Social Media Analysis for Social Good

Olanrewaju Mohammed Tahir Aduragba

Supervised by: Professor Alexandra I. Cristea

A thesis presented for the degree of

Doctor of Philosophy in Computer Science

Department of Computer Science
The University of Durham
United Kingdom

March 2023

In the name of Allah, the Compassionate, the Merciful.

I dedicate this thesis to my beloved parents, Engr. Abubakar Aduragba, Alhaja Aminat Aduragba and my darling wife, Shakirah Mustapha.

Abstract

Data on social media is abundant and offers valuable information that can be utilised for a range of purposes. Users share their experiences and opinions on various topics, ranging from their personal life to the community and the world, in real-time. In comparison to conventional data sources, social media is cost-effective to obtain, is up-to-date and reaches a larger audience. By analysing this rich data source, it can contribute to solving societal issues and promote social impact in an equitable manner. In this thesis, I present my research in exploring innovative applications using Natural language processing (NLP) and machine learning to identify patterns and extract actionable insights from social media data to ultimately make a positive impact on society.

Firsts, I evaluate the impact of an intervention program aimed at promoting inclusive and equitable learning opportunities for underrepresented communities using social media data. Second, I develop EmoBERT, an emotion-based variant of the BERT model, for detecting fine-grained emotions to gauge the well-being of a population during significant disease outbreaks. Third, to improve public health surveillance on social media, I demonstrate how emotions expressed in social media posts can be incorporated into health mention classification using an intermediate task fine-tuning and multi-feature fusion approach. I also propose a multi-task learning framework to model the literal meanings of disease and symptom words to enhance the classification of health mentions. Fourth, I create a new health mention dataset to address the imbalance in health data availability between developing and developed countries, providing a benchmark alternative to the traditional standards used in digital health research. Finally, I leverage the power of pretrained language models to analyse religious activities, recognised as social determinants of health, during disease outbreaks.

Acknowledgements

I am deeply grateful to Professor Alexandra Ioana Cristea, whose exceptional guidance and support have been integral to the completion of this work. Not only have you provided unwavering academic support, but also pastoral care to me and my family. Your willingness to go above and beyond for us has been a constant source of inspiration, and I will always be grateful for your kindness and generosity.

I owe an immeasurable debt of gratitude to the love of my life, Shakirah. She has selflessly put aside her own aspirations to support mine while simultaneously caring for our beloved children, Sumayyah and Safiyyah. Shakirah has been my rock, my constant source of encouragement, love, and understanding. I cannot thank her enough for her unwavering support and the sacrifices she has made on my behalf.

I am also indebted to my parents, Engr. Abubakar Aduragba and Alhaja Aminat Aduragba, my sister Tawakalit, and my brothers, Ahmed and Kabir, who have always been there for me, offering their steadfast support and dedication, even when I was unable to reciprocate. Their constant encouragement, unwavering prayers, and support have been essential to my success. Thank you for all you have done for me.

I would like to express my sincere gratitude to my friends Dr. Ahmed Alamri, Dr. Jialin Yu, Dr. Mohammad Alsheri, Khalil Alsaeed, Seyma Yucer, Muna Almushyti, Laila Al Rajhi, and Zhongtian Sun. Each of you has played a unique role in my academic and personal life during my time at Durham. I am grateful for your support and the memories we have shared.

Finally, I would like to thank the Tertiary Education Fund Nigeria and Kwara State University Nigeria for providing me with the opportunity and financial support to pursue this research. Their investment in my academic and professional development is greatly appreciated.

Contents

Declaration

The work in this thesis is based on research carried out within the Artificial Intelligence and Human Systems (AIHS) group at the Department of Computer Science, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification, and it is all the author's work unless referenced to the contrary below.

Note on Publications Included in This Thesis: Some of the work presented in this thesis have been previously published or is in communication in the following peer-review publications, and is used in the chapters as indicated below:

1. *Aduragba, O. T., and Cristea A. I. 2019.* ***Research on Prediction of Infectious Diseases, their spread via Social Media and their link to Education****. In Proceedings of the 2019 4th International Conference on Information and Education Innovations.* (Published, Contributing to Chapter 2) [6]

2. *Aduragba O. T., Yu J., Cristea A., and Shi L. 2021.* ***Detecting Fine-Grained Emotions on Social Media during Major Disease Outbreaks: Health and Well-being before and during the COVID-19 Pandemic****. In AMIA Annual Symposium Proceedings.* American Medical Informatics Association. (Published, Contributing to Chapter 5) [9]

3. *Aduragba O. T., Yu J., Cristea A. I. 2023.* ***Incorporating Emotions into Health Mention Classification Task on Social Media*** *2023 International Joint Conference on Neural Networks (IJCNN).* IEEE. (In communication, Contributing to Chapter 6)

4. *Aduragba O. T., Yu J., Cristea A. I. and Long Y. 2023.* ***Improving Health Mention Classification Through Emphasising Literal Meanings: A***

Study Towards Diversity and Generalisation for Public Health Surveillance. *In Proceedings of the ACM Web Conference.* (Accepted, Contributing to Chapter 7)

5. *Aduragba O. T., Cristea A. I., Phillips P., Kurlberg J. and Yu J. 2023.* ***Religion and Spirituality on Social Media in the Aftermath of the Global Pandemic***.*å In 2023 International Joint Conference on Neural Networks (IJCNN).* IEEE. (In communication, Contributing to Chapter 8)

6. *Aduragba O. T., Yu J., Cristea A. I., Hardey M., and Black S. 2020.* ***Digital Inclusion in Northern England: Training Women from Underrepresented Communities in Tech: A Data Analytics Case Study***. *In 2020 15th International Conference on Computer Science & Education (ICCSE).* IEEE. (Published, Contributing to Chapter 4) [8]

Note on Publications Not Included in This Thesis: As well as the above papers, the following works have been published during the period of research for this thesis; they have helped my deeper understanding of the work towards this thesis; however, these publications do not fit into the narrative of this thesis and have not been included in the text.

- *Aduragba, O. T., Yu J., Senthilnathan G., and Crsitea A. 2020.* ***Sentence Contextual Encoder with BERT and BiLSTM for Automatic Classification with imbalanced medication tweets.*** *In Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pp. 165-167.

- *Yu J., Aduragba O. T., Sun Z., Black Sue, Stewart C., Shi L., and Cristea A. I. 2020.* ***Temporal Sentiment Analysis of Learners: Public Versus Private Social Media Communication Channels in a Women-in-Tech Conversion Course***. *In 2020 15th International Conference on Computer Science & Education (ICCSE).* IEEE.

- *Yu, J., Cristea, A.I., Harit, A., Sun, Z., Aduragba, O.T., Shi, L. and Al Moubayed, N., 2022, July.* ***Efficient Uncertainty Quantification for Multilabel Text Classification***. *In 2022 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8).* IEEE.

- *Yu, J., Cristea, A.I., Harit, A., Sun, Z., Aduragba, O.T., Shi, L. and Al Moubayed, N., 2022, July.* ***INTERACTION: A Generative XAI Framework for Natural Language Inference Explanations***. *In 2022 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8).* IEEE.

List of Figures

List of Tables

Nomenclature

API Application Programming Interface

BERT Bidirectional Encoder Representations from Transformers

NLP Natural language processing

LSTM Long short-term memory

LDA Latent Dirichlet Allocation

AI Artificial Intelligence

SDG Sustainable Development Goals

HMC Health Mention Classification

AI4SG Artificial Intelligence for Social Good

DNN Deep Neural Network

MLM Masked language modeling

NSP Next sentence prediction

NHMD Nairaland Health Mention Dataset

PLM Pre-trained Language Model

MTL Multi-task learning

NLP4SG Natural Language for Social Good

LDA Latent Dirichlet Allocation

WHO World Health Organisation

BOW Bag of words

SVM Support Vector Machine

Chapter 1

Introduction

Social media has revolutionised how we communicate and exchange information, connecting billions of users who generate and share vast amounts of data on everyday matters [37]. This creates an opportunity to analyse and understand the various aspects of society in unprecedented detail. By leveraging the power of social media data, we can gain valuable insights into the needs, concerns, and behaviours of individuals and communities and use this information to inform policy decisions and positively impact society [267]. The widespread use of social media has allowed greater access to real-time data at a low cost, making it suitable for automatic analysis [57].

This thesis aims to explore different approaches to analyse data derived from social media for applications with social impact. Specifically, I explore novel applications using Natural Language Processing (NLP) and machine learning to process large-scale social media data to address some of our society's most pressing challenges.

The main goal is to contribute to the growing body of knowledge on the role of social media for social good and to demonstrate how this data can be used to improve the lives of individuals and communities.

Although there is no universally agreed definition of social good in the field of Computer Science [241], a common approach for research in this area is to adopt the UN Sustainable Development Goals (SDG), which consists of a set of 17 objectives that

aim to guide society towards a more equitable, prosperous, and sustainable world [135]. Existing research focus on applying Artificial Intelligence (AI) techniques often referred to as Artificial Intelligence for Social Good (AI4SG), that address some of the social issues facing our world in areas such as healthcare [107], education [42], agriculture [231], social welfare, and justice [49]. Social media analysis can be thought of as a sub-field of artificial intelligence that focuses on the automatic analysis of large amounts of text data generated by users on social media platforms [70].

In this thesis, I adopt the definition of the umbrella term AI4SG as defined by Cowls et al [55]: *actions that prevent, mitigate or resolve problems affecting human life or wellbeing, enable socially sustainable developments while not introducing new forms of harm and/or amplifying existing disparities and inequities.* Based on this, I present novel computational methods for social media analysis to understand better and draw actionable insights that deliver positive social impact. I will leverage the huge amount of social media data generated by billions of people and novel NLP methods to analyse such data automatically. When successfully applied, social media analysis has a great potential impact that can ensure healthy lives, promote well-being for individuals and promote inclusive and equitable education.

Building upon the definition of AI4SG by Cowls et al [55], which describes AI4SG as *actions aimed at preventing, mitigating, or resolving issues affecting human life or wellbeing, enabling socially sustainable developments without introducing new forms of harm or amplifying existing disparities and inequities*, I present novel computational methods for social media analysis to gain deeper understanding and extract actionable insights for delivering positive social impact. In this thesis, I harness the voluminous social media data generated by billions of people worldwide, alongside innovative NLP techniques to automatically analyse such data. If successfully applied, social media analysis has the potential to foster inclusive and equitable education, ensure healthy lives and promote well-being for individuals thereby contributing to social good.

1.1 Research Motivation

As social media grows in popularity, it generates enormous amounts of data daily. The sheer volume and variety of user-generated content and the user interaction network create unique opportunities for understanding individuals, communities, and society as a whole. One key advantage of social media data is its real-time nature, providing current information and commentary on significant events and happenings in society. Additionally, social media posts often include geo-location information, which can be valuable for large-scale geographical assessments at the city, state or national level [88].

The contents of social media, including micro blogs on Twitter*, status updates on Facebook†, comments on YouTube‡ and user forums on Reddit§, are mainly in the form of textual data [98]. This offers the potential for applications in NLP [24]. In recent years, NLP has achieved remarkable success across several domains, attracting widespread interest from researchers. NLP applications have been used to solve a wide range of practical problems such as speech recognition [219] and machine translation of low-resource languages [5]. In general, the success of machine learning presents the possibility of utilising the large amount of data generated from social media as a data source for tracking social phenomena.

However, analysing social media data is a non-trivial task and poses challenges for NLP[189]. Social media often uses unstructured nature of language, including misspellings, the adoption of informal words such as slang, and deviations from standard grammar [208]. Additionally, social media posts are often brief, covering a wide range of topics, making it challenging to develop tools specific to this domain or to use contextual information to clarify the meaning of the content. It is therefore crucial to optimise methods to deal with the unique characteristics of social media

*https://twitter.com
†https://www.facebook.com
‡https://www.youtube.com
§https://www.reddit.com

data.

The use of social media analysis in various fields has proven to be impactful in promoting public health and addressing social issues. For example, the analysis of social media data has been used to monitor the spread of diseases and identify potential health risks associated with certain behaviours [99, 235, 187, 11]. Additionally, researchers have utilised social media data to identify instances of discrimination [26, 30] and shed light on systemic issues related to immigration [115, 87, 158]. These applications demonstrate the potential for social media analysis to have a significant positive impact on society.

The primary motivation of the work presented in this thesis is to examine how social media content can be analysed for the benefit of society in two key areas: education and healthcare. In education, I measure the impact of an intervention program that promotes inclusive and equitable learning opportunities for underrepresented communities In healthcare, I propose to monitor health and well-being during disease outbreaks and effectively detect health reports on social media. Overall, I explore ways to address the unique characteristics of social media data to improve the performance of related tasks through a combination of NLP and machine learning techniques.

1.2 Research Questions

In this thesis, I focus on *how social media data can be automatically analysed to address social issues that can lead to positive social impact.* Based on this, I have formulated the following research questions:

- **RQ1:** What is the impact of technology retraining programs for women from underrepresented communities, as reflected on social media?
- **RQ2:** How can social media be used to detect fine-grained emotions during major disease outbreaks?

- **RQ3:** How can health mentions be detected on social media to track health-related conversations?
- **RQ4:** How can the gap in health-related social media data between developed and developing countries be narrowed?
- **RQ5:** How has the COVID-19 pandemic impacted religious and spiritual practices in the UK, as reflected on social media?

The research areas in this thesis were selected to offer a broad coverage within AI4SG, while considering data availability and the topical relevance of the issues at the time of research. For instance, at the commencement of this research, the TechUPWomen* programme was being implemented. This initiative, dedicated to training women from underrepresented demographics in the field of technology, offered a prime opportunity to delve into the domain of education. The examination of this programme allowed for a nuanced exploration of methods to promote inclusive and equitable learning opportunities for communities typically underrepresented in these areas.

Concurrently, the emergence of the COVID-19 pandemic drew the global focus towards public health and disease control. This major disease outbreak prompted a redirection of part of this research towards healthcare, exploring the use of social media analysis for monitoring health and well-being during disease outbreaks and for effective detection of health reports on social media platforms.

These research areas, while broad in their coverage withinAI4SG, are also responsive to pressing societal challenges that were particularly relevant during the course of this research. Detailed exploration of these research questions and the insights gleaned from them can be found in Chapters 4, 5, 6, 7 and 8.

*https://techupwomen.org/

1.3 Research Objectives

To answer the above research questions, I have established the following objectives:

- To assess the impact of training women from underrepresented communities in tech using social media data. This addresses **RQ1** in Chapter 4.
- To develop and evaluate methods for detecting fine-grained emotions on social media and introduce emotion-specific pre-training objectives. This is important to address **RQ2**. This is further explored in Chapter 5
- To investigate how emotions can be incorporated into health mention detection. This addresses **RQ3** and it is explored in Chapter 6.
- To investigate how detection of literal meanings can be used to classify social media posts that contain health mentions. This is intended to answer **RQ3**. This is studied in Chapter 7.
- To create a dataset that aims to narrow the gap in health-related social media data between developed and developing countries. This objective addresses **RQ4** and it is further explored in Chapter 7.
- To explore the use of social media for tracking religious activities during the COVID-19 pandemic. This addresses **RQ5** in Chapter 8.

1.4 Research Contributions

With this thesis, I make the following contributions:

- Measure the impact of a tech retraining programme for women from underrepresented communities (Chapter 4).

- Propose EmoBERT, a new emotion-based variant of the BERT transformer model, able to learn emotion representations and outperform the state-of-the-art (Chapter 5).
- Propose how to *implicitly* and *explicitly* incorporate emotional information into health mention detection on social media to improve performance (Chapter 6).
- Design and develop a new health mention dataset for underserved population (Chapter 7).
- Propose a novel literal emphasised multi-task learning framework for the health mention classification on social media (Chapter 7).
- Introduce a method to track the shift of religious activities online before and during the COVID-19 pandemic (Chapter 8).

1.5 Thesis Outline

In **Chapter 2**, I present an overview of related work to social media analysis. The review of literature is focused on applications of social media for social good. Furthermore, this chapter discusses the ethical considerations of using social media data for research.

Chapter 3 provides a technical background on NLP and machine learning methods, which are necessary for understanding the modelling approaches used in this research.

Chapter 4) tracks the impact of a technology retraining program for women from underrepresented communities by analysing social media content from both a public channel (Twitter) and a private channel (Microsoft Teams). This chapter aims to understand the impact of the retraining program and provide insights on the effectiveness of using social media to communicate and evaluate the program.

Chapter 5 focuses on detecting fine-grained emotions to understand the emotional health and wellbeing of a population during the COVID-19 pandemic. In particular, this chapter presents EmoBERT, a new emotion-based variant of the BERT model that is able to learn emotion representations and outperform the state-of-the-art. Additionally, I conduct a fine-grained emotion analysis of the pandemic's impact on London before and during the pandemic.

Chapter 6 investigates the potential of incorporating emotion information into the task of detecting health information on social media, which is useful for public health surveillance. The chapter explores two different approaches for incorporating emotions: intermediate task fine-tuning (*implicit*) and multi-feature fusion (*explicit*).

In **Chapter 7**, I introduce a new health mention dataset for people from underrepresented communities to address the gap in the availability and quality of health-related social media data between developed and developing countries. This chapter also investigates modeling the literal meanings of disease or symptom words, to improve the performance of detecting health information on social media.

Chapter 8 explores the impact of the COVID-19 pandemic on religious and spiritual practices in the UK using NLP techniques, including language modeling. The chapter aims to understand how these practices, which can be a social determinant of health, have been influenced by the pandemic.

Chapter 2

Related Work

2.1 Social media analysis

Social media platforms have rapidly expanded throughout the world in recent years, attracting billions of users and becoming a prominent source of up-to-date information and commentary on current events in people's lives. Twitter, with over 500 million users sending more than 500 million tweets each day, is just one example of the many social media sites available. People can use social media to report anything happening in their lives, from personal health updates [256] to the latest news on a community's response to a disease outbreak [246].

With the prevalence of mobile communications and user interface design improvement, social media has broken the communication barrier between the real and virtual world [145]. Social media platforms such as Facebook[†], Twitter[‡], LinkedIn[§], YouTube[¶], and Instagram[‖] offer users a range of options for generating and sharing content, including texts, audio and video clips, and pictures. However, the various forms of data and their real-time nature require a suitable approach to process them.

[†] https://facebook.com
[‡] https://twitter.com
[§] https://linkedin.com
[¶] https://youtube.com
[‖] https://instagram.com

Social media analysis involves the automatic analysis of large amounts of data generated by users on social media platforms using associated methodologies [269]. In the last decade, there has been an increasing trend in analysing social media data to draw useful insights about phenomena that affect humans and their communities [39, 77]. Such work leverages techniques from NLP [253, 45] and machine learning [67, 198] to extract valuable information.

2.2 NLP methods for social media analysis

NLP techniques have gained widespread success in processing natural language data [61, 204], leading to their increased popularity in social media research [185]. One of the most commonly used approaches to NLP in social media analysis is text classification, which involves assigning categories to individual social media posts based on the application [188]. This task is essential since it enables efficient analysis of large volumes of real-time social media data that cannot be manually coded by humans [156].

Sentiment analysis is a well-established field in NLP that aims to determine the sentiment expressed in a social media post, whether it is positive, neutral, or negative [163]. This task is particularly useful for studying and analysing people's sentiment about real-world phenomena that affect them, given that one of the primary uses of social media is to express opinions. In [97, 127, 50], sentiment analysis was used to analyse people's responses to public health guidelines, allowing researchers to gain insight into how individuals perceive and react to such guidelines. Furthermore, researchers have also used sentiment analysis to predict the movement of stock prices by analysing people's sentiment towards real-world events and news [59, 180].

Recently, researchers have focused on determining the sentiment towards a specific target entity [163]. Target-based sentiment analysis is based on the notion that a social media post can express sentiment towards multiple aspects [10]. For example,

in [212], the researchers proposed a new dataset and baseline methods to determine sentiments towards different aspects such as transportation, shops and restaurants in urban neighbourhoods. Another similar line of research in this field is stance detection, which focuses on identifying an author's position (favour or against) towards a target [14].

Over time, sentiment analysis has evolved from detecting polarity to identifying more discrete emotions such as sad, fear, trust, or joy in text [223]. Moreover, different events or situations stir a diverse range of emotions in people, and people often turn to social media platforms to express their emotions. Detecting fine-grained emotions has been extensively studied on social media, with most studies focusing on two main emotion categories: Ekman's 6 basic emotions (anger, surprise, disgust, enjoyment, fear, and sadness) [66] and Plutchik's 8 emotions (anger, anticipation, joy, trust, fear, surprise, sadness, and disgust) [195]. In [229, 201, 214], researchers analysed people's emotions towards different public health topics such as treatment, vaccination, and mental health.

Emotion analysis and sentiment analysis are essential areas in social media analysis, and the results of these tasks can be utilised to draw useful insights about different topics. For this reason, part of my work in this thesis focus on analysing emotions and sentiments to improve the understanding of social media data with respect to its applications for social good (see Chapter 4 & 5).

Apart from these traditional NLP tasks, researchers have also focused on classifying social media based on their content. For example, researchers in [101, 218, 20] analysed the contents of social media posts for disaster analysis to determine if they were informative, personal, or not related to disaster. Some researchers have also explored detecting fake news on social media [146, 116, 263]. In this thesis, I explore detecting health related content on social media to support public health surveillance (see Chapter 6 & 7).

2.3 Social media analysis for social good

In recent years, there has been a growing interest in the analysis of social media data for applications with social impact [248]. This trend is echoed by similar efforts in the domains of AI and NLP, which have been denoted as AI4SG [241] and Natural Language for Social Good (NLP4SG) [76] respectively. These broader initiatives are well established and have shown the potential to address crucial societal challenges [233, 36].

One important application of social media analysis for social good is in the field of public health. Approaches vary across different public health themes and social media platforms such as Twitter [65, 110], Reddit [186, 243], Facebook [69, 203], and Instagram [221, 179]. Disease surveillance is one of the primary applications of social media data for health. Surveillance is an essential means of tracking population-level health outcomes. According to the World Health Organisation (WHO), surveillance is "the cornerstone of public health security" [184]. Traditionally, researchers carry out health surveillance by analysing clinical records from clinics, hospitals, mortality data, laboratory reports, and surveys [238]. However, these data sources have lag times of weeks which is crucial in stopping the spread of diseases. Social media platforms have been demonstrated to be early warning systems for disease outbreaks and sensitive to trends that conventional health surveillance methods might otherwise miss [177].

The task of surveillance on social media focuses on extracting health-related posts to estimate cases of diseases or monitor disease spread [119, 268]. Previous work showed that classifying social media posts as being related or not related to a disease or health condition provided promising surveillance results [130]. Early research in this domain focused on estimating common illnesses like influenza-like illnesses [4, 213, 249]. Researchers used tweets to detect influenza epidemics [260, 165], while another study used a combination of non-traditional data sources, including tweets, to forecast Zika incidences [157]. This thesis will extend the boundaries of

these efforts by proposing advanced methods to detect health-related conversations and emotions on social media during disease outbreaks (RQ2 and RQ3).

Furthermore, social media provides novel sources of data that are valuable in combating health challenges across several communities. For example, Public Health England announced the inclusion of internet-based search queries as a means of monitoring influenza-like illnesses (ILI) rates in England. This is as a result of the development of a FluDetector model which combines historical Royal College of General Practitioners (RCGP) ILI data and social media data to estimate daily ILI rates [131, 274].

In the context of COVID-19 pandemic, social media analysis has already made significant contributions towards addressing the public health crisis and bringing about positive social outcomes. Some of the works include studying vaccine hesitancy [181, 22], surveying public attitudes [46, 15], assessing mental health [143, 79], detecting or predicting COVID-19 cases [139, 140, 176] and analysing government responses to the pandemic [129, 144].

Generally, current social media health surveillance research primarily focuses on developed nations, leaving a gap in the understanding of health discourse in developing countries. This research aims to reduce this disparity (RQ4), thereby improving our understanding of global public health narratives and ensuring equitable health data representation.

Social media analysis for social good has also been applied in the field of disaster response and humanitarian aid [155, 182]. By automatically analysing social media data in real-time, researchers can identify areas in need of assistance and coordinate relief efforts more effectively [102]. This can be particularly useful in the aftermath of natural disasters, where timely and accurate information can be critical to saving lives.

In the area of social welfare, social media data has been used to track and predict poverty to plan efficient distribution of wealth within a community [35]. In [41],

researchers investigate the use of social media in college teaching and how it affects educational outcome. Another application in education looked at how social media can improve student retention in Massive Open Online Courses (MOOC) [271]. This thesis seeks to further explore the education domain, particularly how technology retraining programs impact women from underrepresented communities, as reflected on social media (RQ1).

On the other hand, there have been counterproductive actions bringing danger to social media users. For example, a survey of Pew Research Centre revealed that 73 percent of online internet users experienced online harassment, with 40 percent targeted personally*. To promote equality and rights of person online, researchers have equally been analysing social media data to combat detect abusive language, hate speech and cyber bulling online [78].

Following the successful application across several domains including health and education, this thesis will explore analysing social media data to promote healthy lifestyles and well-being for individuals, while also fostering inclusivity and equity in education.

*https://www.pewresearch.org/internet/2015/01/09/social-media-update-2014/

Chapter 3

Methodology

3.1 Overview

This thesis is centered on the analysis of social media data with the aim of promoting positive social impact. To this end, I have conducted various experiments using NLP and machine learning techniques to gain valuable insights from social media data. In this Chapter, I detail the various types of social media data that I analysed, and the process of collecting them. Additionally, I describe the preprocessing steps that were taken to prepare such data prior to performing the relevant tasks. Finally, I explain the text representation approach I employed in this thesis and the ethical considerations that need to be addressed when analysing social media content.

Text Classification The majority of experiments performed in thesis to automatically analyse social media data involve text classification. Text classification is the task of automatically categorising a text document into a predefined set of labels. In a supervised learning scenario, text inputs are represented using embedding techniques (see Section 3.3) and are passed into a classifier to predict the label. The text documents are typically annotated with target labels that the classifier uses for learning. The annotations for text classification tasks vary depending on the task; for instance, in Health Mention Classification (HMC) tasks,

text documents are labelled as either *health mention*, *other mention*, or *non-health*, usually by human annotators. The classification tasks conducted in this thesis include fine-grained emotion detection (see Chapter 5), HMC (see Chapters 6 and 7), and sentiment analysis (see Chapter 4). In the following sections, I will provide a detailed description of the text representation and classification techniques used.

Topic Modelling In Section 4.3.2, I applied topic modelling to social media texts to identify topics of discussion automatically. The role of topic models in text processing is to assemble correlated words into broader themes or subjects and subsequently detect the topics within a given document [225]. These models have seen considerable usage across diverse domains such as health, education, and other research fields that demand the analysis of large volumes of text to uncover inherent patterns and insightful information [199, 68, 81].

At their core, topic models are statistical models employing a generative and parameter estimation process to produce output, given the training data comprised of text documents [154]. The generative process in topic modelling operates under the presumption that a document is generated from a randomly chosen mixture of topics. For each word in the document, a topic is randomly selected from the topic distributions, and a word is then randomly assigned to the corresponding topic. This process is then reversed in the topic model, to estimate the parameters that represent the hidden topic structure most likely to have generated the document. The specific topic model used for identifying topics within social media texts is detailed in section 3.5.

3.2 Social media data

3.2.1 Twitter

Twitter* is a social media platform where its users post short messages (known as "tweets") to share "status updates" on their timelines. Twitter has become very popular since it launched in 2006, with as many as 238 million daily active users as of Q2 2022 †. Twitter users are allowed to send messages up to 280 characters ‡. Tweets mainly contain text, URL links, images, hashtags (words prefixed with # symbol, which is generally used to refer to the topic of the message) and user mentions (means of referring to other Twitter users in a tweet by using the @ symbol e.g. "@username"). Tweets can also provide information about user's geo-location. Users can interact with tweets by liking them or re-tweeting them (i.e. share another user's tweet on their timeline).

Twitter data collection

Like other social media platforms, Twitter is designed to share information with a public audience. Hence, Twitter data can be publicly accessed through the Twitter Application Programming Interface (API) §. The Twitter API offers several features to collect Twitter data, including tweet filtered streaming and tweet search. Users with access to the Twitter API can collect tweets in real time or search historical tweets that match a set of rules. The rules are created to match tweet attributes such as message keywords, hashtags, the location where tweets are posted or tweets from a specific user. In January 2021, Twitter launched a new API platform for academic researchers ¶ that grants access to Twitter's real-time and

*https://twitter.com

†`https://s22.q4cdn.com/826641620/files/doc_financials/2022/q2/Final_Q2'22_Earnings_Release.pdf`

‡Before 2017, Twitter only allowed messages up to 140 characters. Some of the tweet datasets used in this Thesis were collected before 2017.

