum is firstly that it is a publicly available forum and not closed off to non-members. Secondly, it allows for insights within each livestock forum, as individual subforums exist pertaining to advice/discussions. Specific "poultry and waterfowl" and "pig" subforums were deemed to contain accurate textual information related to solving the research questions of this project, as the main purpose of these forums is to converse about livestock issues, biosecurity, regulations etc thus reducing textual noise. An evident limitation to such an approach relates to bias and representativeness within this demographic, in addition to the influence of "key users" within these communities who are able to dominate and influence the narrative. The selection bias is alleviated as the main aim of this project is to examine one particular demographic, and the influence of key users was examined through network.

3.4.2 Twitter

Twitter's API is a web scraping service on a grander scale. It gives structured data to researchers by obtaining scraped data from their clients via API[314]. Until the recent introduction of APIs, which are unique interfaces meant to make communication simpler for apps and servers, web scraping was the only means for a computer to collect information from the internet[249].

Users on Twitter may interact with one another by subscribing to their updates or "following" them[182]. Users build networks such as follower networks, retweet networks, and mention networks through interactions such as reacting to tweets or passing interesting ones via retweets. Network analysis methods can be applied to study these networks to evaluate the effect of members and their social positions on the Twitter social network. Moreover, Twitter data is a rich source of information that can be mined using content analysis to uncover themes and user sentiment towards certain concerns.

Developer-level access was obtained through the platform, allowing researchers to extract historical tweets[315]. An application was made in October 2019 and academic research privileges were granted by Twitter, permitting for the extraction of ten million tweets per month, a query limitation of 1024 characters and a streaming rate of 50 requests every 15 minutes.

These permissions were activated through keyword search queries for Avian influenza and the extraction of user timelines and friend networks. Extracting location has become difficult in Twitter since the option for enabling co-ordinates was by default disabled, therefore free-text matching of the "location" field was performed to limit the tweets to within the UK.

Justification

The use of Twitter's API is ideal for this research because it allows comprehensive network analysis of user interactions such as retweets and mentions, revealing the influence and social positioning of different members. The platform’s rich content also supports effective content analysis, uncovering themes and user sentiment around Avian influenza. With developer-level access to historical tweets and network data, the API facilitates structured data collection on a grand scale. Keyword searches and timeline extractions enable targeted data gathering, while free-text location matching helps focus the study within the UK despite location-sharing restrictions. Therefore, the Twitter API is the most suitable method for this social media analysis.

3.4.3 Google trends

Google Trends may be used to uncover new trends and subjects that are increasing popularity over time[74]. This information can help researchers keep updated on contemporary pertinent topics and drive their research questions. Researchers may use this tool to examine the popularity of different search keywords or subjects over time, as well as to investigate geographical and demographic changes in search trends. Furthermore, by studying search trends connected to certain subjects or events, Google Trends may be utilised to undertake sentiment analysis and can harness insights into popular attitude and opinion on certain subjects. Moreover, another proponent is investigating correlations with data from other sources, such as social media or news items. This can provide a holistic picture of the link between search patterns and other factors.

Additionally, it was used for the descriptive analysis of avian flu related search terms within the UK, corresponding to the same time period as the twitter data extraction, depicting a visual aid of internet search results. The data was mined into CSV format and mapped against the time series avian influenza Twitter extraction to highlight similarities between Google searches and Twitter activity.

Justification

Using Google Trends is particularly valuable for this research because it offers insights into contemporary search trends, helping to identify emerging topics relevant to Avian influenza. By analysing the popularity of search terms over time, it provides a clear picture of public interest and sentiment. Geographic and demographic breakdowns of search trends add a deeper understanding of how interest varies across regions.

The data's compatibility with other sources, such as social media, enables correlation studies with Twitter activity. By mining the data into CSV format and mapping it against Avian influenza tweets over the same period, it offers a comprehensive, visual representation of the relationship between Google searches and Twitter trends. Thus, Google Trends is the optimal tool for revealing public attitudes and behaviour patterns in this study.

3.4.4 Data storage

Storing internet data may be a difficult task because the data is frequently large and requires specific infrastructure to manage[316]. Scalability, performance, affordability, and data security are all essential considerations when selecting a storage solution for internet data. It's also critical to make sure the storage solution works with the data analysis tools and platforms that will be utilised to process and analyse the data.

Table 5 provides an overview of data storage solutions:

Storage type	Tools	Description
Cloud	Amazon S3 Google Cloud storage Microsoft Azure	Provides scalable and cost-effective alternatives for storing massive volumes of Twitter data. These services may be set up to automatically scale storage capacity up and down as needed, and they can also include data management and analysis capabilities.
Relational database	MySQL PostgreSQL Microsoft SQL server	These databases provide great data integrity and are suitable for advanced querying and reporting.
NoSQL	MongoDB Cassandra Couchbase	Designed to handle massive amounts of unstructured or semi-structured data and are suitable for real-time data processing.
Open-source framework	Hadoop	Twitter's Hadoop-based data storage and processing technology is used to store and handle Twitter data.
Local storage	CSV Excel	For smaller-scale applications, local storage, such as hard discs or network-attached storage devices, can be utilised to store internet data. Local storage, on the other hand, may not be scalable or secure enough for larger-scale applications.

TABLE 5: DATA STORAGE OPTIONS

Due to the niche study population in this study, the collected data falls under the rubric of "small data". Textual information, in addition to metadata, was scraped from both the forum and Twitter, and stored into a CSV format in local storage and on the University's cloud server. Prospectively, a NoSQL database should be applied if a dynamic streaming system is adopted as a result of this work, as it is able to handle the unstructured nature of internet-based data.

3.5 Data analysis

The analytical methodology and the justification of adopting such methods is proposed below in table 6, with further explanations regarding the theoretical and practical aspects of each method is also described in this section:

Method	Application	Results context	Justification
Word2Vec	Text mining	Generating word embeddings based on co-occurrence patterns.	Co-occurrence patterns were used to generate word embeddings with Word2Vec. By converting words into vector representations, researchers can examine language usage among small-scale farmers on social media. Furthermore, it aids in identifying essential terms and topics by grasping the context in which words are employed, revealing farmers' concerns. When compared to alternative methods, it's proficiency in generating sophisticated and contextually aware embeddings makes it exceptionally well-suited for nuanced analysis. It is highly proficient in identifying connections between words that conventional frequency-based methods may overlook, offering valuable insights into the specific language and topics of interest within the agricultural sector.
Bigrams		Bigrams to display the most frequent term pairings.	Bigrams were selected because they capture the connections between pairs of words that follow each other, uncovering common phrases and expressions used by small-scale farmers on social media, which may be missed out in traditional unigram analysis.
LDA	Topic modeling	Unsupervised clustering of forum/Twitter data to uncover themes within the text	The selection of LDA topic modelling was due to its effectiveness in identifying themes and patterns in unstructured social media data by grouping words into topics based on co-occurrence. The concerns and issues of small-scale farmers are exposed through this. When compared to clustering, LDA offers a more nuanced insight into the underlying topics driving conversations, which is particularly beneficial for uncovering hidden themes in this community.
K-nearest neighbour (KNN) **Multinomial Naïve Bayes (MNB)** **Decision tree (DT)**	Text classification	User classification of Twitter profiles on manually coded data	Text classification was selected for this study to enable the systematic categorisation of unstructured data, specifically Twitter posts, into meaningful groups. Through manual coding of a dataset, text classification achieves accurate categorisation according to user-defined criteria. Manual coding outperforms automated clustering methods in accurately capturing the nuanced understanding of small-scale farming discourse, making it the superior approach for differentiating farmers' profiles and disease concerns. By aligning with

Logistic regression (LR) **Random Forest (RF)** **Bagging classifier (BGC)** **Support Vector classifier (SVC)**			research objectives, text classification can better capture the distinct aspects of these conversations. Seven popular classification algorithms were selected as their effectiveness has been proven in both epidemiological and social media data.
ARIMA/SARIMA	Time series	Overlapping with APHA confirmed avian influenza cases to measure seasonality and patterns.	ARIMA/SARIMA models were chosen for this study due to their capability in analysing time series data with seasonal patterns, which is crucial for comprehending the seasonality and trends of avian influenza. These models combine time series analysis with APHA-confirmed cases to uncover patterns and anomalies, shedding light on the disease's seasonal fluctuations. The flexibility of ARIMA/SARIMA sets it apart from other methods, as it can effectively model different types of seasonality and cyclic data.
Moran's I **Kernal Density Estimation (KDE)**	Spatial analysis	Measure spatial correlation between tweets and confirmed cases.	Moran's I and Kernel Density Estimation (KDE) were selected for their capability in evaluating spatial patterns and correlations in disease data, providing valuable insights into the geographic distribution of avian influenza-related conversations. Compared to other spatial analysis methods, these tools together offer a more precise understanding of spatial clustering and distribution patterns, enabling a clearer view of the disease's spread and the locations where farmers are most affected.
Pearson's correlation coefficient (PCC)	Correlation analysis	Measure correlation between confirmed cases and daily Twitter activity.	PCC was chosen for its ability to measure the strength and direction of the linear connection between avian influenza cases and daily Twitter activity. Through measuring the correlation, PCC reveals the degree of association between these variables, indicating if increased social media activity corresponds to an increase in confirmed disease cases. In terms of simplicity and sensitivity in detecting linear relationships, it is the optimal choice among other correlation metrics.

Isolation Forest	Outlier detection	Detect spikes in Twitter activity outside a 95% confidence interval calculated through expected averages.	The choice of Isolation Forest was based on its effectiveness in detecting outliers by isolating anomalies, making it well-suited for spotting spikes in Twitter activity. To identify unusual patterns within a 95% confidence interval, this approach utilises a tree-based method. By isolating these anomalies, it can pinpoint significant changes or spikes that may indicate a heightened interest in avian influenza Twitter searches. When compared to other outlier detection methods, this method has distinct advantages, such as efficient handling of high-dimensional data and minimal assumptions about the data distribution.
Djikstra's algorithm	Network analysis	Algorithm to find shortest path between two users in a Twitter network.	The reason Dijkstra's algorithm was chosen is that it is effective at finding the shortest path between users in a Twitter network. It's advantage over other pathfinding algorithms lies in its assurance of the shortest path in graphs with non-negative weights, making it ideal for comprehending the flow of information and the interconnectivity of users in the Smallholder Twitter network. It is highly valuable in identifying influential nodes that impact the spread of information on disease surveillance and other important subjects.
Louvain method	Community detection	Used to detect smaller communities within the Twitter network	The efficiency and accuracy of the Louvain method made it the ideal choice for detecting smaller communities in the Twitter network. When it comes to community detection algorithms, it is highly effective due to its ability to maximise modularity, allowing for the identification of meaningful sub-groups in large networks.

TABLE 6: DATA ANALYSIS SUMMARY

3.5.1 Text mining

3.5.1.1 Word2Vec

The core concept behind Word2Vec is to train a neural network on a vast quantity of text data to learn a vector representation for each word in a corpus of text[317]. The network has been trained to predict the likelihood of a particular word appearing in a given context (i.e., a window of words around the target word). It has been proven to capture semantic and syntactic links between words. Words that are semantically related (for example, "duck" and "chicken") have comparable vector representations in the Word2Vec space. As a result, the application has been shown to be an effective word embedding method for text classification, information retrieval, and language translation, in particular when applied to social media data[318].

3.5.1.2 Bigrams

Bigrams are n-grams in NLP that consist of two adjacent words in a text sequence[319]. Bigrams are often employed in language modelling and text analysis applications including machine translation, text classification, and sentiment analysis. It is formed in a text corpus by sliding a window of two words across the text. They yield the ability to capture some of the local context and interdependence between nearby words in a text.

In addition to bigrams, other forms of n-grams, such as trigrams (three-word sequences) or higher-order n-grams, can be employed in NLP. Nevertheless, as the length of n-grams rises, so does the number of potential combinations, making them computationally costly to utilise in some applications[320]. Bigrams were chosen as they can derive common adjacent pairings amongst the extracted text, which may illuminate potential insights into smallholding message behaviour.

3.5.2 Topic modeling

3.5.2.1 LDA

LDA is a prominent topic modelling approach in NLP that is used to detect latent themes in a text corpus[321]. The technique assumes that each text in the corpus is a composite of numerous topics, with each topic being a probability distribution over a collection of words. LDA's purpose is to find these latent subjects and the words connected with them.

The model starts with a fixed number of topics and then iteratively updates the topic-word distribution and document-topic distribution until convergence. Given the observed data, the model uses a Bayesian technique to estimate the posterior probability distribution of topics. This procedure entails giving a probability distribution across topics to each word in the corpus and then utilising this distribution to assign each word to a single subject[322].

This model was chosen as substantial amounts of literature have purported that it works well on social media data[323], and was used on both the forum and twitter datasets.

3.5.3 Text classification

3.5.3.1 K-nearest neighbour (KNN)

KNN is a basic yet effective ML technique used for classification and regression tasks[324]. The method compares a new data point to existing data points in a dataset and chooses the "k" closest data points based on some distance measure. The majority class or average value of the k-nearest neighbours is then used to forecast the class or value of the new data point.

The KNN algorithm is non-parametric, which means it makes no assumptions about the data's underlying distribution. It's also an instance-based learning method, meaning it doesn't build a model during training but rather the training data is saved and used to create predictions during runtime.

The key benefit of KNN is its ease of use and interpretability as it is resilient to noisy data, which is ideal for social media usage. Nevertheless, the choice of distance measure, the value of k, and the complexity of the data can all have an impact on KNN performance. Finally, It is also computationally intensive for large datasets since the distances between all pairs of data points must be calculated[325].

3.5.3.2 Multinomial Naïve Bayes (MNB)

MNB is a probabilistic model that computes the likelihood that a text belongs to a specific class based on the frequency of terms in the document[326]. Text classification applications such as spam filtering, sentiment analysis, and topic categorisation frequently employ MNB, as it is capable of handling a high number of features (i.e., words) and is computationally efficient, making it suited for huge datasets. It is highly resistant to overfitting, which may be an issue with other classification methods.

One weakness of MNB is that it assumes all characteristics are equally relevant, which is not always the case. It also presumes that characteristics are independent of one another, which may not be true for all NLP related tasks [327].

3.5.3.3 Decision tree (DT)

DTs are a tree-like model that displays a series of decisions and their potential outcomes, with each internal node of the tree reflecting a decision based on a certain data characteristic or attribute, and each leaf node representing a class label or numeric value[328].

An advantage to opting for a DT includes their interpretability and simplicity, as well as the handling of numerical and categorical data, missing values and outliers automatically. To increase their accuracy, decision trees can be expanded to ensemble approaches such as Random Forests and Gradient Boosting. On the other hand, DTs are prone to overfitting, which can result in poor generalisation to new data and are highly sensitive to tiny changes in the training data, which might result in the generation of distinct trees[329].

3.5.3.4 Logistic regression (LR)

LR is a ML and statistical approach that is used to solve binary classification issues by calculating the likelihood of an event happening as a function of one or more independent factors or attributes[330]. The logistic function, also known as the sigmoid function, is used by the method to transfer the linear combination of the characteristics and their related weights to a probability score between 0 and 1. The logistic function may be thought of as the probability of an event occurring.

Similarly to DT, LR provides a number of advantages, including its ease of use and interpretability. Furthermore, it can be modified to handle multiclass classification issues using approaches like one-vs-rest or SoftMax regression. Nevertheless, LR presupposes a linear connection between the independent and dependent variables, which is not necessarily the case in practise. It is also susceptible to outliers and feature collinearity, therefore is unsuitable for issues with extremely nonlinear decision boundaries.

3.5.3.5 Random Forest (RF)

RF is an ensemble learning system that uses several decision trees to increase forecast accuracy and robustness[331]. The technique generates a forest of decision trees, with each tree trained on a fraction of the original data and characteristics. The random selection of data and characteristics prevents overfitting and increases tree variety. During training, the method builds each tree by recursively splitting the data based on feature values, using a criterion like Gini impurity or entropy to identify the appropriate split at each node.

RF offers a number of benefits, including the ability to handle high-dimensional data with linked features, tolerance to noisy data and outliers, and resistance to overfitting. It also gives a measure of feature relevance, which may be used to choose and understand features. However, they can be computationally and memory-intensive, especially for big datasets with many characteristics. Interpreting the different trees and their contributions to the final forecast might also be difficult.

3.5.3.6 Bagging classifier (BGC)

Bagging (Bootstrap Aggregating) is an ensemble learning approach that use numerous base models to increase prediction accuracy and stability[332]. Bagging may be used with a number of different base models, including decision trees, logistic regression, and support vector machines.

The approach generates a number of bootstrapped samples from the original data, each of which is a random subset of the original data with replacement. Each base model is trained on a distinct bootstrapped sample of the data, yielding a diversified set of models capable of capturing various characteristics of the data.

Bagging has various advantages, including the potential to decrease overfitting and variance by integrating varied models, as well as its sensitivity to noise and outliers in the data. It may also be used to measure prediction uncertainty or confidence by estimating the variance of predictions across models. In contrast, it may not increase the performance of highly accurate or stable base models, and it may not be helpful for situations with strongly correlated features or limited datasets. Lastly, it may also be computationally and memory-intensive, particularly for big datasets with several models[333].

3.5.3.7 Support Vector classifier (SVC)

SVC, also known as the Support Vector Machine (SVM) for binary classification, is a ML technique that identifies a hyperplane in a high-dimensional space that optimally separates input points from distinct classes[334]. To convert the input characteristics into a higher-dimensional space where a linear decision boundary may be determined, the algorithm employs a technique known as kernel trick.

The hyperplane is chosen to maximise the margin between the two classes, which is defined as the distance between the hyperplane and the data points closest to the hyperplane in each class.

It has several advantages including the ability to handle nonlinear decision boundaries using kernel functions, data resilience to noise and outliers, and efficacy for high-dimensional data with few samples[335]. It can also offer a measure of feature relevance for use in feature selection and interpretation. However, the kernel function and its parameters can also have a substantial influence on algorithm performance and may need careful adjustment.

3.5.4 Time series analysis

3.5.4.1 ARIMA/SARIMA

ARIMA (Autoregressive Integrated Moving Average) and SARIMA (Seasonal Autoregressive Integrated Moving Average) are time series models that anticipate future values based on previous values[336].

ARIMA is a generalisation of the basic ARMA (Autoregressive Moving Average) model and is composed of three components: autoregressive (AR), integrated (I), and moving average (MA). The AR component models the current value's dependency on previous values, the MA component models the dependence on past mistakes, and the I component models the time series difference or trend. The parameters of ARIMA and SARIMA must be determined using model selection approaches such as autocorrelation and partial autocorrelation plots, as well as the Akaike Information Criterion (AIC). After the parameters are chosen, the model may be used to forecast future values of the time series[337].

They both offer a number of advantages, including the capacity to handle non-stationary and seasonal time series data as well as capture complicated patterns and relationships in the data. They are also often employed in a variety of sectors like as finance, economics, and meteorology.

Unfortunately, when the data is excessively irregular or noisy, these models may not perform well, and determining their parameters may be difficult and time-consuming. Moreover, these models are univariate, which means they only analyse one time series at a time and do not account for external factors or predictors that may influence the time series.

3.5.4.2 Pearson's Correlation Coefficient (PCC)

PCC is a measure of the linear connection between two variables, which are commonly designated as X and Y. It spans from -1 to +1 and indicates the degree to which X and Y are linearly connected, with +1 indicating a perfect positive linear relationship, -1 suggesting a perfect negative linear relationship, and 0 indicating no linear relationship[338].

Frequently used in statistics, data analysis, and ML to measure and evaluate assumptions regarding the link between two variables. It is used to find trends and patterns in data, as well as to analyse the strength and direction of links and their relevance.

It should be emphasised, however, that PCC assesses just the linear link between two variables and may not capture more complicated nonlinear correlations. Moreover, outliers or influential observations may impact the correlation coefficient, making it less robust in such instances. Finally, a strong correlation does not always imply causality between the two variables, hence the results need to be carefully interpreted[339].

3.5.5 Outlier detection

3.5.5.1 Isolation Forest

Isolation Forest is a ML approach for detecting outliers, operating by generating isolation trees, which are binary trees that divide data into two pieces at random feature values[340]. The number of splits necessary to isolate one instance, i.e., an outlier, from the rest of the data is then determined by the algorithm. The more splits that are necessary, the more separated the case is thought to be, and hence the more probable it is an anomaly.

Isolation Forest outperforms other outlier identification techniques in various ways. It is generally quick and scalable, can handle high-dimensional data, and is unaffected by irrelevant or duplicate characteristics. Likewise, it also does not require any data distribution assumptions and can handle both linear and nonlinear interactions.

Conversely, these models may not perform well when the data contains many comparable examples, as it might be difficult to separate such instances from one another. Also, when the number of outliers are relatively low in comparison to the size of the dataset, the approach may be ineffective[341].

3.5.6 Spatio-temporal analysis

3.5.6.1 Moran's I statistic

Bivariate Moran's I is a statistical measure used to analyse any correlations between two spatial datasets[342]. This test is a variation of the univariate Moran's I, which takes a single dataset and assesses whether spatial structures are present. Spatial autocorrelation evaluates the similarity or difference between neighbouring observations in a geographic space.

The statistic varies from -1 to 1. A measurement close to 0 suggests no correlation between the spaces, while a positive number implies a positive correlation, and a negative number implies a negative correlation. In a positive spatial correlation, large numbers of one variable are connected to high values of the other variable in adjacent locations, and the same is true for smaller numbers. A negative spatial correlation, however, implies that large values of one variable are paired with small values of the other variable in close locations, and vice versa.

In the context of epidemiology, Moran’s' I has been applied in global studies examining the spatial autocorrelation of hand, foot and mouth disease in China[343], exploring Visceral leishmaniasis in Iran[344] and mapping Lyme disease cases to user tweets in the UK[345].

3.5.7 Network analysis

3.5.7.1 Dijkstra’s algorithm

This is a greedy algorithm which guarantees the shortest path being found, which is relevant to Twitter networks as they can become extremely large and chaotic[346]. Starting from a source node, the method gradually explores the nearby nodes to discover the shortest path to all other nodes in the graph. Furthermore, it keeps a priority queue of nodes to examine, with the source node having the lowest priority. By choosing the node with the lowest priority at each step, it changes the priorities of its neighbours based on the edge weights. This step is repeated by the algorithm until all nodes have been investigated.

3.5.8 Community detection

3.5.7.1 Louvain method

The Louvain method is a well-known community discovery approach for identifying communities or clusters in a network[347]. The approach is based on maximising a modularity measure, which determines how well nodes within a community are related to nodes in other communities. It may be used to detect groups of people that have similar interests or habits, clusters of genes that perform similar tasks, or subjects in a document corpus.

The Louvain approach outperforms alternative community discovery techniques in various ways[348]. It is quick and scalable, capable of handling massive networks, and capable of detecting communities of various sizes and densities. Therefore, no prior knowledge of the network's topology or the number of communities is required.

3.6 Ethical considerations

Ethical approval for this project was obtained by both The University of Stirling and SRUC prior to the commencement of the project on 1st October 2019, which covered all the relevant publication and experimental chapters succeeding this chapter. All the data has been anonymised and individual users nor their precise locations are discoverable from the subsequent analysis. This thesis is compliant with the research ethics and integrity guidelines set out in https://www.stir.ac.uk/research/research-ethics-and-integrity/.

Compliance with ethical standards required the implementation of rigorous measures due to the ethical concerns surrounding web scraping and social media mining, including privacy, consent, and fair use. These measures were put in place to protect personal data, prevent the abuse of publicly accessible information. .

Ethical recommendations by Taylor et al (2018) were proposed in an in-depth study covering guideline adherence within the UK[316]. The authors stated that the general lack of integration of ethical concepts and guidelines in the corpus of published articles reviewed suggests a lack of awareness among researchers using social media mining in their studies, echoing observations from other areas of 'big data' research. This is consistent with the vast range of ethical recommendations provided by Research councils UK (RCUK) members in respect to the use of social media platforms and the data obtained from them. Given the extremely multidisciplinary character of studies in this field, they discovered that just one RCUK body (Economic and social research council/ESRC) explicitly referred to social media research in its ethical advice, as evidenced by an examination of relevant health-related publications.

3.7 Data validity and reliability

The consistency and stability of the measurements or observations employed in the research are referred to as data reliability[349]. To put it another way, if the research were repeated, would the results be consistent, or would they change greatly? Researchers may utilise approaches such as test-retest reliability, inter-rater reliability, or internal consistency to verify data dependability. The process of repeating a measurement at different times to check if the findings are consistent is known as test-retest reliability. In addition, the process of comparing measures collected by various raters to see if they are consistent is known as inter-rater reliability. The consistency of replies within a series of questions that are designed to test the same concept is referred to as internal consistency.

In contrast, the accuracy and validity of measurements or observations used in research is called data validity[349]. The validity of a study pertains to whether it measures what it purports to measure. Validity is grouped into different forms, including content validity, construct validity, criteria validity, and ecological validity. Research content validity refers to whether all relevant features of the research question are assessed. The construct validity of a study refers to whether it evaluates the theoretical concepts it is designed to test. Compared to external standards, the validity of a standard refers to whether the study measures what it purports to measure. Finally, the generalisability of results to real conditions is referred to as ecological plausibility.

Data reliability and validity are crucial factors when conducting social media research to guarantee that the results correctly represent the activities and views of users[350]. The dynamic nature of social media platforms, shifting user bases, and the enormous volume of data created can all have an impact on data dependability in social media research. Data dependability may be established by ensuring that data gathering procedures are regularly standardised, through means of searching for comparable information with the same keywords or hashtags, utilising the same data collecting techniques, and training programmers to guarantee consistent data coding are all examples of this[351].

Similarly, many factors can influence data validity in social media research, including the representativeness of the sample, the correctness of the data gathered, and the assessment of constructs[351]. Researchers should verify that their sample of social media users is representative of the population they are investigating in order to demonstrate data validity.

Within this project, this was achieved through constructing a thorough keyword dictionary depending on the domain of the task, i.e. Avian influenza related keywords (e.g. "bird flu", H5N1”). Coupled with explicitly denoted date and location filters, the code used to extract the data can be replicated by another researcher whilst achieving the same results. The user classification experimental analysis in the following chapter is the most prominent example of sound data validity methods being utilised in this project. Intercoder agreement /inter-coder reliability was adopted on the smallholding twitter database by 2 coders to ensure consistency in the classification of smallholding profiles. It can be defined is a measure of the consistency or agreement in classifying or categorising data between two or more coders (or raters)[352]. It is a statistical metric that evaluates how well various coders agree on the same coding rules and criteria while studying the same data set, as measured by Cohen's Kappa for this project[353]. As the manual coding of free text is subjective and open to bias, this method mitigates most of this by quantifying the agreement between 0 (no agreement) and 1 (perfect agreement).

To further boost the validity of the results from chapter 4, an unseen dataset was fed through the predictive algorithm to ascertain the effectiveness of the model to classify new Twitter profiles and build a framework for creating an ongoing smallholder database through this process.

The overall validity of the data will be examined in the discussion chapter following the results, as one of the main research questions pertains to whether social media data can be used as a supplement to traditional surveillance data sources, therefore cross-referencing the findings with established results from APHA can illuminate such answers.

3.8 Conclusion

The underlying philosophical ideas and methodological issues that underpin the research are discussed in this chapter. The pragmatic research paradigm is used to grasp the study phenomena, which is followed by a description of the research methodology, strategy, and data collection methodologies. Web scraping of a livestock forum and API calls were used to collect data from Twitter, prior to being analysed through a combination of text mining, topic modelling, time series analysis and network analysis. This sets the theoretical tone for the experimental chapters following this section.

4/. Typifying smallholders

The first experimental chapter presents the findings from the submitted paper entitled *"Text classification of UK smallholding communities through Twitter"*. This chapter sought to build a classification algorithm which was able to differentiate between a smallholding and non-smallholding Twitter account by:

1. Leveraging text classification techniques on user profiles, this section aimed to classify users based on their interests and engagements regarding smallholding activities.
2. Using this model to extract smallholder users from new Twitter profiles and validate the efficacy of the predictive model
3. The results have considerable implications for appreciating small holder farmers' demographics and for the construction of a database of identified users.

The key findings from this chapter were:

- Twitter data was successfully used to accurately identify and categorise smallholder farmers in the UK via a robust manually trained dataset.
- Machine learning algorithms were able to use the 160-character profile description to achieve a high degree of accuracy. Logistic regression proved to yield the best predictive accuracy.
- The successful navigation of several classification algorithms was helped by the manually sorted training dataset.
- Collecting and compiling information to create a database for potential disease surveillance and control was achieved.
- User location data was extracted.
- When dealing with smaller datasets, over-sampling provided the best results.
- Importance of considering the performance of different models under various sampling adjustments.
- The model exhibits great accuracy in identifying non-smallholders and demonstrates satisfactory performance in detecting smallholders.

4.1 Introduction

Building upon the paper, "User Categorisation of UK Smallholding Communities Through Twitter," this chapter seeks to distinguish and sort smallholding farmers in the UK with the use of Twitter data. This analysis shows how using social media data can be used to accurately divide users into distinct communities based on their profile descriptions. Carefully labelling each profile description with relevant annotations enabled the training dataset to be used with machine learning algorithms, which delivered impressive accuracy rates in identification, thus proving the efficacy of merely using the 160-character profile description.

Besides achieving its primary goal, this chapter also helps create a sizable database of smallholder accounts that can be developed for disease surveillance and control. Through collecting user data from farming-related accounts in the UK and running the profile data through the trained models, the model has the ability to classify users as smallholders or non-smallholders. By extracting and mapping the location field from user profiles, which users enter as a free-text option, hotspots of potential disease outbreaks can be identified.

The analysis highlights the disparity of classes in the dataset, and this can be corrected via either under-sampling or over-sampling techniques. When dealing with smaller datasets, over-sampling is the preferred option, yet it is essential to analyse the performance of various models under different sampling adjustments.

4.2 Methodology

The process of the methodology included several important steps, such as gathering data, categorising users, assessing inter-rater agreement, preparing the data, conducting the analysis, selecting features, balancing the dataset, and finally classifying the text.

4.2.1 Data Collection

The process of data collection started with the identification of an appropriate data source. Twitter was chosen due to its broad user base and the accessibility of public data through its Application Programming Interface (API). With the enhanced privileges granted by the developer-level academic account, it was possible to acquire permission to scrape the profile and locations of numerous users. Python was the chosen programming language for gathering the data, with the use of the Tweepy package to link with the Twitter API.

4.2.2 Accessing the Twitter API

Access to the Twitter API was successfully achieved by obtaining a developer-level account. Access to this account enabled the collection of a variety of data from the platform, such as user profile information and the followers they had. The Tweepy package in python was used to communicate with the API, enabling the extraction of up to 1,000 followers per request. A 15 second delay was implemented to prevent the server from being overburdened and comply with the web scraping rules established by Twitter. Furthermore, the API Key, tokens, consumer key and consumer secret tokens were employed in the data collection process.

4.2.3 Building a training dataset

The formation of the initial dataset was accomplished through the synthesis of the 953 members of a notable UK-based smallholder's network (@asmallholder), with details such as the profile name, location, and description for each user included. The user profiles were sorted according to the data in the description field, with profiles containing terms such as "smallholder" or similar being classified as Smallholders (1), and the rest classified as Not Smallholders (0). On occasions, usernames that contained applicable terms were also considered, hence forming a dataset that incorporated 953 distinct users with the allotted tags. To ensure the accuracy of the training dataset, data cleaning was conducted by eliminating profiles with blank descriptions, decreasing the total count of profiles to 774, with smallholders making up 26% of the dataset. Two manual coders classified the training dataset, and the inter-rater agreement was gauged based on this. The dataset structure resembled the example in table 7:

Username	Location	Description	Classification
			1
			0

TABLE 7: TRAINING DATASET STRUCTURE

The username, location and description columns were scraped directly from the profile, with the location column representing the free text location option rather than the now obsolete precise geolocation co-ordinates metadata. A binary classification of 0 or 1 was used for the classification column, where 0 indicated a non-smallholder and 1 a smallholder.

4.2.4 Exclusion of Incomplete Profiles

The inclusion of the description column is essential to this study. Subsequently, users with empty profiles were excluded from the analysis, resulting in a decrease of the aggregate figure of users to 774. Smallholders (coded as 1) accounted for 26% of the remaining users.

4.2.5 Inter-rater Agreement

To appraise the consistency and dependability of the manual coding process, the agreement between the two coders was determined using Cohen's Kappa statistic. This statistical calculation evaluates the real-world agreement between the coders against what would be conceivable by mere chance. This study yielded a Kappa statistic of 0.87, signifying a substantial degree of consensus among the coders. This consensus lends validity to the partitioning and implies that the labels assigned to user profiles are uniform and trustworthy for subsequent examination.

4.2.6 Data Pre-processing

Upon the completion of the annotation procedure, the data was imported into Python for pre-processing. Pre-processing is a fundamental step to ensure that the data is in an appropriate format for analysis and to eliminate noise or irrelevant information. Social media data typically contains non-alphanumeric elements, such as emojis and symbols, which necessitates a comprehensive cleanse prior to analysis. The full list of stop words is displayed in Appendix A.

The Natural Language Toolkit (NLTK) was used, and the profile descriptions experienced the following pre-processing measures explained in table 8:

Pre-processing measure	Description and justification
Excluding non-characters, punctuation, and numbers	This step is critical to avert any unnecessary characters that could detrimentally affect the analysis and to concentrate solely on the textual material. An example of this are emojis often found within user profiles.
Transforming to lowercase	Through transforming all words to lowercase, the analysis becomes case-insensitive, protecting that words aren't treated as separate entities because of distinctions in capitalisation. Usernames often contain a random mix of lower- and uppercase words as users wish to create unique usernames
Deleting URL links	URLs found in the data are discarded in order to protect the accuracy of the results as they do not contribute any pertinent information for

	text analysis. Business owners and academics sometimes embed their business/personal websites into the profile descriptions.
Exclusion of Stopwords	Stopwords are recurrent words that have no consequential importance when investigating text. Removing them helps to reduce the dimensionality of the data and allows for a more focused analysis on relevant words. In this context, custom stopwords were also applied to further remove noise within the text.
Eliminating white space	Surplus white space within the text can trigger variations during assessment. Removing unnecessary spaces ensures a uniform and clean dataset.
Stemming	This process entails reducing words to their basic form, permitting the integration of different variants of the same word, which boosts the performance and precision of the analysis. An example found with livestock text is the root word "feed", which stems from "feeding" and "feeds".
Lemmatisation	Like stemming, this procedure shortens words to their base form but factors in the morphological analysis of the words. For instance, in animal husbandry text, the verb "grazing" could be lemmatized to "graze," and the collective noun "chickens" could be lemmatized to its singular form, "chicken."
Standardisation	Text standardisation/normalisation is converting text into a consistent format, guaranteeing uniformity in the illustration of words and phrases.

TABLE 8: TEXT PRE-PROCESSING WORKFLOW

4.2.7 Data Analysis

Descriptive statistics and word clouds were created for the smallholding and non-smallholding groups to map out the most frequent terms in each cohort. This step enabled the recognition of significant themes and patterns among the groups.

4.2.8 Feature Selection

TF-IDF and Word2Vec vectorizers were both applied to conduct feature selection. The first option was chosen for its capacity to show text information by allotting weights to words depending on their value within the document and the whole collection. The data set was initially examined to count each individual word and rank them by the most frequent to least frequent. By creating a word-to-index mapping, the text instances were transformed into numerical vectors. The TF-IDF vectorisation method was used to create an alternative set of features, in which words with higher frequencies were weighted more heavily. Through this dual approach, the performance of various feature representations could be measured and compared, ultimately leading to the selection of the most appropriate features for this classification task.

The TF-IDF vectoriser was selected as the best option due to its effectiveness in capturing the complexity of text data. By assigning a value to each word based on its importance within the document and the entire dataset, this approach was able to determine the most relevant words and compress the dataset for more effective classification.

4.2.9 Dataset Balancing

As the smallholding class only accounted for 26% of the total frequency, two approaches were used to address this disparity. First, NearMiss (version 1) was used to under-sample the majority label (0), which selects samples from the dominant class with the shortest average distance to the three closest minority label examples. This method guarantees that the under-sampled data set retains some of its original format while lessening the imbalance. Secondly, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to boost the minority class (1) by replicating some of the existing examples and supplementing the dataset. This technique enhances the classifier's effectiveness by boosting the representation of the minor class.

4.2.10 Text Classification

The dataset was divided into an 85:15 ratio for training and testing respectively, and seven classification models were employed. These were:

1. **K-nearest Neighbour (KNN)**
2. **Multinomial Naïve Bayes (MNB)**

3. **Decision Tree (DT)**
4. **Logistic Regression (LR)**
5. **Random Forest (RF)**
6. **Bagging Classifier (BGC)**
7. **Support Vector Classifier (SVC)**

The selection of these models was based on their diversified underlying mechanisms, as addressed in chapter 3, and the possibility of detecting distinct patterns in the dataset. Through analysing the performance of various models via evaluation metrics, the research sought to find the most suitable classifier.

4.2.11 Hyperparameter Tuning

Hyperparameter tuning was performed using the grid search method to optimise the input parameters to enable the best performance for each classification model. This process went through multiple parameter combinations, which made it possible to identify the most effective parameter settings for each algorithm. Adjusting the parameters to the ideal configuration can drastically enhance the model's capability and its generalisation performance[354].

4.2.12 Performance Metrics

Several metrics, including accuracy, precision, recall, and the F1-score, were used to assess the performance of each classification model. These metrics offered insight into the model’s capacity to accurately classify users and assisted in deciding the best model for this study.

4.2.13 Area under the ROC Curve (AUC)

To measure the models' proficiency in distinguishing between the two classes, an area under the Receiver Operating Characteristic (ROC) curve was plotted. The ROC curve is a graphical comparison of the true positive rate (sensitivity) and false positive rate (1-specificity) at different threshold points. When the AUC value is greater, the model is better at distinguishing between the classes, offering another way to measure the performance of a classifier[355].

4.3 Results

4.3.1 Descriptive statistics

Descriptive statistics in figure 12 highlight that only a quarter (26%) of user profiles were deemed to be from smallholders, while the other 74% were from non-smallholders. 953 distinct users were categorised using the description provided in their profile, and on occasions, the textual information within the username was also considered within the training dataset. As incorporating the description column was a vital element to this project, any followers who left this blank were removed from the analysis, reducing the total to 774; 201 smallholders and 573 non-smallholding accounts. This suggests that a varied community in regard to user demographics is active around the account, and it seems to entice not only smallholders but also individuals and organisations from different backgrounds, interests, or roles in agriculture and related domains.

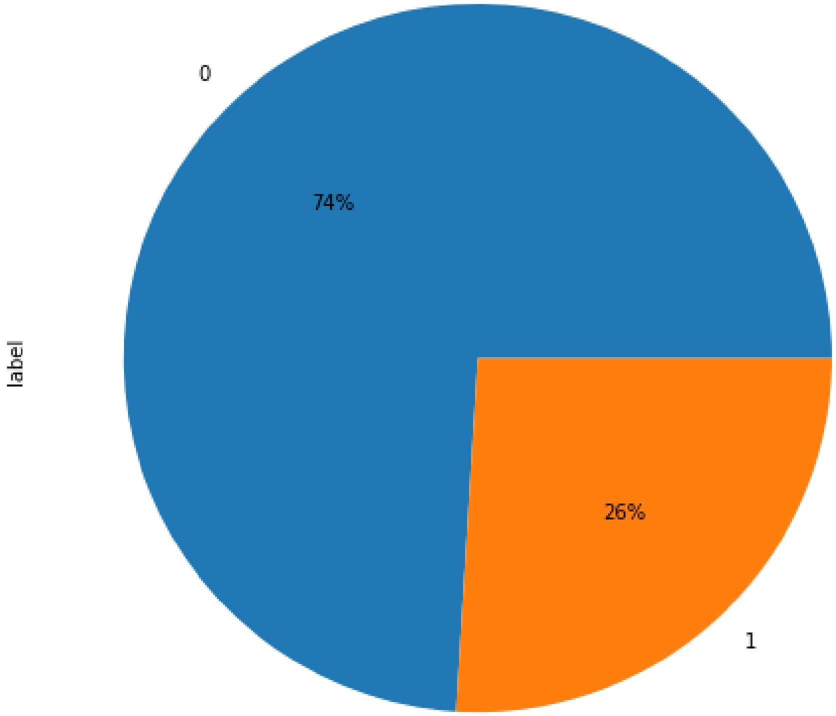


FIGURE 12: PIE CHART OF LABEL DISTRIBUTION

The Venn diagram in figure 13 provides a visual representation of the unique terms found in the two cohorts, as well as the terms that are common to both categories. It can be observed that there are distinct terms exclusive to each category, which might indicate the differences in language or topics discussed the contrasting profiles. The size of each circle corresponds to the number of unique terms, with the larger non-smallholding cohort possessing 2303 terms, compared to only 625 from the smallholding cohort.

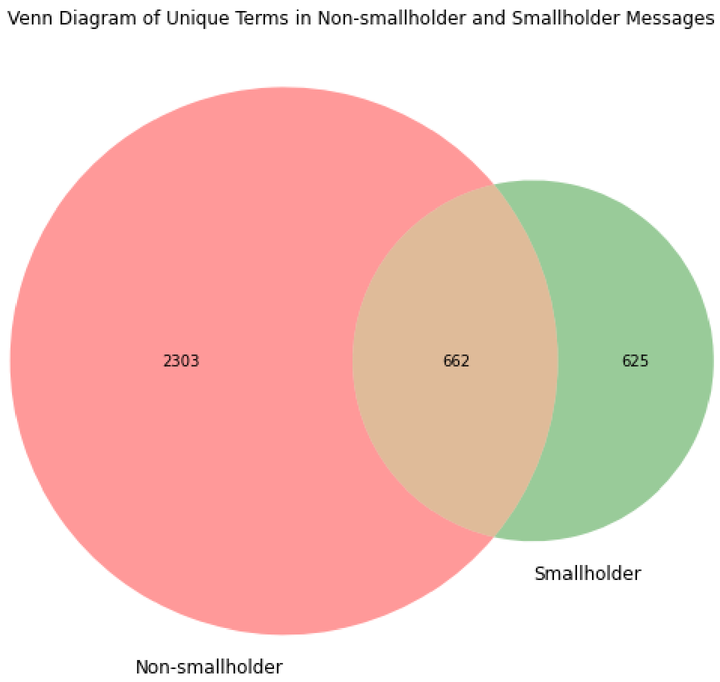


FIGURE 13: VENN DIAGRAM OF PROFILES

A deeper dive into these differences is shown in table 9, portrays the top 5 terms in each respective category, providing insights into the key topics or concepts that characterise each group.

Non-smallholder	Intersection	Smallholder
product	farm	flock
leading	food	coloured
market	sheep	ryelands
bird	smallholder	herd
environment	farmer	regional

TABLE 9: TOP 5 TERMS BY COHORT

Among non-smallholders, the most prominent terms were "product", "market,", "environment" and "sustainability". This shows their interest in the market, environmental protection, and sustainable practices by looking into various terms, reading news, and subscribing to magazines. Additionally, they focus on training and education, likely with the purpose of refining their agricultural techniques or gaining additional knowledge in related fields.

By contrast, the key terms for the smallholder category were "flock" "ryelands" "herd" "regional" "hobby" and "holding”. These terms imply that they prioritise particular breeds, animal husbandry, and local aspects of small-scale farming. The terms point to a focus on local and regional breed varieties, as well as the recreational side of farming on a small scale.

Finally, the intersection reveals terms that both cohorts have in common, demonstrating the shared interests and concerns of the two groups. These terms comprise "farm", "food", "sheep", "smallholder", "farmer” and "breed". This suggests a communal interest in farming procedures, livestock rearing, and food production between both cohorts. The shared terms allude to the commonality of the life and experiences of smallholding farmers in the UK, as well as the broader community.

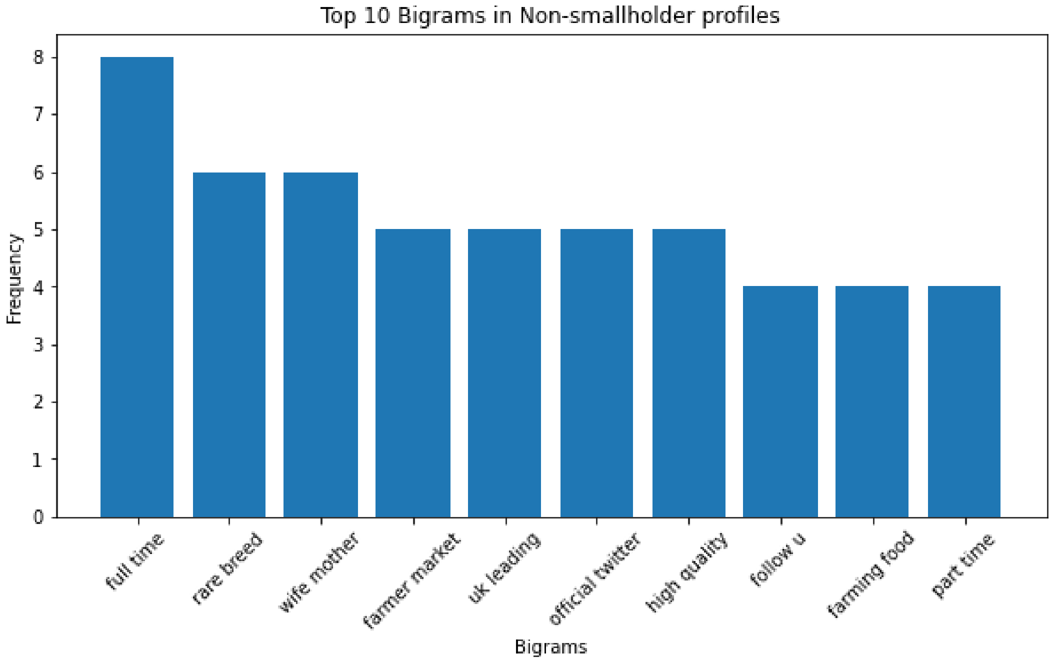


FIGURE 14: NON-SMALLHOLDER BIGRAMS

Figures 14 and 15's bigram analysis uncovers the most frequent pairs of consecutive words in both profile groups, delivering knowledge into widespread phrases, expressions, and topics for each group. Non-smallholder bigrams are associated with farming markets and food production, whereas smallholder bigrams allude to small-scale agriculture and livestock rearing.

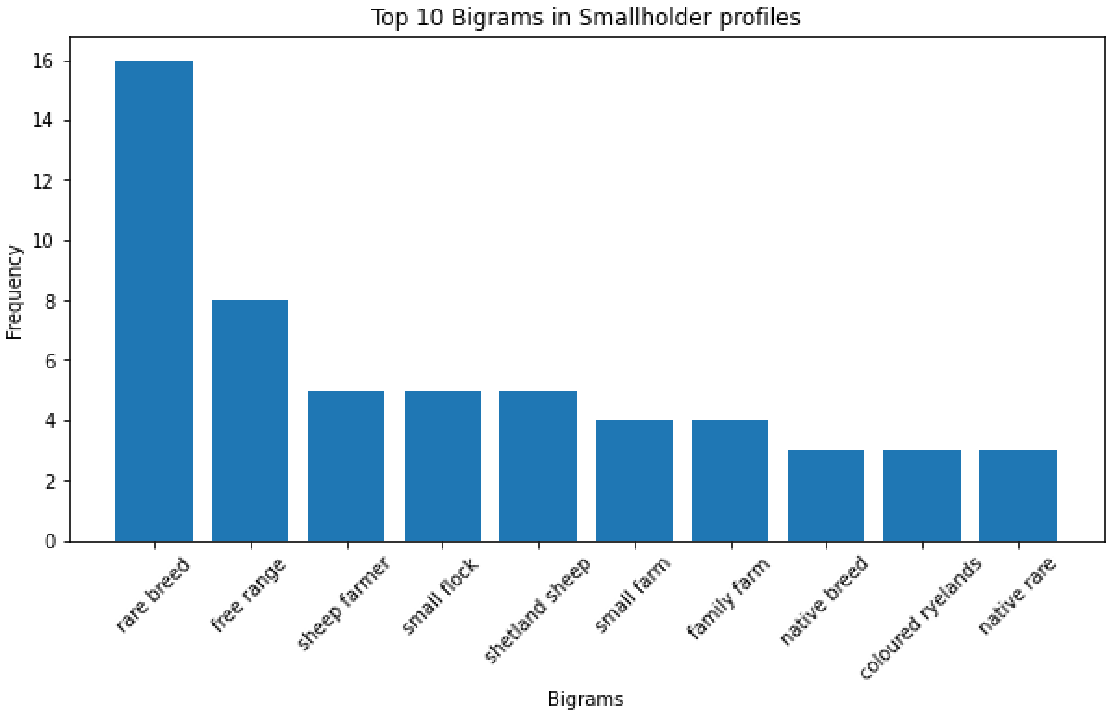


FIGURE 15: SMALLHOLDER BIGRAMS

4.3.2 Inter-rater agreement

The inter-rater agreement was measured by four separate metrics: Cohen's Kappa[353], Fleiss' Kappa[356], Krippendorff's Alpha[357], and Scott's Pi[358]. Cohen's Kappa and Scott's Pi are both measures of agreement between two raters, while Fleiss' Kappa is a variation of Cohen's Kappa that applies to more than two raters. Conversely, Krippendorff's Alpha is a more comprehensive metric, considering different types of data, scales, and the number of raters. Examining the values acquired from the agreement coefficients gave a comprehensive understanding of the stability and accuracy of the raters' assessments, which further confirmed the validity of the examination.

Measure	Value
Kappa	0.8712921959298566
Fleiss	0.8712921959298566
Alpha	0.8710287766708
Scotts	0.870961075241231

TABLE 10: INTERRATER AGREEMENT

The inter-rater agreement analysis shown in table 10, revealed that the two raters had a high level of consistency between them, suggesting a strong reliability in their classification decisions. Cohen's Kappa, Fleiss' Kappa, and Krippendorff's Alpha all yielded nearly identical results, with Kappa at 0.8713, Fleiss at 0.8713, and Alpha at 0.8710. These outcomes demonstrate that the raters had a large amount of agreement, indicating a strong degree of harmony in their categorisations. In addition, Scott's Pi, an alternative measure of agreement, generated a value of 0.8710, which is in close alignment with the other three measures. The similarity in the values of these four agreement coefficients is evidence that the raters' categorisations are reliable and gives weight to the following analysis based on their decisions.

4.3.3 Dataset Balancing

The NearMiss method and SMOTE were both used in order to mitigate the disproportion between smallholders and non-smallholders. Applying these methods created a balanced dataset, leading to more precise and dependable classification outcomes. Table 11 demonstrates the distribution of each dataset after the dataset balancing was complete.

Imbalanced		Undersampling		Oversampling	
Smallholder	Non-smallholder	Smallholder	Non-Smallholder	Smallholder	Non-smallholder
201	573	201	201	573	573

TABLE 11: DATASET FREQUENCY

Table 11 presents the two methods for addressing the class imbalance in a dataset: Under-sampling, and Over-sampling.

By employing the under-sampling method, the dataset is balanced by diminishing the larger class (non-smallholder) to the same size as the smaller class (smallholder), thus resulting in 201 instances for both categories. Addressing the class imbalance concern through this approach could also result in the loss of crucial information from the larger class as several instances are eliminated from the dataset.

On the contrary, the over-sampling approach makes the dataset equal by increasing the minor class (smallholder) to equal the major class (non-smallholder), hence resulting in 573 instances for both classes. This method takes care of the class imbalance problem without having to sacrifice information from the larger class. This can lead to some redundancy or over-fitting in the dataset, as new instances are generated by copying or interpolating from the less populated class.

4.3.4 Text classification

Seven classification models were trained on the 85:15 training/testing split dataset, and their performances were compared using various metrics, including accuracy, precision, recall and F1-score. The results of the classification models are summarized in tables 12-14.

Model	Accuracy	Precision	Recall	F1-score
KNN	0.83	0.79	0.68	0.71
MNB	0.80	0.72	0.73	0.73
DT	0.81	0.74	0.77	0.75
LR	0.84	0.82	0.69	0.72
RF	0.81	0.76	0.64	0.67
BGC	0.80	0.72	0.70	0.71
SVC	0.85	0.83	0.71	0.74

TABLE 12: PERFORMANCE METRICS FOR IMBALANCED DATASET

Model	Accuracy	Precision	Recall	F1-score
KNN	0.74	0.66	0.69	0.67
MNB	0.70	0.65	0.70	0.65
DT	0.81	0.74	0.77	0.75
LR	0.83	0.76	0.79	0.77
RF	0.84	0.77	0.78	0.77
BGC	0.81	0.74	0.75	0.74
SVC	0.81	0.74	0.75	0.74

TABLE 13: PERFORMANCE METRICS FOR UNDERSAMPLED DATASET

Model	Accuracy	Precision	Recall	F1-score
KNN	0.23	0.12	0.50	0.19
MNB	0.73	0.67	0.72	0.67
DT	0.80	0.73	0.76	0.74
LR	0.86	0.83	0.76	0.78
RF	0.84	0.78	0.74	0.75
BGC	0.83	0.76	0.77	0.77
SVC	0.86	0.82	0.77	0.79

TABLE 14: PERFORMANCE METRICS FOR OVERSAMPLED DATASET

The SVC model was the most successful for the Imbalanced dataset, with an accuracy of 0.85 and an F1-score of 0.74. LR followed with an accuracy of 0.84 and an F1-score of 0.72. Compared to the others, The KNN model had the lowest F1-score (0.71) yet still yielded an accuracy of 0.83.

When the under-sampled dataset was applied, the LR and RF models delivered enhanced performance, with accuracy ratings of 0.83 and 0.84, respectively, and both obtaining F1-scores of 0.77. The KNN model witnessed a sharp decrease in performance, with an accuracy rating of 0.74 and an F1-score of 0.67.

Finally, the LR and SVC models for the over-sampled dataset had the highest accuracy of 0.86, surpassing the other models. The SVC model has the highest F1-score (0.79), followed by the LR model with an F1-score of 0.78. The KNN model suffered a significant drop in performance on the over-sampled dataset, with an accuracy of 0.23 and an F1-score of 0.19.

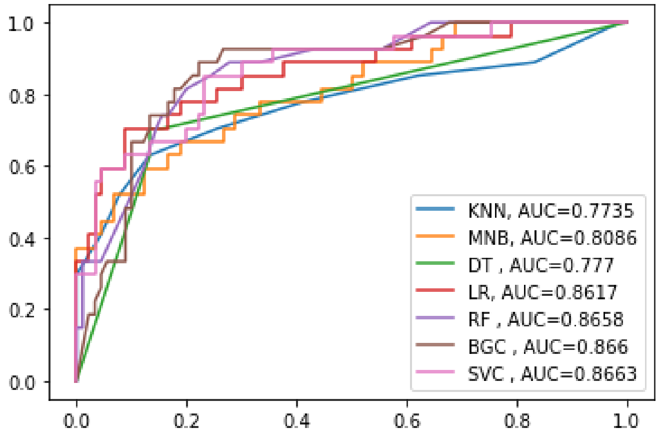


FIGURE 16: AUC IMBALANCED DATASET

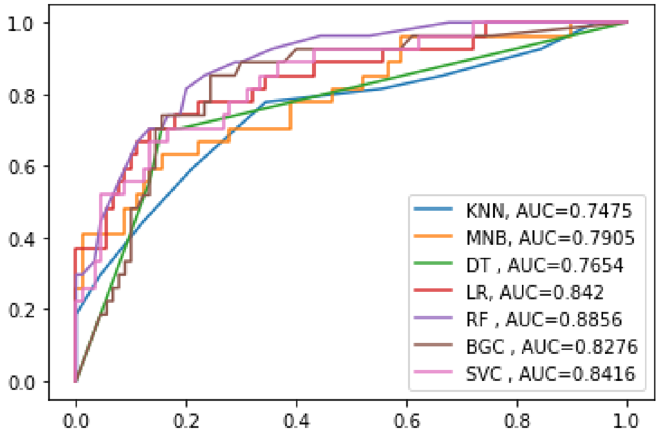


FIGURE 17: AUC UNDERSAMPLING

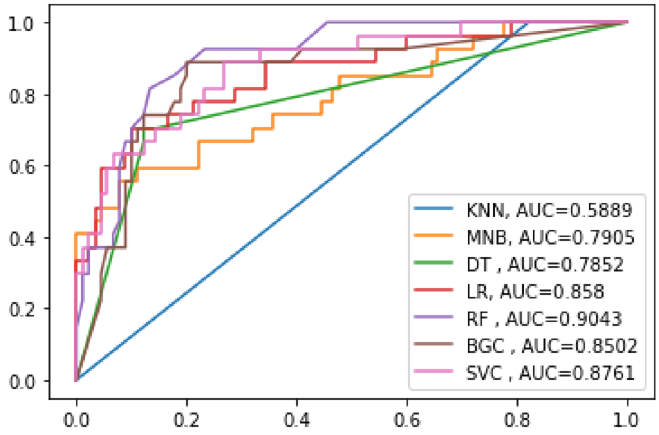


FIGURE 18: AUC OVERSAMPLING

The Area Under the Curve (AUC) is a critical evaluation tool for gauging the performance of classification models across alternate thresholds[355]. The Area Under the Curve (AUC) denotes the probability that a randomly chosen smallholder will be assigned a higher score by the classifier than a randomly chosen non-smallholder. When the AUC rating is 1, it denotes perfect classification, while a score of 0.5 implies the classifier has no advantage over random selection.

Results from figures 16-18 reveal that RF and SVC models persistently obtained the highest AUC scores, implying optimal performance in differentiating between the two classes. This demonstrates that these models are better at accurately classifying smallholder and non-smallholder instances, irrespective of the threshold employed.

Conversely, the KNN model shows the lowest AUC scores among the evaluated classification models, suggesting a weaker proficiency in distinguishing between the two classes. This could point to the DT model being less adept in managing intricate relations between features and classes or being more vulnerable to over-fitting, especially when placed alongside the RF model, which utilises a collection of decision trees.

4.3.5 Model Interpretation

An exploration of the model's feature importance's identified the most crucial terms in predicting the smallholder class. "Chickens", "livestock", "organic", "farming" and "rural" were particularly highlighted in the word clouds. This conclusion affirms the idea that smallholders focus on animal husbandry and routine farming operations.

Ultimately, the success of the classification models is contingent upon the training dataset and the technique used to address class imbalances. The SVC and RF models are uniformly more effective across all datasets, whereas the KNN model's performance is heavily influenced by the chosen strategy. The significance of class imbalance and dataset pre-processing for creating and judging classification models is made clear by these results.

4.3.6 Cross validation

Ten-fold cross-validation was used to evaluate the strength of the models. This process requires the dataset to be split into K parts which all have an equal size, and then the model is trained on K-1 parts and tested on the last one. This procedure is carried out K times, with each part functioning as the test set once. The results of the cross-validation are presented in Table 15.

Model	Avg. Accuracy	Avg. Precision	Avg. Recall	Avg. F1-score	Avg. AUC
K-nearest neighbour (KNN)	0.81	0.77	0.74	0.75	0.84
Multinomial Naïve Bayes (MNB)	0.79	0.75	0.71	0.73	0.81
Decision tree (DT)	0.80	0.76	0.72	0.74	0.83
Logistic regression (LR)	0.86	0.83	0.82	0.82	0.90
Random Forest (RF)	0.84	0.80	0.79	0.79	0.87
Bagging classifier (BGC)	0.82	0.78	0.77	0.77	0.85
Support Vector classifier (SVC)	0.85	0.82	0.81	0.81	0.89

TABLE 15: CROSS-VALIDATION PERFORMANCE METRICS FOR CLASSIFICATION MODELS

Table 15 displays the average performance of different classification models based on cross-validation results. LR produced the highest average accuracy, precision, recall, F1-score, and AUC values, making it the most effective model. Contrastingly, the most metrics showed that MNB had the poorest performance. The results infer that the LR model may the best option for this task, as it continually surpasses the other models on various performance measures.

4.3.7 Model Limitations

Even though most of the models performed well, the accuracy of the results is determined by the quality of the data, especially the user descriptions. Misclassifications can be caused by descriptions that are not precise or complete. Secondly, the model responds to the selection of hyperparameters, requiring precise hyperparameter tuning for peak performance. The model's dependence on the TF-IDF vectorizer could impede its capacity to perceive more complicated semantic connections between words, which could be better managed through more sophisticated methods, such as word embeddings.

4.4 Model Predictions

To test the true performance of the classification model, another prominent Twitter user profile dataset was scraped (username kept anonymous as per the owner’s request). The dataset consisted of 1000 user profiles. However, it was noted that 710 profiles did not have a profile description, and 513 profiles did not have a location specified. These profiles were excluded from further analysis, leaving a refined dataset which was used as input into the LR predictive model. This was then manually categorised, and the model's performance was evaluated using the following confusion matrix displayed in table 16:

	Predicted: Non-Smallholder (0)	**Predicted: Smallholder (1)**
Actual: Non-Smallholder (0)	True Negatives (TN): 493	False Positives (FP): 7
Actual: Smallholder (1)	False Negatives (FN): 2	True Positives (TP): 11

TABLE 16: CONFUSION MATRIX FOR USER PROFILE CATEGORISATION

Performance Metrics		
Metric	**Value**	**Summary of result**
Sensitivity (Recall)	0.846 (84.6%)	High Sensitivity (Recall): The model is good at identifying smallholders, missing only a small number (2 out of 13 smallholders).
Specificity	0.986 (98.6%)	High Specificity: The model is very effective at identifying non-smallholders, making very few mistakes (only 7 false positives out of 500 non-smallholders).
Accuracy	0.982 (98.2%)	High Accuracy: The overall accuracy is high, indicating the model performs well across both classes.
Precision	0.611 (61.1%)	Moderate Precision: The precision is relatively lower, meaning that when the model predicts a profile as a smallholder, there's a 61.1% chance it is correct. This suggests some room for improvement in reducing false positives.
F1 Score	0.710 (71.0%)	Balanced F1 Score: The F1 score shows a good balance between recall and precision, making it suitable for scenarios where both false positives and false negatives are important.

TABLE 17: PERFORMANCE METRICS FOR USER PROFILE CATEGORISATION PREDICTION

Table 17 highlights that the model shows great accuracy in identifying non-smallholders and demonstrates satisfactory performance in detecting smallholders. The current level of precision indicates that there is a need to improve the model's accuracy in predicting smallholders, which can be done by fine-tuning the model or incorporating additional features.

4.5 Discussion

This research has some implications that are applicable to smallholder communities. Accurately recognising these users in online networks allows for targeted outreach and support initiatives to be constructed to meet the particular requirements of this group. Moreover, the understanding of the interests and issues of the community can be used to construct policies and programs that are tailored and disseminated to the correct groups of individuals.

A broad network can promote the exchange of thoughts, experiences, and knowledge between smallholders and other people or organisations. Interchanging concepts can result in progressive solutions, collective efforts, or circulating the most effective techniques that benefit both the farmers and the wider agricultural sector. Non-smallholders in the network may be advocating for particular practices, understanding the significance of small-scale agriculture for food security, environmental sustainability, and community resilience. Furthermore, smallholding accounts could offer useful facts, points of view, or resources that are helpful to those wishing to gain knowledge about operations or considering transitioning to such a lifestyle.

One of the limitations of this study could be attributed to the small training dataset used (>1000), in addition to only incorporating the username and short profile description into the classifier. Additional work could be conducted to also collect user timeline information and use this as another means of information extraction.

This research offers a helpful addition to the increasing number of studies on the utilisation of machine learning techniques to identify distinct users, and it helps reinforce the potential of these approaches to improve the current intelligence held by governmental bodies by incorporating twitter follower networks into surveillance systems.

4.6 Conclusion

This section has displayed the capacity of machine learning models to accurately detect smallholder farmers from their Twitter profiles, which can have profound results for targeting outreach and assistance initiatives. The LR, RF and SVC models yielded the highest average accuracy, precision, recall, F1-score, and AUC values. Nevertheless, it is important to consider the quality of the data, especially user descriptions, as this can influence the reliability of the results. To achieve the best performance, hyperparameter tuning must be precise, and more elaborate techniques, such as word embeddings, can be applied to detect the semantic links between words.

By creating links between smallholders and other stakeholders, it can encourage the flow of information, experiences, and knowledge, potentially leading to creative solutions and collective wisdom.

Future research may include the use of different social networks and more sophisticated NLP techniques. Furthermore, looking into the effect of specific interventions predicated on these classifications could offer beneficial insight into the efficacy of such tactics in supporting the smallholder community.

5/. Social network analysis

The main aim of this chapter is to expand on the outcomes from the published paper entitled "Social Media Network Analysis of Smallholder Livestock Farming Communities in the United Kingdom”. This part of the research investigates the possibility of using social media data, especially Twitter interactions, to perform a network analysis of smallholder farmers. This involved:

1. Pinpointing influential users
2. Visualising community structures and hierarchies.
3. Mapping the user locations

The key findings from this analysis were:

- The scale-free topology of the network was evident in the few hubs with multiple connections and the many nodes with only a few links.
- Four different centrality methods were employed to identify the most influential members of the network, revealing a wide range of topics based on these measures.
- 11 groups were identified by the community detection algorithm.
- The clustering algorithm effectively separated the textual data into three unique clusters, and the t-SNE visualisation provided a clear depiction of the dimensionality reduction methods' effectiveness.
- According to the geolocation analysis, urban areas had the highest concentration of smallholders and their followers, while rural and remote areas had fewer locations, which may be explained by population differences.
- The findings of this research have important implications for policy decisions, industry practices, and future research endeavours. By using various communication channels, APHA can modify their communication strategies to ensure efficient transmission of animal health information to the identified influential users, who can disseminate this to others in the networks.

5.1 Introduction

Social media sites have become a popular method of exchanging news and messages, transforming the way people make contact, interact, and share ideas[359]. These platforms have become an essential part of everyday life for many, offering a virtual area for people to interact with others who have similar interests, backgrounds, or locales. Regarding smallholder farming, social media can be utilised as a valuable asset for grasping the complex social arrangements that form the behaviour and judgment-making procedures of farmers[360]. It provides a chance to examine the relationships, networks, and data flows that influence the adoption of the best practices, access to resources, and the ability to navigate new regulations surrounding their livestock.

Previous studies have shown the usefulness of social media data to comprehend social structures and the flow of data in multiple contexts, such as public health[361], political communication[362], and consumer behaviour[363]. Different social network analysis methods have been used in these studies to discover influential users, outline community structures, and examine the spread of information. Despite this, the use of social media data in livestock farming has not been explored much, and few studies have looked into how online interactions impact these communities in the context of veterinary epidemiology.

Farming systems are typically distinguished by intricate social structures that affect the sharing of data, implementing new technologies, and the adoption of the best techniques[364]. It is essential to comprehend the relationships within and between small-scale farming communities in order to tackle the livestock issues they may encounter. There has been a tremendous increase in the application of social media, which has created a vast amount of data that can offer valuable insights into the interactions, relationships, and information flows within these communities[365].

The distinct backdrop of the United Kingdom's smallholder farming sector provides an ideal setting to assess the potential of social media data in agricultural research[366]. A broad range of production systems in the UK puts smallholder farmers in the face of numerous problems, such as climate change and animal health administration[367]. Examining the online interactions and relationships of these users offers an opportunity to explore how this shapes their decision-making processes and lead to the adoption of better practices.

This research aims to bridge the knowledge gap by conducting a network analysis of small-scale farmers in the UK, making use of Twitter data to identify influential users, define community structures, and analyse the effects of these results on information dispersal.

This section adds to the existing literature on the adoption of social media data in agricultural research, providing insight into livestock farming communities.

5.2 Methodology

5.2.1 Data collection

With the Tweepy library in Python, data was extracted from Twitter, with access being granted through the Twitter API. Developer-level authorisation was obtained, permitting the retrieval of a greater volume of data by means of the API. By scraping the followers of a large smallholding Twitter account (@SmallholdersUK), a dataset of over 20,000 users based in the UK was generated. The final dataset was constructed by extracting user IDs (which were then anonymised), locations, and profile descriptions.

5.2.2 User anonymisation

Individual numerical codes were given to disguise user IDs in the system, preserving user confidentiality. To make sure user privacy was preserved, the initial identifications were cross-referenced with the anonymised data across the analysis.

5.2.3 Network analysis

A network graph was constructed to represent the Twitter follower network. This graph presented users as nodes, with edges signifying follower relationships. To work around computational and time constraints, only the 50 nodes with the highest degree were included in the network.

5.2.4 Network Metrics Evaluation

Degree centrality, betweenness centrality, closeness centrality, and modularity were used to measure various aspects of the network graph and gain a better understanding of its structure and features.

5.2.5 Key Users Identification

For each node in the graph, four metrics were used to identify key users in the smallholder farmers' Twitter network - degree centrality, betweenness centrality, eigenvector centrality, and PageRank. After ordering the nodes by their centrality values, the top 10 nodes with the highest impact were selected. By creating a dataset with the user IDs of the top ten influential users for each centrality metric, it was possible to explore the findings and pinpoint the most prominent users in the network in relation to various criteria.

5.2.6 Data cleaning

NLP methods were used to cleanse and prepare text data. This process entailed bringing in the required libraries and including custom stop words to reduce data noise. The text cleaning function made the text suitable for analysis through standardisation, deleting URLs, mentions, hashtags, and punctuation marks, and transforming the text to the lowercase. Merging user profile descriptions and timelines into one column was the precursor to the text categorisation.

5.2.7 Community Detection

Examining the network layout and distinct clusters was performed by charting the follower network and discovering communities. Both modularity-based algorithms and hierarchical clustering techniques were part of this process. The use of both methods is rationalised because of the distinct perspectives they bring to the fundamental structure of the network. Hierarchical clustering methods create a dendrogram to show how clusters are nested, while algorithms based on modularity seek to maximise a modularity score to spot well-segregated groups[348]. With the combination of both methods, the stability of the discovered communities can be evaluated, and a more in-depth comprehension of the network's community structure can be obtained. Additionally, examining the outcomes of these two approaches can authenticate the conclusions and guarantee a more dependable and exact understanding of the data.

5.2.7.1 Modularity based algorithms

The Louvain technique, which is known for its accuracy and efficiency in finding communities in large networks, was used for community discovery[166]. After calculating the best segmentation, the network was displayed with each node coloured according to its corresponding group.

After discovering communities using the Louvain method, their spread and how often they appear in the network were analysed. A data set was constructed which contained the partition details, each row reflecting a user and the community they are included in.

5.2.7.2 Hierarchical Clustering Methods

Text vectorisation, dimensionality reduction and hierarchical clustering were used to group users depending on their textual descriptions. The Term Frequency-Inverse Document Frequency (TF-IDF) approach was used to transform textual data into numerical features, effectively quantifying the significance of words within the user descriptions while taking into account their frequency across the whole data set. The vast network of followers necessitated a reduction of dimensions, which was accomplished using the singular value decomposition (SVD) technique, aiding a more efficient clustering analysis in the following steps.

The cosine similarity of the reduced-dimension data was used to calculate a distance matrix, which demonstrated the distances between data points and illustrated the differences between user descriptions. This matrix was used as an input for a clustering algorithm, which was able to recognise trends and organise users into groups.

A graph of the t-SNE was plotted in order to show the high-dimensional data and the conclusions of the study of the text dataset. Truncated Singular Value Decomposition (Truncated SVD) was used to reduce the feature space to 100 components prior to t-SNE being employed. T-SNE was applied to reduce the dimensions down to 2, allowing data points to be displayed in a two-dimensional space.

Clustering users hierarchically was used to determine groups with similar descriptions. Minimising the sum of squared differences within clusters with the ward method created a hierarchical representation of the relationships between users.

5.2.8 Geolocation Mapping

In addition to the primary focus of the study, spatial distributions of the users in the network were also examined. This supplementary examination sought to detect any possible geographic patterns or trends among the networks. Converting the free-text location column from the profiles into numerical coordinates was part of the process, enabling the locations to be shown on a heat map.

This was made feasible through the employment of the GeoPY package, which can translate textual location data into numerical latitude and longitude coordinates. The library leverages a range of geocoding services, such as OpenStreetMap, Google Geocoding API, and others to perform the conversion. This is a vital part of plotting a user's location on a map since it provides a uniform way of displaying spatial data when co-ordinates are unavailable in the user profile data.

To construct the choropleth map, the latitude, longitude, and count (if available) of each location inside the UK mapping parameters were considered. The limits were established by the following latitude and longitude constraints: minimum latitude of 49.8, maximum latitude of 60.9, minimum longitude of -10.9, and maximum longitude of 2.0.

The justification behind selecting a choropleth map may provide some insight into the geographic dispersion of users in the network, detect any clusters or high-activity areas, and potentially find regional trends or choices that may be of significance to the community.

5.3 Results

This results section provides a comprehensive exploration of the network, revealing a variety of patterns, observations, and influential factors that impact the network's structure and dynamics. The examination looks at several components, such as its evaluation, identifying influential users, determining the communities, and determining the spatial spread of smallholders and their followers. This exploration not only amplifies our knowledge of the network but also gives useful information about the kinds of users and stakeholders involved in these online networks.

5.3.1 Network analysis

When the follower network was visualised, as seen in figure 19, it was clear that it possessed a scale-free topology, with a few hubs having multiple connections and a large number of nodes having fewer links. Numerous users were associated with these hubs either directly or through other nodes, which acted as the hubs. This finding aligned with the traits of social networks, where a select group of powerful users typically has a significant influence on the entire structure[368].

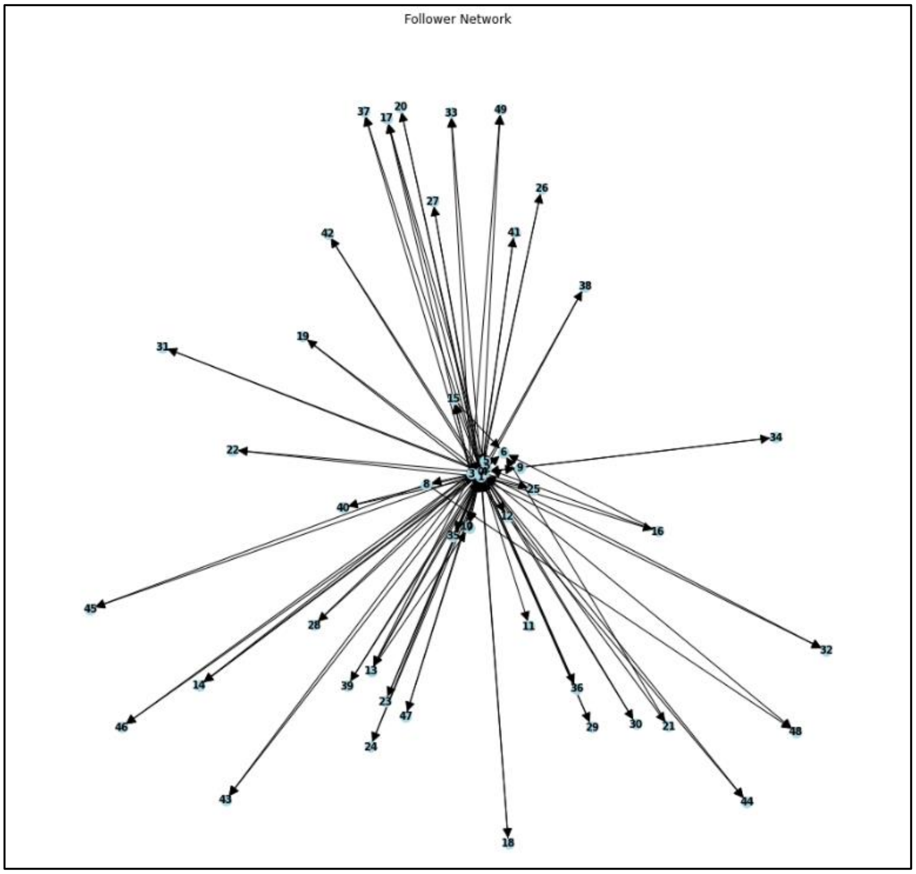


FIGURE 19: TWITTER FOLLOWER NETWORK

5.3.2 Influential users

Four different centrality methods, including page rank, degree, between, and eigenvector centrality, were employed to identify the 10 most influential members of the network. These individuals were then categorised, and the top 3 terms from their profiles were extracted.

In Table 18, the most popular topics in the degree centrality analysis covered themes like employment, nature, and individual interests. In comparison, Table 19's between centrality analysis indicated a more heterogeneous cluster of categories with a higher concentration of scholarly and veterinary themes. Keywords of prominence in this table centred around newsletters, energy, and animals.

Anonymised ID	Top 3 terms
0	['grass', 'video', 'haylage']
3	['day', 'read', 'medium’]
8	['please', 'see', 'health’]
9	['love', 'would', 'like']
10	['know', 'sperm', 'whale']
11	['thing', 'health', 'day']
12	['work', 'consultant', 'junior']
13	['plant', 'wildflowerhour', 'spring']
14	['green', 'need', 'party']
15	['nature', 'dawn', 'chorus']

TABLE 18: DEGREE CENTRALITY

Anonymised ID	Top 3 terms
103	['elephant', 'asian', 'stae']
57	['energy', 'year', 'new']

92	['time', 'god', 'tide']
59	['folk', 'horror', 'bird']
84	['thanks', 'thank', 'ardnamurchan']
17	['newsletter', 'substack', 'note']
100	['yes', 'haha', 'yorkshire']
16	['blossom', 'tulip', 'cherry']
3	['day', 'read', 'medium']
58	['puppy', 'pet', 'great']

TABLE 19: BETWEEN CENTRALITY

Predominately, the themes in the eigenvector centrality (Table 20) highlighted phrases pertaining to personal hobbies, patriotism, and social media. Conversely, the PageRank results (Table 20) displayed topics from politics and conflict to hobbies and health. Anonymised ID 3 was present the most out of the influential metrics which were seen twice in Tables 18 and 19.

Anonymised ID	Top 3 terms
24487	['amp', 'kennedy', 'want']
24488	['follow', 'account', 'drop']
24491	['usmc', 'veteran', 'patriot']
24492	['amp', 'bill', 'three']
24493	['video', 'tiktok', 'check']
24494	['happy', 'love', 'friday']
24495	['order', 'link', 'good']
24496	['marxist', 'care', 'dear']
24497	['let', 'get', 'disloyal']
24498	['god', 'ceo', 'buying']

TABLE 20: EIGENVECTOR CENTRALITY

Anonymised ID	Top 3 terms
21845	['government', 'lying', 'number']
21846	['borrowdale', 'stay', 'award']
21847	['war', 'amp', 'asking']
21848	['rescue', 'mountain', 'men']
21850	['great', 'lakedistrict', 'easter']
21715	['youtube', 'full', 'amp']
21716	['pay', 'people', 'rise']
21720	['metabolic', 'health', 'symposium']

TABLE 21: PAGE RANK

Atypical findings were uncovered among the most prominent users, including the use of terms such as "elephant," "whale," and "Ardnamurchan. As these terms were found amongst the “other” category”, an inference about the demographic of individuals who have an interest within the smallholder domain could be made, in regard to their vested interest within the wider animal field. "Elephant" and "whale" could represent particular health concerns related to larger creatures or broader environmental issues that influence small-scale farming. The reference to "Ardnamurchan" may indicate the relevance of matters impacting smallholder agriculturists in that area of Scotland, potentially due to distinctive cultivation techniques.

5.3.3 Community detection

5.3.3.1 Modularity based algorithms

Figure 20 illustrates the results from the Louvain community detection and has enabled a better understanding of the organisation and the linkages between the users in the network.

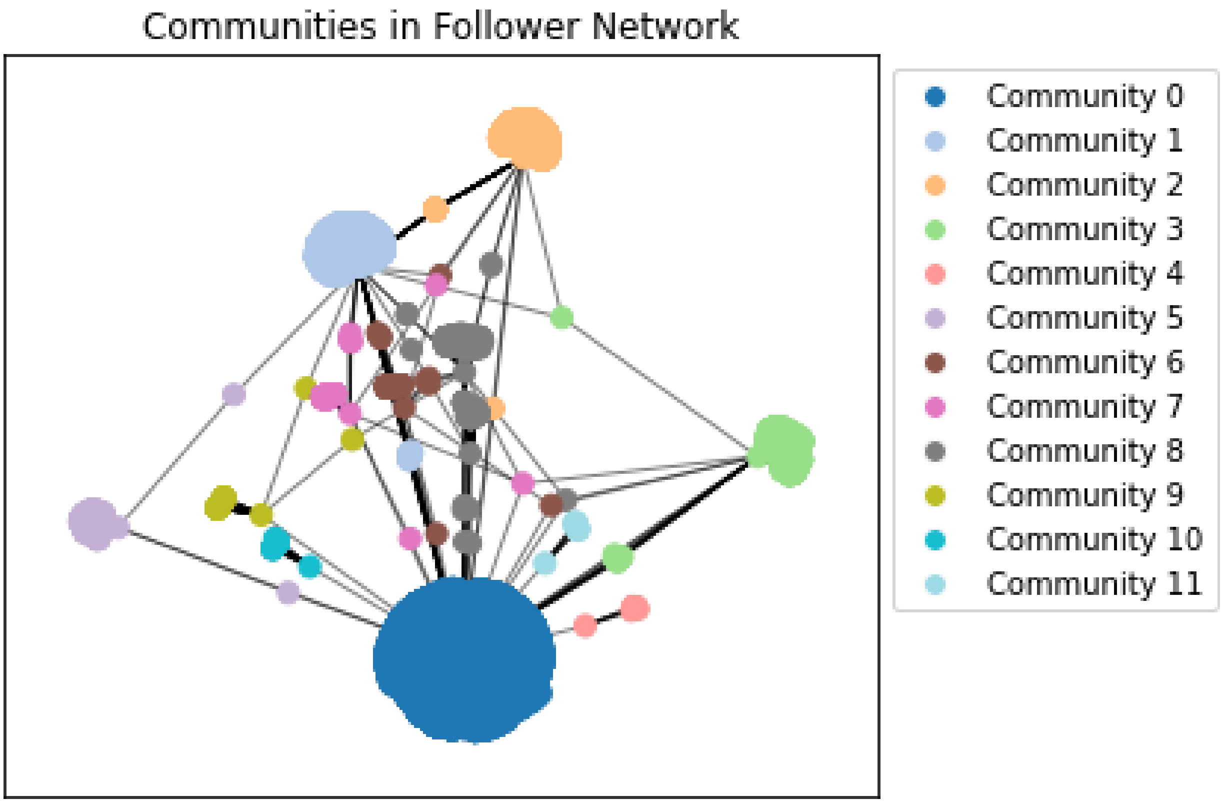


FIGURE 20: COMMUNITY DETECTION

Regarding the connections between individuals, the community detection algorithm discovered 11 groups, with a strong right skew visualised in figure 21. Community 0 has the most expansive network of followers, as it is the biggest smallholding Twitter account. All the other communities have a frequency under 2500, with communities 1, 2 and 3 being the most populous. These communities embody groups of individuals who are united by shared connections or interests, which can be beneficial for grasping the network's internal workings.

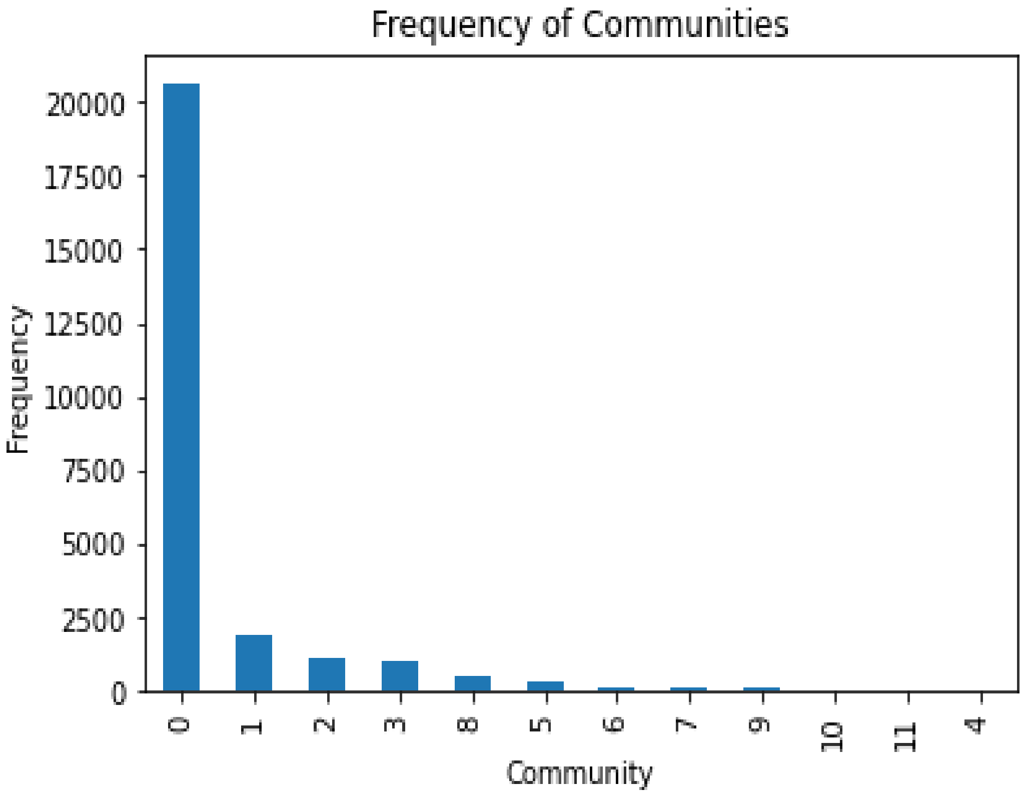


FIGURE 21: FREQUENCY OF COMMUNITIES

5.3.3.2 Hierarchical clustering methods

The dendrogram of the clustering hierarchy which was generated from the hierarchical clustering analysis of the textual dataset is represented in Figure 22. Dimensionality reduction with SVD was implemented to restrict the feature space to 100 components.

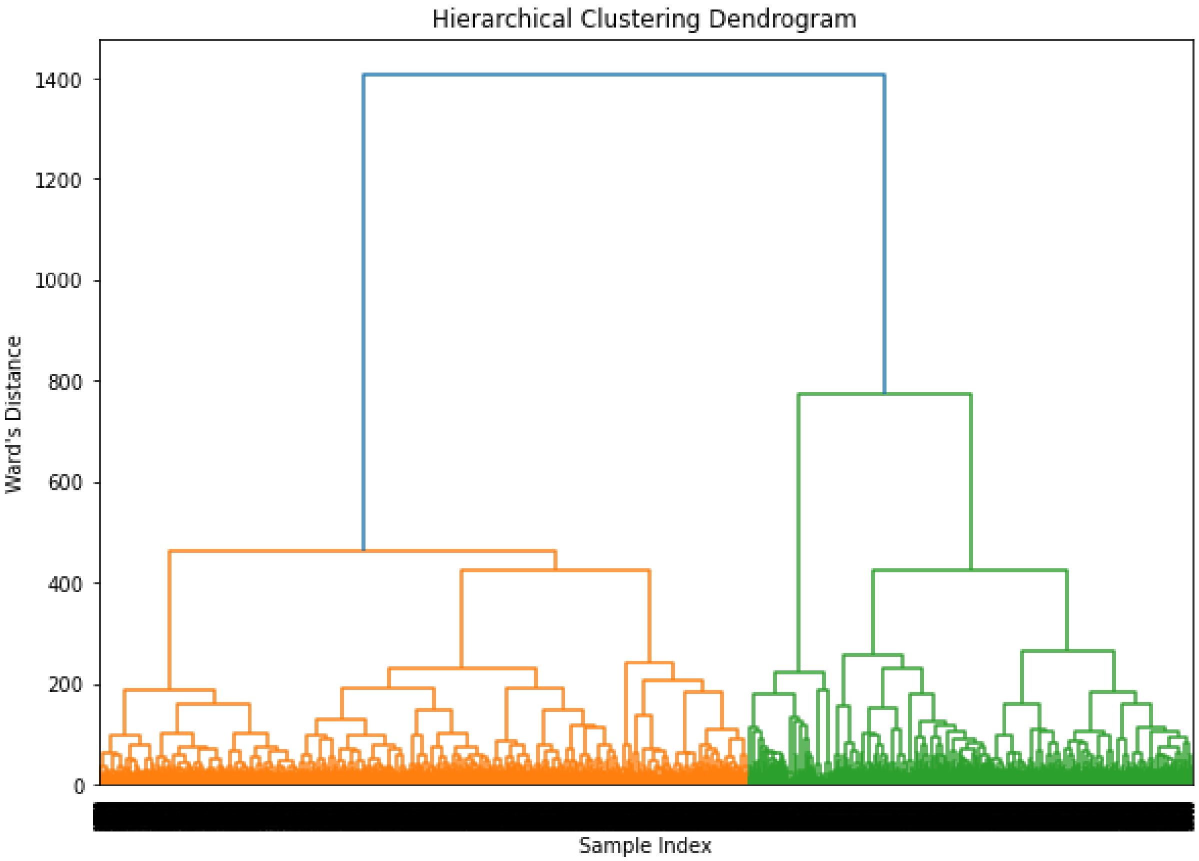


FIGURE 22: CLUSTERING DENDROGRAM

Additionally, this illustrates the hierarchical structure of the data, with each branch representing the merger of two clusters based on their similarity. Ward's distance is plotted on the y-axis of the dendrogram. Moving up the y-axis, clusters become more similar and are unified. Three clusters (n = 3) were created by selecting an appropriate cut-off point for grouping based on an inspection of the dendrogram. The cut-off point was employed to create more distance between clusters while reducing within-cluster variation and the dendrogram's longest vertical line that wasn't intersected by any horizontal lines served as the basis for the decision.

Cluster	Frequency
1	10544
2	1376
3	5805

TABLE 22: CLUSTERING FREQUENCY

Moreover, further analysis was done to investigate the distribution of the data points between the three clusters after adding cluster labels to the dataset. As outlined in table 22, Cluster 1 consists of 10544 data points, Cluster 2 is made up of 1376 data points and Cluster 3 consists of 5805 data points. Based on these results, the clustering algorithm was able to successfully segment the textual data into three distinct clusters, thereby generating a useful dataset division based on the similarity of the feature vectors in the condensed space.

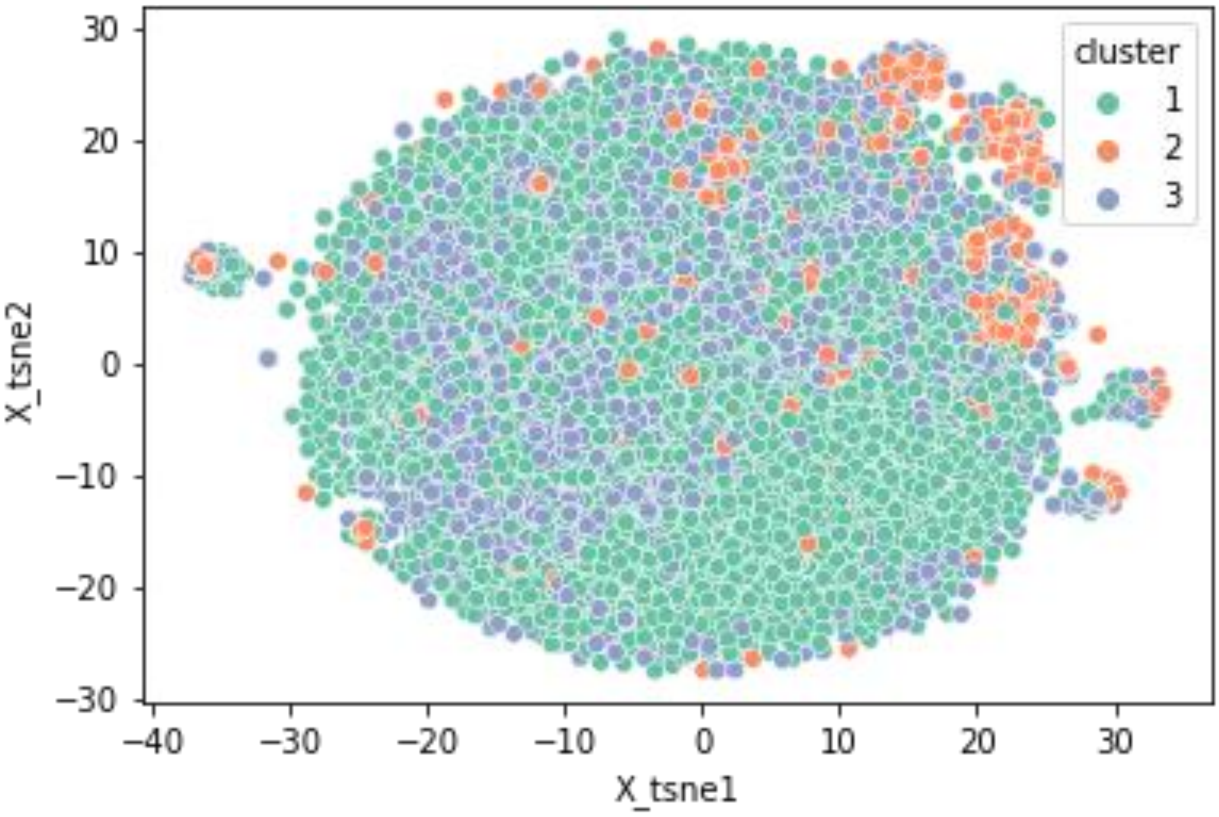


FIGURE 23: T-SNE VISUALISATION

The scatter plot shown in Figure 23 displays the t-SNE visualisation of the high-dimensional data to present the clustering analysis results of the textual dataset.

The cluster of each data sample is represented by its colour. The scatter plot lays out how the clusters are separated in the two-dimensional t-SNE space and gives details on how the data points are spread out in the clusters. The feature vectors in the t-SNE space were close enough that the clustering algorithm could effectively group the textual data into three distinct parts, creating a useful dataset sectioning. Cluster 1 was the most prevalent, covering the majority of the data in a consistent way, whilst cluster 2 was less frequent and more concentrated in the top right corner of the graph. Overall, the fact that the scatterplot was able to distinguish the clusters demonstrates how successful the dimensionality reduction methods were.

5.3.4 Geolocation mapping

The spatial distribution of the studied locations across the UK were visualised using a choropleth map, as depicted in figure 24. Urban areas including Lancashire, Midlands, Devon and Central Scotland had the greatest concentrations of locations. These points appeared the darkest on the map, which suggested a greater number of places compared to other regions.

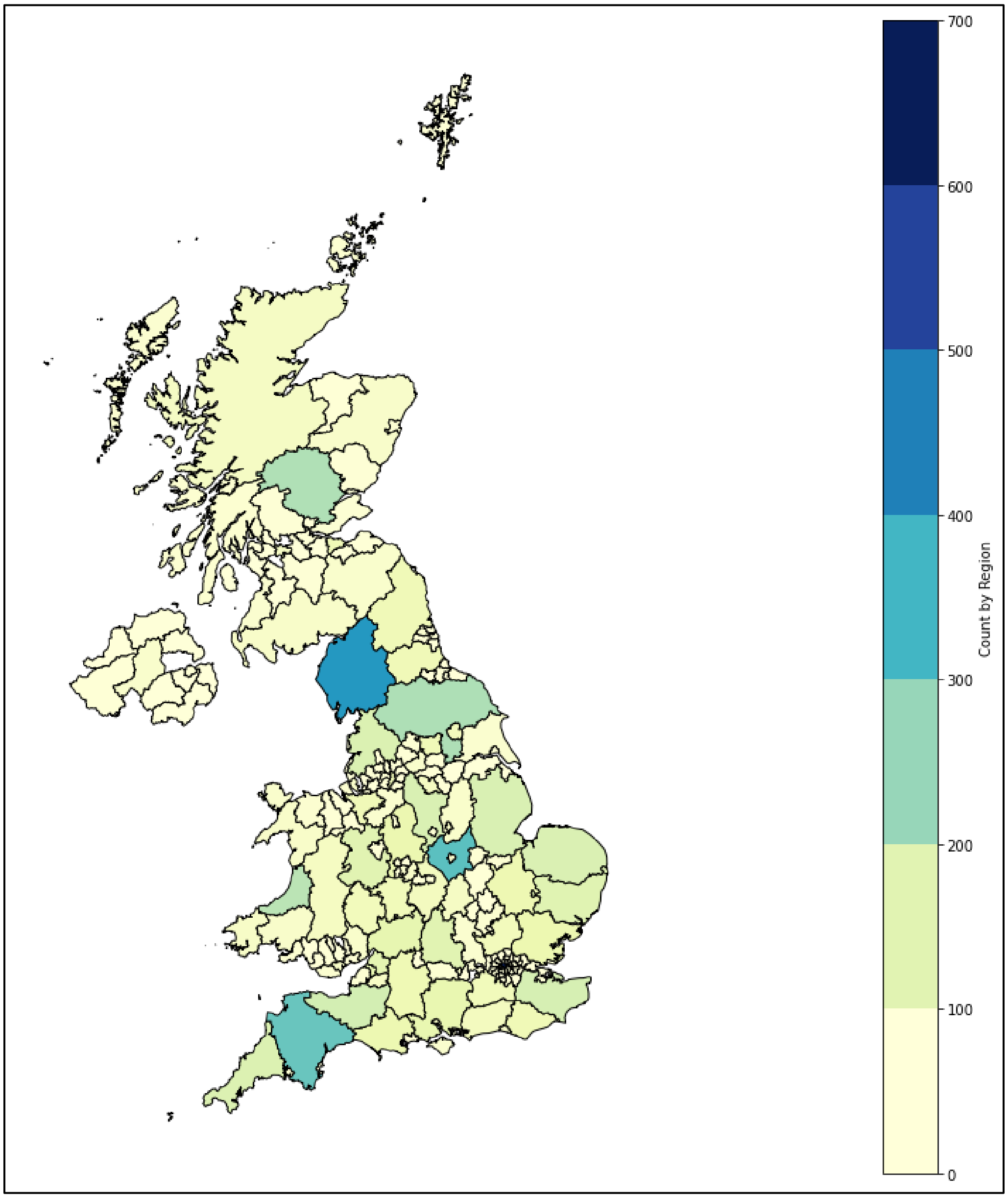


FIGURE 24: CHLOROPLETH MAP OF USER LOCATIONS

Regions such as the Scottish Highlands, western Wales, and certain parts of southwest England, which are rural and remote, had less intense colours on the heatmap, suggesting a lower density of locations. The overall pattern that was observed was that the level of locations decreased as the distance from large cities grew, but there were a few areas with numerous locations in smaller towns and on major transportation routes.

This foundational map analysis yields important facts about the spatial distributions of smallholders and their followers. Investigating the components that are driving the patterns and analysing the variations in density of locations through time could be beneficial.

5.4 Discussion

Previous literature established that the scale-free topology is still present in modern online social networks, which is consistent with the structure of the follower network in figure 19[368]. Additionally, the impact of influential users in driving public opinion and sharing data on social media outlets has been thoroughly studied [324]. The techniques in previous literature are similar to the methods used in the current study, and it points to a few key players in how information is spread amongst the communities.

Community detection in complex networks has been especially important for comprehending the structure and function of social networks[325]. Recent studies have revealed the potential of employing community detection on various modern network datasets, like those related to cooperation networks, citation networks, and online social networks. Within the scope of this investigation, eleven groups were determined from the findings, and further investigation into the population of each society would provide an exhaustive examination of the distinct features and likenesses between each community.

The identification of various groups within the network indicates the presence of multiple user groups that are associated either through links or interests, which could be indicative of specialised farming techniques, localised agricultural trends, or shared experiences among farmers.

A thorough analysis of centrality measures revealed influential users who could have a considerable impact on the spread of information and the implementation of optimal animal health management approaches. Interaction with these users and employing their connections can strengthen the efficacy of animal health initiatives and encourage better practices amongst smallholder farmers.

5.4.1 Limitations and further research

The study explored smallholder communities within the framework of online social networks. Examining only those who follow one account does not accurately represent all smallholders or the range of topics covered. Additionally, the evaluation concentrated on centrality measures that could not accurately illustrate every aspect of sway and data movement in these gatherings. However, the study's data was derived from social media interactions, not considering the full scope of smallholder farmers, particularly those who don’t actively partake in online discussions.

It's important that future research includes larger and more representative sample sizes, and a range of different social network analysis techniques. The community detection analysis we conducted on the given network offers us useful knowledge about the structure and links between users in the realm of smallholder farming, but it is not without its drawbacks. It is necessary to use a more comprehensive hyper tuning of the parameters of the two community detection models, combined with a reliable custom stop word dictionary, to reduce the majority of the noise present in the textual input.

To get a fuller picture, it might be beneficial to incorporate additional data sources, such as interviews or surveys, to gain more insight into the relationships and dynamics between and within communities. To better grasp the special qualities of these communities and their individual components, further research is necessary. Additionally, it is critical to investigate if interventions based on the identified community structures and influential users are successful. This could be done by considering the views of the communities, which can be discerned from tweets from livestock authorities such as APHA, as well as information from local veterinarians.

5.5 Conclusion

Network analysis of smallholder farmers was done using social media data, concentrating on Twitter communications. The main conclusions of the analysis involved identifying influential users, recognising the community structures in the small-scale farming network, and the potential influence these findings have on information dissemination, disease monitoring, and control. Nevertheless, the excitement around these discoveries should be balanced with a practical evaluation of the resources for putting them into action.

Firstly, to maintain the accuracy and relevance of the network analysis, ongoing engagement and regular updates are needed when implementing the research's suggested strategies, like leveraging influential users for information dissemination. As an illustration, the Twitter API data collection and subsequent analysis require ongoing updates to accommodate the dynamic nature of social media and farming communities. This task demands both technical resources and dedicated, skilled personnel. Since the takeover of Twitter, and subsequent rebranding to X by Elon Musk, the ability to access the Developer-level APIs has been drastically affected by moving to a “pay as you use” model, rather than a free tool. Furthermore, the debate surrounding the openness of these identified influential users still remains questionable, because of a reduction in government trust since the COVID-19 pandemic. This will be discussed in greater detail in Chapter 9.

Furthermore, regularly computing efforts are necessary for identifying influential users and community structures. The results reveal that centrality measures can identify key individuals in the network, indicating the importance of ongoing monitoring to track changes in user influence, however, are computationally expensive when performed on large networks. The importance of continuous analysis and human resources in handling social network dynamics is emphasised by this aspect alone.

Moreover, the geolocation mapping reveals that smallholder activities are primarily concentrated in urban regions. In order to effectively use this information, policymakers should devise communication strategies that consider specific geographic and demographic findings from our analysis. The free-text input nature of user locations online presents a potential challenge due to the poor level of granularity, thus conducting small pilot studies in densely populated urban areas would be an ideal way to evaluate the effectiveness of network analysis findings.

To conclude, this research has yielded a valuable understanding of the structure, key users, and community relations inside this population, whilst simultaneously highlighting the possible barriers to large scale adoption. These results show that social media data can be used to improve comprehension of small-scale farming systems and tackle the challenges experienced by these groups. It opens up avenues to explore the field of information dispersal, and how government advice permeates though these identified networks.

6/. Forum scraping

This chapter expands on the findings from the published paper "*Text mining of veterinary forums for epidemiological surveillance supplementation"*. There were 4 main analytical segments applied:

1. Investigating the effectiveness of topic modelling (both static and time series) on a livestock discussion forum regarding smallholder pig and poultry subforums.
2. Anomaly detection to explore unusual spikes in posting activity over time.
3. Geolocation analysis was conducted to determine the locations of users.
4. Comparative analysis was performed to analyse the overlaps in both locations and word occurrences of pig and poultry subforums, in addition to comparisons of locations with confirmed APHA recorded holdings.

The key findings from this analysis were:

- UK smallholder farmers utilise online forums to converse about the everyday management of their animals, requesting advice on matters including nutrition, breeding, husbandry, sales, biosecurity, slaughter, and disposal, as shown by the results from topic modelling.
- Anomaly detection sets the foundation for early warning systems to be created, whereby manually adjusted thresholds may be constructed which allow for any outliers to be flagged and investigated on those specific dates.
- Comparative analysis through geolocation mapping highlighted particular regions with a prevalence of both types of smallholders. Furthermore, bigrams revealed overlaps between the terms discussed, particularly pertaining to the age of the livestock i.e. "Weeks old, months old".
- Geolocation mapping of confirmed APHA holdings with forum extracted locations poses interesting questions regarding intelligence disparities and open avenues for further research.
- Governments can use the information from these forums, combined with the data that they have gathered, to implement interventions for diseases and biosecurity measures.
- There are issues with using internet data for observation, such as selection bias, small sample sizes, and the possibility that a few powerful users might drive the conversation.
- These forums discuss clinical signs without any input from veterinarians, and many conditions require tests to be verified.

6.1 Introduction

Forums are another avenue for data mining and contain a wealth of information which may not be contained within Twitter. This chapter seeks to analyse the potential of making use of these tools, specifically in forum-based intelligence, to interpret conversations amongst the smallholding community.

In the UK, large-scale farms are mandated to share their livestock health data with governmental agencies, yet there is scarce data accessible from small or backyard farms[369]. These smaller farms frequently turn to digital sources such as social media and search engines for instruction on animal husbandry and health. Reports indicate that small-scale farmers are applying digital data with greater frequency to meet their livestock requirements[6]. Yet, there is limited knowledge of what is discussed on these platforms, responses to regulations, and the role influential users have in propagating information. Traditionally, farmers looked to agricultural professionals for direction, however with the growth of social media, they are now turning to their online peer networks.

With the rise of online livestock-related data, text mining and NLP can be used to gain insights into the discussions occurring in these online communities. These tools can discern patterns in unstructured text and create groups of related text segments or topics[321]. The LDA algorithm is especially useful for topic modelling, giving clear outcomes and avoiding problems like over-fitting. By taking data from a static timepoint and adapting this for use on a dynamic basis, the potential to identify disease outbreaks in real-time becomes a viable option.

6.2 Methodology

6.2.1 Justification for forum selection

As mentioned in the methodology section in chapter 3, this particular forum (www.accidentalsmallholder.net) was selected as the primary data source as it’s the most prominent forum for smallholdings in the UK. The website provides information on numerous subjects, from animal husbandry and biosecurity to entrepreneurship and general livestock management. Over 57,000 active participants were registered on the forum during the time of conducting this research. The platform’s large and active community makes it an invaluable resource for gathering smallholder-centric information and consolidating passive intelligence. In contrast to Twitter’s general conversations or opinions on non-smallholding matters, forums are focused on livestock-specific queries, making them a less noisy platform.

6.2.2 Data collection

All the data was sourced from a prominent UK based smallholding forum, https://www.accidentalsmallholder.net/. The methodology, as depicted in Figure 29, was implemented for both subforums, “poultry & waterfowl” and “pigs” respectively. Data extraction from web scraping was accomplished with the RVest package, via the R programming language. This was coupled with the CSS (cascading style sheet) selector function in Mozilla Firefox, which made the extraction of HTML tags within each page possible.

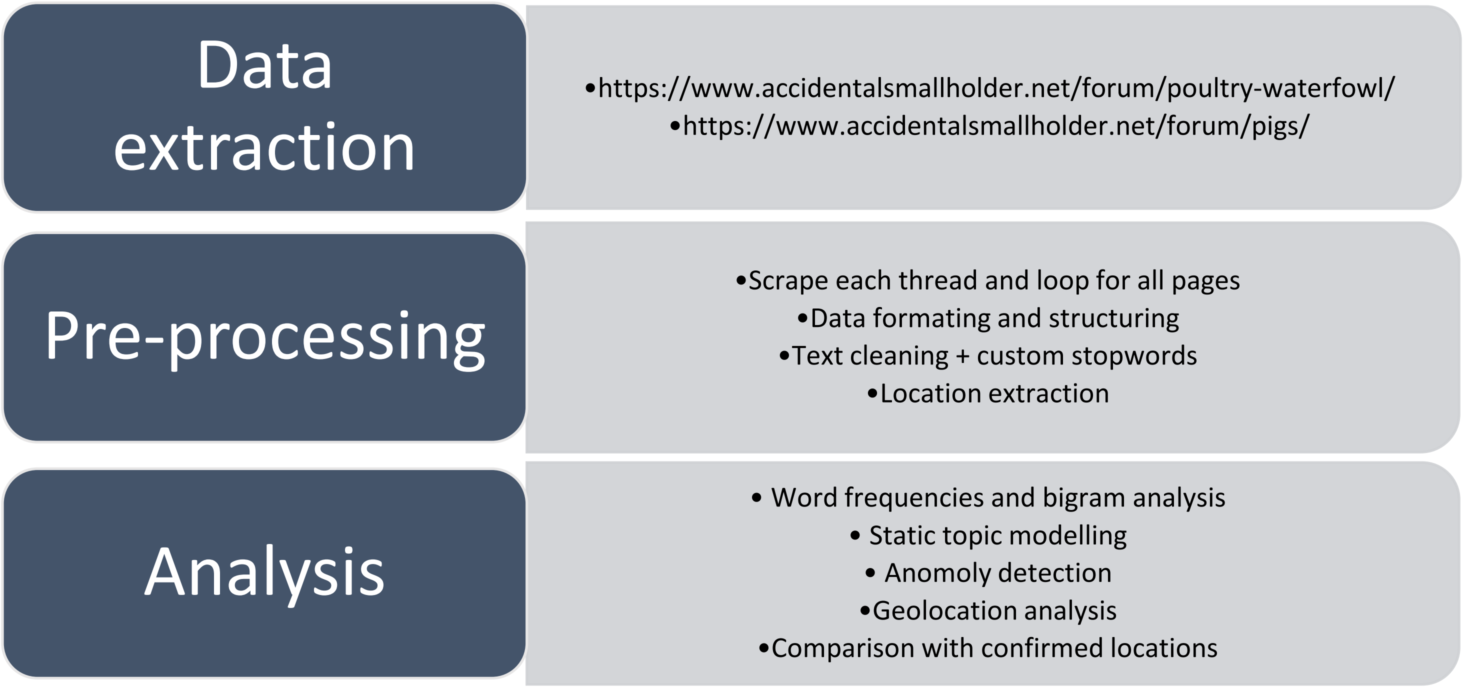

Data extraction
•https://www.accidentalsmallholder.net/forum/poultry-waterfowl/
•https://www.accidentalsmallholder.net/forum/pigs/
Pre-processing
•Scrape each thread and loop for all pages
•Data formating and structuring
•Text cleaning + custom stopwords
•Location extraction
Analysis
• Word frequencies and bigram analysis
• Static topic modelling
• Anomoly detection
•Geolocation analysis
•Comparison with confirmed locations

FIGURE 25: FORUM SCRAPING WORKFLOW

The process of collecting data from the forums was divided into two parts, each of which produced a distinct dataset. The initial step included taking data from each forum page in the poultry & waterfowl section and the pig subsections. Resulting in a dataset that included the date of the post, the title, URL link for each thread, and engagement data such as likes and replies, as seen in table 22.

Significant temporal variations and events that impacted the smallholding community were strategically captured through the selection of the 2017-to-2022-time scale. The six-year duration was deemed to be sufficient enough to observe year-over-year changes and trends. Within this time frame, critical events like the Avian flu epidemic in the UK and the COVID-19 pandemic had a substantial impact on small-scale farming, including supply chains and market demands. 281 URL links were generated from the threads in the pig forum. The poultry data was significantly larger and generated 775 URL links, possibly due to a higher number of poultry keepers in the area and recent outbreaks of Avian flu.

The second stage of extraction involved deriving the entire discussion within each forum post by applying the extracted URL links. Time-outs were implemented in the code to prevent overwhelming the website with massive requests, as it was computationally expensive. Moreover, each page in every forum post was subjected to this procedure, along with the URLs gathered before. The pig forum produced a total of 4191 unique posts, while 5425 were produced from the poultry forum.

Subforum	Thread number	Thread title	Thread originator	Start date	URL link	Last post date	Replies	Views
Poultry and waterfowl								
Pigs								

TABLE 23: FORUM FRONT PAGE DATASET STRUCTURE

Subforum	Thread number	Date	Thread title	Message	User Location
Poultry and waterfowl					
Pigs					

TABLE 24: INDIVIDUAL THREAD DATASET STRUCTURE

Stage two was comprised of extracting the entire discussion from each forum post using the URL links from stage one. This led to a more thorough dataset with every row denoting a single user post from the forum dialogue, as illustrated in table 23. From the pig forum, 4180 distinct entries were produced, and 7140 from the poultry forum.

6.2.3 Data analysis

6.2.3.1 Pre-processing

Text cleaning was applied to each message in the threads. This necessitated deleting any visual depictions and non-alphanumeric symbols, such as emojis. Along with this, punctuation, hyperlinks, and white spaces were also removed. Hyperlinks within the text had to be taken out as they did not have any value to our exploration, usually linking to items that could be bought (e.g. “pre-owned farm equipment”). To make the forums more tailored to the research questions, the NLTK package's common stopwords and custom stopwords (see appendix A) were removed to decrease the excess noise.

Livestock-related data may not be best suited for lemmatization and stemming, despite their common use in NLP to simplify words to their root forms. The context and meaning of discussions in livestock farming depend on specific terminology. These techniques can change the meaning of these terms and potentially obscure important subtleties. In addition, animal-related terminology often includes specialised language, informal phrases, and recently created words that could be incompletely interpreted by these techniques. It might be more insightful to preserve the original language used in discussions and only perform basic text cleaning tasks such as removing stop words, punctuation, correcting spelling errors, and lowercasing the text.

6.2.3.2 Text mining

The analysis calculated the frequencies of words from bigram components, with the 5 most common word-pairings for each component being displayed. Bigram analysis was especially helpful, as they uncovered correlations that unigram analysis would not have detected. Concentrating on the most frequent bigrams enabled a deeper comprehension of the subjects and styles discussed in these online forums. This technique allowed for the acquisition of significant patterns from the unprocessed online discussion text data, resulting in observations that are applicable to both the pig and poultry communities.

6.2.3.3 Topic modelling

The LDA technique was applied to cluster textual data and gathered insights into the topics discussed on this niche forums[370]. LDA is a probabilistic approach for estimating the topic structure of text, determining the likelihood of a specific word relating to a single topic[371]. It estimates the posterior distribution of the Bayesian probability model, which determines the proportions of topic configurations in documents and the word patterns of the topics. The meaning of the topic can be determined by the words with the highest likelihood of belonging to that topic.

The perplexity score was used to calculate the best number of topics, and this was shown graphically on an elbow curve plot[372]. The perplexity score symbolises the variance between the assumed value of the topic probability and the real value of the topic probability. The log-likelihood of a word belonging to a topic in the testing data was used to measure this. Furthermore, the coherence score, which looks at the similarities between words with high probabilities for each topic, was also used as a performance metric.

Through the coherence score, the ideal number of topics was determined by analysing the similarities between words with high probabilities for each topic's words. A model's ideal state is reflected by its high coherence score, displaying meaningful and consistent output, and a low perplexity score, signifying superior predictive accuracy[373]. Low perplexity indicates the model's confidence in its predictions, while high coherence suggests that the generated topics are closely related and easy to understand.

The coherence score is generally regarded as a more reliable and understandable metric for selecting the optimal number of topics in a topic model, as opposed to the perplexity score[374]. Coherence scores gauge semantic similarity among highly scored words in a topic, making it a better metric for human comprehension of a topic. Conversely, perplexity may not always be an accurate statistical measure. Although lower perplexity scores can indicate superior model performance, they may not necessarily correspond to more significant or comprehensible topics. Over-fitting and fragmentation of meaningful topics can be caused by the preference of perplexity score for models with more topics. Since it emphasises semantic interpretability, the coherence score is generally the preferred method for determining the optimal number of topics in topic modelling.

6.2.3.4 Temporal analysis

6.2.3.4.1 Outlier Detection

Outlier detection enabled the identification of abnormal temporal patterns in the time series data. The Isolation Forest algorithm was applied to identify anomalies, and its effectiveness in detecting isolated data points in high-dimensional datasets has proven to be successful[341]. The occurrence of abrupt changes or spikes in topic prevalence can imply significant events or themes during particular time frames.

6.2.3.5 Spatial analysis

A function was created which selects a free-text location as an input and returns latitude and longitude values, filtered to show only locations within the United Kingdom. Through these functions, latitude, longitude, and country information was extracted and stored in new columns within the data structure for each location in the dataset.

To ensure data integrity, entries without location, latitude, longitude, or country information were eliminated.

Extracting geolocation data allowed for the creation of a choropleth map with visual representations of the locations. A geometry-integrated format was used to transform the data structure, resulting in a map centred on the mean latitude and longitude of the dataset.

In addition, bigrams were generated for the top 10 locations identified through a frequency count. The process involved a function that tokenised and cleaned threads into words, then formed bigrams from them. The function was used on all the distinct locations, and a dictionary was created from the resulting bigrams. A list of bigrams (values) was associated with each location (key) in this dictionary. By doing this, it became possible to identify and compare the most frequently discussed word-pairing per location.

6.2.3.6 Comparative analysis

By examining the bigrams and locations of pig and poultry smallholding forums, a unique insight into the smallholding community in the UK is revealed.

Extracting sets of frequently used terms and bigrams from the pig and poultry forums was the first step in discovering overlapping areas between the two. Through the identification of intersections between these sets, common terms and bigrams that were shared by both the pig and poultry forums were discovered.

A Venn diagram was used to visualise the shared bigrams between pig and poultry discussions. By explicitly listing the shared bigrams, a visual representation of the commonalities in language used across both forums was provided.

The geographical locations mentioned in each forum were the next point of interest for the analysis. A comparison was made between the datasets to find common locations, which were extracted and identified. The unique locations were plotted on a map, with distinct colour's being assigned to each cohort, in addition to the overlapping locations.

6.2.4 Location validation

To verify the conclusions of this chapter and analyse the parallels and distinctions between the documented information on pig and poultry facilities in the UK, and the extracted forum location data were made. Official reports from APHA were extracted which contained maps of the holdings and were examined alongside the forum-generated choropleth maps.

6.2.5 Computational requirements

To extract and analyse the data in the methodology section, computationally intensive tasks such as web scraping, text processing, and topic modelling are necessary. Addressing computational resource requirements and model run times is crucial for ensuring efficiency and feasibility. The for loop encountered multiple time-out requests when extracting data from each pig and poultry forum, which prolonged the process to almost three hours. Running the topic models and calculating the optimal topic numbers took approximately 20 minutes to complete during the modeling phase on an Intel Core i7 x64 laptop.

Processing thousands of URL links and retrieving substantial amounts of data in web scraping requires critical memory allocation and bandwidth. To conduct text preprocessing and analysis efficiently using packages like TM, TidyText, and NLTK, especially when working with large datasets, it is recommended to use a multi-core processor for CPU-intensive tasks, thus improving performance. Computational intensity is a challenge when using LDA topic modeling, which is a probabilistic method. This is especially true when optimising parameters like the number of topics, which requires a large amount of CPU and memory resources. A powerful CPU and abundant RAM are essential for achieving optimal machine performance. The speed of LDA computations can be further improved by making use of a GPU. Integrating anomaly detection algorithms, such as Isolation Forest, into temporal analysis may require memory usage and parallel processing for optimal results. Finally, cloud computing services like Amazon Web Services (AWS) or Azure offer an efficient way to manage computational demands and scale as needed. This may be necessary when applying the predictive models created in this thesis on a larger scale, including daily use and application to multiple sub-forums simultaneously.

6.3 Results

This section will provide a complete report of the results achieved from different data analysis methods, including word frequencies, topic modelling, temporal analysis, and spatial analysis. Through these methods, significant findings were uncovered from datasets of both pigs and poultry, encompassing crucial patterns, prevalent themes, temporal fluctuations, and geographical spreads. Interpreting these analyses will lead to a better understanding of the patterns and trends in these smallholding communities.

6.3.1 Word frequencies

An exploratory analysis was performed on the two datasets using word frequencies and bigrams to uncover common terms and themes discussed in the forums.

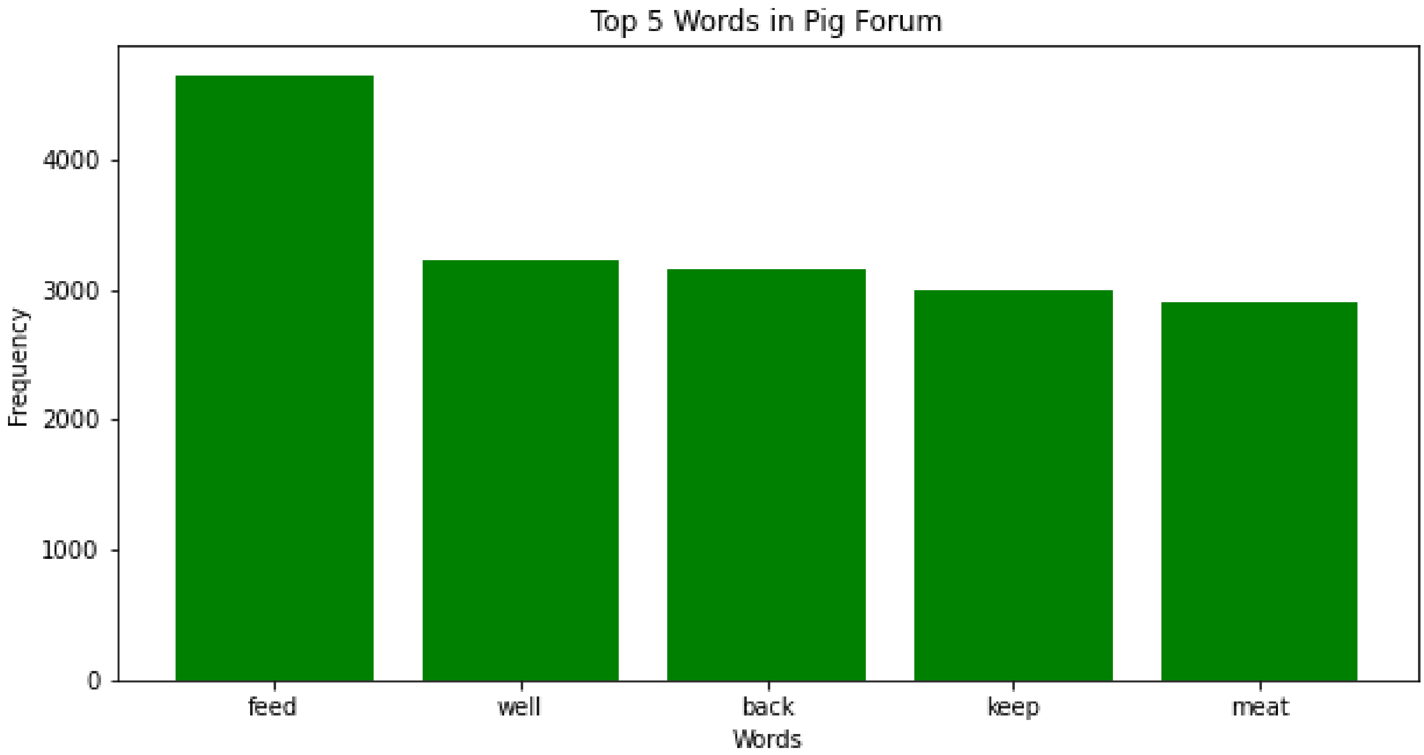


FIGURE 26: PIG WORD FREQUENCIES

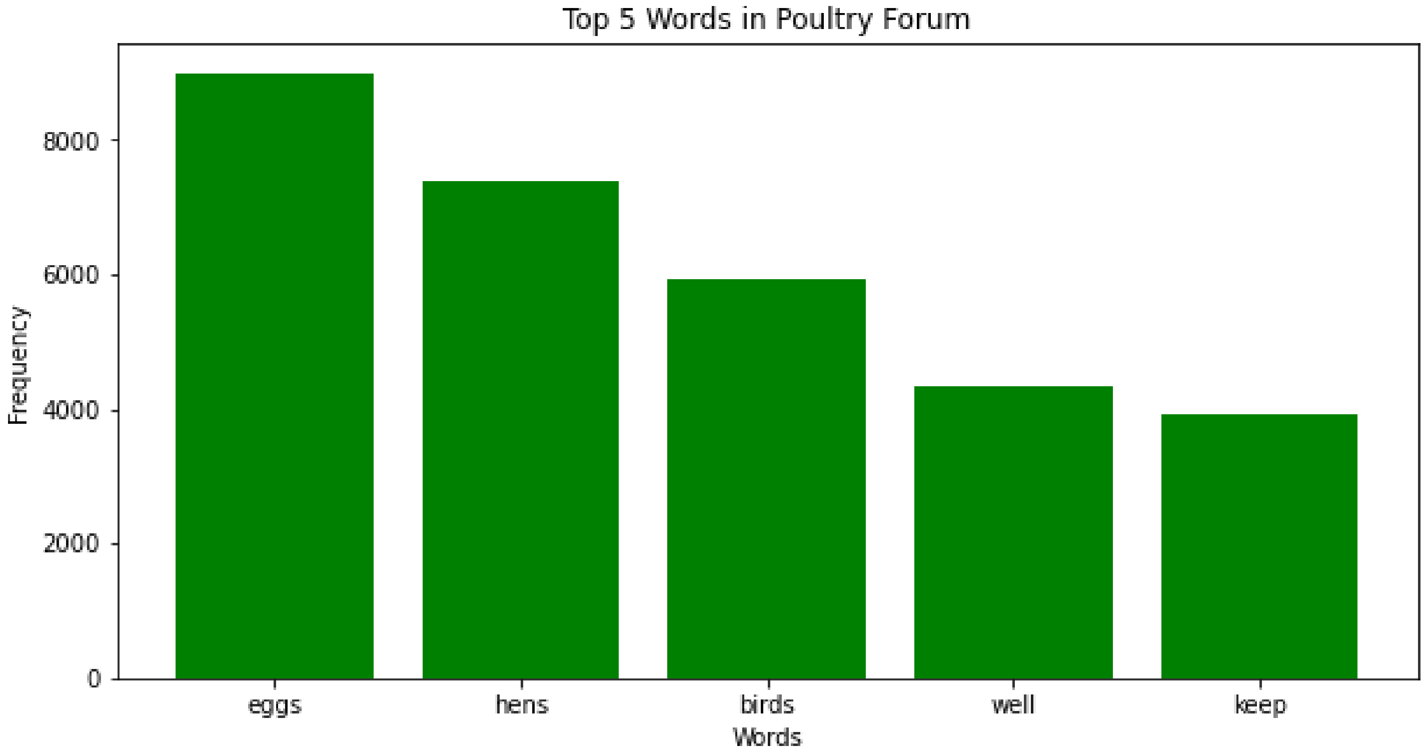


FIGURE 27: POULTRY WORD FREQUENCIES

In Figure 26, a word frequency plot was used to illustrate the frequency distribution of words in the pig forum, highlighting terms such as “feed” and “meat”. The bigram analysis delves deeper into terms that are already recognisable, such as rare breed, slap mark, and electric fence.

As with the pig forum analysis, the most common topic in the poultry word cloud in Figure 27 was animal husbandry. The exploration of this was taken a step further in Figures 32 and 33 with bigram analysis respectively.

6.3.2 Bigram analysis

The results of the bigram analysis are illustrated in figures 28 and 29, with the top 5-word pairings emphasised.

As described by Amalraj et al. (2018), “Kune kune” refers to a breed of domestic pig that is known for its gentle nature and small size[375]. These pig types are likely preferred by smallholder pig farmers because of their manageable size and temperament. The bigram’s high frequency of 442 occurrences supports this claim.

Additionally, “slap mark” is referenced 348 times as the pig’s tattoo for identification. In pig farming, a slap mark is a unique combination of letters and numbers. The Agricultural and Rural economy directorate highlighted the importance of identifying and tracking pigs for smallholder farmers, possibly for regulatory compliance or management purposes[376].

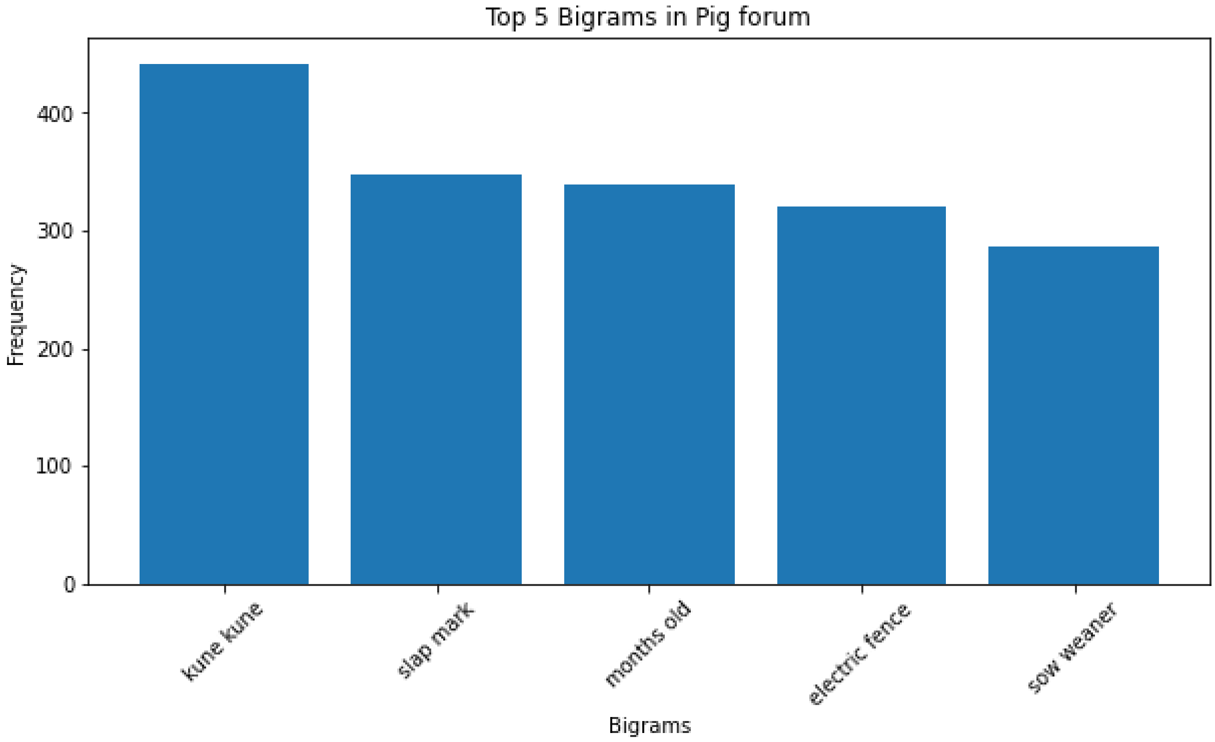


FIGURE 28: PIG BIGRAMS

“Months old” as a bigram, with 339 instances, is most likely used to describe the age of pigs. The influence of age on pig farming is significant as it affects breeding, weaning, and market readiness[377]. When it comes to optimising growth, ensuring health, and making sales or breeding decisions, the age of pigs is a popular discussion topic.

The importance of confinement in pig farming is emphasized with 320 mentions of “electric fence”. Due to their ease of installation and effectiveness in controlling pig movements, smallholder pig farmers commonly opt for electric fences[378]. Effective fencing solutions are crucial in managing and containing pigs, as indicated by the frequency of this bigram. This precautionary step averts losses and liability problems that may arise due to escapes.

A young pig that has been weaned from a sow is called a “Sow weaner,” and this term was mentioned 287 times. It is possible that the term is linked to conversations about breeding and weaning practices. Breeding and weaning piglets are key operational factors for smallholder pig farmers, which can have an impact on herd productivity and financial feasibility, as indicated by the frequency of this bigram[379].

In summary, the bigrams and word cloud illuminate the topics that smallholder pig farmers are most concerned and interested in, including breed selection, animal identification, age management, containment methods, and breeding and weaning practices. Small-scale pig farming's operational efficiency, productivity, and sustainability are likely to be substantially affected by these areas. This establishes the expected terminology in the results of the upcoming topic modelling analysis.

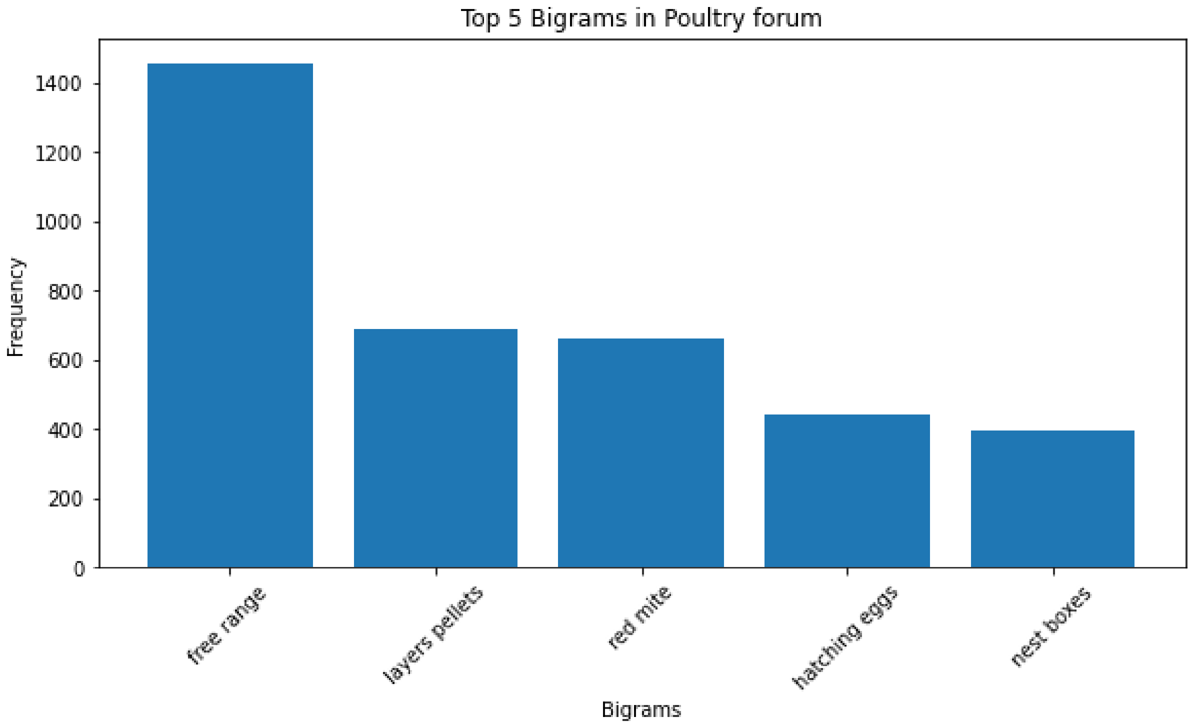


FIGURE 29: POULTRY BIGRAMS

With 1,457 instances, the bigram "free range" ranked first on the list in the poultry dataset results, as shown in Figure 33. Free range is a poultry farming method that allows birds to have unrestricted outdoor access instead of being contained in enclosures. The surge in consumer concern for animal welfare and the perception that free-range items are more organic has led to the increased popularity of this practice[380].

The occurrence of the bigram "Layers-Pellets" 687 times indicates that it is likely a reference to a specific type of feed known as layer pellets. The key to high-quality eggs is the special feed given to laying hens called layer pellets[381].

Hens specifically bred for egg-laying are commonly referred to as "layers". The discussions on layers and pellets are probably centred around the feeding techniques that enhance egg production and meet the dietary needs of laying hens.

The term "red mite" is commonly associated with the poultry red mite, a parasite that frequently infests chickens, with 659 documented cases. These mites can cause irritation, anaemia, or death in infected poultry[382]. A common topic in poultry farming discussions is the prevention and treatment of red mite infestations.

441 instances of the term "Hatching-eggs" suggest discussions about eggs meant for incubation to produce chicks. Proper incubation techniques, maintaining optimal temperature and humidity, and caring for the chicks are all important topics related to hatching eggs.

The bigram "nest boxes" was used 398 times, referring to the designated areas within a chicken coop where hens lay their eggs. Nest box topics may include structure, materials, hen quantity, and sanitation. Ensuring that nest boxes are safe and comfortable is essential in promoting a conducive environment for hens to lay eggs, which affects the quality of the eggs[383]. By analysing these word- pairings, we can gain valuable insights into the interests and priorities of the poultry community. From ethical farming practices to egg production and housing practicalities, various topics are covered.

6.3.3 Topic modelling

After analysing the coherence and perplexity scores for different numbers of topics using an elbow curve, it was determined that the highest coherence score was achieved with 4 topics, with a value of 0.4147, as shown in table 25 and figure 30.

Number of topics	Coherence Score	Perplexity Score
2	0.318705	-7.790232
3	0.313959	-7.705013
4	0.414670	-7.644855
5	0.390472	-7.598243
6	0.312640	-7.565363
7	0.373896	-7.515910
8	0.350921	-7.504429
9	0.362277	-7.456728
10	0.363266	-7.429319
11	0.327299	-7.408429
12	0.357909	-7.361867
13	0.395595	-7.349266
14	0.373376	-7.313781
15	0.381800	-7.297026

TABLE 25: OPTIMUM TOPIC NUMBERS – PIGS

Although the perplexity score continued to decrease as the number of topics increased, indicating better model generalisation, the coherence score, which measures how interpretable the model is to humans, was maximised at 4 topics. As a result, despite the potential for improved model generalisation with more topics, the decision was made to prioritise interpretability and select 4 as the optimal number of topics.

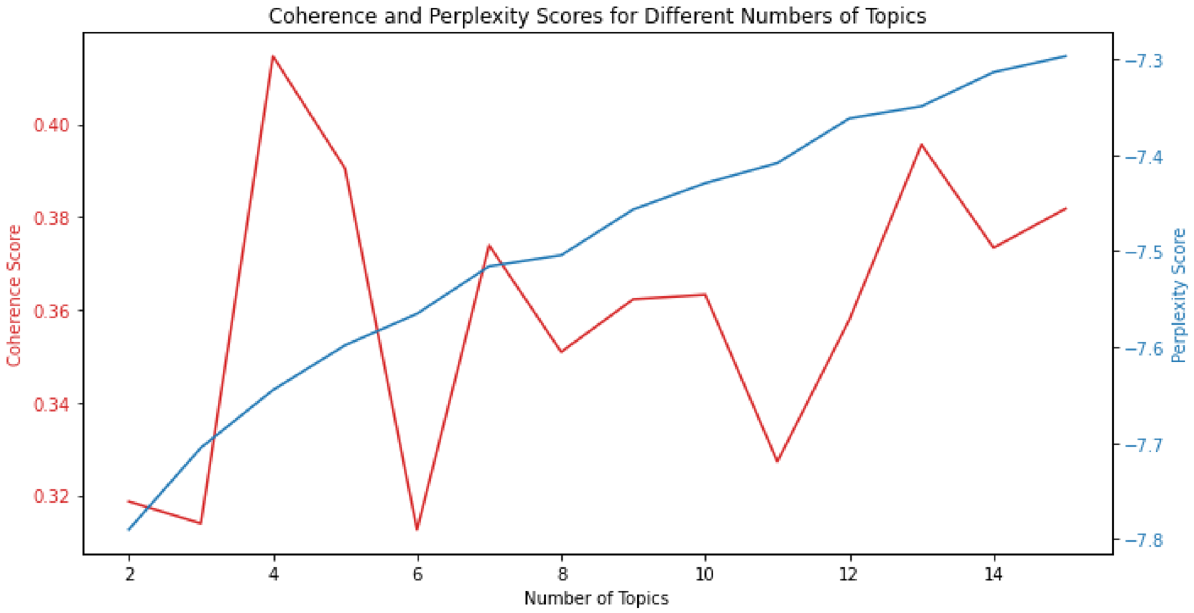


FIGURE 30: ELBOW PLOT – PIGS

The optimal number of topics for the poultry forum was determined to be 14, likely due to its larger corpus, increased user base, and wider range of poultry-related discussions. Furthermore, the 2021 Avian flu outbreak prompted the UK government to impose stricter housing regulations on poultry keepers, in an effort to contain the disease.

The number of topics was plotted against the coherence and perplexity scores using an elbow curve, and it was observed that the coherence score peaked at 14 topics with a value of 0.3529, as displayed in table 26. At 14 topics, the lowest perplexity score (-7.6353) was achieved, indicating a strong ability to generalise in the model. 14 topics were found to strike a balance between model generalisability and human interpretability, as shown by the significant drop in coherence at 15 topics, highlighted in figure 31.

Number of Topics	Coherence Score	Perplexity Score
2	0.266742	-7.934730
3	0.276549	-7.898388
4	0.332996	-7.826787
5	0.321125	-7.826034
6	0.345918	-7.792073
7	0.321604	-7.758866
8	0.334288	-7.736243
9	0.330306	-7.719719
10	0.310561	-7.704776
11	0.317685	-7.699935
12	0.323260	-7.672131
13	0.331440	-7.656394
14	0.352940	-7.635348
15	0.303546	-7.635580

TABLE 26: OPTIMUM TOPIC NUMBERS - POULTRY

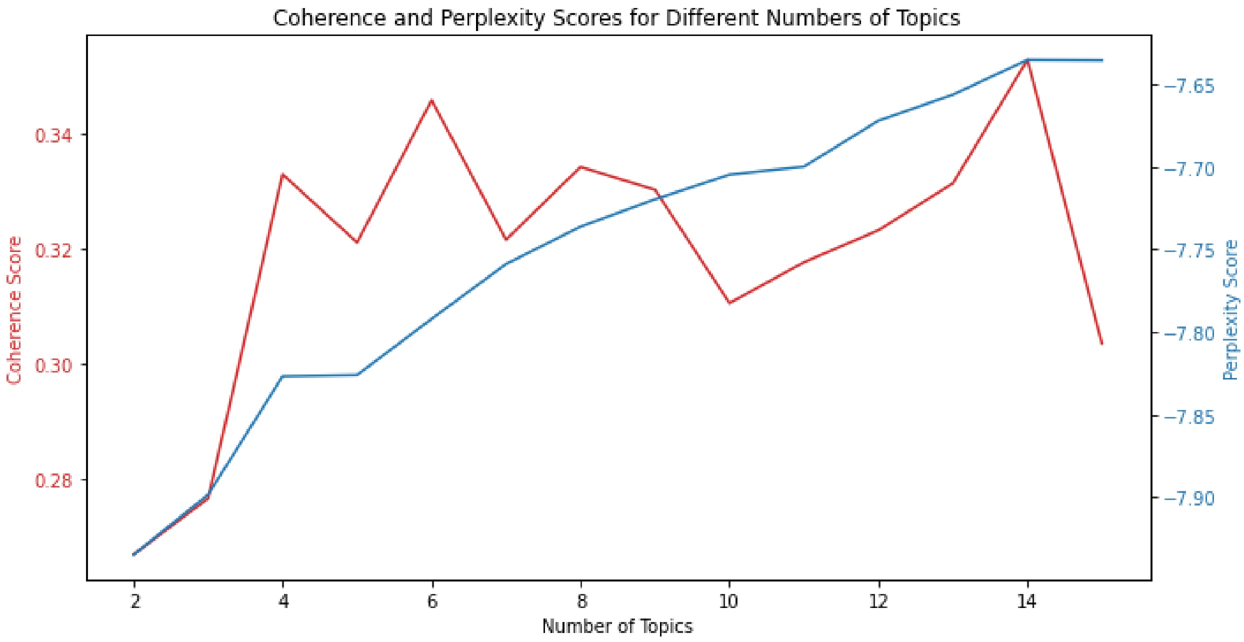


FIGURE 31: ELBOW PLOT - POULTRY

The top terms were depicted, and a topic name was assigned to each topic number based on a feasible name that encapsulates the essence of the top 10 terms, as deemed fit manually. Table 27 displays the topic numbers, the top 10 terms within this topic and the manually assigned topic names.