§https://developer.twitter.com/en/docs/twitter-api

¶https://developer.twitter.com/en/products/twitter-api/academic-research

historical public data. Before then, the Twitter API was only limited to 1% of the overall tweet volume for filtered streaming and a sample of recent tweets published in the past 7 days for tweet search*. This limit could be a challenge, as shown in previous research [167]. However, the academic research access used mainly in this research eliminates this problem by providing more accurate, comprehensive and unbiased Twitter data. All tweets collection are subject to a monthly consumption cap based on the level of access. For Academic research access, the limit is 10 million tweets per month.

In this thesis, I have used the filtered stream and the full-archive search to collect tweets related to my work. I used a set of rules and search queries based on keywords or geo-location coordinates, depending on the performed task. The Twitter API returns Tweet objects in JSON format as a response to our queries. A tweet object contains various information, including a unique identifier (tweet ID), the text content of the tweet, unique identifier of the user who posted the tweet (user ID) and other metadata such as location where the tweet was posted from. Generally, I used mainly the text content of the collected tweets and, in some cases, the location associated with a tweet. I anonymised all user information before using a tweet, and our findings are reported on an aggregate level (see Section 3.6 for ethical considerations).

In some instances, I also used publicly available tweets related to our research. Twitter allows the redistribution of tweets for research purposes. However, the Twitter data sharing policy only permits sharing tweet IDs. I downloaded the actual content of the distributed tweet IDs through the Twitter API.

3.2.2 Reddit

Reddit is another social media platform where discussions are organised based on large communities, called subreddits. Unlike Twitter, discussions are usually

*Searching further back required a Premium API which comes at a high cost

around a specified theme or topic in communities. The communities are often monitored by one or more moderators to ensure that the contents posted by users are relevant to the community. As at January, 2021, there were more than 50 million daily active unique users and over 100,000 active communities*. Users can author posts (called submissions) to a subreddit and other users can respond to the submission with posts (called comments) to form a discussion thread. Reddit has a much larger character limit than Twitter with 10,000 characters for a comment and 40,000 characters for a submission.

Reddit data collection

Reddit like Twitter provides an open API† to access its data. However, I take advantage of a platform, Pushshift‡, that collects and maintains historical data from Reddit. Pushshift is a service provided for researchers that makes it easier than official Reddit API to collect Reddit data [27]. Data can be retrieved from the Pushshift database through queries which are based on search parameters including search term, specific subreddit or a date range. Search results are returned in JSON format. Each data object contains information about the ID of the post, title if post is a submission, body text and other metadata such as timestamps and subreddit where the post is created. For the most part of this thesis, I only collect available submission from the subreddit of interest and used the body text of the collected submissions. All posts containing reference to other users were anonymised (see Section 3.6 for ethical considerations). This research used Reddit data in Chapter 8 to filter tweets related to religious or spiritual activities.

*https://www.redditinc.com/press/
†https://www.reddit.com/dev/api/
‡https://pushshift.io/

3.2.3 Nairaland

Nairaland* is an online community-based forum designed for Nigerians, with a broad discussion platform that receives 33 million visits per month †. The forum, which was created in March 2005, has 2,946,061 registered members and 7,136,617 topics as of October 2022. The discussions on Nairaland are classified into "Nigerian forums" that revolve around topic areas pertinent to the Nigerian community, such as health, politics, and business. Users can create new topics in each forum, which others can respond to with comments. Posts mainly consist of text, with the occasional use of emojis, images, and URL links. There are no limits to the length of posts.

Nairaland data collection

Nairaland does not provide any official APIs for data collection, unlike popular social media platforms like Twitter or Reddit. Therefore, to obtain the data required, I utilised Beautiful Soup‡, a Python library for parsing HTML documents. Relevant HTML tags were identified for each post, and the text in the post, along with the time it was created, and the topic and forum it was created in, were extracted. To ensure anonymity, any names or references to other users in the post were anonymised. This thesis focuses solely on social media texts, hence, no information about the author of the post or images in the post were collected. The ethical considerations surrounding the use of social media data are discussed in 3.6.

3.2.4 Pre-processing

Due to the noisy nature from social media data, I carry out several pre-processing steps before performing any task. The type of pre-processing step depends on the

*`https://www.nairaland.com/`
†https://www.similarweb.com/website/nairaland.com/#overview
‡`https://www.crummy.com/software/BeautifulSoup/bs4/doc/`

performed task. Generally, the pre-processing steps involves reducing noise and transforming the data in to a more meaningful content to improve interpretability or performance of models applied on them. The major pre-processing steps used in this thesis are as follow:

Tokenisation is a fundamental pre-reprocessing step used to split any sequence of text into simpler units called tokens (e.g. words). This process can be non-trivial when handling user-generated texts. There are several tokenisers that have been developed to make tokenising social-media text easier [28].

Replacement of URLS, hashtags and user mentions with common tokens such as `<url>, <hashtag> and <user>`, as these provide little to no information for the language understanding tasks.

Replacement of social media specific symbols such as emojis with their semantic meanings. Specifically, I use the `emoji` package to translate emojis into text strings.

3.3 Text Representation

Representing natural language text as numerical data is a crucial step in NLP that enables machines to understand, analyse and manipulate language data. Representation learning in NLP involves creating numerical representations of text that capture its semantic and syntactic properties. Various methods such as bag-of-words [149], word embeddings [160], and contextualised word representations [193] can be used for text representation, depending on the specific NLP task and available computational resources. The quality of the text representation affects the performance of NLP models, and choosing the appropriate representation technique is essential for accurate and effective language processing. One of the popular techniques for generating text representations that has provided strong baselines is language model pre-training. In the following subsections, I describe the task of language modelling and the language models used in this thesis.

3.3.1 Language Model Pre-training

The goal of a language model is to predict the probability of word (or words) in a sentence. Based on statistical analysis of a given sentence, a language model is capable of looking at a word (or words) and predicting the next word most likely to follow. Or given a sentence with corrupted words, a language model is capable of reconstructing the original sentence. In essence, they are a useful component in NLP applications [142].

A benefit of language modelling is that it can be trained with any text corpus, which is freely available and without requiring hand-labelled annotation. For example, to train an auto regressive language model, the model takes the next word from a running text and use it as a supervision signal to form its prediction task. While performing this task, the language model learns an embedding representation for each word. The main concept behind this is that the embedding vectors of similar words will be in close proximity to one another in the embedding vector space. There are number of pre-training objectives proposed to learn better language representations.

Causal language modelling

Causal language modelling is the most common case for training language models where they are trained to predict the next token given the previous tokens. For example, given an input text, *I saw a cat on a* _____. The task of the model is to predict the word that is likely to come next given the previous words. This is sometimes also referred to as autoregressive language modelling. The training objective is formulated as:

$$P(X) = \sum_{n=1}^{N} log\, P(x_n|x_1, ..., x_{n-1}) \tag{3.1}$$

where X is a sequence of N tokens $(x_1.....x_N)$. The objective of the model is to provide good estimates for $P(X)$. Log probabilities is used to prevent numerical underflow when multiplying raw probabilities together [113]. The language model uses unannotated text from a large corpus to generate the training sequences. Specifically, the model takes the next word in a running text and uses it as label for the learning task.

This concept can be useful in essential language understanding tasks such as text generation, question answering or machine translation [111]. They also have practical applications in areas such as text correction and completion in keyboards [80], automated email response suggestion [117] and speech recognition [170].

One drawback of this training objective is that it is unidirectional i.e. the model only considers tokens at positions less than i to predict the i-th token. This fails to take into account the tokens on the right side of the i-th token to generate a complete contextual representation.

Mask language modelling

To train language models that are bidirectional i.e considers tokens both on the left-side and right-side of i to predict the i-th token, researchers proposed using a fill in the blank task. This was inspired by the cloze task [237]. Instead of trying to predict the next word, the task of the model is to predict the missing words in a sentence with some words removed. For example, given an input text, *I saw a* _____ *on a mat*, the model tries to predict the missing word using the rest of the sentence. Models that are trained with this objective are called Masked Language model (MLM) [61].

Given an original sequence of tokens $X = x_1.....x_N$ and the position of masked tokens $M = m_1, ...m_K$, where K is the number of masked tokens, the objective of MLM is formulated as:

$$P(X_M|X_{-M}) = \frac{1}{k}\sum_{k=1}^{K} log\, P(x_m|X_{-M}) \tag{3.2}$$

where X_M denote the set of masked tokens and X_{-M} denote the set of unmasked tokens in sentence X.

Same as causal language models, MLM uses unlabelled text from a large corpus to generate a training corpus. A random sample of tokens is chosen from each training sequence and used for learning. The chosen token is either replaced with a special token e.g. [MASK] or replaced with another token from the vocabulary or left unchanged.

Due to incorporating bidirectional learning, masked language models capture contexts of words from both left and right sides. As a result, they perform better on natural language understanding tasks, including those undertaken in this thesis: text classification.

Loss function

The standard loss function for training language models is cross-entropy loss [113]. Cross entropy loss is defined as:

$$L(\hat{y}, y) = -\sum_{k=1}^{K} y_k log \hat{y}_k \tag{3.3}$$

Where K is the total number of output classes. Here, the classes are the word tokens in the vocabulary. For causal language modelling, $\hat{y}_k$ is the probability that the model predicts the correct next token as defined in equation 3.1, while for masked language modelling, $\hat{y}_k$ is the probability of predicting the missing tokens as defined in equation 3.2.

Evaluation of Language models

An approach to evaluate the performance of a language model is to test how well it predicts unseen or missing texts. A good language model assigns the highest

probabilities to a correct sentence and low probabilities to an incorrect sentence. This can be measured using perplexity. Perplexity is a measurement of how well a probability model predicts a sample [84]. Given a test sentence T with N words and a language model θ, the perplexity, PP_θ is:

$$PP_\theta(T) = P_\theta(T)^{-\frac{1}{N}} \tag{3.4}$$

A high probability results to lower perplexity values and vice versa. A good language model will be reflective of real language usage. Hence, I use perplexity scores in Chapter 8 to analyse language models trained with tweets to measure faith-related engagements during the COVID-19 pandemic.

However, improvement in perplexity score does not frequently result to improvement on downstream tasks. Thus, perplexity is useful for comparing the capacity of different language models to recognise patterns in sequences, but it is not a good measure of evaluating the performance in natural language understanding tasks [84]. In this thesis, I evaluate the performance of different pre-trained language models based on their performance on text classification tasks in Chapters 5 and 6 in terms of F1 score.

Measuring similarity

The similarity between words or sentences represented as vectors can be calculated by computing the distance between the corresponding vectors. One of the benefits of this is that it can be used to retrieve relevant candidates based on a query when performing a semantic search. The main concept behind semantic search is to embed the query sentence and all documents within a corpus into the same vector space, and the closest embeddings to the query are found. Traditional methods perform this task by lexically matching keywords. However, it is also essential to understand the proper use of the words in context to retrieve the most similar sentences. Transformer-based language models such as BERT [61] consider the full

context of words in a sentence. Hence, they are useful for understanding the intent behind search queries.

A common and effective measure of similarity is cosine similarity, which measures the angle between the vectors' cosines [84]. Thus, the similarity between two documents $D_1 = w_1^1, w_2^1, \ldots, w_m^1$ and $D_2 = w_1^2, w_2^2, \ldots, w_n^2$ can be computed as:

$$sim_{cos}(D_1, D_2) = (\sum_{i=1}^{m} w_i^1) \cdot (\sum_{j=1}^{n} w_j^2) \tag{3.5}$$

D_1 and D_2 are represented with sentence embeddings obtained from a pre-trained language model. It is a common practice to find the k most similar documents to a given document (query). Given a corpus of documents $D_{1:k}$, let $\boldsymbol{D}$ be a matrix in which row i corresponds to the sentence embeddings of document D_i. The similarity between a query $Q = w'_{1:n}$ and each of the documents in $\boldsymbol{D}$ can be computed as:

$$\boldsymbol{s} = \boldsymbol{D} \cdot (\sum_{i=1}^{n} w'_{1:n}) \tag{3.6}$$

$\boldsymbol{s}$ is a vector of similarities, where $\boldsymbol{s}_{[i]}$ is the similarity of q to the i-th document, D_i in the corpus. Using s, the indices corresponding to the k highest values can be used to determine the k most similar documents.

In the absence of unlabelled data, which is commonplace for social media analysis, relevant texts can be retrieved by leveraging posts from other social media sites where posts can be implicitly annotated depending on which community they appeared. In Chapter 8, I retrieved relevant tweets from a large corpus of unlabelled tweets by using posts that appeared in related Reddit communities (subreddits) as query.

3.3.2 Transformer Architecture

The core architecture for many successful pre-trained language models is the Transformer [247]. The Transformer model was proposed to address problems with recur-

rent models regarding information loss when processing longer sentences. Moreover, recurrent models process inputs sequentially, which makes it slow and difficult to parallelise. With the Transformer model, input sequences are passed in parallel to speed up training.

Typically, a Transformer model is made up of stacked identical transformer layers. Each Transformer layer has two sub-layers: the multi-head self-attention sub-layer and the point-wise feed-forward network sub-layer. The key element in the Transformer architecture is self-attention. The concept of attention allows models to focus on the relevant parts of an input sentence as necessary. Self-attention is a variant of attention where each word in the sequence attends to every other word in the same sequence. Thus, the relationship between words in the sequence are captured.

In the self-attention sub-layer, the process of applying attention can be depicted as mapping a query and a set of key-value pairs to an output [247]. Given a sequence of input vectors X = $[x_1...x_n]$, the self-attention sub-layer projects the input vector x_i to a query, q_i and all other input vectors in X including itself as keys $(k_1, ..., k_n)$ to produce the attention weights. These weights show how relevant each input vector in X is with respect to x_i. The output is constructed as weighted some of the values $(v_i, ..., v_n)$. The self-attention process for an input sequence of N tokens is defined as:

$$SelfAttention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3.7}$$

where $Q \in \mathbb{R}^{N \times d_k}$, $K \in \mathbb{R}^{N \times d_k}$, and $V \in \mathbb{R}^{N \times d_k}$ are weight matrices containing the key, query and value vectors for all input vectors X respectively. The dimensions, d_k are the same for all weight matrices.

As an additional improvement, multi-head self-attention layers was introduced to capture the different kinds of relations such as syntactic and semantic relationships among inputs. Essentially, they are multiple instances of the self-attention layer,

with each instance, referred to as *head*, computing its own queries, keys and values. The outputs from each individual k-th self-attention head are then concatenated and linear transformation is applied to get the final output as:

$$MultiHeadAttn(X) = (head_1 \oplus head_2 ... \oplus head_K)W^O \tag{3.8}$$

$$head_i = SelfAttention(Q, K, V) \tag{3.9}$$

where $W_O \in \mathbb{R}^{d \times d}$ is a learnable parameter.

Each Transformer layer has a fully connected network in addition to the self-attention sub-layer, which is applied to each position separately and uniformly. The feed-forward network sub-layer (FFN) consists of two linear transformations with ReLU activation function between them. Given a sequence of vectors $h_1, ..., h_n$,

$$FFN(h_i) = ReLU(h_i W^1 + b^1)W^2 + b^2 \tag{3.10}$$

where W^1, W^2, b^1 and b^2 are parameters.

Residual connections are used to pass information from the self-attention sub-layer to the feed-forward sub-layer. This is performed by adding the input vector of a sub-layer to its output vector before passing it to the next sub-layer. Layer normalisation [19] is also employed as a form of normalisation to improve the training performance of the Transformer model by keeping the values of a hidden layer within a range that facilitates gradient-based training.

Due to the parallel nature of self-attention layers, Transformer models do not have any notion of the positions of the input tokens. To solve this, Transformers use *positional embeddings* to equip the input tokens with their positional information. The positional embeddings are added to the corresponding embedding of each input token, thus enabling the model to capture positional dependencies among input tokens. The positional embeddings are learned together with other parameters during training.

Transformer-based models have been used to achieve state-of-the-art performance across various NLP benchmarks [147], hence all of the models I use for language understanding tasks in this thesis are based on the this architecture.

BERT: Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) [61] is one of the prominent Transformer-based language models. In contrast to unidirectional language models, BERT reads the entire input sequence at the same time and considers both the left and right contexts. Consequently, the self-attention mechanism is applied over the entire input and the resulting vectors capture contextual information from both before and after the target tokens. The focus on generating contextualised representations make them useful for sequence classification tasks I performed in Chapters 5 and 6.

BERT was trained on a large corpus of unlabelled text including the Book Corpus (about 800 million words) and English Wikipedia (over 2.5 billion words). To train bidirectional encoders, BERT used the the *masked language modelling* (3.3.1) as a pre-training objective. The masked language model randomly replaces some tokens with the *[MASK]* token and the objective is to predict the masked tokens while considering tokens from surrounding contexts. In addition to the masked language modelling, BERT also used *next sentence prediction* task. In this task, the model receives pairs of sentences and learns to predict if the second sentence follows the previous or not in the training corpus.

To train BERT, original sentences from the corpora are first tokenized with the WordPiece [264], a subword model and mapped to a sequence of embeddings. The word embeddings for all the tokens are then combined with segment and positional embeddings to pass as input into the transformer. A special classification token (*[CLS]*) is added to the start of every input sequence pair. Another special token *[SEP]* is also added in between the sentences and at the end of the second sentence.

As the state-of-the-art approach for many social media analysis research reviewed in Chapter 2, I use BERT as the baseline model in the experiments performed in Chapters 5 and 6.

RoBERTa: A Robustly Optimized BERT Pretraining Approach

RoBERTa [147] is an optimised variant of the original BERT model that employs an enhanced pre-training strategy. The primary objective of RoBERTa is to surpass the performance of the original BERT model, by altering the pre-training process. This involves expanding the training data size, increasing the batch size, and modifying the approach towards next sentence prediction.

RoBERTa incorporates a wider scope of training data, as compared to BERT. Besides using the English Wikipedia as its data source, just like BERT, RoBERTa also utilises text from Common Crawl, thus enriching its pre-training corpus with more diversified content.

Another significant modification is the removal of the next sentence prediction task from the training objectives. The authors of RoBERTa discovered that the elimination of this task led to improvements in performance across various downstream tasks.

Furthermore, RoBERTa employs a dynamic masking process, deviating from the static masking strategy utilised in BERT. This dynamic approach involves using a different mask for each time a sequence is fed into the model during training. This alteration encourages the model to develop a more robust understanding of the text, further enhancing the performance of RoBERTa over the original BERT model. To further improve results on HMC tasks, I utilised RoBERTa in Chapter 7.

ALBERT: A Lite BERT

The ALBERT model [132] represents an effective solution to challenges such as long training times and GPU memory constraints associated with sizeable models like the original BERT. By significantly reducing the number of parameters, ALBERT facilitates a more manageable training of larger models.

To reduce the number of parameters, the creators of ALBERT employed two innovative strategies: factorised embedding parameterisation and cross-layer parameter sharing. Both strategies were instrumental in addressing the issue of scaling up large pre-trained language models.

Moreover, the ALBERT model swaps out the next sentence prediction objective with a more refined sentence-order prediction. This shift enhances the model's ability to capture the coherence between sentences. As a result, despite its fewer parameters, ALBERT maintains competitive performance on various tasks.

The efficiency of the ALBERT model is particularly useful in settings where computational resources are limited. This is evident in my work with underrepresented communities where data might be sparse, as shown in Chapter 7.

GPT-2: Generative Pre-trained Transformer 2

Generative Pre-trained Transformer 2 (GPT-2)[205], like BERT is a large pre-trained language based on the Transformer architecture. GPT-2 is trained with causal language model 3.3.1 objective hence the focus is on predicting the next token in a sequence. The self-attention mechanism is also based on autoregressive self-attention which means the model only attends to previous tokens up the current one. This makes them applicable for generative tasks like text generation. As such, I follow the approach for training GPT-2 in Chapter 8 to measure faith-related engagements through language use.

GPT-2 was trained on approximately 40GB of text data (over 8 million documents)

crawled from the web. The input sequences passed in to GPT-2 are encoded using Byte Pair Encodings [222]. The model also uses a special token, </*endoftext*/> to mark the end of a text. The input representation passed into the model is the sum of the corresponding token and position embeddings. Finally, each token in the input is processed and passed through all the transformer layers of the model successively.

3.3.3 Other Large Language Models

The field of NLP has continued to evolve rapidly, especially in the area of transformer-based language models. These more recent models build upon their predecessors' strengths, address their limitations, and push the boundaries of what these language models can achieve. Some of the newer models include Longformer[29] which addresses the limitation of processing longer sequences. There are other models such as T5 [206] which uses a slightly different technique in its MLM objective. It can mask and predict multiple words in a sequence which gives the model flexibility.

In the area of autoregressive language models, notable advancements like GPT-3 [38] have emerged. GPT-3 is a massive language model that comprises an astounding 175 billion parameters. The model, which operates on the principle of next-word prediction, was trained using an impressive 45 Terabytes (TB) of text data. While access to GPT-3 has been restricted, efforts have been made to develop open-source large language models like BLOOM [217] which has powerful text generation capabilities. Notably, BLOOM is a multilingual language model with 176 billion parameters. The contributors behind BLOOM have placed a strong emphasis on maintaining transparency throughout the model's training process.

While this thesis concentrates on specific methods of NLP for analyzing social media data for societal benefit, the field is continuously evolving with the development of large language models such as Longformer, T5, GPT-3, and BLOOM. These new models represent both the rapid advancement in NLP and the potential for more

sophisticated analyses in the future. Even though these models were not utilised in this study, they indicate the ongoing growth and relevance of the methodologies explored in this thesis and suggest avenues for further research in social media analysis.

3.4 Transfer learning

Recently, transfer learning in NLP has significantly improved performance on various tasks [61, 266, 147]. In this thesis, I follow the pre-train and finetune paradigm which is a case of transfer learning. The process of language model pre-training can be viewed as developing a universal knowledge that allows model to understand text. This knowledge can then be transferred to a target task. When compared to computer vision, annotated data are limited for NLP tasks. Deep learning models do not generalize well and usually overfit on low resource tasks (smaller datasets). Therefore, language models which are often pre-trained on unlabelled data is suitable for the transfer learning setup as the representations learnt can be useful for a task with limited data.

Given a pre-trained language model and an input sentence, the model's output can be considered as containing contextual embeddings for each token in the input. These contextual embeddings can serve as a general representation of the meaning of the input sentence for text classification tasks. To make use of these general representations for downstream tasks, I apply fine-tuning. Fine-tuning facilitates the transfer of knowledge from pre-trained models to task-specific models. In the following subsections, I discuss the fine-tuning methods employed for adapting pre-trained language models to the text classification tasks performed in this thesis.

Fine-Tuning

Fine-tuning involves taking the representations from the pre-trained models and further training on a supervised task. Formally, given a pre-trained model M with

weights W for a new target task T, M is used to learn a new function $f(\cdot)$ that maps the parameters as:

$$f(W) = W' \tag{3.11}$$

In text classification tasks, an input sequence is often represented with a single unified representation. For example, when using BERT, the *[CLS]* token serves as the overall representation for the entire input sequence. This representation which is sometimes referred to as the sentence embedding is fed into a neural network. Then, a softmax is applied to the output of the network to get the probability distributions over possible classes:

$$\hat{y} = softmax(W' y_{CLS}) \tag{3.12}$$

where W' are the network parameters of the downstream model and y_{CLS} is the representation of the input sequence.

The structure of the neural network can either be a simple architecture such as a linear layer or a more complex architecture such as a bidirectional Long short-term memory (LSTM) [86]. Generally, using a simple linear layer is sufficient to achieve strong performances on the downstream tasks [61]. I experiment with both simple and complex neural networks. In terms of adapting the parameters of the pre-trained model, the parameters can either be left unchanged (freeze) or updated during training (unfreeze).

Intermediate task

Most work apply transfer learning by first pre-training and then fine-tuning. However, it has been recently shown that performing an intermediate task in between can produce better performance in some cases [194, 90]. The intuition behind this is that the relatedness of data-rich tasks can be helpful to the target domain or task. In chapters 5 and 6, I experiment with different approaches to incorporate relevant knowledge to improve performance on the target task.

The first approach is based on an intermediate fine-tuning using large scale in-domain data. I used a fine-tuning method that enhances the masked language model for a specific task. In addition to the random token masking, I propose masking words that are important to a specific task (e.g. emotion words in an emotion detection task) and then model an attention mechanism that focuses on these words.

The second approach I consider is to supplement the pre-trained model with further training on a related data-rich supervised task. I leverage the availability of tasks with abundant labelled data to infuse knowledge into the target task. The relatedness of the intermediate and target tasks can either be in terms of the domain of the data (e.g. Twitter domain) or a shared linguistic phenomenon (e.g. expressing emotions in self-reports of personal health experiences).

Multi-Task learning

In the field of machine learning, it is common practice to train models on a single task. However, this approach may lead to a loss of valuable information from related tasks that could improve performance on the target task [211]. Multi-task learning (Multi-task learning (MTL)) addresses this limitation by allowing a model to learn from related tasks through shared knowledge gained from the training signals of those tasks. This transfer learning paradigm has been shown to improve generalisation and performance on a primary task [270], making it particularly useful in low-resource settings where annotated data is limited [151].

The implementation of MTL can occur in different settings: a learning setting with predetermined set of tasks is simultaneously learned with equal emphasis given to each task [56], or as learning with auxiliary tasks, where supplementary tasks are leveraged, to augment the main learning goal [210]. In essence, any instance where the optimisation process involves more than a single loss function, the process transitions from a single task learning to a multi-task learning [210]. For a detailed

introduction to MTL, readers are referred to [211].

In Chapter 7, I implemented MTL to enhance the performance of HMC tasks by jointly learning two objectives. Specifically, I used one objective for the primary task of HMC and the auxiliary task for detecting the literal usage of disease keywords, both based on the input text. For further details of this architecture, please refer to Section 7.3.5.

3.5 Latent Dirichlet Allocation

Arguably, the most well-known approach for topic modelling is the Latent Dirichlet Allocation (LDA). The LDA [34] approach is particularly effective in extracting valuable insights from datasets. As a robust topic model, LDA has been extensively applied to a wide range of text documents, including social media data, to identify concealed topic structures that provide deeper insights into the data [40, 169]. In the LDA model, a set of documents (D) is assumed to contain a number of topics (K), each represented by a set of words (w). Each document $d \in D$ is modelled as two multinomial distributions: $p(t|d)$, which represents the probability distribution of words in document d assigned to topic t, and $p(w|t)$, which denotes the probability distribution of assignments to topic t over all documents D originating from a word w. The model assigns a Dirichlet prior (α to the multinomial distribution (θd) over K topics, represented as $Dir(\theta d|\alpha)$). Similarly, for topic k, a Dirichlet prior (η) is assigned to the multinomial distributions (βk) over words, represented as $Dir(\beta k|\eta)$ [178]. In Chapter 4, I employed LDA to automatically uncover the themes prevalent in the discourse across social media texts.

3.6 Ethical consideration

As demonstrated in this thesis, social media analysis has many positive and beneficial applications that can lead to positive societal impact. However, there are

ethical concerns attached to using such data [236]. A key issue is relating to the privacy of the users whose social media posts are analysed. It is of utmost importance to protect personal information by taking necessary precautions and measures to ensure data security. To address the ethical concerns in this thesis, I took the following steps:

- The work carried out in this thesis was subjected to ethical review by the Durham University Ethics Review team, and necessary approval was sought to conduct the research in compliance with ethical guidelines.
- In light of the potentially sensitive nature of social media data, which may include personal information regarding individuals' health status, religion, and education, extra precautions were taken during all phases of data collection, processing, and analysis. These measures were informed by Benton et al.'s [32]. In the majority of cases, only the textual content of social media posts were collected and used for analysis. In instances where personal details were included in the text, personal information identifiers - such as names, screen names, and mentions of other users - were anonymised. In Chapter 5, where geo-location information of tweets was utilised, the location filter of the Twitter API was employed to specify a bounding box for the collection of tweets. Notably, no information about the actual coordinates of the tweets - such as longitude and latitude - was processed.
- The data used in this thesis were obtained from social media and were generally publicly available. The datasets developed in this thesis, however, were not made available publicly in their raw form and were released in accordance with the data release guidelines for each platform. In the case of Microsoft Teams, which is a private platform, ethical consent was obtained from all TechUPWomen participants for the use of the data for evaluation purposes, and steps were taken to ensure the privacy and confidentiality of the data. As a result, the Microsoft Teams dataset was not made publicly available.

- All disseminated results and statistics are not linked to individual users. In addition, I aggregated the data at a level that prevents re-identification of individual users.

Chapter 4

Digital Inclusion in Northern England: Training Women from Underrepresented Communities in Tech: A Data Analytics Case Study

This chapter aims to provide insights into the effectiveness of the technology re-training programme and the potential of social media data for evaluating such programmes. Through the analysis of social media content, this study seeks to provide evidence-based recommendations for improving the design and delivery of technology training programmes for underrepresented communities. Part of the work in this Chapter have been published in [8].