Topic number	Top terms	Assigned topic name	Example post
0	keep electric fence well back weaners water feed way know.	Pig containment and care	*"Little weaners currently housed in stable next couple weeks lot comfortable…".*
1	feed sow piglets food well weeks around boar feeding keep	Feeding	*"Piglets weeks old, remained solidly sow, done brilliantly. Sow looking really thin, regardless quality, varied food given"*
2	straw back meat feed abattoir well butcher trailer ark used.	Slaughter Processes and Housing Considerations.	*"The butcher priced labeled meat anticipation selling, came home lucky last private butchery prior Christmas, couple pics priced packs wow abattoir butcher, local butcher made sausages last shoulders jointed left."* *"Anyone ideas using hay instead straw bedding? Eat used eat bad thing, tends mustier bedding used, giving fresh often bed stays smelling"*
3	meat tag slaughter tags abattoir know weaners back people mark	Pig Identification and Processing	*"New keeper of weaners arrived, already have plastic button tags breeder tag, slap-mark, remove ear tags already"*

TABLE 27: TOPIC MODELLING RESULTS - PIGS

Based on the high occurrence of terms like "electric fence", Topic 0 was determined to be connected to containment and upkeep. Feeding was the central theme of Topic 1, with "feed, food and feeding" being prominent, Topic 2 was focused on slaughter processes and housing considerations, and Topic 3 centred around identification and processing.

There is a possibility of term overlap as they may pertain to different dimensions of pig and poultry agriculture. The word "feed" is an essential element in both topic 0 (Pig containment and care) and topic 1 (Feeding) due to its significance in pig care and feeding practices. Given their involvement in pig identification and processing, as well as the slaughter process, abattoirs are a critical component of both topic 2 and topic 3.

Topic number	**Top terms**	**Assigned topic name**	***Example post***
0	birds hens eggs old meat laying keep around cockerel lay	Egg Laying and Breeding	*“Three-year-old hens used to give three eggs, stopping short in winter spring, laid eggs, stopped laying last month, showed mucky bottoms. Asked if wormed yet, reply advised to buy treated feed, bottoms cleared, still no egg. Anything to return to laying”*
1	birds coop hens well eggs feeder sand work old	Equipment and Maintenance	*“Used sand in coop, allows poop to clump easily scooped like cat litter. Environmentally, wood shavings claim sustainable but consider carbon footprint, sand used long term and cleaned easily."*
2	geese bit may waterfowl see keep back run birds	Waterfowl	*“During a flu lockdown, buried under a thick blanket of snow, bought an expensive sack of goose and waterfowl pellets, but geese simply won't eat it”*
3	hatch hens broody chicks water ducks hen eggs used weeks	Hatching and Raising Chicks	*“Home hatched Saxony ducks have started laying. The drake has been seen treading the ducks, hoping the eggs are fertile. We took an egg to the table, assuming they aren't regular yet, but they have laid together since Monday. Is there anything should provide or anything to help a broody hen sit”*
4	care birds eggs water clean hen days well run coop	General Poultry Care	*“Moved to a smallholding, fenced the run, and cleaned and painted the coops to weatherproof them. Nearly ready to introduce the birds. Wondering how many hens to start with, as the coops and run are quite large. Also inquiring about bedding—wood shavings vs. straw and how to manage*

			food and water for the best health and minimal me”
5	eggs ducks hens old well house ones coop	Housing and Shelter	*“Researching keeping ducks and read they can be kept with hens. Anyone successfully use a space in the hen house for ducks? Thinking about Campbells or Appleyard”*
6	rats run hens house keep around rat birds fox mite	Pest and Predator Management	*"To control mites in the coop, focus on cleanliness and regular treatment of the birds and their environment. Dusting with diatomaceous earth helps, as does ensuring the coop is well ventilated and the bedding is changed regularly. For larger predators like foxes, secure fencing with buried edges and a covered run are essential to keep the birds safe”*
7	hens feed well birds cockerel eat	Nutrition and Feed	*“Any advice on feeding? Currently, I feed them layer's pellets which are high in calcium. Should I keep this up, or switch to a breeding ration as they're free-range and don't eat much ready-made feed?"*
8	birds hens eggs breed last white hen	Poultry Breeds	*“"NHB breeders group had a decent season with breeding groups of either pure bred NHBs or pure cocks with Dutch hens. Inferior birds have gone to laying homes, inferior cocks have been eaten.”*
9	hens ducks water hen food care keep feed	Ducks and Waterfowl Care	*“How do people manage to keep ducks and hens together without creating a huge mess? Ducks make everything so wet and muddy. Any specific coop designs or management tips would be greatly appreciated to help keep the coop environment clean and healthy for both types of birds."*

10	birds eggs geese hens back ducks free range	Free-Range Management	*“Debating whether to let my ducks and hens free-range together. I have separate areas for them at night, but during the day, would they benefit from roaming together”* *“For those with free-ranging flocks, how do you manage the risk of predators?”*
11	chicks well eggs raise geese gosling feed weeks	Raising Chicks and Goslings	*“Considering raising chicks and goslings together. I've read that goslings can be raised by hen mothers, but are there special considerations to keep in mind regarding feed and the physical environment to accommodate both?"*
12	eggs hatch goose geese nest incubator hen days keep egg	Incubation and Hatching	*“Advice needed for a broody hen incubating goose eggs: should I let her continue, or would it be better to move the eggs to an incubator to ensure better control over the conditions? The hen has been diligent, but I'm unsure if the temperature and humidity needs for goose eggs are different."”*
13	eggs sell egg selling duck birds people hens	Egg Production and Selling	*“Considering selling eggs from my free-range hens at a local farmer's market. I know I need to keep track of the eggs' freshness and ensure proper handling”*

TABLE 28: TOPIC MODELLING RESULTS - POULTRY

As per the data in Table 28, poultry forums demonstrate a greater scope of themes, including discussions that delve into specific species. Analogous to table 26, there can be an overlap of terms in poultry farming owing to their significance to multiple elements of the subject. The presence of the term “eggs” in several topics can be attributed to the crucial role of egg production in various aspects of poultry farming, including breeding, hatching, and selling. Similarly, due to their crucial function in poultry husbandry, the term “hens” crops up in numerous discussions on the topic.

6.3.4 Anomaly detection

Several anomalies were identified in the results from Figure 32, pointing to dates with unusual pig livestock-related posts or discussions. The presence of numerous high peaks during June and July 2017 may imply a seasonal inclination or transformations in the industry. In the initial phases of the COVID-19 outbreak, there was another notable peak in February 2020 with 30 posts, potentially linked to the collateral impact on pig farming. In addition, intermittent postings were detected during slow periods, potentially attributable to vacations or other happenings. There were more occurrences of these between 2017 and late 2019, with only a few days of low activity after 2020. The occurrence of over 20 posts on different days across different years could be a regular event that sparks more discussions.

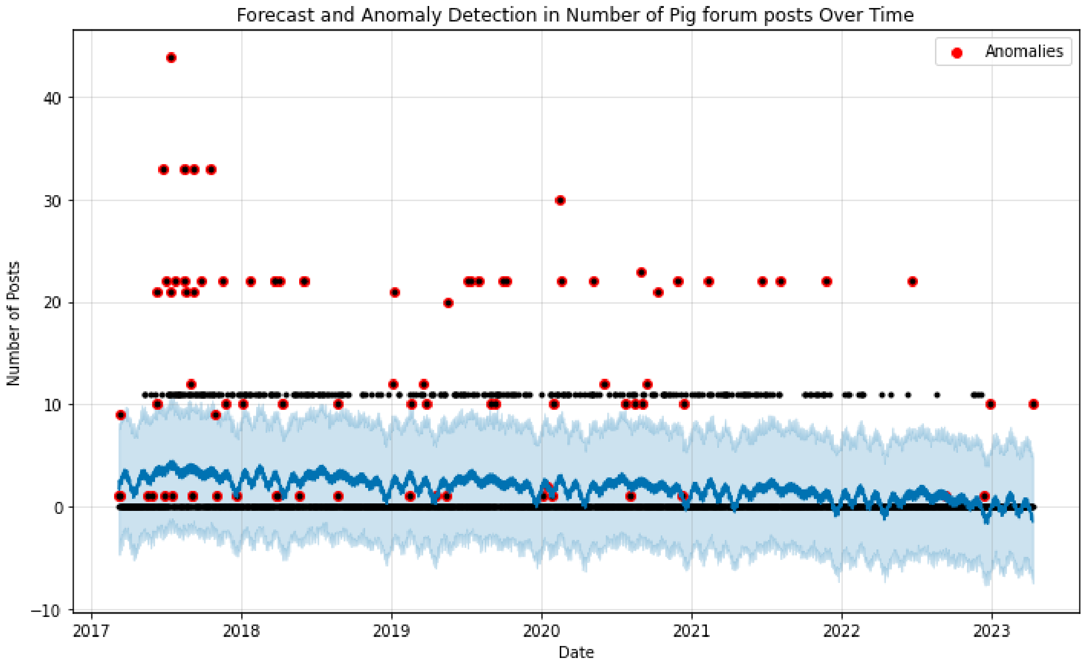


FIGURE 32: ANOMALY DETECTION - PIGS

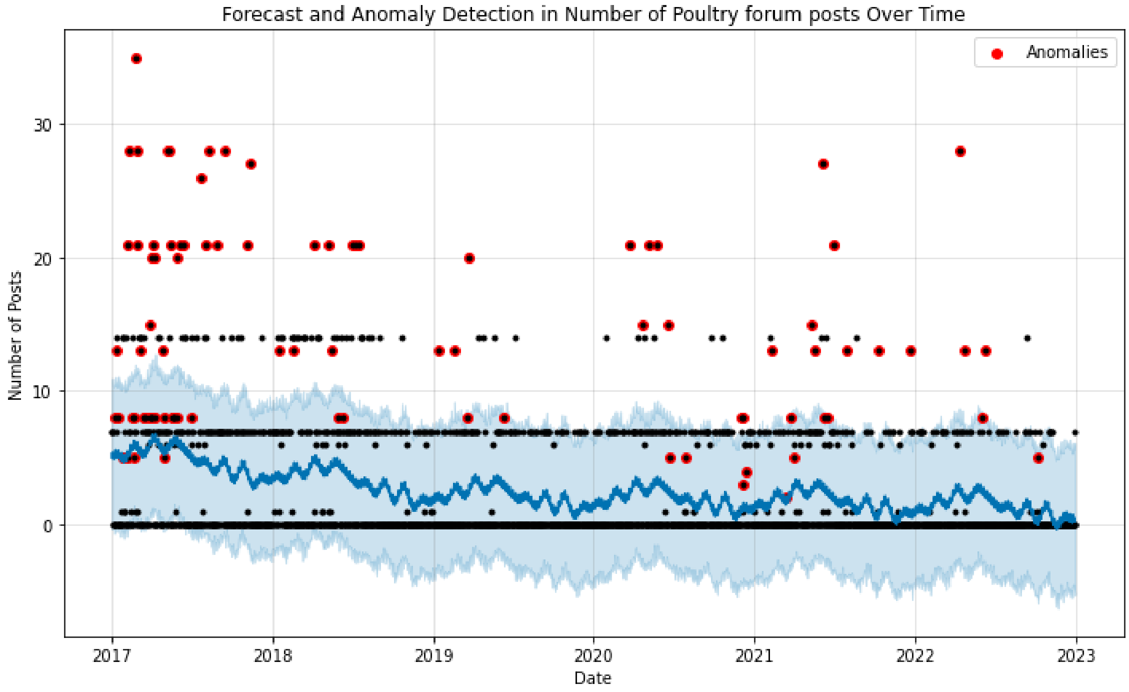


FIGURE 33: ANOMALY DETECTION - POULTRY

The anomaly detection outcomes for the poultry dataset are illustrated in Figure 33. Compared to the pig group, this group experienced more variations in trends and a larger number of anomalies. The spring breeding season could be the reason for the unusually high number of posts during the first quarter of 2017. The COVID-19 pandemic and an influx of newcomers seeking advice may have contributed to the significant increase in posts from March to May 2020.

Conversations about adapting to or recovering from the pandemic or early spring-related discussions may have caused anomalies between February and March 2021. Anomalies between April and June 2022 may be due to seasonal activities and new housing regulations in the poultry industry.

Considering events related to public and animal health is crucial to comprehend these anomalies. Comparing these anomalies to past events can give us valuable information about the link between discussions and real-world happenings.

6.3.5 Spatial analysis

Another element to the analysis was the examination of the spatial distribution of the users who frequently posted in both discussion forums. A visual representation of user distribution was created by mapping these locations, which helped identify regional concentrations and trends within the UK. Understanding the geographical context allowed for a more nuanced view of the smallholder community, connecting discussions of pig and poultry farmers to their surroundings.

Geolocation data integration enhances the potential of a surveillance system. It allows the tracking and prediction of the dissemination of specific topics or concerns across different regions among small-scale farmers, such as the outbreak of diseases, recommended methods, or resource accessibility. Identifying regional concerns/patterns through geographical clustering of topics/sentiment enables targeted interventions. A surveillance system that considers spatial information allows for a more precise, efficient, and prompt response that considers both the problem and the areas it affects the most.

The choropleth map featured in Figure 34 illustrates the population density of pig forum users. Firstly, Torquay, in Southwest England, has the most users, followed by Devon. The ideal conditions for pig farming may exist in these areas due to their mild climate and fertile soil. In the Midlands, Montgomeryshire, and Herefordshire also show a significant proportion of users. Pork products can be efficiently distributed from these areas because of their moderate climate, balanced soils, and central location. The strategic demographic positioning of the Midlands, Montgomeryshire, and Herefordshire supports pig farming due to their physical centrality and a demographic profile characterized by higher education and technological adoption in agriculture. The alignment of demographics with innovative farming practices may be reinforced by local agricultural policies and investment in Agri-tech education, indicating a propensity for embracing technological solutions to improve farming efficiency.

Additionally, SE Wales has smallholder pig farmers due to its fertile land suitable for mixed farming. Appleby-in-Westmorland's farming tradition could be advantageous for pig farming. Apart from its famed Suffolk pig breed, there are also many users from this vicinity. Despite their colder climates, Aberdeenshire, Morayshire, and Hawick in the Scottish Borders have a moderate number of users. Local conditions may have prompted farms in these regions to adapt or specialise in pig breeds suited to the environment. Lastly, London has little smallholder pig farming activity due to its urban nature, high land prices, strict regulations, and reliance on imported pig products.

A more granular breakdown of the most frequent locations was mapped with their corresponding bigrams in order to evaluate the frequent terms associated with each region. Analysing bigrams split by location offers an additional layer of granularity, potentially enabling policy makers to tailor interventions based on regional nuances in discussion themes. The results from this are presented in tables 28 and 29.

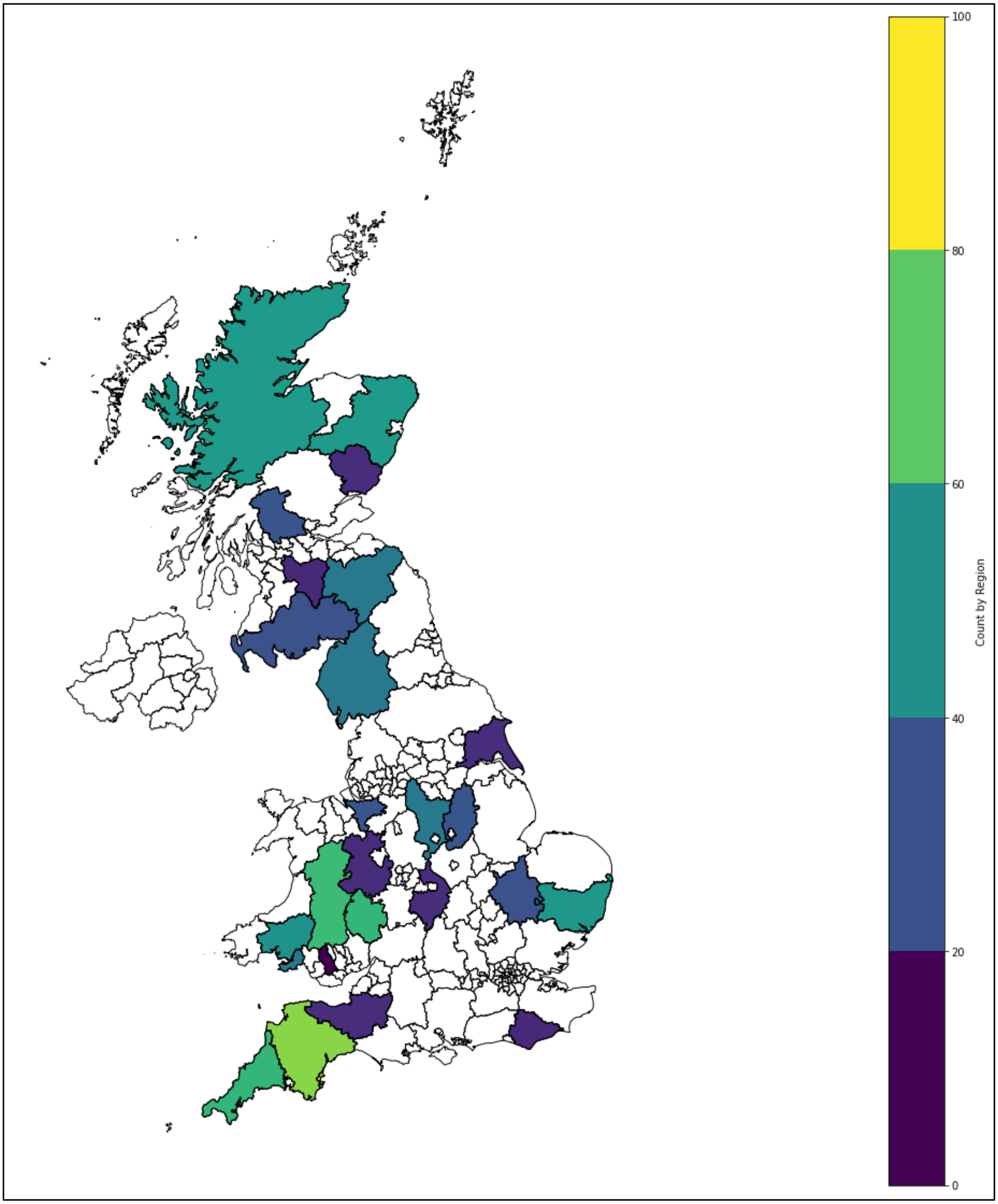


FIGURE 34: PIG FORUM USER MAP

Likewise, figure 35 offers a similar visualisation of UK poultry forum users and their poultry farming activity across various regions. The central belt of Scotland, including Fife, South Lanarkshire, and Clackmannanshire, has the highest concentration of poultry forum users. Although the fertile soil and mild weather are ideal for raising poultry, a demographic analysis uncovers more complexities. Robust agricultural education infrastructure and active engagement in agricultural cooperatives benefit these regions and can boost community-based farming initiatives. The demographic makeup of these regions often consists of a larger rural population that values agriculture and sustains it through policy measures that promote small-scale farming and sustainability.

Wales yielded the second largest concentration of users after Scotland, with Montgomeryshire and Gower providing suitable conditions for poultry farming. Lastly, England displayed a sparser distribution, with the majority of users stemming from Yorkshire and the Midlands, in particular, Leicester, Lincolnshire, and West Yorkshire. This may be due to mixed crop and poultry farming systems within these areas as a result of productive and arable land.

Whilst location provides a visual cue of the geographical distributions, it can’t alone be used as a gauge for correlations in pig and poultry farming activities. Other factors such as demand, market access, cultural preferences, and farming traditions also influence smallholder farming distribution.

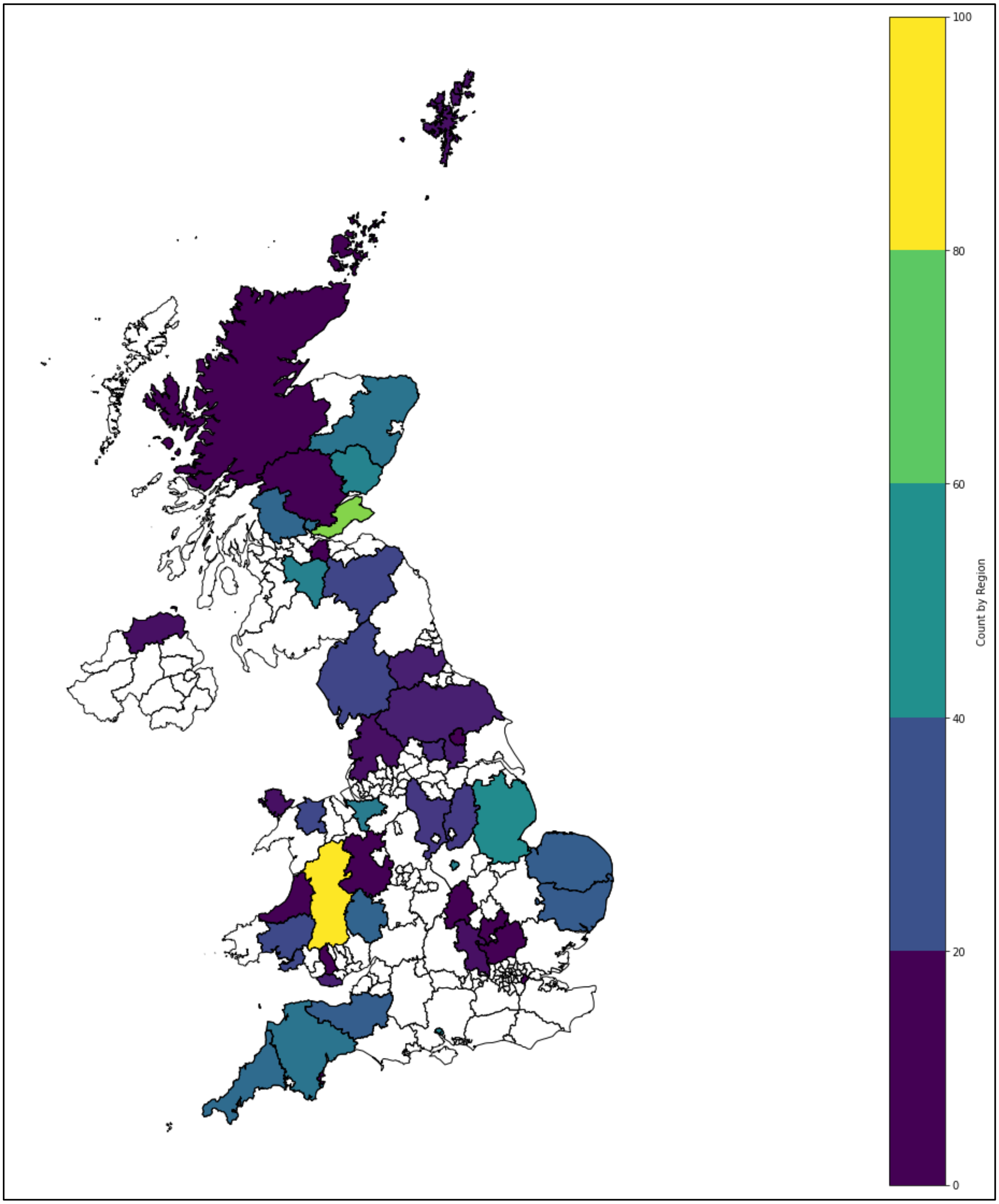


FIGURE 35: POULTRY FORUM USER MAP

Locations	Top two bigrams	Evaluation
Torquay	“Electric- wire”, “many - advice”	The bigram “electric wire” likely refers to the use of electric fencing, a common method of pig containment. “Many, advice” could imply that the forum members in this location are actively seeking or giving advice, indicating a highly engaged community.
Devon	“Butcher - priced”, “always- buy”	The term "butcher, priced" indicates discussion around the costs of butchering, possibly comparing home butchery versus professional services. "Always, buy" could refer to discussions about what supplies or breeds are essential to buy for successful smallholding.
Montgomeryshire	“Old - age”, “traditional - breeds”	The bigrams "old, age" and "traditional, breeds" suggest a concentration on older pigs and heritage pig breeds. This could indicate a community interested in slower, more traditional farming practices.
Herefordshire	“Boned - rolled”, “charged - per”	This suggests a strong focus on the processing and costs associated with pork production, which could imply a community focus on commercial pig farming.
Aberdeenshire	“Water - carrier”, “Abattoir - options”	These bigrams imply conversations around basic pig care needs and slaughtering options, showing a pragmatic focus in this community.
SE Wales	“Many - years”, “years – last”	This could indicate long-term pig keepers sharing their experiences and might suggest a forum full of experienced smallholder

Appleby-in-Westmorland	“Three – weeks”, “large – black”	“Three, weeks” could refer to key milestones in pig development, while “large, black” is a popular heritage breed of pig, showing a community interested in specific pig breeds and their care.
Hawick, Scottish Borders	“Scottish – borders”, “borders – willing”	This points to location-specific discussions and could indicate a community that is regionally focused or dealing with location-specific issues.
Morayshire	“Play – fingers”, “wish – suggest”	These terms could be part of broader discussions or idiomatic expressions prevalent within this community.
Suffolk	“Wild – boar”, “Rare – breed”	Both these terms relate to types of pigs, indicating that there is a strong interest in diverse and potentially non-traditional pig breeds in this community.

TABLE 29: TOP 10 LOCATIONS AND CORRESPONDING BIGRAMS - PIG FORUM

Locations	Top two bigrams	Evaluation
Fife	“Dry – soil”, “Quote – Northfifieduckling”	The term “dry soil” could be associated with discussions around ideal conditions for poultry health or possibly the effects of weather or soil quality on feeding. “Quote, northfifeduckling” suggests a specific user named ‘NorthFifeDuckling’ who is often quoted and could be a prominent and respected member of the community.
South Lanarkshire	“Wee – hen”, “Water – overnight”	Wee, hen” is a colloquial term often used to describe a small or young chicken. This could suggest a lot of discussions revolving around the rearing of young poultry. "Water,

		overnight” could point to discussions around best practices for providing water for poultry, especially during nighttime.
Leicester	“Dual – purpose”, “Large – fowl”	These are clear terms relating to the type of birds being reared – dual-purpose refers to breeds that are good for both egg-laying and meat, while large fowl refers to larger breeds of poultry. This could suggest a local focus on practical, self-sustaining farming.
Gower	“Free – range”, “Wild – birds”	“Free range” could indicate a focus on animal welfare and ethics, and “wild birds” might indicate discussions around avian interactions or potential disease transmission.
Spalding	“Goose – eggs”, “Back – lay”	The terms suggest a strong focus on egg production. “Goose eggs” indicates a possible specialty or interest in geese, while “back, lay” could be referring to issues with hens resuming laying after a hiatus.
Powys	“Hand – plucking”, “Growers – pellets”	These terms could suggest a concentration on traditional methods of poultry care and feeding, indicating a more hands-on approach in this community.
Devon	“Nest – boxes”, “Red – mite”	“Nest, boxes” and “red, mite” appear often, indicating discussions around creating comfortable nesting spaces and dealing with common poultry pests.
Montgomeryshire	“Weight – loss”, “Duck – eggs”	Weight loss in birds could be a concern for farmers here, indicating health or nutrition discussions. “Duck eggs” implies a substantial focus on duck rearing and egg production.
West Yorkshire	“Layers – pellets”, “Soaked – wheat”	“Layers, pellets” and “soaked, wheat” suggest that discussions are around specific types of feed, indicating a strong focus on poultry diet and nutrition in this location.

Clackmannanshire	“Eating – drinking”, “”Food – water”	These terms imply discussions revolving around the basic care needs of poultry.

TABLE 30: TOP 10 LOCATIONS AND CORRESPONDING BIGRAMS – POULTRY FORUM

6.3.6 Comparative analysis

This section presents the results from the comparative analysis between the two forums, with overlaps in common themes and locations being highlighted.

Raising larger, more demanding animals, such as pigs, poses challenges for smallholders in managing health, containment, and meeting regulations. These forums often center around discussions of breed choice, feeding practices, and sustainable husbandry. Conversely, small-scale poultry farmers face a distinct set of obstacles, as they work with smaller animals and potentially larger quantities. Flock management, disease prevention, egg production, and processing are common topics in their discussions.

The common bond between both groups is the responsibility to steward their animals and land, despite differences. They have comparable limitations and advantages concerning factors like market entry, adhering to regulations, available resources, and the unpredictable British weather. Examining the overlap in their discussions and locations allows us to understand their shared experiences and mutual influences.

The overlap in posting frequency between both forums is illustrated in Figure 36. The poultry sub-forum showed the most significant fluctuations, with over 120 posts per day in the first quarter of 2017, followed by a sharp decrease. The pig forum reached its highest number of posts in mid-2017, surpassing 60, and has since remained stable along with the other forums. One reason for this might be that most topics and questions have already been answered in previous years, resulting in fewer discussions among users now. Furthermore, users may have shifted their discussions to other platforms such as Twitter and Facebook.

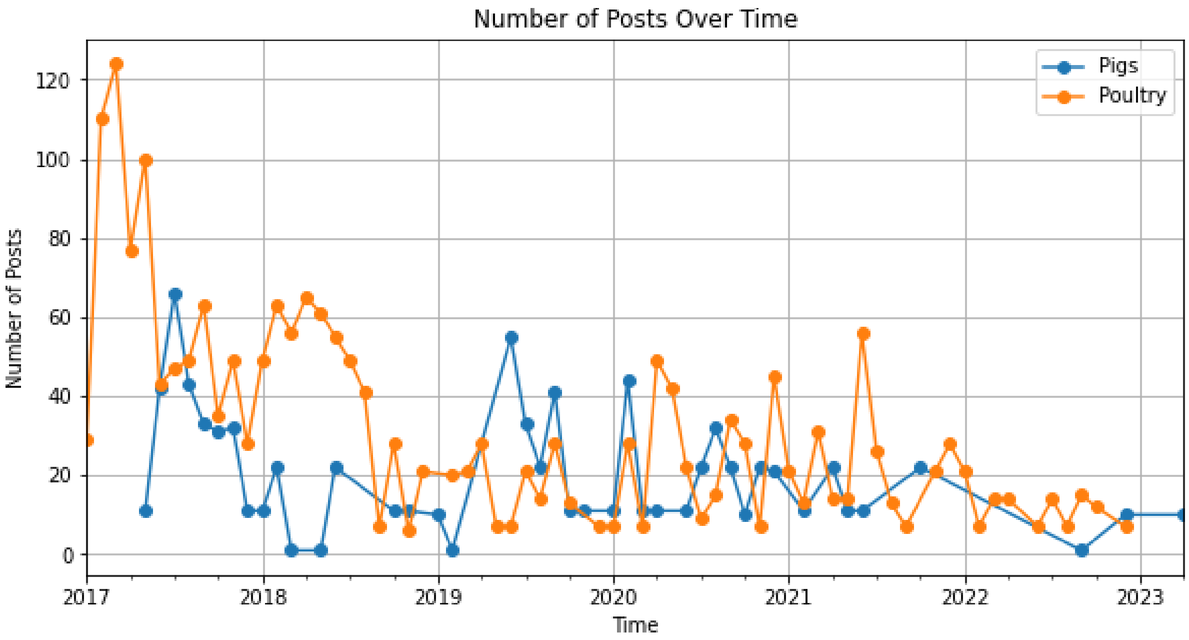


FIGURE 36: TEMPORAL POST FREQUENCY

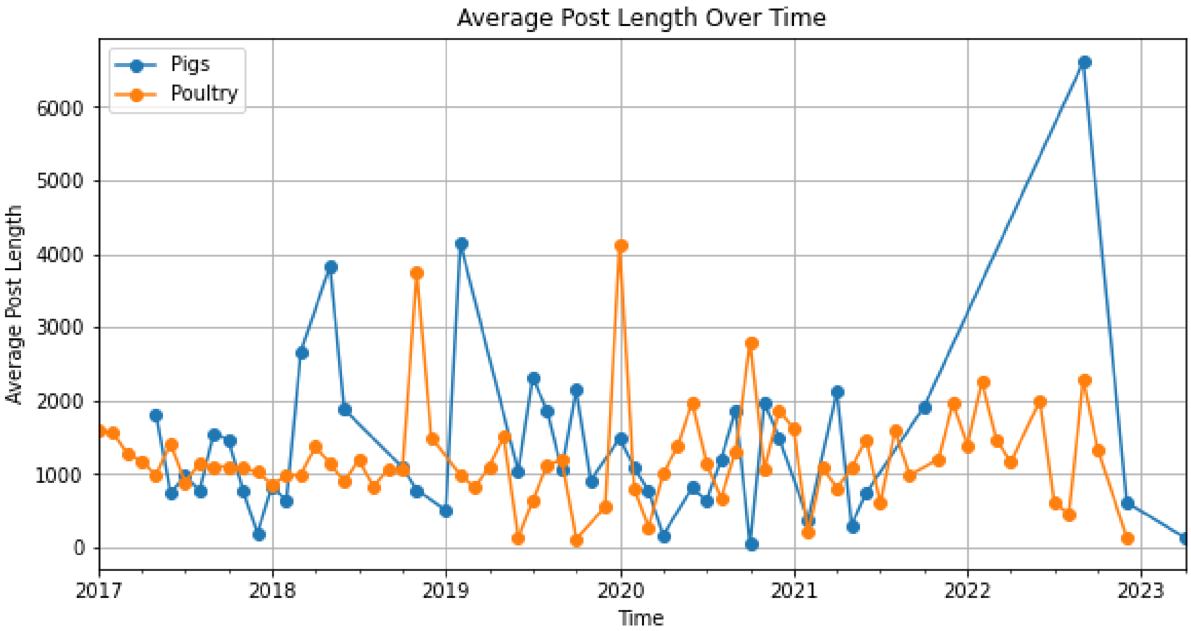


FIGURE 37: AVERAGE POST LENGTH

Figure 47 depicts the average post length from both forums. A significant increase was observed in the pig forum discussions while investigating the evolution of average post length in both the pig and poultry forums from 2021 to 2023, in contrast to the poultry forum's stable average post length.

There might be multiple causes for this pattern. It's possible that the changing circumstances in the smallholding community or farming industry may have led to a shift in the complexity or depth of discussions on the pig forum.

Longer posts and in-depth discussions often occur when users exchange information and advice in response to changes in legislation, disease outbreaks, or shifts in market conditions.

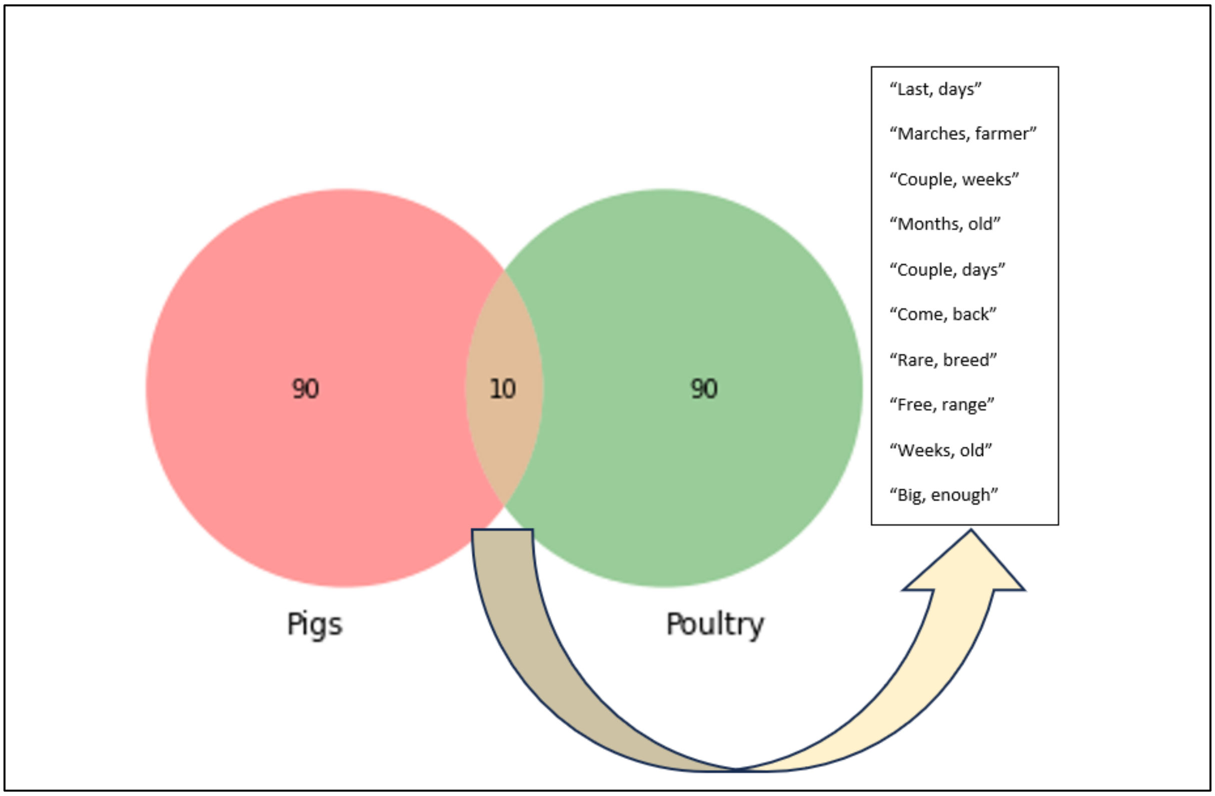


FIGURE 38: BIGRAM VENN DIAGRAM

A further examination into the similarities and differences between the two cohorts was examined in a Venn diagram of the top 100 bigrams, as depicted in figure 38.

The most frequently occurring theme is based on the concept of time. The bigrams "last days", "couple weeks", "couple days" and "months old" indicate that there are frequent conversations about the animals' age and growth rate. Based on this, it can be inferred that the developmental stages of pigs and poultry are of great interest to smallholders, who may need to determine the optimal time for activities such as breeding, selling, or slaughtering.

Additionally, the bigram "free range" indicates a preference for ethical animal farming, which aligns with a presumed dedication to animal welfare and high-quality products. The practical aspects are also emphasised, and the term "big enough" may refer to discussions about animals being suitable for specific purposes like selling or breeding.

Figure 39 displays the unique and overlapping geographical areas between pig and poultry smallholding forums and presents an insight into locations of these livestock keepers. With favourable climates and a mix of urban and rural environments, High Peak and Stirlingshire in Central Scotland, and Devon are potential locations for pig and poultry farming.

Due to their agricultural roots and plentiful resources, Llandeilo in Carmarthenshire, West Cornwall, and Morayshire are ideal for smallholding farming. Furthermore, regions like West Yorkshire and South Shropshire, which have a rich heritage of mixed farming and supportive communities, have access to infrastructure such as local farmer markets and transportation links. The appeal of Aberdeenshire, Suffolk, and Herefordshire to pig and poultry smallholders could be attributed to their fertile lands and favorable climate. Meanwhile, regions like Hawick in the Scottish Borders, Montgomeryshire in Wales, and Carmarthenshire offer supportive local initiatives for small-scale, sustainable farming. Finally, towns such as Torquay offer exclusive benefits such as easy access to marketplaces and tourist attractions, which can help small-scale farmers who sell directly to consumers or those involved in agritourism.

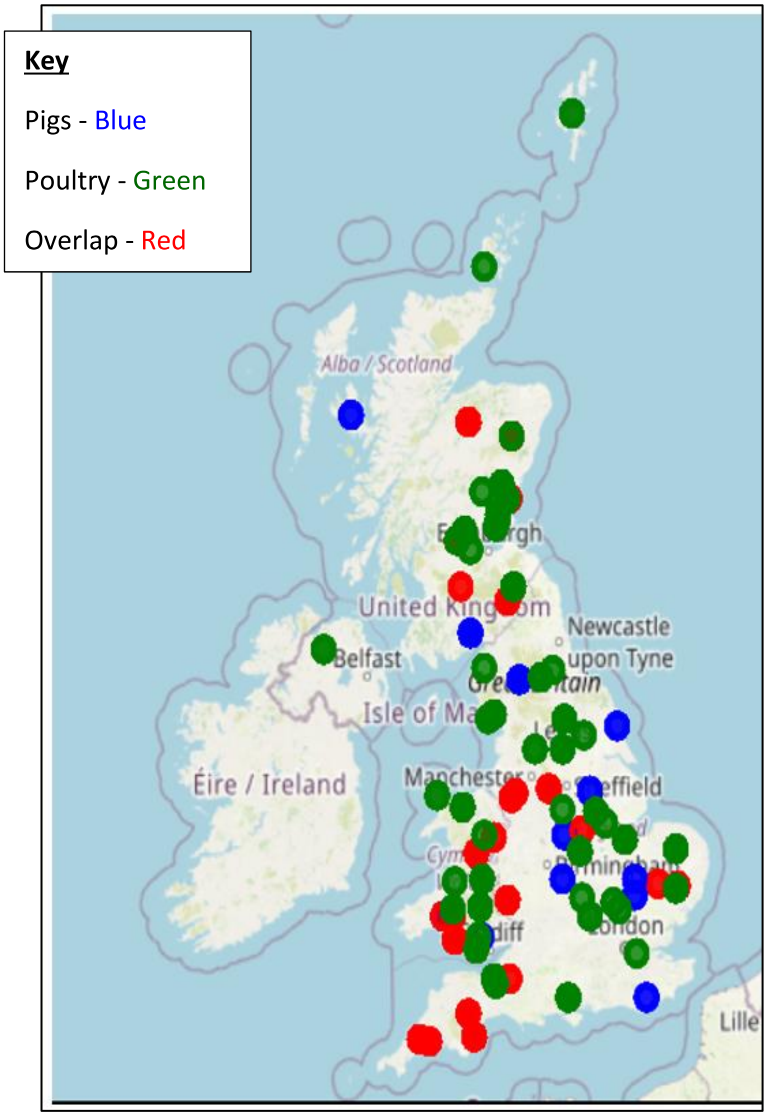
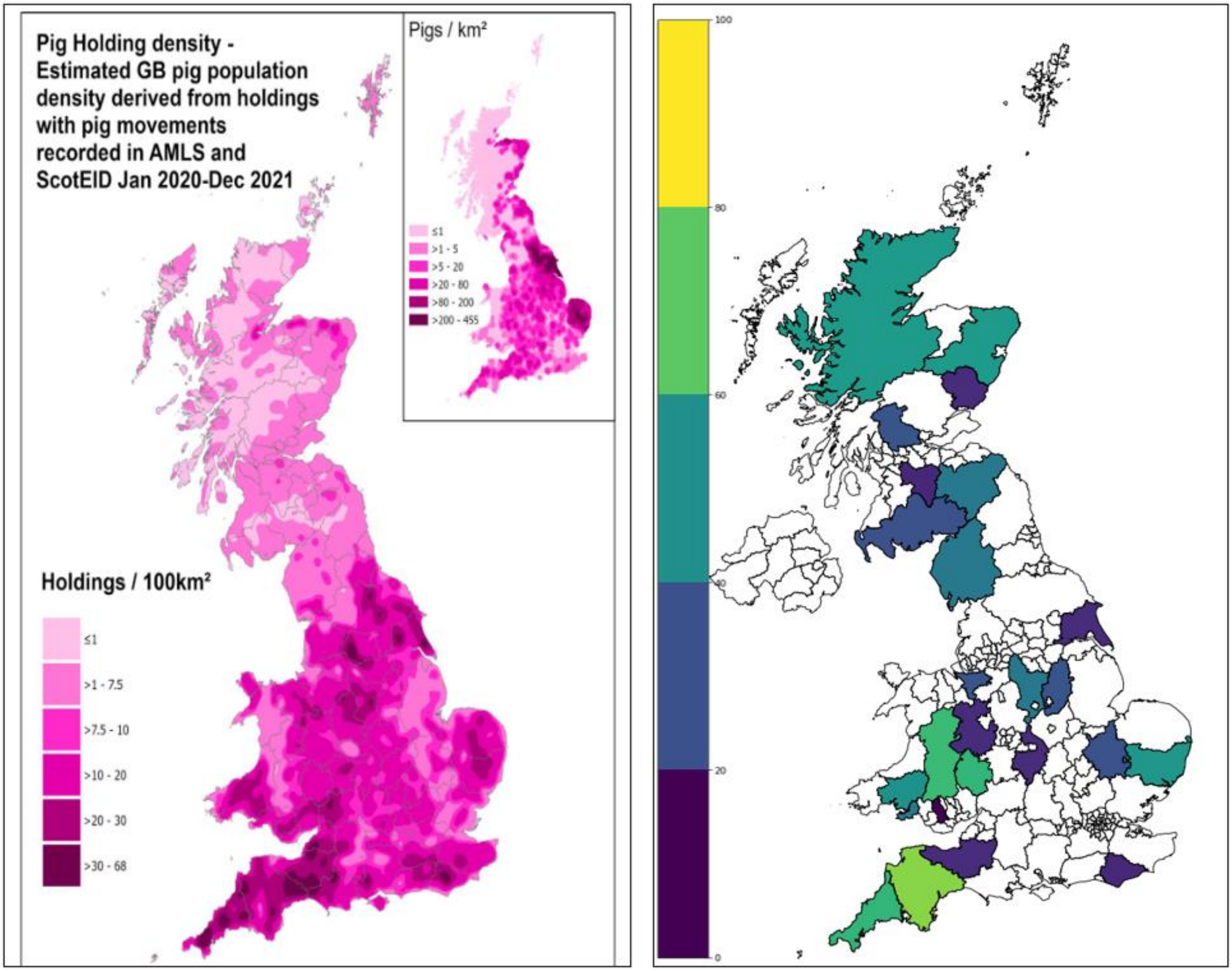


FIGURE 39: GEOLOCATION OVERLAP BETWEEN SUBFORUMS

6.3.7 Comparisons with known holding locations

To further validate these geolocation findings, and to address RQ2 and RQ5 of this project, APHA official pig and poultry holding maps were extracted from official reports conducted in 2023 and 2022 respectively – These are the most recent publications at the time of writing this thesis.

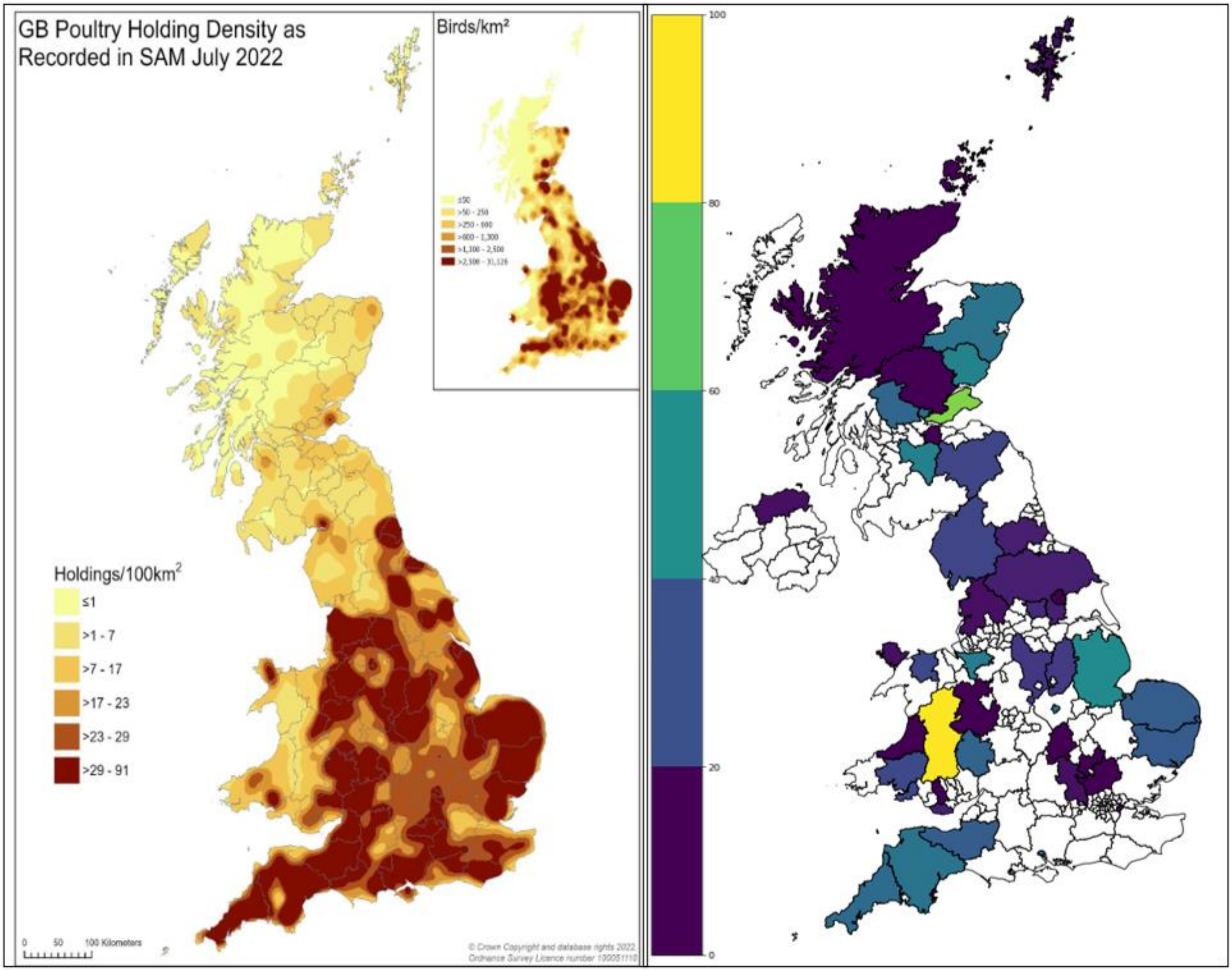

Pig Holding density - Estimated GB pig population density derived from holdings with pig movements recorded in AMLS and ScotEID Jan 2020-Dec 2021
Pigs / km²
≤1
>1 - 5
>5 - 20
>20 - 80
>80 - 200
>200 - 455
Holdings / 100km²
≤1
>1 - 7.5
>7.5 - 10
>10 - 20
>20 - 30
>30 - 68
100
80
60
40
20
0

*SOURCE 1[384]

FIGURE 40: APHA RECORDED PIG HOLDINGS VS FORUM-EXTRACTED LOCATIONS

Figure 40 displays a juxtaposition of the APHA recorded pig holdings and forum extracted locations. Whilst the APHA map shows heavier concentration in the Midlands, Yorkshire and Northeast of England, the forum map displays a greater concentration within Scotland – Both centrally and in the Highlands. Both maps capture the concentration within the Southwest and Southeast of England.

The largest disparity between both maps is visible in the North East, with no recordings being found in the forum users, yet a significant proportion of holdings were recorded. Multiple factors can explain the significant difference between government data on pig holdings and user locations extracted from pig forums in the UK's North East. Pig farmers may have different levels of engagement with online communities, depending on their digital literacy skills and availability of internet access. Moreover, farmers may choose not to share location data out of concern for privacy, especially considering the potential biosecurity hazards linked to revealing farm locations. The collection methods themselves also create variations: government records are gathered systematically and are likely more complete, while forum data depends on voluntary, self-reported information (i.e. free-text locations), making it susceptible to biases and lacks validity.



*SOURCE 2[385]

FIGURE 41: APHA RECORDED POULTRY HOLDINGS VS FORUM-EXTRACTED LOCATIONS

Similarly, figure 41 displays the recorded poultry holdings compared to the forum-extracted locations. An immediate observation is the sparse concentration of holdings within Scotland, especially around the Highlands, which is in contrast to the data collected from the forum. As previously mentioned, recording poultry holdings is substantially more difficult as flocks under 50 do not require registration, therefore the official recordings will be significantly underreported. Both maps show similarities within the midlands, Southeast and Southwest of England.

6.4 Discussion

This chapter offers new insights into the priorities and concerns of small-scale farmers in the UK by analysing forum data. Bigram and topic modeling analysis were used to identify key themes in discussions about pig farming, including breed selection, animal identification, age management, containment strategies, and breeding and weaning practices. The frequent appearance of the word pairing “Kune Kune” indicates that this breed of domestic pig is popular among smallholders, possibly due to its small size and gentle nature. Terms such as “slap mark” for pig identification tattoos and “electric fence” reflect the importance of identification and containment in pig farming.

Discussions in the poultry community focused on free-range methods, layer pellet nutrition, red mite infections, egg hatching, and nest boxes. The frequent use of the term “free range” as a bigram suggests that poultry farming places a strong emphasis on animal welfare and ethical farming practices. Conversations about layer pellets and red mites highlight the critical role of nutrition, health, and parasite prevention in optimising poultry production.

Although the use of topic modeling in the veterinary field is still limited, LDA applications combined with outbreak detection methods have been used to identify disease in UK dogs and have proven to be an accurate predictive tool [386]. However, data on companion animals is much more abundant (200,000+ records) and accurate than data on small-scale livestock, making direct comparisons with our research difficult.

Further work is needed to build on the results of our anomaly detection analysis to develop a practical outbreak detection surveillance system that can detect spikes in posting activity around certain topics. Early warning systems can be created by detecting anomalies in user activity within a given time period, as demonstrated by the significant increase in research during the COVID-19 pandemic[387]. Increases in posting frequency may provide opportunities to respond to disease outbreaks more quickly, as shown by research on endemic disease surveillance[388]. Outlier detection using thresholds can be an effective tool for epidemiological disease surveillance, but a more dynamic source of information such as Twitter may be a better platform for data imputation due to its larger user base and ability to generate more online traffic.

It is important to distinguish between publicly available data such as this forum and more detailed information that is only available to the government, including precise farm locations, demographics, and personal information. APHA regularly updates its surveillance dashboards with information on livestock disease incidence[244]. For avian livestock, this dashboard includes veterinary diagnoses from non-commercial, hobby, and small-scale flocks of chickens. It is updated monthly with a frequency chart showing the main clinical signs corresponding to confirmed diagnoses during that time period. Data can be further filtered by species age and county location. The government has additional information about these dashboards, including the addresses of registered pig and poultry holdings.

While the validity of passive data collection through internet-based mediums is still uncertain, having this information as a supplementary tool can help strengthen existing surveillance tools. Research has shown that smallholders are increasingly obtaining their livestock information through different methods[6], so enhancing current dashboards to include insights from social media can help build a stronger intelligence repertoire.

Passive surveillance through internet data is a method that can be adopted to reduce the knowledge gap between commercial and small-scale flocks. The added benefit of Infoveillance techniques allows researchers to passively “listen in” and collect intelligence from these publicly visible exchanges. Such methods also allow for the distinction between community-based communication exchanges versus single opinion leader who has many followers within the forums and therefore the potential to control the narrative. In addition, this can be taken a step further through the classification of user type whereby the individuals who are posting the most/have the largest following can be examined to determine their characteristics (e.g. livestock size/range, those claiming to be vets).

6.4.1 Confirmed vs Extracted location differences

The differences between recorded pig and poultry holdings data and user locations from the forums indicate notable variations that deserve further discussion. The discrepancies in the dataset may be explained by three key factors: engagement variability, data collection methods, and regional demographic differences, each of which plays a role in the observed patterns.

- **Engagement Variability**: The variability in forum engagement differs significantly across regions. The maps in figures 40 and 41 indicate that areas with active forums don't always match with those exhibiting the most recorded holdings. The presence of this phenomenon may suggest regional variations in the interaction between communities and digital platforms. Engaging in forums may not directly correlate with farm density, but it reflects a proactive attitude towards seeking advice and participating in the community. The data suggests that digital engagement strategies should be adapted to promote participation in farming regions, regardless of the number of livestock. Furthermore, the ability of for discussions to be skewed by influential or the "loudest" users can also impact the levels of location frequencies, as discussed in the previous chapter.

- **Data Collection Methods**: The distinction in data collection methods between government records and forum-derived data is another important factor. Typically, APHA data is collected using structured and mandatory reporting methods, ensuring comprehensive coverage across regions. On the other hand, forum data relies on users voluntarily reporting (i.e. free text imputation), leading to self-selection biases.
- **Demographic and Internet Access Variability**: The data landscape is shaped significantly by the demographic characteristics of forum participants and regional variations in internet accessibility. Forums in regions with younger populations and advanced tech infrastructure may experience higher levels of activity. The skew might result in an overrepresentation of tech-savvy farmers, disregarding older or less technologically inclined individuals who may not engage in digital forums despite owning significant assets. This may be especially apparent within the Scottish Highlands, as indicated by the large differences in both pig and poultry holdings with forum location data in figures 40 and 41.

Further analyses can be conducted on other sub-forums, including those for equine, cattle, and sheep. Additional work can be done to examine the relationships between users themselves through social network analysis. Similar to research on peak detection and sentiment analysis, these methods can also be incorporated as an extension of this study to further strengthen these findings[220].

There are limitations to conducting web scraping and Infoveillance from internet data, including selection bias and small sample sizes. The potential for a few influential users to control the narrative of these forums is often overlooked, and network analysis is the only way to mitigate these effects. Forum data may only represent a portion of the community, as not everyone may participate online. Additionally, while topic modeling is a useful technique, it may not fully capture the essence of the discussions, so human expertise in the field of veterinary epidemiology is necessary to ensure the efficacy and quality of the results.

It is also important to note that animal health information discussed in forums relates to clinical signs observed by their owners. While this information is valuable, it lacks veterinary input. Furthermore, many conditions require laboratory testing for confirmation. In contrast, diagnoses published on APHA surveillance dashboards for various livestock animals have been generated following strict diagnostic criteria.

6.5 Conclusion

This section aimed to demonstrate the use of topic modeling algorithms on veterinary forum data as a starting point for further research on the effectiveness of combining data science techniques in the veterinary field. The results of the topic modeling and the high frequency of bigrams suggest that smallholders are primarily concerned with regulatory compliance and day-to-day management, particularly with regard to specific breeds, animal identification, containment, and feeding practices. The increase in discussions about free-range practices in poultry farming reflects a broader societal trend towards ethical and sustainable farming practices.

These findings have implications for policy making, extension services, and smallholder practices, and highlight the potential of forum data as a valuable resource for understanding and supporting communities.

This section focused on the implementation of one of many available methods in the field of topic modeling. LDA is one of the most commonly used applications in this field and has proven to be a successful tool when applied to public health data[371]. Contemporary models such as Top2Vec and BERTopic have shown to be effective with sparse, unstructured social media data, but further research is needed to compare their performance with other models. A full comparison of all these models applied to this data is beyond the scope of this study.

By applying techniques used in public health and epidemiological studies for human health surveillance, this study lays the groundwork for more detailed research on livestock animals, as similar studies on companion animal surveillance have proven effective[386].

7/. Spatio-temporal analysis

The final experimental chapter expands on the results from the published paper *"Spatio-temporal evaluation of social media as a tool for livestock disease surveillance"* and looks to explore the effectiveness of Twitter as a potential surveillance tool in detecting future outbreaks of Avian Influenza in the UK, through matching tweets with the spatio-temporal epidemiology of both domestic and wild bird outbreaks. By using temporal, geographical, and correlation analyses, the connection between avian influenza tweets and officially verified outbreaks in the United Kingdom in 2021 and 2022 were explored.

The methods used included:

1. Time series analysis to analyse the stationarity, temporal correlations, associations and causality.
2. Pearson correlation coefficient to test the significance between the datasets.
3. Bivariate Moran's I analysis for spatial analysis of tweets and confirmed outbreaks.
4. Google trends for associations between online searches and tweets.

The key findings were:

- Statistically significant but weak correlation between the number of tweets and confirmed outbreaks over time, suggesting that relying only on social media data for monitoring may not be enough.
- Spatial analysis shed light on the overlaps between confirmed outbreaks and tweet locations, providing insights into regionally targeted interventions during outbreaks.
- Social media can be helpful in understanding public sentiment and concerns during outbreaks, but it is not enough on its own. It must be combined with traditional surveillance methods and official data sources for a more accurate and comprehensive approach.
- Outbreak detection and response can be enhanced with advanced data mining techniques and real-time analysis.
- A robust surveillance system is crucial in monitoring and managing disease outbreaks to safeguard public health.

7.1 Introduction

During the summer of 2020, Russia and Kazakhstan detected an emerging outbreak of Avian influenza (AVIN) subtype H5N8[4]. The outbreak was attributed to wild and migratory birds that spread into Northern Europe due to their nomadic nature, particularly waterbirds. The virus was detected in poultry farms across the Netherlands, Germany, Belgium, and France by October 2020. This resulted in the mass culling of livestock[389]. In Late November/Early December 2020, as predicted by epidemiologists, the United Kingdom was affected by the spread and Highly pathogenic AVIN was found in Northallerton, North Yorkshire in rearing turkeys, which led to implementing biosecurity measures by the APHA [390]. Housing measures were enforced as part of protection zones around poultry premises on December 14th, 2020[391]. The enforcements persisted into the new year, with additional outbreaks being verified in numerous UK farms throughout the year, particularly in Autumn 2021. The complexity of constructing resilient surveillance methods for AVIN necessitates the use of dynamic information surveillance tools to enhance the current level of intelligence[21].

AVIN is a zoonotic virus that can be transmitted from wild birds (ducks, geese, swans) to domesticated animals (such as livestock chickens and turkeys) and is categorised in to two strains: Low pathogenic and high pathogenic[4]. LPAI infections in chickens can be asymptomatic, meaning the chickens show no visible signs of illness, namely swollen head or diarrhoea, whereas highly pathogenic AVIN (HPAI) may result in acute illness, and even death in a relatively short span of time. Although it is a zoonotic disease, the risk of transmission to humans is relatively low, with only one reported case in 2021. However, it remains a serious threat to the livestock industry, as infected birds often require mass culling[392]. The economic impact of the outbreak in the UK is estimated to be in the millions, with the commercial poultry industry enduring a large proportion of the losses[335].

Virus transmission is often attributed to feather moulting, dispersion of infected water droplets, and defecation[393]. Ducks and waterfowl are highly likely to carry the virus, highlighting the interconnectedness of transmission between them and wild birds[334]. Although biosecurity measures are implemented to protect the livestock and prevent contact, LPAI can still be transmitted via contaminated faecal material from waterfowl and can quickly mutate into HPAI[336], [337].

The difficulty with current surveillance systems is their inability to account for the mixing of wild and domestic animals, creating a pathway for disease transmission that can only be prevented through sound biosecurity measures [26]. The situation can be made worse by the type of farm, whether it's commercial, single-species, open air, or mixed[394]. For these biosecurity measures to be implemented, farmers and backyard keepers must be alerted of an imminent or highly likely threat through robust early warning signals. The migratory behaviour of wild birds makes it challenging to detect clinical signals in their populations[338]. Culling infected birds can cause long-term damage to farmers due to negative economic effects and loss of reputation. The 2004 outbreak of AVIN (H5N1) caused a financial loss of around £25 billion in a five-year period[339].

According to literature, a more adaptable and dynamic approach to surveillance is necessary to consider the complex relationship between domesticated and wild birds [17]. Non-traditional data forms are being extracted more frequently through social media as a surveillance tool. These studies have examined Twitter search terms in real-time and retrospectively to identify unusual patterns in disease-related search activity[21].

Early disease detection can be aided by modelling spikes in activity from a spatio-temporal context[340]. The literature suggests that this approach can be successful with bigger livestock animals like cattle and pigs, but it remains a challenging task with poultry. Government agencies usually confirm AVIN-related social media signals as outbreaks, which then spread through networks and are often retweeted among farming communities [7]. The occurrence of challenges can give rise to opportunities to analyse the frequencies, sentiments and topics of AVIN-related tweets during spikes, thereby aiding public and animal health planning.

The ability to communicate globally and access a wealth of current information has been made possible by developments in social media platforms[10]. Moreover, the applicability and usefulness of computational methods for text analysis and data mining of internet data have significantly increased across multiple domains. There has been a significant adoption of these methods in the human health field, with particular emphasis on using patient triage comments to anticipate illness/disease progression[49]. In the veterinary domain, its usage is limited due to the lack of dynamic or voluminous data. The efficacy of social media data for AVIN risk surveillance in North America was highlighted by Robertson and Yee in 2016. They used anomaly detection models in a time series manner to detect possible early disease detection cases. Through a spatio-temporal approach, they pinpointed locations with abnormally high Twitter activity regarding AVIN phrases and terms[21].

Researchers first started using Google's Flu Trends tool in the early 21st century to detect influenza in the American population[88]. Based on the occurrence of flu-like symptoms in particular areas, an estimation was made of the number of patients who visit their general practitioner. Although the statistical model overestimated the patient count by not accounting for symptom similarities with other diseases, it demonstrated the potential of employing comparable techniques for detecting outbreaks early. Within the veterinary domain, this would entail possessing a robust dictionary of clinical signs particular to each illness, which could be extracted from tweets and hashtags.

Google was one of the pioneering social media tools used over a decade ago and has been a valuable resource for health researchers tracking influenza using search term frequencies [13]. By monitoring spikes in virus-related search history over time, it's possible to analyse the disease's spread by location in detail. The technique was broadened to include other social media platforms and web scraping methods[108].

Hashtags play a crucial role in the public health field, as they enable the tracking of epidemics, as evidenced during the successful COVID-19 pandemic tracking from March 2020[15]. Researchers used Twitter scraping to monitor public opinions on vaccines, government lockdowns and mask mandates, as well as outbreak hotspots where cases were increasing. Social media surveillance tools proved beneficial during the COVID-19 pandemic and opened new opportunities in veterinary epidemiology[342].

To understand the connection between social media activity and confirmed AVIN cases, time series analysis plays a key role. However, when dealing with time series data, challenges may arise, especially the risk of spurious regression, which can result in misleading statistical relationships if not addressed correctly, as highlighted by Lamsal et al (2022)[395]. The occurrence of spurious regression arises when non-stationary time series data are regressed on each other, indicating significant relationships that are actually caused by underlying trends or seasonality instead of genuine causality. To minimise this risk and guarantee the strength of these results, stationarity checks were conducted before any analysis was performed.

7.2 Methodology

Three different data sets were obtained from various sources for the same time period:

- AVIN-related tweets
- APHA confirmed outbreaks in domestic birds
- APHA confirmed outbreaks in wild birds

The methodology segregated confirmed outbreaks of AVIN into two types: those detected in domestic and wild birds, with the aim of carrying out this distinction to better understand the transmission dynamics of the infection. The disease's tendency to switch between wild birds and domestic birds makes it crucial to comprehend the unique characteristics and patterns of each. Additionally, the differentiation enabled a thorough comparison and correlation analysis of the outbreaks in both populations. The selected methodology aimed to produce a detailed and extensive understanding of the spatial-temporal trends of the outbreak. The full timeline of confirmed outbreaks is displayed in Appendix B.

7.2.1 Data collection

7.2.1.1 Twitter dataset

Developer-level access was obtained, allowing the extraction of up to ten million tweets per month using the Tweepy module to activate the Twitter search API in Python. The extraction of data employed a retrospective approach, and the time filter was set to 1st January 2021 to 31st December 2022.

The parameters for the search terms included the following terms:

- *'bird flu OR avian influenza OR H5N1 OR H7N9 OR H5N8 OR H5Nx OR HPAI OR LPAI OR poultry flu'*

In order to have a balanced number of tweets throughout the year and due to API limitations, 24-time slots were assigned, corresponding to each month, and 500 tweets were gathered for each slot (the Twitter API maximum was between 10 and 500). Large and frequent requests can cause the code to time out or access to be revoked, hence a 15-second timeout between each request was used to prevent this. After retrieving 12000 tweets initially, applying the UK location filter reduced the number to 5843 unique tweets, with 2566 in 2021 and 3277 in 2022.

7.2.1.2 APHA confirmed outbreaks dataset

Extracting the official AVIN confirmed domestic bird outbreaks for 2021 and 2022 involved using the APHA Twitter page. Through a string match, each instance of the statement "Highly pathogenic avian influenza has been confirmed" was extracted.

In total, there were 300 outbreak occurrences, with 82 in 2021 and 218 in 2022. 295 outbreak events were analysed after removing 5 observations with unconvertible location data.

7.2.1.3 Wild bird outbreaks dataset

The wild bird outbreak dataset for the years 2021 and 2022 was downloaded directly from the UK government website [85]. The dataset's format comprised of week number, location, number of findings, species involved, and total number. The total number of outbreaks for 2021 was 159 and 754 for 2022, with a combined total of 913 confirmed cases.

7.2.1.4 Geolocation extraction

The Python GeoPy package was used to convert free text location data into geographic coordinates (latitude and longitude). The geocoding process was completed utilising OpenStreetMap's geocoding service. Through the conversion of the text-based location data into standardised geographic coordinates, spatial analysis and mapping of the data was made possible. By applying this to all three datasets, a more complete comprehension of the spatial distribution of tweets and their link to confirmed disease outbreaks was enabled.

7.2.2 Data analysis

The relationship between Twitter activity and outbreaks was investigated through a spatio-temporal analysis.

7.2.2.1 Temporal analysis

The related AVIN activity was depicted over a period of time using a time series approach that emphasised relevant periods of activity through spikes in frequency.

The daily aggregation of tweets was done to match the timeline of confirmed cases, highlighting the correlation between them. The temporal structure was determined using standard time series modelling techniques, including summary statistics, ARIMA, SARIMA, and Box-Jenkins.

The Pearson correlation coefficient (PCC) was used to calculate the strength and statistical significance of the correlation between tweet counts and confirmed outbreak counts.

Results from Google trends was also incorporated into the analysis to provide an additional data perspective around the online prevalence of AVIN-related search queries. The same search parameters as previously mentioned in section 7.2.1.1 were used for this analysis also.

7.2.2.1.1 Mitigating spurious regression

Spurious regression occurs when non-stationary time series data falsely display strong statistical relationships[396]. High R-squared values and significant t-statistics may not accurately represent a true causal relationship if the series contains underlying trends or seasonal patterns that are unaccounted for. Ensuring stationarity of all-time series data used in the analysis is crucial to avoid spurious regression. Stationarity checks were performed using the Augemented Dicky-Fuller (ADF).

7.2.2.2 Spatial analysis

The connection between tweets and outbreaks was investigated through spatial data analysis. The data was grouped by date and type, and then the number of instances for each group was counted. Kernel density estimation (KDE) plots were employed to visualise the density of observations on a map, thus providing insights into the distribution patterns of the events.

For each unique event type found in the daily count data, a KDE plot was created. These event types included Tweets and outbreaks Distinct KDE plots were generated for each event type by plotting longitude and latitude data against each other. Each plot was assigned a specific colour to differentiate between the different events. Tweets were assigned 'Blue', while confirmed cases were designated 'Red'.

Furthermore, an analysis of local bivariate Moran's I was performed to assess the spatial autocorrelation between the standardised tweet count and standardised case count of AVIN[342]. Unlike global Moran's I, a local version of Moran's I is interpreted as a z-score, without being limited to the -1 to 1 range.

7.2.3 Comparison between APHA confirmed outbreaks and reported wild bird outbreaks

The mean counts of outbreaks in livestock and wild bird populations were compared using a two-sample independent t-test. The test produced a T-statistic, which displayed the sample data's variation, and a p-value that indicates the likelihood of observing such a difference under the null hypothesis of identical average values.

The comparison of the two datasets was constrained by limitations, primarily due to differences in their temporal resolution. Daily data collection allowed for a detailed understanding of the disease's spread among domestic birds. In contrast, the dataset for wild birds was compiled weekly and marked as week 1 to week 52. The difference in data collection frequency between the two datasets was problematic, as it could have hidden the detailed timing of disease outbreaks in wild bird populations. In addition, there's a possibility of bias towards detecting outbreaks more immediately in domestic birds due to frequent data collection, which could introduce disparities.

The conversion of week numbers into a date format in the wild birds dataset was undertaken. The median day of the week was selected as the proxy date, i.e. the 4th day of a 7 day week was used transforming “week 1” in 2021 to “04-01-2021”. By employing comparable date structures, this process facilitated a standardised evaluation of the temporal features in the wild birds dataset against the other two datasets.

7.3 Results

7.3.1 Temporal analysis

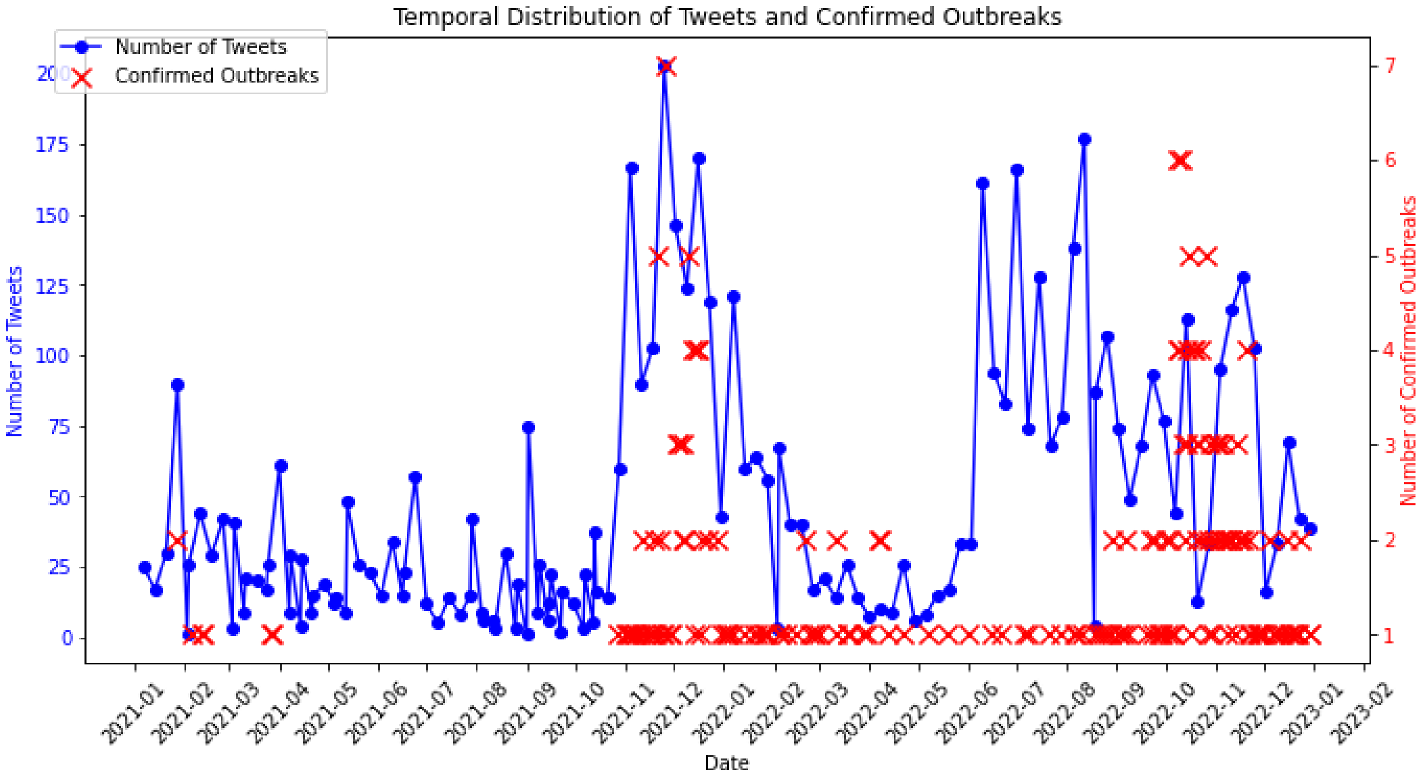


FIGURE 42: TEMPORAL ANALYSIS OF TWEETS AND CONFIRMED OUTBREAKS

The timing of AVIN tweets and confirmed cases reveals fascinating connections between social media discussions and actual occurrences. Overlaying both datasets can uncover temporal trends and correlated spikes between confirmed outbreaks and tweeting activity.

Figure 42 highlights that between October 2021 and January 2022, there was a notable increase in tweets related to AVIN. The number of tweets increased correspondingly with the sharp rise in outbreaks. Online discussions about AVIN mirrored public concern and awareness about the situation.

Throughout the timeline, the number of tweets fluctuated irregularly. There was a substantial decrease in the trend after January 2022, coinciding with a decrease in outbreaks. Social media discourse has a reactive nature to real-world occurrences, which could be the reason for the decreased public interest and concern as the situation improved.

There was a second and distinct rise in avian flu-related tweets from August to December 2022, which was also a noteworthy period. The confirmed cases of avian flu climbed simultaneously during this period. The volume of related discussions on social media platforms and the prevalence of a health crisis has a dynamic relationship that is pointed out by this recurrent pattern.

The analysis highlights how the temporal distribution of AVIN tweets closely matches the occurrence of confirmed cases in the UK. The synchronicity is notably high during two periods of increased activity - October 2021 to January 2022, and August to December 2022. Valuable insights for public health monitoring, response strategies, and public engagement efforts can be gained from the close alignment of social media discourse and actual occurrences.

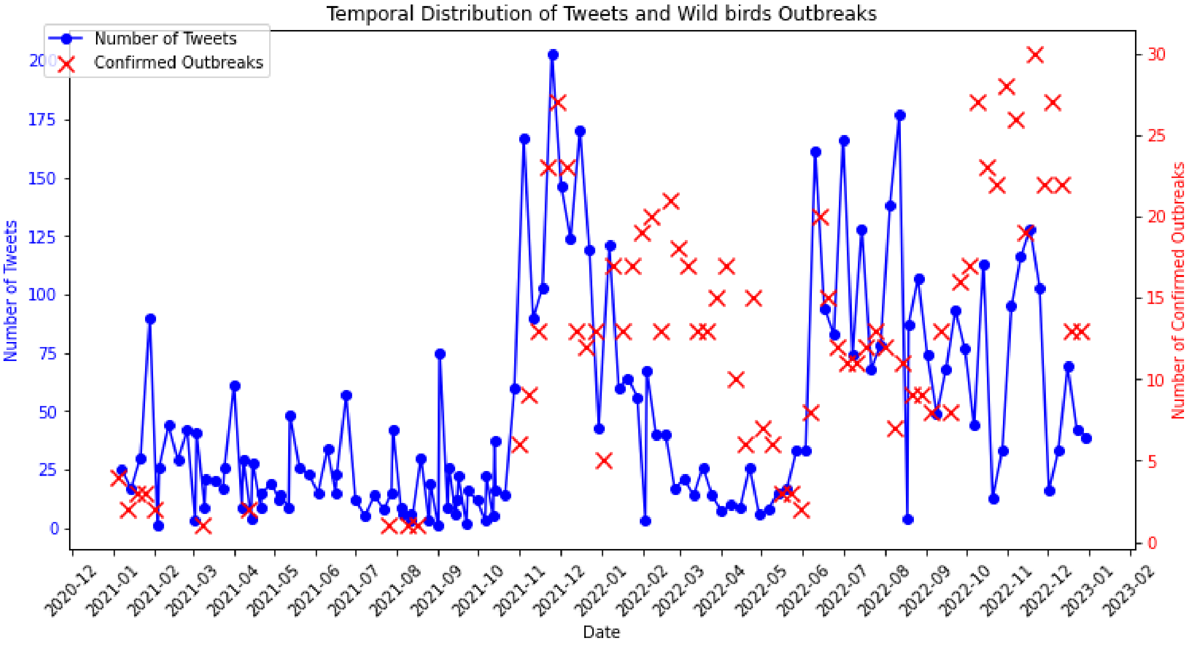


FIGURE 43: TEMPORAL DISTRIBUTION OF TWEETS AND WILD BIRD OUTBREAKS

Similarly, for the wild birds dataset, increases in outbreaks visually correspond to the spikes in tweet frequency, as displayed in figure 43. The number of tweets increased significantly from October to December 2021. The surge in social media activity has coincided with a significant increase in outbreaks of AVIN among wild birds. It can be inferred that there is a connection between the escalating situation in the wild bird population and the surge of online discussions.

The volume of AVIN-related tweets has decreased significantly since December 2021, in line with the reduction in confirmed wild bird outbreaks. In this context, Twitter discussions and disease incidence among wild birds share a strong correlation.

A second phase of substantial activity is observed during the period spanning June 2022 to December 2022. Similar to the initial occurrence, the surge in AVIN related tweets is closely linked to an increase in outbreaks among wild birds.

The recurring pattern underscores how social media discourse responds to changes in real-world situations, such as the spread of disease in wild birds.

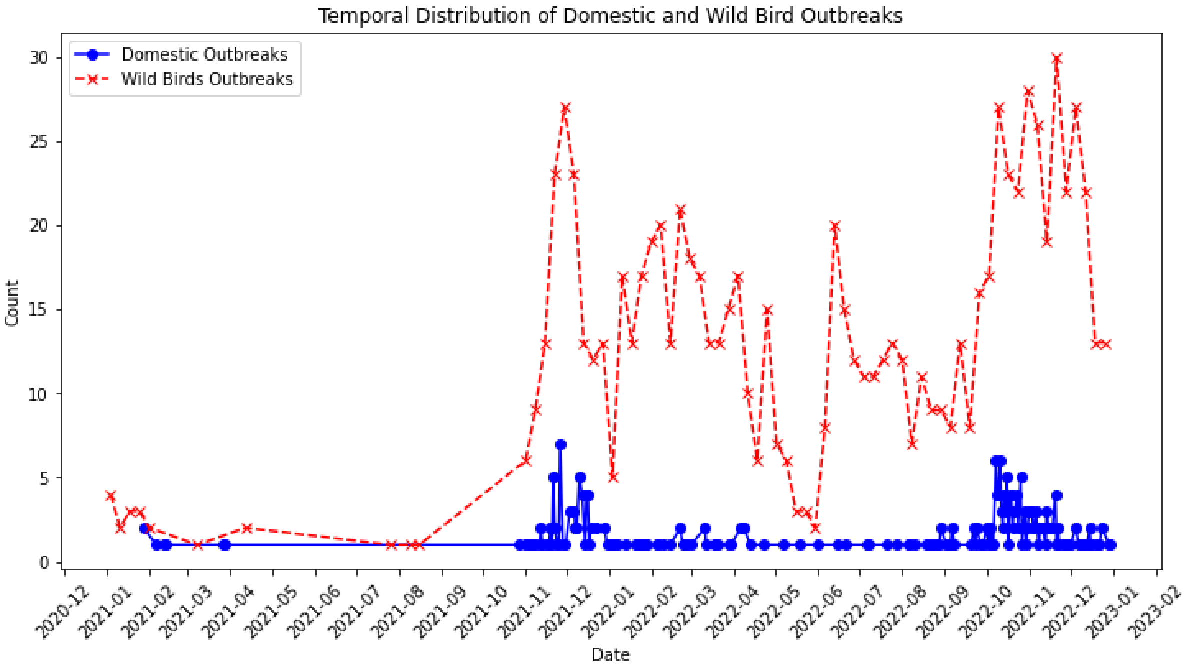


FIGURE 44: SUPERIMPOSITION OF DOMESTIC AND WILD BIRD OUTBREAKS

Figure 44 shows an overlap of the frequency counts of both domestic and wild bird cases depicts similar trends in outbreaks, with November - December 2021, February – April 2022 and September 2022 – December 2022 being the 3 main time epochs which show related spikes between the two datasets. The count of wild bird cases is considerably greater throughout the two years.

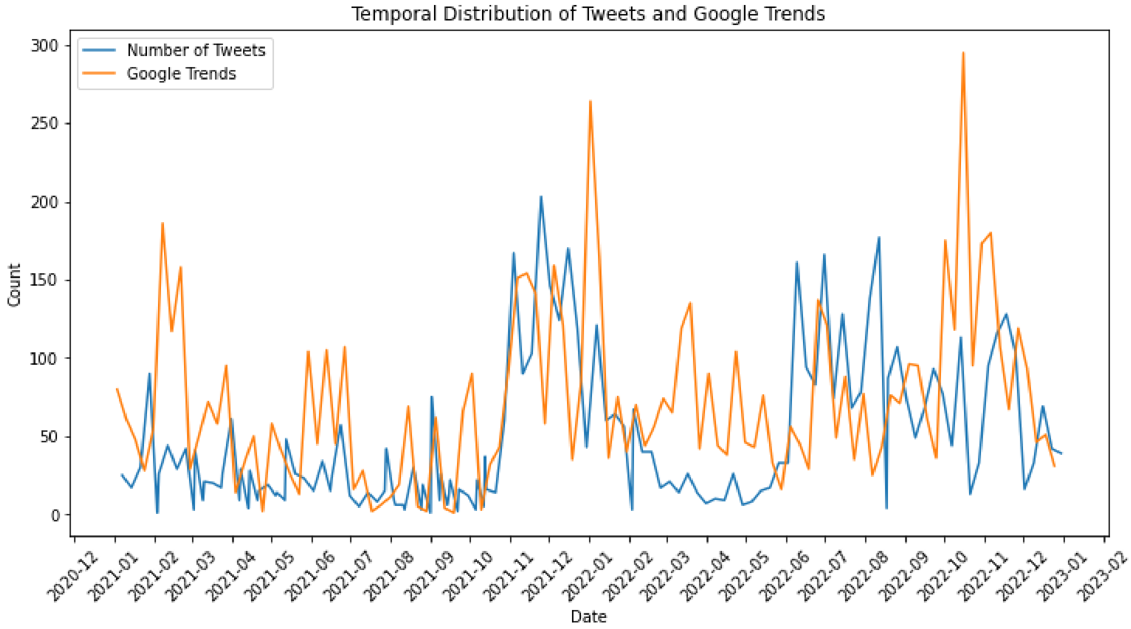


FIGURE 45: GOOGLE TRENDS - TWEETS AND SEARCH RESULTS

Bringing in Google trends adds an additional layer of insight into the results, demonstrated by the plot in figure 45. Whilst it provides a predominantly surface level supplement to the analysis, witnessing the parallels between online search activity and tweets offers not only an additional dataset, but also the ability to understand the interconnectedness between online search activity and tweets. From the plot above, this interconnectedness is noticeable as both lines follow similar trends over time. One visible aspect is the slightly early spike in tweeting activity, which is then followed immediately by the increase in Google searches, perhaps indicating that users engage on Twitter first, and then follow this with online searches that are related to the same discussions.

This analysis may indicate that the topic of the tweets or the level of user engagement increased in 2022. From these results, we can infer that the frequency of discussions increased considerably possibly due to numerous factors. Firstly, there was a larger concentration of outbreaks in late 2021, which could have caused a latent effect. Furthermore, the distribution of outbreaks in 2022 was more equally distributed across the year, which led to increased media coverage and public awareness of the disease, whereas 2021 displayed more of a sparse distribution. Moreover, there were new scientific findings about the risks of HPAI, including the possibility of the virus mutating and becoming more transmissible to humans. Coupled with the continued economic impacts of COVID-19, the outbreak also proved to have a significant economic impact, driving up the cost of poultry and causing supply chain disruptions[397].

7.3.2 Time series analysis

7.3.2.1 Checks for Stationarity

ADF Test Results for AVIN Tweets per Day	Value
Test Statistic	-2.8837
p-value	0.0473
Used Lags	20
Number of Observations	702
Critical Values	
1%	-3.4397
5%	-2.8657
10%	-2.5690
Stationary	True

TABLE 31: ADF TEST RESULTS FOR AVIN TWEETS PER DAY

Tabel 31 displays the results from the stationarity tests. The ADF test statistic for the daily tweet counts related to AVIN is -2.8837. In continuation, with a corresponding p-value of 0.0473, which is below the commonly employed threshold of 0.05, the results are statistically significant. This suggests that we can reject the null hypothesis of a unit root at a significance level of 5%, providing evidence that the series is stationary.

To address autocorrelation, the test incorporated 20 lags of the dependent variable, and there was a total of 702 observations. The test statistic's critical values at various significance levels (1%, 5%, and 10%) are -3.4397, -2.8657, and -2.5690, respectively. Further evidence that the series is stationary is provided by the test statistic falling between the critical values for the 1% and 5% significance levels.

Based on these findings, it can be stated that the AVIN tweet data is in a stationary state, ensuring that the time series analyses and models are built on a strong foundation. This will decrease the chance of false correlations and enhance the dependability of our results.

7.3.2.2 ARIMA

The findings of the time series analysis are explored through Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. The plots in figure 46 are indicative of statistical relationships, trends, and cyclic patterns that might exist within the tweet data.

To the ACF and PACF plots, horizontal lines indicating the significance level and shaded blue areas representing 95% confidence intervals are added. Statistical significance is attributed to correlation values that lie beyond these boundaries, and they are not likely to be due to chance. The correlations of note imply a link between the observation and its earlier observations, referred to as lagged observations.

Upon examining the ACF plot specifically, significant autocorrelations are seen at the first six lags. The observation suggests a possible positive trend in the data, indicating that higher tweet volumes on one day could lead to higher tweet volumes on subsequent days.

Fewer significant partial autocorrelations are revealed by the PACF plot. The initial four lags are categorised as such. Even when accounting for other days, there appears to be a connection between tweet counts on a particular day and those on the days before it.

A peak at lag 0 is expected in the ACF, as an observation is perfectly correlated with itself. In addition, the second most prominent peak was detected at lag 3. The volume of tweets on a particular day has the closest correlation with the volume of tweets from three days earlier. This could potentially suggest a recurring pattern or cycle in the data, where a surge or decline in a tweets appears to repeat itself every three days.

Upon examining the PACF, once again the peak is at lag 0, which is typical for the same explanation provided earlier. At lag 1, the second highest peak was detected, which is noteworthy. The number of tweets on a specific day is most closely correlated with the number of tweets from the day before, when controlling for other days' impact. A potential short-term dependency in the tweet data could mean that the tweet volume on the current day is greatly influenced by the events or discussions from the day before.

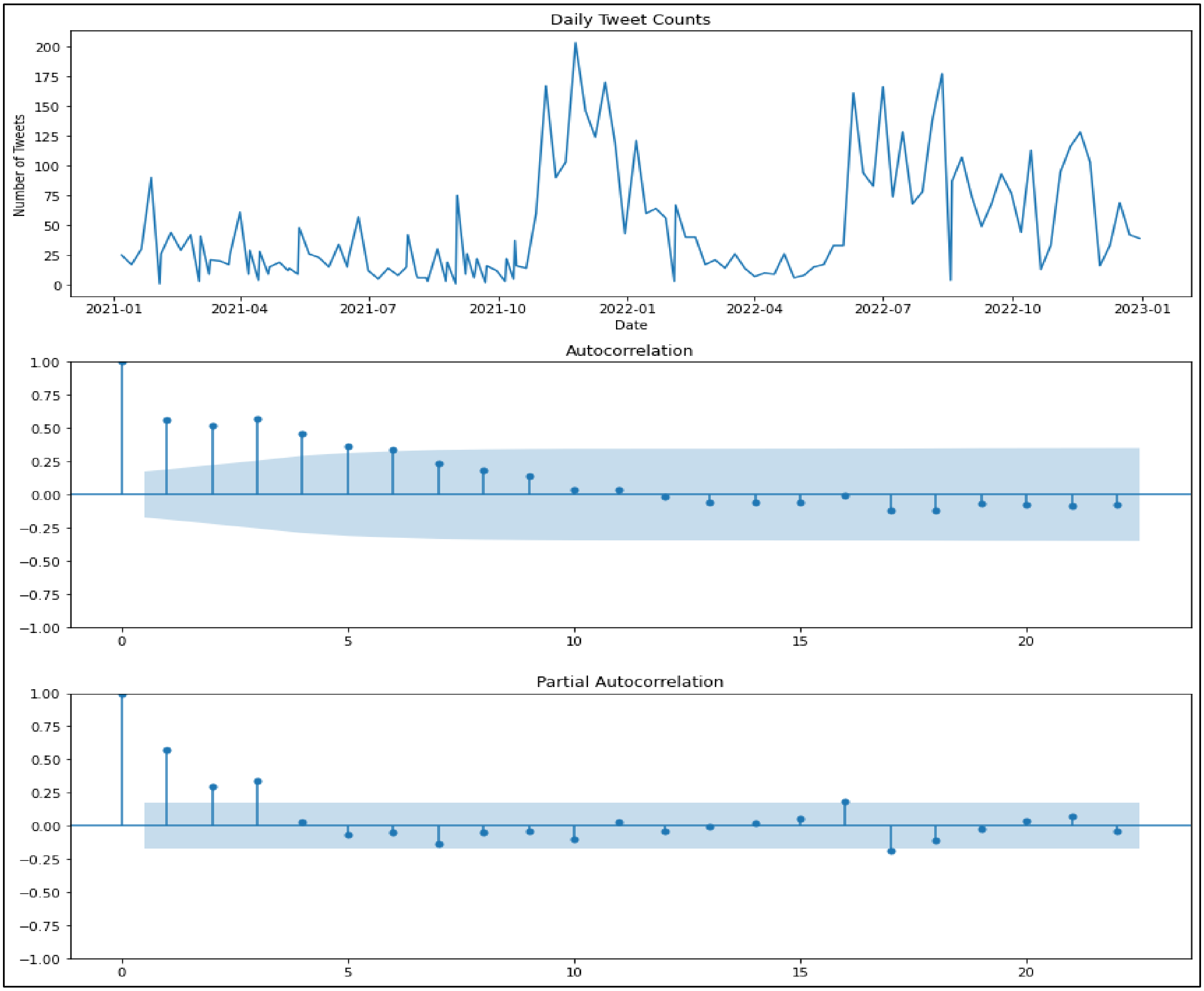


FIGURE 46: AUTOCORRELATION AND PARTIAL AUTOCORRELATION PLOTS

Both an immediate (one-day lag) and a slightly longer-term (three-day lag) relationship were found in the tweet volumes, according to the results. These patterns could aid in modelling and predicting future tweet volumes. Additional analysis is required to grasp the potential causes and consequences of these cycles that last for three and one day.

Measure	Value
Model	ARIMA(1, 1, 1)
Log Likelihood	-634.307
AIC	1274.615

BIC	1283.147
HQIC	1278.081
Covariance type	OPG

TABLE 32: ARIMA MODEL MEASURES

The log likelihood value of -634.307 generated by the ARIMA model measures the probability of observing the dataset with the model's parameters. While the figure is negative, it's important to note that smaller negative values indicate a better model fit and higher likelihood. The measures and corresponding values are displayed in table 32.

The model has a reliable forecasting ability for avian flu related tweet volumes, due to its balance between simplicity and fit, which is indicated by its relatively low AIC, BIC, and HQIC values.

Parameter	Coefficient	Std Error	Z – score	P >|z|	0.025	0.975
AR.L1	-0.1039	0.144	-0.723	0.470	-0.386	0.178
MA.L1	-0.5385	0.115	-4.688	0.000	-0.764	-0.313
SIGMA2	1270.9371	92.438	13.749	0.000	1089.761	1452.113

TABLE 33: ARIMA CO-EFFICIENTS

The results in table 33 reveals information about the coefficients estimated by the ARIMA model, their standard errors, Z-scores, p-values, and 95% confidence intervals.

- Negligible correlation between consecutive tweet counts is indicated by a coefficient of -0.1039 for the first-order autoregressive parameter, AR(1). The correlation is not significant statistically since the p-value of 0.470.
- On the other hand, the coefficient of the first-order moving average parameter (MA(1)) is -0.5385 with a significant p-value of 0.000, as it rectifies the errors from previous forecasts. According to this, the model corrects around 54% of the error from prior estimations.
- SIGMA2 represents the variance of the noise term, estimated at 1270.9371, indicating unpredictability in the data. Given the p-value of 0.000, this estimate is statistically significant.

Hence, the model suggests that the number of tweets on a particular day doesn't rely significantly on the tweet count of the previous day. The model has an active error correction mechanism that corrects over half of the previous prediction's errors, suggesting a high degree of randomness in the number of tweets.

Test	Value
Ljung-Box (L1) (Q)	0.03
Prob(Q)	0.86
Jarque-Bera (JB)	88.27
Prob(JB)	0.00
Heteroskedasticity (H)	4.86
Prob(H) (two-sided)	0.00
Skew	0.54
Kurtosis	6.94

TABLE 34: RESIDUAL ANALYSIS

According to table 34, the Ljung-Box test for autocorrelation yielded a result of 0.03 in the residual analysis of the ARIMA model. The lack of significant autocorrelation in the residuals shows that the model has successfully accounted for the time-based connections in the tweet sequence.

On the other hand, normality is contradicted by the Jarque-Bera test, which yielded a value of 88.27 and a p-value of 0.00. It is suggested that the residuals do not adhere to normality in their distribution.

Furthermore, evidence of heteroscedasticity is found as the test reveals a value of 4.86 and a p-value of 0.00, indicating a non-constant variance in the residuals. Moreover, a positive skewness of the residuals (0.54) is observed, which suggests a right-skewed distribution. The kurtosis measure of 6.94 indicates a heavy-tailed or leptokurtic distribution, as it is higher than the normal distribution's value of 3. Despite the lack of significant autocorrelation, the residuals show signs of heteroscedasticity and non-normality.

7.3.3 Correlation analysis

7.3.3.1 Tweets and domestic bird cases

The daily tweet count and the number of domestic bird cases have a correlation coefficient of -0.202, as revealed by the correlation analysis. The p-value of 0.00087 suggests a statistically significant but weak negative correlation. The weak negative correlation suggests a minor inverse connection where there's a slight decrease in domestic bird cases with an increase in tweet count, and vice versa. Despite the correlation, there is inconsistency in the relationship due to its weakness.

7.3.3.2 Tweets and wild bird cases

There was a moderate negative correlation (-0.413) observed between the daily tweet count and the confirmed wild bird case count in. From this moderate correlation, it can be implied that an increase in wild bird cases is associated with a decrease in tweets. The correlation's statistical significance was demonstrated by a p-value of 0.00.

7.3.3.3 Domestic birds and wild bird cases

Finally, a weak negative correlation between the daily domestic bird cases and wild bird cases was shown in the analysis, with a correlation coefficient of -0.272. A weak association exists between an increase in domestic bird cases and a decrease in wild bird cases, and the reverse is also true. It's likely that other factors are affecting these variables, given the weak correlation. Similarly to the previous results, the p-value was also deemed to be statistically significant.

Whilst these results offer an insight into the associations between the datasets, it is important to note that these correlations do not necessarily imply causation for any of the results.

7.3.4 Spatial analysis

Spatial analysis was conducted to examine the distribution and density of tweets across different regions, areas with higher levels of public concern and awareness about the disease were identified. Additionally, by comparing these patterns with the locations of confirmed cases, the association between public discourse and the actual spread of the disease was evaluated. This section will present the spatial analysis results and discuss their implications for understanding the UK's AVIN dynamics.

7.3.4.1 Location frequency tables

Location	Tweet frequency
United Kingdom	615
London	405
England	238
Scotland	219
Wales	91
Northern Ireland	54
North West, England	42
Edinburgh	42
Somerset	40
South West, England	40

TABLE 35: TOP 10 LOCATIONS – TWEETS

Table 35 displays the top 10 locations as stated in the user profiles, with the highest frequency of tweets, originating from the broad category of "United Kingdom", totalling 615 tweets.

Focusing on specific regions, London dominates with the highest tweet frequency, with 405 tweets. The high population density of the capital city may explain the higher volume of tweets and possibly the presence of more Twitter users.

With 238 tweets, "England" ranks third in terms of tweet frequency as a generalised location. Scotland is closely following with 219 tweets, highlighting how the disease affects various countries in the UK. Furthermore, Wales and Northern Ireland register lower frequencies of tweets, with 91 and 54 tweets respectively, which could be a result of their smaller population sizes in comparison to England and Scotland.

Notable tweet frequencies can be found in the North West and South West regions of England, both surpassing 40 tweets. Likewise, this count is met by Edinburgh, the capital of Scotland.

Location	Confirmed case frequency
Attleborough	13
Thirsk	11
Dereham	9
Alford	9
Mundford	6
Wymondham	4
Much Hoole	4
Redgrave	4
Heybridge	4
Taverham	4

TABLE 36: TOP 10 LOCATIONS - CONFIRMED OUTBREAKS

Similarly, In the analysis of confirmed domestic bird cases, several localities emerge as areas with the highest frequencies. With 13 confirmed cases, Attleborough was the top location for cases among domestic birds during the reviewed period.

Moreover, the number of confirmed cases in Thirsk was 11, while Dereham and Alford have nine each. These localities, including Attleborough, have a high percentage of confirmed cases, which suggests a clustered spread of the disease in these areas. Finally, Mundford reported a moderate frequency of six cases, and four confirmed cases have been reported in Wymondham, Much Hoole, Redgrave, Heybridge, and Taverham each.

Location	Wild bird case frequency
Aberdeenshire	28
Highland	24
Fife	23
Moray	22
Dumfries and Galloway	17
Cheshire East	16
Dorset	15
Scottish borders	15
West Lancashire borough	13
East Riding of Yorkshire	13

TABLE 37: TOP 10 LOCATIONS - WILD BIRD OUTBREAKS

The presence of migratory bird routes, bird population densities, and virus-friendly habitats in certain areas may increase the likelihood of outbreaks, as indicated by the geographical distribution of wild bird cases highlighted in table 37.

Scotland was shown to have the highest concentration of wild bird cases, with Aberdeenshire, the Highlands, Fife, Moray and Dumfries & Galloway comprising of the top 5 locations respectively. The Scottish Borders also registered 15 cases.

Northern England was the second highest region, with Cheshire East, West Lancashire and East Riding of Yorkshire confirming a total of 42 between them.

7.3.4.2 Spatial distribution

Figures 47 and 48 display the KDE plots that visualise the spatial distribution of tweets and confirmed. For the tweets and confirmed cases, there appears to be a significant concentration and overlap in London, with more scarce distributions towards the North of England and crossing the border into Scotland. The West of Scotland and Northern Ireland show no overlap between the co-ordinates.

A significant overlap is observed, as seen by the greater coverage between the two lines, in comparison to the rest. There is no overlap in tweeting activity for Moray in Scotland, as well as Northern Ireland as previously shown in the results. A significant overlap in areas indicates that people are tweeting about cases in their locality. Conversely, in regions with minimal overlap, tweets may be more geographically scattered, or confirmed cases may be concentrated in areas with less social media involvement.

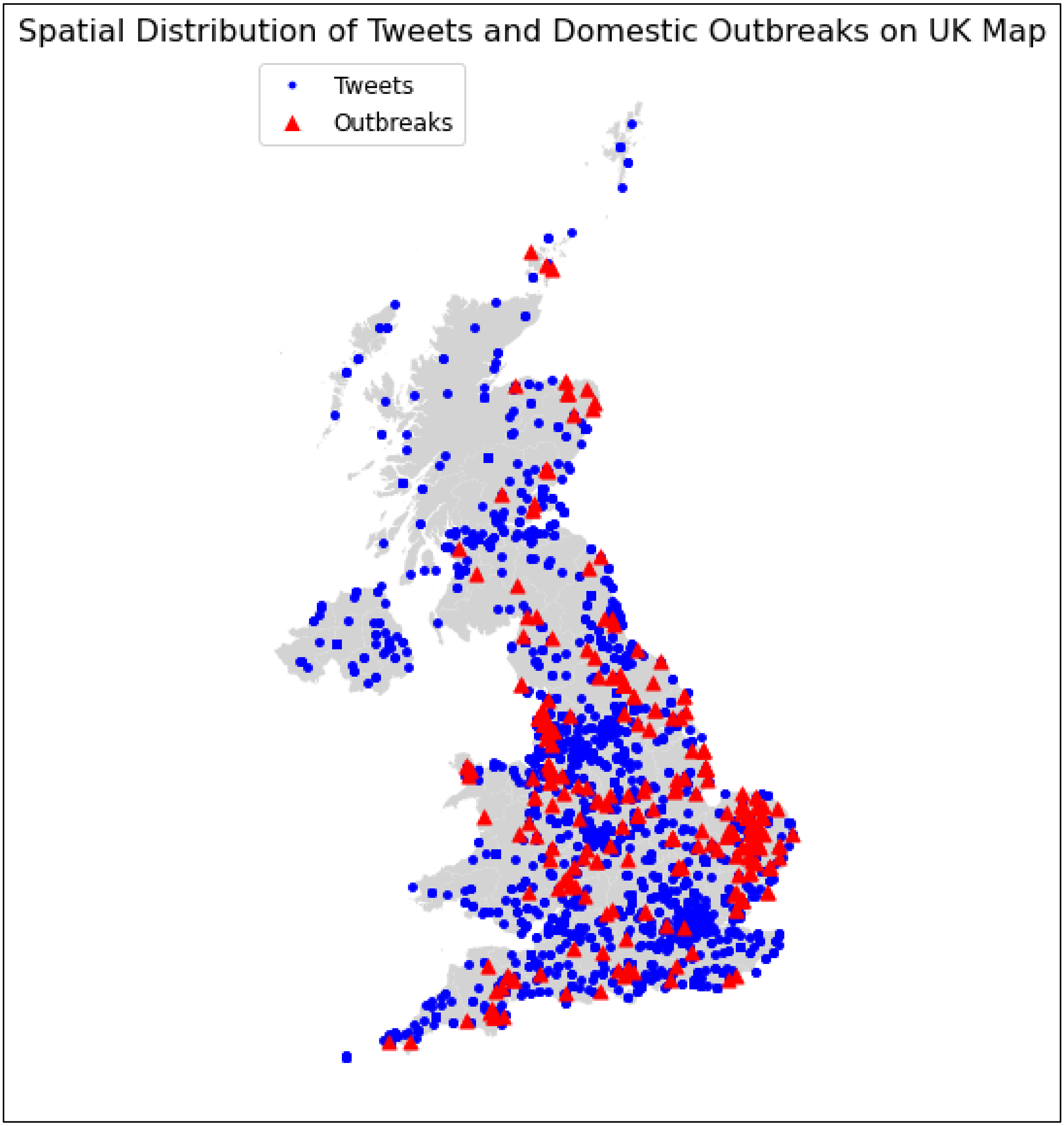

Spatial Distribution of Tweets and Domestic Outbreaks on UK Map
● Tweets
▲ Outbreaks

FIGURE 47: KDE PLOT FOR TWEETS AND CONFIRMED OUTBREAKS

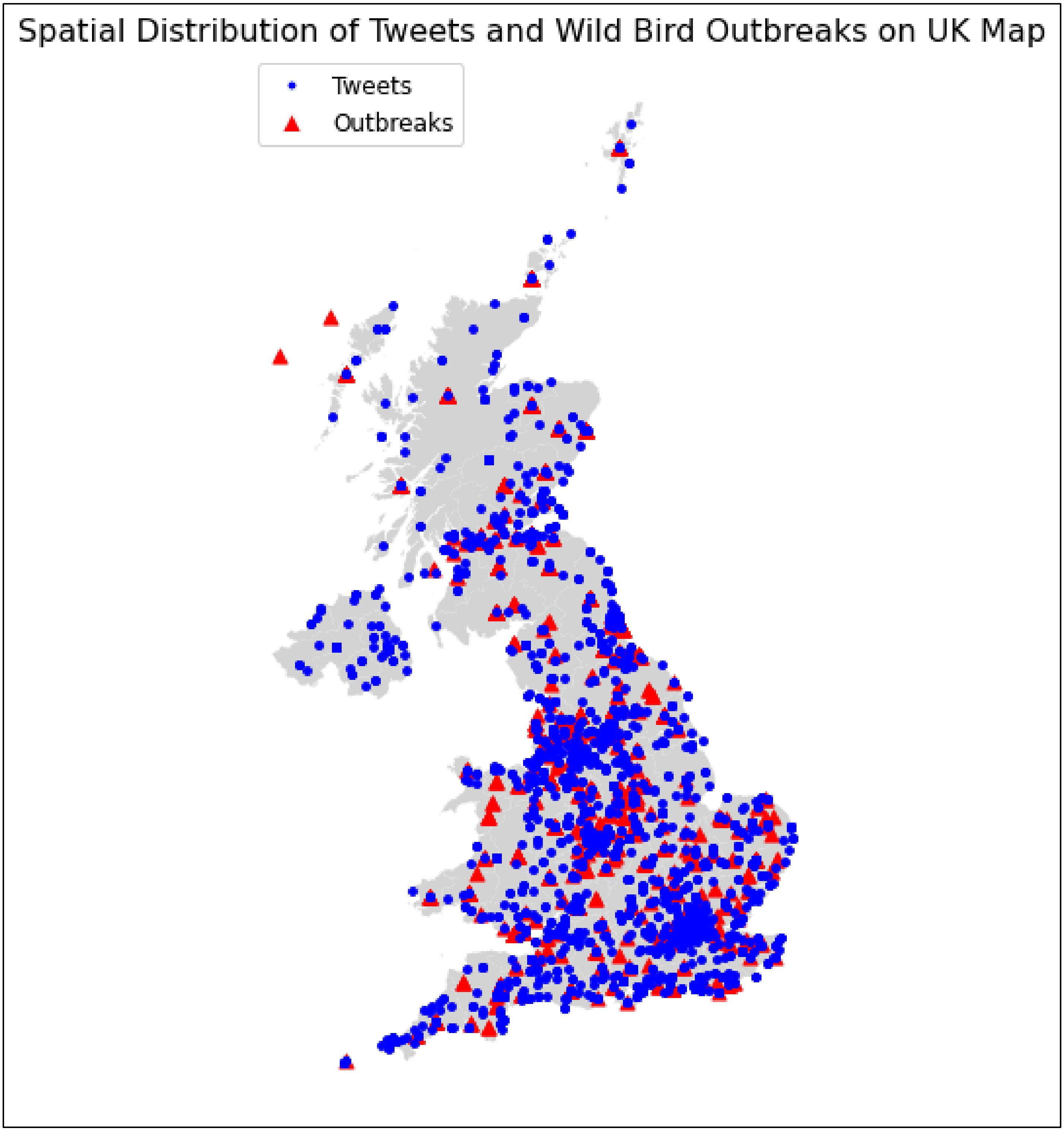

Spatial Distribution of Tweets and Wild Bird Outbreaks on UK Map
● Tweets
▲ Outbreaks

FIGURE 48: KDE PLOT FOR TWEETS AND WILD BIRD OUTBREAKS

7.3.4.3 Spatial clustering

Figures 49 and 50 depict the cluster maps for tweets mapped against confirmed cases and wild bird cases respectively. Figure 49 reveals that the bivariate Moran's I statistic results for tweet count and case count ranged from -1.5 to 0.08, indicating varying levels of spatial autocorrelation. In Northeast England, the presence of negative values (e.g. -1.5) indicates that regions with high tweet volumes are usually linked to low incidence of cases and vice versa. High tweet and case counts show a correlation in regions like the Midlands, as evidenced by the prevalence of positive values of 0.08. The absence of any clear spatial correlation between the two variables is indicated by values near zero. Although broad, these findings suggest a potential difference or agreement between public opinion and real outbreak locations.

Similarly, figure 50 provides a broader spatial distribution of data by correlating tweets with wild bird cases and produces more statistics than figure 49. Furthermore, there was a wide variation in values, ranging from -1.7 in Northwest England to over 10 in certain areas of Scotland. Areas with a high tweet count usually have a low wild bird case count and vice versa, indicating an inverse relationship. Tweet count might not match actual cases among wild birds, implying areas of misinformation or lack of awareness.

On the other hand, strong positive spatial correlation is indicated by high positive values, such as 10.667. Regions with high tweet counts indicate high wild bird case counts. Greater public awareness and discussion might be due to such an observation in regions with larger disease outbreaks.

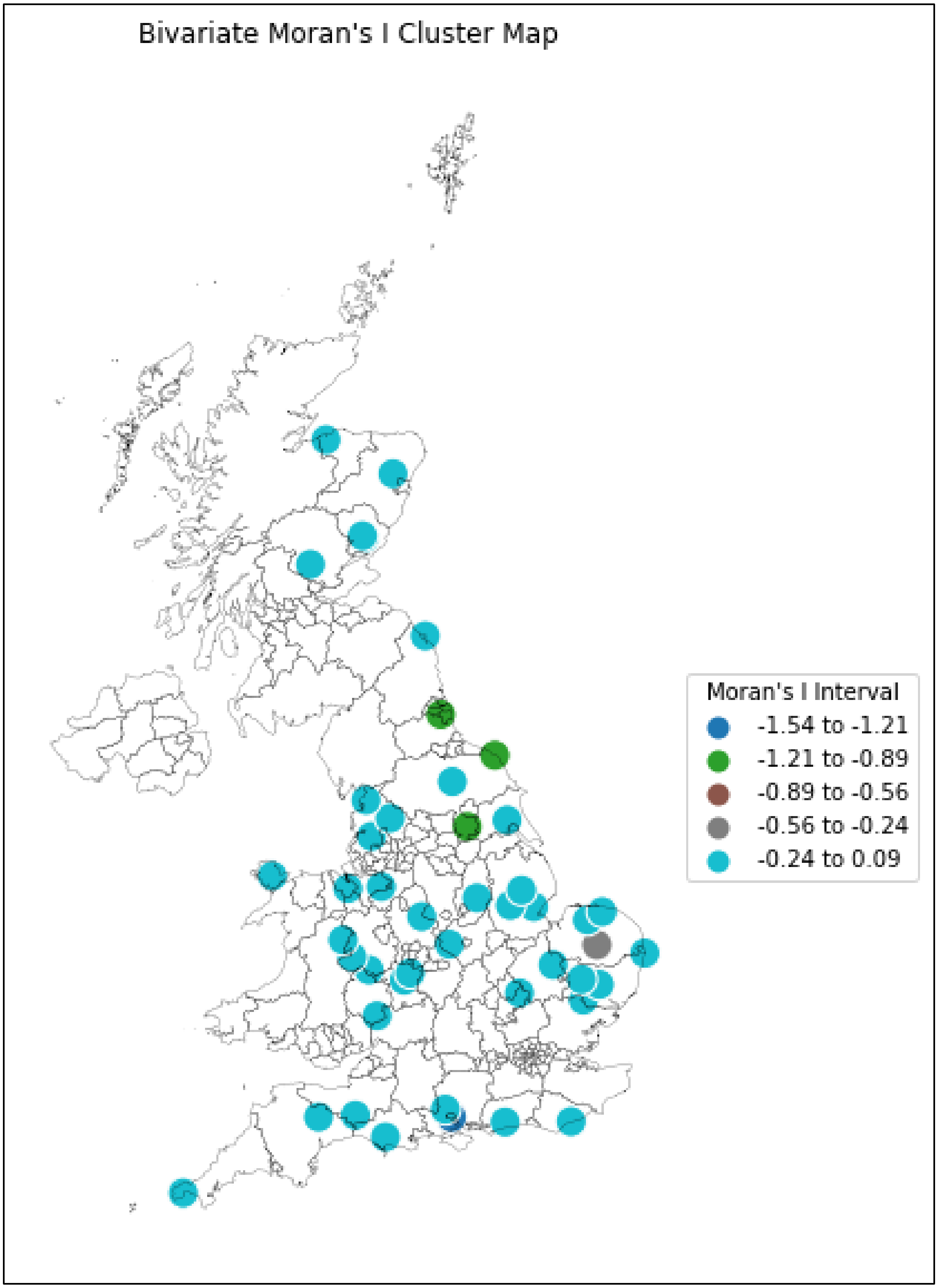


FIGURE 49: BIVARIATE MORAN'S I CLUSTER MAP - TWEETS AND CONFIRMED DOMESTIC OUTBREAKS

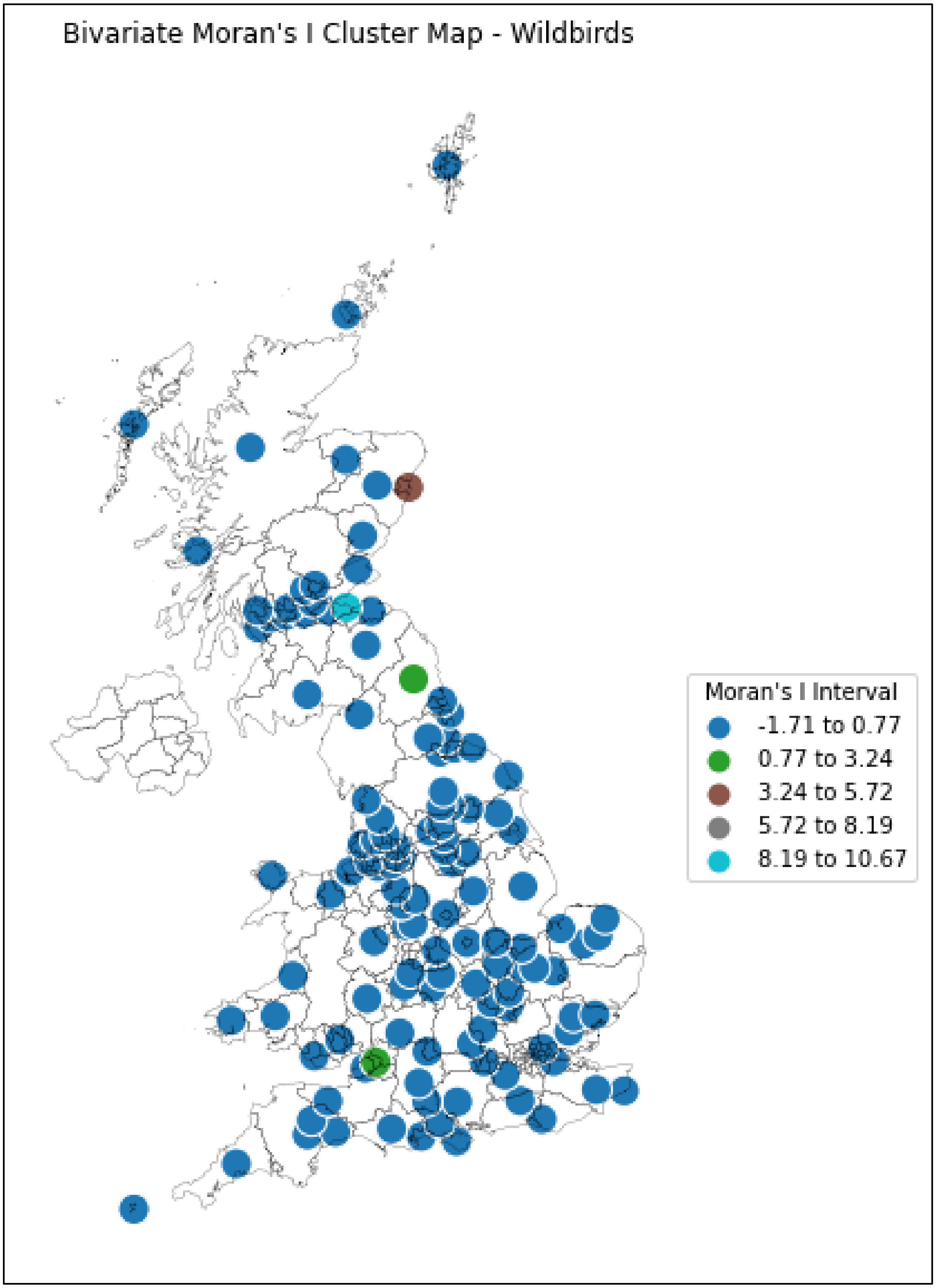


FIGURE 50: BIVARIATE MORAN'S CLUSTER MAP – TWEETS AND CONFIRMED WILD BIRD OUTBREAKS

7.3.5 Comparison between domestic and wild birds

A statistical T-test was performed to test the hypothesis that there is a statistically significant difference between the number of cases in domestic bird populations compared to wild bird populations.

The T-statistic value of -7.48 suggests a significant difference between the mean of the case count for the domestic bird population and the wild bird population. The mean for domestic bird population is below that of wild bird population as suggested by the negative value.

The P-value is less than 0.05 hence the result is statistically significant. The null hypothesis, which suggested that there was no difference in the mean count of avian flu cases between domestic and wild bird populations, can be rejected based on this result. There are various factors that could be indicated by this, such as differing susceptibility to the disease, conditions that affect disease spread, or other ecological factors.

The DBSCAN algorithm's output shows a range of -1 to 4 for the spatial clustering of domestic bird cases, as seen in figure 51. Clusters 3 and 4 are widespread in the London region, with most other areas across the nation being considered noise points, labelled as -1.

Spatial clustering indicates a higher density of domestic bird cases in London, which justifies clusters 3 and 4. The high population density and possible contact with nearby bird populations could be a reason for increased cases in this area. However, the prevalence of '-1' labels in the remaining regions denotes a low density of cases or infrequent events that could not be clustered into distinct groups. These areas may denote rural or sparsely populated regions where domestic bird cases are less frequent.

In contrast, the results in figure 52 reveal a broader range of spatial clustering for wild bird cases, spanning from -1 to 49. The regions of London, Southwest England, and North West England are identified as yielding the highest category (45-49), suggesting they have a high concentration of cases.

Scotland and the North of England exhibit a more diverse cluster distribution, which includes labels ranging from 29 to 36, as well as '-1'. The variation could indicate regions with both high density of cases and isolated, sporadic cases.

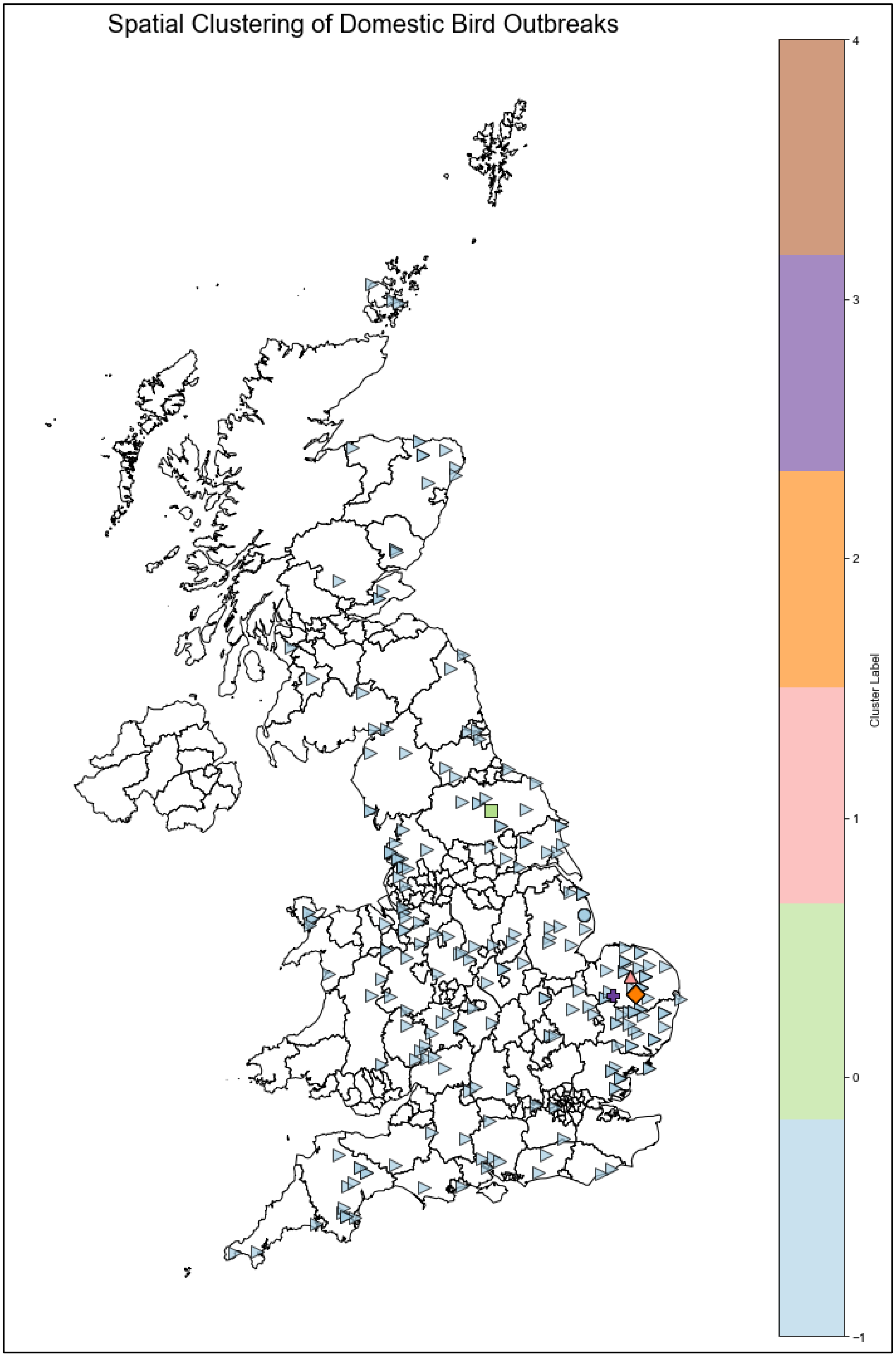


FIGURE 51: SPATIAL CLUSTERING - DOMESTIC BIRD OUTBREAKS

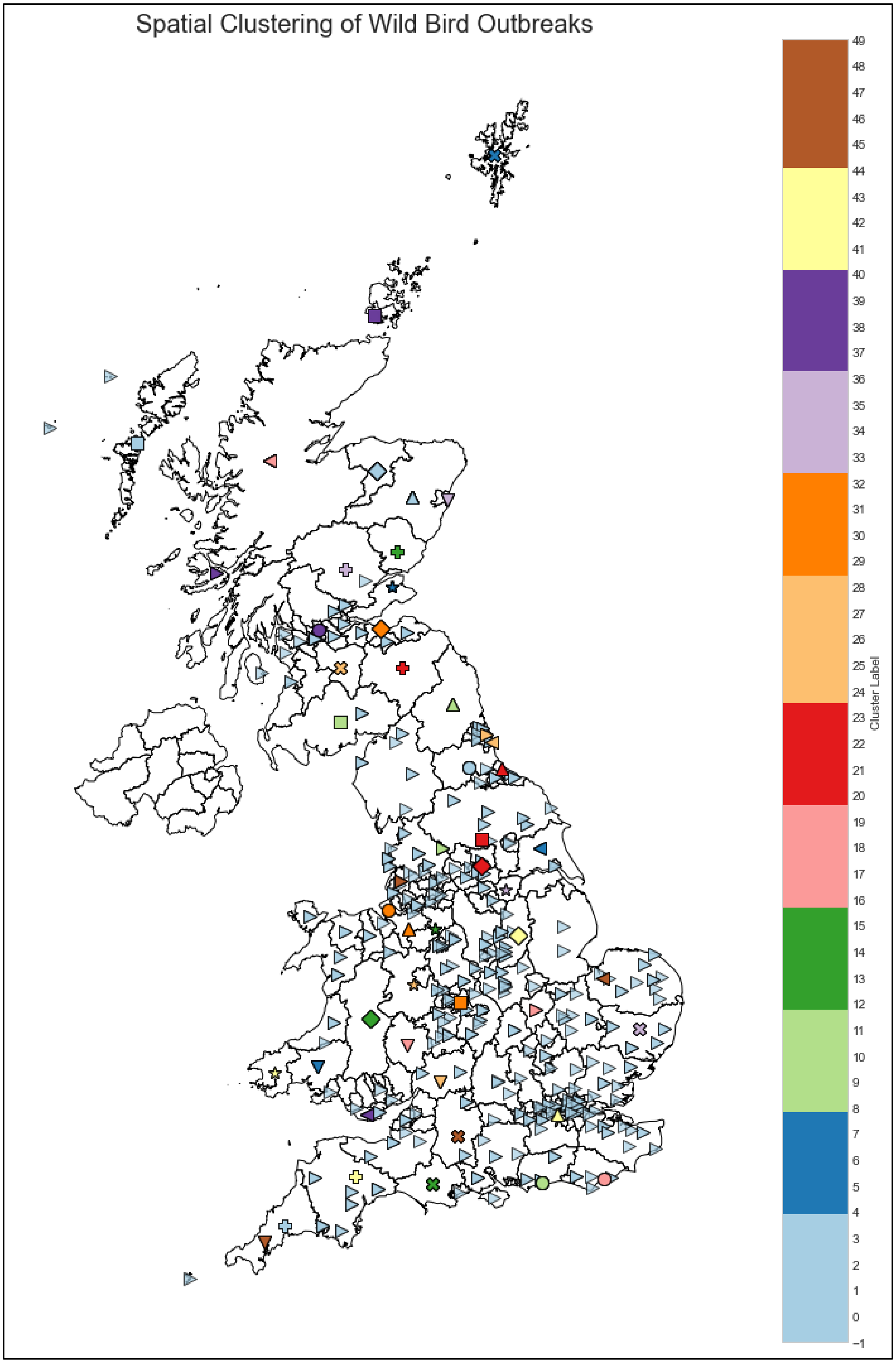


FIGURE 52: SPATIAL CLUSTERING – WILD BIRD OUTBREAKS

7.4 Discussion

By applying spatio-temporal analysis, this chapter was able to establish a link between AVIN related tweets and confirmed cases in the UK in 2021 and 2022. The purpose was to test if social media, specifically Twitter, could be used as a surveillance tool for disease outbreaks by conducting temporal, spatial, and correlation analysis.

The temporal examination revealed that certain periods of heightened tweeting activity coincided with confirmed AVIN cases. The Pearson correlation coefficient showed that the correlation between the quantity of tweets and the number of confirmed cases was weak. The implication is that relying on tweet counts may not be a trustworthy approach to tracking outbreaks. The public's response to information saturation might have caused the negative correlation between cases and the number of related tweets. With the rise of cases, the media and public discourse have increased, causing a sudden increase in tweets related to the topic. Despite the increasing case count, the persistent prevalence of the disease might cause the public to be less engaged with information related to AVIN, resulting in a decrease in the number of tweets. The negative correlation observed could be attributed to "compassion fatigue" or "crisis fatigue," which is a common phenomenon[398].

The spatial analysis revealed the distribution of tweets and confirmed cases across regions. The analysis showed significant overlap in the spatial distribution of tweets and confirmed cases, especially in London. Overlapping areas indicate that individuals may be tweeting about the disease in their community. This may indicate an increase in public awareness in these regions. However, most regions exhibited a weak negative spatial correlation based on the bivariate Moran's I analysis. In addition, spatial clustering identified particular high-density areas for avian flu cases among both domestic and wild bird populations.

A significant statistical difference in the number of cases was noticed when comparing domestic and wild birds. Differing susceptibilities to the disease, conditions impacting disease spread, or other ecological factors could all play a role in this. If domestic birds, such as chickens or turkeys, come into contact with wild birds or their faeces, they run the risk of contracting the virus, which can cause outbreaks in domestic flocks. Wild birds and domestic ones can come into contact directly, or indirectly, like when domestic birds eat contaminated grain that wild birds have been nearby. Domestic bird outbreaks pose a triple threat: they can damage poultry health, production, and economic outcomes. Understanding the interplay between wild bird and domestic bird cases is crucial for studying AVIN epidemiology. By monitoring the virus's spread among wild bird populations, farmers can anticipate potential risks to their domestic flocks, enabling them to implement prompt and efficient disease control measures.

Previous research suggests that Twitter is the most effective social media platform for obtaining useful data on a dynamic time series basis[399]. The 280-character constraint encourages messages that are both informative and concise, eliminating unnecessary noise from the content. Hashtags accelerate information flow and create communities of similar-minded people through shared topics (i.e. #avian flu UK). Based on the tweets analysed in this study, the results illustrate that most of the content pertains to public health communication and promoting awareness. Following on from a confirmed case, there is a noticeable increase in the number of topics discussed, which is evidenced by spikes in frequency, therefore making it difficult to act as an early warning tool without further augmentation.

With this awareness, animal health planning could be significantly improved. "Awareness-oriented" messages differ from "outbreak-oriented" messages in terms of their volatile fluctuations and short-term frequencies[21]. Social media played a crucial role in promoting public health awareness during the 2014 Ebola outbreak in West Africa[19]. Better comprehension of content and improvement in veterinary health guidelines can be achieved by analysing the data extracted from these messages.

7.4.1 Causality

Understanding how public discourse and real-world events interact relies on determining the causal relationship between social media activity (tweets) and confirmed AVIN outbreaks. The findings from this chapter indicate that the main causal flow is from real-world AVIN outbreaks to social media activity. Essentially, tweets are primarily a reaction to the news of confirmed outbreaks rather than being the catalyst for news creation.

7..4.1.1 Evidence from Temporal Analysis

The analysis of tweets and outbreaks strongly supports this causal relationship. Tweet volumes saw significant increases that closely correlated with spikes in outbreaks during two crucial periods: October 2021 to January 2022 and August to December 2022. The alignment suggests that social media activity intensifies as the number of confirmed outbreaks grow, indicating that tweets are a reaction to the increasing number, rather than vice versa.

Moreover, a marked increase in AVIN-related tweets coincided with a significant uptick in outbreaks during the first period (October 2021 to January 2022). The decrease in tweet volumes aligns with the decline in outbreaks, reinforcing the idea that social media reflects real-world events. The observed pattern suggests a reactive response, with increased public concern and awareness triggered by news of rising AVIN outbreaks, resulting in a higher number of tweets.

7..4.1.2 Insights from Google Trends

The direction of causality is further supported by the inclusion of Google Trends data, as shown in Figure 45. According to the trends, whenever there is a spike in tweeting, there is a corresponding uptick in Google searches. According to this sequence, users tend to start by engaging with social media, where they react to news and discussions about AVIN outbreaks, and subsequently rely on online searches to gather further details. Tweets create awareness and inspire more research, but the initial motivation to tweet or search is the news of outbreaks.

7..4.1.3 Statistical Correlation and Temporal Lag

By examining ACF and PACF plots (Figure 46), it becomes evident that there are significant autocorrelations at short lags in the time series analysis, implying that higher tweet volumes on one day tend to result in higher volumes on subsequent days. Nevertheless, these patterns do not indicate that tweets are the cause of the surge in confirmed outbreaks. Rather, they show an ongoing public response to the news on the AVIN outbreak.

The finding from the correlation analysis reveals a weak negative correlation between daily tweet counts and confirmed AVIN outbreaks, providing additional evidence that tweets are not the primary influencers of news. However, this inverse relationship could indicate that the public becomes less sensitive to ongoing outbreaks, resulting in a decrease in tweet activity as the news loses its novelty.

7..4.1.4 Causality Summary

AVIN social media activity is influenced by outbreak confirmations, as demonstrated by this chapter. The primary basis for the evidence is:

- The alignment of spikes in tweet volumes with increases in confirmed cases during specific periods reveals the temporal synchronisation of tweets with news about AVIN outbreaks.
- By observing the sequence of tweeting and subsequent Google searches, it becomes evident that social media reactions are a response to news of confirmed cases.
- The negative correlation between tweet volumes and outbreak counts is weak, indicating that social media activity reflects public awareness and concern rather than influencing the news.

Having this comprehension is vital for using social media effectively in monitoring public health and engaging with the community, as it plays a significant role in reflecting and amplifying the public's response to real-life events.

7.4.2 Limitations

The evaluation of this analysis's limitations is essential. The data only pertains to tweets and may not reflect the general public sentiment or concerns regarding the disease. Additionally, the geographic spread of tweets could be impacted by variables like population density and internet availability, which might result in an exaggerated number of tweets from metropolitan regions. The results are based on a particular time frame and may not consider changes in public interest or awareness over time.

The study assumed that tweets reflect public awareness or response, but this may not be accurate. Another concern is the accuracy of disease reporting, particularly in wild birds. Causation can't be inferred from the observed correlations, according to the findings.

Although there are limitations, spatial analysis can still provide valuable information to aid public health authorities and policymakers in targeting their communication strategies and interventions more effectively. Focusing on areas with a higher tweet frequency can help them address public concerns and promote outbreak prevention and control measures more effectively.

7.4.3 Further work

The current study could be improved by conducting additional research that incorporates advanced text analysis techniques like LDA and sentiment analysis to analyse AVIN-related tweets. Employing LDA can assist in identifying the underlying subjects discussed in tweets, thereby enhancing the comprehension of the public discourse on AVIN outbreaks. By adopting this approach, it can uncover thematic patterns and trends that may not be readily visible through simple keyword analysis, similar to the work conducted by Egger and Yu (2022)[371]. Moreover, sentiment analysis can be used to measure the emotional tone of the tweets, providing valuable information about public sentiment and its potential influence on tweet activity and trends. The integration of these techniques can enable future studies to conduct a more comprehensive analysis of social media data, resulting in enhanced accuracy and depth in monitoring and predicting outbreaks. Implementing this enhanced approach would not only result in a more comprehensive dataset for analysis but also aid in the development of more sophisticated predictive models, as highlighted by Lamsal et al (2022)[395].

To improve the study's predictive capabilities, it would be beneficial to include Granger causality tests and Autoregressive Integrated Moving Average with Exogenous Variables (ARIMAX). ARIMAX models in addition to text analysis, thus establishing a cause-and-effect connection between social media activity and outbreak events[395]. By examining latent variables derived from social media data, these tests offer valuable insights into how public discourse and actual outbreak occurrences are connected over time[400].

Additionally, ARIMAX models, which incorporate exogenous variables like tweet counts, present a more robust method for analysing time series data, by accounting for external factors (i.e. The concurrent COVID-19 pandemic during the study time period), these models enhance the accuracy of predictions.

7.5 Conclusion

This study emphasises that using social media data as the sole surveillance tool for disease outbreaks has its drawbacks. Although relying solely on social media for surveillance may lead to an imprecise portrayal of the confirmed cases' spread, it can serve as a valuable resource for comprehending the public's outlook and apprehensions during an epidemic.

The study framework was limited to a micro-level, focusing on a single disease type, timescale, and location. This presents opportunities for further analysis at a global level and the development of a larger training dataset. Robertson and Yee's work is a prime example of global level analysis with successful adaptations of such methods already established[21].

Further investigation is required to validate these discoveries and optimise social media's role in disease surveillance. This involves considering the representation of social media data and improving disease reporting accuracy. Advanced data mining techniques and real-time analysis can further enhance the precision and promptness of outbreak detection and response. A strong and comprehensive surveillance system is critical in preserving public health by efficiently monitoring and managing disease outbreaks.

8/. Consolidation of results

The key discoveries from the results sections are compiled in this chapter, outlining how they align with the research aims of the project. It clarifies the justification for every experimental chapter and exemplifies the common themes that are interwoven, uniting the individual parts into a cohesive narrative.

8.1 Inter-chapter Connections

8.1.1. Chapter 4 - Typifying Smallholders

The opening chapter demonstrates the potential of utilising text classification algorithms to classify small-scale farmers on social media. Machine learning techniques can be effectively employed to target specific demographics, who may not be well-represented in traditional registration and surveillance systems, as shown by the high accuracy achieved in classifying Twitter profiles.

8.1.2. Chapter 5 - Social Network Analysis

The focus of this chapter shifts to the network structure of smallholder communities on Twitter, following on from the classification work. This analysis adds a layer of understanding to smallholders' online interactions by identifying influential users and visualising user connections. The network perspective is in line with the text classification discussed in Chapter 4. Offering a deeper understanding of how these communities operate and the influential users who can control the narrative within these groups.

8.1.3 Chapter 6 - Forum Scraping

Employing web scraping, text mining, and anomaly detection, this chapter delves into livestock farming forums, moving beyond Twitter. By uncovering key themes and topics within online discussions, the previous two chapters are complemented with more nuanced insights into concerns and practices. The significance of human validation is underscored through the combination of traditional and internet-based surveillance methods.

8.1.4. Chapter 7 - Spatio-temporal Analysis

By using the example of Avian Influenza, the final chapter evaluates the potential of social media in tracking livestock diseases and ties together with the preceding work. By connecting social media discourse, public sentiment, and actual disease occurrences, the spatial and temporal analyses create a bridge. Despite finding a weak correlation between tweets and confirmed cases, the chapter highlights the importance of using integrated and validated surveillance approaches, which builds on themes from previous chapters.

8.2 Addressing Research Questions

	RQ1	RQ2	RQ3	RQ4	RQ5
Chapter 4	🟩	🟩	🟥	🟥	🟩
Chapter 5	🟩	🟩	🟩	🟥	🟩
Chapter 6	🟥	🟩	🟩	🟥	🟩
Chapter 7	🟥	🟩	🟥	🟩	🟩

FIGURE 53: RESEARCH OBJECTIVES ADDRESSED BY EACH CHAPTER.

RQ1. Is it possible to differentiate between a smallholder and non-smallholder user based on their online profile and content? The results obtained in chapter 4 directly deal with this question. By means of a training dataset that was manually annotated, the text classification model was able to attain an accuracy of over 80% in categorising Twitter profiles as smallholder or non-smallholder. This proves the feasibility of distinguishing smallholder and non-smallholder users based on their online content. The analysis was exclusively carried out on the user profile descriptions, although further analysis can be done by considering the user's recent user timeline. Furthermore, the validation of the predictive model on a new profile displayed exceptional ability to identify a non-smallholder profile, but moderate ability to identify a smallholder – indication further refinement and keyword refinement is required to improve this metric.

RQ2. What are the differences between the social media extracted database and the current knowledge database? The research in all four chapters investigates the distinctive viewpoints gained from social media data that may not be accessible in traditional databases. In particular, the chapters on text mining and social network analysis highlight the abundance of context-specific data accessible through online platforms, which provide a more detailed understanding of behaviours and concerns. Social media data and traditional surveillance systems were compared with official data sources in the spatio-temporal analysis, highlighting their differences and potential complementarity. Chapter 6 showcased the geolocation disparities between APHA conformed pig and poultry holdings, compared with the forum-extracted user locations. This snapshot picture allows for further research into why some stark differences were apparent (i.e. Northeast England).

RQ3. How do these communities exchange information and what is the information being discussed? The detailed examination of smallholder communities' online information exchange is provided in the chapters on social network analysis and text mining of veterinary forums. Key themes, words, and topics are uncovered in their discussions, revealing distinct communities within farming sectors. Understanding the dynamics of information flow within online communities is further enhanced by studying influential users and community structure.

RQ4. Can social media sources be used as an early predictor of disease? The spatio-temporal evaluation chapter is the sole section that examines this issue from the Avian Influenza viewpoint. The research revealed a weak correlation between the number of tweets and confirmed outbreaks but indicated that social media may offer valuable insights into public concerns and sentiments during outbreaks. Nevertheless, as with all the other chapters, it underscores the importance of combining social media with traditional surveillance methods to improve the precision and monitoring, underlining the intricacy of relying solely on social media as an early warning. The direction of causality highlighted that users reacted to the news of an outbreak after the outbreak had been confirmed.