4.1 Introduction

There is a growing global, as well as UK-based, attention on narrowing the gender gap and improving participation of women and members of underrepresented groups

in computer science [216]. Yet, the gender inequality in technology-related fields still very much exist [220]. According to the Higher Education Statistics Agency (HESA) , only about 18% of the students in higher education studying computer science were women in 2017/2018, with under 1% increase from the previous year. In contrast, the computing and information technology industry has been growing exponentially, showing an urgent national need for people in technology-related fields [226]. Due to persistent hiring challenges in the technology industry, bootcamps have sprung up to provide a fast-track entry into technology roles, while costing less in terms of tuition and time. Additionally, there is an increase in the understanding that diversity is a strength in any community, in general, and in technology, in particular [258]. Thus, underrepresented groups are particularly interesting for the technology industry. Whilst 'coding bootcamps' have appeared [251], training participants in technology roles, their effectiveness is not always clear or measurable. Especially difficult is the measuring of new features introduced during the programme, to inform further potential changes. Moreover, there has been very limited data analytics performed on the retraining programmes for women in tech(nology) during the running of the programme.

In this Chapter, I present a case study on a technology retraining program for women from underrepresented communities in Northern England. The overall aim of the program aligns with the United Nations SDG goal on education - *ensure inclusive and equitable quality education and promote lifelong learning opportunities for all*[*]. The program's impact is analysed by employing data analytics techniques on social media content from both public (Twitter) and private (Microsoft Teams) channels. My contribution in this Chapter includes measuring the impact of a technology retraining programme for women from underrepresented groups by using social media data. Through extensive analysis, I have demonstrated the benefit of social media analysis in evaluating and informing social interventions.

[*]https://sdgs.un.org/goals/goal4

4.2 TechUPWomen

TechUPWomen* is a programme funded by the Institute of Coding† that targets women from minority groups based in the Midlands and North of England and puts them through free online training sessions for gaining technology skills, as well as offering four residential weekends for motivation, catching up and networking. Similar to coding bootcamps, the programme is developed in close collaboration with industrial partners; participants are assigned one-to-one mentors and each participant is guaranteed an interview with one of the industry partners after the programme. To accommodate the diversity of needs of the women, who traditionally have multiple roles and responsibilities, the intervention programme is mostly delivered online, while providing opportunity for face-to-face meetings during the course of the programme. The programme maintains a very active social media presence for participants to be engaged in collective learning processes, as well as to be exposed to relevant industry networks. Some of the social media tools used for the TechUP programme include LinkedIn, Twitter, Microsoft Teams discussions, blog and Instagram.

To understand how special intervention methods would support women transition into technology roles, in the TechUPWomen programme, this study uses social media analysis methods, to investigate the participants' temporal activities on social media, to measure the impact of the programme to increase women participation in computer science as well as to support knowledge transfer into computer science roles. The methodology used in this study involves the calculation of important descriptive analytics. This begins with the statistical analysis of TechUPWomen-related social media posts and their evolution over time. In addition, NLP techniques, such as topic modelling and sentiment analysis, were employed to gain insights into the social media exchanges and online learning platform discussions

*https://techupwomen.org/
†https://instituteofcoding.org

utilised throughout the program. These methods provide a valuable toolset for understanding the impact of the program on its participants.

4.3 Materials and Methods

4.3.1 Data Collection

This study analysed data collected during the TechUPWomen programme using multiple sources. Firstly, a Twitter dataset was compiled using the Twitter streaming API and the official programme hashtags (*#TechUpWomen* and *#TUW2019*) from July 2019 to January 2020. Additionally, participant engagement on Microsoft Teams discussions were collected during the same period. Microsoft Teams was used as a collaborative platform for sharing learning materials, assignments, and communication between programme participants and support staff. The platform had over 200 users, including learners, coordinators, and mentors, with 100 being active users. However, data had to be manually extracted and exported from the general Microsoft Teams conversation as the platform does not have automatic extraction or export options.

4.3.1.1 Pre-processing

Data pre-processing techniques were applied to the data directly collected from both Twitter and Microsoft Teams. In particular, Twitter language differs from text in books and articles, and because of the text limit, texts are often shortened, and they also include distinctive uses, such as URLS, repeated letters, @ for usernames, # for hashtags and emoticons. Thus, it is important to pre-process and normalise these texts (see Section 3.2.4). To preserve privacy of users on Teams and Twitter, I replaced usernames with a *<user>* token. I further applied simple pre-processing techniques, such as stemming, to remove tenses and plurals from the endings of

words (e.g., inspired, inspiring => inspire), and stop-word filtering to remove words that were frequent but did not contain useful information (e.g., and, the).

4.3.2 Topic Modelling

To uncover the underlying topics in the studied datasets, I employed topic modelling using LDA [34]. LDA method was chosen for topic modelling because of its widespread use in the literature and its ability to handle large and diverse datasets [13]. The main idea behind LDA is to assume that the mixture of words in a document originates from a set of latent topics, which in turn come from a fixed probability distribution over the vocabulary. By estimating this distribution and the topic mixtures for each document, insights can be gained into the main themes and patterns of discussion in the datasets.

To implement LDA, I used the Scikit-learn library, a widely-used Python package for machine learning and data analysis [190]. I used the default hyperparameter settings for the LDA algorithm in Scikit-learn, which are known to perform well in practice. Specifically, I set the number of topics as a hyperparameter, which determines the granularity of the topic models and thus requires careful consideration [152]. To choose this parameter, I employed qualitative judgement based on the coherence and interpretability of the resulting topics [2].

4.3.3 Sentiment Analysis

Transformer-based models have become the state-of-the-art methods for automated sentiment polarity detection in texts [171]. To determine the sentiment polarity of texts from both Twitter and Microsoft Teams, I used a publicly available RoBERTa-based model called SiEBERT [93]. SiEBERT was fine-tuned and evaluated for sentiment analysis on 15 diverse text datasets to enhance the generalisation of sentiment annotations across different types of texts, making it particularly suited for the studied datasets (Twitter and Microsoft Teams) in this Chapter.

The fine-tuning step for SiEBERT follows the process outlined in section 3.4. The creators used a learning rate of $2e-5$, and set the number of training epochs to 3, warmup steps to 500, and weight decay to 0.01 [93]. SiEBERT outperformed the vanilla BERT model on the benchmark sentiment analysis dataset, Stanford Sentiment Treebank 2 (SST-2). SiEBERT was used to automatically annotate the sentiment of all texts from both Twitter and Microsoft Teams.

4.4 Results and Discussion

4.4.1 Twitter and Microsoft Teams Activity Analysis

Table 7.3 shows the number of engagements (including tweets, retweets and comments) from Twitter and Microsoft Teams between July 2019 (programme start) and January 2020 (programme end). The Twitter dataset has a significantly larger number of posts and users compared to Microsoft Teams. The average length of posts on Twitter is also shorter, with an average of 28 tokens per post. On the other hand, Microsoft Teams has a smaller number of posts and users, but the average length of posts is longer at 40 tokens per post. One possible explanation for the difference in the number of posts between these two platforms is the overall popularity of each platform. Twitter is known as a widely-used social media platform, while Microsoft Teams was primarily used for communication and collaboration within the participants and mentors. This accounted for the higher number of posts and users on Twitter compared to Microsoft Teams.

Additionally, the difference in the average length of posts between these two platforms could be due to the intended use of each platform. Twitter is known for its short, concise posts, while Microsoft Teams may be used for more detailed communication and discussion, leading to longer posts. Overall, these statistics suggest that Twitter is a more widely-used platform with shorter posts, while Microsoft Teams is used by a smaller number of users for more detailed communications.

Dataset	Statistics
Twitter	
No of posts	6,990
No of users	947
Avg. len of posts	28 tokens
Microsoft Teams	
No. of posts	1,461
No. of users	181
Avg. len of posts	40 tokens

Table 4.1: Dataset statistics

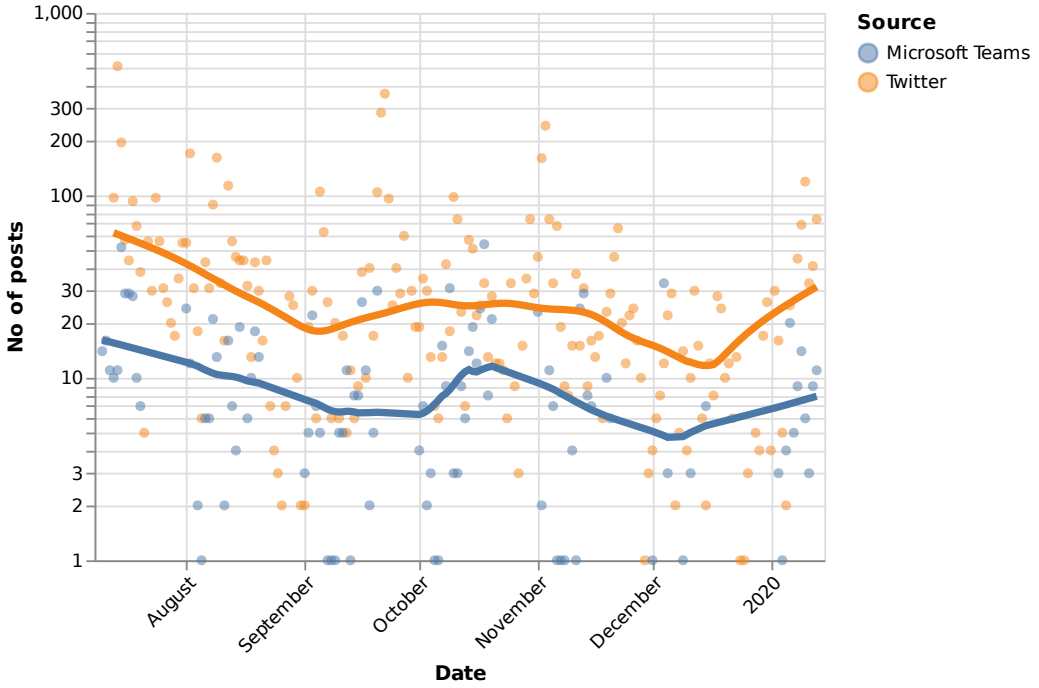


Figure 4.1: Comparison of daily posts on Twitter and Microsoft Teams

Figure 4.1 illustrates the engagement trends observed in both Twitter and Microsoft Teams datasets during the course of the programme. Notably, the engagement patterns observed in both sources were similar throughout the study period. Specifically, the number of posts on both platforms showed an increase before and during residential weekends (i.e. 14 July 2019, 22 September 2019, 2 November 2019 and 20 January 2021), indicating higher levels of engagement. However, while the highest number of posts were recorded only during the residential sessions for Twitter, Microsoft Teams showed a higher number of comments on other days as

No of Topics	Coherence score - Twitter	Coherence score - Microsoft Teams
2	0.286	0.258
4	0.308	0.3
5	0.313	0.305
8	0.320	0.311
10	0.324	0.320

Table 4.2: Coherence Scores for Different Numbers of Topics in LDA Topic Modelling

well, such as when assignments were due (e.g., 20 September 2019 and 18 October 2019). The graph further demonstrates that discussions were sustained on both platforms over the course of the programme.

4.4.2 Topic Analysis

In the topic modelling experiment, I examined the output of LDA with different numbers of topics (2, 4, 5, 8, and 10) and assessed the resulting topics' semantic relevance to the studied datasets. I computed the coherence scores for these various numbers of topics, with the results presented in Table 4.2. The maximum coherence scores were observed with 10 topics, registering values of 0.344 for tweets and 0.32 for Microsoft Teams chats. It should be noted that the coherence scores were relatively low across all tested numbers of topics.

While measuring topic coherence is a common approach to selecting the best number of topics, prior research [168] has shown that qualitative evaluation of topic models can also be useful. This dual approach provided a more robust justification for the selected numbers of topics, especially given the observed overlap in the Twitter topics. After careful analysis, I determined that using 10 topics for the Twitter dataset and 5 topics for the Microsoft Teams dataset produced the most relevant themes.

The results of the LDA analysis are presented in Tables 4.3 and 4.4 for the Twitter and Microsoft Teams datasets, respectively. The first column in each table shows the assigned topic labels and their frequency counts, and the second column shows the top 20 most representative words for each topic. The tables reveal a variety of

Topic label	Top Weighted Words
Women in technology (12.64%)	*trend, #womenintech, #tuwres, #womeninstem, wonderful, #womenwhocode, talk, great, you, inspire, hear, happy, twitter, minute, want, speaker, watch*
Feminism (7.96%)	*girl, do, well, amazing, #womenintech, womenwhocode, whoop, ctrlyourfuture, real, instituteofcode, omg, #tuw, #digitalskillsmatter, ioc, kat, hey, module*
Appreciation (12.37%)	*thank, much, make, love, last, time, get, always, hard, hope, hey, enjoy, think, woohoo, work, one, night*
Mentorship (11.32%)	*woman, change, check, life, welcome, midland, amazing, wait, awesome, take, residential, look, forward, see, meet, mentor, delight*
Learning (8.51%)	*code, #tuwres, #tuw, new, let, ltd, round, support, like, video, meeting, assignment, feel, need, challenge, hour*
Residential Weekend (9.25%)	*today, #tuwres, great, python, day, photo, get, ella, brilliant, residential, code, final, #tuw, thank, course, time, fab*
Congratulatory remarks (13.38%)	*congratulation, graduate, nottingham, first, weekend, congratulations, #womenintech, womenwhocode, future, programme, super, tech, #tuw, proud, huge, world, part, good*
Graduation (10.65%)	*thank, sci, durham, uni, please, #tuwres, team, graduation, amazing, vote, apply, website, yet, board, blog, celebrate, see, available*
Diversity and inclusion (7.62%)	*tech, #womenintech, woman, amazing, late, #tuwres, check, #tuw, #womeninstem, read, post, incredible, retrain, diverse, yay, weekend, background, north*
Technology award (6.32%)	*techup, #tuwres, beautiful, via, vote, uon, impactawards, inspiration, open, thank, share, finalist, #tuw, day, peopleschoice, good, residential, techforgood*

Table 4.3: Topics with representative words for Tweets.

themes related to the TechUPWomen programme across both data sources. Notably, both platforms contain discussions about *Learning* and *Residential weekend*.

In the Twitter data (Table 4.3), additional topics related to the programme's goals, such as *Women in technology* and *Diversity and inclusion*, were also identified. These topics suggest that discussions on Twitter were focused on the overall aim of the TechUPWomen programme. Additionally, discussions about accomplishments during the programme, such as *Appreciation*, *Graduation*, and *Technology awards*, were found. The overlap in Twitter topics reflects the nature of the discussions on the platform, where different topics can often intersect and interact.

Topic label	Top Weighted Words
Coding (18.34%)	*guest, file, python, know, try, work, use, hand, also, go, think, get, code, activity, exercise, run, import, topic, probably, agree*
Learning (18.75%)	*video, get, assignment, slide, need, really, thank, else, please, share, time, link, right, would, know, ve, want, i*
Residential weekend (14.99%)	*great, thank, everyone, see, smile, weekend, hello, forward, look, guy, like, meeting, saturday, meet, hope, well, today, tomorrow, year, hey*
Learning support (20.26%)	*add, term, techup, yes, laugh, twitter, list, korpet, please, pathway, link, assignment, smile, new, account, tech, email, one, need*
Peer support (27.65%)	*thank, good, work, much, course, do, thanks, think, really, use, get, nt, one, agile, useful, way, python, help*

Table 4.4: Topics with representative words for Microsoft Teams chat.

On Microsoft Teams (Table 4.4), the discussions were more focused on learning experiences, including *Peer support*, *Learning support* and *Coding*. These topics suggest that the conversations on Microsoft Teams were more centered on the learning and educational aspects of the TechUPWomen programme.

The LDA analysis provides insight into the main topics discussed during the TechUPWomen programme on both Twitter and Microsoft Teams. This understanding helps to grasp the focus and goals of the programme and the types of conversations among participants.

4.4.3 Sentiment Analysis results

The results of the temporal sentiment analysis are presented as the frequency of polarised tweets on a daily basis. Figure 4.2 reveals interesting trends regarding the sentiment polarity of the tweets throughout the TechUPWomen programme on Twitter. Overall, the analysis suggests that the majority of the tweets on Twitter regarding the programme are positive. Furthermore, the trend of positive tweets appears to be increasing towards the end of the programme. This is a potentially positive indication of increasing positivity towards the programme as

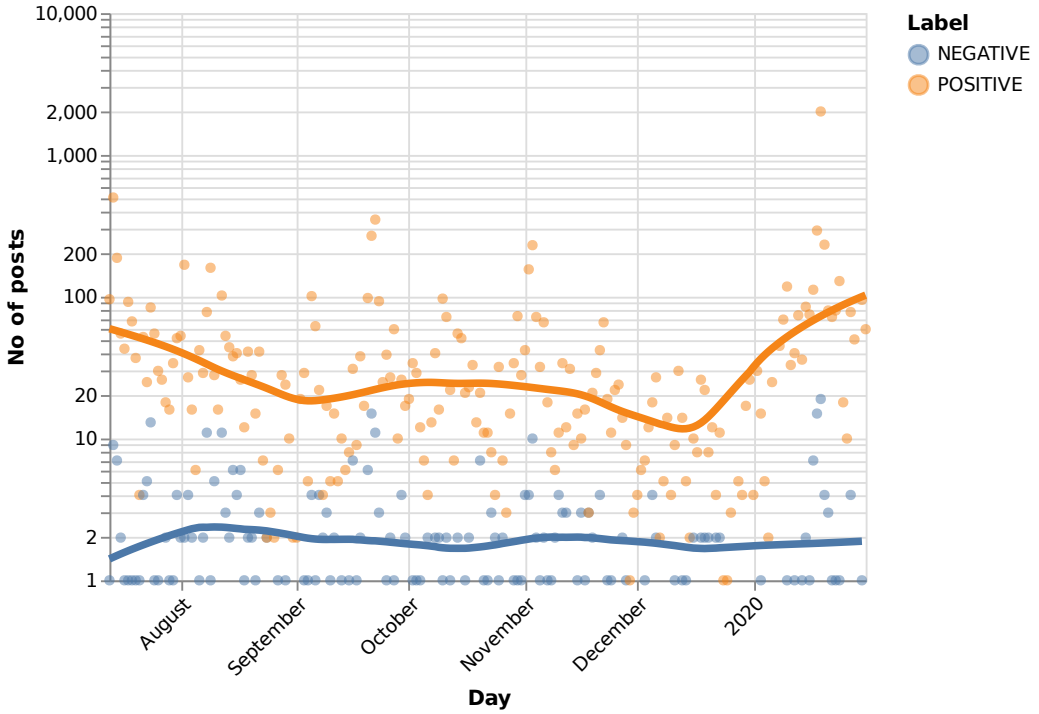


Figure 4.2: The daily frequency of polarised posts on Twitter from July 2019 to January 2020

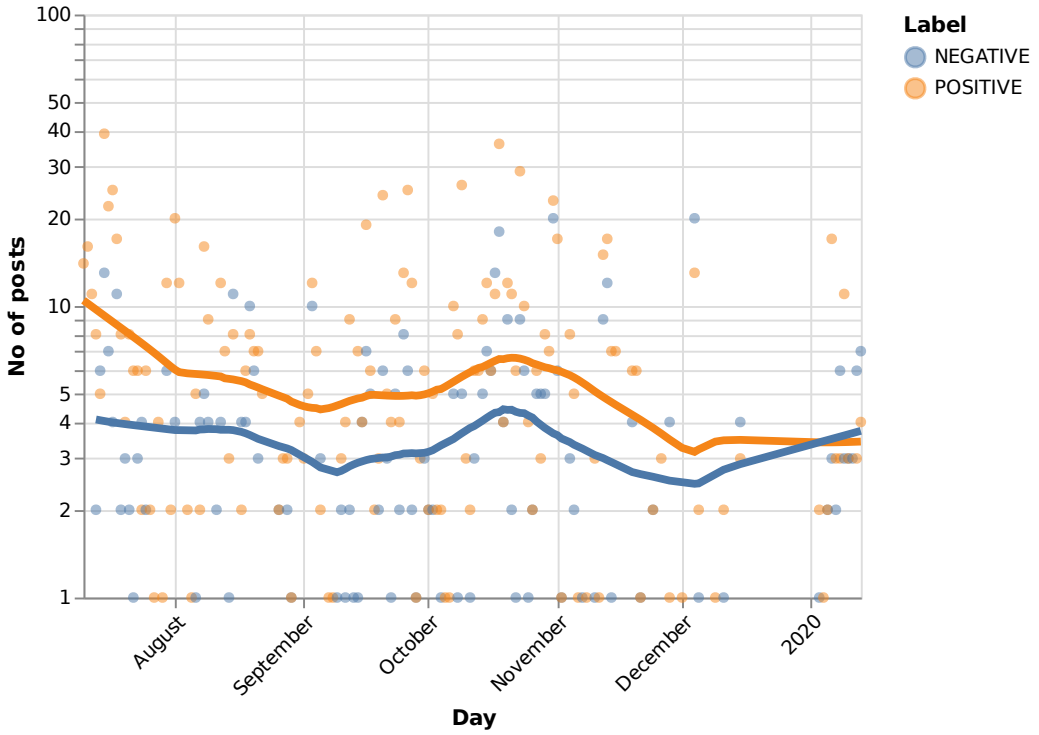


Figure 4.3: The daily frequency of polarised posts on Microsoft Teams from July 2019 to January 2020

the programme progresses. Interestingly, the trend of negative tweets is relatively stable throughout the duration of the programme, with only minor fluctuations.

Moving on to the Microsoft Teams dataset (Figure 4.3), which is a private channel, there is a discernible difference in the trend of sentiments expressed on the platform. The sentiment analysis indicates that more positive sentiment is equally expressed on both the public channel (Twitter) and the private channel (Microsoft Teams), whereas negative sentiment is more frequently expressed through the private channel. It is noteworthy that negative sentiment is more frequently expressed on Microsoft Teams when compared to Twitter. This higher negative sentiment expression could be attributed to the fact that people tend to express their emotions, particularly negative emotions, more freely when they are closer to their audience. Nonetheless, the number of positive posts still outweighs the number of negative posts on Microsoft Teams.

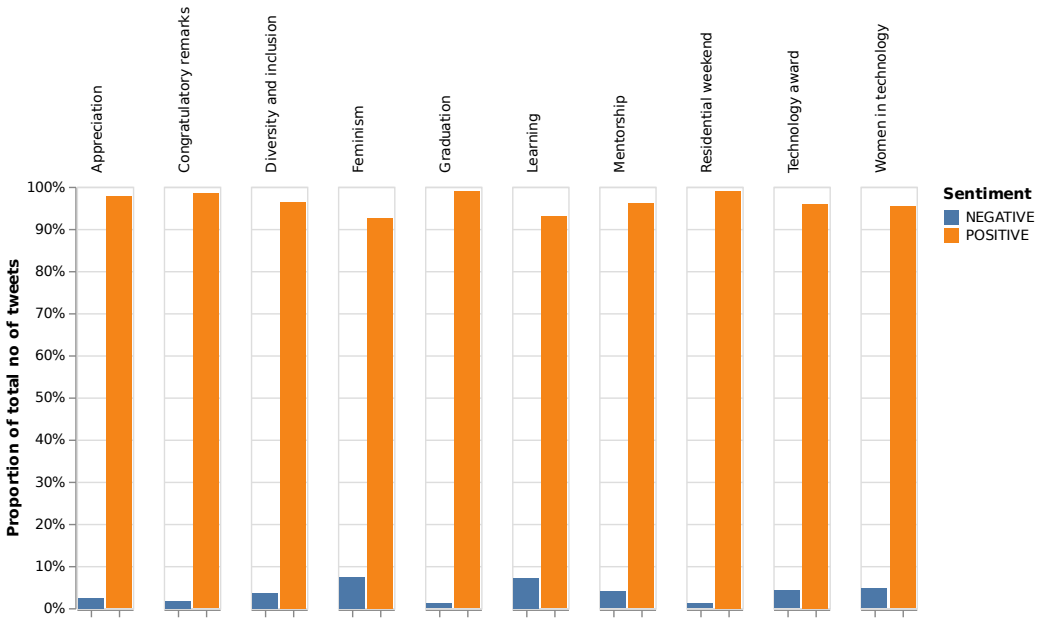


Figure 4.4: Sentiment distribution of identified topics on Twitter

Furthermore, I conducted an analysis of the sentiment distribution over the topics identified (section 4.4.2 on both Twitter and Microsoft. This is particularly valuable in determining which aspects of the programme were successful and which require improvement. As depicted in Figure 4.4, the majority of the posts on Twitter and

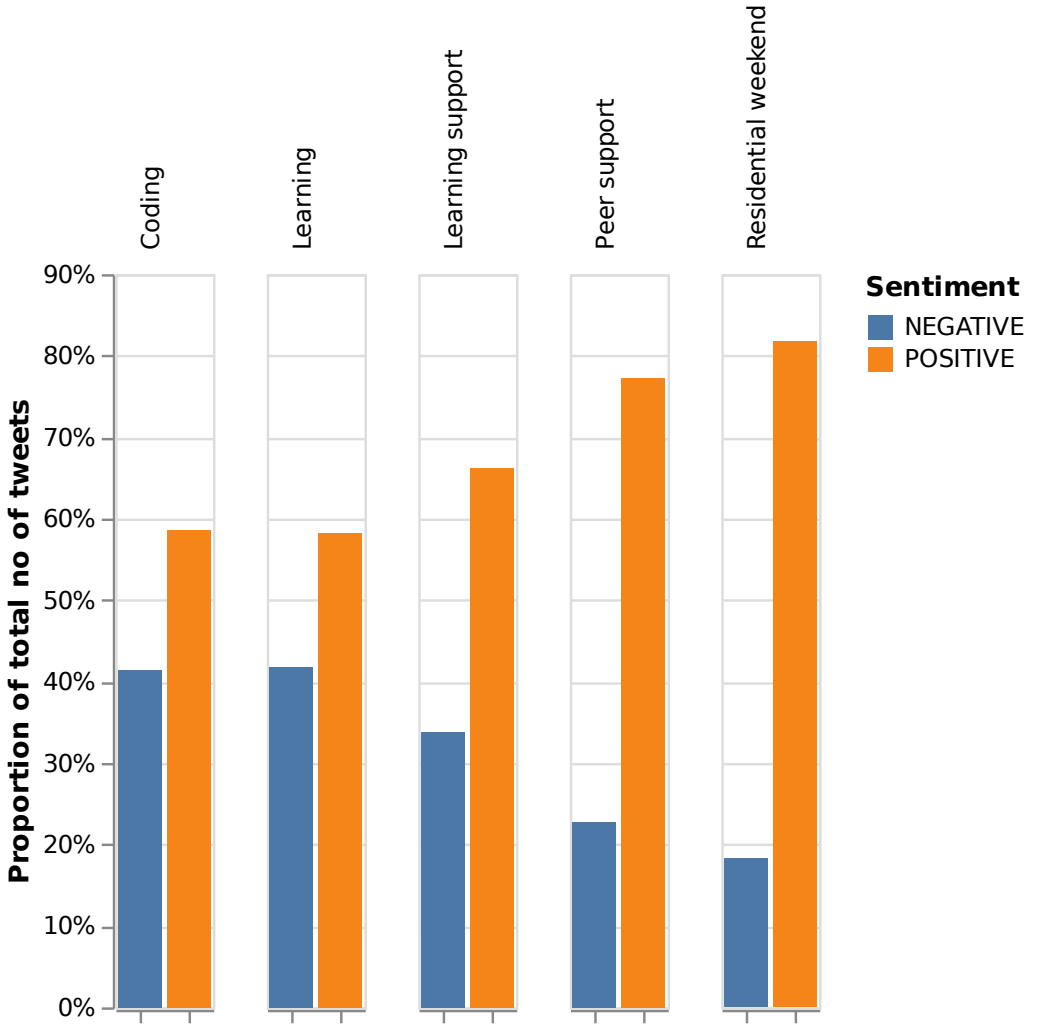


Figure 4.5: Sentiment distribution of identified topics on Microsoft Teams

Microsoft are positive, with the highest proportion of discussions related to the *residential weekend*.

The sentiment distribution also confirms the distinct differences in the levels of sentiment between Twitter and Microsoft Teams, with more negative posts on the latter. Specifically, discussions related to *coding* and *learning* on Microsoft Teams generated the most number of posts categorised as negative.

4.4.4 Error Analysis

I performed a comprehensive error analysis on the sentiment predictions generated by the SiEBERT model. By drawing a diverse sample of misclassifications from both positive and negative sentiments, I have grouped the errors into the following categories:

- Emotional Nuance Misunderstanding: The model sometimes fails to interpret the mixed or nuanced emotional content in the posts correctly. In cases like this, the model might categorise sentiment as purely negative or positive when it is actually a mix. For example: *@<user> I can't believe this is the last time ill be travelling to an official #techupwomen weekend... I'm really going to miss..* was predicted as a negative sentiment.

- Sarcasm and Irony Misinterpretation: These are instances where the model fails to recognise and correctly interpret sarcasm or irony, leading to a misinterpretation of the sentiment. For instance, in this post, it led to it being misclassified as negative: *Who knew that my knowledge of Madagascan animals would serve me well on the Introduction to Machine Learning module @TechUpWomen... more niche mammals please! .....*

- Contextual Misunderstandings: The model sometimes relies on surface-level lexical features and it fails to understand the broader context of a statement. Often, sentiment is determined not just by the words used, but also the wider context, which the model may not fully grasp. For example: *Does data structures ever end?? #techupwomen really want to get started on the advanced programming stuff* was predicted as a positive post and *We are absolutely not crying over here. There .... just something in our eyes #Tuwres4 #techupwomen* was predicted as a negative post.

4.5 Discussion

This chapter has explored the impact of a retraining programme, TechUPWomen, for women from underrepresented groups, as reflected on social media. By utilising both topic modelling and sentiment analysis, the engagement patterns and sentiment polarity of the participants were studied during the programme, in order to answer **RQ1**. The findings indicate that the TechUPWomen programme generated a significant impact, with a large amount of engagement on both Twitter and Microsoft Teams. While Twitter had a higher number of posts and users, the average post length was shorter compared to Microsoft Teams, which could be due to the different purposes of the two platforms.

The engagement on both platforms followed a similar pattern, with peak activity occurring during residential weekends and when assignments were due. Through LDA analysis, diverse themes related to the programme were identified, including discussions about learning, residential weekends, and programme goals on Twitter, and learning experiences on Microsoft Teams. These themes provided insights into the focus and goals of the programme and the types of conversations that took place among participants.

Regarding the sentiment analysis, both public (Twitter) and private (Microsoft Teams) channels showed similar trends, with more positive sentiment expressed during the programme. Specifically, discussions around the programme's *residential weekend* were the most positive. The analysis of the programme using social media data has provided useful insights that can be leveraged to design *inclusive and equitable* programmes that *promote lifelong learning opportunities for all.*

Chapter 5

Detecting Fine-Grained Emotions on Social Media during Major Disease Outbreaks

By examining the emotions expressed on social media during major disease outbreaks, valuable insights can be provided to policymakers, to assist in delivering support that positively impacts society. Specifically, I analyse the emotions of a significant population on Twitter during the global pandemic, COVID-19. The findings of this Chapter have been published in [9].

5.1 Introduction

In this chapter, I shift from previous applications of social media analysis in education to health, effectively addressing **RQ2**. *How can social media be used to detect fine-grained emotions during major disease outbreaks*? Here, I focus on emotion analysis, to gain a deeper understanding of the emotional states of a population, during major public health crises. Building on the main thesis's emphasis on using social media data for social good, this exploration into the emotional responses during health crises illustrates the benefit of utilising social media data and NLP

techniques to understand and respond to critical societal issues. By harnessing social media to detect emotions during disease outbreaks, this chapter contributes to the main goal of AI for social good, fostering healthy lives, and promoting well-being, as articulated in the primary introduction.

Disease outbreaks have remained a problem over many years. Most recently, the Coronavirus disease (COVID-19) pandemic has left people across the globe extremely vulnerable. As of March 2021, over 100 million people from more than 200 countries have been infected [209]. To suppress the virus, governments worldwide have introduced restrictions to human movement and social gatherings. These measures, while necessary, have caused widespread disruption to human lives and led to stress, economic hardships, and uncertainties about the future [89], stirring a diverse range of emotional responses, such as anxiety, sadness, and anger [64].

In the absence of face-to-face undertakings and meetings, people have resorted to expressing their feelings on social media. Platforms like Twitter have become essential outlets for self-reported thoughts and feelings during public health emergencies [183]. Consequently, user-generated content has proven useful in monitoring public perceptions and sentiments during past disease outbreaks, such as the Ebola and Zika epidemics [96, 18], and the current COVID-19 pandemic [150].

The study of emotion detection has become important in this context and has been extensively applied in various settings, including online health-related forums [123] and social media sites [60]. Emotion detection in public health emergencies, however, is particularly challenging, due to the lack of annotated data. Previous studies often used existing resources developed for general domains, risking a bias towards domain-general contexts. Some studies attempted to solve this problem by automatically annotating user-generated data with cues specific to such corpora, like emojis, emoticons, and hashtags [254, 94]. However, research on the usage of emojis as emotion labels in public health emergencies is limited. *This motivated my focus on analysing emojis in tweets, under the assumption that a tweet expresses an emotion if it contains an emoji.*

Detecting fine-grained human emotions from text is an even more daunting task, due to limited manually annotated data. Researchers have turned to emotional cues, such as emoticons and hashtags from texts, for distant supervision, serving as emotion labels, to build powerful deep learning models and predict fine-grained emotions accurately. For example, hashtags, such as *#sad*, *#angry* and *#happy*, have been used to automatically annotate general Twitter data with fine-grained emotions and train models to learn useful text representations in an emotional context [1].

Existing research has demonstrated that pre-trained language models achieve state-of-the-art performances on natural language processing tasks, including text classification [62], named entity recognition [48] and question answering [95]. These models' standard workflow involves initial pre-training on a large amount of unlabelled corpus data using a language model loss function, then fine-tuning the pre-trained model on labelled data to adapt to a specific downstream task. Despite their success in sentiment and emotion analysis [44, 141], these models often fail to consider task-specific objectives that may improve performance.

Recent research efforts have enriched models with sentiment information, by masking sentiment words and predicting the masked words [240]. This approach has shown success in capturing rich sentiment knowledge. Other works aimed to incorporate sentiment knowledge by masking emoticons in a text and predicting if the masked token was an emoticon or not [272]. Yet, most of these techniques mainly focus on identifying positive or negative sentiments, without providing fine-grained emotional understanding.

Although pre-trained language models have excelled in a diverse range of NLP tasks [61], they often fail to consider knowledge essential for tasks related to the determination of emotions, valence, or affective states from text [121]. For instance, studies have shown that learning sentiment-specific knowledge during pre-training can enhance text sentiment understanding, thereby improving sentiment analysis performance [234, 240]. However, there is limited research on incorporat-

ing emotion-specific knowledge for fine-grained emotion detection. As sentiment analysis is closely related to emotion detection, I argue that integrating emotional knowledge into pre-trained models will yield better performance for fine-grained emotion detection. This Chapter explores these concepts and aims to infuse a range of emotions into model training, specifically targeting the improvement of fine-grained emotion detection.

My contributions in this Chapter are twofold: (1) I propose EmoBERT, a new emotion-based variant of the BERT transformer model, able to learn emotion representations and outperform the state-of-the-art; (2) I provide a fine-grained analysis of the pandemic's effect in a major location, London, comparing specific emotions (annoyed, anxious, empathetic, sad) before and during the epidemic.

5.2 Model

The success of Deep Neural Network (DNN) has provided the ability to learn useful representations from large data (with or without annotation). DNNs have been widely used in NLP [83]. The captured knowledge can be leveraged and used in downstream tasks. As a result, pre-trained language models, such as BERT [61], have been empirically successful across various tasks. BERT utilised Transformer networks [247] and was trained on large amounts of unlabelled data from the Bookcorpus* (containing 800 million words) and English Wikipedia (containing 2.5 billion words). Other variants of BERT, such as XLNet [266] and RoBERTa [147], have been proposed since the launch of BERT. Thus, a BERT-based architecture, infused with emotional knowledge, is employed for the task of fine-grained emotion detection (detecting a specific emotion e.g. sad or anxious) in tweets. The aim to instil in the model a strong inductive power and learn useful emotional knowledge. Unlabelled tweets data related to the target domain is exploited to learn this knowledge.

*https://huggingface.co/datasets/bookcorpus

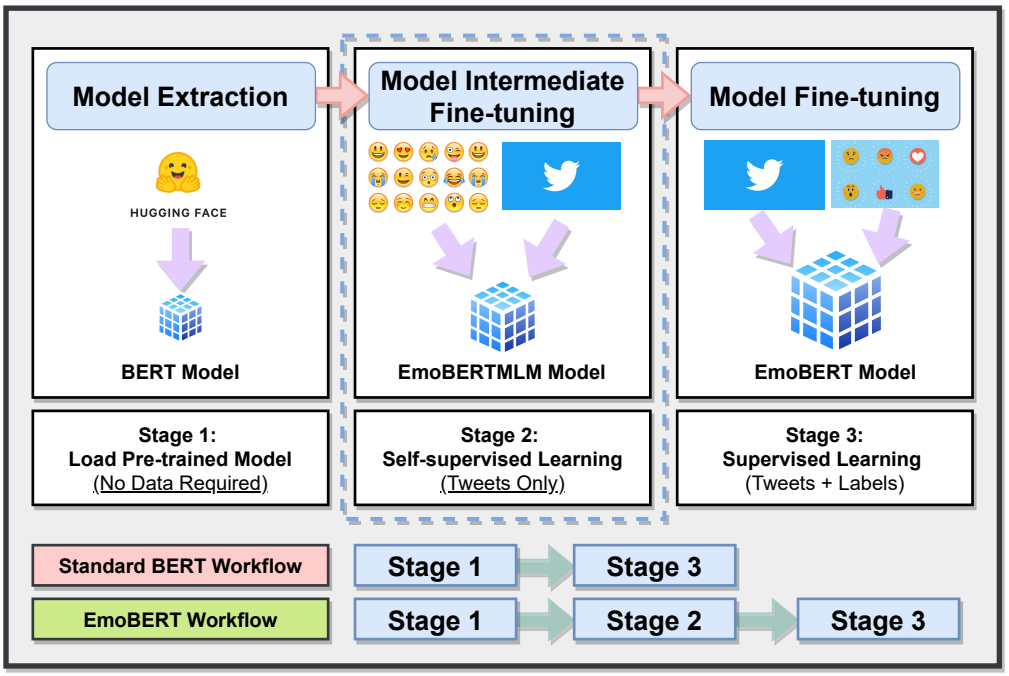


Figure 5.1: EmoBERT Architecture

Recently, models such as BERTweet [175] have been developed, purposefully trained on domain-specific text, in this case, English tweets. However, the model proposed in this study stands out, as it combines both domain adaptation (to health-related tweets) and task adaptation (for emotion detection). This combination has been shown to yield the best performance for downstream tasks [90]. Furthermore, the solution proposed in this Chapter precedes the proposal of BERTweet.

By adapting the model to the data closer to the distribution of the target data and downstream task, the model will be capable of learning emotion-related bias. The approach employed in this chapter follows the standard BERT workflow, pre-training and fine-tuning. However, the fine-tuning step involves an intermediate step that helps incorporate emotional knowledge into our model (see Figure 5.1).

5.2.1 Pre-Training

Pre-trained language models have become popular in recent years, because they became capable of learning knowledge that is useful for transfer learning in NLP. One of the best known pre-trained language models, as aforementioned, is BERT.

It was pre-trained using two language modelling objectives: (1) Masked language modeling (MLM), where randomly masked tokens are predicted, and (2) Next sentence prediction (NSP), predicting two input sentences that are next to each other. The pre-trained BERT could then be used to fine-tune on downstream tasks, such as text classification. Due to limited computational power, *BERT-base*, a version of BERT which contains 12 transformer layers is used to initialise EmoBERT. It is expected the general knowledge captured from pre-training BERT will be useful for the fine-tuning step.

5.2.2 Emotion Knowledge Enhanced Fine-tuning

The first fine-tuning stage is the major novelty of the proposed approach. The goal in this intermediate fine-tuning step is to enhance the model with emotional knowledge for emotion detection. Formally, given the target domain D_T, the source domain D_S, and the tweet representations $X_s \in D_S$ belonging to the source domain, the aim is to learn the target domain tweet representations X_T via modelling their marginal probability distribution $P(X_T)$ over the target domain D_T with explicit modelling of emotion knowledge. Note that both X_S and X_T consist of N tweets and can be denoted in the general form as $X = \{x_1, x_2, ..., x_N\}$. At this stage, BERT is fine-tuned, by using the MLM objective to recover emotion information, while learning about the tweet representations X_T and the distribution of the domain data $P(X_T)$. The intermediate fine-tuning task is based on two concepts: (1) extraction of tweets with emotion emojis, and (2) emotion word masking.

5.2.3 Extraction of Tweets with Emotion Emojis

As discussed, it is assumed that if a tweet contains an emotion emoji, then it carries some emotional information. Facial expression emojis allow Twitter users to express their feelings with non-verbal elements in a tweet [128]. Although emojis in texts may not always reflect the emotions experienced by users, recent works have

shown that they can still be used to classify the emotional content of a text [71, 261] accurately. For example, *sad* emojis such as 😔 and 😫 show a strong correlation with sad emotions in tweets, whilst anger emojis such as 😠 and 😖 are strongly associated with anger emotions[261]. Our use of tweets with emotion emojis has two advantages. Firstly, Twitter is very noisy, and it contains many tweets that do not express any emotion. This way, spurious tweets can be eliminated and tweets with, arguably, emotional information are preserved. Secondly, as discussed above, emotional cues from the emojis present can be leveraged and the information used to enhance the detection of emotions from tweets in the model. Emojis from the "Smileys & Emotion" category, based on the official Unicode emoji set*, and part of the most commonly used emojis on Twitter†, are utilised to filter unlabelled tweets datasets. This leads to emotion specific biases being implicitly induced as the training is carried out only on tweets containing at least one emoji from the emoji shortlist.

5.2.4 Emotion Word Masking

I use the same MLM objective function as in the BERT paper [61] to fine-tune the pre-trained model extracted from the Huggingface library [259]. The MLM objective was to predict the original tokens, given the masked token. BERT's standard masking strategy involved randomly sampling and selecting 15% of the input tokens and then replacing 80% of these sampled tokens with a special masked token [MASK]. Another 10% were replaced with a random token, while the remaining 10% of the sampled tokens were kept unchanged. In addition to the standard MLM for BERT, the proposed approach uses emotion word masking (see Figure 5.2) for learning emotional knowledge in the text. This masking process is different from the standard MLM in BERT. Since the emotion words are likely to impact on the emotion expressed in a tweet, a higher masking probability (50%) is employed

*https://unicode.org/emoji/charts/full-emoji-list.html
†as indicated in https://emojitracker.com

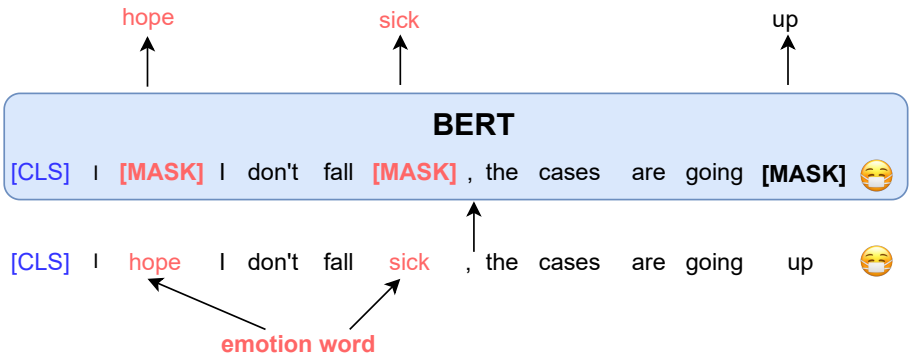


Figure 5.2: **Emotion word masking**: recognises the emotion information in a tweet and masks it; then, the pre-trained model can attempt to recover this information.

for masking emotion words.

To determine emotion words, the NRC Word-Emotion Association Lexicon (a.k.a. EmoLex) [164] was used. EmoLex consists of 14,182 crowdsourced English words associated with eight basic emotions: *anger, anticipation, disgust, fear, joy, sadness, surprise and trust* [196]. Whilst there are other lexical emotion resources, such as DepecheMood [230], I chose EmoLex due to its large size and coverage of broad emotional dimensions [82]. Additionally, past works such as [229] have used this lexicon as a prior association of emotions, to detect emotions in texts automatically. An initial examination of the lexicon showed that some lexicon words are associated with no emotion category mentioned above. Such words were removed, since they were deemed not useful for the emotion-word masking and select only words associated with at least one emotion from the list. A total of 4,463 words from the lexicon were used after the elimination process.

BERT used the WordPiece algorithm [264] as its tokenisation algorithm, to deal with words that are out of its vocabulary, by splitting them into sub-word tokens present in its vocabulary. Given this, it is expected that some emotional words would be out of BERT's vocabulary. For example, the emotion word '*somatic*' could be split into '*so*' and '*##matic*' by the BERT tokeniser. In the proposed model, I define the masked emotion word as the sub-word tokens that correspond to the original emotion word.

Since emotion words appear in emotional contexts, the proposed model aims to capture implicit emotion knowledge representations and preserve information that could be useful in detecting the emotions expressed in a tweet. For the intermediate fine-tuning task, a collection of original (no retweets) English tweets collected in April, 2020, related to COVID-19, using keywords such as *coronavirus, corona, covid, covid-19, coronaoutbreak, 2019nCoV, pandemic, epidemic, wuhanandlockdown* and *sars-cov-2* were used. Consequently, the training tweet datasets were filtered to contain at least one emoji from the emoji selection and remove all duplicate tweets. In total, the training dataset contained 1,540,983 tweets after the data filtering process. Furthermore, all tweets were pre-processed, by replacing all Twitter usernames and URLs with the common tokens: *<user>* and *<url>*, respectively.

5.2.5 Emotion Detection Fine-tuning

In the final step, I fine-tune EmoBERT on the downstream task: fine-grained emotion detection. Outputs from the emotion knowledge-enhanced fine-tuned model are trained to classify the emotion of a tweet. Following the fine-tuning setting in the original BERT paper [61],the last state vector of the classification token [CLS] is used as input and fed it into a feed-forward neural network, to predict the respective emotion.

5.3 Experiment

I used a publicly available annotated dataset, SenWave*, providing fine-grained emotion labels of tweets during the COVID-19 pandemic [265], to evaluate the proposed model. This dataset consists of 10,000 English tweets, labelled with 10 categories: *optimistic, thankful, empathetic, pessimistic, anxious, sad, annoyed, denial, official report* and *joking*. I considered 3 emotions (*annoyed (anger), anxious*

*instructions on accessing data can be found at https://github.com/gitdevqiang/SenWave

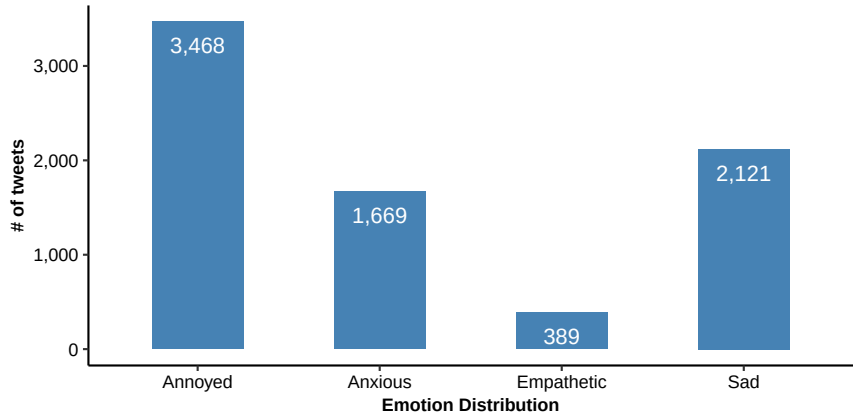


Figure 5.3: Emotion Distribution

and *sad*), which are part of the fundamental emotions from Plutchik's model [196], and an additional emotion (*empathetic*), which has been shown to be expressed in discussions about chronic illnesses in online health communities [138]. I believe the chosen emotions are related to individuals' mental well-being and can provide some insights into the public mood during the COVID-19 pandemic. Figure 5.3 shows the distribution of the chosen emotions from the annotated dataset.

Since a tweet can be annotated with more than one emotion, 4 binary classification tasks: *annoyed/non-annoyed*, *anxious/non-anxious*, *empathetic/non-empathetic* and *sad/non-sad* were created, to determine if an emotion is expressed in a tweet. To create the negative samples for an emotion category, I sample an equal amount from the other emotion categories. Following this, I shuffled the positive and negative samples and split them into training, validation, and test sets using an 80/10/10 split. While the number of positive and negative samples was matched and balanced, the resulting splits were not guaranteed to have the same number of samples. Nonetheless, this approach created a more natural distribution of labels in each set. To reduce the likelihood of sampling bias when selecting negative samples, I performed negative sampling (with replacement) ten times for each emotion category, generating ten dataset samples per category.

For each emotion category, I used the ten dataset samples to construct ten sets of training, validation, and test data. The models were trained using all data samples and evaluated by reporting the mean performance across the ten data samples

Table 5.1: Emotion detection results averaged across 10 dataset samples. The numbers are percentages. Best results are in **bold**. Precision - P, Recall - Re and F1 score - F1.

Model	Annoyed			Anxious			Empathetic			Sad		
	Pr	**Re**	**F1**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
BERT	0.78	0.80	0.78	0.68	0.73	0.69	0.83	0.74	0.77	0.71	0.77	0.73
XLNet	0.74	**0.83**	0.77	0.63	**0.81**	0.69	0.73	0.67	0.69	0.69	0.77	0.72
EmoBERT	**0.81**	0.82	**0.81**	**0.72**	0.73	**0.72**	**0.84**	**0.79**	**0.81**	**0.71**	**0.84**	**0.76**

per category. During training, I employed binary cross-entropy loss to align with the target class distribution, and used the Adam optimizer with a learning rate of $5e-5$. Based on prior empirical experiments, I set the number of epochs to 4 and the batch size to 16 for the final model.

To evaluate the performance of EmoBERT, I compared it with two widely-used pre-trained models: BERT [61] and XLNet [266]. These models have achieved state-of-the-art results on a variety of natural language processing tasks, making them strong baselines for comparison purposes.

5.4 Results and Discussion

Table 5.1 presents the results obtained using EmoBERT, compared with generally pre-trained language models, i.e. BERT and XLNet. EmoBERT achieves higher performance across all the considered emotion categories. On average, EmoBERT outperforms other state-of-the-art approaches by at least 3% F1 score. This shows that incorporating emotion-specific knowledge in pre-trained language models is effective for detecting fine-grained emotions.

5.4.1 Significance Test

To determine if the results are statistically significant, I performed a paired T-test to test the differences between the model results for all 10 runs per emotion category. The result confirms that the improved F1 score results from EmoBERT

over BERT are statistically significant at $p < 0.05$ across all emotions, except for the anxious emotion.

5.4.2 Tracking Emotional Toll of COVID-19 on Twitter

To further illustrate the power of the proposed model, I conducted a focused analysis of the impact of the COVID-19 pandemic on well-being and emotions in London, United Kingdom (UK), using EmoBERT.Specifically, I examined a collection of geo-located tweets in London during March 2020 and compared them with the same period in 2019. To obtain these tweets, I utilised the new Twitter API for academic research, which grants access to the full Twitter archive*. Initially, I collected all tweets geo-located in the UK and extracted tweets that contained "London" in the full name of the place information†. I excluded non-English tweets, replies, and retweets from the collected tweets. The resulting sample consisted of 361,384 tweets for March 2020 and 352,678 tweets for March 2019. I chose this period because it corresponds to the onset of the COVID-19 pandemic in the UK and provides a clear delineation between pre-pandemic and pandemic periods.

Each tweet from the respective periods is classified into one of four emotions (*annoyed, anxious, empathetic*, and *sad*). As the evaluation of the model is performed on 10 distinct dataset samples, an ensemble of 10 classifier models is generated for each respective emotion, and the average of their probabilities is utilized to predict the final class labels. For all emotions, each tweet is classified as expressing the emotion or not (*e.g., sad/non-sad*). The probability of the positive class membership of the emotion expressed in each tweet is used to measure the emotion present in tweets on day d. To achieve this, I employed logistic regression classifiers and define the probability of emotion p_d present in tweets on day d as follows:

*https://developer.twitter.com/en/solutions/academic-research/products-for-researchers

†https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/geo#place

$$p_d = \frac{1}{1 + exp(-x_d)} \tag{5.1}$$

where:

$$x_d = \frac{1}{N_d} \sum_{t=1}^{N_d} \ln(\frac{p_{td}}{1 - p_{td}}) \tag{5.2}$$

Where x_d denotes the proportion of emotion present in tweets on day d. N_d is the number of tweets on day d and p_{td} represents the probability of emotion expressed in tweet t on day d. We take the mean of the logit function in equation 5.2, because the means of skewed variables are not necessarily representative of those variables. Probabilities tend to exhibit such skewness, because they are bounded, so it is often cleaner to do algebraic manipulations on an unbounded scale, such as logit, then back-transform [106].

Figure 5.4 shows the comparisons of the emotions for each respective year. Generally, the proportion of tweets expressing emotions in 2019 show a similar trend over time. There is no notable change in the emotions for that period. Similarly, the tweets expressing *annoyed* and *sad* emotions in 2020 are consistent during the period. Interestingly, the tweets expressing the *anxious* emotion rise sharply after the 5th of March (after the first COVID-19 related death was announced in London)* before levelling off and declining around the 20th of March 2020. The tweets expressing *empathetic* emotion rise slightly in the middle of the month and begin to decrease after about 10 days.

Although the trend in tweets expressing *annoyed* and *sad* emotions over time is similar in 2019 and 2020, the level of emotions is slightly higher in 2020 than in 2019.

I measured the effect size using Cohen's d to determine the difference between each respective year's emotions [209] and report *p-value* after Benjamini-Hochberg p-correction. On average, more tweets expressed *annoyed* (Cohens $d = 0.21$, $p = 0.05$), *anxious* ($d = 1.99$, $p < 0.05$), *empathetic* ($d = 0.73$, $p < 0.05$) and *sad* (d

*https://www.bbc.co.uk/news/uk-england-london-56271001

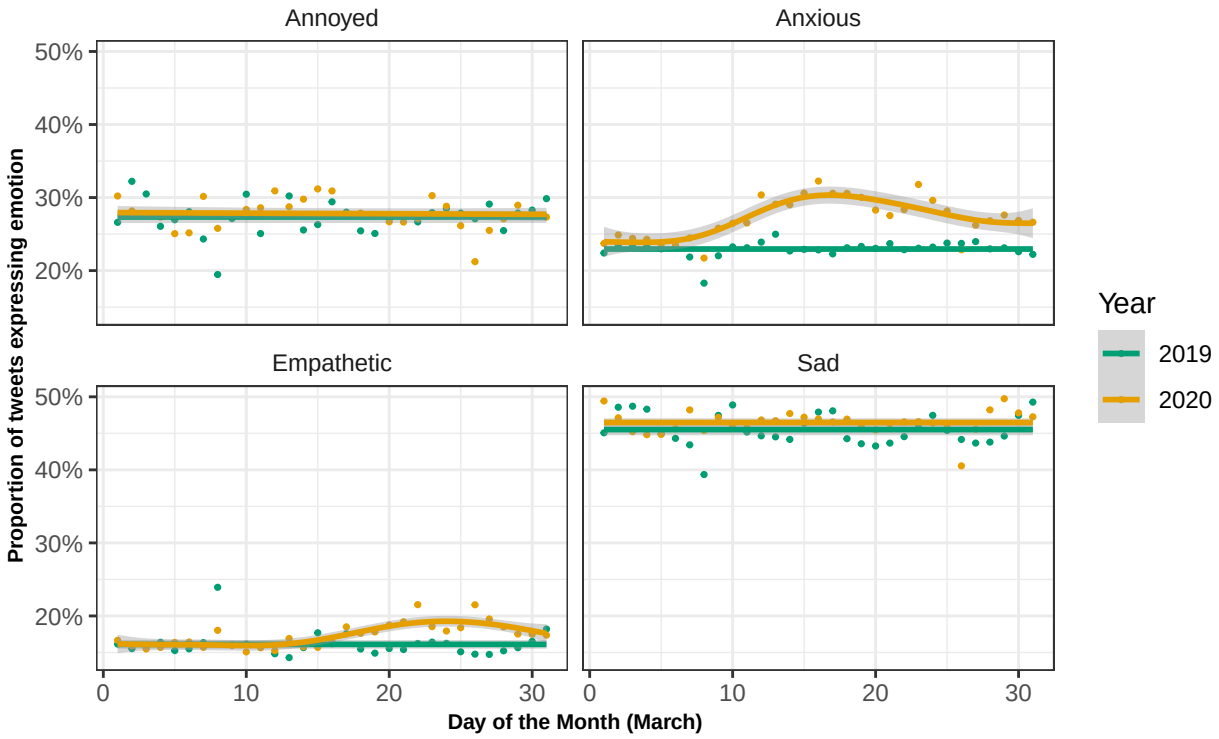


Figure 5.4: The trend of emotions from the 1st March through 31st March for years 2019 and 2020

$= 0.5$, $p < 0.05$) emotions in 2020, compared to 2019, suggesting that there is a difference in the expression of emotions in tweets as a result of the COVID-19 pandemic.