RQ5. Is it possible to extract locations of smallholders and/or disease outbreak hotspots? Similar to RQ2, all four chapters included elements of location extraction and geolocation mapping to plot the spatial distribution of the users. The spatio-temporal analysis delves deeper into the confirmed disease cases' spatial patterns and related tweets, mapping the intersections between them, in contrast to the basic spatial analysis presented in the first three chapters. Demonstrating spatial correlation and regional targeting, this chapter supports the idea of extracting smallholder or disease hotspot locations through social media data. Improved data mining techniques could enhance the precision of location-based analysis, as indicated in the chapter. As all locations were extracted from a free-text option in the user profiles, it should be noted that the validity around the use of such methods is still ambiguous, particularly when compared with the precise confirmed case locations data from APHA and DEFRA.

8.3 Conclusion

Through the four chapters, a narrative is woven that progresses from identifying and classifying smallholders online to understanding their community structures and information exchange patterns, and ultimately, evaluating social media's potential and limitations in epidemiological surveillance.

The chapters have common themes - social media's impact on traditional surveillance, targeting underrepresented smallholder communities, and combining data sources for better livestock disease understanding. This is discussed in greater detail in the following discussion chapter.

By consolidating the results, the interconnected nature of the research questions is effectively addressed, and common themes are highlighted across the chapters. This lays the groundwork for the following discussion chapter, which will delve deeper into these findings and place them in context within the wider scope of veterinary epidemiology and livestock health.

9/. Discussion

A comprehensive study of small-scale farming communities in the UK was conducted, specifically examining the integration of social media surveillance, data science techniques, and small-scale livestock farming. By exploiting unconventional data sources and pioneering methodologies, the study aimed to uncover valuable insights into the challenges, dynamics, and prospects of social media mining pertaining to smallholder pig and poultry farming.

The research employed text classification, social network analysis, web scraping, and spatio-temporal analysis to produce a comprehensive set of findings that are of great benefit to both theoretical understanding and practical applications in the field.

During the ensuing discussion, connections with existing literature will be made, and the methodological approach will be reflected on to contextualise and interpret these findings, whilst considering the broader implications of the work. The section aims to explore the interconnectedness of the results from the social media, smallholdings, surveillance, and data science experimental chapters by analysing the common themes and threads that run through them.

The following structure will be used for the discussion:

- Themes and common threads: The recurring aspects which run through the entire project.
- Interpretation of findings: A comprehensive review of the primary outcomes, corresponding to the research questions.
- Theoretical implications: Determined through an examination of existing theories and literature.
- Practical applications: 3 distinct frameworks for dashboards and reproducible applications are put forward.
- Misinformation and digital pitfalls in social media data: Tackling the common data issues with social media and the procedures put in place to curb these effects.
- Limitations and Future Research: Reviewing the study's boundaries and identifying potential avenues for future research.

9.1 Common themes and threads

"Themes guide research with abstract concepts, while common threads bind chapters and findings together"

This section explores the themes and common threads which permeate across the four results chapters. This will depict their influence on shaping the research's progression and their contribution in addressing the core research questions. Understanding these interconnections enhances our knowledge of smallholder farming in the UK, their obstacles, and the creative use of data science techniques for potential solutions. Throughout the multifaceted exploration of these communities and their role in veterinary epidemiology surveillance, guiding principles have included both broad themes and specific common threads.

The central themes of this research are the criticality of understanding smallholder communities, their role in the spread of diseases, the potential of social media in disease surveillance, and the overall implications for public and animal health. The central ideas represented by these themes set the stage for the research questions and objectives to be addressed.

Similarly, common threads weave the different parts of the research together into a unified whole, complementing the themes. Specific methodologies, such as text classification, network analysis, and spatio-temporal evaluation, are included. Recurring patterns include biases, insufficient sample sizes, and the need for human validation as challenges. Highlighting an innovative yet grounded approach, the study emphasises the integration of social media analysis with traditional surveillance methods.

9.1.1 Themes – Guiding concepts

1/. Enhancing Traditional Surveillance: Modern farming is continuously evolving, driven by technological advancements and data-driven insights, despite its deep roots in time-tested practices. This research acknowledges the significance of traditional surveillance systems in monitoring and understanding livestock health and farming practices. However, it also highlights areas that can be improved. By incorporating digital tools and methodologies, such as spatio-temporal analysis and text classification, the study sought to enrich traditional surveillance. Instead of replacing existing practices, the aim is to enrich them, providing a more complete understanding of the complexities and swift changes in modern farming.

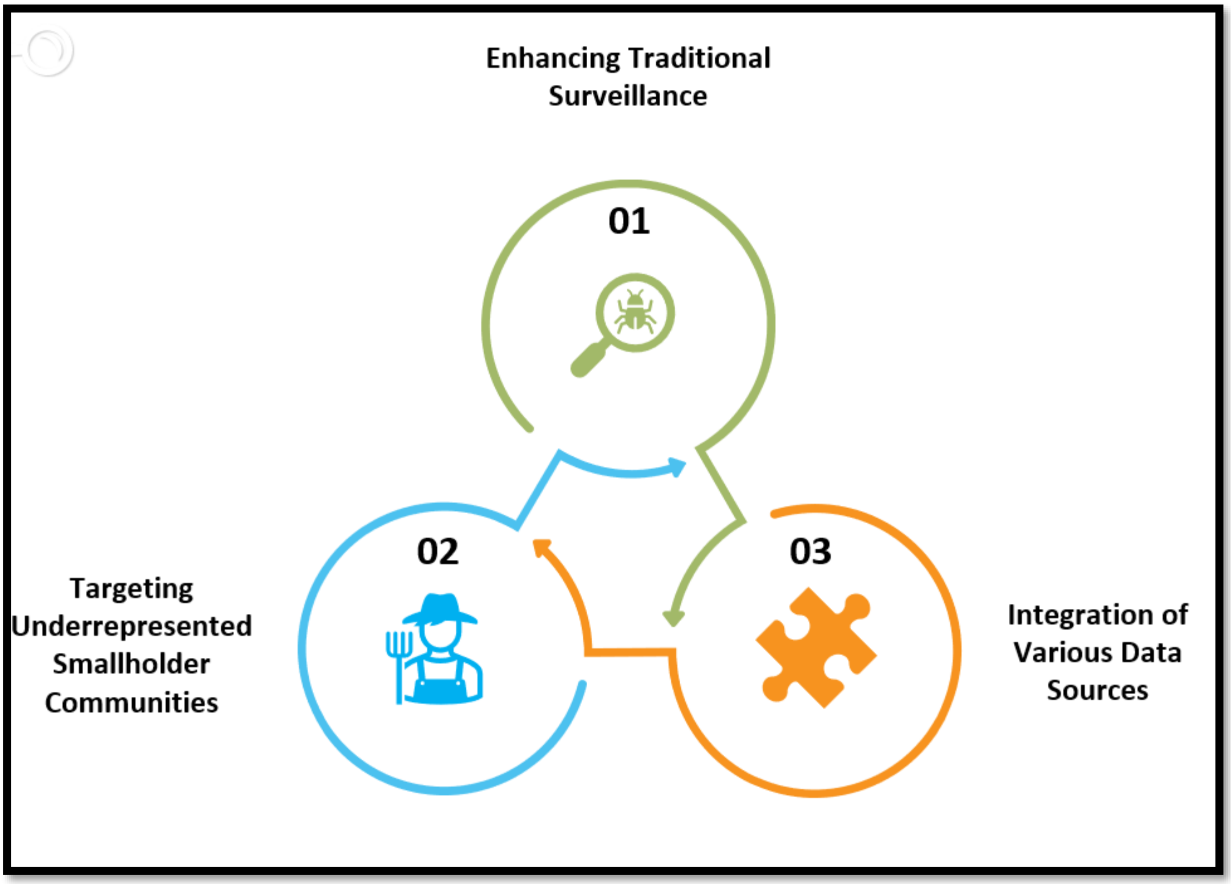

Enhancing Traditional Surveillance
01
02
Targeting Underrepresented Smallholder Communities
03
Integration of Various Data Sources

FIGURE 54: COMMON THEMES

2/. Targeting Underrepresented Smallholder Communities: The main goal of this research was to bring attention to communities that are often overlooked. Although smallholder farmers play a significant role in agriculture, they are often underrepresented in broader discussions and datasets. This study highlights these communities in every chapter, be it analysing disease outbreaks or exploring social network interactions. Despite each results chapter having a different goal, the central focus of each was to build our current intelligence around smallholding communities.

3/. Integration of Various Data Sources: In today's digital age, data exists in various forms and across multiple platforms. The integration of multiple data sources is emphasised in this research. The study provides a comprehensive interpretation by examining both social media interactions and structured data sets from organisations such as APHA and DEFRA. Although it's challenging, the merging of data results in a more enriched and multi-dimensional understanding of small-scale farming. The significance of developing a comprehensive narrative lies in having not only a large amount but also a diverse and high-quality data. To achieve the research objectives, the analysis exploited Google trends, Twitter, and Forum data. Incorporating multiple data sources concurrently could be an extension to detect overlapping discussions and users.

9.1.2 Threads – Connecting concepts

1/. Methodological Consistency: Throughout the research, there is a clear dedication to methodological rigour and consistency. Despite the distinct focus of each chapter, there are common methodologies that are frequently employed, including text classification, topic modelling, social network analysis, and spatio-temporal evaluation. While grounded in data science, these techniques are not applied in isolation. They have been tailored to fit the individual requirements of small-scale livestock farming communities. The use of a consistent methodological approach guarantees that the obtained insights are strong and contextually applicable.

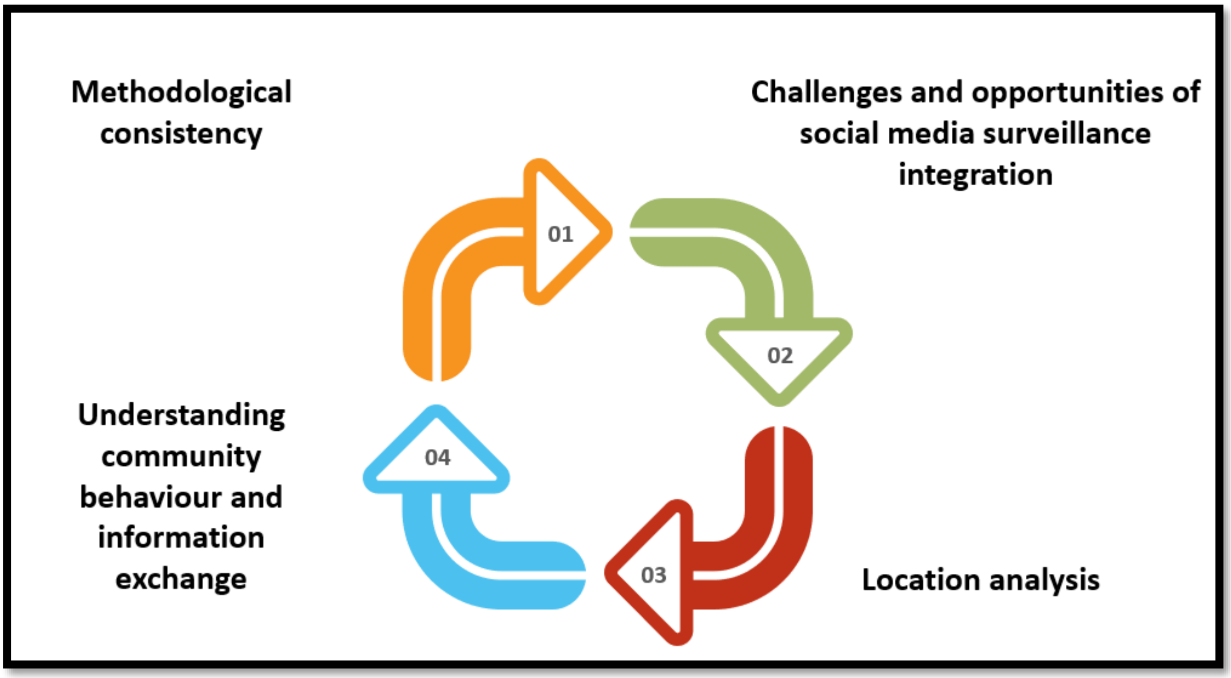

Methodological consistency
Challenges and opportunities of social media surveillance integration
01
02
03
04
Understanding community behaviour and information exchange
Location analysis

FIGURE 55: COMMON THREADS

2/. Challenges and Solutions in Surveillance Integration: Acknowledging the transformative role of digital platforms is necessary to explore modern farming. Throughout the thesis, there is a continuous discussion about merging social media data with conventional surveillance systems. There are some challenges that come with this integration. The issue of data validation and the correlation between digital conversations and on-ground cases arises as a major concern. Chapter 7 highlights some of the current deficiencies faced with social media integration of Avian flu outbreaks and proposes solutions on how to deal with the current disparities facing Infoveillance.

3/. Understanding Community Behaviour and Information Exchange: The study centres on smallholder farmers and their digital interactions. Every chapter explores the complexities of online interactions, information exchange, and identifying influencers in digital ecosystems in its own way. A composite image of the community's behaviour is stitched together by this thread, stressing communication patterns, shared concerns, and knowledge dissemination.

4/. Location and Disease Outbreak Analysis: The research narrative hinges on geographical analysis. The geographical component is always present, whether it's locating smallholder farms or pinpointing disease outbreak hotspots. The research creates a spatial narrative by mapping these locations and overlaying them with other datasets, indicating regions of concern or success. Targeted interventions can be informed by geospatial insights, ensuring that resources and efforts are directed where they are most required.

9.2 Interpretation of findings

Despite each results section containing its own individual discussion, it is important to highlight the main findings and implications in the context of the entire project, rather than as an individual entity. This links in with the previous section highlighting the common themes and threads permeating throughout the results.

9.2.1 Typifying smallholders

A better understanding of the smallholder community in the UK is offered by the results from chapter 4. By classifying Twitter users based on their profile descriptions, the complexities and subtleties within these communities become more apparent. This categorisation not only gives a demographic overview but also provides an insight into the interests and concerns on a greater level.

The wider aim of this chapter was to create a database which can be used alongside the current intelligence we have around this demographic. The predictive model generated by this analysis builds the foundation for a data pipeline to be built, which inputs users into the algorithm, hence generating an output of “smallholder or not” based on user profile descriptions. I discuss the practicalities of this in greater detail in the upcoming sections.

One of the significant findings is the disparity of classes within the dataset, which underscores the heterogeneous nature of the smallholder community. These insights have implications for tailoring interventions and resources, specific to the needs of various sub-groups within the larger community. As explained in the introduction, determining the nature of smallholders is not merely a binary classification task and contains many nuances which need to be examined through specialist livestock domain knowledge.

Non-smallholders have a significant role in these networks. As the discussion suggests, their presence might be motivated by a desire to advocate for smallholding practices or a genuine curiosity about the intricacies of small-scale farming. This dynamic suggests a two-way information street, where smallholding accounts aren't just recipients but active contributors, enriching the discourse with their lived experiences and insights. However, the results from this chapter partially allow for the initial classification of segregating the over-arching smallholder cohort from the rest of the population, setting up further explorations which delve into the intricacies within the cohort itself (i.e. crofter, hobbyist etc).

The chapter showcases a highly valuable machine learning application in identifying relevant users based on their Twitter profiles. The wider implications matter as much as the accuracy and precision of these models. With machine learning, the livestock industry can achieve excellent identification results, as shown by the superior performance metrics of models, such as RF and LR.

9.2.2 Social network analysis

Chapter 5 provides insights into the social dynamics among smallholder networks on Twitter. Through the exploration of connections, interactions, and influential nodes, the study uncovers the patterns of information dissemination and the key people driving the discussions. The analysis of common word pairings and frequent phrases provides a window into the prevalent topics and concerns within the community. These findings have significant implications for targeted communication strategies, where identifying and engaging with influential nodes can amplify the reach and impact of messages.

This builds from the findings of the previous chapter, by delving directly into the community and extracting the relationships amongst users. This is the unique aspect of this chapter and has been reinforced by its adaption into a peer-reviewed paper.

The network displayed a distinctive scale-free topology, as revealed by the findings. In this structure, there are only a few nodes that are highly connected while the majority of nodes have fewer connections. A small number of influential entities play a crucial role in shaping public discourse and enabling information sharing. The study found eleven distinct communities in this expansive network, indicating diverse interests and group affiliations.

The results suggest that spatial analysis can reveal patterns in the spread of disease that may not be apparent through temporal analysis alone. Disease hotspots can be identified through spatial analysis, which in turn helps contain an outbreak before it gets out of hand. This approach accurately captures the geographical trends shaped by migratory bird routes or areas of high-density poultry farming, and the diverse spread of the disease. By identifying high-risk areas, this approach aids in the efficient allocation of resources, targeting surveillance and control measures. In addition, by analysing the spatial dimension, we can gain a better understanding of the relationship between environmental factors, such as landscape, climate, and ecosystems, and the spread of disease.

To fully comprehend the network demographics, it's crucial to address this deficiency by adopting user-categorisation of profiles of the influential users (as performed in chapter 4).

9.2.3 Forum scraping

Chapter 6 explored a different aspect of digital presence. Due to their specificity and focus, forums often facilitate detailed discussions and knowledge exchanges. Overlapping common themes and locations in the results suggests that the community shares concerns and interests. The granular and specific insights from this data extraction method are invaluable for understanding the depth of expertise and concerns within the community.

In-depth insights into the priorities and concerns of the pig farming sector were revealed through discussions on the topic. The significance of breed choices was emphasised by the results, with "Kune Kune" being a preferred option. Due to the breed's size being appropriate for small-scale farming and its friendly nature, such inclination could be attributed. Pig farming operations were clearly visible in the discussions about "slap mark" and "electric fence," indicating the practical aspects of animal husbandry.

Similarly, the topic of ethical farming practices were apparent in the discussions on poultry farming. The use of "free range" not only supports contemporary farming values, but also displays a commitment to animal well-being. Health-centric practices for optimal poultry care were highlighted in discussions about nutritional aspects, including layer pellets and red mites.

Moreover, the outcomes concerning anomaly detection established a foundation for an upgraded system to alert about disease outbreaks. The system is intended to function as an early warning mechanism by identifying irregularities in user activity, especially in those who experience sudden increases in conversation. Drawing comparisons to methodologies used during global events like

COVID-19, it emphasises the need for anticipatory actions in the realm of livestock farming. This is discussed in further details in the upcoming sections.

9.2.4 Spatio-temporal analysis

The final experimental chapter's results are crucial for linking online discussions to real-world events. The study highlights the potential of real-time surveillance by analysing Twitter discussions and temporal patterns in relation to confirmed disease cases. The correspondence between Twitter activity and disease outbreaks like Avian Influenza reveals that online platforms have the potential to be early warning systems. Moreover, the spatial analysis can be used to guide targeted interventions by identifying regions with higher levels of public concern and ensuring that resources and information are directed towards them.

By also taking a temporal perspective, the study uncovered moments when Twitter discussions about avian flu corresponded with actual outbreaks. These parallels imply that the public is more vigilant and involved in addressing emerging health threats. While the Pearson correlation analysis revealed a weak relationship between the volume of tweets and validated cases, it cautioned that depending only on tweet frequencies might not offer a full understanding of outbreaks.

Additionally, the spatial analysis revealed interesting correlations between tweet locations and verified cases, with urban areas standing out as a major hotspot for tweeting. Communities might be actively discussing and sharing information on outbreak incidents in these intersections. Nevertheless, when taking a wider approach into account, guided by the bivariate Moran's I assessment, it becomes apparent that there is a generally weak negative spatial link, emphasising the complexity of drawing direct geographical inferences.

Lastly, the analysis brought to light differences in cases between domestic and wild birds. These variations raise questions about the varying susceptibility to diseases, or other external variables that could cause such discrepancies.

9.3 Theoretical implications

The findings reveal the intricate dynamics of the UK's smallholder community, exposing complex behaviours, perceptions, and relationships. Through established theoretical frameworks, the results validate current concepts and present new viewpoints. These theoretical frameworks, when viewed in conjunction with the research results, create a compelling narrative. Although the results largely confirm the principles of the HBM and Social Physics, as mentioned in the Introduction chapter, they also highlight the complexities present in real-life situations. Digital footprints shed light on how health beliefs, perceptions, and social interactions are interlinked, as seen through Twitter and forums. The study not only connects theory and practice but also highlights the necessity of interdisciplinary methods in confronting challenges in small-scale agriculture.

9.3.1 Health belief model

According to the HBM, individuals are more likely to take health-related actions if they perceive a threat and value the actions aimed at reducing the threat. The results align with this theory in various aspects:

- The rapid influx of tweets during disease outbreaks suggests that people perceive the situation to be both serious and potentially affecting them. Online discussions serve as an important source of cues to action, which are key to the tenets of the HBM.
- The collective belief in the benefits of humane farming is evident in the discussions about ethical farming practices, with a particular emphasis on "free range". The model suggests that perceived benefits play a significant role in shaping health-related behaviours, and this aligns with it.
- The discussions often revolve around challenges in smallholder farming, which leads to a perception of barriers. Effective interventions depend on overcoming these barriers.

9.3.2 Social physics

The integration of the social physics framework can provide a predictive tool for future actions by quantifying patterns of user interactions, which offers deep insights into group behaviour. The principles of this theory are reiterated by the findings from the social network analysis and forum scraping chapters.

- The social network's scale-free topology suggests the presence of a small number of highly influential nodes that exert behavioural influence. This serves as an illustration of Social Physics' concept that a small number of influencers can shape group behaviour.
- The diversity of human interaction patterns is evidenced by the discovery of eleven distinct communities in the social network, each with their unique interests and affiliations. According to social physics, these patterns can provide predictive insights when the communities are thoroughly examined. The extension into the user classification work within this chapter also builds on this notion, by examining distinct clusters to gauge the inter and intra cluster connections within the overall dataset.

- The granularity of discussions on forums and active exchanges highlights that collective knowledge impacts decision-making within communities, relating to the concept of information flow within the model.
- As mentioned in section 1.6.11, to understand the changes in networks and activities, temporal analysis is necessary. Temporal analysis in chapters 6 and 7 reinforce this particular aspect of social physics as an examination into the fluctuations of messaging activity was studied.

9.3.3 Social Media Surveillance and Epidemiology

The use of social media as a surveillance tool for disease monitoring contributes to the increasing literature on digital epidemiology. The literature review discussed social media's role in public health surveillance, but Chapter 7's spatio-temporal analysis provided proof of its potential and limitations. By focusing on web scraping and Twitter API calling, this aligns with grounding theoretical implications in practical research techniques. The aggregation of the findings bolsters the background section laid out in section 1.6.5 and contributes to the growing body of knowledge around the shifting paradigms of surveillance through the incorporation of internet-based data.

A thorough analysis of the findings strongly supports the claim that social media can effectively be utilised for monitoring disease trends, establishing trigger limits, and operating as early warning systems. The points below demonstrate how this work supports these assertions:

*10.3.3.1 **Checking the thresholds for triggering actionable responses***

The study's approach, specifically the use of anomaly detection methods like the Isolation Forest algorithm, shows that it is possible to set specific thresholds for acting based on identified anomalies. Chapter 7's spatio-temporal analysis revealed a strong connection between spikes in social media activity and outbreaks of animal diseases. The presence of these spikes serves as an early indication, validating the application of social media data to establish thresholds that prompt further investigation and intervention. This is in line with established practices in epidemiological surveillance, where predetermined thresholds prompt public health interventions[401]. One issue which was detected in Chapter 7 pertained to the direction of causality, with the tweet responses following a confirmed outbreak rather than during or prior, hence setting message frequency thresholds concerning outbreaks may not be most appropriate method.

Nevertheless, actioning certain hashtags during an epidemic/pandemic and spreading awareness around spotting clinical signs through the lens of social media. An example of this would be converting the APHA clinical signs documentation to TikTok/YouTube/Instagram short videos, supplemented with specific hashtags such as *#swollenheadswansheffield*[402]. Adding the clinical sign, the species type and a location can mitigate the noise associated with such messages and allow more appropriate message thresholds to be put in place.

10.3.3 2 ***Applying real-time monitoring in practical scenario***

Using web scraping and Twitter's API together provides a practical solution for monitoring diseases in real-time. According to the results, it was possible to extract time-sensitive and location-specific information, allowing for the swift identification of potential disease hotspots on an aggregate scale as co-ordinates are not available. For example, the grouping of keywords associated with symptoms and outbreaks in particular geographic areas highlights the possibility of Twitter/X as a real-time source of epidemiological information[403]. The ability to detect and contain disease outbreaks in real-time is vital but is made difficult by the lack of location precision and social media noise, thus requires other avenues of supplementary information.

10.3.3.3 ***Strengthening surveillance with community identification***

By applying community detection algorithms, distinct clusters of influential users were identified within the network of Twitter interactions. The surveillance system becomes more effective by identifying these influential nodes, as they can amplify alerts and information within their networks. This method enhances the effectiveness and scope of surveillance efforts, ensuring the swift and efficient distribution of vital information. Creating a database of such users can be of great value, especially to disseminate public health information, but once again may be met with some pushback and surveillance aversion.

10.3.3.4 ***Support for the efficacy of an early warning system***

Table 16 presents the results, while Table 17 showcases the performance metrics for user profile categorisation, both highlighting the models' high accuracy and precision in classifying smallholder accounts. The capacity to precisely recognise and classify users using their profiles and activities implies that comparable methods could be used to identify and track disease-related discussions. The models' high sensitivity and specificity suggest they could be effective in an early warning system. In situations where diseases spread quickly, timely detection and response are particularly important, but additional validation of the predictive model needs to occur to bolster the mediocre performance of categorising smallholding only accounts.

10.3.3.5 Limitations

Although the results show promise, they also underscore the need to address certain limitations. Relying on the quality and completeness of user-generated content poses a major limitation. Including additional data sources and using advanced NLP techniques could improve the surveillance system's reliability by reducing misclassifications caused by ambiguous or incomplete descriptions. The integration of forum discussions and other social media platforms should be a focus of future research to enhance the comprehensiveness of the surveillance network.

9.3.4 Data science in the livestock domain

Employing data science techniques, such as topic modelling, social network analysis, and anomaly detection, provides a new approach to veterinary and agricultural research. Combining data science methods with the techniques used in Chapters 6 and 7 shows that these methods are adaptable and effective. The convergence of theory and practice sheds light on new ways to use data science in unconventional and niche research fields.

9.3.5 Enhancing digital literacy

The results from chapter 6 highlight the depth of discussions and knowledge exchange within digital platforms, and this opens up opportunities to guide initiatives aimed at enhancing digital literacy among smallholder communities. Such programs can help ensure that they can effectively navigate and benefit from online resources. Similar digitalisation initiatives have proven to be effective in Sub-Saharan Africa, and substantial research has shown the positive uptake from small-scale communities in adopting such measures[404].

9.3.6 Ethical Considerations in Digital Research

The ethical considerations stated in the methodology section, such as informed consent, confidentiality, and privacy during data collection, are relevant to larger discussions in digital research ethics. These findings expand the theoretical discussion on responsible data usage, underscoring the crucial role of ethics in modern research. Responsible digital research's theoretical understanding gains depth with the alignment of ethical practice and literature review's emphasis on public engagement.

Building upon the literature review in section 2.2.6, the existing body of work emphasises that ethical considerations determine whether users find it socially acceptable for the government to use social media for surveillance. The willingness to adopt these methods is affected by trust in the research process, perceived benefits and risks, and transparent data handling. Figure 56 reiterates the key considerations by authorities to be actioned before large-scale digital surveillance implementation can occur.

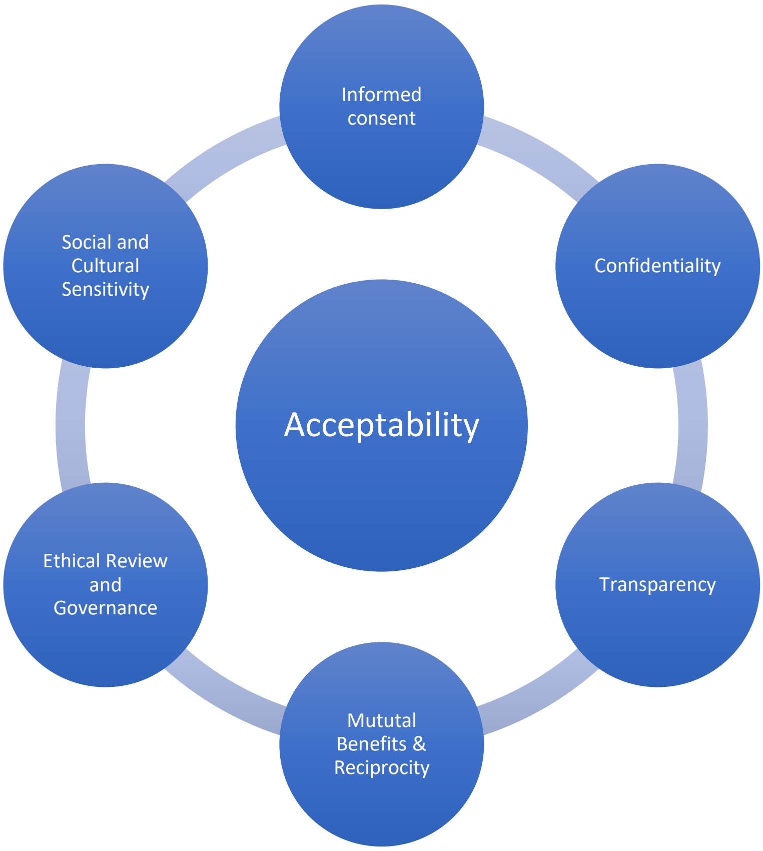

Informed consent
Confidentiality
Transparency
Mututal Benefits & Reciprocity
Ethical Review and Governance
Social and Cultural Sensitivity
Acceptability

FIGURE 56: CONSIDERATIONS FOR THE SOCIAL ACCEPTABILITY OF DIGITAL EPIDEMIOLOGY

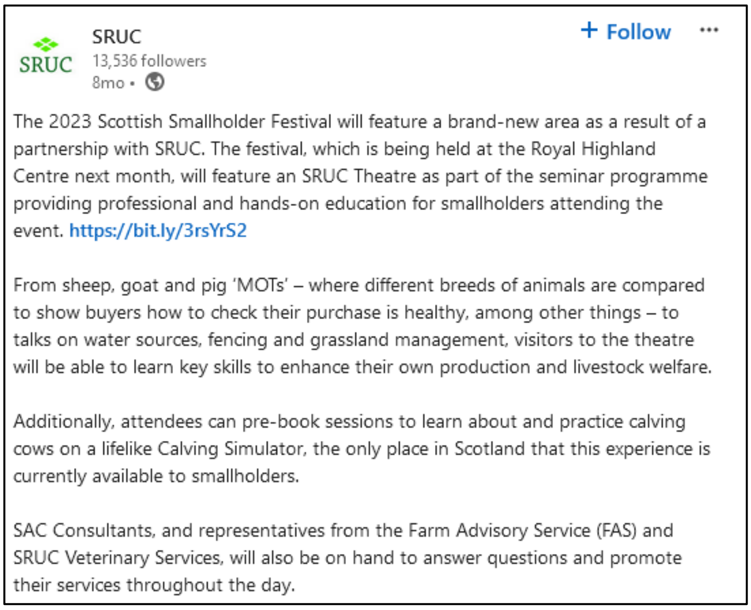
9.3.6.1 Supporting Social Acceptability

In order to promote the social acceptability of these methods, various strategies can be put into place, as highlighted in table 38:

Method	Explanation	Resources & Evidence of feasibility
Community Workshops	Conducting community workshops can greatly improve acceptance by smallholders about digital surveillance and involving them in the development process. An example could be APHA or SRUC hosting workshops to showcase the usage of social media data in this study for identifying disease outbreaks, explaining the methodology and addressing concerns. Creating interactive sessions allows smallholders to ask questions and contribute, promoting collaboration.	https://www.ruralnetwork.scot/protect-your-flock-best-practice-biosecurity-webinar-smallholders-poultry-and-other-captive-birds https://festival.scot/rural-skills/ https://www.linkedin.com/posts/sruc_the-2023-scottish-smallholder-festival-will-activity-7112391713918214144-nZyd/
Feedback Loops	Establishing feedback loops allows concerns and suggestions addressed, maintaining trust and engagement. Possible feedback mechanisms in this study could involve surveys or focus group discussions, allowing smallholders to share their experiences with the surveillance system and propose enhancements. Demonstrating commitment to community involvement and responsiveness would involve incorporating this feedback into system updates.	https://aphascience.blog.gov.uk/2018/01/23/how-we-know-where-your-animals-are/ https://www.ruralbrexit.scot/future-policy/farmer-intentions-survey-a-comparative-analysis-of-island-and-mainlands-smallholder-farmers-and-crofters/

Ethical Training	It is crucial to train researchers in ethical best practices and make them aware of the social dynamics of the communities they study. Topics covered in training sessions may include informed consent, data anonymisation, and community engagement. To improve the ethical conduct of the research team, this study suggests including workshops on effective communication and understanding smallholders' perspectives in the ethical training.	https://www.coventry.ac.uk/research/about-us/research-events/2023/ethics-in-ai-and-surveillanc/ https://www.gov.uk/government/publications/public-health-england-approach-to-surveillance/public-health-england-approach-to-surveillance#training
Collaborative Research	Collaborative research involving smallholder organisations can build trust and a sense of ownership – Linking back with the transparency and mutual benefit element of figure 56. To illustrate, this study could work together with key stakeholders to jointly create the surveillance system, guaranteeing it addresses the community's specific needs and concerns- Reaching out to influential users, like the LinkedIn profile posted, would be a great way to prototype such approaches.	https://www.leeds.ac.uk/global/dir-record/profiles/21731/giving-smallholders-a-stronger-voice-through-collaborative-research https://www.linkedin.com/in/rosemary-champion-1658b5148?originalSubdomain=uk

TABLE 38: METHODS FOR PROMOTING SOCIAL ACCEPTABILITY



SRUC
13,536 followers
8mo •

The 2023 Scottish Smallholder Festival will feature a brand-new area as a result of a partnership with SRUC. The festival, which is being held at the Royal Highland Centre next month, will feature an SRUC Theatre as part of the seminar programme providing professional and hands-on education for smallholders attending the event. https://bit.ly/3rsYrS2

From sheep, goat and pig 'MOTs' – where different breeds of animals are compared to show buyers how to check their purchase is healthy, among other things – to talks on water sources, fencing and grassland management, visitors to the theatre will be able to learn key skills to enhance their own production and livestock welfare.

Additionally, attendees can pre-book sessions to learn about and practice calving cows on a lifelike Calving Simulator, the only place in Scotland that this experience is currently available to smallholders.

SAC Consultants, and representatives from the Farm Advisory Service (FAS) and SRUC Veterinary Services, will also be on hand to answer questions and promote their services throughout the day.

FIGURE 57: SRUC PROMOTION OF SMALLHOLDER FESTIVAL 2023

Figure 57 provides a clear illustration of the community workshops and collaborative research mentioned in table 38. SRUC collaborates closely with Scottish smallholdings and veterinary practices, making them an ideal intermediary between government authorities and this community.

9.4 Practical implications and recommendations

This section will create 3 separate blueprints on the practical applications of the findings into real-world scenarios using live data:

1. **Building a smallholder database**
2. **A tool for text analysis of discussions**
3. **An early warning system for outbreak detection**

The models and testing for each have been achieved through this analysis and they have been coded in such a way that they can be easily reproduced.

9.4.1 Building a smallholder database

The first component details the creation of a specialised database of smallholder farms in the UK through the use of machine learning algorithms and data from Twitter and/or forums. This approach had a unique feature in which a pre-trained predictive model, constructed in chapter 4, was utilised to classify Twitter accounts into two categories - smallholder farmers and others. Twitter's large user base and real-time content make it a valuable and untapped source of data on smallholder farming practices, concerns, and issues.

The predictive model is a vital tool in constructing the smallholder farming database, along with the model's implementation and validation. The process to put this framework into production has been depicted in the figure below, comprising of 8 steps:

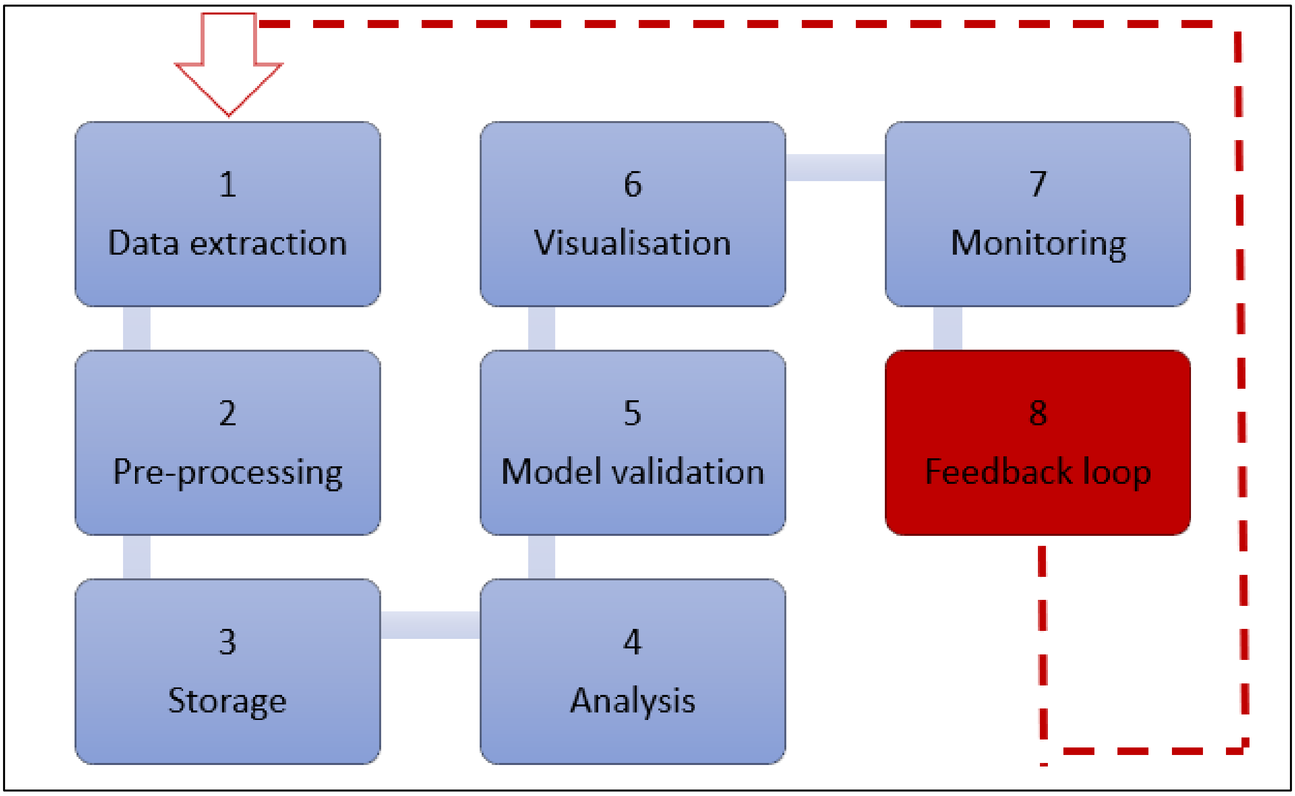


FIGURE 58: SMALLHOLDER DATABASE PRODUCTION PIPELINE

1. Data Extraction:

Tool: Tweepy (for Twitter) and Beautiful Soup/CSS extractor (for Forums)

- Tweepy, a Python library that facilitates access to the Twitter API, can be used to extract tweets, retweets, user profiles, and other relevant metadata. To extract forum posts, user profiles, and threads, use Beautiful Soup, a Python library specifically designed for web scraping. In addition, manually identifying the information to scrape using a CSS extractor can also be performed. The code has been set-up to accept any new twitter username input and any sub-forum link from www.accidentalsmallholder.net.

2. Data Pre-processing:

Tool: Pandas and NLTK

- Using Pandas, a Python library designed for data manipulation, you can easily clean and structure the raw data. The process of preparing data involves several key steps, including removing duplicates, managing missing values, and converting data types. NLTK (Natural Language Toolkit) allows for text normalisation tasks like tokenisation, stemming, and the elimination of stop words. Additional stop words have been implemented within the smallholder domain due to large amounts of noise, which can be found in Appendix A. An extensive livestock-specific stop words list was implemented for the pigs and poultry sub-forums, and similar preliminary work must be conducted depending on the livestock data in question.

3. Data Storage:

Tool: PostgreSQL

- The cleaned and structured data can be saved in an open-source relational database system, such as PostgreSQL. With this approach, data accuracy is preserved, and accessibility is made easier.

4. Data Analysis and Model Training:

Tool: Scikit-learn and TensorFlow

- Scikit-learn library offers a range of machine learning algorithms for data classification and categorisation. TensorFlow is a viable option for those who prioritise deep learning models. This can be applied on structured data to re-train the predictive model from Chapter 4 and make it more accurate and robust.

5. Model Validation and Iteration:

Tool: Scikit-learn and Keras

- After training the model, it's important to validate its predictions against a test dataset. Metrics like accuracy, precision, recall, and F1-score can be used to assess the effectiveness of the model. Keras provides functionalities for model validation when working with deep learning models. To improve results, the model should be iterated and refined based on its performance.

6. Visualisation and Reporting:

Tool: Tableau and Power BI

- Employ Tableau or Power BI to create visually engaging presentations of the data insights extracted from the database. Develop dashboards that highlight the trends, patterns, and essential metrics of the smallholder community. APHA use tableau for the livestock surveillance dashboards, as shown in figure 59, thus mirroring this may be the best option.

7. Continuous Monitoring and Updating:

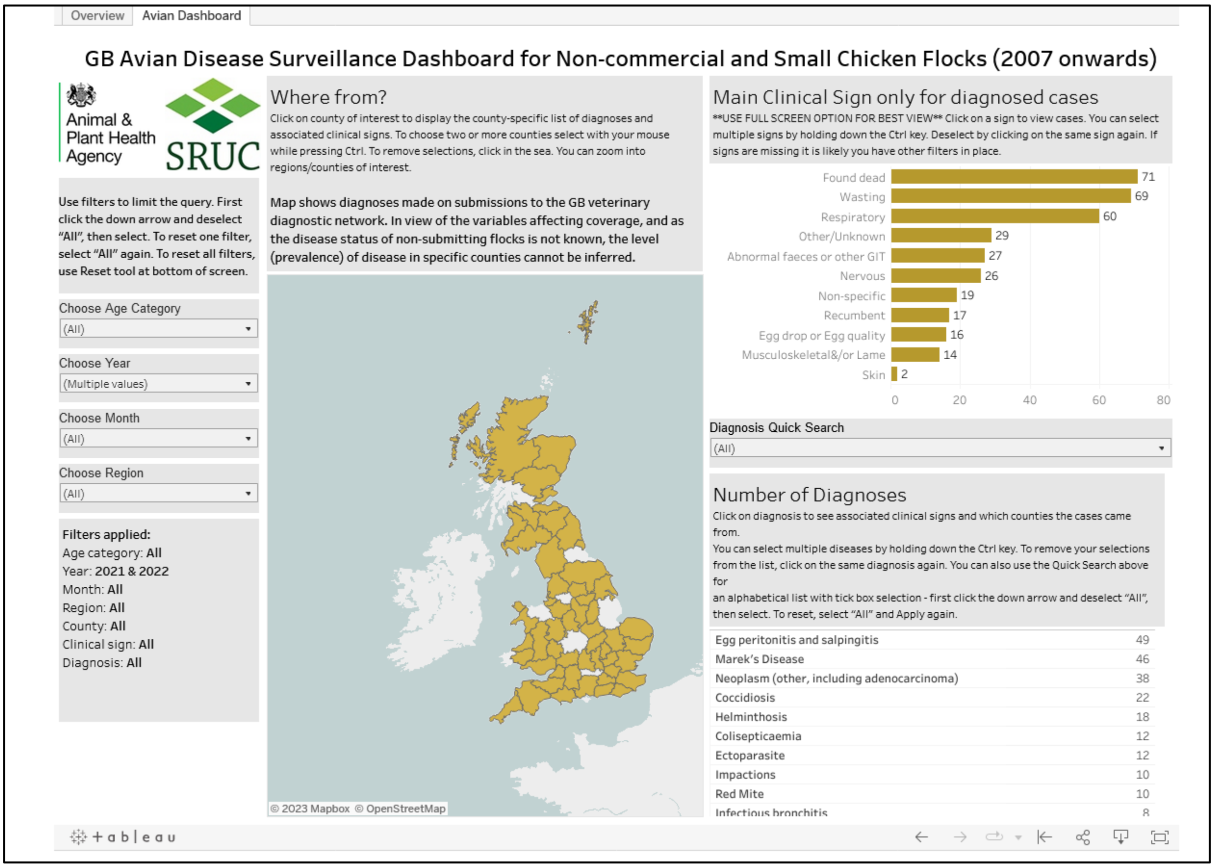
Tool: Apache Airflow

- With Apache Airflow, the tasks of scheduling and monitoring workflows becomes easier, to ensure that the pipeline stays up to date with new data, while also re-training the model and updating the database and dashboards in real-time.

8. Feedback Loop:

Tool: Custom Web Interface

- Create a personalised web platform that allows stakeholders to share their opinions on the database findings. Using this feedback, the model and pipeline can be improved even further.



GB Avian Disease Surveillance Dashboard for Non-commercial and Small Chicken Flocks (2007 onwards)

Use filters to limit the query. First click the down arrow and deselect "All", then select. To reset one filter, select "All" again. To reset all filters, use Reset tool at bottom of screen.

Choose Age Category
(All)

Choose Year
(Multiple values)

Choose Month
(All)

Choose Region
(All)

Filters applied:
Age category: **All**
Year: **2021 & 2022**
Month: **All**
Region: **All**
County: **All**
Clinical sign: **All**
Diagnosis: **All**

Where from?

Click on county of interest to display the county-specific list of diagnoses and associated clinical signs. To choose two or more counties select with your mouse while pressing Ctrl. To remove selections, click in the sea. You can zoom into regions/counties of interest.

Map shows diagnoses made on submissions to the GB veterinary diagnostic network. In view of the variables affecting coverage, and as the disease status of non-submitting flocks is not known, the level (prevalence) of disease in specific counties cannot be inferred.

Main Clinical Sign only for diagnosed cases

USE FULL SCREEN OPTION FOR BEST VIEW Click on a sign to view cases. You can select multiple signs by holding down the Ctrl key. Deselect by clicking on the same sign again. If signs are missing it is likely you have other filters in place.

Diagnosis Quick Search
(All)

Number of Diagnoses

Click on diagnosis to see associated clinical signs and which counties the cases came from.
You can select multiple diseases by holding down the Ctrl key. To remove your selections from the list, click on the same diagnosis again. You can also use the Quick Search above for
an alphabetical list with tick box selection - first click the down arrow and deselect "All", then select. To reset, select "All" and Apply again.

Egg peritonitis and salpingitis	49
Marek's Disease	46
Neoplasm (other, including adenocarcinoma)	38
Coccidiosis	22
Helminthosis	18
Colisepticaemia	12
Ectoparasite	12
Impactions	10
Red Mite	10
Infectious bronchitis	8

FIGURE 59: APHA AVIAN SURVEILLANCE DASHBOARD

9.4.1.1 Practicality and approach

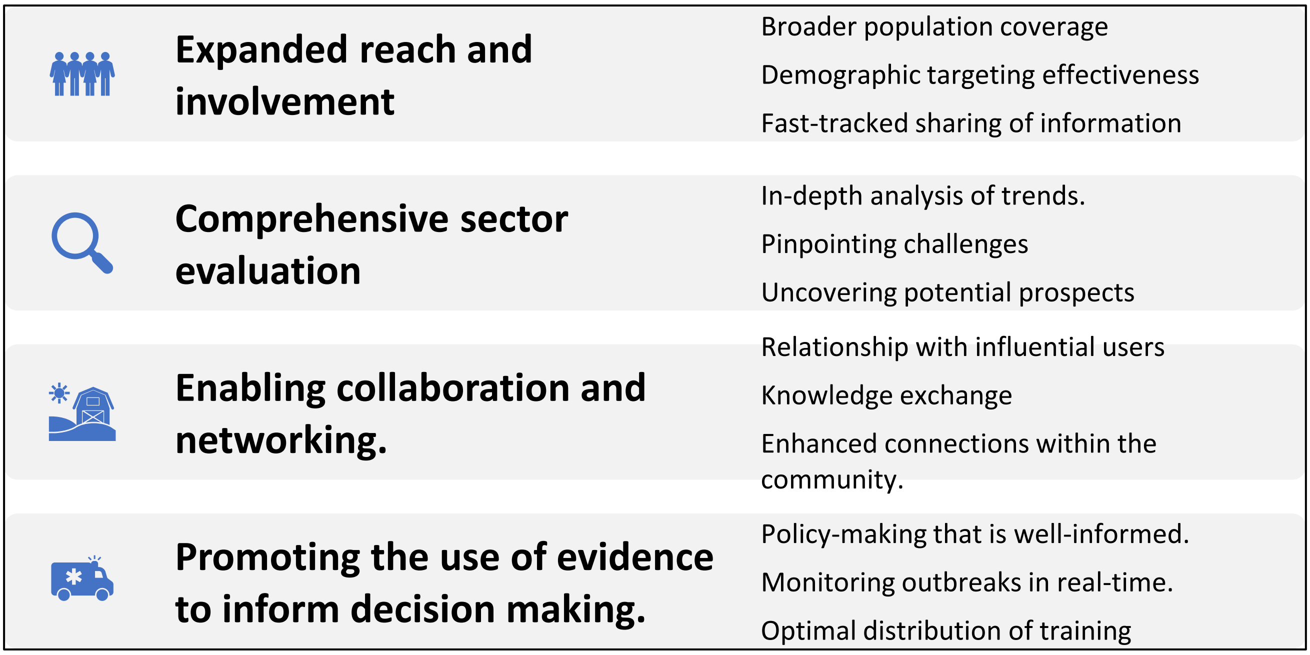
Developing a database like this requires a deep understanding of both machine learning and the relevant livestock domain. The success of this endeavour is inextricably linked to the pre-trained predictive model from Chapter 4. Twitter's dynamic content and vast user base provide a wealth of information on practices, concerns, and interactions. The study categorised Twitter accounts into smallholder farmers and non-farmers, simplifying the extensive Twitter data into a precise dataset.

9.4.1.2 Real-time application

The database creation process stands out due to its dynamic nature, which allows for flexibility and adaptability. The predictive model is a dynamic tool that adapts to changing data. Communities are constantly changing, and the database is updated accordingly to stay current. Stakeholders can always stay up to date with real-time insights thanks to this dynamic aspect.

9.4.1.3 Potential benefits

Such a database offers a multitude of possible benefits. The existing work conducted by Correia-Gomes and Sparks (2020) may be expanded on a grander scale given the knowledge accrued from this database, as they targeted poultry smallholders with a questionnaire pertaining to their livestock, through a mixture of social media, network courses about poultry welfare and via leaflets[6]. They only received 176 responses which met their selection criteria, therefore the opportunity to distribute such questionnaires via this constructed smallholder database could prove to be far more effective, in regard to reaching a larger population, reaching the correct demographics and expedite the information sharing process. With the rich dataset at their disposal, researchers can delve into the sector's trends, challenges, and opportunities. With the database, the smallholder community can establish connections that can help them collaborate and exchange ideas, further strengthening their networks. A breakdown of the potential benefits can be seen in figure 60:



Expanded reach and involvement

Broader population coverage

Demographic targeting effectiveness

Fast-tracked sharing of information

Comprehensive sector evaluation

In-depth analysis of trends.

Pinpointing challenges

Uncovering potential prospects

Enabling collaboration and networking.

Relationship with influential users

Knowledge exchange

Enhanced connections within the community.

Promoting the use of evidence to inform decision making.

Policy-making that is well-informed.

Monitoring outbreaks in real-time.

Optimal distribution of training

FIGURE 60: POTENTIAL BENEFITS OF SMALLHOLDER DATABASE

9.4.1.4 Potential issues

Nevertheless, creating and maintaining this database presents some challenges. Due to the vastness of Twitter data, irrelevant or misleading information can easily slip in and create noise. Although the predictive model is strong, there is always the possibility of error. Classifications must be validated and refined on a regular basis to ensure their accuracy and relevance. Additionally, questions of data privacy and consent become paramount ethical concerns. The over-arching limitations of such a task is discussed in greater detail in the upcoming sections. Figure 61 captures the aforementioned limitations throughout the results discussion chapters.

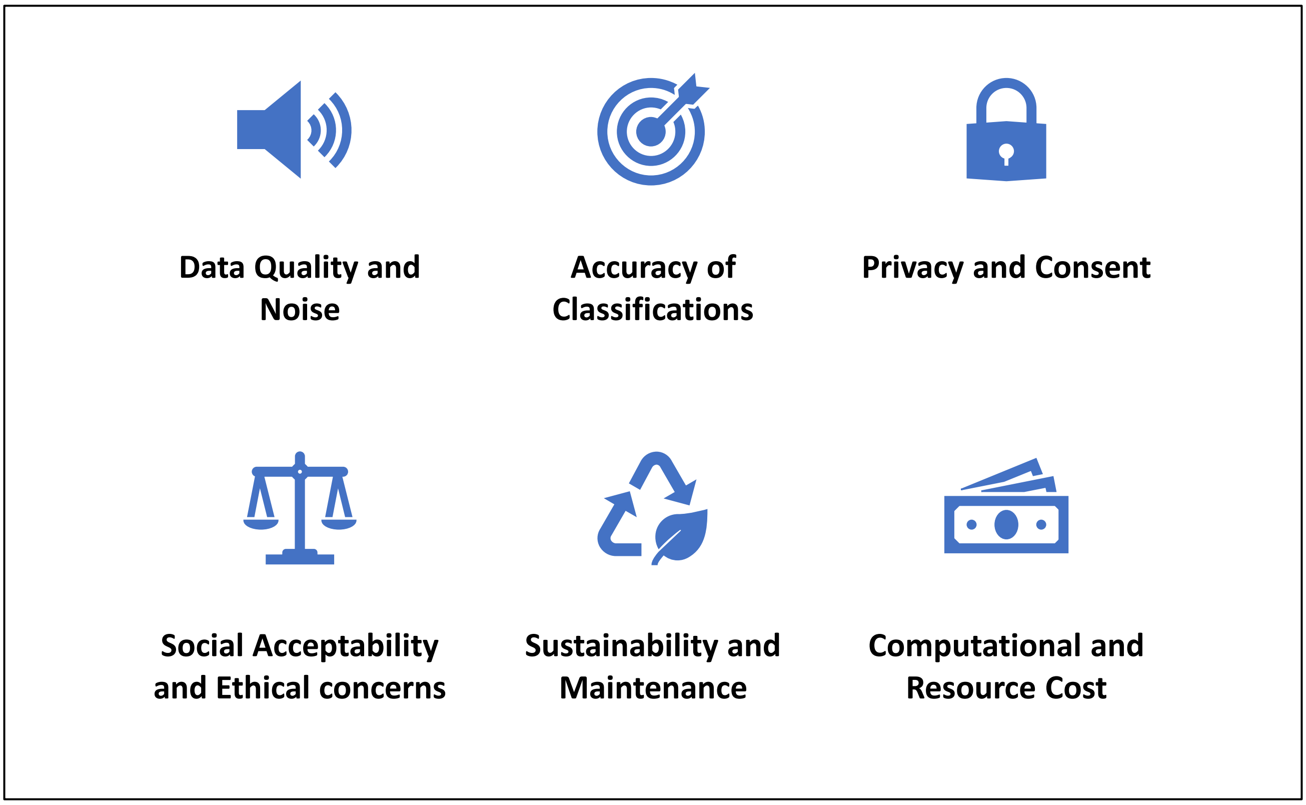

Data Quality and Noise
Accuracy of Classifications
Privacy and Consent
Social Acceptability and Ethical concerns
Sustainability and Maintenance
Computational and Resource Cost

FIGURE 61: POTENTIAL LIMITATIONS OF SMALLHOLDER DATABASE

9.4.2 Dynamic pipeline for passive intelligence

In contrast to the previous section, this pipeline will employ multiple unsupervised machine learning methods, though topic modelling and community detection. Online discussions reveal the dynamic nature of smallholder activities, highlighting the need for a real-time data pipeline. The pipeline will provide real-time insights on relevant discussions in the UK by applying the methodologies and insights gained from Chapters 5 and 6. It will focus specifically on Twitter discussions and contributions from the www.accidentalsmallholder.net forum.

The following steps are the framework for creating such a pipeline:

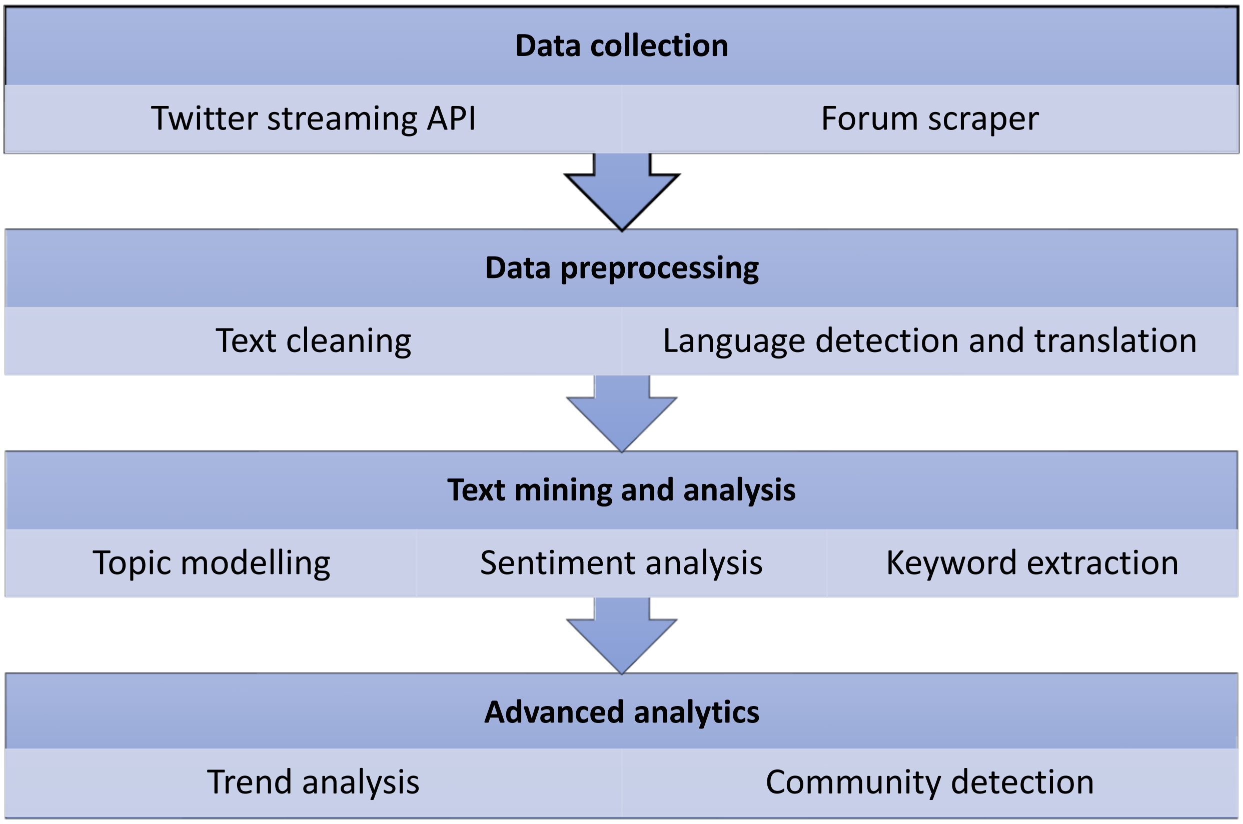


FIGURE 62: DYNAMIC PIPELINE FOR PASSIVE INTELLIGENCE

9.4.2.1 Data collection

Twitter Streaming API:

- Continuously fetch tweets from identified smallholder accounts using the Twitter Streaming API. To deal with the overwhelming amount of data on Twitter, apply filters to capture only tweets relevant to smallholder activities, using keywords, hashtags, or mentions as indicators. This pre-processing stage has been laid out in all of the results chapters, to only capture the information pertaining to smallholders, pigs, poultry and Avian influenza, depending on the objective of the chapter.

Forum Scraper:

- Use the web scraper outlined in chapter 6 to consistently monitor the forum and extract new discussion threads and posts, ensuring they comply with the platform's terms of service. Time-out requests must be enforced. Additional subforums can be accessed if the URL link is changed within the code.

9.4.2.2 Data preprocessing

Text Cleaning:

- Noise is a concern when dealing with raw data from both sources. Before processing, it's necessary to clean up the data by removing URLs, special characters, and irrelevant emojis. The text can be cleansed and standardised by using regular expressions and NLP libraries, and the custom stop words list provided in the results should be adopted.

Language Detection and Translation:

- Twitter's worldwide audience means that tweets may be written in a variety of languages. Language detection should be implemented to identify non-English content, which can then be translated to English for consistent analysis. The “lang = eng” filter was applied in chapter 7, which retains only those tweets written in English.

9.4.2.3 Analysis

Topic Modelling:

- By implementing the LDA technique as in chapters 6 and 7, the primary topics within the discussions can be identified. This will provide insights into current trends, concerns, or innovations among smallholders.

Sentiment Analysis:

- By using NLP models as in chapter 8, it's possible to assess the sentiment of discussions and understand the community's mood or reactions to specific events.

Keyword Extraction:

- Quickly identify discussion focal points by extracting and ranking frequent terms or phrases. This can be paired with bigram analysis and word clouds to be included in the dashboard.

9.4.2.4 Advanced analytics

Trend Analysis:

- As employed in chapters 5 and 7, observing the development of specific themes or moods over time can be beneficial for additional insights. Such trends assist in understanding the effects of external events, such as COVID-19 on Avian influenza search results.

Community Detection:

- Graph-based algorithms, as in the Louvain's algorithm in chapter 5, can be applied to Twitter data to detect user communities that are interconnected, which can help identify potential influencers or niche interest groups.

9.4.2.5 Visualisation and reporting

Interactive Dashboard:

- Similar to the previous section, a dynamic dashboard using platforms such as Tableau or PowerBI can be incorporated to display the insights that have been extracted in a visually appealing manner. There are several features, such as a sentiment timeline, prevalent topics word cloud, and influential users' network visualisation.

Alert Mechanism:

- This may relate to tracking a certain topic over time related to pig containment, and once the frequency of the discussion pertaining to this topic exceeds a particular frequency, this triggers an alert.

9.4.2.6 Scalability and maintenance

- The scalability of the pipeline should be carefully considered before extending it to other platforms or regions. Regular updates to models and algorithms ensure that they remain relevant.
- Continuously monitoring the pipeline for any potential issues with data acquisition or processing to maintain reliable intelligence.

9.4.3 Early warning signal system

Research reinforces that early warning signal systems have the power to transform outbreak prediction and detection[387]. Chapter 7 underscores the importance of detecting anomalies as the fundamental element of this early warning mechanism. The tool's primary aim is to detect potential avian or swine flu outbreaks as early as possible by employing anomalies in spatial and temporal datasets, which may aid in effective containment and management, but may be configured for any other disease given the right keyword filters and preliminary text cleaning.

Including the dualistic approach of both wild and domestic bird data in Avian flu veterinary surveillance can lead to a more thorough approach. Regular testing and active surveillance of wild bird populations, especially migratory waterfowl, can serve as an early warning system for possible outbreaks. By combining this approach with a comprehensive social media system for detecting unusual frequencies of deaths, surveillance efforts can be further enhanced. The integration of data on wild bird movements, population densities, and disease prevalence with domestic bird population data is a crucial step in understanding how diseases may spread between bird populations. Geospatial analysis can be used to create maps of outbreaks, pinpointing areas that are at high risk and guiding surveillance measures.

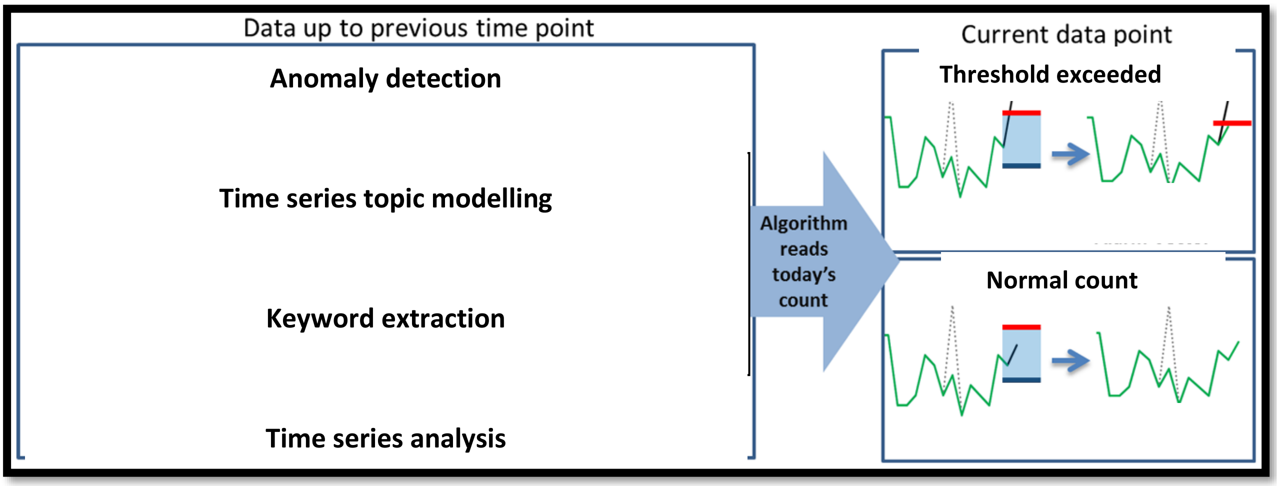


FIGURE 63: EARLY WARNING SIGNAL SYSTEM PROTOTYPE

9.4.3.1 Application of Anomaly Detection

The "Forum Scraping" chapter detailed a methodology that was successful in identifying differences in online discussions. This approach allows for real-time monitoring of discussions, enabling the detection of sudden spikes in discussion frequencies or irregular patterns. The detection of an anomaly triggers a cross-referencing process with the trained models to determine its significance. Thresholds need to be enforced in order to measure the fluctuations in message counts and may require significant tuning and experimentation to ascertain a sufficient figure. These may also vary from species to species depending on the sub-forum in question.

9.4.3.2 Keyword-Driven Monitoring

The system is designed to continuously monitor and extract specific keywords and phrases from discussions using the methodologies described in chapters 5 and 6. By priming the system with avian and pig disease-related terms, it can scan for any anomalies or patterns in their mentions, providing an initial detection layer.

9.4.3.3 Geographical Contextualisation

By merging the keyword-driven insights with spatio-temporal techniques from chapter 7, the system can pinpoint the geographic context of these mentions. This technology can be incredibly useful in detecting sudden increases in the conversation about "avian flu" symptoms in a particular region, allowing the local health and agricultural authorities to take quick action. It analyses the development of these discussions, identifying whether the concerns escalate or remain steady over time. By understanding the potential trajectory of an outbreak, this capability can help prevent further spread.

9.4.3.4 Temporal Analysis

With the developed time series analysis methodologies, the system's capabilities are greatly enhanced. By tracking the temporal progression of these discussions, it can identify whether there's a sudden escalation or a prolonged concern over time, providing insights into the potential trajectory of an outbreak. In addition, Chapter 7 outlines that the system uses analysis techniques that consider both space and time concurrently, providing contextual information about location and time.

9.4.3.5 Data validation

As explained in chapter 7, an outstanding feature is the capability to compare online discussion trends with confirmed disease outbreaks. By cross-referencing, the system adds a layer of verification that ensures the alerts generated are supported by concrete data, ultimately increasing reliability. Validating online discussions can be done by comparing the trends with confirmed cases of diseases, if available. The system's reliability is improved by ensuring that the alerts generated are based on real-world data. Confirmed cases both APHA and DEFRA would need to be scraped in real-time and act as an additional data source which is refreshed on the same time scale as the Twitter and forum data.

9.4.3.6 Real-time alerts

Real-time data integration is a major feature, enabling the constant ingestion, processing, and analysis of data from sources such as Twitter and forums. The system processes the incoming data and actively works to comprehend any irregularities. By applying techniques from Chapters 6 and 7, it is possible to identify emerging topics, track temporal trends, and highlight significant terms. This ensures alerts are contextually enriched, facilitating a comprehensive understanding of potential threats.

9.4.3.7 Summary

In conclusion, a robust alert mechanism framework has been presented, considering the thresholds identified in the research. If discussions exceed these limits, relevant parties are alerted through real-time notifications to ensure a quick response.

With the help of the research methodologies and insights, this early warning system could become a highly effective tool in current livestock and smallholder surveillance. With its focus on dynamic data, geographical and temporal analytics, and keyword-driven monitoring, this approach aims to prevent large-scale outbreaks of avian and porcine diseases.

Nevertheless, it is vital to acknowledge some important precautions to moderate the optimism surrounding the adoption of such a system. Firstly, it is crucial to acknowledge the limitations and potential biases of using Twitter data as a broad and timely dataset. Twitter does not provide a complete picture of the situation due to the absence of some crucial tech-averse smallholders and stakeholders. Moreover, the existence of false information or unrelated conversations can generate interference that makes it difficult to accurately detect real dangers – This is discussed in further detail in section 9.6.

Secondly, there is need to thoroughly evaluate how tweets and actual disease outbreaks are connected. In Chapter 7, the connection between social media activity and AVIN outbreaks was explored, with the results illuminating that an increase in tweeting about symptoms or outbreaks occurs after the first wave of outbreaks. Therefore, although social media can indicate emerging issues, it may not always give the earliest warning and consequently curtails the large-scale implementation of early warning system tools, as shown in figure 63.

Furthermore, depending on social media data brings up doubts regarding the reliability and authenticity of the information being employed. Critical decisions may not rely on unverified social media data, according to governments and public health authorities. Although the research shows the potential of these tools, it is crucial to combine them with traditional APHA surveillance methods and expert assessments for a comprehensive approach to disease monitoring and response.

9.5 Realistic adoption of social media surveillance

An important statement which must be emphasised by this project is surrounding the reality of such a tool being implemented by the authorities. The UK government is highly unlikely to switch to unverified social media surveillance for public health monitoring because accuracy, reliability, and legal compliance are crucial. Rigorous testing and standardisation of established surveillance frameworks ensure strong data collection and analysis, surpassing the capabilities of social media data alone[67]. However, social media monitoring can serve as a useful additional resource, offering immediate insights and early alerts that can strengthen conventional approaches. Nevertheless, the results from this project underline that extensive human validation is necessary to verify the accuracy and relevance of social media data, ensuring its effective integration into broader surveillance strategies.

The use of social media by authorities has been found to have multiple benefits, including increased transparency, participation, and collaboration, which aligns with the framework set out by the UK Open Government Initiative[405].

Nonetheless, there are notable obstacles to implementing it. The barriers can be divided into internal factors (i.e., leadership, management, funding) and external factors (i.e. communication infrastructure, ICT capabilities, legal and political influences). This classification is consistent with the Technology-organisation-environment Framework (TOE) framework[406].