To understand more deeply the distribution of emotions in tweets related to the COVID-19 pandemic, I further extract, from the tweets geo-located in London, those containing the following keywords: *coronavirus, corona, covid, covid-19, coronaoutbreak, 2019nCoV, pandemic, epidemic, wuhan* and *lockdown* and analyse them in-depth. I build a matrix of the emotion distributions with the most frequent hashtags that appear in the filtered tweets. Hashtags are strong indicators used to provide context, emotions or topics related to a tweet. Figure 5.5 shows the heatmap of the distribution of emotions and the top 40 hashtags. The distribution of emotions is calculated as the number of instances in which the emotion is expressed in a tweet with the hashtag, divided by the total number of tweets containing that hashtag from the body of extracted tweets [85]. As can be seen in Figure 5.5, tweets expressing *sad* and *anxious* emotions are prevalent in the hasht-

Figure 5.5: Top 40 hashtags for COVID-19 related London tweets in March 2020

ags used in the COVID-19-related tweets. Hashtags in tweets expressing *empathetic* emotions are common in campaigns about staying safe and staying at home, while hashtags in tweets expressing *annoyed* emotions are more related to the crisis and chaos in the UK, as a result of the pandemic.

The majority of the hashtags used in tweets expressing the *sad* emotion are #isolation, #quarantine, #londonlockdown and #coronapocalypse. London is a travel hub and, understandably, the tweets are expressing sadness about travel restrictions, such as quarantine, isolation and lockdown in London. The tweets also express sadness about what some people have regarded as the 'end of the world' because of the pandemic. Amongst tweets with the #coronavirusoutbreak, #covid-19, #coronavirusupdates and #coronaoutbreak hashtags, the *anxious* emotion is preponderant. This suggests that a significant amount of tweets are expressing anxiety about the COVID-19 pandemic and particularly around the updates provided on the virus.

The majority of tweets with hashtags such as #coronavirusuk, #coivd-19uk and #coronacrisisuk express the *annoyed* emotion predominantly. Among these, hashtags related to the coronavirus crisis in the UK are most common. It can be seen that #staysafe, #stayhomesavelives and #nhs hashtags, which are related to the UK government official campaign advising people to 'stay home, save lives and protect the NHS (National Health Service)', appear more frequently in tweets expressing the *empathetic* emotion.

5.5 Discussion

In this Chapter, I proposed a new method for analysing the emotional health and well-being of both global and local populations, specifically in London, by utilising EmoBERT, a novel model that incorporates emotion representation into the cutting-edge BERT model. As the current pre-training objectives of BERT do not consider knowledge relevant for emotion detection tasks, I developed a new pre-training objective to induce emotion-specific bias into the original BERT model, ultimately outperforming current state-of-the-art methods. Evaluation results demonstrated significant improvements in emotion detection, enabling the proposed EmoBERT to *efficiently detect fine-grained emotions during major disease outbreaks*, as presented in response to **RQ2**.

Furthermore, I selected emotions related to health and health communities to showcase a methodology for an in-depth comparison of social media emotions both before and during the COVID-19 pandemic. I also demonstrated how these selected emotions could be used to understand the individual topics that are likely to evoke them. Applying this methodology locally to London, a major location expected to be strongly affected, confirmed the significant impact of the pandemic on emotional well-being. Future research can apply the defined methodology to other specific areas. However, limitations of this approach include the use of only one (freely available) social media platform, Twitter, which may reflect the emotions of only a specific section of the population. Therefore, incorporating data from other social media sites such as Facebook or Weibo, as well as data from the World Health Organisation (WHO) or local health authorities, could potentially increase accuracy. In terms of future work, it would be worth expanding the scope of the emotional analysis across a broader timeframe. Longitudinal studies that track emotional trends over extended periods could provide a deeper understanding of how communities emotionally respond and adapt to prolonged crises, such as the COVID-19 pandemic.

Overall, the proposed method of analysing emotional health and well-being using EmoBERT has the potential to provide valuable insights into how emotions and health are related, which can lead to the development of effective interventions and policies that promote social good. To aid reproducibility of the work in this Chapter, I release the implementation of the proposed models here - `https://github.com/tahirlanre/EmoBERT`.

Chapter 6 continues this work by focusing on emotions expressed in health-related social media posts and improving the detection of health mentions on social media.

Chapter 6

Incorporating Emotions into Health Mention Classification Task on Social Media

Detecting health mentions on social media is crucial for complementing existing health surveillance systems. However, the task of annotating data for detecting health mentions at a large scale can be challenging. To address this, I propose a framework for incorporating affective features into the HMC task.

Additionally, I evaluate the approach on 5 HMC-related datasets from different social media platforms, including three from Twitter, one from Reddit, and another from a combination of social media sources.

6.1 Introduction

In this chapter, I build on the findings of Chapter 5 that emotions are often expressed in tandem with discussions of health on social media. With this understanding, I aim to leverage the emotions expressed in social media posts to improve the performance of classifying health mentions on social media, to answer **RQ3**. *How can health mentions be detected on social media to track health-related con-*

versations? By connecting emotions to health discourse, this chapter extends the thesis's overarching ambition to employ social media data for social good. It adheres to the principal goal of understanding human well-being through social media analysis, offering another facet of how NLP techniques can contribute to positive social impact.

Social media platforms such as Twitter and Reddit are increasingly used by people to share personal health experiences. Their widespread availability, ease of accessibility, and the near real-time nature of the data they generate make them invaluable for public health surveillance. However, the large volume, rapid generation rate, and unstructured nature of these data pose significant challenges. Moreover, potential biases in such data may also exist [75]. Despite these challenges, social media data have shown significant applications in areas such as health informatics, public health, and medical research [65].

A critical step in harnessing social media data for public health surveillance involves detecting content related to health reports, a task known as HMC [33]. The HMC task aims to develop algorithms and models that can accurately identify and classify health mentions in a text, enabling automated analysis and interpretation of large volumes of health-related data. In this task, text documents are analyzed, and any mentions of health-related entities such as diseases, symptoms, treatments, and other medical concepts are identified and labeled according to a predefined set of categories. This task is challenging, due to the complex nature of natural language and the wide range of health-related entities that can be mentioned in a text. To illustrate, consider the post "Every time I wrap gifts it looks good until I rapidly develop Parkinson's in both of my hands" which is a health mention, while "Congratulations to Coach Parkinson on receiving a contract extension through the 2021-2022 season! #JagsROAR" is not. While the former implies the author's affliction with Parkinson's disease, the latter merely refers to a person named Parkinson.

Previous work on HMC using social media has primarily focused on Twitter posts

[108, 118]. Twitter is popular for public health applications due to the public availability of its contents through the Twitter API. However, the 280-character limit on tweets can make it difficult to distinguish different contexts. More recently, Reddit, with its longer posts and moderated discussions, has been collected for HMC tasks [172]. Like Twitter, Reddit data is publicly accessible through the Reddit API. Other dedicated online health forums, such as *AskaPatient* and *patient.info*, also exist for health experience discussions.

Various NLP techniques have been employed to improve performance on HMC tasks. These range from methods that use contextual to non-contextual word representations [33, 118]. Past research has also considered modeling the literal or figurative usage of disease or symptoms words within texts expressing personal health experiences [105]. Another body of work investigated using a combination of user behavioural information, such as sentiment and emotion, with other features [172].

Different from the above works, my approach considers the relationship between the self-reporting of personal health experiences and the emotions expressed in these reports. As the act of discussing personal health experiences often triggers an emotional response, incorporating these emotional responses can potentially improve the performance of our target task, i.e., HMC in social media texts.

In this chapter, I explore emotions conveyed in social media texts describing personal health experiences. The aim is to improve HMC by implicitly incorporating emotional knowledge into the target task through an intermediate emotion detection task. I hypothesise that leveraging the inductive bias from the emotion detection task will improve HMC performance over baseline methods. To improve results further, I also propose explicitly combining HMC-specific and affective features to model the relationship between emotions and health mentions.

I evaluate the proposed approaches on five datasets sourced from popular social media platforms such as Twitter and Facebook, as well as online health communit-

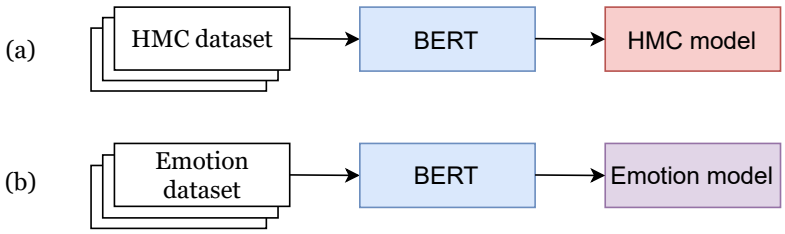


Figure 6.1: The structure for (a) HMC model and (b) Emotion detection model

ies. These datasets vary in size and characteristics. I also explore the potential benefits of focusing solely on negative emotions. Finally, I compare the performance of cross-task transfer from respective HMC models to a direct transfer from an emotion model, to understand the relative effectiveness of these approaches.

6.2 Methodology

Firstly, I describe the models for HMC and emotion detection (see Figure 6.1). Following that, I present the framework for incorporating emotions into the task of HMC with two different types of enhancements. Both enhancements aim to enrich the neural representations learned by BERT with emotional knowledge.

6.2.1 Health Mention Classification

Social media posts related to health mentions are usually extracted using symptoms or disease-related keywords. On the other hand, people on social media often use slang and varied representations of a word, which contribute to a high noisiness of social media posts. Thus, the presence of a symptom or disease word does not necessarily mean it is health-related. State-of-the-art approaches have been proposed to model contextual relationships between words in a text [61]. Such models represent every word dependent on the particular context of occurrence. Incorporating contextual information is essential for language understanding tasks, even more so for correctly classifying health-related social media posts.

Table 6.1: Distribution of labels and examples for each HMC dataset

Dataset	Label	Size	Text
FLU2013	Flu infection (positive)	1,280	*<user> Ugh. I'm getting a flu shot (hopefully) in about half an hour. :(Sorry yours is being ugly!*
	Flu awareness (negative)	1,342	*I hope Is there some kind of flu going around? It's like everyone's getting sick all of a sudden. Weird.*
PHM2017	Self-mention	306	*Officially now a cancer patient (1991)*
	Other mention	516	*<user> set a goal after her #stroke: walk in high heels again <url> #2health #ForOurHearts <url>*
	Awareness	1,278	*#Stroke threatens millions of lives. Learn the signs: <url> #ForOurHearts <url>*
	Non-health mention	2,483	*You are Alzheimer's mascot you master of socialism <url>*
SELF2020	No self-disclosure	2,954	*There is an otosclerosis community FB page which is quite helpful.*
	Possible self-disclosure	2,586	*Im basically taking one day at a time. I guess some viruses are unknown to medicine. So is what it is.*
	Clear self-disclosure	1,010	*Dementia and its Genetic Markers, many are known, however that may not mean you will end up with a problem. I have a Congenital Short Term Memory Defect from Birth, and I have had to relatives who died from Dementia.*
ILL2021	Negative	18,435	*Brain 'pacemaker' could prevent tremors and seizures for Parkinson's and epilepsy sufferers*
	positive	3,872	*'I'm not OK': <user> gets emotional talking about 5-year-old son's cancer battle*
RHMD	Figurative mentions	3,430	*Addiction to a Toy **As a kid, I was always addicted to this one toy called a slinky. I would spend hours and hours just fiddling with it. It seemed so satisfying to me. Whenever I would lose it, I would go into a depressing state for days and days, until I found it again. is it just me who has an addiction to a specific type of toy**.*
	Non-health mentions	2,586	*Court let Merck hide secrets about a popular drug's risks - Lawsuits claim baldness drug Propecia causes sexual problems and depression. The judge sealed evidence suggesting the maker downplayed the side effects.*
	Health mentions	3,360	*I was diagnosed with Asperger's, OCD, Major depressive, and PTSD while I was inpatient. Ask me anything I was inpatient for 6 days due to homicidal thoughts and urges towards those who had hurt me emotionally and physically. And I put that hatred on others who did nothing wrong. In Inpatient I was diagnosed with Asperger's, OCD, and later after Outpatient, PTSD. I was abused by my mother, and three friends over the years. Physically and Mentally. Ask me anything.*

To train the health mention classification model, I leverage a context-sensitive model, BERT [61]. BERT is fine-tune on the respective HMC dataset. Fine-tuning BERT on task-specific corpus often yields performance gains on downstream tasks [61]. I refer to the model fine-tuned on HMC dataset as $BERT_{HMC}$.

6.2.2 Emotion Detection

To capture fine-grained emotions, I leverage existing datasets annotated with emotions. I consider two publicly available emotion datasets - *GoEmotions (GE)* [58] and *SemEval18 - Emotions (SE)* [162]. *GoEmotions* is a benchmark emotion dataset containing 58,000 Reddit comments manually annotated with 6 Ekman emotions (*anger, disgust, fear, joy, sadness, surprise*) and *neutral* for experiments. *SemEval18 - Emotions* comprises 10,896 tweets manually annotated with 11 emotion labels – (*anger, disgust, anticipation, fear, joy, love, optimism, pessimism, sadness, surprise and trust*), each of which is a binary label that denotes the presence of a specific emotion.

Based on standard practice in NLP and as shown in Chapter 5, I fine-tuned BERT [61] on the emotion dataset to learn general emotion representations. The domain-specific nature of emotion expressed in social media texts made this step crucial. I refer to the emotion model fine-tuned on *GoEmotions* and *SemEval18* as $BERT_{GE}$ and $BERT_{SE}$ respectively.

6.2.3 Emotion Incorporation Framework

Studies have shown that social media users typically express a range of emotions when posting about personal health updates [138]. Building on this, I aim to capture the emotion spectrum when people post about their personal health experiences on social media. We consider two approaches to incorporate emotions into HMC. Both approaches aim to enrich the neural representations learned by BERT with emotional knowledge.

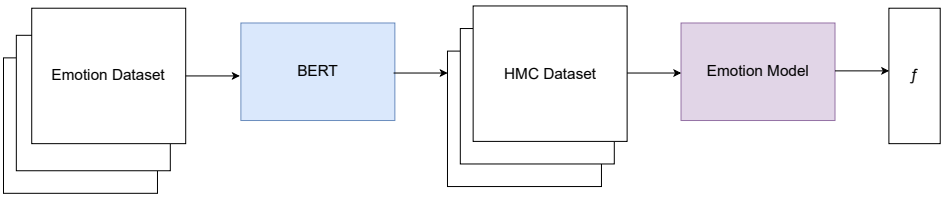


Figure 6.2: *Implicit* emotion incorporation with intermediate task fine-tuning

6.2.4 Intermediate Task Fine-tuning Approach

Recent work has shown that initially fine-tuning on an intermediate task before fine-tuning on a target task of interest improves the performance of pre-trained models [194]. Notwithstanding, the effectiveness of this approach depends highly on the intermediate task that is applied [43]. The intuition behind intermediate fine-tuning is that if both tasks are related, the linguistic knowledge learnt in the intermediate task can contribute to understanding the target task. Following this observation, I hypothesize that emotion detection tasks can assist the task of HMC.

To improve upon the baselines, I explore intermediate fine-tuning as a means of *implicitly* incorporating an emotion-specific inductive bias into the target task. I follow the emotion detection approach (6.2.2) described above to serve as an intermediate task. The intermediate fine-tuning step *implicitly* learns affective features that could be helpful for the target task. Specifically, I use the fine-tuned emotion model parameters to initialise a new BERT model and then fine-tune on the HMC task. The approach is illustrated in Figure 6.2.

6.2.5 Multi-Feature Fusion Approach

Research has demonstrated improved performance on HMC tasks when sentiment and emotional features are combined [33, 172]. They were generated, however, using emotion lexicons. Pre-trained language models capture better emotions expressed in social media texts due to their success in natural language understanding tasks [215]. I hypothesise that by combining emotional information, the HMC specific sentence encoder could be guided to detect the nuances of reporting personal

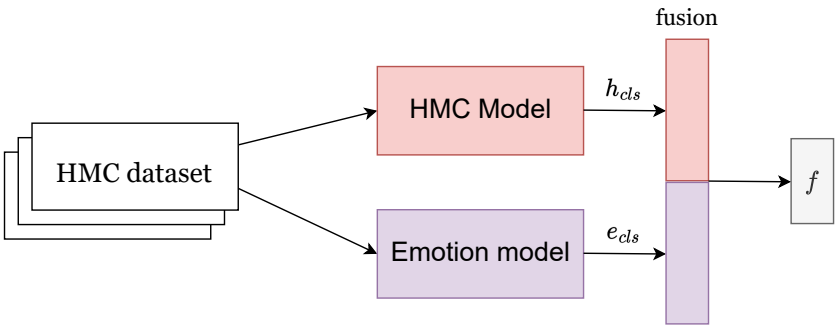


Figure 6.3: *Explicit* emotion incorporation with multi-feature fusion

health experiences on social media. To achieve this, I *explicitly* combine affective and HMC-specific linguistic features. I extract HMC-specific features from the HMC model and then these representations are fused with affective features extracted from the emotion model. In this approach, emotional information is incorporated *explicitly* via the extracted affective features. The approach is illustrated in Figure 6.3.

6.3 Experimental Setup

6.3.1 Datasets

I explore a variety of HMC-related datasets from different social media platforms to study the general applicability of the proposed approach. I use three Twitter datasets - *FLU2013* [130], *PHM2017* [118], *ILL2021* [119], one Reddit dataset - *RHMD* [172] and one from a combination of Facebook, Reddit, Twitter and *patient.info* - *SELF2020* [244]. These datasets are annotated for classifying mentions of health-related concepts in social media text (e.g. *health mention/non-health mention* or *flu infection/Flu awareness*). Table 6.1 presents the summary of all HMC datasets.

FLU2013 This dataset was created by [130] to distinguish between reports of actual Flu infections and awareness. Each tweet in the dataset was manually

annotated as either a *flu report* (positive) or a *flu awareness* (negative). At the time of this study, only 2,622 tweets were available to download, which is about 58% of the original dataset.

PHM2017 Another existing dataset which focuses on more than one disease and condition was constructed by Karisani and Agichtein [118]. In the corpus, they collected English tweets related to Alzheimer's disease, heart attack, Parkinson's disease, cancer, depression, and stroke and manually annotated them in terms of *self-mention*, *other-mention*, *awareness* and *non-health*. At the time of this study, only 4,987 tweets were available to download, which is about 69% of the original dataset.

SELF2020 This dataset consists of health-related posts covering a range of health issues collected from online communities, including *patient.info* and social media platforms (Facebook, Reddit and Twitter) [244]. The dataset consists of 6,550 posts annotated as either *no self-disclosure*, *possible self-disclosure* or *clear self-disclosure*. The majority (88.1%) of the posts are from patient.info, thus the dataset contains phrases and sentences that are mostly longer than the Twitter-based datasets.

ILL2021 The ILL2021 dataset is an illness report dataset related to three different health conditions: Parkinson's disease, cancer, and diabetes [119]. The dataset is annotated to detect if a tweet mentions the health condition and contains a health report (*positive*) or not (*negative*). 22,307 tweets (98% of the original dataset) were available for download at the time of this study.

RHMD The RHMD dataset focuses on Reddit posts only [172]. The posts contain keywords related to up to 15 diseases and symptoms such as Headache, OCD and Allergic. In total, the dataset contains 10,015 unique posts. They are labelled

with *figurative mention, non-personal health mention and health mention.* In terms of length, the posts are longer than the Twitter-based datasets.

6.3.2 Model Architecture

I describe the approach to fine-tune BERT for both HMC and emotion detection. Given an input sequence, I use a WordPiece tokenizer to tokenize the input as described in [61]. The tokenizer adds two special tokens *[CLS]* and *[SEP]* to the input sequence, and the tokenized input is represented as:

$$X = [x_{[CLS]}, x_1, x_2, ..., x_n, x_{[SEP]}] \tag{6.1}$$

where x_t is the contextualised embedding of the t-th token in a sequence of n symbols. The tokenized input is then passed into the BERT model to yield a sequence of hidden states as follows:

$$H = [h_{[CLS]}, h_1, h_2, ..., h_n, h_{[SEP]}] \tag{6.2}$$

I consider the hidden vector $h_{[CLS]} \in \mathbb{R}^{768}$ from the last hidden layer as the aggregate sequence representation for both HMC and emotion models. The representation is then passed through a linear output layer for prediction:

$$\tilde{y_i} = W_y h^i_{[CLS]} + b_y \tag{6.3}$$

where W_y and b_y are learnable network parameters and $\tilde{y_i}$ is the network output. For the multi-feature approach (Section 6.2.5), the representations from both HMC and emotion models are directly concatenated* to form a combined representation. Then, the fused representations are fed into a linear layer for prediction:

$$\tilde{y_i} = W_y(h^i_{[CLS]} \oplus e^i_{[CLS]}) + b_y \tag{6.4}$$

*Initial experiments with max pooling and self-attention gave worse results.

where $h^i_{[CLS]}$ and $e^i_{[CLS]}$ are the extracted features from the HMC and emotion models, respectively. The HMC and emotion models are fine-tuned at the same time during training.

6.3.3 Model optimisation

For HMC tasks, which are single-label classification problems, the output uses a softmax activation and the network is optimised with a cross-entropy loss. For emotion tasks, which are multi-label classification problems, we use a sigmoid activation and optimise the network with a binary cross entropy loss.

6.3.4 Baselines

To serve as the baseline, I used the HMC model described above (section 6.2.1). Specifically, *bert-base-uncased* was employed, which comprises of 12 bidirectional transformer encoders with 768 hidden layers, 12 self-attention heads and has a total parameter count of 110M. The performance of emotion incorporation approaches are compared to the baseline.

6.3.5 Training

For all our experiments, I trained the models with minibatch gradient descent using the Adam optimizer [125]. I used a batch size of 128 (except for PHM2017*, batch size = 64). The number of epochs is set to 3 with a learning rate of $2e^{-5}$.

Dataset split The dataset splits were not provided by the dataset distributors hence I created my own splits. For each dataset, I performed a *80%/10%/10%* split randomly to create the train, validation and test sets respectively. To train the models, the training set was used, while the validation set was used to select

*Initial experiments with batch size = 128 gave low performance.

hyperparameters, and the test set was used to evaluate the performance of our models. The dataset splits I used for the experiments are presented in Table 6.2.

Table 6.2: Summary statistics of the dataset splits

Dataset	Train	Validation	Test	All
FLU2013	2,098	263	264	2,622
PHM2017	3,667	459	460	4,583
SELF2020	5,241	656	656	6,550
ILL2021	17,846	2,232	2,232	22,307
RHMD	8,013	1,002	1,003	10,015

Evaluation Following previous works on HMC [33, 172], I evaluated each model's performance using F1 macro score and report the results on the test set. To account for variability, I run each model five times with different seeds and report the average results over these five runs. I reported the average performance across 5 runs with different seeds on the test set.

6.4 Results and Discussion

The results are presented in Table 6.3. To determine whether the improvements are statistically significant, I use a two-sample t-test to compare the F1 scores. I assert significance if $p < 0.05$ under a two-sample t-test with the vanilla BERT model. Both of the approaches to incorporating emotional information boost performance across the HMC datasets. I also find that most of the gains are on the HMC datasets with limited samples.

For the intermediate task fine-tuning approach, I observe that fine-tuning on either emotion dataset improves performance over the respective HMC task in the majority of cases. On some HMC datasets, such as FLU2013 and SELF2020, there was at least a 3% increase in the performance. The BERT model fine-tuned on *Sem-Eval18* emotion data ($BERT_{SE}$) yields the most improvements on all but one dataset, compared to the BERT model fine-tuned on GoEmotions ($BERT_{GE}$), which only obtained better improvement on PHM2017. Though, the vanilla BERT achieved

better results on PHM2017 when compared to both $BERT_{GE}$ and $BERT_{SE}$. I note that $BERT_{SE}$ performs significantly better on RHMD which is a Reddit-based dataset, than $BERT_{GE}$, which is trained on an emotion dataset from the same domain (Reddit).

Overall, the results for the *multi-feature* approach show the benefit of combining both health mention representations and emotional information. The *multi-feature* approach consistently improves on the baseline and *intermediate task fine-tuning* across all HMC datasets. The performance on the SELF2020 data shows the most significant improvement, up nearly 7 F1 points when emotion features are generated using $BERT_{SE}$ and up more than 7 F1 points when emotion features are generated with $BERT_{GE}$. Similarly to the results from the first approach, emotion-based models trained with *Sem-Eval18* achieve the best performance in most cases.

Table 6.3: F1 macro score for the health mention classification task. **Bold** denotes the highest score and * denotes statistical significance. The average of five random seeds is used for all scores.

Model	FLU2013	PHM2017	SELF2020	ILL2021	RHMD
Baseline					
BERT$_{HMC}$	82.18	81.66	70.13	91.25	80.76
Intermediate Task Fine-tuning					
BERT$_{GE}$	82.18	81.29	73.02*	91.32	80.91
BERT$_{SE}$	85.15	80.93	74.02*	91.38	81.91*
Multi-Feature Fusion					
BERT$_{HMC}$ + **BERT**$_{GE}$	85.85*	83.59*	**77.28***	91.85*	82.64*
BERT$_{HMC}$ + **BERT**$_{SE}$	**86.08***	**83.9***	76.50*	**91.88***	**82.77***

6.4.1 Effect of negative emotions

Although a direct relationship has not been established, negative emotions have been associated with social media references to personal health [255]. For example, tweets about colonoscopies were found to express more negative sentiment on average [159]. Another study showed that users post more frequently when symptoms are worse, raising concerns about bias towards negative emotions [54]. As a result, I investigate the effect of using only texts annotated with negative emotions

to fine-tune our emotion model. I use a subset of the emotion dataset with only negative emotions (and neutral). For the GoEmotions data, I follow the negative emotions defined by the authors [58] i.e. *anger, disgust, fear, sadness, neutral*. For Sem-Eval18 - emotions data set, I used the following labels as negative emotions: *anger, disgust, fear, pessimism, sadness*. Table 6.4 shows the result when I used only negative emotions.

Results The results show no significant gain when using only negative emotions over using all emotions (positive, negative and neutral). This applies to both our approaches. The performance on most tasks deteriorated moderately for the *intermediate task fine-tuning*. While there are improvements for our *multi-feature* approach, these are relatively small and insignificant. This result shows no additional benefit to incorporating only negative emotions. Instead, taking advantage of the full spectrum of emotions might be more helpful.

Table 6.4: F1 macro score for the health mention classification task. **Bold** denotes the highest score and * denotes statistical significance. The average of five random seeds is used for all scores.

Model	FLU2013	PHM2017	SELF2020	ILL2021	RHMD
Baseline					
$\textbf{BERT}_{HMC}$	82.18	81.66	70.13	91.25	80.76
Intermediate Task Fine-tuning					
$\textbf{BERT}_{GE-neg}$	82.41	81.0	72.25*	91.39	81.67*
$\textbf{BERT}_{SE-neg}$	84.32	81.59	73.37*	91.24	81.19
Multi-Feature Fusion					
$\textbf{BERT}_{HMC}$ + $\textbf{BERT}_{GE-neg}$	86.07*	83.07*	76.94*	**91.93***	**82.57***
$\textbf{BERT}_{HMC}$ + $\textbf{BERT}_{SE-neg}$	**86.23***	**83.28***	**77.33***	91.91*	82.42*

6.4.2 Cross-HMC Task Transfer

As part of the study in this Chapter, I compare the performance of using an emotion fine-tuned model to a model fine-tuned on a specific HMC dataset and cross-transfer to another HMC dataset. For example, I fine-tune a HMC model using PHM2017 and further fine-tune it on a target dataset, FLU2013.

Results I present the results obtained in Table 6.5. Here, I denote the best results between $BERT_{GE}$ and $BERT_{SE}$ as $BERT_{emotion}$. In some cases ($FLU2013$ and $PHM2017$), the model, $BERT_{emotion}$ fine-tuned on a emotion dataset, leads to better results than models fined-tuned on another HMC dataset. On $SELF2020$, $ILL2021$ and $RHMD$, the performance of the $BERT_{emotion}$ is very close to the best-performing fine-tuned models on HMC datasets. These findings demonstrate that publicly available emotion datasets can be used to enhance performance on HMC tasks in the case where manually annotated HMC datasets are scarce.

Table 6.5: F1 macro score for the health mention classification task. **Bold** denotes the highest score. The average of five random seeds is used for all scores.

Model	FLU2013	PHM2017	SELF2020	ILL2021	RHMD
$\mathbf{BERT}_{emotion}$	**85.15**	**81.29**	74.02	91.38	81.91
$\mathbf{BERT}_{FLU2013}$	-	79.53	73.34	91.18	**82.22**
$\mathbf{BERT}_{PHM2017}$	84.98	-	**75.86**	**91.82**	81.73
$\mathbf{BERT}_{SELF2020}$	83.77	78.15	-	91.57	81.57
$\mathbf{BERT}_{ILL2021}$	84.52	77.99	74.16	-	82.0
$\mathbf{BERT}_{RHMD}$	84.09	80.55	73.88	91.69	-

6.5 Discussion

In this chapter, I showed that, as per the initial hypothesis, health mentions discussion contains emotional content, which can be exploited to improve health mention classification tasks. I proposed to incorporate emotions into HMC in two ways: (1) by implicitly adding affective features through intermediate fine-tuning on emotion detection task; and (2) by explicitly combining affective and HMC-specific features from both emotion and HMC models. Overall, I found that both approaches increased performance on the target task, with the *explicit* addition of affective features offering the highest gains (multi-feature fusion). The benefits cut across all HMC datasets, demonstrating the generalisation and robustness of the proposed approach. As such, this approach is to answer **RQ3**. Detecting health mentions on social media has the potential to improve public health surveillance systems for

ensuring healthy lives and promoting well-being.

I also investigated if there is any relationship between negative emotions and health mentions. The results show that there is no significant effect on the performance of HMC when only considering negative emotions for learning an emotion model. We further show that transferring emotion models to HMC tasks offers competitive performance to cross-HMC-task transfer. It suggests that in the absence of annotated data for HMC tasks, data-rich emotion tasks can be used to improve results. To aid reproducibility of the work in this Chapter, I release the implementation of the proposed approaches here - `https://github.com/tahirlanre/Emotion_PHM`.

In the upcoming chapter, an alternative approach for HMC will be investigated. This approach involves distinguishing between literal and non-literal meanings of disease keywords, with the aim of improving the detection of health mentions on social media. Additionally, I will describe a new health mention dataset generated from a context that is very different from those of the traditional benchmarks used in HMC domain.

Chapter 7

Improving Health Mention Classification Through Emphasising Literal Meanings: A Study Towards Diversity and Generalisation for Public Health Surveillance

Here, I examine data from *Nairaland.com*, an online forum widely used by Nigerians to discuss a variety of topics, including healthcare. Using this data, I create a health mention dataset to study the generalisability of health-related data from other social media sources and locations, addressing **RQ3**. *How can the gap in health-related social media data between developed and developing countries can be narrowed*? This dataset is a key component of the overall goal in this thesis, which is to analyse social media data for social good and extract actionable insights that can lead to positive impacts on the health and well-being of individuals and communities.

Furthermore, I propose a multi-task framework to improve the detection of health mentions by emphasising the literal meanings of disease words, contributing to **RQ3**. *How can health mentions be detected on social media to track health-related conversations*? The proposed framework and analysis of the data from Naira-land.com will provide valuable insights into the generalisation of health-related data from different sources and locations, and how this can be used to improve public health surveillance in developing countries. The work in this chapter has been been accepted for publication in The Web Conference 2023 [7].

7.1 Introduction

Chapters 5 and 6 have demonstrated the potential of harnessing social media data for public health surveillance. However, the data used in these studies were primarily from social media sites such as Twitter and Reddit, which are more popular in developed countries. In this chapter, I expand on existing work on HMC by examining data from developing countries, which often bear a disproportionate burden during public health emergencies. This study is aligned the thesis's primary objective of utilising social media analysis to address societal challenges. By incorporating data from areas that are frequently overlooked, this research enhances the understanding of global health trends and contributes to the broader mission of employing of inclusive and equitable well-being.

Public health emergencies have become a significant global concern, due to their detrimental impact on economic growth, stability, and the overall quality of life of the population. This threat to public health is especially pronounced in low and middle-income countries where, for instance, half of human mortality in Africa is attributed to infectious diseases[73]. These health crises exacerbate already high unemployment rates, thereby adversely affecting economic productivity.

Mitigating the effects of public health emergencies necessitates the collection, analysis, and interpretation of health-related data for surveillance purposes [227].

Given the widespread use of social media, these platforms provide a wealth of real-time data at a low cost, making them suitable for digital public health surveillance [273]. Users often discuss personal experiences concerning various health-related topics, from prescription drugs to symptoms and disease experiences. Aggregated and analysed, these data can provide population-level insights that can contribute significantly towards achieving the third Sustainable Development Goal, ensuring good health and well-being [16].

Previous works on harnessing social media data for public surveillance have covered various applications, including detecting and monitoring outbreaks [275], adherence to public health guidelines [148], and tracking health and well-being during global pandemics [9]. Social media data offer a real-time source of information that can serve as an early detection system for disease outbreaks, often revealing patterns overlooked by traditional health surveillance techniques, such as the analysis of hospital clinical records, laboratory reports, or surveys [136]. To utilise social media data effectively for public health surveillance, identifying content related to health reports, a task known as HMC, is critical [33].

However, most research on public health surveillance, including HMC, has primarily focused on social media platforms prevalent in developed nations [6], leaving other online data sources popular within underrepresented communities, particularly in low- and middle-income countries, largely unexplored [197]. This oversight potentially leads to data bias and undermines the goals of diversity and generalisation for public health surveillance. In this chapter, I address this bias by focusing on creating a dataset used predominantly by people from underrepresented communities, thus addressing the gap in the availability and quality of health-related data between developed and developing countries.

The dataset is constructed from Nairaland, a dedicated online forum for Nigerians, and covers health conditions such as HIV/AIDS, malaria, stroke and tuberculosis, which account for 27% of the disease burden from communicable diseases in Nigeria [250]. This dataset, referred to as Nairaland Health Mention Dataset (NHMD),

aims to provide better diversity coverage of vulnerable populations and generalisation for HMC tasks in a global public health surveillance setting.

The approaches used for health mention classification range from simple techniques like WESPAD - Word Embedding Space Partitioning and Distortion [118] - to deep learning approaches such as Long Short-Term Memory Networks (LSTM) [108] and bidirectional LSTM [33, 172]. More recently, pre-trained contextual language models such as BERT [61] and ELMo [192] have been employed for health mention detection.

In this chapter, I propose to explore the detection of literal use of disease words as an auxiliary task in a multi-task setting to improve HMC primary task performance. Previous literature has demonstrated that combining linguistic phenomena such as figurative or literal usage of a disease word can enhance the performance on HMC tasks [105, 172]. However, these approaches mostly focus on extracting these linguistic phenomena as features and using them in combination with task-specific features, or the tasks have been trained independently. In contrast, my approach involves jointly learning the literal usage of a disease word. This literal usage detection task predicts whether a disease word in a given post is used literally or not, following the method presented in [153].

7.2 Nairaland Health Mention Dataset

In this section, I present the health mention dataset from the largest Nigerian online community, Nairaland: NHMD. I detail the data collection and filtering, annotation procedures, and present an analysis of the dataset.

7.2.1 Data Collection and Filtering

There are no publicly accessible health mention datasets for underserved populations at the moment. Thus, developing such a dataset is crucial for the equality

and diversity of the health mention research community. I selected Nairaland, since it is the most popular online community used by Nigerians [21]. The web forum is the most visited indigenous site in Nigeria and the ninth most visited site in the country*. A detailed description about Nairaland and the data collection process has been provided in section 3.2.3.

I selected forums where health-related topics are likely to be discussed, i.e. Health and Politics forums. I retrieved all the posts in these forums from March 2005, when Nairaland was created, until April 2022. In total, I collected 20,995,525 posts from both forums.

The posts were filtered to include only relevant disease keywords such as HIV, AIDS, tuberculosis, malaria, and stroke.. I apply length filtering to only include posts between 3 to 120 tokens long (length matching with existing HMC datasets). I further sample randomly from the remaining posts for annotation, while maintaining the distribution across diseases. This resulted in 7,763 posts - an acceptable number, slightly more than the datasets introduced by Karisani and Agichtein [118].

To preserve users' privacy, all usernames or references to names were replaced with the <USER> token. I also removed any website links, emails or phone numbers from all posts.

7.2.2 Data Annotation

I hired two annotators to label the dataset. The annotators are Nigerian undergraduate students fluent in English and their local language, with one studying a health-related course. They also are proficient in Nigerian Pidgin, an English-based creole language spoken across Nigeria (see section 7.2.3). Additionally, these annotators are well-versed in Nigerian culture and humour, which is vital for understanding contexts.

*https://techcabal.com/2021/10/11/the-next-wave-wrestling-us-cyber-dominance/

Table 7.1: Example of annotations and corresponding label descriptions

Sample post	Disease	Label	Description
i am HIV + and to tell u there's no need to commit suicide or what have u.all u have to do is get committed to taking your drugs religiously ,eat well and stay healthy.	HIV/AIDS	Health mention	The post contains a health mention using a disease term. The author of the post or someone has a certain disease, or has corresponding symptoms
If you are that knowledgeable about Tuberculosis..You should know that being infected with the bacteria is not the same thing as being a Tuberculosis patient.	Tuberculosis	Other mention	The post contains the disease term but does not mention a specific person health. Discuss disease or symptom or discuss prevention of disease or symptoms in general.
OP's English fit give pesin Malaria sef.	Malaria	Non health mention	The post contains the disease terms used metaphorically, departing from the literal meaning, not aligning with commonsense, mock usage, or sarcastic expression

I adopt the annotation guidelines in [33] and define 3 classes: *health mention*, *other mention* and *non health mention*. See Table 7.1 for examples in each class along with their respective annotation descriptions. Each post can only be annotated with a single label based on annotation agreements. I asked the authors to skip any instance they were unsure about. The dataset is annotated in two steps: the preparation step and the production step. In the preparation step, 100 posts were annotated to establish guidelines for quality control and training. The annotators were then instructed to annotate the same batch of 100 posts. Both achieved at least 70% agreement with the annotations from the preparation batch. For the 30% dataset with non-agreement reached, I manually went through the examples with the annotations and discussed the mislabelled instances to ensure they understood the label categories fully. I emphasised the significance of basing assessments solely on the details expressly contained in a given post and avoiding any further assumptions. In the production step, I sent the whole dataset to the annotators with the first annotated examples.

I consider only posts that both annotators have labelled. For instances where both annotators disagreed, I first consider the level of disagreement between annotators. For example, suppose one of the annotators selects *non health mention*, and the other annotator selects *health mention*. In that case, I assume this instance is

Table 7.2: Inter-Annotator Agreement across diseases

Disease	**Kappa (κ)**
HIV/AIDS	0.6033
Malaria	0.6517
Stroke	0.7806
Tuberculosis	0.6866

difficult, and I discard the post. For cases with a smaller annotation difference, I forward these to a third annotator (a Nigerian with a Bachelor's degree qualification) to label and determine the final label, based on the majority vote. In the event there is no majority, I remove the post.

I measure the inter-annotator agreement using Cohen's kappa [51]. The average Cohen's kappa across the entire dataset is $\kappa = 0.67$. According to [133], the score indicates a strong agreement between the annotators. We also calculate the Kappa score (κ) per disease, to verify the agreement across the diseases (see Table 7.2). As can be seen, the agreement is consistent across diseases, with stroke-related posts having the highest agreement. This suggests that no disease-specific posts are more difficult to annotate than others.

7.2.3 Dataset Analysis

In this section, I conduct an extensive analysis of the proposed dataset on the following aspects: data statistics and language distribution.

Dataset Statistics

Table 7.3 shows the statistics of the dataset. I observe that the majority of the posts (64%) are labelled as *other mention*. This is the overall trend across diseases, except for stroke, where posts labelled as *non health mention* are the majority. The label distribution is similar to a popular public HMC dataset created by Briddle et al. [33].

Table 7.3: Dataset Statistics

Disease	Health mention	Other mention	Non health mention	Total
HIV/AIDS	221	2,855	1,061	4,137
Malaria	820	1,742	298	2,860
Stroke	90	288	295	673
Tuberculosis	17	54	17	93
Total	1,148	4,939	1,676	7,763

In terms of coverage of diseases, posts related to HIV/AIDS and Malaria are the majority, with 54% (4137/7763) and 37% (2860/7763) of the posts, respectively. Tuberculosis-related posts are the least represented, at only 1% (93/7763). Posts related to stroke account for 8% (673/7763) of the posts. The uneven distribution of posts across the diseases shows the focus of the discussion on Nairaland on HIV/AIDS over other diseases considered in this research.

Language Distribution

Nigeria is a multilingual society, and English is the common language adopted as the official language to enhance communication. However, the contact of indigenous languages with the English language has led to the development of Nigerian Pidgin [239]. Nigerian Pidgin is spoken widely across Nigeria, and it has been suggested that it makes communication easier on the Nairaland forum [242]. To determine the proportion of the datasets that contain Nigerian Pidgin, I use Franc*, a Language Identification Tool trained on 403 languages, including Nigerian Pidgin. Franc has shown superior performance on the Nigerian Pidgin dataset [5]. Of the 7,763 posts, 1,527 (20%) are in Nigerian Pidgin, while 6,233 posts (80%) are in English†. Although Franc can detect other major Nigerian languages, I observe that none of the posts was identified as any of the widely spoken languages: Hausa, Yoruba or Igbo. I suspect this is because the forums are meant for a general audience, and the posters use languages that are widely understood in Nigeria, such as English and Pidgin English.

*https://github.com/wooorm/franc

†The tool could not determine the language of 3 posts.

Table 7.4: Train, validation and test splits per class

Label	Train	Validation	Test
Health mention	923	112	113
Other mention	3,950	491	496
Non health mention	1,335	173	168

Dataset Split

I randomly split the dataset into training (80%), validation (10%) and test (10%) sets to promote reproducibility and facilitate comparisons between existing HMC models. The training set was used to train the proposed model, while the validation set was used to choose hyperparameters, and the test set was used to evaluate the performance of the models. I make the dataset split publicly available * and the breakdown of the splits is provided in Table 7.4.

7.3 Experiments

In this section, I detailed the experiments, including baseline models, proposed methods, evaluation metrics, hyperparameter search, results and discussions.

7.3.1 Baseline Models

Several machine learning models have been applied to the task of HMC [118, 105], and current state-of-the-art models for HMC tasks are based on Pre-trained Language Model (PLM) [122, 173]. I consider the following PLMs as our baseline models: BERT, ROBERTa and ALBERT.

- **Bag of words**: I experiment with traditional Bag of words (BOW) and used Support Vector Machine (SVM) as a classifier [109]. This non-neural baseline model will be compared with more advanced models to illustrate the effectiveness of the applied techniques in the context of the HMC task.

*Data available at `https://github.com/tahirlanre/nairaland_hmc`

- **BERT**: Bidirectional Encoder Representations from Transformers [61] is a language model pre-trained on unlabelled English texts Transformers [247]. The pre-training objective of BERT focuses on learning contextualised representations of words that can be useful for downstream applications. BERT has achieved exceptional performance across many natural language understanding tasks [61, 9].

- **RoBERTa**: Robustly optimised BERT approach (RoBERTa) [147] is a descendent of BERT introduced with modified pre-training objectives to create a more robust model. RoBERTa outperformed BERT on several NLP benchmark tasks [147].

- **ALBERT**: A Lite BERT (ALBERT) [132] was proposed to reduce the size of parameters and lower memory consumption. ALBERT has 12 million parameters compared to BERT, which has 110 million parameters. This is well-suited for low-resource settings where computing memory is limited.

Given the relatively small proportion of posts in Nigerian Pidgin (20% of the total), I only considered English-based pre-trained language models. Research has demonstrated that English fine-tuned language models tend to outperform those fine-tuned on other Nigerian languages for downstream tasks in Nigerian Pidgin due to the inherent lexical and structural resemblances between Nigerian Pidgin and English [126, 137].

7.3.2 Datasets

I use the three publicly available HMC datasets for our experiments:

- **PHM2017**: this dataset is a collection of English tweets related to Alzheimer's disease, heart attack, Parkinson's disease, cancer, depression, and stroke [118]. The dataset contains 4,987 instances labelled with either *personal health mention*, *awareness*, *other mention* and *non health mention*.

- **HMC2019**: This dataset is an extension of the PHM2017 dataset. The creators [33] of the dataset added tweets related to four additional health condition: cough, fever, headache and migraine. The dataset contains 14,051 posts labelled as *health mention*, *other mention* and *figurative mention*.
- **RHMD**: Unlike the other datasets, this dataset is a Reddit-based data that covers 15 disease or symptom terms [172]. Generally, the posts in this dataset are longer than the Twitter-based datsets. The dataset consists of 10,015 posts annotated as either *figurative mention*, *non-pesornal health mention* or *health mention*.
- **NHMD**: The proposed dataset in this paper with a detailed description in section 7.2.

7.3.3 Label Mapping

There are slight differences in the labels used for these datasets. Thus, I map the original labels to three classes: *health mention*, *other mention* and *non health mention* to create a uniform label distribution using the annotation descriptions provided in each dataset.

7.3.4 Compared Method: Fine-Tuning PLMs

Before introducing the novel multi-task learning approach, I start with the standard fine-tuning approach with acPLM, shown as the single-task learning method in Figure 7.1. In this section, I describe the fine-tuning method, using BERT as an example; all other models presented in section 7.3.1 can be interchangeably adopted. From this point, I represent data in its vectorised form and ignore the number suffix, for clarity.

For the fine-tuning approach, given data $(\boldsymbol{x}, \boldsymbol{y})$, I first pass $\boldsymbol{x}$ through the PLM, here, a BERT model, and retrieve its contextual representation $\boldsymbol{h}$:

Table 7.5: Main Results Between Baselines and the Proposed Framework. P - Precision, R - Recall and F_1 - Macro F_1

	NHMD			PHM2017			HMC2019			RHMD		
Model	**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1
Single-task Learning Models (baseline)												
BOW - SVM	75.58	61.34	65.25	75.13	65.41	68.65	78.92	75.67	76.71	70.67	70.52	70.16
BERT	80.56	76.29	78.08	86.51	84.12	85.23	89.47	88.91	89.12	80.47	80.31	80.32
RoBERTa	82.94	80.55	81.25	83.00	86.02	84.33	90.00	89.35	89.58	81.27	80.84	80.87
ALBERT	81.7	78.44	79.86	84.41	84.24	84.29	88.26	87.49	87.84	78.32	78.38	78.23
Multi-task Learning Models (proposed)												
BERT-MTL	81.75 ↑	78.62 ↑	79.98 ↑	86.53 ↑	84.19 ↑	85.28 ↑	89.65 ↑	89.17 ↑	89.35 ↑	80.69 ↑	80.34 ↑	80.43 ↑
RoBERTa-MTL	83.08 ↑	81.08 ↑	81.91 ↑	85.14 ↑	86.51 ↑	85.77 ↑	90.46 ↑	89.95 ↑	90.16 ↑	81.60 ↑	81.10 ↑	81.18 ↑
ALBERT-MTL	82.04 ↑	79.65 ↑	80.74 ↑	85.67 ↑	84.83 ↑	85.16 ↑	88.65 ↑	87.77 ↑	88.18 ↑	78.44 ↑	78.50 ↑	78.39 ↑

$$\boldsymbol{h} = BERT(\boldsymbol{x}) \tag{7.1}$$

Then I directly map the contextual representation $\boldsymbol{h}$ to its label $\boldsymbol{y}$ through an affine transformation:

$$\hat{\boldsymbol{y}} = Softmax(\boldsymbol{W_1} * \boldsymbol{h} + \boldsymbol{b_1}) \tag{7.2}$$

Finally, I calculate the cross-entropy loss between the prediction label $\hat{\boldsymbol{y}}$ and the ground truth label $\boldsymbol{y}$. I denoted this loss from fine-tuning the model as the HMC loss:

$$L_{ft} = L_{HMC} \tag{7.3}$$

Although the method of fine-tuning PLM is very intuitive and simple, it is the state-of-the-art method for HMC tasks and is considered as a very strong baseline method to compare with.

7.3.5 Proposed Method: Literal Emphasised Multi-task Learning

In this section, I present the proposed novel multi-task learning framework for HMC task, as shown in Figure 7.1. I propose to explicitly model the *literal* usage of a disease word in the text context as an auxiliary task.

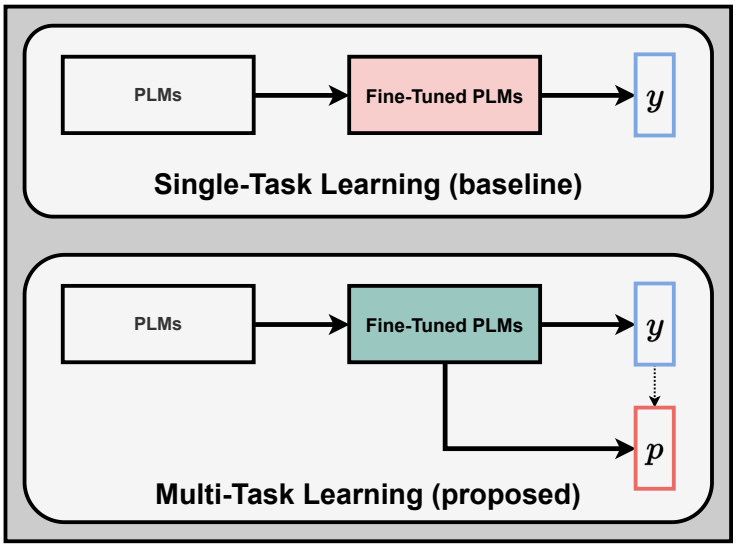


Figure 7.1: Multitask learning framework to emphasise literal meanings as an auxiliary task (demonstrated as the red block) for personal health mention classification tasks.

Pseudo-Literal Label Generation

Since the literal label is unknown, I create a pseudo literal label $\boldsymbol{p}$ (as demonstrated with a dashed arrow in Figure 7.1) for a given text pair $(\boldsymbol{x}, \boldsymbol{y})$ from its existing label $\boldsymbol{y}$, based on the following rule: if a sentence is originally labelled as either *health-mention* or *other mention* in $\boldsymbol{y}$, then I assume the disease word is labelled as literally in $\boldsymbol{p}$. Otherwise, if the original label $\boldsymbol{y}$ is *non health*, I assume the use of the disease word is labelled as non-literal in $\boldsymbol{p}$. Assigning the pseudo-literal label essentially uses the same amount of data in its original form, without introducing any additional human labeling process.. With the pseudo-literal label induced, I convert the original dataset from $(\boldsymbol{x}, \boldsymbol{y})$ to the following form $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{p})$.

Literal Emphasised Multi-task Learning

In this section, I describe the multi-task learning method, same as in section 7.3.4, using BERT as an example; all other models presented in section 7.3.1 can be interchangeably adopted.

For the multi-task learning approach, given data $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{p})$, I first pass $\boldsymbol{x}$ through the PLM - here a BERT model - and retrieve its contextual representation $\boldsymbol{h}$:

$$\boldsymbol{h} = BERT(\boldsymbol{x}) \tag{7.4}$$

For diseases or symptoms that are multi-word expressions such as heart attack, I take the average contextual representation. Then I directly map the contextual representation $\boldsymbol{h}$ to its label $\boldsymbol{y}$ through an affine transformation and to its pseudo label $\boldsymbol{p}$ through a complex non-linear transformation:

$$\begin{aligned} \hat{\boldsymbol{y}} &= Softmax(\boldsymbol{W_1} * \boldsymbol{h} + \boldsymbol{b_1}) \\ \hat{\boldsymbol{p}} &= \sigma(\boldsymbol{W}_3(tanh(\boldsymbol{W}_2 * \boldsymbol{h} + \boldsymbol{b}_2) + \boldsymbol{b}_3) \end{aligned} \tag{7.5}$$

Finally, I calculate the cross-entropy loss between the prediction label $\hat{\boldsymbol{y}}$ and the ground truth label $\boldsymbol{y}$, I denoted this loss from multi-task learning model as the HMC loss, , L_{HMC}. I calculate the cross-entropy loss between the prediction label $\hat{\boldsymbol{p}}$ and the ground truth label $\boldsymbol{p}$, I denoted this loss from multi-task learning model as the Literal loss, $L_{literal}$:

$$L_{mtl} = L_{HMC} + \lambda * L_{literal} \tag{7.6}$$

Where λ is a tunable weight hyperparameter that controls the importance placed on the auxiliary task.

7.3.6 Evaluation Metrics

I evaluated the performance of the models on the dataset using precision, recall and F1-score following previous work on HMC [118, 33], and with all performance reported on the test set. To account for variability, I perform five independent runs using different seeds for each model and report the average results over five runs.

7.3.7 Hyperparameter Selection

I select the best hyperparameters based on the average validation F1-score across 5 seeds. The range of hyperparameters is summarised as follows: batch size ε {16, 32}, learning rate ε {1e-5, 2e-5, 3e-5, 5e-5}, loss weight parameter λ ε [0.0, 1.0] for multi-task experiments. The optimal value for λ was 0.8. All models were trained for 5 epochs using the Adam optimiser. I use a dropout of 0.2 for all model.

7.3.8 Results and Discussion

Table 7.5 presents the results of the single-task and multi-task models. The table shows precision, recall and F1 scores for the test set. In terms of the single-task models, the Transformer based models performed substantially above the non-neural SVM with BOW representations. RoBERTa is superior to the other PLMs on all datasets except PHM2017 where BERT achieved the best performance. ALBERT (lite version of BERT), a smaller model, achieved significantly better results than BERT on NHMD. This is a promising result, particularly in low-income countries like Nigeria, with limited access to powerful computing resources.

In general, the multi-task models, where I jointly model the literal usage of disease words with the HMC task, consistently outperform their corresponding single-task models across all datasets in terms of precision, recall and F1 scores. The improvements are generally statistically significant based on the Wilcoxon test ($p < 0.01$) over five runs with random seeds. I speculate that this is because the model learns to identify the context in which the disease word is used to determine whether a text is a health mention. The RoBERTa-based multi-task model, RoBERTa-MTL achieves the best performance across all datasets.

For the proposed method, the NHMD dataset had the highest gains, with the improvement ranging from 0.7% - 1.9% on the F1 score. I suggest that the literal meaning of the disease words, e.g. HIV/AIDs and malaria, used to collect NHMD contains more information that is beneficial to the HMC task. I also note that

these words are less likely to be used in figurative contexts when compared to other disease or symptoms words, such as headache, and depression, used in the other datasets. This phenomenon can be justified from an information theory perspective: if one event is less likely to happen, it generates more information when it happens. Hence, modelling the health mention dataset (NHMD) when the disease keywords are less likely to be used in figurative contexts results in most information gain and better improvements in performance. On the remaining datasets, the performance improvements of the multi-task models over the single-task models are also substantial. For instance, on the PHM2017 dataset, RoBERTa-MTL and Albert-MTL improve on their corresponding single-task models by at least 0.9% on the F1 score. Overall, the results demonstrate the feasibility and generalisation of the proposed approach model if a disease word is used literally or not.

7.4 Further Analysis

7.4.1 Domain Shift and Generalisation

The domain of existing HMC datasets varies in terms of where they are extracted from (e.g. Twitter and Reddit), the disease or health condition they focus on (e.g. cancer, heart attack, HIV/AIDS) and their target population (e.g. mainly based on text from developed countries). The distinction in the data domain imposes high selection bias and potentially leads to domain shift. The domain shift in data harms the generalisation of the models when tested on an out-of-distribution dataset in a text classification setting [252]. For public health classification tasks, it is critical for the systems to react to unseen diseases from other domains [134]. To address this challenge, previous research has proposed to use domain adaptation to leverage datasets from related domains [103]. Domain adaptation is particularly useful in public health research, where the availability of labelled data is limited, as a result of the cost or expert annotators and sudden-onset of a public health emergency, such as the global COVID-19 pandemic. Nevertheless, the generalisation performance

of models is expected to drop under domain shift due to underlying distributional shifts [232]. Recent work by Harrigian et al. [92] showed that mental health models generalise poorly across multiple social media platforms. To this end, I evaluate the domain generalisation for current HMC datasets and discuss whether the created dataset, NHMD alleviates this issue.

7.4.2 Analysis Setting

For the domain adaptation, I explore the following settings:

Single-Source (In-Domain) -> Single-Target (In-Domain)

In this setting, I perform the standard fine-tuning, as described in section 7.3.4, using single-source data and report results based on its corresponding in-domain single target dataset (i.e. I report results on the test split of PHM2017 when it is trained on the training split of PHM2017). The results for this experiment are denoted as 'S(I) -> S(I)' in Table 7.6.

Multiple-Source (In-Domain) - Single-Target (In-Domain):

In this setting, I again perform the fine-tuning, as described in section 7.3.4, using multiple-source data and report results for each individual target dataset (i.e. I report results on the test split of PHM2017 when it is trained on a combination of training split of PHM2017, HMC2019, RHMD and NHMD). The results for this experiment are denoted as 'MI) -> S(I)' in Table 7.6.