9.5.1 Implementation prospects and barriers

Whilst the potential factors and barriers to mass adoption have been highlighted in previous section, various government initiatives have seamlessly integrated digital surveillance. The Centre for Disease Control (CDC) in the USA has previously utilised social media data, specifically Twitter, to track flu outbreaks alongside traditional surveillance techniques[17]. In a similar fashion, Singapore's Health Promotion Board relies on social media analytics to promptly handle potential health crises and monitor public health trends[407]. Both of these case studies illuminated that small-scale adoption through pilot studies and surveys had to be undertaken first, followed by gradually building the computational resource and staff to create such surveillance outlets on a larger scale.

It is essential to comprehend the veterinary epidemiology landscape in the UK, with support from DEFRA and APHA. Disease monitoring and control across different farming operations are coordinated by these organisations. The comparison of holding locations in chapter 6 could potentially act as another piece of supplementary information that APHA to publish alongside their annual reports.

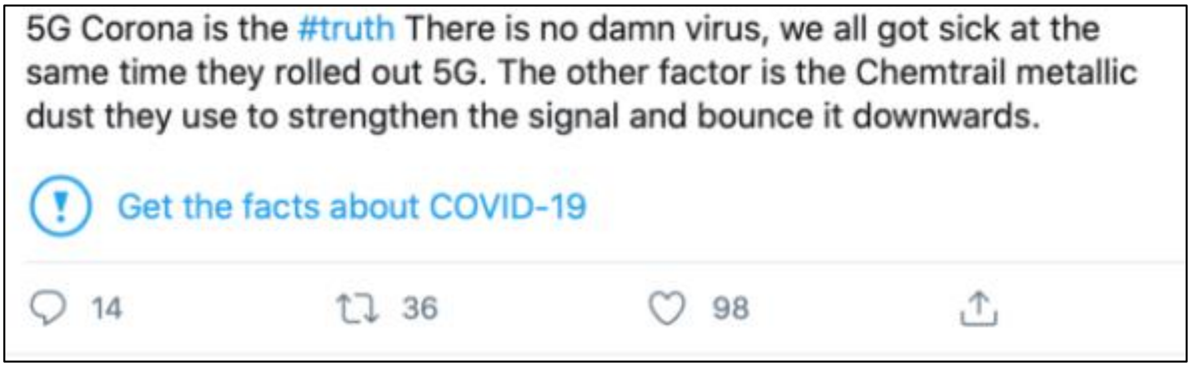
A synthesis of current literature underscores that the successful adoption of new surveillance methods by UK authorities requires the following:

1. Validation and standardisation are essential for maintaining consistency and comparability with established protocol – I.e. Location measured by co-ordinates/postcode/area compared with the free-text locations within social media profiles. Standardisation of such discrepancies must be addressed.
2. Follow both UK and EU regulations, including GDPR and Data Protection Act 2018, which cover areas such as animal welfare, data protection, and biosecurity.
3. Seamlessly integrate into current APHA systems, improving without causing disruption. Figure 58 cold have an additional tab which displays the extracted social media pig/poultry holdings from the same time period and overlays this.
4. Provide an affordable option that can replace or enhance existing approaches. Budget restrictions and time constrains of feasibility studies have to be considered and factored into any potential implementation.

Once more, effective surveillance relies on the cooperation of both authorities and the public, particularly small-scale farmers, as demonstrated by the successful implementation in Singapore. The key to building a reciprocal relationship lies in transparent surveillance practices, educating and involving smallholders, and setting up feedback mechanisms for collaborative efforts.

9.6 Misinformation and digital pitfalls in social media data

With the rise of the digital age, researchers face significant challenges due to the potential for misinformation, fake news, and the activities of internet trolls and bots on social media platforms[408]. The significant challenge of online communities lies in the potential distortion of narrative due to bots, trolls, and superfluous users/information. These entities can affect the reliability of data by manipulating discussions, spreading misinformation, and creating confusion on these platforms[409]. These challenges were present in the research findings of this project, particularly the noise element in user profiles pertaining to the user classification chapter.



5G Corona is the #truth There is no damn virus, we all got sick at the same time they rolled out 5G. The other factor is the Chemtrail metallic dust they use to strengthen the signal and bounce it downwards.

FIGURE 64: EXAMPLE OF COVID-19 MISINFORMATION TWEET

The spread of misinformation, whether intentional or not, can impact perceptions, introduce biases, and undermine the credibility of research, as highlighted in figure 64[410]. Bots and internet trolls, designed for specific purposes or to cause disruptions, can add unnecessary information to datasets, which can skew true human interactions and emotions. Furthermore, echo chambers can exist within online communities, whereby users are exposed primarily to information that supports their existing beliefs, reinforcing their views and decreasing exposure to diverse perspectives[411].

9.6.1 Mitigating misinformation

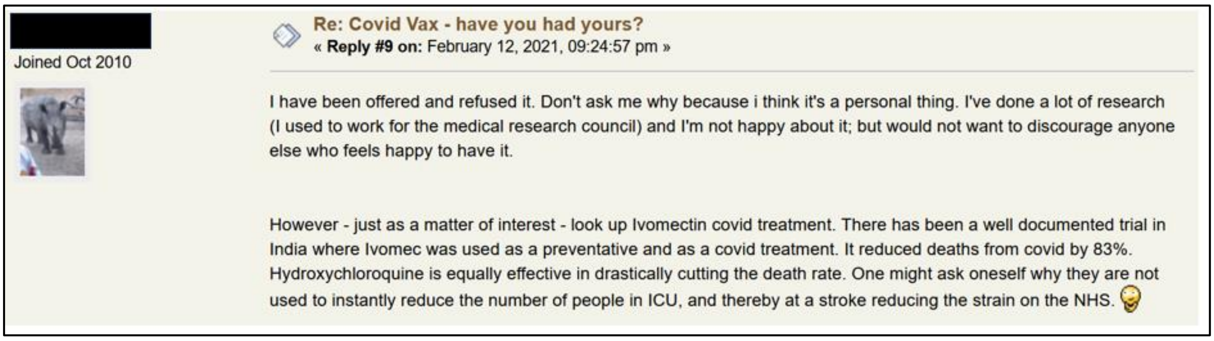
Despite this, there is a silver lining in the specific context of this research. Due to the specialised discussions and close-knit dynamics of smallholder farmers in the UK, misinformation and bot interference face limited appeal and opportunity. In-depth conversations about livestock management or farming practices are less likely to be initiated by trolls or automated bots. The community's nuanced conversations and interactions act as a safeguard against insincere or automated accounts trying to infiltrate.

To tackle some of the misinformation problems, methods such as content moderation, user verification, and algorithmic filtering can be employed[412]. Within this research, data preprocessing techniques were used to carefully identify relevant data sources, such as probing key smallholder accounts and specific forums. A strong analytical framework distinguished the importance and genuineness of information, reducing any unnecessary content through keyword filtering and user posting frequency statistics. The integration of these strategies sought to mitigate potential distortions in the online narrative.

In order to minimise digital risks, the research focuses on authentic and relevant conversations by targeting particular keywords, interests, and engagements linked to smallholding activities. The emphasis was on important discussions, which were made possible through the use of techniques like text mining and user classification, resulting in the search being focused on genuine conversations that revolved around livestock-related topics.

9.6.2 Barriers to successful mitigation

Even with optimistic mitigation strategies, the presence of anti-vax sentiments in the smallholder community remains a significant challenge that must be addressed[413]. Such beliefs can significantly hinder the ability of social media surveillance to effectively monitor diseases, as stated by Donadeu et al (2019). Anti-vax sentiments can be attributed to a lack of trust in conventional medicine and government authorities, a preference for anecdotal evidence, and the reinforcement of these views within echo chamber, as depicted in figure 65. The effects of this have been significantly increased since the inception of the COVID-19 pandemic, resulting in large anti-establishment and anti-science movements online[414].



[redacted]
Joined Oct 2010

Re: Covid Vax - have you had yours?
« **Reply #9 on:** February 12, 2021, 09:24:57 pm »

I have been offered and refused it. Don't ask me why because i think it's a personal thing. I've done a lot of research (I used to work for the medical research council) and I'm not happy about it; but would not want to discourage anyone else who feels happy to have it.

However - just as a matter of interest - look up Ivomectin covid treatment. There has been a well documented trial in India where Ivomec was used as a preventative and as a covid treatment. It reduced deaths from covid by 83%. Hydroxychloroquine is equally effective in drastically cutting the death rate. One might ask oneself why they are not used to instantly reduce the number of people in ICU, and thereby at a stroke reducing the strain on the NHS.

FIGURE 65: VACCINE HESITANCY POST FROM SMALLHOLDER FORUM

These sentiments have a deep impact on social media surveillance. Misinformation about animal health practices and vaccine efficacy can be spread through anti-vax beliefs, which may affect the accuracy of data collected via social media, making it considerably more difficult to implement the tools and recommendations set out in Section 9.4.

In order to address this, it's vital for authorities to adopt a cautious and empathetic stance. It is crucial to engage with the smallholder community through open and respectful conversation and focus on the mutual benefits aspects highlighted in figure 56. By informing and correcting misconceptions, some element trust can be established through vaccine education.

Moreover, to counteract anti-vax sentiments, it is important to collaborate with trusted community figures and smallholder associations to spread accurate information, hence making the influential user identification element of this project so crucial. Robust data collection and analysis methods are essential to filter out misinformation and prioritize verified sources, ensuring the integrity of the surveillance system. By addressing these challenges head-on, the research can better navigate the complexities of digital misinformation in the smallholder community.

Although the study acknowledges these pitfalls, it emphasises the exceptional characteristics and methodologies that minimise these risks. This reinforces that within the field of digital surveillance research; context, rigorous methodologies, and continual reflection are crucial.

9.6 Limitations

To contextualise the findings, guide future studies, and establish grounded conclusions, it's essential to understand the limitations of this research project. The subsequent part of the study explains the limitations that are part of understanding demographics, interactions, and networks through social media data.

9.6.1 Scope and population representation

The study's main focus was smallholder farmers in the UK who use Twitter and particular forums. Due to this emphasis, a considerable portion of farmers who prefer conventional communication or have low digital involvement were inadvertently excluded. A partial or biased understanding of the community's concerns, behaviours, and practices may result from this. Additionally, the sampling technique chosen to represent the population may lead to potential sampling bias.

9.6.2 Methodological constraints

- Text-based Limitations: Platforms such as Twitter, with a character limit, might oversimplify intricate issues, according to the study. An obstacle in online forums was separating authentic insights from opinions.
- Data-related Issues: Factors such as the preference of platforms affecting the demographic skewing, changes in platforms influencing user behaviour, and the abundance of data presenting opportunities and challenges were observed.
- Models and Tools: The study's unique needs might not have been fully met as the pre-trained models and specific data extraction tools failed to capture some nuances. Also, the analysis of community detection on the network gives helpful insights on the relationships

and structure, but it has its limitations. The two community detection models require a more comprehensive parameter hyper tuning implementation. This needs to be coupled with a strong custom stop word dictionary to eliminate most of the noise present in the text input.

9.6.3 Geolocation challenges

A noteworthy point that pertains to RQ5 is that extracting precise geolocation data is challenging when dealing with social media data, particularly from platforms such as Twitter. Free-text descriptions of user locations are frequently ambiguous, misleading, or whimsical. There are multiple challenges in converting these textual descriptions into accurate geographical coordinates.

- **Ambiguity**: Multiple places can be referred to by many location descriptions. For example, a location described as "Perth" could refer to Perth in Scotland or Australia, and this becomes very difficult to decipher. This was witnessed numerous times in the dataset from chapter 7.
- **Accuracy**: The user's given location might not accurately identify their exact geographic position. "London" may be mentioned by a user, even if they're situated in a suburb or separate neighbourhood within the city. London was the most popular tweeting location for the chapter 7, which further illuminates this point.
- **Missing Data**: As the Twitter geolocation option was disabled since the data mining scandals of the past 5 years, missing data becomes an ever-present issue. Not all users provide location data, and those that do, might provide broader regions (e.g., "South UK") rather than specific towns or cities.
- **Fictitious or Non-Standard Entries:** Some users, either for privacy reasons or creative expression, might list fictional places, jokes, or abstract concepts as their locations, e.g., "Middle Earth", "Everywhere and Nowhere", “Gaia”, “God’s green earth” are all examples which appeared prior to the location filtering.

Although focusing on the UK allowed for a specific understanding in that region, it also presented potential limitations:

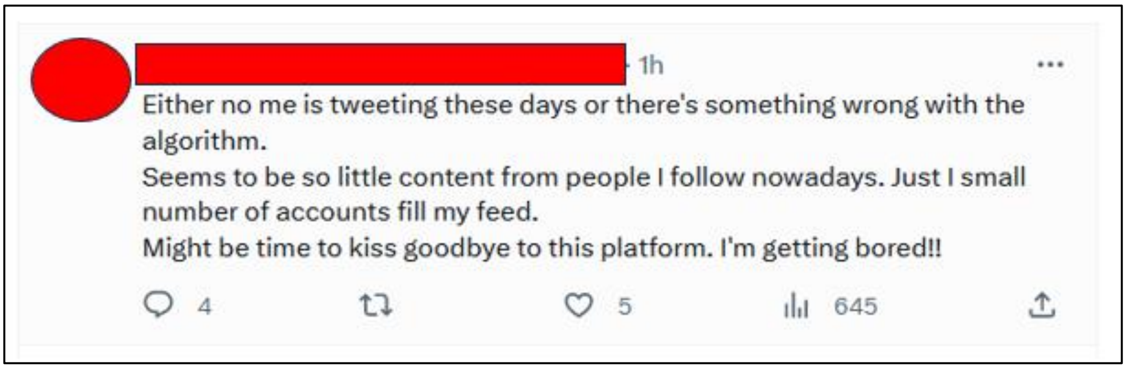
- **Cultural Specificity:** The study's exclusive focus on the UK might cause it to overlook important global trends, practices, or challenges that could be useful and informative.
- **Comparative Insights:** Limiting the analysis to the UK could hinder the ability to draw comparative insights with other regions or countries because of differences in practices, challenges, or digital behaviours.

- **Regional Variations within the UK**: Regional variations in farming practices, challenges, and digital adoption exist within the diverse UK itself. Sometimes a broad focus can miss the nuances within a region.

9.6.4 Data sources

The research failed to consider the wealth of information available on platforms like blogs, YouTube channels, and e-commerce sites. These platforms had the opportunity to present a more intricate depiction of digital interactions and concerns. Focusing on specific platforms may have hidden significant insights from platforms such as Instagram or LinkedIn. Changes in ownership of social media sites have led to limitations on API requests, making it uncertain whether studies like this can be replicated in the future. However, the emergence of newer platforms, such as “Threads”[415], could curtail such uncertainties and adds another stream of potential data within the Infoveillance field.

The disillusionment of users on certain platforms is a significant concern, particularly because of algorithm bias favouring particular political or social perspectives. Users may flee to other platforms or depart from social media entirely, should this issue persist. A tweet from a smallholding account has been provided below as an example of capturing this particular sentiment.



[redacted] [redacted] · 1h

Either no me is tweeting these days or there's something wrong with the algorithm.
Seems to be so little content from people I follow nowadays. Just I small number of accounts fill my feed.
Might be time to kiss goodbye to this platform. I'm getting bored!!

FIGURE 66: TWEET SNAPSHOT OF NEGATIVE USER SENTIMENT

9.6.5 Theoretical and practical issues

The application of theoretical frameworks to this research may have overlooked the intricacies of smallholder community online behaviour, despite providing structure. Due to the lack of research in this field, validating theories is difficult. Moreover, the conversion of theoretical models into practical tools and applications is acknowledged as a difficulty that requires additional validation, particularly if the outcomes from this project are implemented dynamically. Furthermore, the community may object to being included in surveillance systems or targeted interventions.

Another issue which is often overlooked in textual analysis of internet data are the difficulties with language interpretation pertaining to sarcasm, nuance, hyperbole and slang. Smallholding communities may adopt their own vernacular and dialect within online conversations, which can cause a problem for the unsupervised machine learning models. Furthermore, as this research focusses on the UK, it has not considered additional languages such as Welsh and Gaelic, which may contain an even greater level of insights. Finally, translation tools may result in the loss of important details in discussions.

9.6.6 Spatial, temporal and exogenous factors

The research recognised that spatio-temporal findings don't imply causation. Additionally, worldwide occurrences like health crises or policy reforms could have impacted online narratives and behaviours during the research period, just as in the case of COVID-19 occurring during the study period of this research.

9.6.7 Ethical considerations

The methodology section elaborates on how the usage of publicly available online data posed ethical issues such as privacy, consent, and potential misinterpretation. Navigating potential bias from echo chambers, users omitting failures in their online presentation, and platform-specific nuances was necessary. Lastly, the exponential growth of AI is leading to an abundance of big data that requires ongoing discussions about informed consent and ethical participation to achieve a universal ethical standard.

9.7 Further research

The comprehensive findings of this research offer new avenues for exploration and application. Based on the current outcomes, the proposed directions for future investigations are as follows:

1. Extending the Reach of Digital Platforms:

This study was primarily concentrated on Twitter and specific forums, but in the future, other platforms, including Instagram, LinkedIn, or even Pinterest, could be considered. By way of example, Instagram offers practical farming insights with its visual-centric approach, while LinkedIn illuminates the professional and entrepreneurial side of small-scale farming. To enhance future research, larger and more diverse sample sizes, as well as a broader range of social network analysis methods, should be considered. Online images and videos also offer the opportunity to apply computer vision algorithms, which may add an even greater level of insights into the existing surveillance systems.

2. In-depth Analysis of Underrepresented Groups:

The current research pointed out the multifaceted nature of the smallholder community. Future studies may narrow in on these sub-sets, potentially exploring more extensively the rituals, obstacles, and goals of certain groups, like crofters or hobbyists, as implied in Chapter 4.

Whilst this research has been centred primarily on smallholder farming in the UK, it is imperative to reflect on the broader global framework surrounding this demographic. Factors such as internet access, social media platforms, and cultural practices can lead to different small-scale farming systems in developing regions such as West Africa[416]. Despite regional differences, the significant role of influential individuals in spreading precise information may be universal.

Online communities could have similar structures and relationships in various regions. It is necessary to conduct parallel studies that compare populations across different regions, including both developed and developing nations.

3. Enhanced Models for Better Classification:

Chapter 4 discusses the Twitter user classification predictive model, which could be improved through advanced deep learning techniques or more diverse datasets. The outcome may be increased accuracy in classifications and more detailed demographic analyses. In addition, this model only focussed on the user profile descriptions, thus a model which also factors in the user timeline may develop better results.

4. Influencer Engagement and Social Dynamics:

In Chapter 5, the presence of influential nodes in the network is highlighted by the scale-free topology. To study their content, reach, and impact, future research could concentrate on these influencers. Developing tactics to interact with these nodes could enhance the spread of specific information. As they have significant influence over subsections of the community, frequent engagement and communication with these individuals may serve as a more efficient method of information dissemination to the rest of the community.

To improve the understanding of these communities and their unique features, a greater depth of research is required. Additionally, evaluating the effectiveness of interventions that use the identified community structures and influential users is crucial. Incorporating the opinions and sentiments from livestock authorities such as APHA and veterinarians based on their tweets could be one possible way to achieve this. The partially completed results from chapter 8 create a foundation for such research to occur, delving into the sentiments behind government legislation.

5. Ethical and Humane Farming Practices:

The contemporary global trend towards ethical farming aligns with Chapter 6's focus on "free range". There is potential for further exploration into the economic, environmental, and social landscape in relation to small-scale practices.

6. Predictive Outbreak Systems:

Anomaly detection insights from Chapter 6 present an opportunity to create real-time surveillance systems. Utilising machine learning and AI, such systems could detect early signs of outbreaks and enable prompt interventions. Section 9.4 creates a blueprint for this to be put into practice.

7. Spatio-temporal explorations:

Geospatial analysis potential was revealed in Chapter 7. Integrating more granular data, possibly at the village or town level, could help reveal hyper-local trends and concerns in future research. Moreover, temporal patterns may be associated with other noteworthy events (e.g., policy changes, global events) to discern their impact on discussions.

8. Incorporating Feedback Loops:

Integrating feedback mechanisms is crucial as tools and models are created, such as the one proposed in section 9.4. Real-time feedback from users, particularly those from smallholder groups, can enhance the accuracy of the tool's predictions over time.

9. Bridging Online and Offline Worlds:

Digital footprints are insightful, but combining them with on-ground surveys, interviews, or focus group discussions can provide a more comprehensive understanding of the group’s issues and goals.

10. Cross-Species Analysis

The research can be extended to different regions, species, and farming practices for broader generalisability and comparative insights. In addition to the smallholder forum mentioned in chapter 6, there are numerous sub-forums dedicated to other species, all of which offer a comparable level of information to the one I used. A dashboard that is divided by species type can be a powerful tool in the future, giving real-time insights into animal discussions.

11. Collaborative Efforts:

Finally, the creation of platforms where insights from this research are shared, discussed, and acted upon could be facilitated by engaging with governmental bodies, such as APHA and DEFRA, and the commercial farming community. By collaborating, researchers can ensure that their findings have a positive impact on the daily lives of smallholders.

10/. Conclusion

The research delved deep into the workings of smallholder communities in the UK, and how they engage with digital platforms. By effectively using non-traditional data sources, particularly Twitter and a prominent livestock forum, the primary research questions were successfully addressed. With access to these platforms, one could gain a more nuanced understanding of the cultural and social factors that shape this population. By applying specific data science techniques, it was found that valuable insights could be extracted and interpreted from these platforms, providing unique perspectives that are often overlooked by traditional research methods. intervention to ensure accuracy and validity. A more integrative approach to studying this demographic was highlighted with the merging of digital data sources and traditional agricultural research.

The initial objectives and research questions were successfully achieved. By leveraging text classification, topic modelling, social network analysis, and spatio-temporal analysis, a better understanding of the conversational patterns and interests was achieved. Social media's potential as a veterinary surveillance tool was evaluated, indicating novel approaches for disease control measures. Actively promoting the use of digital technologies and social media through hosting seminars, conferences and events is an avenue which can be undertaken to foster a stronger uptake of digital literacy amongst these populations.

These findings carry significant implications. The study recommends a potential change in how traditional small-scale communities are perceived, proposing that policymakers (APHA and DEFRA), veterinarians, and agricultural experts (SRUC) should adopt a new perspective for monitoring and supporting these communities. The study found that social media could be a useful supplementary tool alongside traditional data avenues to enhance surveillance methods. However, it requires significant human intervention to ensure data quality and validity, coincided with the appropriate livestock-specific data preprocessing.

The challenges in achieving comprehensive and accurate data result from social media's vast and ever-changing nature, notably the changes in ownership of Twitter, which has resulted in modifications to data scraping regulations. A concentration on English-language platforms could have unintentionally excluded non-English speaking individuals in the UK. In addition, biases surrounding population selection, sample sizes and a focus on a specific forum all need to be considered when drawing inferences from these results.

Future research avenues were identified in light of these findings and limitations. The possibility of exploring multilingual social media platforms and integrating technological advancements, such as AI-driven sentiment analysis, might be necessary. These methodologies have the potential to be applied to other livestock sectors, broadening the scope of their implications beyond just the pig and poultry domain.

In conclusion, this research employed an interdisciplinary approach to investigate smallholder communities in the UK, analysing their discussions and interactions through the lens of social media. The significance of this research is expected to grow as the boundaries between the digital and physical realms continue to blur, guiding interdisciplinary studies in the future.

10.1 Final reflection

This thesis blends traditional and modern research methods, combining small-scale agriculture, social media integration, innovative data science approaches, and current disease surveillance issues pertaining to gaps in intelligence. It showcases how the intersection of separate disciplines can yield research that is coherent, diversified, and contemporary. In addition, it highlights the significance of interdisciplinary exploration in shedding light on niche populations in a world that is rapidly becoming more digital and interconnected. By bridging the gap between academic theory and practical application, this research has the potential to make a real difference in small-scale farming surveillance and builds a foundation for these methods to be replicated. The insights and methodologies presented here offer a significant contribution to veterinary epidemiology, non-commercial agriculture, and the broader landscape of data-driven research.

References

[1] C. Correia-Gomes, M. K. Henry, H. K. Auty, and G. J. Gunn, ‘Exploring the role of small-scale livestock keepers for national biosecurity-The pig case’, *Prev Vet Med*, vol. 145, pp. 7–15, Sep. 2017, doi: 10.1016/j.prevetmed.2017.06.005.

[2] ‘Animal welfare’, GOV.UK. Accessed: Feb. 01, 2023. [Online]. Available: https://www.gov.uk/guidance/animal-welfare

[3] APHA, ‘Poultry (including game birds): registration rules and forms’, GOV.UK. Accessed: Dec. 27, 2022. [Online]. Available: https://www.gov.uk/government/publications/poultry-including-game-birds-registration-rules-and-forms

[4] A. Amirgazin *et al.*, ‘Highly pathogenic avian influenza virus of the A/H5N8 subtype, clade 2.3.4.4b, caused outbreaks in Kazakhstan in 2020’, *PeerJ*, vol. 10, p. e13038, Mar. 2022, doi: 10.7717/peerj.13038.

[5] C. Correia-Gomes, M. K. Henry, A. Reeves, and N. Sparks, ‘Management and biosecurity practices by small to medium egg producers in Scotland’, *Br Poult Sci*, vol. 62, no. 4, pp. 499–508, Aug. 2021, doi: 10.1080/00071668.2021.1894635.

[6] C. Correia-Gomes and N. Sparks, ‘Exploring the attitudes of backyard poultry keepers to health and biosecurity’, *Preventive Veterinary Medicine*, vol. 174, p. 104812, Jan. 2020, doi: 10.1016/j.prevetmed.2019.104812.

[7] I. Andretta *et al.*, ‘Environmental Impacts of Pig and Poultry Production: Insights From a Systematic Review’, *Front. Vet. Sci.*, vol. 8, Oct. 2021, doi: 10.3389/fvets.2021.750733.

[8] A. Arvidsson, K. Fischer, E. Chenais, J. Kiguli, S. Sternberg-Lewerin, and K. Ståhl, ‘Limitations and opportunities of smallholders’ practical knowledge when dealing with pig health issues in northern Uganda’, *PLOS ONE*, vol. 18, no. 6, p. e0287041, Jun. 2023, doi: 10.1371/journal.pone.0287041.

[9] CPRE, ‘CPRE’s Vision for the future of farming: Pig and poultry farming’, CPRE. Accessed: Apr. 03, 2024. [Online]. Available: https://www.cpre.org.uk/resources/cpres-vision-for-the-future-of-farming-pig-and-poultry-farming/

[10] AHDB, ‘Disease surveillance for pig farmers | AHDB’. Accessed: Apr. 03, 2024. [Online]. Available: https://ahdb.org.uk/knowledge-library/disease-surveillance-for-pig-farmers

[11] L. V. Alarcón, A. Allepuz, and E. Mateu, ‘Biosecurity in pig farms: a review’, *Porcine Health Management*, vol. 7, no. 1, p. 5, Jan. 2021, doi: 10.1186/s40813-020-00181-z.

[12] A. E. Aiello, A. Renson, and P. N. Zivich, ‘Social Media– and Internet-Based Disease Surveillance for Public Health’, *Annual Review of Public Health*, vol. 41, no. 1, pp. 101–118, 2020, doi: 10.1146/annurev-publhealth-040119-094402.

[13] J. M. Barros, J. Duggan, and D. Rebholz-Schuhmann, ‘The Application of Internet-Based Sources for Public Health Surveillance (Infoveillance): Systematic Review’, *Journal of Medical Internet Research*, 003 2020, doi: 10.2196/13680.

[14] G. K. Sandaka and B. N. Gaekwade, ‘Sentiment Analysis and Time-series Analysis for the COVID-19 vaccine Tweets’, p. 37, 2021.

[15] A. Mavragani, ‘Infodemiology and Infoveillance: Scoping Review’, *J Med Internet Res*, vol. 22, no. 4, p. e16206, Apr. 2020, doi: 10.2196/16206.

[16] L. A. Boden, H. Auty, A. Reeves, G. Rydevik, P. Bessell, and I. J. McKendrick, ‘Animal Health Surveillance in Scotland in 2030: Using Scenario Planning to Develop Strategies in the Context of “Brexit”’, *Frontiers in Veterinary Science*, vol. 4, 2017, Accessed: Jan. 30, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fvets.2017.00201

[17] A. E. Wilson, C. U. Lehmann, S. N. Saleh, J. Hanna, and R. J. Medford, ‘Social media: A new tool for outbreak surveillance’, *Antimicrobial Stewardship & Healthcare Epidemiology*, vol. 1, no. 1, p. e50, ed 2021, doi: 10.1017/ash.2021.225.

[18] S. Kandula and J. Shaman, ‘Reappraising the utility of Google Flu Trends’, *PLOS Computational Biology*, vol. 15, no. 8, p. e1007258, Aug. 2019, doi: 10.1371/journal.pcbi.1007258.

[19] L. Hossain, D. Kam, F. Kong, R. T. Wigand, and T. Bossomaier, ‘Social media in Ebola outbreak’, *Epidemiol Infect*, vol. 144, no. 10, pp. 2136–2143, Jul. 2016, doi: 10.1017/S095026881600039X.

[20] L. Alsudias and P. Rayson, ‘Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic With Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study’, *JMIR Medical Informatics*, vol. 9, no. 9, p. e27670, Sep. 2021, doi: 10.2196/27670.

[21] C. Robertson and L. Yee, ‘Avian Influenza Risk Surveillance in North America with Online Media’, *PLOS ONE*, vol. 11, no. 11, p. e0165688, Nov. 2016, doi: 10.1371/journal.pone.0165688.

[22] J. S. Brownstein, C. C. Freifeld, B. Y. Reis, and K. D. Mandl, ‘Surveillance Sans Frontières: Internet-based emerging infectious disease intelligence and the HealthMap project’, *PLoS Med*, vol. 5, no. 7, p. e151, Jul. 2008, doi: 10.1371/journal.pmed.0050151.

[23] F. C. Dórea and C. W. Revie, ‘Data-Driven Surveillance: Effective Collection, Integration, and Interpretation of Data to Support Decision Making’, *Frontiers in Veterinary Science*, vol. 8, 2021, Accessed: Dec. 27, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fvets.2021.633977

[24] G. Eysenbach, ‘Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet’, *Journal of Medical Internet Research*, vol. 11, no. 1, p. e1157, Mar. 2009, doi: 10.2196/jmir.1157.

[25] M. Grandjean, ‘A social network analysis of Twitter: Mapping the digital humanities community’, *Cogent Arts & Humanities*, vol. 3, no. 1, p. 1171458, Dec. 2016, doi: 10.1080/23311983.2016.1171458.

[26] I. Himelboim, M. A. Smith, L. Rainie, B. Shneiderman, and C. Espina, ‘Classifying Twitter Topic-Networks Using Social Network Analysis’, *Social Media + Society*, vol. 3, no. 1, p. 2056305117691545, Jan. 2017, doi: 10.1177/2056305117691545.

[27] J.-W. Seol, K.-Y. Jeong, and K.-S. Lee, ‘Follower Classification through Social Network Analysis in Twitter’, in *Grid and Pervasive Computing*, J. J. (Jong H. Park, H. R. Arabnia, C. Kim, W. Shi, and J.-M. Gil, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013, pp. 926–931. doi: 10.1007/978-3-642-38027-3_108.

[28] F. Firdaniza, B. N. Ruchjana, D. Chaerani, and J. Radianti, ‘Information Diffusion Model in Twitter: A Systematic Literature Review’, *Information*, vol. 13, no. 1, Art. no. 1, Jan. 2022, doi: 10.3390/info13010013.

[29] A. De Salve, P. Mori, B. Guidi, L. Ricci, and R. D. Pietro, ‘Predicting Influential Users in Online Social Network Groups’, *ACM Trans. Knowl. Discov. Data*, vol. 15, no. 3, p. 35:1-35:50, Apr. 2021, doi: 10.1145/3441447.

[30] D. Barrett, ‘The Potential for Big Data in Animal Disease Surveillance in Ireland’, *Frontiers in Veterinary Science*, vol. 4, 2017, Accessed: Sep. 26, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fvets.2017.00150

[31] C. Stieg, ‘How this Canadian start-up spotted coronavirus before everyone else knew about it’, CNBC. Accessed: Feb. 01, 2023. [Online]. Available: https://www.cnbc.com/2020/03/03/bluedot-used-artificial-intelligence-to-predict-coronavirus-spread.html

[32] A. R. Daughton *et al.*, ‘Mining and Validating Social Media Data for COVID-19-Related Human Behaviors Between January and July 2020: Infodemiology Study’, *J Med Internet Res*, vol. 23, no. 5, p. e27059, May 2021, doi: 10.2196/27059.

[33] S. Amin, M. I. Uddin, D. H. alSaeed, A. Khan, and M. Adnan, ‘Early Detection of Seasonal Outbreaks from Twitter Data Using Machine Learning Approaches’, *Complexity*, vol. 2021, p. e5520366, Mar. 2021, doi: 10.1155/2021/5520366.

[34] R. M. Anholt, J. Berezowski, I. Jamal, C. Ribble, and C. Stephen, ‘Mining free-text medical records for companion animal enteric syndrome surveillance’, *Preventive Veterinary Medicine*, vol. 113, no. 4, pp. 417–422, Mar. 2014, doi: 10.1016/j.prevetmed.2014.01.017.

[35] S. Milusheva, R. Marty, G. Bedoya, S. Williams, E. Resor, and A. Legovini, ‘Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning’, *PLOS ONE*, vol. 16, no. 2, p. e0244317, Feb. 2021, doi: 10.1371/journal.pone.0244317.

[36] ‘Precision livestock farming’. Accessed: Feb. 03, 2023. [Online]. Available: https://www.sruc.ac.uk/research/research-facilities/beef-sheep-research-facility/beef-sheep-research-facilities/precision-livestock-farming/

[37] B. Contiero, G. Cozzi, L. Karpf, and F. Gottardo, ‘Pain in Pig Production: Text Mining Analysis of the Scientific Literature’, *J Agric Environ Ethics*, vol. 32, no. 3, pp. 401–412, Jun. 2019, doi: 10.1007/s10806-019-09781-4.

[38] Y.-W. Kim, S.-J. Yoon, and I.-H. Oh, ‘The economic burden of the 2009 pandemic H1N1 influenza in Korea’, *Scand J Infect Dis*, vol. 45, no. 5, pp. 390–396, May 2013, doi: 10.3109/00365548.2012.749423.

[39] S. Fan and C. Rue, ‘The Role of Smallholder Farms in a Changing World’, in *The Role of Smallholder Farms in Food and Nutrition Security*, S. Gomez y Paloma, L. Riesgo, and K. Louhichi, Eds., Cham: Springer International Publishing, 2020, pp. 13–28. doi: 10.1007/978-3-030-42148-9_2.

[40] H. Ritchie and M. Roser, ‘Farm Size and Productivity’, *Our World in Data*, Jul. 2022, Accessed: Feb. 05, 2023. [Online]. Available: https://ourworldindata.org/farm-size

[41] Addland, ‘Addlands’ Guide to Smallholdings’, Arbtech. Accessed: Jan. 04, 2023. [Online]. Available: https://arbtech.co.uk/a-guide-to-smallholdings/

[42] C. Onoselase, ‘Seventy-first Annual Report to Parliament on Local Authority Smallholdings in England (1 April 2020 – 31 March 2021)’, 2020.

[43] C. Nast, ‘The new wave of UK agritourism’, House & Garden. Accessed: Feb. 05, 2023. [Online]. Available: https://www.houseandgarden.co.uk/gallery/agritourism-in-the-uk

[44] ‘Farming Investment Fund’, GOV.UK. Accessed: Feb. 05, 2023. [Online]. Available: https://www.gov.uk/guidance/farming-investment-fund

[45] W. N. S. Zondo and J. T. Ndoro, ‘Attributes of Diffusion of Innovation’s Influence on Smallholder Farmers’ Social Media Adoption in Mpumalanga Province, South Africa’, *Sustainability*, vol. 15, no. 5, Art. no. 5, Jan. 2023, doi: 10.3390/su15054017.

[46] ‘UK pig numbers and holdings | AHDB’. Accessed: Feb. 06, 2023. [Online]. Available: https://ahdb.org.uk/pork/uk-pig-numbers-and-holdings

[47] ‘Red meat’s £1.3bn boost to the Scottish economy value to economy | Pig World’. Accessed: Feb. 05, 2023. [Online]. Available: https://www.pig-world.co.uk/news/red-meats-1-3bn-boost-to-the-scottish-economy-value-to-economy.html

[48] ‘Pig resources for farmers from Farm Advisory Service’, FAS. Accessed: Feb. 05, 2023. [Online]. Available: https://www.fas.scot/livestock/pigs/

[49] ‘71st annual report to Parliament on smallholdings in England, 1 April 2020 to 31 March 2021’, GOV.UK. Accessed: Feb. 05, 2023. [Online]. Available: https://www.gov.uk/government/publications/71st-annual-report-to-parliament-on-smallholdings-in-england/71st-annual-report-to-parliament-on-smallholdings-in-england-1-april-2020-to-31-march-2021

[50] B. J. Hoye, V. J. Munster, H. Nishiura, M. Klaassen, and R. A. M. Fouchier, ‘Surveillance of Wild Birds for Avian Influenza Virus’, *Emerg Infect Dis*, vol. 16, no. 12, pp. 1827–1834, Dec. 2010, doi: 10.3201/eid1612.100589.

[51] 'Disease prevention for livestock and poultry keepers', GOV.UK. Accessed: Feb. 05, 2023. [Online]. Available: https://www.gov.uk/guidance/disease-prevention-for-livestock-farmers

[52] DEFRA, 'Compulsory Poultry Registration Form - Keeper of 50 or More Birds'. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/981417/IRA81.pdf

[53] E. Chenais *et al.*, 'Smallholders' perceptions on biosecurity and disease control in relation to African swine fever in an endemically infected area in Northern Uganda', *BMC Veterinary Research*, vol. 15, no. 1, p. 279, Aug. 2019, doi: 10.1186/s12917-019-2005-7.

[54] C. U. Nwafor and I. C. Nwafor, 'Communication networks used by smallholder livestock farmers during disease outbreaks: Case study in the Free State, South Africa', *Open Agriculture*, vol. 7, no. 1, pp. 808–819, Jan. 2022, doi: 10.1515/opag-2022-0119.

[55] H. De Los Reyes, 'What Qualifies As a Smallholding In the UK?', 🐝 BootstrapBee.com. Accessed: Aug. 06, 2023. [Online]. Available: https://bootstrapbee.com/smallholding/what-qualifies-as-a-smallholding-in-the-uk

[56] B. Cousins, 'What is a "smallholder"? Class- analytic perspectives on small- scale farming and agrarian reform in South Africa', in *Reforming Land and Resource Use in South Africa*, Routledge, 2010.

[57] L.-A. Sutherland, C. Barlagne, and A. P. Barnes, 'Beyond "Hobby Farming": towards a typology of non-commercial farming', *Agric Hum Values*, vol. 36, no. 3, pp. 475–493, Sep. 2019, doi: 10.1007/s10460-019-09930-5.

[58] K. S. Havyas, 'Millennials and Gen Z To Revolutionize The Future of Farming', Beegle Agritech. Accessed: Aug. 06, 2023. [Online]. Available: https://www.linkedin.com/pulse/millennials-gen-z-revolutionize-future-farming-havyas-k-s

[59] M. R. A. Mamun, 'Food and the city: urban agriculture and the new food revolution', *Urban Geography*, Feb. 2014, Accessed: Aug. 06, 2023. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/02723638.2013.864477

[60] K. Runacres, 'Homesteading vs. Smallholding: Observations from Both Sides of the Pond', Homestead.org. Accessed: Aug. 06, 2023. [Online]. Available: https://www.homestead.org/lifestyle/homesteading-vs-smallholding/

[61] Crofting commission, 'What is Crofting? | Crofting Commission', crofting.scotland.gov.uk. Accessed: Aug. 06, 2023. [Online]. Available: https://www.crofting.scotland.gov.uk/what-is-crofting

[62] Nature.scot, 'Crofting', NatureScot. Accessed: Mar. 20, 2023. [Online]. Available: https://www.nature.scot/professional-advice/land-and-sea-management/managing-land/farming-and-crofting/types-farming/crofting

[63] A. Phiri, G. T. Chipeta, and W. D. Chawinga, 'Information behaviour of rural smallholder farmers in some selected developing countries: A literature review', *Information Development*, vol. 35, no. 5, pp. 831–838, Nov. 2019, doi: 10.1177/0266666918804861.

[64] 'Bovine Viral Diarrhoea Virus'. Accessed: Feb. 03, 2023. [Online]. Available: https://www.epicscotland.org/our-research/prioritising-disease-risks/disease-risks/bovine-viral-diarrhoea-virus/

[65] 'QMS Pigs Assurance Scheme | Quality Meat Scotland'. Accessed: Feb. 03, 2023. [Online]. Available: https://www.qmscotland.co.uk/pig-standards

[66] 'The Review of Veterinary Surveillance - Final Report'.

[67] UKSF, 'The UK approach to animal health surveillance', 2019. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/869173/uksf-animal-health-surveillance.pdf

[68] N. Bollig, L. Clarke, E. Elsmo, and M. Craven, 'Machine learning for syndromic surveillance using veterinary necropsy reports', *PLOS ONE*, vol. 15, no. 2, p. e0228105, Feb. 2020, doi: 10.1371/journal.pone.0228105.

[69] A. P. Barnes *et al.*, 'Influencing incentives for precision agricultural technologies within European arable farming systems', *Environmental Science & Policy*, vol. 93, pp. 66–74, Mar. 2019, doi: 10.1016/j.envsci.2018.12.014.

[70] 'Public Health Priorities for Scotland'.

[71] L. Madoff, 'Rapid reporting of emerging disease outbreaks using unofficial sources: Lessons from ProMED'.

[72] J. George, B. Häsler, E. Komba, C. Sindato, M. Rweyemamu, and J. Mlangwa, 'Towards an integrated animal health surveillance system in Tanzania: making better use of existing and potential data sources for early warning surveillance', *BMC Veterinary Research*, vol. 17, no. 1, p. 109, Mar. 2021, doi: 10.1186/s12917-021-02789-x.

[73] L. Donelle *et al.*, 'Digital technology and disease surveillance in the COVID-19 pandemic: a scoping review protocol', *BMJ Open*, vol. 11, no. 10, p. e053962, Oct. 2021, doi: 10.1136/bmjopen-2021-053962.

[74] A. Mavragani and G. Ochoa, 'Google Trends in Infodemiology and Infoveillance: Methodology Framework', *JMIR Public Health Surveill*, vol. 5, no. 2, p. e13439, May 2019, doi: 10.2196/13439.

[75] J.-P. Allem *et al.*, 'Topics of Nicotine-Related Discussions on Twitter: Infoveillance Study', *J Med Internet Res*, vol. 23, no. 6, p. e25579, Jun. 2021, doi: 10.2196/25579.

[76] C. E. Koppeschaar *et al.*, 'Influenzanet: Citizens Among 10 Countries Collaborating to Monitor Influenza in Europe', *JMIR Public Health Surveill*, vol. 3, no. 3, p. e66, Sep. 2017, doi: 10.2196/publichealth.7429.

[77] 'The associations of active and passive social media use with well-being: A critical scoping review - Patti M Valkenburg, Irene I van Driel, Ine Beyens, 2022'. Accessed: Feb. 05, 2023. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/14614448211065425

[78] C. L. Snelson, 'Qualitative and Mixed Methods Social Media Research: A Review of the Literature', *International Journal of Qualitative Methods*, vol. 15, no. 1, p. 1609406915624574, Dec. 2016, doi: 10.1177/1609406915624574.

[79] T. Mackey *et al.*, 'Machine Learning to Detect Self-Reporting of Symptoms, Testing Access, and Recovery Associated With COVID-19 on Twitter: Retrospective Big Data Infoveillance Study', *JMIR Public Health and Surveillance*, vol. 6, no. 2, p. e19509, Jun. 2020, doi: 10.2196/19509.

[80] M. Hernández-Jover, L. Hayes, R. Woodgate, L. Rast, and J.-A. L. M. L. Toribio, 'Animal Health Management Practices Among Smallholder Livestock Producers in Australia and Their Contribution to the Surveillance System', *Frontiers in Veterinary Science*, vol. 6, p. 191, 2019, doi: 10.3389/fvets.2019.00191.

[81] V. Rodríguez-Prieto *et al.*, 'Systematic review of surveillance systems and methods for early detection of exotic, new and re-emerging diseases in animal populations', *Epidemiology & Infection*, vol. 143, no. 10, pp. 2018–2042, Jul. 2015, doi: 10.1017/S095026881400212X.

[82] M. Mort, I. Convery, J. Baxter, and C. Bailey, 'Psychosocial effects of the 2001 UK foot and mouth disease epidemic in a rural population: qualitative diary based study', *BMJ*, vol. 331, no. 7527, p. 1234, Nov. 2005, doi: 10.1136/bmj.38603.375856.68.

[83] C. L. Gibbons *et al.*, 'Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods', *BMC Public Health*, vol. 14, no. 1, p. 147, Feb. 2014, doi: 10.1186/1471-2458-14-147.

[84] C. P. D. Birch, M. Bakrania, A. Prosser, D. Brown, S. M. Withenshaw, and S. H. Downs, 'Difference in differences analysis evaluates the effects of the badger control policy on bovine tuberculosis in England', *Sci Rep*, vol. 14, no. 1, p. 4849, Feb. 2024, doi: 10.1038/s41598-024-54062-4.

[85] GOV.UK, 'Bird flu (avian influenza): cases in wild birds', GOV.UK. Accessed: May 01, 2023. [Online]. Available: https://www.gov.uk/government/publications/avian-influenza-in-wild-birds

[86] S. Yousefinaghani, R. Dara, Z. Poljak, T. M. Bernardo, and S. Sharif, ‘The Assessment of Twitter’s Potential for Outbreak Detection: Avian Influenza Case Study’, *Sci Rep*, vol. 9, no. 1, p. 18147, Dec. 2019, doi: 10.1038/s41598-019-54388-4.

[87] S. Ryu, C. Han, S. T. Ali, C. Achangwa, B. Yang, and S. Pei, ‘Impact of public health and social measures on hand-foot-mouth disease transmission and prediction of upcoming season after relaxation of COVID-19 control measures’. Sep. 27, 2022. doi: 10.21203/rs.3.rs-1999622/v1.

[88] V. Suarez-Lledo and J. Alvarez-Galvez, ‘Prevalence of Health Misinformation on Social Media: Systematic Review’, *Journal of Medical Internet Research*, vol. 23, no. 1, p. e17187, Jan. 2021, doi: 10.2196/17187.

[89] S. Quach, P. Thaichon, K. D. Martin, S. Weaven, and R. W. Palmatier, ‘Digital technologies: tensions in privacy and data’, *J. of the Acad. Mark. Sci.*, vol. 50, no. 6, pp. 1299–1323, Nov. 2022, doi: 10.1007/s11747-022-00845-y.

[90] J. A. Karl, R. Fischer, E. Druică, F. Musso, and A. Stan, ‘Testing the Effectiveness of the Health Belief Model in Predicting Preventive Behavior During the COVID-19 Pandemic: The Case of Romania and Italy’, *Frontiers in Psychology*, vol. 12, 2022, Accessed: Mar. 31, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.627575

[91] I. M. Rosenstock, ‘The Health Belief Model and Preventive Health Behavior’, *Health Education Monographs*, vol. 2, no. 4, pp. 354–386, Dec. 1974, doi: 10.1177/109019817400200405.

[92] P. Gryffin, W. Chen, and N. Erenguc, ‘Knowledge, Attitudes and Beliefs of Meditation in College Students: Barriers and Opportunities’, *EDUCATION*, vol. 2, no. 4, pp. 189–192, Mar. 2014, doi: 10.12691/education-2-4-2.

[93] J. S. Sulat, Y. S. Prabandari, R. Sanusi, E. D. Hapsari, and B. Santoso, ‘The Validity of Health Belief Model Variables in Predicting Behavioral Change’, *Health Education*, vol. 118, no. 6, pp. 499–512, 2018, doi: 10.1108/HE-05-2018-0027.

[94] A. Khani-Jeihooni, M. Manouchehri, M. Bahmandoost, and Z. Khiyali, ‘Effect of Educational Intervention Based on the Health Belief Model on Preventive Behaviors Against Influenza A (H1N1) among Students’, *J Educ Community Health*, vol. 7, no. 2, Art. no. 2, Jun. 2020, doi: 10.29252/jech.7.2.97.

[95] T. Yamagishi, H. Hashimoto, and J. Schug, ‘Preferences versus strategies as explanations for culture-specific behavior’, *Psychol Sci*, vol. 19, no. 6, pp. 579–584, Jun. 2008, doi: 10.1111/j.1467-9280.2008.02126.x.

[96] M. Salathé, ‘Digital epidemiology: what is it, and where is it going?’, *Life Sci Soc Policy*, vol. 14, no. 1, p. 1, Jan. 2018, doi: 10.1186/s40504-017-0065-7.

[97] N. Mirin, H. Mattie, L. Jackson, Z. Samad, and R. Chunara, ‘Data Science in Public Health: Building Next Generation Capacity’, *Harvard Data Science Review*, vol. 4, no. 4, Oct. 2022, doi: 10.1162/99608f92.18da72db.

[98] P. Schröder-Bäck, P. Duncan, W. Sherlaw, C. Brall, and K. Czabanowska, ‘Teaching seven principles for public health ethics: towards a curriculum for a short course on ethics in public health programmes’, *BMC Medical Ethics*, vol. 15, no. 1, p. 73, Oct. 2014, doi: 10.1186/1472-6939-15-73.

[99] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, ‘Definitions, methods, and applications in interpretable machine learning’, *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22071–22080, Oct. 2019, doi: 10.1073/pnas.1900654116.

[100] D. Jakhar and I. Kaur, ‘Artificial intelligence, machine learning and deep learning: definitions and differences’, *Clinical and Experimental Dermatology*, vol. 45, no. 1, pp. 131–132, Jan. 2020, doi: 10.1111/ced.14029.

[101] J. Bao and Q. Xie, ‘Artificial intelligence in animal farming: A systematic literature review’, *Journal of Cleaner Production*, vol. 331, p. 129956, Jan. 2022, doi: 10.1016/j.jclepro.2021.129956.

[102] S. Fuentes, C. G. Viejo, E. Tongson, and F. R. Dunshea, ‘The livestock farming digital transformation: implementation of new and emerging technologies using artificial

intelligence', *Animal Health Research Reviews*, vol. 23, no. 1, pp. 59–71, Jun. 2022, doi: 10.1017/S1466252321000177.

[103] M. Abdulkareem and S. E. Petersen, 'The Promise of AI in Detection, Diagnosis, and Epidemiology for Combating COVID-19: Beyond the Hype', *Frontiers in Artificial Intelligence*, vol. 4, 2021, Accessed: Feb. 09, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2021.652669

[104] H. Coelho, 'Review of Social Physics: How Good Ideas Spread--the Lessons from a New Science'. Accessed: Jul. 05, 2023. [Online]. Available: https://www.jasss.org/19/1/reviews/2.html

[105] A. Pentland, *Social Physics: How Good Ideas Spread— The Lessons from a New Science*. 2014. Accessed: Jul. 05, 2023. [Online]. Available: https://www.goodreads.com/book/show/18079689-social-physics

[106] M. Jusup *et al.*, 'Social physics', *Physics Reports*, vol. 948, pp. 1–148, Feb. 2022, doi: 10.1016/j.physrep.2021.10.005.

[107] F. Ryan, A. Fritz, and D. Impiombato, 'TikTok privacy concerns and data collection', Australian Strategic Policy Institute, 2020. Accessed: May 12, 2024. [Online]. Available: https://www.jstor.org/stable/resrep26120.7

[108] S. Liang and J. Wolfe, 'Getting a Feel of Instagram Reels: The Effects of Posting Format on Online Engagement', *J Stud Res*, vol. 11, no. 4, Nov. 2022, doi: 10.47611/jsrhs.v11i4.3600.

[109] C. Stokel-Walker, 'Why is Twitter becoming X?', *New Scientist*, vol. 259, no. 3449, p. 9, Jul. 2023, doi: 10.1016/S0262-4079(23)01398-2.

[110] A. Petutschnig, 'IJGI | Free Full-Text | Evaluating the Representativeness of Socio-Demographic Variables over Time for Geo-Social Media Data', 2021, Accessed: May 12, 2024. [Online]. Available: https://www.mdpi.com/2220-9964/10/5/323

[111] M. V. M. de Lima and G. Z. Laporta, 'Evaluation of the Models for Forecasting Dengue in Brazil from 2000 to 2017: An Ecological Time-Series Study', *Insects*, vol. 11, no. 11, p. 794, Nov. 2020, doi: 10.3390/insects11110794.

[112] F. P. Rocha and M. Giesbrecht, 'Machine learning algorithms for dengue risk assessment: a case study for São Luís do Maranhão', *Comp. Appl. Math.*, vol. 41, no. 8, p. 393, Nov. 2022, doi: 10.1007/s40314-022-02101-z.

[113] K. Roster, C. Connaughton, and F. A. Rodrigues, 'Predicting Dengue Fever in Brazilian Cities'. bioRxiv, p. 2021.02.17.430949, Feb. 18, 2021. doi: 10.1101/2021.02.17.430949.

[114] A. M. Campbell, M.-F. Racault, S. Goult, and A. Laurenson, 'Cholera Risk: A Machine Learning Approach Applied to Essential Climate Variables', *Int J Environ Res Public Health*, vol. 17, no. 24, p. 9378, Dec. 2020, doi: 10.3390/ijerph17249378.

[115] Y. Wang, Z. Cao, D. Zeng, X. Wang, and Q. Wang, 'Using deep learning to predict the hand-foot-and-mouth disease of enterovirus A71 subtype in Beijing from 2011 to 2018', *Sci Rep*, vol. 10, no. 1, Art. no. 1, Jul. 2020, doi: 10.1038/s41598-020-68840-3.

[116] D. O. Oyewola, E. G. Dada, and S. Misra, 'Machine learning for optimizing daily COVID-19 vaccine dissemination to combat the pandemic', *Health Technol.*, vol. 12, no. 6, pp. 1277–1293, Nov. 2022, doi: 10.1007/s12553-022-00712-4.

[117] T. M. Lincoln *et al.*, 'Taking a machine learning approach to optimize prediction of vaccine hesitancy in high income countries', *Sci Rep*, vol. 12, no. 1, Art. no. 1, Feb. 2022, doi: 10.1038/s41598-022-05915-3.

[118] E. Robertson *et al.*, 'Predictors of COVID-19 vaccine hesitancy in the UK household longitudinal study', *Brain Behav Immun*, vol. 94, pp. 41–50, May 2021, doi: 10.1016/j.bbi.2021.03.008.

[119] V. Mhasawade, Y. Zhao, and R. Chunara, 'Machine learning and algorithmic fairness in public and population health', *Nat Mach Intell*, vol. 3, no. 8, Art. no. 8, Aug. 2021, doi: 10.1038/s42256-021-00373-4.

[120] T. Grote and G. Keeling, ‘Enabling Fairness in Healthcare Through Machine Learning’, *Ethics Inf Technol*, vol. 24, no. 3, p. 39, Aug. 2022, doi: 10.1007/s10676-022-09658-7.

[121] P. Noor, ‘Can we trust AI not to further embed racial bias and prejudice?’, *BMJ*, vol. 368, p. m363, Feb. 2020, doi: 10.1136/bmj.m363.

[122] A. Allen *et al.*, ‘A Racially Unbiased, Machine Learning Approach to Prediction of Mortality: Algorithm Development Study’, *JMIR Public Health Surveill*, vol. 6, no. 4, p. e22400, Oct. 2020, doi: 10.2196/22400.

[123] G. M. Clarke, S. Conti, A. T. Wolters, and A. Steventon, ‘Evaluating the impact of healthcare interventions using routine data’, *BMJ*, vol. 365, p. l2239, Jun. 2019, doi: 10.1136/bmj.l2239.

[124] A. Talaei-Khoei and J. M. Wilson, ‘Identifying people at risk of developing type 2 diabetes: A comparison of predictive analytics techniques and predictor variables’, *Int J Med Inform*, vol. 119, pp. 22–38, Nov. 2018, doi: 10.1016/j.ijmedinf.2018.08.008.

[125] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, ‘Risk prediction of cardiovascular disease using machine learning classifiers’, *Open Med (Wars)*, vol. 17, no. 1, pp. 1100–1113, Jun. 2022, doi: 10.1515/med-2022-0508.

[126] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, ‘Natural language processing applied to mental illness detection: a narrative review’, *npj Digit. Med.*, vol. 5, no. 1, Art. no. 1, Apr. 2022, doi: 10.1038/s41746-022-00589-7.

[127] D. Plana, D. L. Shung, A. A. Grimshaw, A. Saraf, J. J. Y. Sung, and B. H. Kann, ‘Randomized Clinical Trials of Machine Learning Interventions in Health Care: A Systematic Review’, *JAMA Network Open*, vol. 5, no. 9, p. e2233946, Sep. 2022, doi: 10.1001/jamanetworkopen.2022.33946.

[128] M. Abo-Tabik, Y. Benn, and N. Costen, ‘Are Machine Learning Methods the Future for Smoking Cessation Apps?’, *Sensors (Basel)*, vol. 21, no. 13, p. 4254, Jun. 2021, doi: 10.3390/s21134254.

[129] K. K. Mak, K. Lee, and C. Park, ‘Applications of machine learning in addiction studies: A systematic review’, *Psychiatry Res*, vol. 275, pp. 53–60, May 2019, doi: 10.1016/j.psychres.2019.03.001.

[130] S. Neethirajan, ‘The role of sensors, big data and machine learning in modern animal farming’, *Sensing and Bio-Sensing Research*, vol. 29, p. 100367, Aug. 2020, doi: 10.1016/j.sbsr.2020.100367.

[131] P. Ezanno *et al.*, ‘Research perspectives on animal health in the era of artificial intelligence’, *Veterinary Research*, vol. 52, no. 1, p. 40, Mar. 2021, doi: 10.1186/s13567-021-00902-4.

[132] R. P. Smith, C. Gavin, D. Gilson, R. R. L. Simons, and S. Williamson, ‘Determining pig holding type from British movement data using analytical and machine learning approaches’, *Preventive Veterinary Medicine*, vol. 178, p. 104984, May 2020, doi: 10.1016/j.prevetmed.2020.104984.

[133] M. Taneja, J. Byabazaire, N. Jalodia, A. Davy, C. Olariu, and P. Malone, ‘Machine learning based fog computing assisted data-driven approach for early lameness detection in dairy cattle’, *Computers and Electronics in Agriculture*, vol. 171, p. 105286, Apr. 2020, doi: 10.1016/j.compag.2020.105286.

[134] T. Bobbo, S. Biffani, C. Taccioli, M. Penasa, and M. Cassandro, ‘Comparison of machine learning methods to predict udder health status based on somatic cell counts in dairy cows’, *Sci Rep*, vol. 11, no. 1, Art. no. 1, Jul. 2021, doi: 10.1038/s41598-021-93056-4.

[135] A. D. Boursianis *et al.*, ‘Internet of Things (IoT) and Agricultural Unmanned Aerial Vehicles (UAVs) in smart farming: A comprehensive review’, *Internet of Things*, vol. 18, p. 100187, May 2022, doi: 10.1016/j.iot.2020.100187.

[136] G. Machado, M. R. Mendoza, and L. G. Corbellini, ‘What variables are important in predicting bovine viral diarrhea virus? A random forest approach’, *Vet Res*, vol. 46, no. 1, p. 85, 2015, doi: 10.1186/s13567-015-0219-7.

[137] B. Larison, K. Y. Njabo, A. Chasar, T. Fuller, R. J. Harrigan, and T. B. Smith, ‘Spillover of pH1N1 to swine in Cameroon: an investigation of risk factors’, *BMC Veterinary Research*, vol. 10, no. 1, p. 55, Mar. 2014, doi: 10.1186/1746-6148-10-55.

[138] J. Bradley and S. Rajendran, ‘Increasing adoption rates at animal shelters: a two-phase approach to predict length of stay and optimal shelter allocation’, *BMC Vet Res*, vol. 17, no. 1, p. 70, Feb. 2021, doi: 10.1186/s12917-020-02728-2.

[139] M. P. Romero *et al.*, ‘Decision tree machine learning applied to bovine tuberculosis risk factors to aid disease control decision making’, *Preventive Veterinary Medicine*, vol. 175, p. 104860, Feb. 2020, doi: 10.1016/j.prevetmed.2019.104860.

[140] P. S. Basran and R. B. Appleby, ‘The unmet potential of artificial intelligence in veterinary medicine’, *American Journal of Veterinary Research*, vol. 83, no. 5, pp. 385–392, May 2022, doi: 10.2460/ajvr.22.03.0038.

[141] J. S. Jones-Diette, R. S. Dean, M. Cobb, and M. L. Brennan, ‘Validation of text-mining and content analysis techniques using data collected from veterinary practice management software systems in the UK’, *Prev Vet Med*, vol. 167, pp. 61–67, Jun. 2019, doi: 10.1016/j.prevetmed.2019.02.015.

[142] D. T. Heinze, M. L. Morsch, and J. Holbrook, ‘Mining free-text medical records.’, *Proc AMIA Symp*, pp. 254–258, 2001.

[143] J. Rodríguez *et al.*, ‘A text-mining based analysis of 100,000 tumours affecting dogs and cats in the United Kingdom’, *Sci Data*, vol. 8, no. 1, Art. no. 1, Oct. 2021, doi: 10.1038/s41597-021-01039-x.

[144] E. D. Costa *et al.*, ‘Text Mining Analysis to Evaluate Stakeholders’ Perception Regarding Welfare of Equines, Small Ruminants, and Turkeys’, *Animals : an Open Access Journal from MDPI*, vol. 9, no. 5, May 2019, doi: 10.3390/ani9050225.

[145] H.-J. Lin, P. C.-Y. Sheu, J. J. P. Tsai, C. C. N. Wang, and C.-Y. Chou, ‘Text mining in a literature review of urothelial cancer using topic model’, *BMC Cancer*, vol. 20, no. 1, p. 462, May 2020, doi: 10.1186/s12885-020-06931-0.

[146] A. Zuliani *et al.*, ‘Topics and trends in Mountain Livestock Farming research: a text mining approach’, *Animal*, vol. 15, no. 1, p. 100058, Jan. 2021, doi: 10.1016/j.animal.2020.100058.

[147] A. Zunic, P. Corcoran, and I. Spasic, ‘Sentiment Analysis in Health and Well-Being: Systematic Review’, *JMIR Med Inform*, vol. 8, no. 1, p. e16023, Jan. 2020, doi: 10.2196/16023.

[148] S. Boon-Itt and Y. Skunkan, ‘Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study’, *JMIR Public Health Surveill*, vol. 6, no. 4, p. e21978, Nov. 2020, doi: 10.2196/21978.

[149] J. J. Myszewski, E. Klossowski, K. M. Schroeder, and C. A. Schroeder, ‘Utilization of sentiment analysis to assess and compare negative finding reporting in veterinary and human literature’, *Research in Veterinary Science*, vol. 148, pp. 27–32, Nov. 2022, doi: 10.1016/j.rvsc.2022.04.010.

[150] S. Sosa, C. Sueur, and I. Puga-Gonzalez, ‘Network measures in animal social network analysis: Their strengths, limits, interpretations and uses’, *Methods Ecol Evol*, vol. 12, no. 1, pp. 10–21, Jan. 2021, doi: 10.1111/2041-210X.13366.

[151] M. Valeri and R. Baggio, ‘Social network analysis: organizational implications in tourism management’, *International Journal of Organizational Analysis*, vol. 29, no. 2, pp. 342–353, Jan. 2020, doi: 10.1108/IJOA-12-2019-1971.

[152] T. W. Valente and S. R. Pitts, ‘An Appraisal of Social Network Theory and Analysis as Applied to Public Health: Challenges and Opportunities’, *Annual Review of Public Health*, vol. 38, no. 1, pp. 103–118, 2017, doi: 10.1146/annurev-publhealth-031816-044528.

[153] L. P. Doan *et al.*, ‘Social network and HIV/AIDS: A bibliometric analysis of global literature’, *Frontiers in Public Health*, vol. 10, 2022, Accessed: Feb. 10, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2022.1015023

[154] B. Govoeyi *et al.*, ‘Social network analysis of practice adoption facing outbreaks of African Swine Fever’, *Preventive Veterinary Medicine*, vol. 179, p. 105008, Jun. 2020, doi: 10.1016/j.prevetmed.2020.105008.

[155] A. Cottica, A. Hassoun, M. Manca, J. Vallet, and G. Melançon, ‘Semantic Social Networks: A Mixed Methods Approach to Digital Ethnography’, *Field Methods*, vol. 32, no. 3, pp. 274–290, Aug. 2020, doi: 10.1177/1525822X20908236.

[156] C. Wetherell, ‘Historical Social Network Analysis*’, *International Review of Social History*, vol. 43, no. S6, pp. 125–144, Dec. 1998, doi: 10.1017/S0020859000115123.

[157] I. de Freslon, B. Martínez-López, J. Belkhiria, A. Strappini, and G. Monti, ‘Use of social network analysis to improve the understanding of social behaviour in dairy cattle and its impact on disease transmission’, *Applied Animal Behaviour Science*, vol. 213, pp. 47–54, Apr. 2019, doi: 10.1016/j.applanim.2019.01.006.

[158] S. Vishnu, J. Gupta, and S. P. Subash, ‘Social network structures among the livestock farmers vis a vis calcium supplement technology’, *Information Processing in Agriculture*, vol. 6, no. 1, pp. 170–182, Mar. 2019, doi: 10.1016/j.inpa.2018.07.006.

[159] A. Albizua, E. M. Bennett, G. Larocque, R. W. Krause, and U. Pascual, ‘Social networks influence farming practices and agrarian sustainability’, *PLOS ONE*, vol. 16, no. 1, p. e0244619, Jan. 2021, doi: 10.1371/journal.pone.0244619.

[160] ‘PMAP - Policy brief 8’. Accessed: Feb. 10, 2023. [Online]. Available: https://www.pik-potsdam.de/~wrobel/mediation-platform/pbs/pb8/strengths_and_weaknesses.html

[161] X. Su *et al.*, ‘A Comprehensive Survey on Community Detection With Deep Learning’, *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2022, doi: 10.1109/TNNLS.2021.3137396.

[162] Z. Yang, R. Algesheimer, and C. J. Tessone, ‘A Comparative Analysis of Community Detection Algorithms on Artificial Networks’, *Sci Rep*, vol. 6, no. 1, Art. no. 1, Aug. 2016, doi: 10.1038/srep30750.

[163] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, ‘Community detection in Social Media’, *Data Min Knowl Disc*, vol. 24, no. 3, pp. 515–554, May 2012, doi: 10.1007/s10618-011-0224-z.

[164] P. Chunaev, ‘Community detection in node-attributed social networks: A survey’, *Computer Science Review*, vol. 37, p. 100286, Aug. 2020, doi: 10.1016/j.cosrev.2020.100286.

[165] N. Barbieri, F. Bonchi, and G. Manco, ‘Influence-Based Network-Oblivious Community Detection’, in *2013 IEEE 13th International Conference on Data Mining*, Dallas, TX, USA: IEEE, Dec. 2013, pp. 955–960. doi: 10.1109/ICDM.2013.164.

[166] V. A. Traag, L. Waltman, and N. J. van Eck, ‘From Louvain to Leiden: guaranteeing well-connected communities’, *Sci Rep*, vol. 9, no. 1, Art. no. 1, Mar. 2019, doi: 10.1038/s41598-019-41695-z.

[167] M. Yan and C. Guoqiang, ‘Label Propagation Community Detection Algorithm Based on Density Peak Optimization’, *Wireless Communications and Mobile Computing*, vol. 2022, p. e6523363, Apr. 2022, doi: 10.1155/2022/6523363.

[168] J. Zeng and H. Yu, ‘A Distributed Infomap Algorithm for Scalable and High-Quality Community Detection’, in *Proceedings of the 47th International Conference on Parallel Processing*, Eugene OR USA: ACM, Aug. 2018, pp. 1–11. doi: 10.1145/3225058.3225137.

[169] H. M. Alash and G. A. Al-Sultany, ‘Enhanced Twitter Community Detection using Node Content and Attributes’, in *2021 1st Babylon International Conference on Information Technology and Science (BICITS)*, Apr. 2021, pp. 5–10. doi: 10.1109/BICITS51482.2021.9509873.

[170] D. Joshi and T. Patalia, ‘Community Detection Methods and Tools for Various Complex Network’, *International Journal of Engineering Research and Technology*, vol. 13, no. 6, p. 1386, Jun. 2020, doi: 10.37624/IJERT/13.6.2020.1386-1390.

[171] A. Clauset, M. E. J. Newman, and C. Moore, ‘Finding community structure in very large networks’, *Phys. Rev. E*, vol. 70, no. 6, p. 066111, Dec. 2004, doi: 10.1103/PhysRevE.70.066111.

[172] S. Jiang and S. Tokdar, ‘Consistent Bayesian Community Detection’, Jan. 2021. Accessed: Feb. 12, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Consistent-Bayesian-Community-Detection-Jiang-Tokdar/084204313fe6d86872430154570fae9442ef0d27

[173] G. Bello, J. Hernandez-Castro, and D. Camacho, ‘Detecting discussion communities on vaccination in twitter’, *Future Generation Computer Systems*, vol. 66, pp. 125–136, Jul. 2016.

[174] D. Surian, D. Q. Nguyen, G. Kennedy, M. Johnson, E. Coiera, and A. G. Dunn, ‘Characterizing Twitter Discussions About HPV Vaccines Using Topic Modeling and Community Detection’, *Journal of Medical Internet Research*, vol. 18, no. 8, p. e6045, Aug. 2016, doi: 10.2196/jmir.6045.

[175] J. D. Featherstone, G. A. Barnett, J. B. Ruiz, Y. Zhuang, and B. J. Millam, ‘Exploring childhood anti-vaccine and pro-vaccine communities on twitter – a perspective from influential users’, *Online Social Networks and Media*, vol. 20, p. 100105, Nov. 2020, doi: 10.1016/j.osnem.2020.100105.

[176] X. Yuan, R. J. Schuchard, and A. T. Crooks, ‘Examining Emergent Communities and Social Bots Within the Polarized Online Vaccination Debate in Twitter’, *Social Media + Society*, vol. 5, no. 3, p. 205630511986546, Jul. 2019, doi: 10.1177/2056305119865465.

[177] A. Cossard, G. D. F. Morales, K. Kalimeri, Y. Mejova, D. Paolotti, and M. Starnini, ‘Falling into the Echo Chamber: The Italian Vaccination Debate on Twitter’, *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, pp. 130–140, May 2020, doi: 10.1609/icwsm.v14i1.7285.

[178] A. J. Lam and C. Cheng, ‘Utilizing Tweet Content for the Detection of Sentiment-Based Interaction Communities on Twitter’, in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2018, pp. 682–691. doi: 10.1109/DSAA.2018.00088.

[179] A. Kanavos, I. Perikos, I. Hatzilygeroudis, and A. Tsakalidis, ‘Emotional community detection in social networks’, *Computers & Electrical Engineering*, vol. 65, pp. 449–460, Jan. 2018, doi: 10.1016/j.compeleceng.2017.09.011.