Single-Source (In-Domain) - Single-Target (Out-Domain):

In this setting, I perform out-of-domain experiments by training a model on a single HMC dataset (source), e.g. PHM2017 and test on another HMC dataset (target), e.g. NHMD. For this experiment, the aim is to understand and quantify the effect

Table 7.6: Macro F1 score for the domain adaptation experiments.

	PHM2017	**HMC2019**	**RHMD**	**NHMD**
In Domain Generalisation				
S(I) -> S(I)	85.28	90.69	80.87	81.25
M(I) -> S(I)	84.79 ↓	89.66 ↓	84.13 ↑	79.05 ↓
Out Domain Generalisation				
S(I) -> S(O)				
PHM2017	-	76.42	80.31	67.51
HMC2019	80.77	-	68.66	58.93
RHMD	74.04	80.59	-	58.34
NHMD	76.32	72.00	67.33	-
Average	77.04	76.34	72.10	61.59
M(I) -> S(O)	79.8 ↑	77.98 ↑	69.92 ↓	63.2 ↑

of out-domain generalisation on HMC tasks with unseen examples. The results for this experiment are denoted as 'S(In) ->S(O)' in Table 7.6.

Multiple Source (In Domain) - Single Target (Out Domain):

In this setting, I perform another set of out-of-domain experiments by training a model on a single HMC dataset (source), e.g. PHM2017 and test on another HMC dataset (target), e.g. NHMD. For this experiment, the aim is to understand and quantify the effect of out-domain generalisation on HMC tasks with unseen examples. The results for this experiment are denoted as 'S(In) ->S(O)' in Table 7.6.

7.4.3 Analysis Results and Discussion

Table 7.6 shows the results of the domain adaptation experiments using the RoBERTa model, which is the overall best performing architecture based on the results presented in Table 7.5. I report the average F1 score from five independent runs using different seeds.

In the first part of the table, I examined the in-domain generalisation of the datasets: from *Single-Source (In-Domain) -> Single-Target (In-Domain)* and *Multiple-Source (In-Domain) -> Single-Target (In-Domain)* experiment, in most cases, the in-domain generalisation performance drops statistically significantly ($p < 0.01$ based on the Wilcoxon test), with the exception of the RHMD dataset.

In the second part of the table, I examine the out-domain generalisation of the datasets: for *Single-Source (In-Domain) -> Single-Target (Out-Domain)*; I present results with source dataset used on the left column with respect to its out-domain test datasets. Additionally, I report the mean average of the out-domain performance as an indication on the average generalisation performance. For *Multiple-Source (In-Domain) -> Single-Target (Out-Domain)*, I combine all the datasets as training and test their out-domain performance on each single target dataset. The results suggest that, in most cases, the out-domain generalisation performance improves statistically significantly ($p < 0.01$ based on the Wilcoxon test) with the exception of the RHMD dataset. I also observe that models trained with Twitter-based datasets (PHM2017 & HMC2019) transfer to the Reddit-based dataset (RHMD) better than models trained on the proposed dataset (NHMD) in the *Single-Source (In-Domain) -> Single-Target (Out-Domain)* setting. The negative transfer to NHMD from other datasets is notably higher, with a 14 - 23% decrease in the F1 score. The results are similar in the reverse direction, except for PHM2017, where the transfer from NHMD performs better than RHMD. These results demonstrate the importance of our dataset, which aims to mitigate the data selection bias in HMC tasks.

In summary, I can confidently claim that the proposed dataset, NHMD, imposes a better diversity coverage of vulnerable populations and generalisation for HMC tasks in a global public health surveillance setting.

Table 7.7: LIWC feature correlations across classes for all datasets, sorted by Pearson correlation (r).

PHM2017						**HMC2019**					
Health mention		*Other mention*		*Non health mention*		*Health mention*		*Other mention*		*Non health mention*	
LIWC category	r	LIWC category	r	LIWC category	r	LIWC category	r	LIWC category	r	LIWC category	r
health	0.092	pronoun	0.426	health	0.247	Clout	0.218	pronoun	0.389	prep	0.180
Lifestyle	0.072	det	0.299	emo_neg	0.153	Culture	0.092	verb	0.333	adj	0.101
Clout	0.051	focuspast	0.292	Clout	0.143	Lifestyle	0.092	auxverb	0.282	Clout	0.100
sexual	0.050	verb	0.261	prep	0.135	curiosity	0.063	Authentic	0.278	Drives	0.084
curiosity	0.042	Authentic	0.213	tone_neg	0.134	attention	0.058	det	0.263	tone_pos	0.077
Culture	0.037	adverb	0.193	curiosity	0.113	sexual	0.051	focuspast	0.205	curiosity	0.074
attention	0.034	auxverb	0.192	Drives	0.098	socrefs	0.038	adverb	0.188	death	0.061
prep	0.017	socrefs	0.138	cogproc	0.067	socbehav	0.03	focuspresent	0.176	Lifestyle	0.051
tone_pos	0.016	swear	0.117	socbehav	0.056	health	0.026	acquire	0.124	time	0.049
want	0.010	conj	0.116	death	0.054	tone_pos	0.015	negate	0.120	tone_neg	0.047
adj	0.009	focuspresent	0.110	attention	0.034	cogproc	0.008	swear	0.108	feeling	0.046
RHMD						**NHMD**					
Health mention		*Other mention*		*Non health mention*		*Health mention*		*Other mention*		*Non health mention*	
LIWC category	r	LIWC category	r	LIWC category	r	LIWC category	r	LIWC category	r	LIWC category	r
Clout	0.287	pronoun	0.306	Authentic	0.144	sexual	0.190	pronoun	0.151	auxverb	0.162
health	0.192	verb	0.205	negate	0.120	Clout	0.148	focuspast	0.119	cogproc	0.116
Culture	0.079	auxverb	0.159	conj	0.108	Culture	0.081	time	0.091	focuspast	0.110
Lifestyle	0.064	Authentic	0.149	adj	0.099	cogproc	0.049	feeling	0.082	focuspresent	0.110
socrefs	0.055	focuspresent	0.134	prep	0.097	det	0.046	socbehav	0.081	verb	0.101
curiosity	0.043	focuspast	0.131	tone_pos	0.091	prep	0.044	swear	0.071	prep	0.098
sexual	0.042	det	0.13	cogproc	0.085	health	0.043	socrefs	0.067	health	0.096
attention	0.028	adverb	0.119	curiosity	0.066	swear	0.039	tone_pos	0.066	conj	0.071
prep	0.019	conj	0.085	feeling	0.065	negate	0.038	acquire	0.063	quantity	0.070
death	0.004	feeling	0.065	Drives	0.063	socrefs	0.038	Drives	0.063	adj	0.064
socbehav	0.004	socrefs	0.063	lack	0.059	death	0.034	Authentic	0.062	Authentic	0.060

7.4.4 Linguistic Analysis

Understanding the underlying language variations can highlight the differences between the datasets. Thus, I conducted a further analysis, by comparing the topic distribution of the collected posts in the proposed datasets, NHMD, with three other popular public HMC datasets (PHM2017, HMC2019 and RHMD), based on the LIWC package [191]*. I report the Pearson correlation of the top 10 topics for each label in Table 7.7. A LIWC feature value measures the proportion of words used across posts in a specific label matching a given LIWC dimension. The version of LIWC (LIWC-22) I used covers over 100 language dimensions. All correlations associated with datasets labelled as *other-mention* or *non-health* were found to be statiscally significant ($p - value < 0.05$). In addition, the top 5 topics within datasets labelled as *health-mention* also showed statistical significance ($p - value < 0.05$).

I note some similarities in the topics prevalent across all the HMC datasets. For example, *Health, e.g. illness* related topics are present in *health mention* posts for all datasets. This is unsurprising, as I expect the latter to cover health discussions.

*https://www.liwc.app/

Word use related to other physical and health dimensions, *e.g. sexual*, are also prevalent, but they associate more with *health mention* posts in NHMD. However, I also note some differences between the proposed dataset, NHMD, and the remaining datasets. For instance, topics related to *Lifestyle (e.g. home, work, money)*, *Perception (e.g. attention)* and *Motives (e.g. curiosity)* are present in PHM2017, HMC2019 and RHMD. However, in NHMD, I note more use of words related to *Negations, e.g. not, nothing*, *Determiners (i.e. det), e.g. articles and numbers* and association with *Quantities (e.g. all, one, more)*.

7.5 Discussion

In this Chapter, I have constructed and released NHMD: a new benchmark dataset for underrepresented communities, extracted based on four prevalent diseases (HIV/AIDS, Malaria, Stroke and Tuberculosis) in Nigeria. The novel manually annotated dataset was collected from a dedicated web forum for Nigerians, effectively addressing **RQ4**. Extensive analysis on its transferability and generalisation capacity suggests that the dataset contributes to the domain generalisation of the HMC task.

Furthermore, I propose a novel multi-task learning approach combining HMC with literal keyword use identification. Thus, addressing **RQ3**. The experimental results demonstrate that the approach outperforms the state-of-the-art baseline approaches. Implications include the potential to improve HMC with *literal* identification as an auxiliary task; and also highlight the importance of introducing and using a dataset from the wider community, especially underrepresented groups, to ensure fairness, robustness and generalisation for public health surveillance.

Future work can consider the concatenation of normalised counts of linguistic features from LIWC to the BERT representations. This could further enrich the information harnessed from the data and enhance the performance of the model. Moreover, techniques such as data augmentation [72] or data resampling [257],

which can address the high-class imbalance observed in the data, may also be explored to enhance model robustness.

To aid reproducibility of the work in this Chapter, I release the data and the implementation of the proposed models here - `https://github.com/tahirlanre/nairaland_hmc`.

In the next Chapter, I will explore the impact of public health crises such as the COVID-19 pandemic on religious and spiritual activities in the UK.

Chapter 8

Religion and Spirituality on social media in the Aftermath of the Global Pandemic

In Chapter 5, I have examined the emotional toll of the COVID-19 pandemic on people's health and wellbeing. The pandemic has also caused an unprecedented shift in the way we live, work, and interact with one another. From the closure of places of worship to the increase in online religious and spiritual practices, the COVID-19 pandemic has affected every aspect of our lives.

In this chapter, I examine the impact of the pandemic on religious and spiritual practices in the UK, effectively addressing **RQ5**. *How has the COVID-19 pandemic impacted religious and spiritual practices in the UK, as reflected on social media*? Using state-of-the-art NLP techniques, this chapter explores how the pandemic has affected these practices, which are considered social determinants of health.

8.1 Introduction

Religion is considered a social determinant of health [100]. While poorly understood, it is suggested that religion can affect health at the individual level through

health practices and social ties, as well as providing systems of meaning and feelings of strength to cope with stress and adversity [23, 207]. There is evidence to suggest a positive association between religious belief and happiness [104]. Religion is also recognised as an aspect of cultural competence in healthcare and can provide health guidance and social support [17, 31].

Moreover, research has found that religious beliefs and practices, specifically religious service attendance, can promote better physical and mental health and lower mortality rates, including lower rates of depression, myocardial infarction, and cardiovascular death [124, 224]. Religious coping mechanisms, such as a sense of control and relaxation, may also contribute to better health outcomes [262]. Additionally, religious involvement through social connections, specifically service attendance, can provide avenues of development and support, promoting better health and wellbeing [47].

Before discussing the research findings, it is essential to provide definitions of the terms religiosity and spirituality used in this chapter. Religiosity refers to an individual's devotion to religion, characterised by a belief in God and commitment to follow established principles [3, 200, 174]. Spirituality, on the other hand, is defined by the the World Psychiatric Association as an awareness of something beyond ordinary observation and perception [245].

This Chapter focuses on tracking and quantifying the shift of religious and spiritual practices from offline to online mode by leveraging social media data. By doing this, it echoes the broader themes of this thesis, which emphasises leveraging social media data for social good, especially in healthcare contexts. This alignment illustrates the multifaceted applications of social media analysis, showcasing its potential to provide valuable insights across a diverse spectrum of human activities and societal needs. Specifically, I investigate six religion-related activities: *choir, corporate worship, meditation prayer, reflecting on nature* and *yoga*. These activities are widely recognised as significant by experts in Theology in the UK. Due to the lockdowns, large-scale ethnographic methodologies such as in-person interviews

and focus groups were not feasible. Additionally, since the study examines the move of religious activity online, Twitter, a major social media platform, is chosen as the primary data source. Social media analysis provides a valuable resource for studying engagement with religion-related activities at the population level. With the application of machine learning models, language expressed on social media can be used to carry out sociolinguistic analysis such as analysing polarisation between atheists and theists [12]. Twitter also allow users to share geographical information in a tweet, which is useful in this case to collect tweets posted by users in the UK.

8.2 Materials and methods

8.2.1 Data collection and preprocessing

Twitter

I gathered English tweets that were geo-located in the UK between July and September of 2020. This time period is particularly significant since it corresponds to the early days of the pandemic, after the implementation of various regulations, but before the availability of vaccines. In addition, I also collected tweets from the pre-pandemic period during the same months of the year (July-September 2019). The Twitter API for Academic Research * was used to collect the tweets, providing access to the entire archive of tweets published on Twitter. The total number of collected tweets was $20,927,967$. The breakdown of tweets for each period is provided in Table 8.1.

Reddit

Given the challenge of collecting appropriate tweets related to religious or spiritual activities, I turned to Reddit to extract relevant tweets. Reddit is a popular

*https://developer.twitter.com/en/use-cases/do-research/academic-research

Table 8.1: Statistics about the UK tweets.

	Year	
Month	**2019**	**2020**
July	4,078,800	3,834,890
August	4,053,235	3,659,652
September	4,029,085	3,658,281
Total	**12,161,120**	**11,152,823**

social media platform with a wide range of discussion-based communities known as subreddits. These communities are focused on specific topics such as sports, mental health, and many others. Discussions in a subreddit typically start with a post made by a user, followed by comments from other users. As of January 2021, Reddit has over 50 million daily active users and more than 100,000 active communities, primarily in the UK, US, and Canada*. To select the appropriate subreddit, I look for the one that contains the most posts related to the specific religious activity. Although many subreddits focus on the topics of interest, I choose the one with the largest number of posts for the religious-related activities I am interested in.

I collected all submissions from the start of 2011 through the end of 2020 for the subreddits of interest. This approach allowed me to obtain pre-pandemic as well as pandemic-related information from July to September 2020. To extract the posts, I used the Pushshift API † to access the Pushshift Reddit dataset published by [27].

8.2.2 Extracting tweets related to religious and spiritual activities

To infer the labels of posts based on their subreddit communities, I first identify those that have a clear relationship to religious or spiritual activity. For instance, I assign a label to a post based on its appearance in a relevant subreddit, such as

*https://www.redditinc.com/press
†https://github.com/pushshift/api

Table 8.2: Top 3 tweets based on cosine similarity to respective subreddits.

Subreddit	Tweet	cos score
r/Meditation	do you want to meditate better? :) if so, then these carefully selected meditation quotes from ⟨user⟩ should help. and be sure to read the intro story... it's both insightful and entertaining! #spirituality #mindfulness #meditation.	0.8668
	⟨user⟩ .... daily meditation is a life changer. been meditating for over 2 years now and there is so many benefits. if you want to have a quick read about my thoughts on this... (4/5 minute read)	0.8332
	This a good read. acknowledgement that sometimes it is hardest to meditate when you would most benefit from it because there are times your mind just won't settle in to it! #mindfulness #meditate #thursday-thoughts	0.8243
r/PrayerRequests	can we pray for you? just a reminder that prayer is the driving force behind everything that we do!we would love the chance to pray for you, so please feel free to message your prayer requests via direct message and as a church we will stand with you in prayer!	0.8056
	⟨user⟩ i'll pray for you sis if you need prayer 24/7 then it doesn't matter! you ask away sister	0.7997
	please could you pray for me as i'm going through some persecution at home. i'm the only christian in my family	0.7951
r/yoga	fully endorse this. been doing yoga on and off for 35 years, daily (injury permitting) with ⟨user⟩ for about 8	0.8461
	anyone for yoga?	0.8358
	after years of searching i think i've finally found the right yoga for me	0.8217

r/yoga[*]. I limited my post collection to only those subreddits that satisfied two main criteria:

- They are focused on a specific religious activity, such as *r/Meditation* (religious meditation). This first criterion establishes a clear link between the subreddit and the religious activity, enabling us to implicitly annotate the Reddit posts according to the subreddits in which they appeared.
- They appear to be the largest, most general subreddits dedicated to that religious activity. This second criterion allows us to focus on the general concepts related to a religious activity.

By applying these criteria, I extracted posts from *r/Meditation*, *r/PrayerRequests* and *r/yoga* to represent meditation, prayer and yoga activities respectively. Finding appropriate subreddits for the remaining activities that matched the criteria proved challenging due to the nature of those activities.

To extract relevant text from a large-scale unsupervised corpus, I used an approach that was previously used in research [63]. This method involves extracting activity-related tweets from the unlabelled tweets corpus by embedding all tweets and posts from the relevant subreddit (e.g., *r/Meditation*) in a shared vector space. The embedding process involves using a robust sentence encoder to convert each tweet into a sentence embedding. I used MPNet [228], a pre-trained sentence embedding model that is designed to produce similar representations for sentences with similar meanings. To extract tweets related to each activity, I constructed embeddings that represent each activity using the same MPNet model. I then used these embeddings as queries to extract the most similar tweets based on cosine similarity. To create the query embeddings, I calculated the average sentence embeddings of all posts in the subreddit related to an activity. Since there may be underlying distribution shifts from Reddit to Twitter [202], I selected the top 100 tweets for each activity

[*] `https://www.reddit.com/r/yoga/`

initially to test the relevance of the results. This step is essential to ensure that the retrieved tweets are relevant to an activity.

Examples of the top 3 tweets for each subreddit are shown in Table 8.2. I then use these tweets to query the unlabelled corpus to retrieve more relevant tweets. I use a threshold based on the cosine similarity score to extract the most similar tweets. The threshold (meditation = 0.61, prayer = 0.61, yoga = 0.55) set for each activity was determined after a manual inspection of the results.The threshold selected are subjective and they are based on how relevant the tweets were considered to be (by myself) for the respective activities. Yoga is considered here, arguably, less important for religious activities, when compared to meditation and prayer. Examples of tweets reflecting the specific activities are shown in Table 8.3. In future work, by adopting zero-shot or few-shot learning strategies, larger models might prove beneficial for detecting relevant tweets.

8.2.3 Measuring change in religious and spiritual activities

This study uses language models to explore how language expressions related to religious or spiritual activities have changed during different periods, specifically before and during the COVID-19 pandemic. By training language models on tweets from specific periods, I can measure changes in the language use patterns in Twitter conversations and gain insights into changes in religious and spiritual activities. The language models trained on each period can be considered as representations of their corresponding conversations, providing a way to understand the evolution of these activities over time.

To train the language models, I follow the details of GPT-2 [205]. The training objective of GPT-2 is to predict the next word, given all of the previous words within a given text (see Section 3.3.2 for more details). The language models are trained with all the collected tweets from each month.

I estimate shifts in people's religious activities by comparing the likelihood of

Table 8.3: Example of tweets filtered based on cosine similarity to the top-k tweets. Arrows indicate whether cos score is higher (up) or lower (down) than the threshold.

Activity	Tweet	cos score
Meditation	⟨user⟩ dear doc .. i am huge fan of your podcast , specially mindfulness. i have a question. being indian, yoga/pranayam is an integral part if my life. however when i do meditation i have observed that i feel angrey and irritated whole day. this puts me off	0.7725 ↑
	find a quiet spot either in your garden, balcony, local green space or even by a window and join our meditation in nature session via zoom on tuesday 11 august, 9 - 9:45am for more information and to register, please email ⟨user⟩ #natureconnection ⟨url⟩	0.6585 ↑
	⟨user⟩ what in the heck is going on with his sword	-0.0684 ↓
Prayer	⟨user⟩ ⟨user⟩ wishing you strength to carry on!	0.7005 ↑
	⟨user⟩ ⟨user⟩ may Allah bless you with good health and happiness.⟨user⟩	0.6134 ↑
	⟨user⟩ no!! because apparently christmas is on hold!!! how rude! xx	0.1391 ↓
Yoga	both classes are on as usual on bank holiday monday! have a brilliant long weekend and see you on the mat on monday! #yoga #mensnakedyogalondon #naturist ⟨url⟩	0.5845 ↑
	definitely need to do some yoga tomorrow to ease my back and neck pain	0.7539 ↑
	congratulations to everyone receiving their a level results today! there's lots of useful advice here: ⟨url⟩ ⟨url⟩	0.1056 ↓

phrases indicating performing a religious activity between language models trained on tweets from the pre-COVID-19 period and those trained on tweets during COVID-19. To evaluate the language models, I curate a set of phrases that reflect people performing a specific religious activity. The phrases are based on questionnaires developed by experts in Theology * to explore the spiritual life and religious activity of the UK. For example, the phrase *"I am doing yoga"* corresponds to the yoga question item from the questionnaire. I distinguish between performing an activity offline and online by appending *"online"* or *"via [Zoom/Microsoft Teams/Google Meet]* to the original phrase. For example, "I am doing yoga via Zoom" will represent performing yoga online.

Given the diverse nature of natural language use on social media platforms like Twitter, it is crucial to account for the many ways users may express identical sentiments, or describe similar activities. Therefore, recognising paraphrases is integral to this analysis, ensuring a comprehensive and precise representation of religious or spiritual activities. By employing a paraphrase generation model, this methodology retrieves paraphrases for each activity phrase, significantly expanding the range of detectable expressions. For example, 'I am reflecting on nature' can be rephrased as 'I reflect on nature'. Thus, by incorporating paraphrases, this analysis provides a more robust and inclusive methodology for estimating shifts in religious activities. The complete list of phrases are provided in the appendix A.

I adapt the approach described in [25] to measure the difference in religion-related engagement between two corresponding months in different years (i.e. July 2019 vs July 2020, August 2019 vs August 2020 and September 2019 vs September 2020). I measure how likely a language model generate a phrase using token perplexity. Token perplexity is the inverse log joint probability of the test set, normalised by the number of word tokens in the test set, as assigned by the language model [112]. A lower perplexity score implies more confidence in predicting a sequence of words. I then perform a significance test using a Student's two-tailed test with the mean

*https://www.dur.ac.uk/resources/digitaltheology/PressReleasereOnlineChurch.pdf

perplexity differences of all phrases related to an activity from the same month before COVID-19 (2019) and during COVID-19 (2020). I report the change in activity engagement as the t-value of the test. A negative t-value indicates that an activity is discussed (or performed) less than the previous year, while a positive t-value suggests that an activity is discussed (or performed) more than the previous year. The change is statistically significant if the corresponding p-value < 0.05. The mean perplexity, $\bar{x}_t$ of a set of activity phrases for a period, t is defined as follows:

$$\bar{x}_t = \frac{1}{n} \sum_{i=1}^{n} PP(s_i) \tag{8.1}$$

where PP is the perplexity, and s_i is the activity phrase.

8.2.4 Analysis of tweets related to religious and spiritual activities

I use the extracted tweets from section 8.2.2 to understand how religious activities have changed during the pandemic. For brevity, I consider tweets from July - September 2019 as the pre-COVID-19 period and tweets from July - September 2020 as the COVID-19 period. As a first approach, I compare the frequency of activity-related tweets from the pre-COVID-19 period to the COVID-19 period. I performed a paired T-test to determine if the change is statistically significant. I reject the null hypothesis if $p < 0.05$. In addition, I measure the effect size using Cohen's d to determine the difference between the number of tweets from respective periods. $d = 0.2, 0.5, 0.8$ are considered as a small, medium, and large effect sizes, respectively [52].

As a second approach, I employed the log odds ratio with informed Dirichlet priors [166, 114] to extract the lexical correlates of tweets relevant to religion-related activities between two periods: before COVID-19 and during COVID-19. This method has been used in several analyses of linguistic differences in social media

texts [120, 74]. Other techniques such as Pointwise Mutual Information (PMI) and TF-IDF have been used for similar tasks; however, the log odds ratio has been shown to outperform these methods [166, 114]. I use a word cloud to visualise the most significant tokens from different periods. Tokens that appear less than 10 times are excluded. I aggregate all the pre-COVID-19 tweets and COVID-19 tweets, creating two corpora for each activity. I then extract all tokens from a period and calculate the log odds ratio by contrasting them to all tokens from another period. Log odd ratios are estimated using Z-score. A higher score indicates that the token is more significant within a corpus than the contrasting corpus.

8.3 Results and Discussion

8.3.1 Shift in religion-related engagements

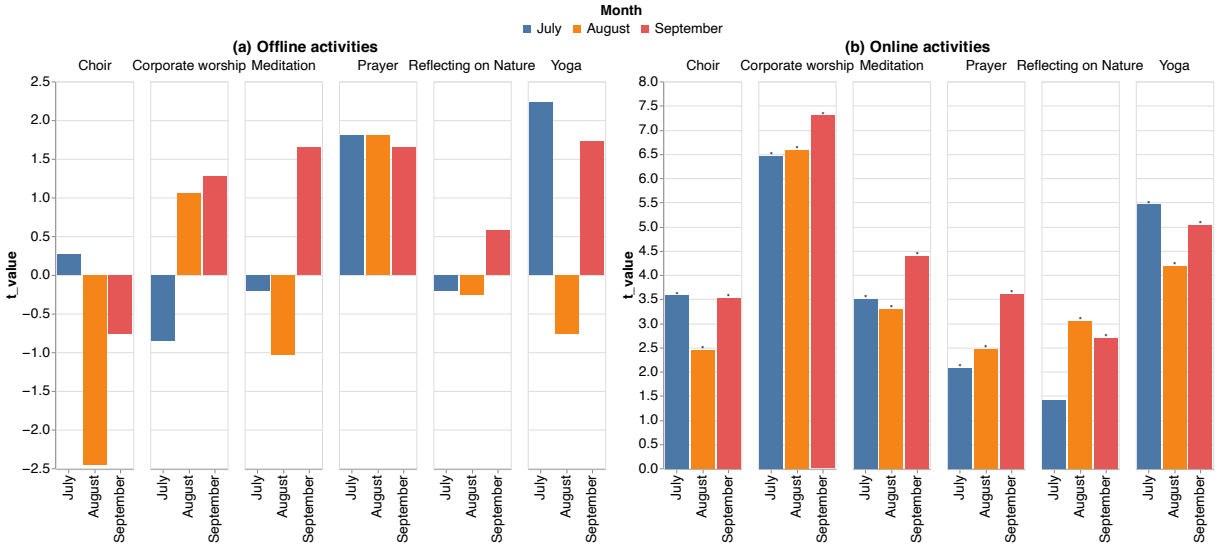


Figure 8.1: Online (left) and offline (right) - Twitter

Fig 8.1 summarises the change effect of engagement with religion-related activities, both offline and online. The shift in engagements varies for offline and online activities. For offline activities (Fig 8.1a), engagement appears to increase (i.e. t-value is positive) from pre-COVID-19 to during COVID-19, indicating that there is more engagement. Prayer, yoga and corporate worship appear to follow a similar

trend bar one month where there is a negative effect (i.e. t-value is negative) on the engagement from pre-COVID-19 and during COVID-19. In terms of Choir and reflecting on nature, most of the change effects across the months are negative, indicating lesser engagement with these activities when compared with the pre-COVID-19 period. The negative effect is most likely due to restrictions by the UK government to prevent the spread of COVID-19, while some of the positive effects might be due to some relaxation of the rules around that period. The change effect for all these activities is not significant ($p < 0.05$).

Fig 8.1 summarises the change effect of engagement with religion-related activities, both offline and online. The shift in engagements varies for offline and online activities. For offline activities (Fig 8.1a), religion-related engagement appears to increase (i.e. t-value is positive) from pre-COVID-19 to during COVID-19, indicating more participation. Prayer, yoga and corporate worship seem to follow a similar trend bar one month where there is a negative effect (i.e. t-value is negative). For Choir and reflecting on nature, most change effects across the months are negative, indicating lesser engagement with these activities compared with the pre-COVID-19 period. This result is similar to the one obtained from the survey results. The change effect for all these activities is insignificant ($p < 0.05$).

For online activities (Fig 8.1b), the trends presented in the results signal the shift in engagement with religious activities online by the increasing usage of online words (such as Zoom, Youtube, and virtual) within the context of religious activity discussions. All of these changes are significant ($p < 0.05$) except for one month (July) for reflecting on nature. The most significant increase is in corporate worship, where the change effective for all the explored months is the highest.