[180] G. Amati *et al.*, ‘Topic Modeling by Community Detection Algorithms’, in *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks*, in OASIS ’21. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 15–20. doi: 10.1145/3472720.3483622.

[181] M. Benabdelkrim, C. Robardet, and J. Savinien, ‘Leveraging semantic for community mining in multilayer networks’, *Proceedings of the Canadian Conference on Artificial Intelligence*, Jun. 2021, doi: 10.21428/594757db.d7f76fa5.

[182] M. Benabdelkrim, J. Savinien, and C. Robardet, ‘Finding interest groups from Twitter lists’, in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, Brno Czech Republic: ACM, Mar. 2020, pp. 1885–1887. doi: 10.1145/3341105.3374077.

[183] El. Akachar, B. Ouhbi, and B. Frikh, ‘A new algorithm for detecting communities in social networks based on content and structure information’, *International Journal of Web Information Systems*, vol. 16, no. 1, pp. 79–93, Jan. 2019, doi: 10.1108/IJWIS-06-2019-0030.

[184] H. N. Win and K. T. Lynn, ‘Community detection in Facebook with outlier recognition’, in *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, Jun. 2017, pp. 155–159. doi: 10.1109/SNPD.2017.8022716.

[185] E. Ferrara, ‘A large-scale community structure analysis in Facebook’, *EPJ Data Sci.*, vol. 1, no. 1, Art. no. 1, Dec. 2012, doi: 10.1140/epjds9.

[186] F. Alimadadi, E. Khadangi, and A. Bagheri, ‘Community detection in facebook activity networks and presenting a new multilayer label propagation algorithm for community

detection', *Int. J. Mod. Phys. B*, vol. 33, no. 10, p. 1950089, Apr. 2019, doi: 10.1142/S0217979219500899.

[187] B. P. Chamberlain, J. Levy-Kramer, C. Humby, and M. P. Deisenroth, 'Real-time community detection in full social networks on a laptop', *PLOS ONE*, vol. 13, no. 1, p. e0188702, Jan. 2018, doi: 10.1371/journal.pone.0188702.

[188] D. Salz, N. Benavides, and J. Li, 'Hidden Community Detection in Online Forums', 2020.

[189] A. Shen and K.-P. Chow, 'Community Detection in a Web Discussion Forum During Social Unrest Events', in *Advances in Digital Forensics XVIII*, G. Peterson and S. Shenoi, Eds., in IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing, 2022, pp. 169–185. doi: 10.1007/978-3-031-10078-9_10.

[190] J. Wu, 'How WeChat, the Most Popular Social Network in China, Cultivates Wellbeing', *Master of Applied Positive Psychology (MAPP) Capstone Projects*, Sep. 2014, [Online]. Available: https://repository.upenn.edu/mapp_capstone/65

[191] S. M. Albladi and G. R. S. Weir, 'User characteristics that influence judgment of social engineering attacks in social networks', *Human-centric Computing and Information Sciences*, vol. 8, no. 1, p. 5, Feb. 2018, doi: 10.1186/s13673-018-0128-7.

[192] S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky, 'Community Interaction and Conflict on the Web', in *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 2018, pp. 933–943. doi: 10.1145/3178876.3186141.

[193] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, 'Detection of Depression-Related Posts in Reddit Social Media Forum', *IEEE Access*, vol. 7, pp. 44883–44893, 2019, doi: 10.1109/ACCESS.2019.2909180.

[194] H. Hu, Z. Sun, F. Wang, L. Zhang, and G. Wang, 'Exploring influential nodes using global and local information', *Sci Rep*, vol. 12, no. 1, Art. no. 1, Dec. 2022, doi: 10.1038/s41598-022-26984-4.

[195] M. M. Tulu, R. Hou, and T. Younas, 'Identifying Influential Nodes Based on Community Structure to Speed up the Dissemination of Information in Complex Network', *IEEE Access*, vol. 6, pp. 7390–7401, 2018, doi: 10.1109/ACCESS.2018.2794324.

[196] M. Hu, 'Cambridge Analytica's black box', *Big Data & Society*, vol. 7, no. 2, p. 2053951720938091, Jul. 2020, doi: 10.1177/2053951720938091.

[197] A. Arora, S. Bansal, C. Kandpal, R. Aswani, and Y. Dwivedi, 'Measuring social media influencer index- insights from facebook, Twitter and Instagram', *Journal of Retailing and Consumer Services*, vol. 49, pp. 86–101, Jul. 2019, doi: 10.1016/j.jretconser.2019.03.012.

[198] S. A. Alsaif, A. Hidri, and M. S. Hidri, 'Towards Inferring Influential Facebook Users', *Computers*, vol. 10, no. 5, Art. no. 5, May 2021, doi: 10.3390/computers10050062.

[199] A. Zareie, A. Sheikhahmadi, and M. Jalili, 'Identification of influential users in social networks based on users' interest', *Information Sciences*, vol. 493, pp. 217–231, Aug. 2019, doi: 10.1016/j.ins.2019.04.033.

[200] N. Alotaibi and D. Rhouma, 'A review on community structures detection in time evolving social networks', *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, Part B, pp. 5646–5662, Sep. 2022, doi: 10.1016/j.jksuci.2021.08.016.

[201] C. Wang and H. Zhao, 'The Impact of COVID-19 on Anxiety in Chinese University Students', *Frontiers in Psychology*, vol. 11, 2020, Accessed: Feb. 13, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01168

[202] W. Karoui, N. Hafiene, and L. Ben Romdhane, 'Machine learning-based method to predict influential nodes in dynamic social networks', *Soc. Netw. Anal. Min.*, vol. 12, no. 1, p. 108, Aug. 2022, doi: 10.1007/s13278-022-00942-4.

[203] G. Zhao, P. Jia, C. Huang, A. Zhou, and Y. Fang, 'A Machine Learning Based Framework for Identifying Influential Nodes in Complex Networks', *IEEE Access*, vol. 8, pp. 65462–65471, 2020, doi: 10.1109/ACCESS.2020.2984286.

[204] S. Jain and A. Sinha, 'Identification of influential users on Twitter: A novel weighted correlated influence measure for Covid-19', *Chaos Solitons Fractals*, vol. 139, p. 110037, Oct. 2020, doi: 10.1016/j.chaos.2020.110037.

[205] A. Dhokar, L. Hlaoua, and L. B. Romdhane, 'Tweet Contextualization Approach Using a Semantic Query Expansion', *Procedia Computer Science*, vol. 192, pp. 387–396, Jan. 2021, doi: 10.1016/j.procs.2021.08.040.

[206] L. Alsudias and P. Rayson, 'Classifying Information Sources in Arabic Twitter to Support Online Monitoring of Infectious Diseases', p. 9, 2019.

[207] K. Roitero, C. Bozzato, S. Mizzaro, and G. Serra, 'Twitter goes to the Doctor: Detecting Medical Tweets using Machine Learning and BERT', 2020.

[208] F. Alhayan, D. Pennington, and S. Ayouni, 'Twitter use by the dementia community during COVID-19: a user classification and social network analysis', *OIR*, Apr. 2022, doi: 10.1108/OIR-04-2021-0208.

[209] F. C. Dórea, B. J. McEwen, W. B. McNab, J. Sanchez, and C. W. Revie, 'Syndromic Surveillance Using Veterinary Laboratory Data: Algorithm Combination and Customization of Alerts', *PLOS ONE*, vol. 8, no. 12, p. e82183, Dec. 2013, doi: 10.1371/journal.pone.0082183.

[210] A. Kim, T. Miano, R. Chew, M. Eggers, and J. Nonnemaker, 'Classification of Twitter Users Who Tweet About E-Cigarettes', *JMIR Public Health Surveill*, vol. 3, no. 3, p. e63, Sep. 2017, doi: 10.2196/publichealth.8060.

[211] Z. Gilani, E. Kochmar, and J. Crowcroft, 'Classification of Twitter Accounts into Automated Agents and Human Users', in *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Jul. 2017, pp. 489–496.

[212] K. E. Daouadi, R. Z. Rebaï, and I. Amous, 'Organization vs. Individual: Twitter User Classification.', 2019.

[213] P. Skelsey, 'Forecasting Risk of Crop Disease with Anomaly Detection Algorithms', *Phytopathology®*, vol. 111, no. 2, pp. 321–332, Feb. 2021, doi: 10.1094/PHYTO-05-20-0185-R.

[214] 'Forecasting Risk of Crop Disease with Anomaly Detection Algorithms | Phytopathology®'. Accessed: Dec. 19, 2022. [Online]. Available: https://apsjournals.apsnet.org/doi/full/10.1094/PHYTO-05-20-0185-R?rfr_dat=cr_pub++0pubmed&url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&

[215] M. J. Paul *et al.*, 'SOCIAL MEDIA MINING FOR PUBLIC HEALTH MONITORING AND SURVEILLANCE', in *Biocomputing 2016*, Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC, Jan. 2016, pp. 468–479. doi: 10.1142/9789814749411_0043.

[216] A. Culotta, 'Towards detecting influenza epidemics by analyzing Twitter messages', in *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, Washington D.C., District of Columbia: ACM Press, 2010, pp. 115–122. doi: 10.1145/1964858.1964874.

[217] 'TWITTER IMPROVES SEASONAL INFLUENZA PREDICTION':, in *Proceedings of the International Conference on Health Informatics*, Vilamoura, Algarve, Portugal: SciTePress - Science and and Technology Publications, 2012, pp. 61–70. doi: 10.5220/0003780600610070.

[218] D. Sharpe, R. Hopkins, R. L. Cook, and C. W. Striley, 'Using a Bayesian Method to Assess Google, Twitter, and Wikipedia for ILI Surveillance', *Online J Public Health Inform*, vol. 9, no. 1, p. e026, May 2017, doi: 10.5210/ojphi.v9i1.7604.

[219] Y. Chen, Y. Zhang, Z. Xu, X. Wang, J. Lu, and W. Hu, 'Avian Influenza A (H7N9) and related Internet search query data in China', *Sci Rep*, vol. 9, no. 1, p. 10434, Jul. 2019, doi: 10.1038/s41598-019-46898-y.

[220] C. Comito, D. Falcone, and D. Talia, 'A Peak Detection Method to Uncover Events from Social Media', in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2017, pp. 459–467. doi: 10.1109/DSAA.2017.69.

[221] M. S. Mredula, N. Dey, M. S. Rahman, I. Mahmud, and Y.-Z. Cho, 'A Review on the Trends in Event Detection by Analyzing Social Media Platforms' Data', *Sensors*, vol. 22, no. 12, Art. no. 12, Jan. 2022, doi: 10.3390/s22124531.

[222] L. E. Charles-Smith *et al.*, 'Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review', *PLOS ONE*, vol. 10, no. 10, p. e0139701, Oct. 2015, doi: 10.1371/journal.pone.0139701.

[223] P. Healy, G. Hunt, S. Kilroy, T. Lynn, J. P. Morrison, and S. Venkatagiri, 'Evaluation of peak detection algorithms for social media event detection', in *2015 10th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, Nov. 2015, pp. 1–9. doi: 10.1109/SMAP.2015.7370090.

[224] K. D. Mandl *et al.*, 'Implementing Syndromic Surveillance: A Practical Guide Informed by the Early Experience', *J Am Med Inform Assoc*, vol. 11, no. 2, pp. 141–150, 2004, doi: 10.1197/jamia.M1356.

[225] A. J. Elliot and H. E. Hughes, 'Real-time Syndromic Surveillance'. [Online]. Available: https://extranet.who.int/kobe_centre/sites/default/files/pdf/WHO%20Guidance_Research%20Methods_Health-EDRM_4.9.pdf

[226] A. Chiolero and D. Buckeridge, 'Glossary for public health surveillance in the age of data science', *J Epidemiol Community Health*, vol. 74, no. 7, pp. 612–616, Jun. 2020, doi: 10.1136/jech-2018-211654.

[227] 'The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic | PLOS ONE'. Accessed: Dec. 09, 2022. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0019467

[228] E. Chen, K. Lerman, and E. Ferrara, 'Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set', *JMIR Public Health and Surveillance*, vol. 6, no. 2, p. e19273, May 2020, doi: 10.2196/19273.

[229] A. Szpiro, 'Health Surveillance and Diagnosis for Mitigating a Bioterror Attack'.

[230] L. Samaras, E. García-Barriocanal, and M.-A. Sicilia, 'Comparing Social media and Google to detect and predict severe epidemics', *Sci Rep*, vol. 10, no. 1, Art. no. 1, Mar. 2020, doi: 10.1038/s41598-020-61686-9.

[231] 'Syndromic surveillance: systems and analyses', GOV.UK. Accessed: Dec. 09, 2022. [Online]. Available: https://www.gov.uk/government/collections/syndromic-surveillance-systems-and-analyses

[232] J. Wells *et al.*, 'Real-time surveillance of severe acute respiratory infections in Scottish hospitals: an electronic register-based approach, 2017–2022', *Public Health*, vol. 213, pp. 5–11, Dec. 2022, doi: 10.1016/j.puhe.2022.09.003.

[233] 'Manual 5 : Surveillance and Epidemiology', O.I.E (World Orgnisation for Animal Health), 2017. doi: 10.20506/standz.2796.

[234] M. Hernández-Jover, B. J. Phiri, L. Stringer, and M. Martínez Avilés, 'Editorial: Developments in Animal Health Surveillance', *Frontiers in Veterinary Science*, vol. 7, 2021, Accessed: Feb. 06, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fvets.2020.637364

[235] F. C. Dórea *et al.*, 'Drivers for the development of an Animal Health Surveillance Ontology (AHSO)', *Preventive Veterinary Medicine*, vol. 166, pp. 39–48, May 2019, doi: 10.1016/j.prevetmed.2019.03.002.

[236] A. Chaskopoulou, C. I. Dovas, S. C. Chaintoutis, J. Kashefi, P. Koehler, and M. Papanastassopoulou, 'Detection and Early Warning of West Nile Virus Circulation in Central Macedonia, Greece, Using Sentinel Chickens and Mosquitoes', *Vector-Borne and Zoonotic Diseases*, vol. 13, no. 10, pp. 723–732, Oct. 2013, doi: 10.1089/vbz.2012.1176.

[237] J. Berezowski *et al.*, 'One Health Surveillance: perceived benefits and workforce motivations', *Rev Sci Tech*, vol. 38, no. 1, pp. 251–260, May 2019, doi: 10.20506/rst.38.1.2957.

[238] I. R. Mremi, C. Sindato, C. Kishamawe, S. F. Rumisha, S. I. Kimera, and L. E. G. Mboera, 'Improving disease surveillance data analysis, interpretation, and use at the district level in Tanzania', *Glob Health Action*, vol. 15, no. 1, p. 2090100, doi: 10.1080/16549716.2022.2090100.

[239] F. Brauer, ‘Mathematical epidemiology: Past, present, and future’, *Infect Dis Model*, vol. 2, no. 2, pp. 113–127, Feb. 2017, doi: 10.1016/j.idm.2017.02.001.

[240] P. Jones *et al.*, ‘SAVSNET: Collating Veterinary Electronic Health Records for Research and Surveillance’, *Online Journal of Public Health Informatics*, vol. 8, Mar. 2016, doi: 10.5210/ojphi.v8i1.6543.

[241] D. E. Chamberlain, S. Gough, H. Vaughan, J. A. Vickery, and G. F. Appleton, ‘Determinants of bird species richness in public green spaces’, *Bird Study*, vol. 54, no. 1, pp. 87–97, Mar. 2007, doi: 10.1080/00063650709461460.

[242] J. Goplin and M. Benz, ‘North Dakota Electronic Animal Health Surveillance System’.

[243] P. K. Maciejewski, B. Zhang, S. D. Block, and H. G. Prigerson, ‘An empirical examination of the stage theory of grief’, *JAMA*, vol. 297, no. 7, pp. 716–723, Feb. 2007, doi: 10.1001/jama.297.7.716.

[244] A. and P. H. Agency, ‘APHA Vet Gateway: Livestock disease surveillance dashboards’. Accessed: Dec. 27, 2022. [Online]. Available: http://apha.defra.gov.uk/vet-gateway/surveillance/scanning/disease-dashboards.htm

[245] R. Gupta, V. Mohanty, A. Y. Balappanavar, P. Chahar, K. Rijhwani, and S. Bhatia, ‘Infodemiology for oral health and disease: A scoping review’, *Health Information & Libraries Journal*, vol. 39, no. 3, pp. 207–224, 2022, doi: 10.1111/hir.12453.

[246] N. Calleja *et al.*, ‘A Public Health Research Agenda for Managing Infodemics: Methods and Results of the First WHO Infodemiology Conference’, *JMIR Infodemiology*, vol. 1, no. 1, p. e30979, Sep. 2021, doi: 10.2196/30979.

[247] J. M. Gorman and D. A. Scales, ‘Leveraging infodemiologists to counteract online misinformation: Experience with COVID-19 vaccines’, *Harvard Kennedy School Misinformation Review*, Feb. 2022, doi: 10.37016/mr-2020-92.

[248] ‘Biggest social media platforms 2022’, Statista. Accessed: Feb. 07, 2023. [Online]. Available: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[249] S. Rennie, M. Buchbinder, E. Juengst, L. Brinkley-Rubinstein, C. Blue, and D. L. Rosen, ‘Scraping the Web for Public Health Gains: Ethical Considerations from a “Big Data” Research Project on HIV and Incarceration’, *Public Health Ethics*, vol. 13, no. 1, pp. 111–121, Mar. 2020, doi: 10.1093/phe/phaa006.

[250] J. Stern, S. Georgsson, and T. Carlsson, ‘Quality of web-based information about the coronavirus disease 2019: a rapid systematic review of infodemiology studies published during the first year of the pandemic’, *BMC Public Health*, vol. 22, no. 1, p. 1734, Sep. 2022, doi: 10.1186/s12889-022-14086-9.

[251] Y. Wang *et al.*, ‘Understanding and neutralising covid-19 misinformation and disinformation’, *BMJ*, vol. 379, p. e070331, Nov. 2022, doi: 10.1136/bmj-2022-070331.

[252] L. Clarke, ‘Covid-19: Who fact checks health and science on Facebook?’, *BMJ*, vol. 373, p. n1170, May 2021, doi: 10.1136/bmj.n1170.

[253] M. Himelein-Wachowiak *et al.*, ‘Bots and Misinformation Spread on Social Media: Implications for COVID-19’, *J Med Internet Res*, vol. 23, no. 5, p. e26933, May 2021, doi: 10.2196/26933.

[254] F. Cascini *et al.*, ‘Social media and attitudes towards a COVID-19 vaccination: A systematic review of the literature’, *eClinicalMedicine*, vol. 48, Jun. 2022, doi: 10.1016/j.eclinm.2022.101454.

[255] P. Soares *et al.*, ‘Factors Associated with COVID-19 Vaccine Hesitancy’, *Vaccines (Basel)*, vol. 9, no. 3, p. 300, Mar. 2021, doi: 10.3390/vaccines9030300.

[256] ‘Identifying and Tracking SARS-CoV-2 Variants — A Challenge and an Opportunity | NEJM’. Accessed: Feb. 07, 2023. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMp2103859

[257] J. T. Brooks and J. C. Butler, ‘Effectiveness of Mask Wearing to Control Community Spread of SARS-CoV-2’, *JAMA*, vol. 325, no. 10, pp. 998–999, Mar. 2021, doi: 10.1001/jama.2021.1505.

[258] ‘How social distancing, mobility, and preventive policies affect COVID-19 outcomes: Big data-driven evidence from the District of Columbia-Maryland-Virginia (DMV) megaregion | PLOS ONE’. Accessed: Feb. 07, 2023. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0263820

[259] L. Rajmil, A. Hjern, P. Boran, G. Gunnlaugsson, O. Kraus de Camargo, and S. Raman, ‘Impact of lockdown and school closure on children’s health and well-being during the first wave of COVID-19: a narrative review’, *BMJ Paediatr Open*, vol. 5, no. 1, p. e001043, May 2021, doi: 10.1136/bmjpo-2021-001043.

[260] M. M. Mello and C. J. Wang, ‘Ethics and governance for digital disease surveillance’, *Science*, vol. 368, no. 6494, pp. 951–954, May 2020, doi: 10.1126/science.abb9045.

[261] J. P. West and J. S. Bowman, ‘Electronic Surveillance at Work: An Ethical Analysis’, *Administration & Society*, vol. 48, no. 5, pp. 628–651, Jul. 2016, doi: 10.1177/0095399714556502.

[262] L. Silici, J. Knox, A. Rowe, and S. Nanthikesan, ‘Evaluating Transformational Adaptation in Smallholder Farming: Insights from an Evidence Review’, in *Transformational Change for People and the Planet: Evaluating Environment and Development*, J. I. Uitto and G. Batra, Eds., in Sustainable Development Goals Series. , Cham: Springer International Publishing, 2022, pp. 187–202. doi: 10.1007/978-3-030-78853-7_13.

[263] A. S. Cohn *et al.*, ‘Smallholder Agriculture and Climate Change’, *Annual Review of Environment and Resources*, vol. 42, no. 1, pp. 347–375, 2017, doi: 10.1146/annurev-environ-102016-060946.

[264] S. Bhatia *et al.*, ‘Using digital surveillance tools for near real-time mapping of the risk of infectious disease spread’, *npj Digit. Med.*, vol. 4, no. 1, Art. no. 1, Apr. 2021, doi: 10.1038/s41746-021-00442-3.

[265] D. Collins, ‘Cambridge Analytica: Data Privacy - Hansard - UK Parliament’. Accessed: May 07, 2024. [Online]. Available: https://hansard.parliament.uk/commons/2018-03-19/debates/2015B5CE-9F99-4B8D-B195-57C51AB4FD0C/CambridgeAnalyticaDataPrivacy

[266] M. Westerlund, D. Isabelle, and S. Leminen, ‘The Acceptance of Digital Surveillance in an Age of Big Data’, *Technology Innovation Management Review*, vol. 11, no. 3, pp. 32–44, 2021, doi: http://doi.org/10.22215/timreview/1427.

[267] F. D. Davis, ‘Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology’, *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989, doi: 10.2307/249008.

[268] M. Nguyen *et al.*, ‘Using the technology acceptance model to explore health provider and administrator perceptions of the usefulness and ease of using technology in palliative care’, *BMC Palliative Care*, vol. 19, no. 1, p. 138, Sep. 2020, doi: 10.1186/s12904-020-00644-8.

[269] A. A. AlQudah, M. Al-Emran, and K. Shaalan, ‘Technology Acceptance in Healthcare: A Systematic Review’, *Applied Sciences*, vol. 11, no. 22, Art. no. 22, Jan. 2021, doi: 10.3390/app112210537.

[270] I. Ajzen, ‘The theory of planned behavior’, *Organizational Behavior and Human Decision Processes*, vol. 50, no. 2, pp. 179–211, Dec. 1991, doi: 10.1016/0749-5978(91)90020-T.

[271] V. Pedrinelli, A. Rossi, and M. A. Brunetto, ‘Theory of Planned Behavior applied to the choice of food with preservatives by owners and for their dogs’, *PLOS ONE*, vol. 19, no. 1, p. e0294044, Jan. 2024, doi: 10.1371/journal.pone.0294044.

[272] H. Si, ‘IJERPH | Free Full-Text | Application of the Theory of Planned Behavior in Environmental Science: A Comprehensive Bibliometric Analysis’, 2019, Accessed: May 07, 2024. [Online]. Available: https://www.mdpi.com/1660-4601/16/15/2788

[273] D. Kapgen and L. Roudart, ‘A Multidisciplinary Approach to Assess Smallholder Farmers’ Adoption of New Technologies in Development Interventions’, *Eur J Dev Res*, vol. 35, no. 4, pp. 974–995, Aug. 2023, doi: 10.1057/s41287-022-00548-8.

[274] A. Adams and E. T. Jumpah, ‘Agricultural technologies adoption and smallholder farmers’ welfare: Evidence from Northern Ghana’, *Cogent Economics & Finance*, vol. 9, no. 1, p. 2006905, Jan. 2021, doi: 10.1080/23322039.2021.2006905.

[275] S. Islam, B. Janghel, D. Pandey, H. Khan, and A. Sahu, ‘Iot And Smart Farming: A Comprehensive Analysis Of The Indian Scenario’, vol. 18, no. 2, 2021.

[276] J. Kaur, S. M. Hazrati Fard, M. Amiri-Zarandi, and R. Dara, ‘Protecting farmers’ data privacy and confidentiality: Recommendations and considerations’, *Front. Sustain. Food Syst.*, vol. 6, Oct. 2022, doi: 10.3389/fsufs.2022.903230.

[277] J. Henderson, ‘How to support small-holder farmers through equitable data and digital technology’. Accessed: May 09, 2024. [Online]. Available: https://www.data4sdgs.org/blog/how-support-small-holder-farmers-through-equitable-data-and-digital-technology

[278] S. van der Burg, L. Wiseman, and J. Krkeljas, ‘Trust in farm data sharing: reflections on the EU code of conduct for agricultural data sharing’, *Ethics Inf Technol*, vol. 23, no. 3, pp. 185–198, Sep. 2021, doi: 10.1007/s10676-020-09543-1.

[279] S. Borthakur and R. Chhatpar, ‘How can data-sharing partnerships enhance access to finance for smallholder farmers? - IDH - the Sustainable Trade Initiative’. Accessed: May 09, 2024. [Online]. Available: https://www.idhsustainabletrade.com/news/how-can-data-sharing-partnerships-enhance-access-to-finance-for-smallholder-farmers/

[280] D. Smart *et al.*, ‘What Influences Parents and Practitioners’ Decisions to Share Personal Information within an Early Help (Social Care) Context? Implications for Practice in Sharing Digital Data across Sectors’, *The British Journal of Social Work*, vol. 52, no. 4, pp. 2146–2165, Jun. 2022, doi: 10.1093/bjsw/bcab167.

[281] D. Fiocco, ‘Agtech: Breaking down the farmer adoption dilemma | McKinsey’. Accessed: May 12, 2024. [Online]. Available: https://www.mckinsey.com/industries/agriculture/our-insights/agtech-breaking-down-the-farmer-adoption-dilemma

[282] A. Kerasidou and C. (Xaroula) Kerasidou, ‘Data-driven research and healthcare: public trust, data governance and the NHS’, *BMC Medical Ethics*, vol. 24, no. 1, p. 51, Jul. 2023, doi: 10.1186/s12910-023-00922-z.

[283] M. Sheehan *et al.*, ‘Trust, trustworthiness and sharing patient data for research’, *Journal of Medical Ethics*, vol. 47, no. 12, pp. e26–e26, Dec. 2021, doi: 10.1136/medethics-2019-106048.

[284] J. van Heek, K. Arning, and M. Ziefle, ‘The Surveillance Society: Which Factors Form Public Acceptance of Surveillance Technologies?’, in *Smart Cities, Green Technologies, and Intelligent Transport Systems*, M. Helfert, C. Klein, B. Donnellan, and O. Gusikhin, Eds., Cham: Springer International Publishing, 2017, pp. 170–191. doi: 10.1007/978-3-319-63712-9_10.

[285] R. Roop, M. Weaver, A. P. Fonseca, and M. Matouq, ‘Innovative Approaches in Smallholder Farming Systems to Implement the Sustainable Development Goals’, in *SDGs in the Americas and Caribbean Region*, W. Leal Filho, N. Aguilar-Rivera, B. Borsari, P. R. B. de Brito, and B. Andrade Guerra, Eds., Cham: Springer International Publishing, 2022, pp. 1–28. doi: 10.1007/978-3-030-91188-1_70-1.

[286] C. R. McGowan *et al.*, ‘Community-based surveillance of infectious diseases: a systematic review of drivers of success’, *BMJ Global Health*, vol. 7, no. 8, p. e009934, Aug. 2022, doi: 10.1136/bmjgh-2022-009934.

[287] J. Gaventa and R. McGee, ‘The Impact of Transparency and Accountability Initiatives’, *Development Policy Review*, vol. 31, no. s1, pp. s3–s28, 2013, doi: 10.1111/dpr.12017.

[288] T. Singh *et al.*, ‘Social Media as a Research Tool (SMaaRT) for Risky Behavior Analytics: Methodological Review’, *JMIR Public Health Surveill*, vol. 6, no. 4, p. e21660, Nov. 2020, doi: 10.2196/21660.

[289] L. Klerkx, ‘Digital and virtual spaces as sites of extension and advisory services research: social media, gaming, and digitally integrated and augmented advice’, *The Journal of Agricultural*

Education and Extension, vol. 27, no. 3, pp. 277–286, May 2021, doi: 10.1080/1389224X.2021.1934998.

[290] M. Ofori and O. El-Gayar, ‘Drivers and challenges of precision agriculture: a social media perspective’, *Precision Agric*, vol. 22, no. 3, pp. 1019–1044, Jun. 2021, doi: 10.1007/s11119-020-09760-0.

[291] ‘Journal of Medical Internet Research - Social Media Use for Health Purposes: Systematic Review’. Accessed: Feb. 08, 2023. [Online]. Available: https://www.jmir.org/2021/5/e17917/

[292] S. Rimadias, N. Alvionita, and A. P. Amelia, ‘Using Social Media Marketing to Create Brand Awareness, Brand Image, and Brand Loyalty on Tourism Sector in Indonesia’, *The Winners*, vol. 22, no. 2, Art. no. 2, Sep. 2021, doi: 10.21512/tw.v22i2.7597.

[293] S.-C. Chen and C.-P. Lin, ‘Understanding the effect of social media marketing activities: The mediation of social identification, perceived value, and satisfaction’, *Technological Forecasting and Social Change*, vol. 140, pp. 22–32, Mar. 2019, doi: 10.1016/j.techfore.2018.11.025.

[294] R. M. Smedley and N. S. Coulson, ‘A practical guide to analysing online support forums’, *Qualitative Research in Psychology*, vol. 18, no. 1, pp. 76–103, Jan. 2021, doi: 10.1080/14780887.2018.1475532.

[295] E. Turcan and K. McKeown, ‘Dreaddit: A Reddit Dataset for Stress Analysis in Social Media’. arXiv, Oct. 31, 2019. doi: 10.48550/arXiv.1911.00133.

[296] W. He, W. Zhang, X. Tian, R. Tao, and V. Akula, ‘Identifying customer knowledge on social media through data analytics’, *Journal of Enterprise Information Management*, vol. 32, no. 1, pp. 152–169, Jan. 2018, doi: 10.1108/JEIM-02-2018-0031.

[297] A. M. S. Osman, ‘A novel big data analytics framework for smart cities’, *Future Generation Computer Systems*, vol. 91, pp. 620–633, Feb. 2019, doi: 10.1016/j.future.2018.06.046.

[298] N. Mheidly and J. Fares, ‘Leveraging media and health communication strategies to overcome the COVID-19 infodemic’, *J Public Health Pol*, vol. 41, no. 4, pp. 410–420, Dec. 2020, doi: 10.1057/s41271-020-00247-w.

[299] C. M. Pulido, L. Ruiz-Eugenio, G. Redondo-Sama, and B. Villarejo-Carballido, ‘A New Application of Social Impact in Social Media for Overcoming Fake News in Health’, *International Journal of Environmental Research and Public Health*, vol. 17, no. 7, Art. no. 7, Jan. 2020, doi: 10.3390/ijerph17072430.

[300] M. Tizzani *et al.*, ‘Integrating digital and field surveillance as complementary efforts to manage epidemic diseases of livestock: African swine fever as a case study’, *PLOS ONE*, vol. 16, no. 12, p. e0252972, Dec. 2021, doi: 10.1371/journal.pone.0252972.

[301] D. Bradley, ‘Understanding farmer motivations: Very small and small farms’, 2021.

[302] T. Newsroom, ‘Half population dream of living the “good life” on a smallholding’, Farming Life. Accessed: Sep. 24, 2023. [Online]. Available: https://www.farminglife.com/business/half-population-dream-of-living-the-good-life-on-a-smallholding-136733

[303] C. Sauter-Louis *et al.*, ‘African Swine Fever in Wild Boar in Europe—A Review’, *Viruses*, vol. 13, no. 9, p. 1717, Aug. 2021, doi: 10.3390/v13091717.

[304] T. Porphyre, L. A. Boden, C. Correia-Gomes, H. K. Auty, G. J. Gunn, and M. E. Woolhouse, ‘How commercial and non-commercial swine producers move pigs in Scotland: a detailed descriptive analysis’, *BMC Veterinary Research*, vol. 10, no. 1, p. 140, Jun. 2014, doi: 10.1186/1746-6148-10-140.

[305] J. I. Alawneh *et al.*, ‘Description of the pig production systems, biosecurity practices and herd health providers in two provinces with high swine density in the Philippines’, *Preventive Veterinary Medicine*, vol. 114, no. 2, pp. 73–87, May 2014, doi: 10.1016/j.prevetmed.2014.01.020.

[306] C. Sibona, S. Walczak, and E. W. Baker, ‘A Guide for Purposive Sampling on Twitter’, *Communications of the Association for Information Systems*, vol. 46, no. 1, May 2020, doi: 10.17705/1CAIS.04622.

[307] F. B. Thomas, ‘The Role of Purposive Sampling Technique as a Tool for Informal Choices in a Social Sciences in Research Methods’, no. 5, 2022.

[308] P. Ranganathan and R. Aggarwal, ‘Study designs: Part 1 – An overview and classification’, *Perspectives in Clinical Research*, vol. 9, no. 4, p. 184, Dec. 2018, doi: 10.4103/picr.PICR_124_18.

[309] C. Kivunja and A. B. Kuyini, ‘Understanding and Applying Research Paradigms in Educational Contexts’, *IJHE*, vol. 6, no. 5, p. 26, Sep. 2017, doi: 10.5430/ijhe.v6n5p26.

[310] J. Schoonenboom and R. B. Johnson, ‘How to Construct a Mixed Methods Research Design’, *Kolner Z Soz Sozpsychol*, vol. 69, no. Suppl 2, pp. 107–131, 2017, doi: 10.1007/s11577-017-0454-1.

[311] J. Mao, ‘Social media for learning: A mixed methods study on high school students’ technology affordances and perspectives’, *Computers in Human Behavior*, vol. 33, pp. 213–223, Apr. 2014, doi: 10.1016/j.chb.2014.01.002.

[312] K. Slupinska, ‘Secondary Observation as a Method of Social Media Research: Theoretical Considerations and Implementation’. Accessed: Feb. 22, 2023. [Online]. Available: https://ideas.repec.org/a/ers/journl/vxxiiiy2020ispecial2p502-516.html

[313] M. Khder, ‘Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application’, *IJASCA*, vol. 13, no. 3, pp. 145–168, Dec. 2021, doi: 10.15849/IJASCA.211128.11.

[314] A. Campan, T. Atnafu, T. M. Truta, and J. Nolan, ‘Is Data Collection through Twitter Streaming API Useful for Academic Research?’, in *2018 IEEE International Conference on Big Data (Big Data)*, Dec. 2018, pp. 3638–3643. doi: 10.1109/BigData.2018.8621898.

[315] ‘Twitter API for Academic Research | Products’. Accessed: Feb. 24, 2023. [Online]. Available: https://developer.twitter.com/en/products/twitter-api/academic-research

[316] Y. Roh, G. Heo, and S. E. Whang, ‘A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective’, *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1328–1347, Apr. 2021, doi: 10.1109/TKDE.2019.2946162.

[317] B. Jang, I. Kim, and J. W. Kim, ‘Word2vec convolutional neural networks for classification of news articles and tweets’, *PLOS ONE*, vol. 14, no. 8, p. e0220976, Aug. 2019, doi: 10.1371/journal.pone.0220976.

[318] B. Shi, J. Zhao, and K. Xu, ‘A Word2vec Model for Sentiment Analysis of Weibo’, in *2019 16th International Conference on Service Systems and Service Management (ICSSSM)*, Jul. 2019, pp. 1–6. doi: 10.1109/ICSSSM.2019.8887652.

[319] M. García, S. Maldonado, and C. Vairetti, ‘Efficient n-gram construction for text categorization using feature selection techniques’, *Intelligent Data Analysis*, vol. 25, no. 3, pp. 509–525, Jan. 2021, doi: 10.3233/IDA-205154.

[320] I. E. Tiffani, ‘Optimization of Naïve Bayes Classifier By Implemented Unigram, Bigram, Trigram for Sentiment Analysis of Hotel Review’, *Journal of Soft Computing Exploration*, vol. 1, no. 1, Art. no. 1, Oct. 2020, doi: 10.52465/joscex.v1i1.4.

[321] U. Chauhan and A. Shah, ‘Topic Modeling Using Latent Dirichlet allocation: A Survey’, *ACM Comput. Surv.*, vol. 54, no. 7, p. 145:1-145:35, Sep. 2021, doi: 10.1145/3462478.

[322] P. Kherwa and P. Bansal, ‘Topic Modeling: A Comprehensive Review’, *EAI Endorsed Transactions on Scalable Information Systems*, vol. 7, no. 24, Jul. 2019, Accessed: Feb. 25, 2023. [Online]. Available: https://eudl.eu/doi/10.4108/eai.13-7-2018.159623

[323] L. Zhang, ‘Data And Content Analysis For Social Network Using LDA Text Model’, *J. Phys.: Conf. Ser.*, vol. 1213, no. 2, p. 022035, Jun. 2019, doi: 10.1088/1742-6596/1213/2/022035.

[324] W. Xing and Y. Bei, ‘Medical Health Big Data Classification Based on KNN Classification Algorithm’, *IEEE Access*, vol. 8, pp. 28808–28819, 2020, doi: 10.1109/ACCESS.2019.2955754.

[325] M. N. A. H. Sha’abani, N. Fuad, N. Jamal, and M. F. Ismail, ‘kNN and SVM Classification for EEG: A Review’, in *InECCE2019*, A. N. Kasruddin Nasir, M. A. Ahmad, M. S. Najib, Y. Abdul Wahab, N. A. Othman, N. M. Abd Ghani, A. Irawan, S. Khatun, R. M. T. Raja Ismail, M. M.

Saari, M. R. Daud, and A. A. Mohd Faudzi, Eds., in Lecture Notes in Electrical Engineering. Singapore: Springer, 2020, pp. 555–565. doi: 10.1007/978-981-15-2317-5_47.

[326] S. Xu, Y. Li, and Z. Wang, 'Bayesian Multinomial Naïve Bayes Classifier to Text Classification', in *Advanced Multimedia and Ubiquitous Engineering*, J. J. (Jong H. Park, S.-C. Chen, and K.-K. Raymond Choo, Eds., in Lecture Notes in Electrical Engineering. Singapore: Springer, 2017, pp. 347–352. doi: 10.1007/978-981-10-5041-1_57.

[327] M. Abbas, A. Kamran, Memon, A. A. Jamali, Saleemullah Memon, and Anees Ahmed, 'Multinomial Naive Bayes Classification Model for Sentiment Analysis', 2019, doi: 10.13140/RG.2.2.30021.40169.

[328] B. Charbuty and A. Abdulazeez, 'Classification Based on Decision Tree Algorithm for Machine Learning', *JASTT*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[329] J.-S. Lee, 'AUC4.5: AUC-Based C4.5 Decision Tree Algorithm for Imbalanced Data Classification', *IEEE Access*, vol. 7, pp. 106034–106042, 2019, doi: 10.1109/ACCESS.2019.2931865.

[330] K. Shah, H. Patel, D. Sanghvi, and M. Shah, 'A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification', *Augment Hum Res*, vol. 5, no. 1, p. 12, Mar. 2020, doi: 10.1007/s41133-020-00032-0.

[331] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, 'A comparison of random forest variable selection methods for classification prediction modeling', *Expert Systems with Applications*, vol. 134, pp. 93–101, Nov. 2019, doi: 10.1016/j.eswa.2019.05.028.

[332] Y. Wang, J. Liu, and L. Feng, 'Text length considered adaptive bagging ensemble learning algorithm for text classification', *Multimed Tools Appl*, Feb. 2023, doi: 10.1007/s11042-023-14578-9.

[333] M. R. Choirulfikri, Adiwijaya, and A. A. Suryani, 'Comparison of Bagging and Boosting in Imbalanced Multilabel of Al-Quran Dataset', in *2022 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS)*, Nov. 2022, pp. 01–05. doi: 10.1109/ICADEIS56544.2022.10037462.

[334] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, 'A comprehensive survey on support vector machine classification: Applications, challenges and trends', *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.

[335] A. I. Kadhim, 'Survey on supervised machine learning techniques for automatic text classification', *Artif Intell Rev*, vol. 52, no. 1, pp. 273–292, Jun. 2019, doi: 10.1007/s10462-018-09677-1.

[336] U. M. Sirisha, M. C. Belavagi, and G. Attigeri, 'Profit Prediction Using ARIMA, SARIMA and LSTM Models in Time Series Forecasting: A Comparison', *IEEE Access*, vol. 10, pp. 124715–124727, 2022, doi: 10.1109/ACCESS.2022.3224938.

[337] T. C. Nokeri, 'Forecasting Using ARIMA, SARIMA, and the Additive Model', in *Implementing Machine Learning for Finance: A Systematic Approach to Predictive Risk and Performance Analysis for Investment Portfolios*, T. C. Nokeri, Ed., Berkeley, CA: Apress, 2021, pp. 21–50. doi: 10.1007/978-1-4842-7110-0_2.

[338] P. Sedgwick, 'Pearson's correlation coefficient', *BMJ*, vol. 345, p. e4483, Jul. 2012, doi: 10.1136/bmj.e4483.

[339] R. A. Armstrong, 'Should Pearson's correlation coefficient be avoided?', *Ophthalmic and Physiological Optics*, vol. 39, no. 5, pp. 316–327, 2019, doi: 10.1111/opo.12636.

[340] D. Xu, Y. Wang, Y. Meng, and Z. Zhang, 'An Improved Data Anomaly Detection Method Based on Isolation Forest', in *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, Dec. 2017, pp. 287–291. doi: 10.1109/ISCID.2017.202.

[341] J. Lesouple, C. Baudoin, M. Spigai, and J.-Y. Tourneret, 'Generalized isolation forest for anomaly detection', *Pattern Recognition Letters*, vol. 149, pp. 109–119, Sep. 2021, doi: 10.1016/j.patrec.2021.05.022.

[342] L. Anselin, ‘Global Spatial Autocorrelation (2)’. Accessed: May 01, 2023. [Online]. Available: https://geodacenter.github.io/workbook/5b_global_adv/lab5b.html

[343] H. Zhang, L. Yang, L. Li, G. Xu, and X. Zhang, ‘The epidemic characteristics and spatial autocorrelation analysis of hand, foot and mouth disease from 2010 to 2015 in Shantou, Guangdong, China’, *BMC Public Health*, vol. 19, no. 1, p. 998, Jul. 2019, doi: 10.1186/s12889-019-7329-5.

[344] D. Adham, E. Moradi-Asl, A. Dorosti, and S. Khaiatzadeh, ‘Spatial autocorrelation and epidemiological survey of visceral leishmaniasis in an endemic area of Azerbaijan region, the northwest of Iran’, *PLOS ONE*, vol. 15, no. 8, p. e0236414, Aug. 2020, doi: 10.1371/journal.pone.0236414.

[345] J. S. P. Tulloch, R. Vivancos, R. M. Christley, A. D. Radford, and J. C. Warner, ‘Mapping tweets to a known disease epidemiology; a case study of Lyme disease in the United Kingdom and Republic of Ireland’, *Journal of Biomedical Informatics*, vol. 100, p. 100060, Jan. 2019, doi: 10.1016/j.yjbinx.2019.100060.

[346] I. Blekanov, S. S. Bodrunova, and A. Akhmetov, ‘Detection of Hidden Communities in Twitter Discussions of Varying Volumes’, *Future Internet*, vol. 13, no. 11, Art. no. 11, Nov. 2021, doi: 10.3390/fi13110295.

[347] S. Shirazi, H. Baziyad, N. Ahmadi, and A. Albadvi, ‘A New Application of Louvain Algorithm for Identifying Disease Fields Using Big Data Techniques’, *Journal of Biostatistics and Epidemiology*, vol. 5, no. 3, pp. 183–193, 2019, doi: 10.18502/jbe.v5i3.3613.

[348] S. Mukerjee, ‘A systematic comparison of community detection algorithms for measuring selective exposure in co-exposure networks’, *Sci Rep*, vol. 11, no. 1, Art. no. 1, Jul. 2021, doi: 10.1038/s41598-021-94724-1.

[349] R. Heale and A. Twycross, ‘Validity and reliability in quantitative studies’, *Evidence-Based Nursing*, vol. 18, no. 3, pp. 66–67, Jul. 2015, doi: 10.1136/eb-2015-102129.

[350] K. Jordan, ‘Validity, reliability and the case for participant-centred research: Reflections on a multi-platform social media study’, *International Journal of Human-Computer Interaction*, vol. 34, no. 10, Art. no. 10, 2018.

[351] N. A. Khan, M. Azhar, M. N. Rahman, and M. J. Akhtar, ‘Scale development and validation for usage of social networking sites during COVID-19’, *Technol Soc*, vol. 70, p. 102020, Aug. 2022, doi: 10.1016/j.techsoc.2022.102020.

[352] C. O’Connor and H. Joffe, ‘Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines’, *International Journal of Qualitative Methods*, vol. 19, p. 1609406919899220, Jan. 2020, doi: 10.1177/1609406919899220.

[353] M. L. McHugh, ‘Interrater reliability: the kappa statistic’, *Biochem Med (Zagreb)*, vol. 22, no. 3, pp. 276–282, Oct. 2012.

[354] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, ‘Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data’, *Ecological Modelling*, vol. 406, pp. 109–120, Aug. 2019, doi: 10.1016/j.ecolmodel.2019.06.002.

[355] ‘Classification: ROC Curve and AUC | Machine Learning’, Google Developers. Accessed: Jan. 04, 2023. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

[356] T. R. Nichols, P. M. Wisner, G. Cripe, and L. Gulabchand, ‘Putting the Kappa Statistic to Use: Putting the Kappa Statistic to Use’, *Qual Assur J*, vol. 13, no. 3–4, pp. 57–61, Jul. 2010, doi: 10.1002/qaj.481.

[357] K. Krippendorff, ‘Computing Krippendorff’s Alpha-Reliability’, *Departmental Papers (ASC)*, Jan. 2011, [Online]. Available: https://repository.upenn.edu/asc_papers/43

[358] K. Wombacher, *The SAGE Encyclopedia of Communication Research Methods*. SAGE Publications, Inc, 2017. doi: 10.4135/9781483381411.

[359] K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, and S. Nerur, ‘Advances in Social Media Research: Past, Present and Future’, *Inf Syst Front*, vol. 20, no. 3, pp. 531–558, Jun. 2018, doi: 10.1007/s10796-017-9810-y.

[360] R. Sinha and J. K. Madsen, ‘Driving behavior change among farmers & fishers: Do social networks matter?’, *PLOS Water*, vol. 2, no. 2, p. e0000095, Feb. 2023, doi: 10.1371/journal.pwat.0000095.

[361] R. Fernández-Peña, M.-A. Ovalle-Perandones, P. Marqués-Sánchez, C. Ortego-Maté, and N. Serrano-Fuentes, ‘The use of social network analysis in social support and care: a systematic scoping review protocol’, *Systematic Reviews*, vol. 11, no. 1, p. 9, Jan. 2022, doi: 10.1186/s13643-021-01876-2.

[362] A. Russo, V. Miracula, and A. Picone, ‘Network analysis on political election; populist vs social emergent behaviour’, arXiv.org. Accessed: May 08, 2023. [Online]. Available: https://arxiv.org/abs/2301.05668v1

[363] J. Xhema, ‘Effect of Social Networks on Consumer Behaviour: Complex Buying’, *IFAC-PapersOnLine*, vol. 52, no. 25, pp. 504–508, Jan. 2019, doi: 10.1016/j.ifacol.2019.12.594.

[364] M. R. Sarker, M. V. Galdos, A. J. Challinor, and A. Hossain, ‘A farming system typology for the adoption of new technology in Bangladesh’, *Food and Energy Security*, vol. 10, no. 3, p. e287, 2021, doi: 10.1002/fes3.287.

[365] M. Andreotta *et al.*, ‘Analyzing social media data: A mixed-methods framework combining computational and qualitative text analysis’, *Behav Res*, vol. 51, no. 4, pp. 1766–1781, Aug. 2019, doi: 10.3758/s13428-019-01202-8.

[366] G. Cusworth and J. Dodsworth, ‘Using the “good farmer” concept to explore agricultural attitudes to the provision of public goods. A case study of participants in an English agri-environment scheme’, *Agric Hum Values*, vol. 38, no. 4, pp. 929–941, Dec. 2021, doi: 10.1007/s10460-021-10215-z.

[367] C. M. Godde, D. Mason-D’Croz, D. E. Mayberry, P. K. Thornton, and M. Herrero, ‘Impacts of climate change on the livestock food supply chain; a review of the evidence’, *Glob Food Sec*, vol. 28, p. 100488, Mar. 2021, doi: 10.1016/j.gfs.2020.100488.

[368] P. Fronczak, ‘Scale-Free Nature of Social Networks’, in *Encyclopedia of Social Network Analysis and Mining*, R. Alhajj and J. Rokne, Eds., New York, NY: Springer, 2018, pp. 2300–2309. doi: 10.1007/978-1-4939-7131-2_248.

[369] ‘Register land you use to keep livestock’, GOV.UK. Accessed: Dec. 27, 2022. [Online]. Available: https://www.gov.uk/guidance/register-land-you-use-to-keep-livestock

[370] N. Widmar, C. Bir, J. Lai, and C. Wolf, ‘Public Perceptions of Veterinarians from Social and Online Media Listening’, *Vet Sci*, vol. 7, no. 2, p. 75, Jun. 2020, doi: 10.3390/vetsci7020075.

[371] R. Egger and J. Yu, ‘A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts’, *Front Sociol*, vol. 7, p. 886498, May 2022, doi: 10.3389/fsoc.2022.886498.

[372] J. Gan and Y. Qi, ‘Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an Example’, *Entropy (Basel)*, vol. 23, no. 10, p. 1301, Oct. 2021, doi: 10.3390/e23101301.

[373] E. Zvornicanin, ‘When Coherence Score Is Good or Bad in Topic Modeling? | Baeldung on Computer Science’. Accessed: Jun. 30, 2023. [Online]. Available: https://www.baeldung.com/cs/topic-modeling-coherence-score

[374] K. Kumar, ‘Evaluation of Topic Modeling: Topic Coherence | DataScience+’. Accessed: Jul. 12, 2023. [Online]. Available: https://datascienceplus.com/evaluation-of-topic-modeling-topic-coherence/

[375] A. Amalraj *et al.*, ‘Health and management of hobby pigs : a review’, *VLAAMS DIERGENEESKUNDIG TIJDSCHRIFT*, vol. 87, no. 6, Art. no. 6, 2018.

[376] Agricultural and Rural economy directorate, ‘Livestock identification and traceability: guidance’. Accessed: Jul. 06, 2023. [Online]. Available: http://www.gov.scot/publications/livestock-identification-and-traceability-guidance/

[377] C. Abel, ‘Pig Maturity Age For Breeding And Slaughter’, Savvy Farm Life. Accessed: Jul. 06, 2023. [Online]. Available: https://savvyfarmlife.com/pig-maturity-age/

[378] EFSA Panel on Animal Health and Welfare (AHAW) *et al.*, ‘African swine fever and outdoor farming of pigs’, *EFSA Journal*, vol. 19, no. 6, p. e06639, 2021, doi: 10.2903/j.efsa.2021.6639.

[379] B. Harlizius, P. Mathur, and E. F. Knol, ‘Breeding for resilience: new opportunities in a modern pig breeding program’, *Journal of Animal Science*, vol. 98, no. Supplement_1, pp. S150–S154, Aug. 2020, doi: 10.1093/jas/skaa141.

[380] H. J. Bray and R. A. Ankeny, ‘Happy Chickens Lay Tastier Eggs: Motivations for Buying Free-range Eggs in Australia’, *Anthrozoös*, vol. 30, no. 2, pp. 213–226, Apr. 2017, doi: 10.1080/08927936.2017.1310986.

[381] N. K. Sakomura, M. D. P. Reis, N. T. Ferreira, and R. M. Gous, ‘Modeling egg production as a means of optimizing dietary nutrient contents for laying hens’, *Animal Frontiers*, vol. 9, no. 2, pp. 45–51, Apr. 2019, doi: 10.1093/af/vfz010.

[382] D. Temple *et al.*, ‘Assessment of laying-bird welfare following acaricidal treatment of a commercial flock naturally infested with the poultry red mite (Dermanyssus gallinae)’, *PLOS ONE*, vol. 15, no. 11, p. e0241608, Nov. 2020, doi: 10.1371/journal.pone.0241608.

[383] K. M. Hartcher and B. Jones, ‘The welfare of layer hens in cage and cage-free housing systems’, *World’s Poultry Science Journal*, vol. 73, no. 4, pp. 767–782, Dec. 2017, doi: 10.1017/S0043933917000812.

[384] APHA, ‘Livestock Demographic Data Group: Pig population report – 2023’. 2024. [Online]. Available: http://apha.defra.gov.uk/documents/surveillance/diseases/lddg-pig-pop-report-23.pdf

[385] APHA, ‘Poultry - Livestock population density maps for GB, using July 2022 data’, 2024, [Online]. Available: http://apha.defra.gov.uk/documents/surveillance/diseases/lddg-pop-report-avian2022.pdf

[386] P.-J. M. Noble, C. Appleton, A. D. Radford, and G. Nenadic, ‘Using topic modelling for unsupervised annotation of electronic health records to identify an outbreak of disease in UK dogs’, *PLoS ONE*, vol. 16, no. 12, p. e0260402, Dec. 2021, doi: 10.1371/journal.pone.0260402.

[387] J. Botz *et al.*, ‘Modeling approaches for early warning and monitoring of pandemic situations as well as decision support’, *Frontiers in Public Health*, vol. 10, 2022, Accessed: Jul. 06, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2022.994949

[388] P. U. Eze, N. Geard, I. Mueller, and I. Chades, ‘Anomaly Detection in Endemic Disease Surveillance Data Using Machine Learning Techniques’, *Healthcare*, vol. 11, no. 13, Art. no. 13, Jan. 2023, doi: 10.3390/healthcare11131896.

[389] S. Kevany, ‘Avian flu has led to the killing of 140m farmed birds since last October’, *The Guardian*, Dec. 09, 2022. Accessed: Sep. 28, 2023. [Online]. Available: https://www.theguardian.com/environment/2022/dec/09/avian-flu-has-led-to-the-killing-of-140m-farmed-birds-since-last-october

[390] L. Bisdounis, ‘Bird flu 2022: Dealing with the UK’s largest ever outbreak’, Nov. 2022, Accessed: Dec. 27, 2022. [Online]. Available: https://lordslibrary.parliament.uk/bird-flu-2022-dealing-with-the-uks-largest-ever-outbreak/

[391] ‘Bird flu: rules in disease control and prevention zones in England’, GOV.UK. Accessed: Dec. 27, 2022. [Online]. Available: https://www.gov.uk/guidance/avian-influenza-bird-flu-cases-and-disease-control-zones-in-england

[392] CDC, ‘Avian Influenza A Virus Infections in Humans’, Centers for Disease Control and Prevention. Accessed: Dec. 27, 2022. [Online]. Available: https://www.cdc.gov/flu/avianflu/avian-in-humans.htm

[393] M. Jonges, J. van Leuken, I. Wouters, G. Koch, A. Meijer, and M. Koopmans, ‘Wind-Mediated Spread of Low-Pathogenic Avian Influenza Virus into the Environment during Outbreaks at Commercial Poultry Farms’, *PLOS ONE*, vol. 10, no. 5, p. e0125401, May 2015, doi: 10.1371/journal.pone.0125401.

[394] J. Artois *et al.*, ‘Avian influenza A (H5N1) outbreaks in different poultry farm types in Egypt: the effect of vaccination, closing status and farm size’, *BMC Veterinary Research*, vol. 14, no. 1, p. 187, Jun. 2018, doi: 10.1186/s12917-018-1519-8.

[395] R. Lamsal, A. Harwood, and M. R. Read, ‘Twitter conversations predict the daily confirmed COVID-19 cases’, *Applied Soft Computing*, vol. 129, p. 109603, Nov. 2022, doi: 10.1016/j.asoc.2022.109603.

[396] E. Baumohl and S. Lyocsa, ‘Stationarity of Time Series and the Problem of Spurious Regression’, *SSRN Journal*, 2009, doi: 10.2139/ssrn.1480682.

[397] L. Lebastard and R. Serafini, ‘Understanding the impact of COVID-19 supply disruptions on exporters in global value chains’, no. 105, Mar. 2023, Accessed: Sep. 28, 2023. [Online]. Available: https://www.ecb.europa.eu/pub/economic-research/resbull/2023/html/ecb.rb230322~5c08629152.en.html

[398] J. Zhan, ‘The Relevance of Compassion Fatigue in Social Media Discourse on the Russia-Ukraine Crisis’, presented at the 2022 5th International Conference on Humanities Education and Social Sciences (ICHESS 2022), Atlantis Press, 2022, pp. 298–311. doi: 10.2991/978-2-494069-89-3_35.

[399] K. W. Ng, F. Mubang, L. O. Hall, J. Skvoretz, and A. Iamnitchi, ‘Experimental evaluation of baselines for forecasting social media timeseries’, *EPJ Data Sci.*, vol. 12, no. 1, Art. no. 1, Dec. 2023, doi: 10.1140/epjds/s13688-023-00383-9.

[400] A. Bouteska, P. Hajek, M. Z. Abedin, and Y. Dong, ‘Effect of twitter investor engagement on cryptocurrencies during the COVID-19 pandemic’, *Research in International Business and Finance*, vol. 64, p. 101850, Jan. 2023, doi: 10.1016/j.ribaf.2022.101850.

[401] B. Zareie, J. Poorolajal, A. Roshani, and M. Karami, ‘Outbreak detection algorithms based on generalized linear model: a review with new practical examples’, *BMC Med Res Methodol*, vol. 23, p. 235, Oct. 2023, doi: 10.1186/s12874-023-02050-z.

[402] APHA, ‘Bird flu (avian influenza): how to spot and report it in poultry or other captive birds’, GOV.UK. Accessed: May 28, 2024. [Online]. Available: https://www.gov.uk/guidance/avian-influenza-bird-flu

[403] J. M. Lane, D. Habib, and B. Curtis, ‘Linguistic Methodologies to Surveil the Leading Causes of Mortality: Scoping Review of Twitter for Public Health Data’, *J Med Internet Res*, vol. 25, p. e39484, Jun. 2023, doi: 10.2196/39484.

[404] A.-R. Abdulai, ‘Toward digitalization futures in smallholder farming systems in Sub-Sahara Africa: A social practice proposal’, *Frontiers in Sustainable Food Systems*, vol. 6, 2022, Accessed: Aug. 23, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fsufs.2022.866331

[405] Central Digital Data Office, ‘UK National Action Plan for Open Government 2021-2023’, GOV.UK. Accessed: May 29, 2024. [Online]. Available: https://www.gov.uk/government/publications/uk-national-action-plan-for-open-government-2021-2023/uk-national-action-plan-for-open-government-2021-2023

[406] J. Baker, ‘The Technology–Organization–Environment Framework’, in *Information Systems Theory: Explaining and Predicting Our Digital Society, Vol. 1*, Y. K. Dwivedi, M. R. Wade, and S. L. Schneberger, Eds., New York, NY: Springer, 2012, pp. 231–245. doi: 10.1007/978-1-4419-6108-2_12.

[407] K. Roystonn *et al.*, ‘Exploring views and experiences of the general public’s adoption of digital technologies for healthy lifestyle in Singapore: a qualitative study’, *Front Public Health*, vol. 11, p. 1227146, Sep. 2023, doi: 10.3389/fpubh.2023.1227146.

[408] S. C. Rhodes, 'Filter Bubbles, Echo Chambers, and Fake News: How Social Media Conditions Individuals to Be Less Critical of Political Misinformation', *Political Communication*, vol. 39, no. 1, pp. 1–22, Jan. 2022, doi: 10.1080/10584609.2021.1910887.

[409] K. Shu, S. Wang, D. Lee, and H. Liu, Eds., *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*. in Lecture Notes in Social Networks. Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-42699-6.

[410] M. C. Wagner and P. J. Boczkowski, 'The Reception of Fake News: The Interpretations and Practices That Shape the Consumption of Perceived Misinformation', *Digital Journalism*, vol. 7, no. 7, pp. 870–885, Aug. 2019, doi: 10.1080/21670811.2019.1653208.

[411] M. Cinelli, 'The echo chamber effect on social media | PNAS', PNAS. Accessed: Aug. 12, 2023. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2023301118

[412] S. Hakak, W. Z. Khan, S. Bhattacharya, G. T. Reddy, and K.-K. R. Choo, 'Propagation of Fake News on Social Media: Challenges and Opportunities', in *Computational Data and Social Networks*, S. Chellappan, K.-K. R. Choo, and N. Phan, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 345–353. doi: 10.1007/978-3-030-66046-8_28.

[413] M. Donadeu, N. Nwankpa, B. Abela-Ridder, and B. Dungu, 'Strategies to increase adoption of animal vaccines by smallholder farmers with focus on neglected diseases and marginalized populations', *PLoS Negl Trop Dis*, vol. 13, no. 2, p. e0006989, Feb. 2019, doi: 10.1371/journal.pntd.0006989.

[414] Q. G. To *et al.*, 'Anti-vaccination attitude trends during the COVID-19 pandemic: A machine learning-based analysis of tweets', *Digit Health*, vol. 9, p. 20552076231158033, Feb. 2023, doi: 10.1177/20552076231158033.

[415] 'Threads', Threads. Accessed: Aug. 20, 2023. [Online]. Available: https://www.threads.net/

[416] R. Eeswaran, A. P. Nejadhashemi, A. Faye, D. Min, P. V. V. Prasad, and I. A. Ciampitti, 'Current and Future Challenges and Opportunities for Livestock Farming in West Africa: Perspectives from the Case of Senegal', *Agronomy*, vol. 12, no. 8, Art. no. 8, Aug. 2022, doi: 10.3390/agronomy12081818.

Appendix A – Custom stop words list

Pigs –

'pigs','need','much','like','could','thanks','also','think','first','one','two','time','re','pig','hi', 'hello', 'want', 'obtain', 'look', 'hii','ive','got','use','number','get','would','day','one','good','year','thought','year'

Poultry –

'poultry','chicken','need','much','like','could','thanks','also','think','first','one','two','time','re','chickens','hi', 'hello', 'want', 'obtain', 'look', 'hii','ive','got','use','number','get','would','day','one','good','year','thought','year'

Appendix B – Avian flu timeline

SOURCE: HTTPS://WWW.GOV.UK/GUIDANCE/AVIAN-INFLUENZA-BIRD-FLU#FULL-PUBLICATION-UPDATE-HISTORY

Date	Event
31 December 2021	Highly pathogenic avian influenza (HPAI) H5N1 was confirmed in captive birds at a premises near Eton, Windsor & Maidenhead, Berkshire. A 3km Captive Bird (Monitoring) Controlled Zone has been put in place around the premises.
31 December 2021	The 3km Protection Zones declared around premises near Leeming Bar, Hambleton, North Yorkshire and second premises near Leeming Bar, Hambleton, North Yorkshire have ended and the areas that formed the Protection Zone now form part of the Surveillance Zone.
30 December 2021	Highly pathogenic avian influenza (HPAI) H5N1 was confirmed in birds at premises near Mablethorpe, East Lindsey, Lincolnshire and near Louth, East Lindsey, Lincolnshire. A 3km Protection Zone and 10km Surveillance Zone has been put in place around each of the premises.
29 December 2021	Highly pathogenic avian influenza (HPAI) H5N1 was confirmed in a non-commercial flock of birds at a premises near Romsey, Test Valley, Hampshire. A 3km Protection Zone and 10km Surveillance Zone has been put in place around the premises.
28 December 2021	The Surveillance Zones around premises near Kirkham, Fylde, Lancashire and near Salwick, Fylde, Lancashire have been revoked.
24 December 2021	Highly pathogenic avian influenza (HPAI) H5N1 was confirmed in commercial poultry at a premises near North Somercotes, East Lindsey, Lincolnshire and near Watlington, King's Lynn and West Norfolk, Norfolk. A 3km Protection Zone and 10km Surveillance Zone has been put in place around each of the premises.
24 December 2021	3km Protection Zone has ended and the 10km Surveillance Zone surrounding a premises at Wells-next-the-Sea, Norfolk has been revoked.
22 December 2021	Updated biosecurity posters for chicken, turkey, duck, geese and game bird keepers.
20 December 2021	Highly pathogenic avian influenza was confirmed in non-

	commercial birds at a second premises near Helsby, Cheshire and in captive birds (non-poultry) at a premises near Near Buckfastleigh, Devon.
20 December 2021	Updated to reflect a revocation of the Surveillance Zone near Pokesdown, Bournemouth, Christchurch and Poole.
18 December 2021	Following successful completion of disease control activities and surveillance within the zones, the 3km Protection Zone surrounding a premises near Pokesdown, Bournemouth, Christchurch and Poole ended on 20 December 2021. The area merged with the 10km Surveillance Zone surrounding this premises
17 December 2021	Highly pathogenic avian influenza (HPAI) H5N1 has been confirmed in poultry at a ninth premises near Alford, East Lindsey, Lincolnshire and at a premises near Helsby, Cheshire West & Chester, Cheshire. All birds on the infected premises will be humanely culled. A 3km Protection Zone and 10km Surveillance Zone has been put in place around each of the premises.
16 December 2021	Highly pathogenic avian influenza (HPAI) H5N1 was confirmed in poultry at an eighth premises near Alford, East Lindsey, Lincolnshire. Protection and Surveillance zones have been put in place. In a different case, avian influenza H5N1 was confirmed in birds at a premises near Frinton-on-Sea, Tendring, Essex on the 11 November 2021. Following successful completion of disease control activities and surveillance within the zones, the 3km Protection Zone has ended and 10km Surveillance Zone declared on 12 December 2021 has now been revoked. Local movement restrictions have now been removed but the Avian Influenza Prevention Zone (AIPZ) measures remain in place.
16 December 2021	Highly pathogenic avian influenza (HPAI) H5N1 was confirmed in poultry at a sixth premises near Alford, East Lindsey, Lincolnshire, at a ninth premises near Thirsk, Hambleton, North Yorkshire, and at a seventh premises near Alford, East Lindsey, Lincolnshire. A 3km Protection Zone and 10km Surveillance Zone has been put in place around each of the premises.
15 December 2021	Highly pathogenic avian influenza (HPAI) H5N1 was confirmed in birds

	near Market Bosworth, Hinckley and Bosworth, Leicestershire. A 3km Protection Zone and 10km Surveillance Zone has been put in place around the premises.
14 December 2021	Highly pathogenic avian influenza (HPAI) H5N1 was confirmed in poultry near Wem, North Shropshire, Shropshire. A 3km Protection Zone and 10km Surveillance Zone have been put in place around each of the premises.
14 December 2021	Highly pathogenic avian influenza (HPAI) H5N1 was confirmed in birds at a fourth and fifth premises near Alford, East Lindsey, Lincolnshire and at a second premises near Pocklington, East Yorkshire, East Riding of Yorkshire. A 3km Protection Zone and 10km Surveillance Zone has been put in place around the premises.
14 December 2021	How to register for Defra's 'stop the spread' webinars.
12 December 2021	Highly pathogenic avian influenza (HPAI) H5N1 was confirmed in poultry at a premises near Middleton-in-Teesdale, County Durham, Durham. A 3km Protection Zone and 10km Surveillance Zone has been put in place around the premises.
11 December 2021	Highly pathogenic avian influenza (HPAI) H5N1 confirmed in birds at second and third premises near Alford, East Lindsey, Lincolnshire.
10 December 2021	Highly pathogenic avian influenza (HPAI) H5N1 confirmed in birds at second premises near Willington, South Derbyshire, Derbyshire and at a premises near Alford, East Lindsey, Lincolnshire. Avian influenza H5N1 confirmed in birds at a premises near Washington, Sunderland, Tyne & Wear.
9 December 2021	Highly pathogenic avian influenza (HPAI) H5N1 confirmed in birds at premises near Clifford, Hereford and South Herefordshire, Herefordshire and at premises near Highworth, Swindon, Wiltshire.
8 December 2021	Highly pathogenic avian influenza (HPAI) H5N1 was confirmed in birds at a premises near Aspatria, Allerdale, Cumbria and a premises near Annan, Dumfriesshire, Dumfries and Galloway, Scotland. A 3km Protection Zone and 10km Surveillance Zone have been put in place around each of the premises.
7 December 2021	Highly pathogenic avian influenza confirmed in birds at an eighth premises near Thirsk, Hambleton, North Yorkshire and at a

	premises near Sudbury, Babergh, South Suffolk.
6 December 2021	Highly pathogenic avian influenza confirmed in birds at premises near Pocklington, East Yorkshire, at a fourth premises near Barrow upon Soar, Charnwood, Leicestershire, premises in Aughnacloy, County Tyrone and premises in Broughshane, County Antrim.
5 December 2021	Highly pathogenic avian influenza confirmed in poultry at a third premises near Barrow upon Soar, Charnwood, Leicestershire.
3 December 2021	Highly pathogenic avian influenza confirmed in birds at a seventh premises near Thirsk, Hambleton, North Yorkshire and at a premises near Newent, Forest of Dean, Gloucestershire.
3 December 2021	Highly pathogenic avian influenza (HPAI) H5N1 was confirmed in birds at premises near Richmond, Richmondshire, North Yorkshire, England, near Crickhowell, Powys, Wales and near Gretna, Dumfriesshire, Dumfries and Galloway, Scotland.
2 December 2021	Revocation of the Surveillance Zone near Chirk, Wrexham, Wales.
30 November 2021	Updated to reflect that avian influenza H5N1 has been confirmed in birds at a premises near Leominster, North Herefordshire, Herefordshire; a premises near Tutbury, East Staffordshire, Staffordshire and a sixth premises near Thirsk, Hambleton, North Yorkshire. A 3km Protection Zone and 10km Surveillance Zone has been put in place around each of the premises.
29 November 2021	Updated to reflect that highly pathogenic avian influenza (HPAI) H5N1 was confirmed in birds at a second premises near Barrow upon Soar, Charnwood, Leicestershire on the 30 November 2021. A 3km Protection Zone and 10km Surveillance Zone has been put in place around the premises.
28 November 2021	Updated to reflect that housing measures have come into force across the UK.
27 November 2021	Updated to reflect that highly pathogenic avian influenza (HPAI) H5N1 was confirmed in commercial poultry at a fifth premises near Thirsk, Hambleton, North Yorkshire on the 28 November 2021.
26 November 2021	Following completion of disease control activities and additional surveillance at the premises near Droitwich Spa, Wychavon, Worcestershire, the Captive Bird (Monitoring) Controlled Zone B has been revoked. All other

	restrictions, including Avian Influenza Prevention Zone remain in force.
26 November 2021	Updated to reflect that Highly pathogenic avian influenza (HPAI) H5N1 has been confirmed in birds at a fourth premises near Thirsk, Hambleton, North Yorkshire. A 3km Protection Zone and 10km Surveillance Zone has been put in place around the premises.
26 November 2021	Further testing has confirmed that the avian influenza strain in birds at the premises near Thirsk, Hambleton, North Yorkshire as highly pathogenic (HPAI H5N1). The Temporary Control Zones have been revoked and replaced by 3km Protection Zone and 10km Surveillance Zone around the premises.
26 November 2021	Updated to reflect that following completion of disease control activities and surveillance within the zones, the 3km Protection Zone declared surrounding a premises near Chirk, Wrexham, Wales was revoked on 26 November 2021 and the areas merged with the 10km Surveillance Zone for the relevant areas of Wales and England.
25 November 2021	Updated to reflect that Avian influenza H5N1 has been confirmed in birds at a premises near Clitheroe, Ribble Valley, Lancashire. Further testing is underway to confirm the pathogenicity of the strain in this case. 3km and 10km Temporary Control Zones have been put in place around the premises.
25 November 2021	Updated to reflect that avian influenza H5N1 has been confirmed in birds at a third premises near Thirsk, Hambleton, North Yorkshire and at a premises near Poulton le Flyde, Wyre, Lancashire.
25 November 2021	Updated the biosecurity guidance document and biosecurity self-assessment checklist.
24 November 2021	Updated to reflect that Avian influenza H5N1 has been confirmed in birds at a premises near Barrow upon Soar, Charnwood, Leicestershire. Further testing is underway to confirm the pathogenicity of the strain. 3km and 10km Temporary Control Zones have been put in place around the premises.
24 November 2021	Added further details of the measures that will apply in the AIPZ in England from 00:01 on 29 November 2021.
24 November 2021	Updated to reflect that new housing measures will come into force at 00:01 on Monday 29 November 2021. This

	means that it will be a legal requirement for all bird keepers across the UK to keep their birds indoors and to follow strict biosecurity measures in order to limit the spread of the disease.
22 November 2021	Further testing has confirmed highly pathogenic strain (HPAI H5N1) in birds at 3 premises: 2 premises near Thirsk, Hambleton, North Yorkshire and a second premises near Leeming Bar, Hambleton, North Yorkshire. 3km Protection Zone and 10km Surveillance Zone have been declared around each of the premises.
21 November 2021	Updated to reflect that avian influenza H5N1 was confirmed in birds at a second premises near Thirsk, Hambleton, North Yorkshire and at a second premises near Leeming Bar, Hambleton, North Yorkshire. 3km and 10km Temporary Control Zones have been put in place around each of the premises. Further testing has also confirmed that avian influenza in birds at a premises near Wells-next-the-sea, North Norfolk, Norfolk is a highly pathogenic strain (HPAI H5N1). 3km Protection Zone and 10km Surveillance Zone have been declared around the premises.
20 November 2021	Updated latest situation: avian influenza H5N1 was confirmed in birds at a premises near North Fambridge, Maldon, Essex. 3km and 10km Temporary Control Zones have been put in place around the premises. Further testing has also confirmed that avian influenza in birds at a premises near Silecroft, Copeland, Cumbria is a highly pathogenic strain (HPAI H5N1). 3km Protection Zone and 10km Surveillance Zone have been declared around the premises.
19 November 2021	Updated to reflect that avian influenza H5N1 has been confirmed in birds at a premises near Mouldsworth, Cheshire West and Chester, Cheshire. 3km and 10km Temporary Control Zones have been put in place around the premises.
19 November 2021	Updated to reflect that avian influenza H5N1 was confirmed in birds at a premises near Silecroft, Copeland, Cumbria. 3km and 10km Temporary Control Zones have been put in place around the premises. Further testing has also confirmed that avian influenza in birds at a premises near Pokesdown, Bournemouth,

	Christchurch and Poole is a highly pathogenic strain (HPAI H5N1). 3km Protection Zone and 10km Surveillance Zone have been declared around the premises.
17 November 2021	Updated to reflect that avian influenza H5N1 has been confirmed in birds at a premises near Pokesdown, Bournemouth, Christchurch and Poole. 3km and 10km Temporary Control Zones have been put in place around the premises. Further testing has also confirmed that avian influenza in birds at a premises near Willington, South Derbyshire, Derbyshire is a highly pathogenic strain (HPAI H5N1). The Temporary Control Zones have been revoked and replaced by a 3km Protection Zone and 10km Surveillance Zone.
16 November 2021	Updated to reflect that further testing has confirmed highly pathogenic strain (HPAI H5N1) in birds at a premises near Kirkham, Fylde, Lancashire. 3km Protection Zone and 10km Surveillance Zone have been declared around the premises. Avian influenza H5N1 was also confirmed in birds at a premises near Willington, South Derbyshire, Derbyshire. 3km and 10km Temporary Control Zones have been put in place around the premises.
15 November 2021	Updated to reflect that avian influenza H5N1 has been confirmed in birds at a premises near Kirkham, Fylde, Lancashire. Further testing is underway to confirm the pathogenicity of the strain. 3km and 10km Temporary Control Zones have been put in place surrounding the premises.
14 November 2021	Added link to avian influenza and game birds guidance on the Game Farmers Association website.
13 November 2021	Updated to reflect that further testing has confirmed highly pathogenic strain (HPAI H5N1) at premises near Leeming Bar, Hambleton, North Yorkshire and premises near Salwick, Fylde, Lancashire. 3km Protection Zone and 10km Surveillance Zones have been put in place surrounding these premises. Issued general licence for the movement of samples for salmonella testing from premises in the Protection Zone or Surveillance Zone.
12 November 2021	Updated to reflect that avian influenza H5N1 has been confirmed in birds at a premises near Leeming Bar, Hambleton, North Yorkshire, England.

	Further testing is underway to confirm the pathogenicity of the strain. 3km and 10km Temporary Control Zones have been put in place surrounding the premises.
12 November 2021	Updated to reflect that avian influenza H5N1 has been confirmed in birds at a premises near Salwick, Fylde, Lancashire, England. Further testing is underway to confirm the pathogenicity of the strain. 3km and 10km Temporary Control Zones have been put in place surrounding the premises.
11 November 2021	Further testing has now confirmed highly pathogenic strain (HPAI H5N1) at premises near Frinton-on-sea, Tendring, Essex, England. A 3km Protection Zone and 10km Surveillance Zone have been put in place surrounding the premises.
10 November 2021	Updated to reflect that avian influenza H5N1 has been confirmed in birds at a premises near Frinton-on-Sea, Tendring, Essex, England. Further testing is underway to confirm the pathogenicity of the strain. 3km and 10km Temporary Control Zones have been put in place surrounding the premises. Also updated the AIPZ section.
8 November 2021	Updated guidance to reflect that the UK is no longer free from avian influenza under the World Organisation for Animal Health (OIE) rules.
7 November 2021	Updated to reflect that the case of avian influenza H5N1 confirmed at a premises near Alcester, Bidford, Warwickshire to be a highly pathogenic strain (HPAI H5N1) after further testing.
5 November 2021	Avian influenza H5N1 has been confirmed in a small poultry unit at a premises near Alcester, Bidford, Warwickshire.
5 November 2021	Updated to reflect changes to the bird gatherings general licence. From 8 November 2021 no gatherings of poultry, galliforme birds or anseriforme birds are permitted.
4 November 2021	Updated to reflect that the case of avian influenza in the Angus constituency in Scotland to be a highly pathogenic strain (HPAI H5N1) after further testing.
3 November 2021	Added links to guidance on avian influenza cases and disease control zones in England. Moved the movement controls and licences details for England cases to this guidance.

2 November 2021	Updated guidance to reflect that an Avian Influenza Prevention Zone (AIPZ) has been declared across Great Britain. A case of avian influenza H5N1 has also been confirmed in birds at a premises in the Angus constituency in Scotland.
1 November 2021	Updated to reflect that further testing has confirmed a case of H5N1 Avian Influenza at a premises near Chirk, Wrexham, Wales to be a highly pathogenic strain (HPAI H5N1).
1 November 2021	Updated to reflect that a case of H5N1 Avian Influenza has been confirmed at a premises near Wrexham, Wales.
29 October 2021	Updated to reflect that the 3km Protection Zone and 10km Surveillance Zone have been amended and 3km and 10km Captive Bird (Monitoring) Controlled Zones have been declared around the rescue centre near Droitwich Spa, Wychavon, Worcestershire.
29 October 2021	Updated the 'Trade, import and export issues' section to reflect that the UK retains its World Animal Health Organisation (OIE) disease free status.
28 October 2021	Updated the meat from poultry within a Protection Zone in England general licences.
27 October 2021	Added details on general and specific licensing for Protection and Surveillance Zones in the movement controls and licences section including application requesting exemption from restrictions (EXD100) form.
26 October 2021	Updated to reflect that further testing has confirmed highly pathogenic strain (HPAI H5N1) at a rescue centre near Droitwich Spa, Wychavon, Worcestershire. The 3km and 10km Temporary Control Zone have been revoked and replaced by a 3km Protection and a 10km Surveillance Zone around the premises. Added section of risk levels.
15 October 2021	Updated to reflect that a new case of avian influenza H5N1 has been confirmed in birds at a rescue centre near Droitwich Spa, Worcestershire.
8 September 2021	Updated the section on 'How to spot avian influenza'.
3 September 2021	Updated the section on 'Meat from poultry within a Protection Zone in England'.

9 June 2021	Updated to reflect that the UK is now free from avian influenza.
15 May 2021	Updated Rules on meat produced from poultry and farmed game birds originating in the Protection Zone guidance document
14 May 2021	Updated to reflect that the Avian Influenza Prevention Zone (AIPZ) has been lifted.
1 May 2021	Updated to reflect that the risk of avian influenza in poultry has now been reduced to 'low'.
22 April 2021	Updated to reflect that the Surveillance Zone surrounding the premises near Uttoxeter, East Staffordshire was revoked. Local movement restrictions have been removed but the Avian Influenza Prevention Zone (AIPZ) measures remain in place.
21 April 2021	Updated to reflect that the Protection Zone surrounding the premises near Uttoxeter, East Staffordshire has been revoked and the area merged with the respective Surveillance Zone.
2 April 2021	Updated to reflect that certain bird gatherings can take place in GB provided you notify the APHA and you meet the requirements of the General Licence.

1 April 2021	Updated to reflect that avian influenza H5N8 was confirmed in 2 captive peregrine falcons at a private residential premises near Skelmersdale, West Lancashire on 31 March 2021. Further testing has confirmed this to be Highly Pathogenic Avian Influenza (HPAI H5N8).
31 March 2021	Updated the 'Biosecurity and preventing welfare impacts in poultry and captive birds' publication and biosecurity checklist.
30 March 2021	Updated the Avian Influenza Prevention Zone (AIPZ) declaration and changes to the avian influenza risk levels.
30 March 2021	Replaced the 'Rules on meat produced from poultry and farmed game birds originating in the Protection Zone' document.
29 March 2021	Updated the latest situation section to reflect the housing measures component of the Avian Influenza Prevention Zone (AIPZ) will still be lifted at 11:59 pm on 31 March 2021.
28 March 2021	Updated to reflect that after further laboratory testing, the case of avian influenza H5N8 in broiler chickens at a commercial premises near Uttoxeter, East

	Staffordshire has been confirmed as a highly pathogenic strain. A 3km Protection and 10km Surveillance Zone has been put in place around the infected premises.
27 March 2021	Updated to reflect that after further laboratory testing, the case of avian influenza (H5N3) in turkey breeders near Winsford, Cheshire West and Chester, has been confirmed as a low pathogenic strain. A 1km Temporary Movement Restriction Zone has been replaced with a 1km low pathogenic avian influenza Restricted Zone.
26 March 2021	Updated to reflect that avian influenza H5N8 (pathogenicity to be confirmed) was confirmed in broiler chickens at a commercial premises near Uttoxeter, East Staffordshire. A 3km and 10km Temporary Control Zone has been declared around the premises.
23 March 2021	Updated to reflect that avian influenza H5N3 was confirmed in turkey breeders near Winsford, Cheshire West and Chester. A 1km Temporary Movement Restriction Zone has been declared around the premises.
19 March 2021	Updated the 'How to prepare for when your free-range birds can be let outside again' guidance document.
15 March 2021	Updated to reflect that the risk of avian influenza has reduced to ‘medium’ and housing restrictions end on 31 March 2021.
11 March 2021	Added information around H5N8 influenza of avian origin being detected in 3 seals and a fox that died at a wildlife rehabilitation centre in England.
11 March 2021	Updated the Public Health England advice on the 'risk level' and 'advice to the public' sections.
2 March 2021	The Surveillance Zone around a premises near Redcar, Redcar and Cleveland has been revoked.
2 March 2021	Uploaded General licence for the movement of mammals, except equines, on foot from and returning to the same premises in the Protection or Surveillance Zone where poultry or other captive birds are kept
1 March 2021	Updated to reflect that the Protection Zone surrounding the premises near Redcar has been revoked and the area merged with the Surveillance Zone.
16 February 2021	Added general licence EXD 296(HPAI)(E) for the movement of

	captive birds, other than poultry, from a premises located in an area free of restrictions or in an AIPZ to a premises located within a Surveillance or Protection Zone in England.
12 February 2021	Updated the guidance under the 'Bird fairs, markets, shows and other gatherings' and 'Pigeons or birds of prey' headings.
10 February 2021	Updated the 'Cases in Scotland' section following confirmation of an additional case of avian influenza in Scotland.
8 February 2021	Updated Rules on meat produced from poultry and farmed game birds originating in the Protection Zone guidance
6 February 2021	Updated to reflect that after further laboratory testing, the case of avian influenza near Redcar, Redcar and Cleveland, has been confirmed as a highly pathogenic strain (HPAI H5N8). A 3km Protection Zone and 10km Surveillance Zone have been declared around the premises.
5 February 2021	Avian influenza H5N8 (pathogenicity to be confirmed) was confirmed in laying chickens at a small commercial premises near Redcar, Redcar and Cleveland on 6 February 2021.
1 February 2021	Updated to reflect that the Surveillance Zone around a premises near Watton, Breckland, Norfolk has been revoked.
30 January 2021	Updated the 'Rules on meat produced from poultry and farmed game birds originating in the Protection Zone' guidance.
28 January 2021	Updated to reflect that following successful completion of disease control activities and surveillance within the zone, the Surveillance Zone around a premises near Exmouth, West Devon, Devonhas been revoked.
27 January 2021	Updated to reflect that the Surveillance Zone around a premises near Attleborough, Breckland, Norfolk has been revoked. Added section on case in Wales.
23 January 2021	Updated to reflect that the Protection Zone surrounding the premises near Watton, Breckland, Norfolk have been revoked and the areas merged with the Surveillance Zone.
22 January 2021	Updated to reflect that the Surveillance Zone declared on 19 January around the second premises near Attleborough, Breckland, Norfolk has been revoked. Local movement restrictions have now been removed but the Avian Influenza Prevention Zone (AIPZ) measures remain in place.

21 January 2021	Updated the 'Rules on meat produced from poultry and farmed game birds originating in the Protection Zone' guidance.
20 January 2021	Updated to reflect that the Protection Zone surrounding the premises near Exmouth, East Devon, Devon have been revoked and the areas merged with the Surveillance Zone. Added section on cases in Northern Ireland.
19 January 2021	Updated to reflect that the Surveillance Zone around a premises near Gillingham, North Dorset, Dorset has been revoked.
	Added the general licence for the movement of horses for the purpose of exercise from a premises in the Surveillance Zone where poultry or other captive birds are kept. Updated the rules on meat produced from poultry and farmed game birds originating in the Protection Zone. The Protection Zones surrounding both the second and third premises near Attleborough, Breckland, Norfolk have been revoked and the Surveillance Zone around a premises near Hawes in North Yorkshire has been revoked.
17 January 2021	Updated to reflect that the surveillance zone near Hawes, Richmondshire, North Yorkshire has been revoked.
16 January 2021	Following successful completion of disease control activities and surveillance within the zones, the Surveillance Zone declared on 8 January around a premises near Willington in South Derbyshire has been revoked.
15 January 2021	Following successful completion of disease control activities and surveillance, the Surveillance Zone declared on 9 January around a premises near Attleborough in Norfolk and the Surveillance Zone declared on 10 January around a premises near King’s Lynn in Norfolk have both been revoked.
11 January 2021	Updated the 'Biosecurity and preventing welfare impacts in poultry and captive birds' guidance document.
10 January 2021	The Protection Zone declared on 19 December around a premises near Gillingham in Dorset has been revoked and the areas merged with the Surveillance Zone. Updated the rules on meat produced from poultry and farmed game birds originating in the Protection Zone guide.
9 January 2021	The Protection Zone declared on 22

	December around a premises near Hawes in North Yorkshire and the Protection Zone declared on 5 December around a premises near King’s Lynn in Norfolk have been revoked and the areas merged with the Surveillance Zones.
9 January 2021	The Protection Zone declared on 4 December 2020 around the first premises near Attleborough in Norfolk has been revoked and the area merged with the Surveillance Zone.
8 January 2021	The Protection Zone declared on 15 December 2020 around a premises near Willington, South Derbyshire, Derbyshire was revoked on 8 January 2021 and the area merged with the Surveillance Zone.
7 January 2021	Updated the guidance on 'Rules on meat produced from poultry and farmed game birds originating in the Protection Zone'.
6 January 2021	Updated to reflect that the Surveillance Zones declared around 2 premises near Northallerton in North Yorkshire have been revoked. Local movement restrictions have been removed but the Avian Influenza Prevention Zone (AIPZ) measures remain in place.

Appendix C – Conference poster

Poultry in motion:

Utilising social media data for veterinary epidemiological surveillance

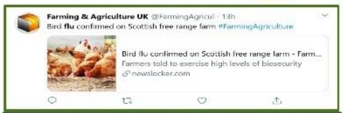
A. Samuel Munaf[1,2], Kevin Swingler[1], Franz Brülisauer[3], Anthony O'Hare[1], George Gunn[2], and Aaron Reeves[2]

[1] Division of Computing Science and Mathematics, University of Stirling, [2] Epidemiology Research Unit, Department of Veterinary and Animal Sciences, Northern Faculty, Scotland's Rural College (SRUC) ,
[3] SRUC Veterinary Services , Scotland's Rural College (SRUC)

Project Overview

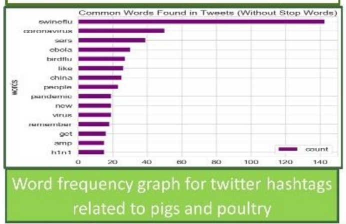
To gain insights about Smallholder pig and poultry farmers through the application of social media data sources in order to:

- To better inform current knowledge of location, demographics and interaction networks
- Achieve early detection of disease
- To compare any trends identified to known animal health related events



Farming & Agriculture UK @FarmingAgricul · 13h
Bird **flu** confirmed on Scottish free range farm #FarmingAgriculture

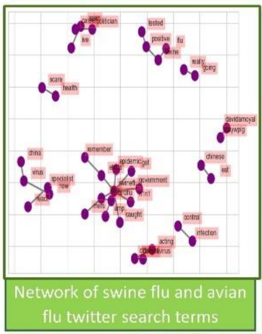
Bird flu confirmed on Scottish free range farm - Farm...
Farmers told to exercise high levels of biosecurity
newslocker.com



Word frequency graph for twitter hashtags related to pigs and poultry

Approach

- Extracting messages over various time periods based on spatial constraints and keyword filters
- Applying web scraping techniques to derive user profile information, location, posts and interactions
- Sentiment analysis to ascertain user feelings regarding outbreak/public health information/news
- Text mining of posts
  - Natural language processing, topic modelling and word frequencies
- Network analysis of:
  - Keyword search terms
  - User message interactions
  - Information dispersion via friends/followers
- Image classification of disease

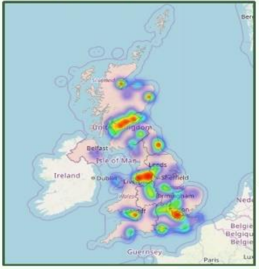


Network of swine flu and avian flu twitter search terms

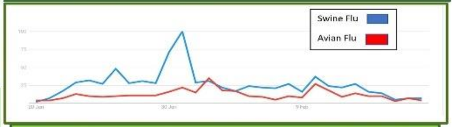
What is Infoveillance?

"The science of distribution and determinants of information via the internet with the ultimate aim to inform public health and public policy"

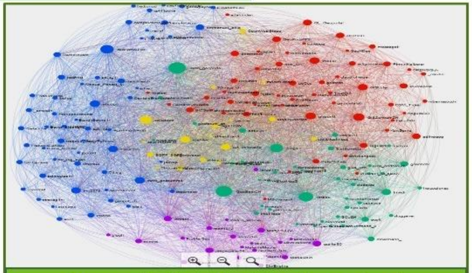
(Eysenbach, G 2009)



Geolocation map of swine flu related tweets in the UK



Google Trends: Search queries past 30 days (UK)



An example of a twitter interactions network (Different colours represent communities within the overall network)

Could the corona virus have come from pigs?

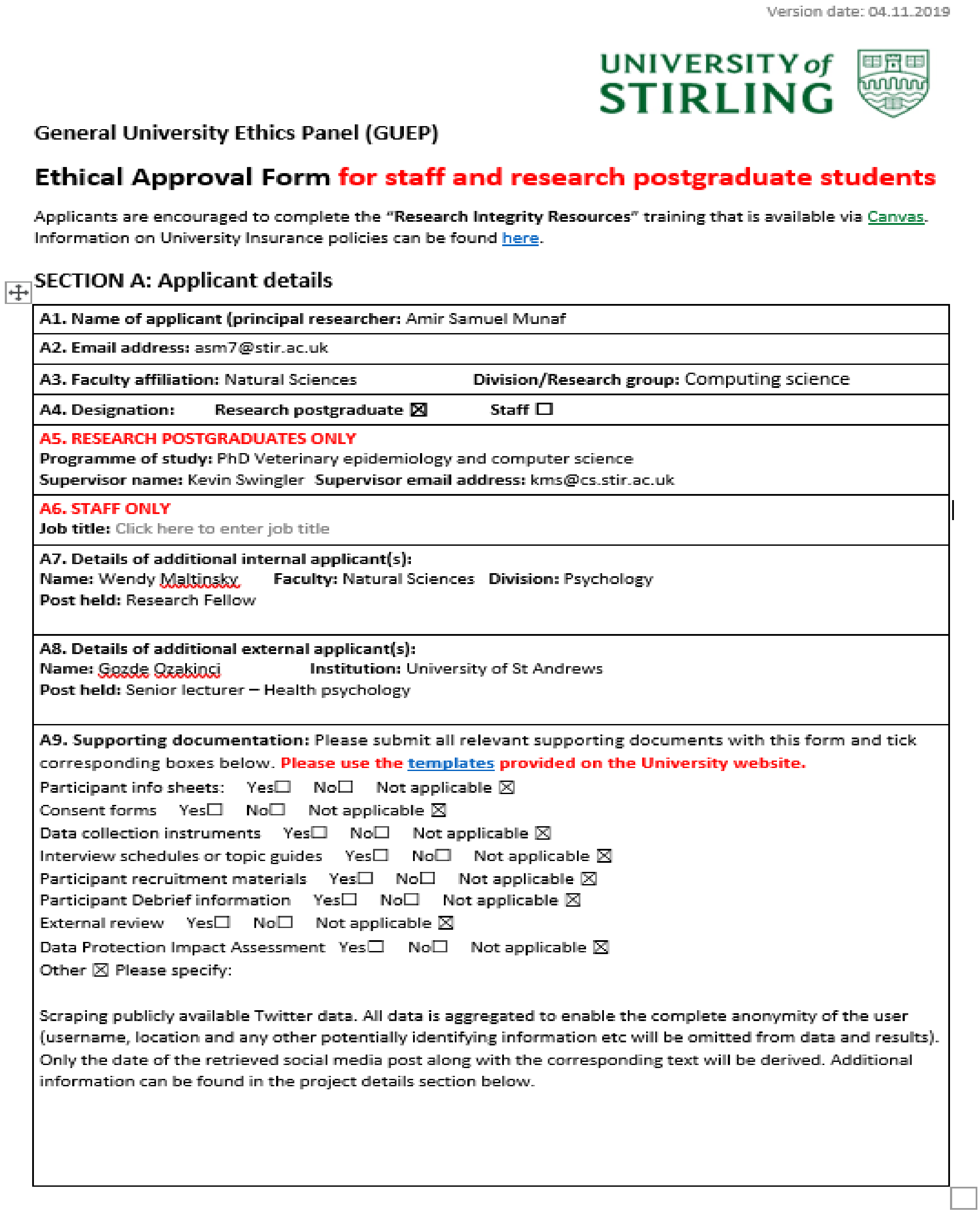
Preliminary findings

- Since the outbreak of African swine fever in early 2019, smallholders have raised concerns regarding animal welfare and have been seeking advice on Biosecurity methods
- Prevalence of many closed private groups within Facebook – Makes accessing information difficult
- Forums seem to provide greater levels of relevant information as they possess pig and poultry specific sub-forums
- A noticeable increase over the past few months relating to posted images of potentially diseased livestock

Samuel Smallholder-Phd *@SmallholderPhd* *asm7@stir.ac.uk* *(+44) 1786 467452*

BE THE DIFFERENCE

Appendix D – Ethical approval



Version date: 04.11.2019

UNIVERSITY of STIRLING

General University Ethics Panel (GUEP)

Ethical Approval Form for staff and research postgraduate students

Applicants are encouraged to complete the "**Research Integrity Resources**" training that is available via Canvas. Information on University Insurance policies can be found here.

SECTION A: Applicant details

A1. Name of applicant (principal researcher: Amir Samuel Munaf

A2. Email address: asm7@stir.ac.uk

A3. Faculty affiliation: Natural Sciences **Division/Research group:** Computing science

A4. Designation: Research postgraduate ☒ Staff ☐

A5. RESEARCH POSTGRADUATES ONLY
Programme of study: PhD Veterinary epidemiology and computer science
Supervisor name: Kevin Swingler **Supervisor email address:** kms@cs.stir.ac.uk

A6. STAFF ONLY
Job title: Click here to enter job title

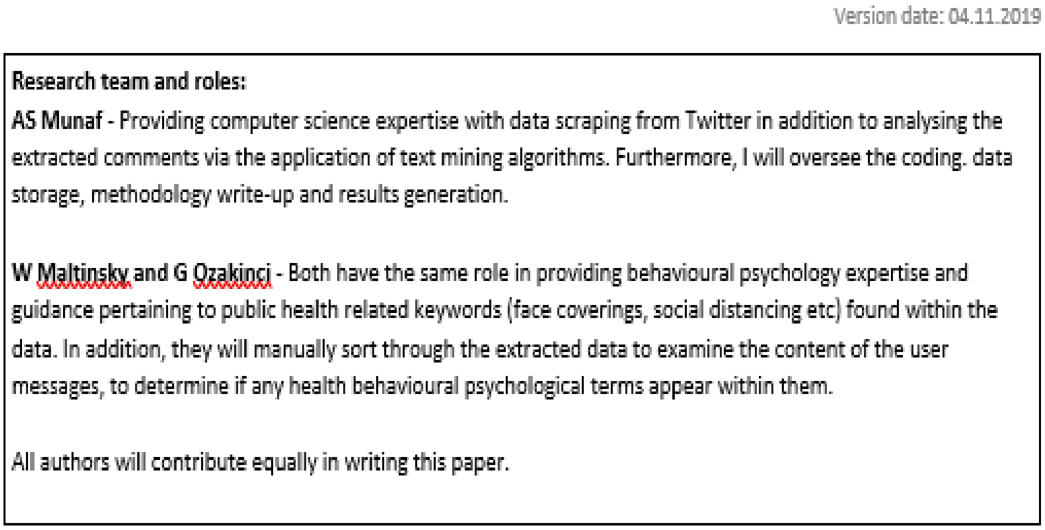
A7. Details of additional internal applicant(s):
Name: Wendy Maltinsky **Faculty:** Natural Sciences **Division:** Psychology
Post held: Research Fellow

A8. Details of additional external applicant(s):
Name: Gozde Ozakinci **Institution:** University of St Andrews
Post held: Senior lecturer – Health psychology

A9. Supporting documentation: Please submit all relevant supporting documents with this form and tick corresponding boxes below. **Please use the templates provided on the University website.**

Participant info sheets: Yes☐ No☐ Not applicable ☒
Consent forms Yes☐ No☐ Not applicable ☒
Data collection instruments Yes☐ No☐ Not applicable ☒
Interview schedules or topic guides Yes☐ No☐ Not applicable ☒
Participant recruitment materials Yes☐ No☐ Not applicable ☒
Participant Debrief information Yes☐ No☐ Not applicable ☒
External review Yes☐ No☐ Not applicable ☒
Data Protection Impact Assessment Yes☐ No☐ Not applicable ☒
Other ☒ Please specify:

Scraping publicly available Twitter data. All data is aggregated to enable the complete anonymity of the user (username, location and any other potentially identifying information etc will be omitted from data and results). Only the date of the retrieved social media post along with the corresponding text will be derived. Additional information can be found in the project details section below.



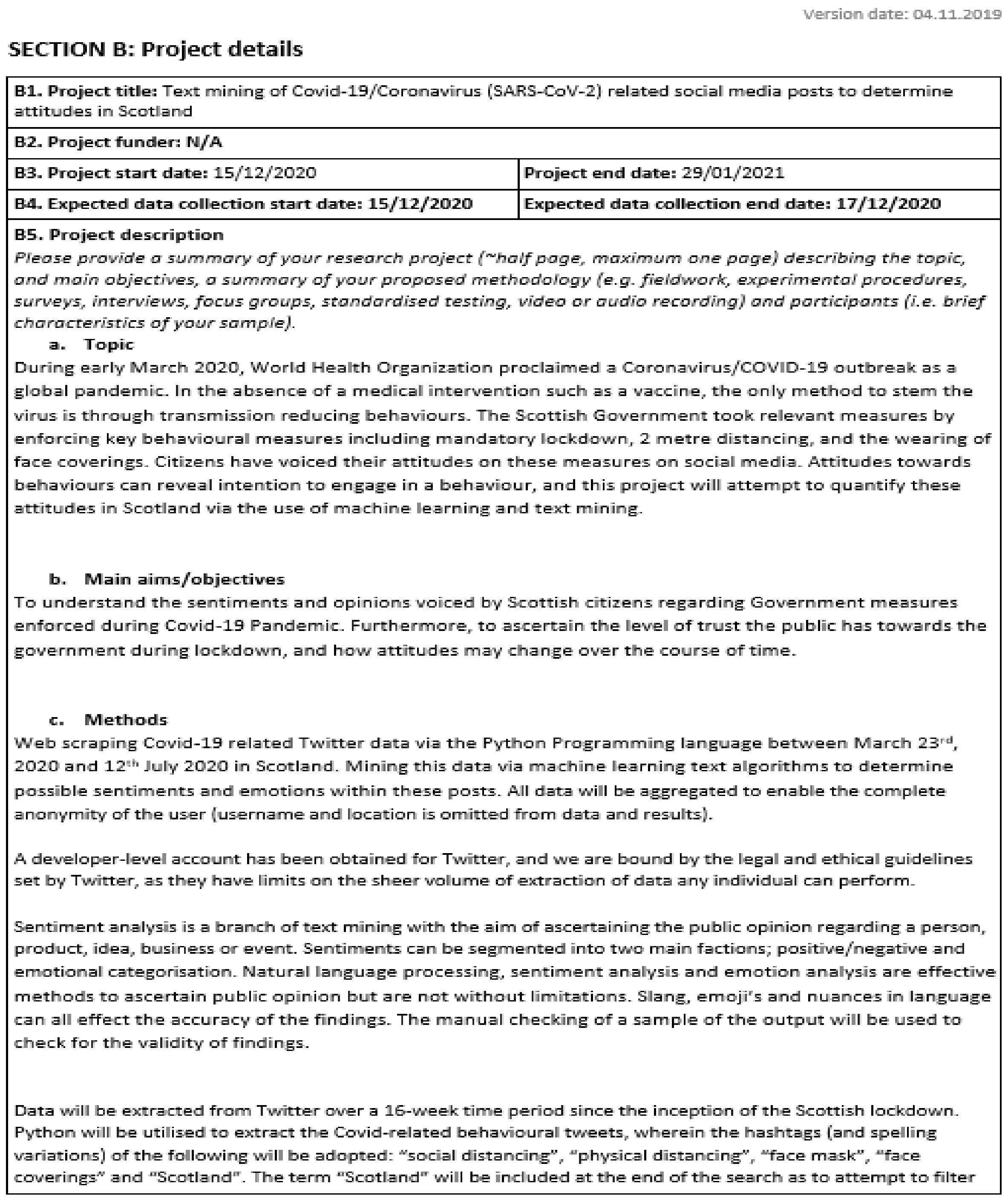
Version date: 04.11.2019

Research team and roles:

AS Munaf - Providing computer science expertise with data scraping from Twitter in addition to analysing the extracted comments via the application of text mining algorithms. Furthermore, I will oversee the coding. data storage, methodology write-up and results generation.

W Maltinsky and G Ozakinci - Both have the same role in providing behavioural psychology expertise and guidance pertaining to public health related keywords (face coverings, social distancing etc) found within the data. In addition, they will manually sort through the extracted data to examine the content of the user messages, to determine if any health behavioural psychological terms appear within them.

All authors will contribute equally in writing this paper.



Version date: 04.11.2019

SECTION B: Project details

B1. Project title: Text mining of Covid-19/Coronavirus (SARS-CoV-2) related social media posts to determine attitudes in Scotland	
B2. Project funder: N/A	
B3. Project start date: 15/12/2020	**Project end date:** 29/01/2021
B4. Expected data collection start date: 15/12/2020	**Expected data collection end date: 17/12/2020**

B5. Project description

Please provide a summary of your research project (~half page, maximum one page) describing the topic, and main objectives, a summary of your proposed methodology (e.g. fieldwork, experimental procedures, surveys, interviews, focus groups, standardised testing, video or audio recording) and participants (i.e. brief characteristics of your sample).

a. Topic

During early March 2020, World Health Organization proclaimed a Coronavirus/COVID-19 outbreak as a global pandemic. In the absence of a medical intervention such as a vaccine, the only method to stem the virus is through transmission reducing behaviours. The Scottish Government took relevant measures by enforcing key behavioural measures including mandatory lockdown, 2 metre distancing, and the wearing of face coverings. Citizens have voiced their attitudes on these measures on social media. Attitudes towards behaviours can reveal intention to engage in a behaviour, and this project will attempt to quantify these attitudes in Scotland via the use of machine learning and text mining.

b. Main aims/objectives

To understand the sentiments and opinions voiced by Scottish citizens regarding Government measures enforced during Covid-19 Pandemic. Furthermore, to ascertain the level of trust the public has towards the government during lockdown, and how attitudes may change over the course of time.

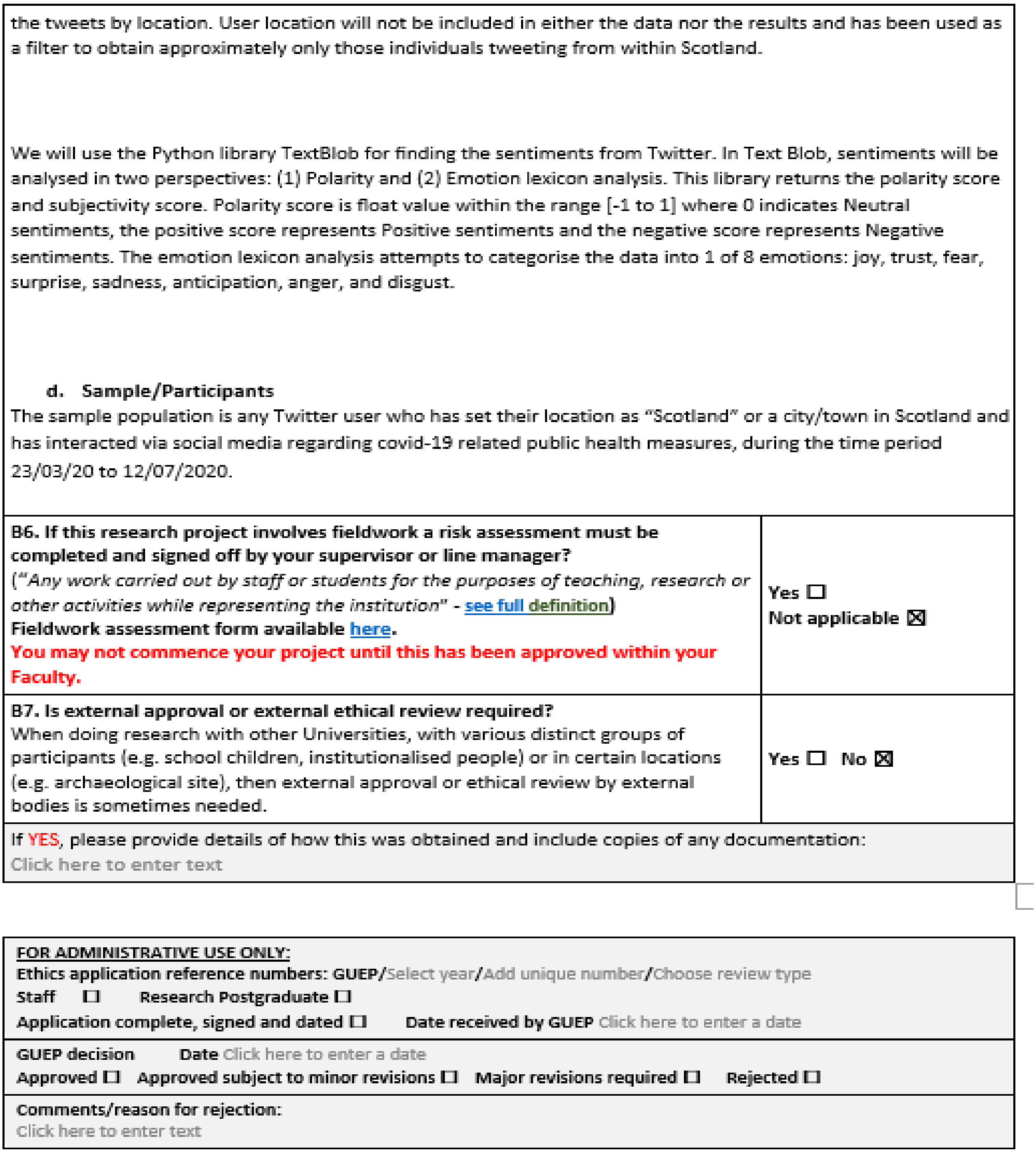
c. Methods

Web scraping Covid-19 related Twitter data via the Python Programming language between March 23rd, 2020 and 12th July 2020 in Scotland. Mining this data via machine learning text algorithms to determine possible sentiments and emotions within these posts. All data will be aggregated to enable the complete anonymity of the user (username and location is omitted from data and results).

A developer-level account has been obtained for Twitter, and we are bound by the legal and ethical guidelines set by Twitter, as they have limits on the sheer volume of extraction of data any individual can perform.

Sentiment analysis is a branch of text mining with the aim of ascertaining the public opinion regarding a person, product, idea, business or event. Sentiments can be segmented into two main factions; positive/negative and emotional categorisation. Natural language processing, sentiment analysis and emotion analysis are effective methods to ascertain public opinion but are not without limitations. Slang, emoji's and nuances in language can all effect the accuracy of the findings. The manual checking of a sample of the output will be used to check for the validity of findings.

Data will be extracted from Twitter over a 16-week time period since the inception of the Scottish lockdown. Python will be utilised to extract the Covid-related behavioural tweets, wherein the hashtags (and spelling variations) of the following will be adopted: "social distancing", "physical distancing", "face mask", "face coverings" and "Scotland". The term "Scotland" will be included at the end of the search as to attempt to filter



the tweets by location. User location will not be included in either the data nor the results and has been used as a filter to obtain approximately only those individuals tweeting from within Scotland.

We will use the Python library TextBlob for finding the sentiments from Twitter. In Text Blob, sentiments will be analysed in two perspectives: (1) Polarity and (2) Emotion lexicon analysis. This library returns the polarity score and subjectivity score. Polarity score is float value within the range [-1 to 1] where 0 indicates Neutral sentiments, the positive score represents Positive sentiments and the negative score represents Negative sentiments. The emotion lexicon analysis attempts to categorise the data into 1 of 8 emotions: joy, trust, fear, surprise, sadness, anticipation, anger, and disgust.

d. Sample/Participants

The sample population is any Twitter user who has set their location as "Scotland" or a city/town in Scotland and has interacted via social media regarding covid-19 related public health measures, during the time period 23/03/20 to 12/07/2020.

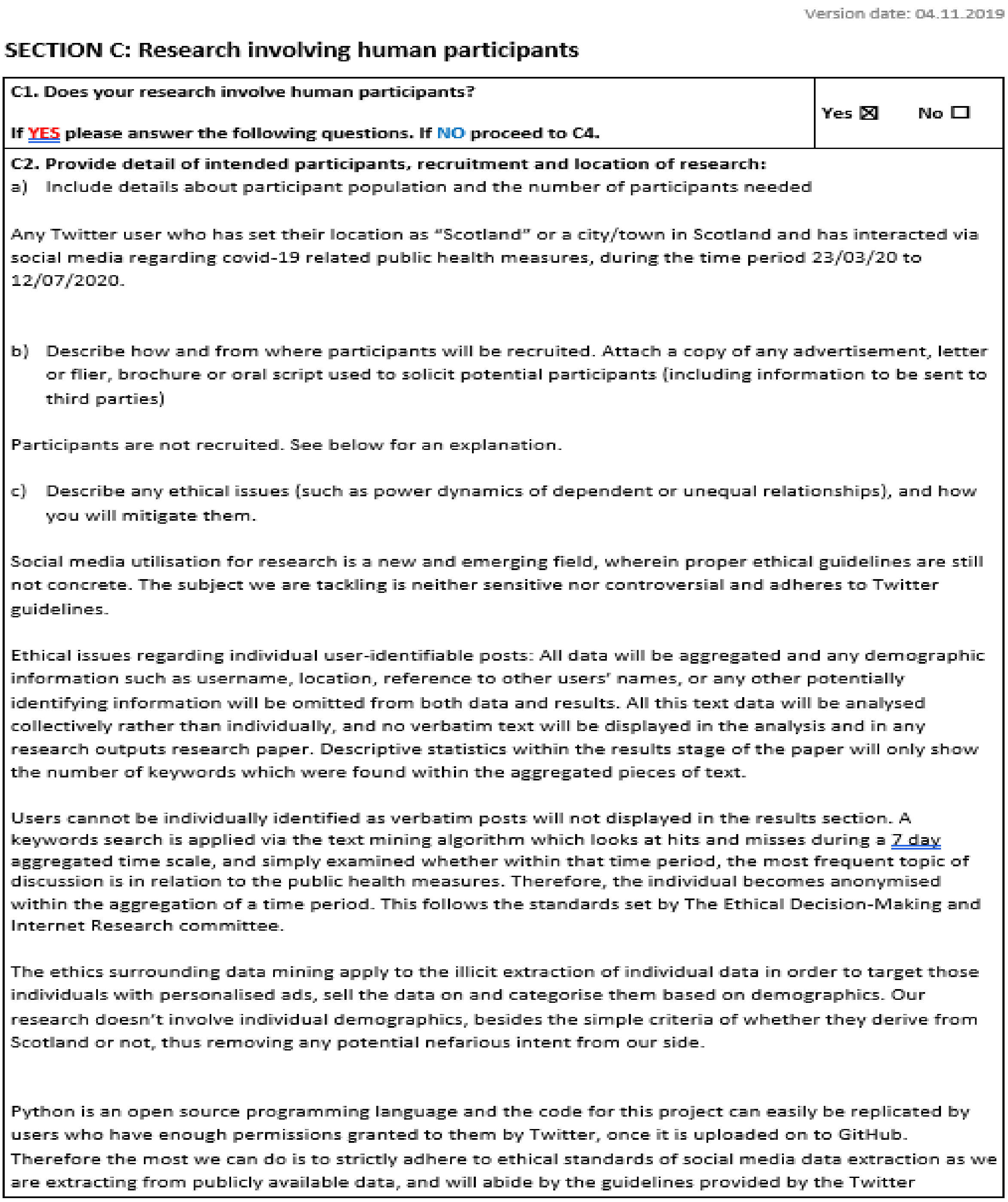
B6. If this research project involves fieldwork a risk assessment must be completed and signed off by your supervisor or line manager? (*"Any work carried out by staff or students for the purposes of teaching, research or other activities while representing the institution"* - see full definition) **Fieldwork assessment form available here.** **You may not commence your project until this has been approved within your Faculty.**	**Yes** ☐ **Not applicable** ☒
B7. Is external approval or external ethical review required? When doing research with other Universities, with various distinct groups of participants (e.g. school children, institutionalised people) or in certain locations (e.g. archaeological site), then external approval or ethical review by external bodies is sometimes needed.	**Yes** ☐ **No** ☒

If YES, please provide details of how this was obtained and include copies of any documentation:
Click here to enter text

FOR ADMINISTRATIVE USE ONLY:
Ethics application reference numbers: GUEP/Select year**/**Add unique number**/**Choose review type
Staff ☐ **Research Postgraduate** ☐
Application complete, signed and dated ☐ **Date received by GUEP** Click here to enter a date

GUEP decision **Date** Click here to enter a date
Approved ☐ **Approved subject to minor revisions** ☐ **Major revisions required** ☐ **Rejected** ☐

Comments/reason for rejection:
Click here to enter text



Version date: 04.11.2019

SECTION C: Research involving human participants

C1. Does your research involve human participants? **If YES please answer the following questions. If NO proceed to C4.**	**Yes** ☒ **No** ☐

C2. Provide detail of intended participants, recruitment and location of research:

a) Include details about participant population and the number of participants needed

Any Twitter user who has set their location as "Scotland" or a city/town in Scotland and has interacted via social media regarding covid-19 related public health measures, during the time period 23/03/20 to 12/07/2020.

b) Describe how and from where participants will be recruited. Attach a copy of any advertisement, letter or flier, brochure or oral script used to solicit potential participants (including information to be sent to third parties)

Participants are not recruited. See below for an explanation.

c) Describe any ethical issues (such as power dynamics of dependent or unequal relationships), and how you will mitigate them.

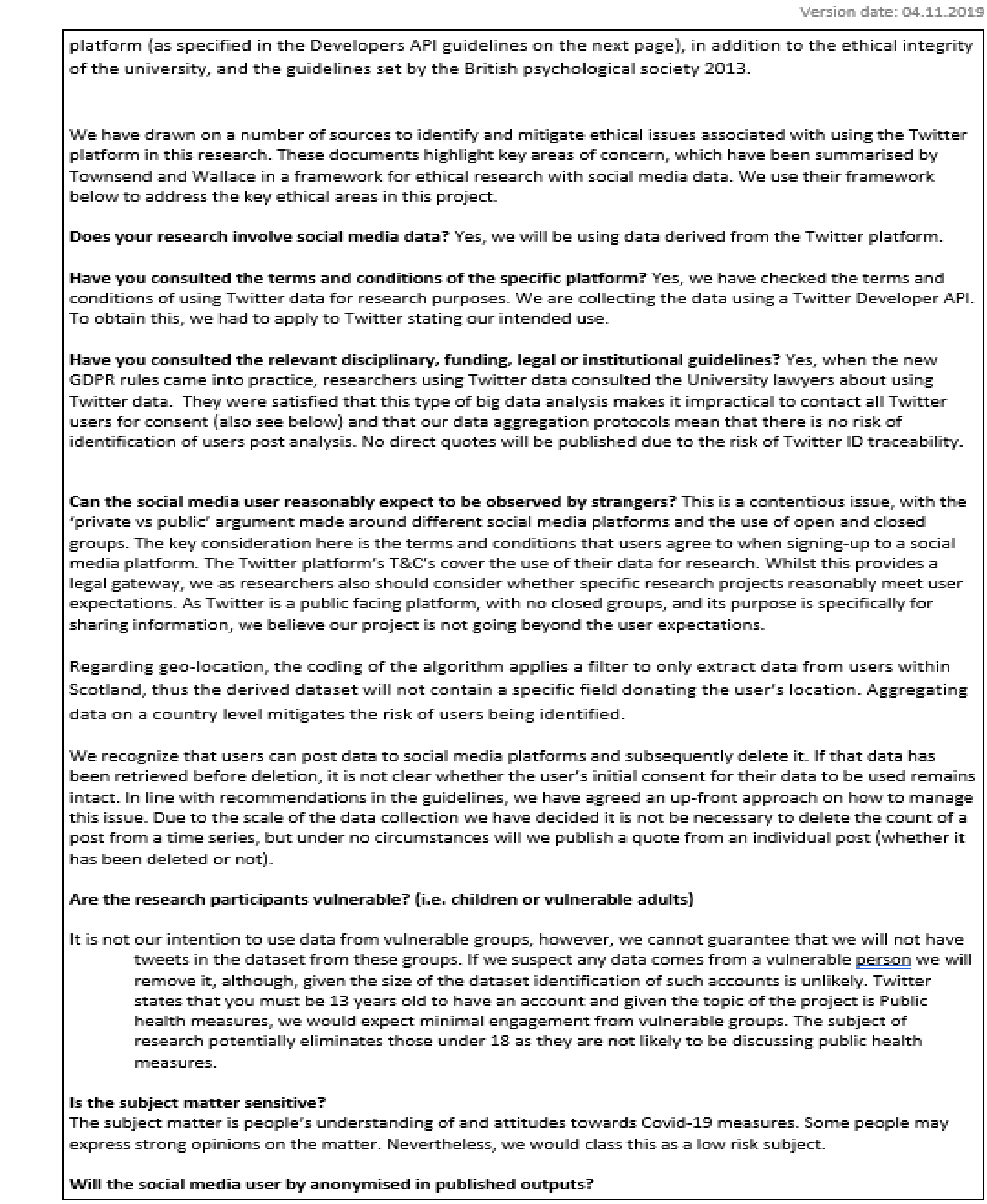
Social media utilisation for research is a new and emerging field, wherein proper ethical guidelines are still not concrete. The subject we are tackling is neither sensitive nor controversial and adheres to Twitter guidelines.

Ethical issues regarding individual user-identifiable posts: All data will be aggregated and any demographic information such as username, location, reference to other users' names, or any other potentially identifying information will be omitted from both data and results. All this text data will be analysed collectively rather than individually, and no verbatim text will be displayed in the analysis and in any research outputs research paper. Descriptive statistics within the results stage of the paper will only show the number of keywords which were found within the aggregated pieces of text.

Users cannot be individually identified as verbatim posts will not displayed in the results section. A keywords search is applied via the text mining algorithm which looks at hits and misses during a 7 day aggregated time scale, and simply examined whether within that time period, the most frequent topic of discussion is in relation to the public health measures. Therefore, the individual becomes anonymised within the aggregation of a time period. This follows the standards set by The Ethical Decision-Making and Internet Research committee.

The ethics surrounding data mining apply to the illicit extraction of individual data in order to target those individuals with personalised ads, sell the data on and categorise them based on demographics. Our research doesn't involve individual demographics, besides the simple criteria of whether they derive from Scotland or not, thus removing any potential nefarious intent from our side.

Python is an open source programming language and the code for this project can easily be replicated by users who have enough permissions granted to them by Twitter, once it is uploaded on to GitHub. Therefore the most we can do is to strictly adhere to ethical standards of social media data extraction as we are extracting from publicly available data, and will abide by the guidelines provided by the Twitter



Version date: 04.11.2019

platform (as specified in the Developers API guidelines on the next page), in addition to the ethical integrity of the university, and the guidelines set by the British psychological society 2013.

We have drawn on a number of sources to identify and mitigate ethical issues associated with using the Twitter platform in this research. These documents highlight key areas of concern, which have been summarised by Townsend and Wallace in a framework for ethical research with social media data. We use their framework below to address the key ethical areas in this project.

Does your research involve social media data? Yes, we will be using data derived from the Twitter platform.

Have you consulted the terms and conditions of the specific platform? Yes, we have checked the terms and conditions of using Twitter data for research purposes. We are collecting the data using a Twitter Developer API. To obtain this, we had to apply to Twitter stating our intended use.

Have you consulted the relevant disciplinary, funding, legal or institutional guidelines? Yes, when the new GDPR rules came into practice, researchers using Twitter data consulted the University lawyers about using Twitter data. They were satisfied that this type of big data analysis makes it impractical to contact all Twitter users for consent (also see below) and that our data aggregation protocols mean that there is no risk of identification of users post analysis. No direct quotes will be published due to the risk of Twitter ID traceability.

Can the social media user reasonably expect to be observed by strangers? This is a contentious issue, with the 'private vs public' argument made around different social media platforms and the use of open and closed groups. The key consideration here is the terms and conditions that users agree to when signing-up to a social media platform. The Twitter platform's T&C's cover the use of their data for research. Whilst this provides a legal gateway, we as researchers also should consider whether specific research projects reasonably meet user expectations. As Twitter is a public facing platform, with no closed groups, and its purpose is specifically for sharing information, we believe our project is not going beyond the user expectations.

Regarding geo-location, the coding of the algorithm applies a filter to only extract data from users within Scotland, thus the derived dataset will not contain a specific field donating the user's location. Aggregating data on a country level mitigates the risk of users being identified.

We recognize that users can post data to social media platforms and subsequently delete it. If that data has been retrieved before deletion, it is not clear whether the user's initial consent for their data to be used remains intact. In line with recommendations in the guidelines, we have agreed an up-front approach on how to manage this issue. Due to the scale of the data collection we have decided it is not be necessary to delete the count of a post from a time series, but under no circumstances will we publish a quote from an individual post (whether it has been deleted or not).

Are the research participants vulnerable? (i.e. children or vulnerable adults)

It is not our intention to use data from vulnerable groups, however, we cannot guarantee that we will not have tweets in the dataset from these groups. If we suspect any data comes from a vulnerable person we will remove it, although, given the size of the dataset identification of such accounts is unlikely. Twitter states that you must be 13 years old to have an account and given the topic of the project is Public health measures, we would expect minimal engagement from vulnerable groups. The subject of research potentially eliminates those under 18 as they are not likely to be discussing public health measures.

Is the subject matter sensitive?
The subject matter is people's understanding of and attitudes towards Covid-19 measures. Some people may express strong opinions on the matter. Nevertheless, we would class this as a low risk subject.

Will the social media user by anonymised in published outputs?

Yes, the Twitter datasets will be aggregated, and no identifiable information will be published. (e.g. quotations) in project outputs.

Can you publish or share the dataset?
Only anonymised datasets will be published in accordance with the ESRC data sharing regulations.

References

- The University of Stirling's guidance on 'Ethical considerations for internet – mediated research'
- Social Media Research Group (2016) Using social media for social research: An introduction
- Townsend, L. and Wallace, C. (no date) Social Media Research: A Guide to Ethics
- The Ethical Decision-Making and
  Internet Research committee (http://aoir.org/reports/ethics2.pdf).

d) Describe the setting in which the research will take place (e.g. online, University campus, schools, institutions, public centres, overseas, etc)
Online - using the Twitter API to download anonymised statements related to Covid made during the study period.

e) Describe any incentive (e.g. tokens, financial payment) participants may receive for participation
Click here to enter text
Not needed as the data is already in the public domain and is free to extract within the Twitter guidelines

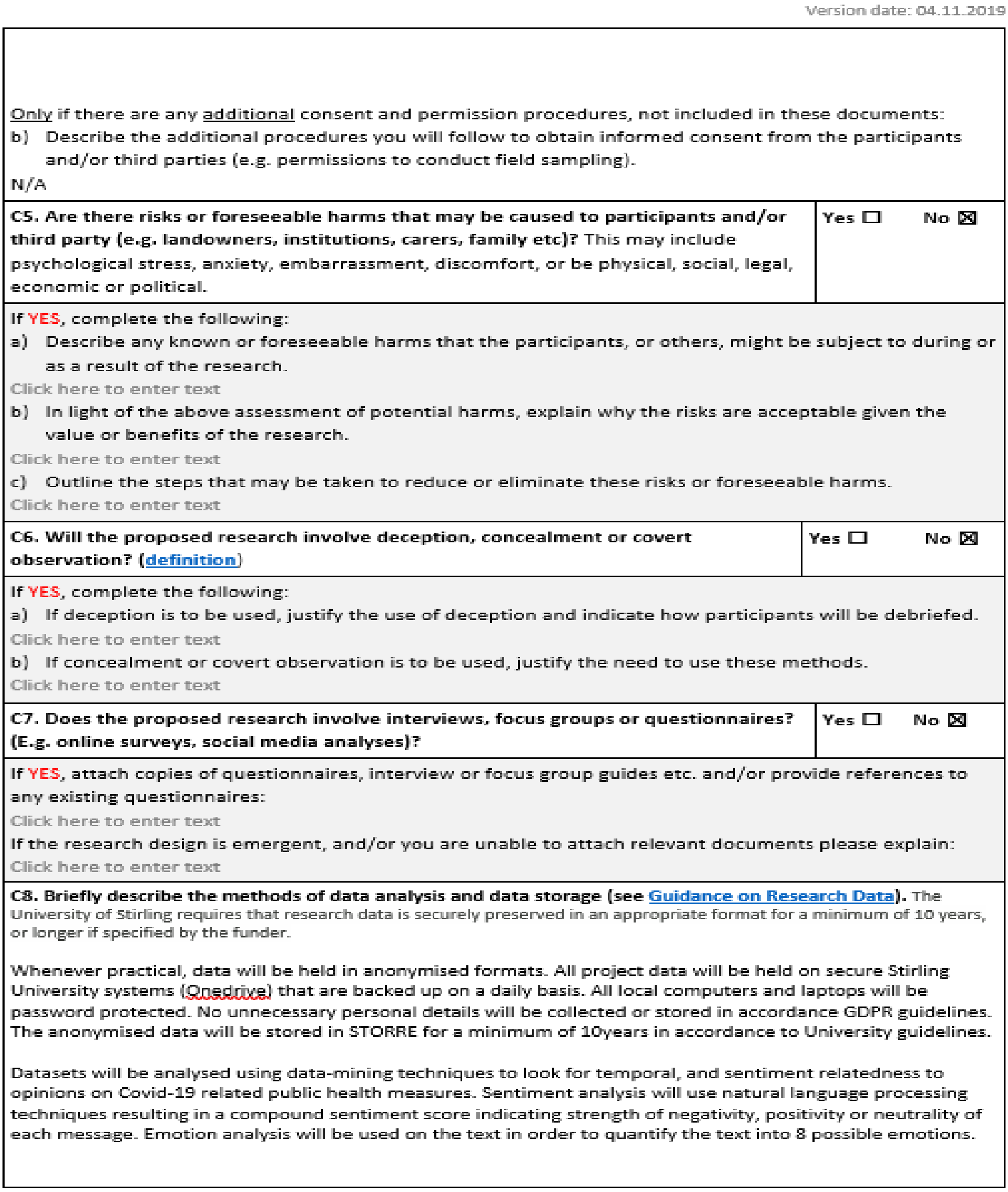
C3. Does your proposed research involve vulnerable groups?
This usually means individuals aged under 18, and/or protected adults (i.e. an individual aged 16 or over in receipt of one or more registered care services; health services; community care services; or welfare services. University of Stirling students aged 16 or over are not considered to be a vulnerable group.)
If YES, membership of the Protecting Vulnerable Groups (PVG) scheme, or a Disclosure may be required. If applicable, provide confirmation or explain how will this be obtained? You must not start the research until you become a PVG member/get a Disclosure.
Click here to enter text

If you will NOT be applying for a PVG, explain how you will ensure the safety of those involved in the research who are in this category (e.g. describe the particular ethical issues involved and how you will address these or explain within ethical (e.g. British Psychological Society) guidance.)
Click here to enter text

Yes ☐
No ☒

C4. Consent and permission procedures
Attach all relevant documents, including participant information and consent sheets, scripts for oral consents, a debriefing document and evidence of permission (if applicable/required, see templates).
a) If written consent will not be obtained, justify it here.

Big data analysis makes it impractical to contact all Twitter users for consent (>thousands of Tweets). The Twitter platform's T&C's state that users' data could be used for research, so there is a level of consent given. We will ensure no individual will be identifiable (by name or location) in the published outputs. See section for C2.c. for more details.



Version date: 04.11.2019

Only if there are any additional consent and permission procedures, not included in these documents:

b) Describe the additional procedures you will follow to obtain informed consent from the participants and/or third parties (e.g. permissions to conduct field sampling).

N/A

C5. Are there risks or foreseeable harms that may be caused to participants and/or third party (e.g. landowners, institutions, carers, family etc)? This may include psychological stress, anxiety, embarrassment, discomfort, or be physical, social, legal, economic or political.	**Yes** ☐ **No** ☒

If YES, complete the following:

a) Describe any known or foreseeable harms that the participants, or others, might be subject to during or as a result of the research.

Click here to enter text

b) In light of the above assessment of potential harms, explain why the risks are acceptable given the value or benefits of the research.

Click here to enter text

c) Outline the steps that may be taken to reduce or eliminate these risks or foreseeable harms.

Click here to enter text

C6. Will the proposed research involve deception, concealment or covert observation? (definition)	**Yes** ☐ **No** ☒

If YES, complete the following:

a) If deception is to be used, justify the use of deception and indicate how participants will be debriefed.

Click here to enter text

b) If concealment or covert observation is to be used, justify the need to use these methods.

Click here to enter text

C7. Does the proposed research involve interviews, focus groups or questionnaires? (E.g. online surveys, social media analyses)?	**Yes** ☐ **No** ☒

If YES, attach copies of questionnaires, interview or focus group guides etc. and/or provide references to any existing questionnaires:

Click here to enter text

If the research design is emergent, and/or you are unable to attach relevant documents please explain:

Click here to enter text

C8. Briefly describe the methods of data analysis and data storage (see Guidance on Research Data). The University of Stirling requires that research data is securely preserved in an appropriate format for a minimum of 10 years, or longer if specified by the funder.

Whenever practical, data will be held in anonymised formats. All project data will be held on secure Stirling University systems (Onedrive) that are backed up on a daily basis. All local computers and laptops will be password protected. No unnecessary personal details will be collected or stored in accordance GDPR guidelines. The anonymised data will be stored in STORRE for a minimum of 10years in accordance to University guidelines.

Datasets will be analysed using data-mining techniques to look for temporal, and sentiment relatedness to opinions on Covid-19 related public health measures. Sentiment analysis will use natural language processing techniques resulting in a compound sentiment score indicating strength of negativity, positivity or neutrality of each message. Emotion analysis will be used on the text in order to quantify the text into 8 possible emotions.

Data will contain the time the post was created and the text. Data (via a CSV format) will then uploaded onto a secure University of Stirling server Sharepoint group which only Myself, Wendy and Gozde have access to. No personal data are collected therefore none will be shared with the research team. The code used to extract and analyse the data will be uploaded to Github for reproducibility purposes. None of the data collected will be shared on Github. SharePoint is Tier D-compliant. This includes the following standards: ISO 27001, ISO 27018, SSAE16 SOC 1 and SOC 2, HIPAA, and EU Model Clauses (EUMC). SharePoint and Teams enforce team-wide and organization-wide two-factor authentication, single sign-on through Active Directory, and encryption of data in transit and at rest. The data will be stored in SharePoint and are backed by SharePoint encryption.
C9: How will the results from this study (including feedback to participants) be disseminated? Results disseminated amongst only those involved in project via Sharepoint. Results will be written up and may be presented at academic conferences and in a research publication.

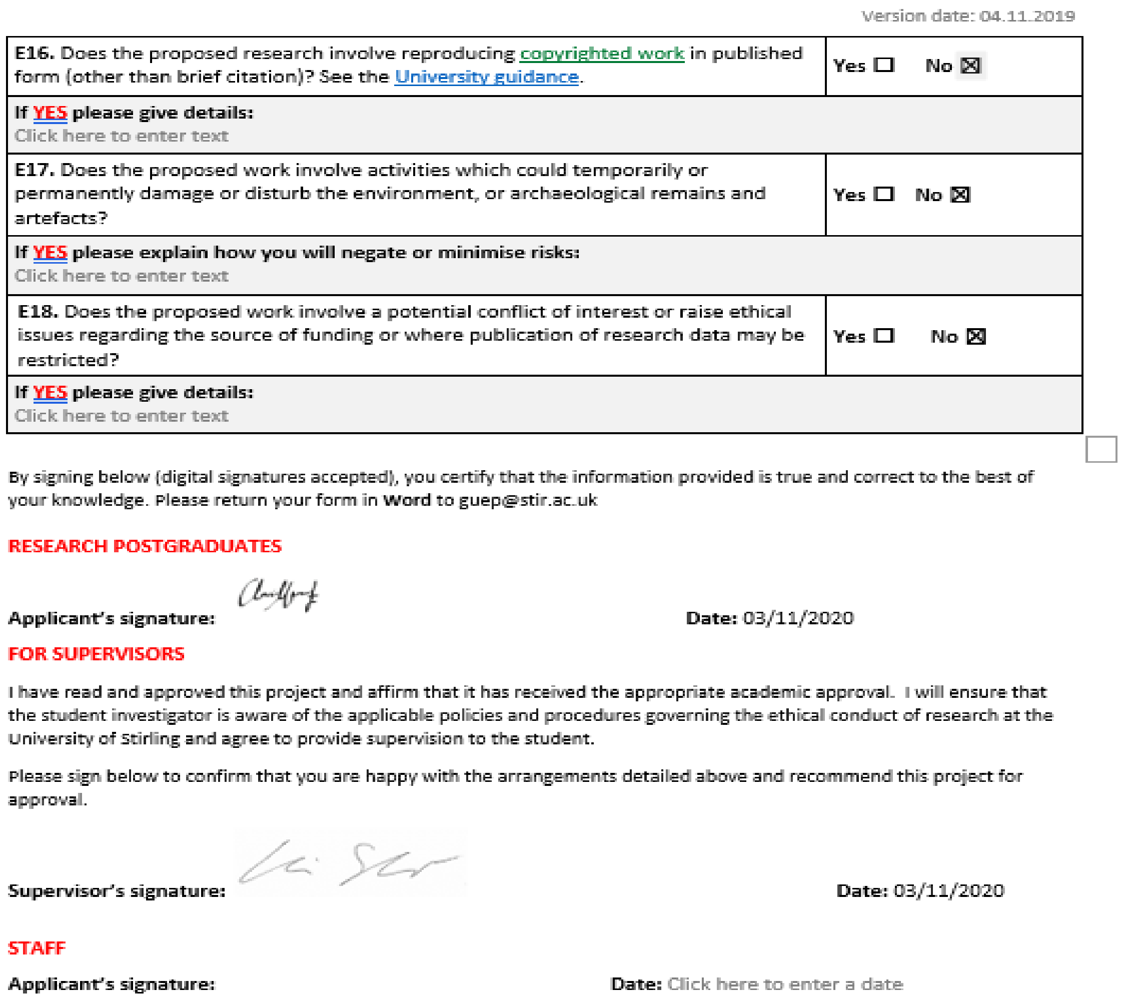
SECTION D: Research involving or impacting on animals

D1. Does your research involve animals?	**Yes ☐ No ☒**
If YES please also submit an application to the University AWERB (click here) – these applications can run in parallel.	

SECTION E: Data protection, copyright and other considerations

Applicants must confirm that they have read and understood the University's guidance on GDPR and that the necessary steps have been considered to protect the data of the participants of your research	**Yes ☒**
E1. Does your proposed research involve vulnerable groups? This usually means individuals aged under 18, and/or protected adults (i.e. an individual aged 16 or over in receipt of one or more registered care services; health services; community care services; or welfare services. University of Stirling students aged 16 or over are not considered to be a vulnerable group.)	**Yes ☐ No ☒**
E2. Does the research involve large scale processing of criminal convictions or special categories of personal data (health, racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, sex life or sexual orientation)?	**Yes ☐ No ☒**
E3. Does the processing of personal data involve new technologies or novel applications of existing technologies	**Yes ☐ No ☒**
E4. Are you processing biometric or genetic data?	**Yes ☐ No ☒**
E5. Are you combining or matching personal data obtained from multiple sources?	**Yes ☐ No ☒**
E6. Are you tracking individual's geolocation?	**Yes ☐ No ☒**
E7. Are you using personal data in a way that could significantly affect or have an impact on an individual?	**Yes ☐ No ☒**

E8. Are you processing personal data that has been obtained from the data subject, without their knowledge?	**Yes** ☐ **No** ☒
E9. Could the research jeopardise the physical health or safety of individuals?	**Yes** ☐ **No** ☒
E10. Are you doing systematic monitoring of publicly accessible areas?	**Yes** ☐ **No** ☒
E11. Does the proposed research involve profiling or automated decision-making on a large scale where significant decisions are made impacting on people? (Profiling is the automated analysis of personal data to classify people into different groups)	**Yes** ☐ **No** ☒
If you answer YES to any of the questions E1 to E11 you will need to complete a Data Protection Impact Assessment, the form and guidance for this are available here. A copy of your completed form should be attached to this GUEP application for the Committee's information.	
E12. Does the proposed research involve the remote acquisition of data from or about human participants using the internet and its associated technologies (e.g. online surveys, social media analyses)? If so please ensure you have read the remote acquisition of data guidance.	**Yes** ☒ **No** ☐
If YES please give details: Text analysis of twitter data	
E13. Does the research involve collecting, or accessing records of, personal or confidential information concerning identifiable individuals.	**Yes** ☒ **No** ☐
If YES: a) Describe how the anonymity of participants and the confidentiality of data will be protected and the specific methods to be used for this (e.g. data coding systems), both during the research and in the dissemination of findings. The collection of unnecessary personal data will be minimised, and only the time/date and text of the post will be derived in the dataset. Data will be stored and analysed securely. Data sets will be aggregated at the earliest possible stage of analysis and no individuals will be identifiable from the project outputs. b) Who will have access to identifiable information? Describe any potential use of the identifiable data by others. Only the 3 members of the research team will have access to the raw tweet data c) Indicate if there are any conditions under which privacy or confidentiality cannot be guaranteed (e.g. focus groups; mandatory reporting) or, if confidentiality is not an issue in this research, explain why. No, post analyses, all data will be aggregated to a level that will not allow individuals to be identified.	
E14. Does the proposed research involve the recording of participants through the use of audio-visual methods?	**Yes** ☐ **No** ☒
If YES please give details: Click here to enter text	
E15. Will external contractor be involved (e.g. transcription services, interpreters, fieldworkers)?	**Yes** ☐ **No** ☒
If YES comment on their compliance with ethical requirements and data protection legislation: Click here to enter text	



Version date: 04.11.2019

E16. Does the proposed research involve reproducing copyrighted work in published form (other than brief citation)? See the University guidance.	**Yes** ☐ **No** ☒
If YES please give details: Click here to enter text	
E17. Does the proposed work involve activities which could temporarily or permanently damage or disturb the environment, or archaeological remains and artefacts?	**Yes** ☐ **No** ☒
If YES please explain how you will negate or minimise risks: Click here to enter text	
E18. Does the proposed work involve a potential conflict of interest or raise ethical issues regarding the source of funding or where publication of research data may be restricted?	**Yes** ☐ **No** ☒
If YES please give details: Click here to enter text	

By signing below (digital signatures accepted), you certify that the information provided is true and correct to the best of your knowledge. Please return your form in **Word** to guep@stir.ac.uk

RESEARCH POSTGRADUATES

Applicant's signature: **Date:** 03/11/2020

FOR SUPERVISORS

I have read and approved this project and affirm that it has received the appropriate academic approval. I will ensure that the student investigator is aware of the applicable policies and procedures governing the ethical conduct of research at the University of Stirling and agree to provide supervision to the student.

Please sign below to confirm that you are happy with the arrangements detailed above and recommend this project for approval.

Supervisor's signature: **Date:** 03/11/2020

STAFF

Applicant's signature: **Date:** Click here to enter a date