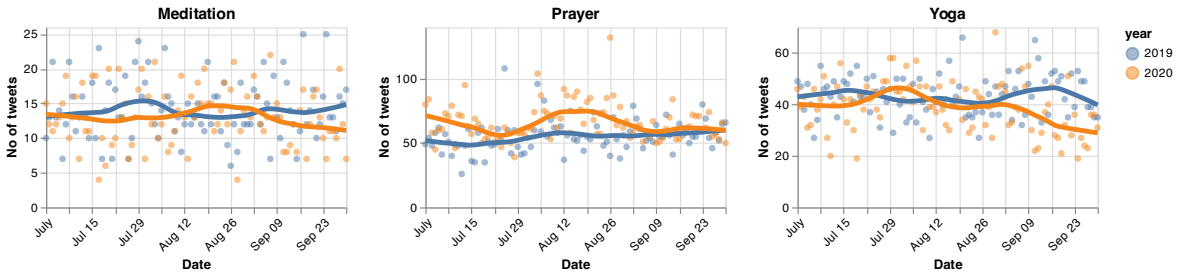


Figure 8.2: Daily activity related tweets over July 1 - September 30 for years 2019 and 2020

8.3.2 Comparisons between tweets from pre-COVID-19 and COVID-19 periods

Fig 8.2 shows the frequency of tweets that are related to a specific activity for pre-COVID-19 and COVID-19 periods. On average, there is an increase in the number of prayer related tweets (*Cohen's d*, p-value < 0.05) from pre-COVID-19 period to COVID-19 period. In contrast, the frequency of tweets related to meditation (*Cohen's* $d = 0.25$, p-value > 0.05) and yoga (*Cohen's* $d = 0.48$, p-value < 0.05) are generally lower during COVID-19 when compared to pre-COVID 19 period.

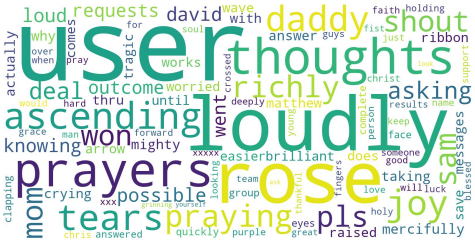
Figs 8.3, 8.4 and 8.5 show the top 100 most representative words for each period. For meditation-related tweets before COVID-19 (Fig 8.3a)), offline-related words such as retreat, centre, and park are present. Some words (e.g. buddhism, buddha) used in the tweets indicate relation to religion. For prayer-related tweets pre COVID-19 (Fig 8.4a)), some of the most common words (e.g. soul, praying, faith, christ) are related to religious practices.

The most important words in meditation-related tweets during COVID-19 are displayed in Fig 8.3b). The presence of terms such as online, zoom, recording, virtual, and youtube indicates that this activity is probably being done online. Similarly, for prayer-related tweets during COVID-19 (Fig 8.4b)), some of the most influential words are link, join, mixlr, which are related to practising online. As expected, there are also words associated with COVID-19 (e.g. covid, safe), which indicate discussion about the pandemic in prayer-related tweets. For yoga-related tweets

a) Pre COVID-19 b) During COVID-19

Figure 8.3: Most representative words of pre COVID-19 and during COVID-19 for meditation-related tweets

(Fig 8.5b) during the pandemic, online-related words (including zoom, online, live, link) are also common. Overall, there is a change in the language used for tweets about prayer and meditation, prayer and yoga during the pandemic to words associated with online engagement (such as online, zoom, and virtual). This is consistent with the results comparing the language models (section 8.3.1) from pre-COVID to during COVID-19. The presence of religion-related words shows the relevance of these activities to spirituality.



a) Pre COVID-19 b) During COVID-19

Figure 8.4: Most representative words of pre COVID-19 and during COVID-19 for prayer-related tweets

a) Pre COVID-19 b) During COVID-19

Figure 8.5: Most representative words of pre COVID-19 and during COVID-19 for yoga-related tweets

8.4 Discussion

In conclusion, this Chapter has explored the impact of the COVID-19 pandemic on religious activities through social media analysis to address **RQ5**. As people are unable to engage in traditional religious practices due to the pandemic, they have turned to online platforms to perform their religious activities. The analysis of Twitter data has revealed a significant increase in online religious activities, including prayer and yoga, during the pandemic. Moreover, the analysis also highlighted a surge in prayer-related tweets during this period. However, it is essential to recognise that Twitter's demographics are somewhat limited [161], and further studies are necessary to gain a comprehensive understanding of religious expression in the digital age. The code used for the analysis in this Chapter is available at `https://github.com/tahirlanre/covid19-online-religion`.

Chapter 9

Conclusion

Social media platforms have become an integral part of our daily lives, generating vast amounts of data that can be harnessed to promote social good. However, effectively analysing this data to address societal problems and deliver positive social impact is a challenging task. In this thesis, I have presented novel computational approaches for the responsible and ethical analysis of social media data that can promote inclusive and equitable education, healthy lives and well-being. This chapter summarises the main findings of the thesis and discusses the limitations of social media analysis for social good.

9.1 Main Findings

This thesis first contribution lies in the domain of education, more specifically in the area of **tracking the impact of an intervention programme using social media data** in Chapter 4 to respond to **RQ1**:

- I have used TechUPWomen, as a case study to measure the impact of technology retraining programme for women from underepresented groups using social media data.
- I have employed LDA to analyse the topics discussed during the programme on both public (Twitter) and private (Microsoft Teams) channels. This

showed insights into the topics that resonated with program participants.

- I used a BERT model to explore and compare the sentiment expressed during the retraining program across both channels.

Secondly, I have shifted from previous applications of social media analysis in education to health to answer **RQ2: detecting fine-grained emotions on social media during major disease outbreaks** in Chapter 5:

- I have proposed a new *deep transfer learning framework* that models emotional representations invariant to specific entities. As part of this framework, I developed *EmoBERT*, a model that incorporates emotion-specific knowledge into BERT.
- Through experimentation and comparison against several state-of-the-art approaches, I illustrated the importance of integrating emotional knowledge in pre-trained language models for predicting fine-grained emotions (i.e. detecting a specific emotion, e.g. 'sad', rather than its presence).
- I conducted the first study on how the COVID-19 pandemic has affected public emotions on Twitter users in London, United Kingdom, comparing emotions (*annoyed, anxious, empathetic* and *sad*) expressed in tweets from March 2020 with the same period in 2019.
- I also separately perform an analysis of the hashtags mostly used in tweets expressing the selected emotions.

Thirdly, I have developed an approach to **incorporate emotions into HMC task on social media, which is useful to support public health surveillance on social media** in Chapter 6 to answer **RQ3**:

- I proposed two approaches - *intermediate task fine-tuning* and *multi-feature fusion* to incorporate emotional information to effectively classify health mentions on social media.

- I investigated if there is any relationship between negative emotions and health mentions that leads to performance gains on HMC tasks.
- I compared cross-transfer between HMC tasks with limited annotation data to the transfer learning framework I proposed and observed competitive performance when data-rich emotion detection tasks are directly transferred to HMC tasks.

Furthermore, I have proposed a multi-task learning framework to model and distinguish the **literal usage of disease and symptom words from non-literal usage** to improve the performance of HMC in Chapter 7 to answer **RQ3**:

- I proposed a novel literal emphasised multi-task learning framework for the HMC task and achieve state-of-the-art performance across various HMC datasets.
- Leveraging the contextualised word representations of disease or symptoms words in different contexts, I modelled the differences between literal and non-literal usage.
- I evaluated the proposed multi-task framework on several HMC datasets and showed its effectiveness in detecting health mentions on social media.

Fourthly, I have made a significant step towards improving **diversity and generalisation for public health surveillance on social media** by creating the first health mention dataset from a web forum used in a developing country in Chapter 7 to answer **RQ4**:

- I constructed *Nairaland health mention dataset (NHMD)*, a new dataset collected from a dedicated web forum for Nigerians. NHMD consists of 7,763 manually labelled posts extracted based on four prevalent diseases (HIV/AIDS, Malaria, Stroke and Tuberculosis) in Nigeria.

- To the best of my knowledge, NHMD is the first health mention dataset for underseved populations that is publicly available - thus addressing the missing distribution of existing publicly available HMC datasets and mitigating the data bias problem.
- Using NHMD, I studied the generalisation ability of HMC models across existing HMC datasets, and investigate the language variations across the datasets.

Finally, I have examined the **influence of the COVID-19 pandemic on religious and spiritual practices, which are recognised as social determinant of health** in Chapter 8 to answer **RQ5**:

- I leveraged the power of pre-trained language models to track the shift of religious practices from before COVID-19 and during COVID-19.
- I present the first - to the best of my knowledge - study on the influence of the COVID-19 pandemic on religious activities in the UK.
- I provided quantitative evidence on the shift of religious and spiritual practices from offline to online mode.

9.2 Limitation and Future Works

A challenge to utilising social media data for research stems from the issue of sampling bias [53]. It is crucial to acknowledge that social media data may not always offer a comprehensive representation of the general population [53]. Researchers have discovered that racial, ethnic, and socioeconomic differences in the adoption of social media exist [91]. Moreover, challenges arise in the selection of content. Currently, most contents are chosen through filters such as keywords or geo-location, which can lead to potential relevant tweets or posts being overlooked.

For instance, in Chapter 5 and Chapter 8, geo-location is utilised to filter tweets in London and the UK, respectively. However, relevant tweets may be missed if the users do not include geo-location information in their tweets. Similarly, in Chapter 7, keywords are employed to select disease-related posts on Nairaland.com, but potentially relevant posts could be missed if they did not use the specific keywords I utilised for filtering in their post.

The scope of this thesis primarily focuses on specific social media platforms for analysis. However, it is imperative to recognise that these platforms differ in terms of content and users. Prior research has highlighted that models trained on one platform may perform poorly on another platform, despite conducting the same task [92]. Therefore, future research should explore domain adaptation techniques to improve the generalisability of models across diverse social media platforms.

An area where improvement can be made in this thesis is in fostering engagement with the community and clinical experts. It will be useful to engage domain experts, such as practising clinicians in the process of annotation and labelling to improve the quality of annotations. This could offer an external validation of the approaches employed, which can subsequently elevate performance in real-world situations.

Further limitations and avenues open to further research have been proposed throughout the thesis. These insights underscore the evolving nature of the field and provide a direction for future investigations.

Bibliography

[1] M. Abdul-Mageed and L. Ungar. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1067. URL `https://aclanthology.org/P17-1067`.

[2] R. Abebe, S. Giorgi, A. Tedijanto, A. Buffone, and H. A. A. Schwartz. Quantifying community characteristics of maternal mortality using social media. In *Proceedings of The Web Conference 2020*, pages 2976–2983, 2020.

[3] R. Abrori, A. Zulfatillah, and H. Bullah. Religion and employees fraud prevention: With moderation of spirituality, leadership and oganizational culture. *Journal of Auditing, Finance, and Forensic Accounting*, 10(2), 2022.

[4] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. Predicting flu trends using twitter data. In *2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS)*, pages 702–707. IEEE, 2011.

[5] I. Adebara, A. Elmadany, M. Abdul-Mageed, and A. A. Inciarte. Afrolid: A neural language identification tool for african languages. *arXiv preprint arXiv:2210.11744*, 2022.

[6] O. T. Aduragba and A. I. Cristea. Research on prediction of infectious diseases, their spread via social media and their link to education. In *Proceedings of the 2019 4th International Conference on Information and Education Innovations*, pages 38–42, 2019.

[7] O. T. Aduragba, J. Yu, A. Cristea, and Y. Long. Improving health mention classification through emphasising literal meanings: A study towards diversity and generalisation for public health surveillance. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30-May 4, 2023, Austin, TX, USA*. ACM, 2018. doi: 10.1145/3543507.3583877. URL `https://doi.org/10.1145/3543507.3583877`.

[8] O. T. Aduragba, J. Yu, A. I. Cristea, M. Hardey, and S. Black. Digital inclusion in nothern england: Training women from underrepresented communities in tech: A data analytics case study. In *2020 15th International Conference on Computer Science & Education (ICCSE)*, pages 162–168. IEEE, 2020.

[9] O. T. Aduragba, J. Yu, A. Cristea, and L. Shi. Detecting fine-grained emotions on social media during major disease outbreaks: Health and well-being before and during the covid-19 pandemic. American Medical Informatics Association, 2021.

[10] M. S. Akhtar, A. Kumar, A. Ekbal, C. Biemann, and P. Bhattacharyya. Language-agnostic model for aspect-based sentiment analysis. In *International Conference on Computational Semantics*, 2019.

[11] H. Al-Dmour, A. Salman, M. Abuhashesh, R. Al-Dmour, et al. Influence of social media platforms on public health protection against the covid-19 pandemic via the mediating effects of public health awareness and behavioral changes: integrated model. *Journal of medical Internet research*, 22(8): e19996, 2020.

[12] Y. Al Hariri, W. Magdy, and M. K. Wolters. Atheists versus theists: Religious polarisation in arab online communities. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–28, 2021.

[13] R. Albalawi, T. H. Yeap, and M. Benyoucef. Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, 3:42, 2020.

[14] E. Allaway, M. Srikanth, and K. McKeown. Adversarial learning for zero-shot stance detection on social media. In *North American Chapter of the Association for Computational Linguistics*, 2021.

[15] D. Allington, B. Duffy, S. Wessely, N. Dhavan, and J. Rubin. Health-protective behaviour, social media usage and conspiracy belief during the covid-19 public health emergency. *Psychological Medicine*, pages 1 – 7, 2020.

[16] Y. M. Asi and C. Williams. The role of digital health in making progress toward sustainable development goal (sdg) 3 in conflict-affected populations. *International journal of medical informatics*, 114:114–120, 2018.

[17] H. G. Atkinson, D. Fleenor, S. M. Lerner, E. Poliandro, and J. Truglio. Teaching third-year medical students to address patients' spiritual needs in the surgery/anesthesiology clerkship. *MedEdPORTAL*, 14:10784, 2018.

[18] E. J. Avery. Public information officers' social media monitoring during the zika virus crisis, a global health threat surrounded by public uncertainty. *Public Relations Review*, 43(3):468–476, 2017.

[19] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[20] P. Babvey, G. Gongora-Svartzman, C. Lipizzi, and J. E. Ramírez-Márquez. Content-based user classifier to uncover information exchange in disaster-motivated networks. *PLoS ONE*, 16, 2021.

[21] I. R. Badmus, S. A. Okaiyeto, and L. K. Mustapha. Agora for the diaspora: Exploring the use of nairaland online forum for political deliberations among nigerian emigrants. *The Nigerian Journal of Communication*, 16(1):191–210, 2019.

[22] K. Baecher, A. Boutwell, N. Gunawardhana, D. M. Tebit, and J. A. Dionne. 1938. covid-19 vaccine hesitancy among adults who rely on social media for health care information in cameroon, africa. *Open Forum Infectious Diseases*, 2022.

[23] A. M. Baer, L. I. Tovar, and R. A. Chaney. Considering on the bigger picture of public health: Student reflections on university mission. *Pedagogy in Health Promotion*, 6(2):142–147, 2020.

[24] T. Baldwin. Social media: friend or foe of natural language processing? In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 58–59, 2012.

[25] S. Barikeri, A. Lauscher, I. Vulić, and G. Glavaš. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.151. URL `https://aclanthology.org/2021.acl-long.151`.

[26] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63, 2019.

[27] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn. The

pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

[28] C. Baziotis, N. Pelekis, and C. Doulkeridis. DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2126. URL `https://aclanthology.org/S17-2126`.

[29] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[30] J. A. Benítez-Andrades, Á. González-Jiménez, Á. López-Brea, J. Aveleira-Mata, J.-M. Alija-Pérez, and M. T. García-Ordás. Detecting racism and xenophobia using deep learning models on twitter data: Cnn, lstm and bert. *PeerJ Computer Science*, 8, 2022.

[31] K. L. Bentley-Edwards, P. A. Robbins, L. T. Blackman Carr, I. Z. Smith, E. Conde, and W. A. Darity Jr. Denominational differences in obesity among black christian adults: Why gender and life stage matter. *Journal for the Scientific Study of Religion*, 60(3):498–515, 2021.

[32] A. Benton, G. Coppersmith, and M. Dredze. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 94–102, 2017.

[33] R. Biddle, A. Joshi, S. Liu, C. Paris, and G. Xu. Leveraging sentiment distributions to distinguish figurative from literal health reports on twitter. In Y. Huang, I. King, T. Liu, and M. van Steen, editors, *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 1217–1227. ACM / IW3C2, 2020. doi: 10.1145/3366423.3380198. URL `https://doi.org/10.1145/3366423.3380198`.

[34] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 601–608. MIT Press, 2001. URL `https://proceedings.neurips.cc/paper/2001/hash/296472c9542ad4d4788d543508116cbc-Abstract.html`.

[35] J. Blumenstock, G. Cadamuro, and R. On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

[36] E. Bondi-Kelly, L. Xu, D. Acosta-Navas, and J. Killian. Envisioning communities: A participatory approach towards ai for social good. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021.

[37] D. M. Boyd and N. B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, 13(1):210–230, 2007.

[38] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

[39] S. Bucci, M. Schwannauer, and N. Berry. The digital revolution and its impact on mental health care. *Psychology and psychotherapy*, 92 2:277–297, 2019.

[40] A. E. Cano Basave, Y. He, and R. Xu. Automatic labelling of topic models learned from twitter by summarisation. Association for Computational Linguistics (ACL), 2014.

[41] Y. Cao, H. Ajjan, and P. Hong. Using social media applications for educational outcomes in college teaching: A structural equation analysis. *British Journal of Educational Technology*, 44(4):581–593, 2013.

[42] C. Catalan Aguirre, N. Gonzalez Castro, C. Delgado Kloos, C. Alario-Hoyos, and P. J. Muñoz Merino. Conversational agent for supporting learners on a mooc on programming with java. 2021.

[43] T.-Y. Chang and C.-J. Lu. Rethinking why intermediate-task fine-tuning works. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 706–713, Punta Cana, Dominican Republic, nov 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.61. URL `https://aclanthology.org/2021.findings-emnlp.61`.

[44] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2005. URL `https://aclanthology.org/S19-2005`.

[45] D. D. Chaudhari and A. V. Pawar. A systematic comparison of machine learning and nlp techniques to unveil propaganda in social media. *J. Inf. Technol. Res.*, 15:1–14, 2022.

[46] 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*, 6, 2020.

[47] Y. Chen and T. J. VanderWeele. Associations of religious upbringing with subsequent health and well-being from adolescence to young adulthood: An outcome-wide analysis. *American journal of epidemiology*, 187(11):2355–2364, 2018.

[48] H. Cho and H. Lee. Biomedical named entity recognition using deep neural networks with contextual information. *BMC bioinformatics*, 20(1):1–11, 2019.

[49] A. G. Chowdhury, R. Sawhney, R. Shah, and D. Mahata. # youtoo? detection of personal recollections of sexual harassment on social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2527–2537, 2019.

[50] E. M. Clark. Applications in sentiment analysis and machine learning for identifying public health variables across social media. 2019.

[51] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[52] J. Cohen. Quantitative methods in psychology: A power primer. In *Psychological bulletin*. Citeseer, 1992.

[53] R. B. Correia, I. B. Wood, J. Bollen, and L. M. Rocha. Mining social media data for biomedical signals and health-related behavior. *Annual review of biomedical data science*, 3:433–458, 2020.

[54] N. S. Coulson. How do online patient support communities affect the experience of inflammatory bowel disease? an online survey. *JRSM short reports*, 4(8):2042533313478004, 2013.

[55] J. Cowls, A. Tsamados, M. Taddeo, and L. Floridi. A definition, benchmark and database of ai for social good initiatives. *Nature Machine Intelligence*, 3 (2):111–115, 2021.

[56] M. Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020. URL `https://arxiv.org/abs/2009.09796`.

[57] J. A. de Bruijn, H. de Moel, B. Jongman, M. C. de Ruiter, J. Wagemaker, and J. C. Aerts. A global database of historic and real-time flood events based on social media. *Scientific data*, 6(1):1–12, 2019.

[58] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.372. URL `https://aclanthology.org/2020.acl-main.372`.

[59] A. Derakhshan and H. Beigy. Sentiment analysis on stock social media for stock price movement prediction. *Eng. Appl. Artif. Intell.*, 85:569–578, 2019.

[60] S. Desai, C. Caragea, and J. J. Li. Detecting perceived emotions in hurricane disasters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5290–5305, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.471. URL `https://aclanthology.org/2020.acl-main.471`.

[61] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

[62] J. Du, Q. Chen, Y. Peng, Y. Xiang, C. Tao, and Z. Lu. Ml-net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association*, 26(11):1279–1285, 2019.

[63] J. Du, E. Grave, B. Gunel, V. Chaudhary, O. Celebi, M. Auli, V. Stoyanov, and A. Conneau. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.426. URL `https://aclanthology.org/2021.naacl-main.426`.

[64] J. Dyer and B. Kolic. Public risk perception and emotion on twitter during the covid-19 pandemic. *Applied Network Science*, 5(1):1–32, 2020.

[65] O. Edo-Osagie, B. De La Iglesia, I. Lake, and O. Edeghere. A scoping review of the use of twitter for public health research. *Computers in biology and medicine*, 122:103770, 2020.

[66] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6:169–200, 1992.

[67] F. Es-Sabery, K. Es-Sabery, J. Qadir, B. S. de Abajo, A. Hair, B. G.-Z. Soto, and I. de la Torre-Díez. A mapreduce opinion mining for covid-19-related tweets classification using enhanced id3 decision tree classifier. *IEEE Access*, 9:58706–58739, 2021.

[68] A. Ezen-Can, K. E. Boyer, S. Kellogg, and S. Booth. Unsupervised modeling for understanding mooc discussion forums: a learning analytics approach. In *Proceedings of the fifth international conference on learning analytics and knowledge*, pages 146–150, 2015.

[69] R. Fagnani, B. dos Santos Bueno, R. M. Itida, J. A. Galhardo, and R. L. Vanot. A novel approach in public health surveillance: searching the illegal dairy trade in facebook. *International journal of environmental health research*, pages 1–11, 2022.

[70] A. A. Farzindar and D. Inkpen. Natural language processing for social media. *Synthesis Lectures on Human Language Technologies*, 13(2):1–219, 2020.

[71] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1169. URL `https://aclanthology.org/D17-1169`.

[72] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, 2021.

[73] F. Fenollar and O. Mediannikov. Emerging infectious diseases in africa in the 21st century. *New Microbes and New Infections*, 26:S10–S18, 2018.

[74] A. Field and Y. Tsvetkov. Unsupervised discovery of implicit gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 596–608, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.44. URL `https://aclanthology.org/2020.emnlp-main.44`.

[75] R. M. Filho, J. M. Almeida, and G. L. Pappa. Twitter population sample bias and its impact on predictive outcomes: a case study on elections. pages 1254–1261, 2015.

[76] P. Fortuna, L. Pérez-Mayos, A. G. T. AbuRa'ed, J. Soler-Company, and L. Wanner. Cartography of natural language processing for social good (nlp4sg): Searching for definitions, statistics and white spots. *Proceedings of the 1st Workshop on NLP for Positive Impact*, 2021.

[77] C. Fuchs. Culture and economy in the age of social media. 2015.

[78] B. Gambäck and U. K. Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3013. URL `https://aclanthology.org/W17-3013`.

[79] J. Gao, P. Zheng, Y. Jia, H. Chen, Y. Mao, S. Chen, Y. Wang, H. Fu, and J. ming Dai. Mental health problems and social media exposure during covid-19 outbreak. *PLoS ONE*, 15, 2020.

[80] S. Ghosh and P. O. Kristensson. Neural networks for text correction and completion in keyboard decoding. *arXiv preprint arXiv:1709.06429*, 2017. URL `https://arxiv.org/abs/1709.06429`.

[81] S. Ghosh, P. Chakraborty, E. O. Nsoesie, E. Cohn, S. R. Mekaru, J. S. Brownstein, and N. Ramakrishnan. Temporal topic modeling to assess associations between news trends and infectious disease outbreaks. *Scientific reports*, 7 (1):40841, 2017.

[82] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas. Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69:214–224, 2017.

[83] Y. Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

[84] Y. Goldberg. Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1):1–309, 2017.

[85] S. D. Gollapalli, P. Rozenshtein, and S.-K. Ng. ESTeR: Combining word co-occurrences and word associations for unsupervised emotion detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1043–1056, Online, 2020. Association for Computational Linguistics.

doi: 10.18653/v1/2020.findings-emnlp.93. URL `https://aclanthology.org/2020.findings-emnlp.93`.

[86] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18 (5-6):602–610, 2005.

[87] T. Grover, E. Bayraktaroglu, G. Mark, and E. H. R. Rho. Moral and affective differences in us immigration policy debate on twitter. *Computer Supported Cooperative Work (CSCW)*, 28(3):317–355, 2019.

[88] O. Gruebner, S. R. Lowe, M. Sykora, K. Shankardass, S. Subramanian, and S. Galea. Spatio-temporal distribution of negative emotions in new york city after a natural disaster as seen in social media. *International journal of environmental research and public health*, 15(10):2275, 2018.

[89] S. C. Guntuku, G. Sherman, D. C. Stokes, A. K. Agarwal, E. Seltzer, R. M. Merchant, and L. H. Ungar. Tracking mental health and symptom mentions on twitter during covid-19. *Journal of general internal medicine*, 35(9):2798–2800, 2020.

[90] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

[91] E. Hargittai. Potential biases in big data: Omitted voices on social media. *Social Science Computer Review*, 38:10 – 24, 2018.

[92] K. Harrigian, C. Aguirre, and M. Dredze. Do models of mental health based on social media data generalize? In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3774–3788, 2020.

[93] J. Hartmann, M. Heitmann, C. Siebert, and C. Schamp. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 2022.

[94] S. A. Hayati and A. O. Muis. Analyzing incorporation of emotion in emoji prediction. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 91–99, Minneapolis, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1311. URL `https://aclanthology.org/W19-1311`.

[95] Y. He, Z. Zhu, Y. Zhang, Q. Chen, and J. Caverlee. Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4604–4614, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.372. URL `https://aclanthology.org/2020.emnlp-main.372`.

[96] L. Hossain, D. Kam, F. Kong, R. Wigand, and T. Bossomaier. Social media in ebola outbreak. *Epidemiology & Infection*, 144(10):2136–2143, 2016.

[97] N. Hu. Sentiment analysis of texts on public health emergencies based on social media data mining. *Computational and Mathematical Methods in Medicine*, 2022, 2022.

[98] X. Hu and H. Liu. Text analytics in social media. In *Mining text data*, pages 385–414. Springer, 2012.

[99] J. Huang, H. Zhao, and J. Zhang. Detecting flu transmission by social sensor in china. In *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*, pages 1242–1247. IEEE, 2013.

[100] E. L. Idler. *Religion as a social determinant of public health*. Oxford University Press, USA, 2014.

[101] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier. Practical extraction of disaster-relevant information from social media. *Proceedings of the 22nd International Conference on World Wide Web*, 2013.

[102] M. Imran, P. Mitra, and C. Castillo. Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1638–1643, Portorož, Slovenia, 2016. European Language Resources Association (ELRA). URL `https://aclanthology.org/L16-1259`.

[103] M. Imran, P. Mitra, and J. Srivastava. Cross-language domain adaptation for classifying crisis-related short messages. *arXiv preprint arXiv:1602.05388*, 2016.

[104] B. Ishaq, L. Østby, and A. Johannessen. Muslim religiosity and health outcomes: A cross-sectional study among muslims in norway. *SSM-Population Health*, 15:100843, 2021.

[105] A. Iyer, A. Joshi, S. Karimi, R. Sparks, and C. Paris. Figurative usage detection of symptom words to improve personal health mention detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1142–1147, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1108. URL `https://aclanthology.org/P19-1108`.

[106] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[107] D. Jiang, M. Hao, F. Ding, J. Fu, and M. Li. Mapping the transmission risk of zika virus using machine learning models. *Acta Tropica*, 185:391–399, 2018.

[108] K. Jiang, S. Feng, Q. Song, R. A. Calix, M. Gupta, and G. R. Bernard. Identifying tweets of personal health experience through word embedding and lstm neural network. *BMC bioinformatics*, 19(8):67–74, 2018.

[109] T. Joachims. A statistical learning learning model of text classification for support vector machines. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 128–136, 2001.

[110] S. E. Jordan, S. E. Hovet, I. C.-H. Fung, H. Liang, K.-W. Fu, and Z. T. H. Tse. Using twitter for public health surveillance from monitoring and prediction to public response. *Data*, 4(1):6, 2018.

[111] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016. URL `https://arxiv.org/abs/1602.02410`.

[112] D. Jurafsky and J. H. Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

[113] D. Jurafsky and J. H. Martin. *Speech and language processing. Vol. 3.* 2014.

[114] D. Jurafsky, V. Chahuneau, B. R. Routledge, and N. A. Smith. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, 2014.

[115] L. Kaati, A. Shrestha, K. Cohen, and S. Lindquist. Automatic detection of xenophobic narratives: A case study on swedish alternative media. In *2016 IEEE conference on intelligence and security informatics (ISI)*, pages 121–126. IEEE, 2016.

[116] R. K. Kaliyar, A. Goswami, and P. Narang. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia Tools and Applications*, 80:11765 – 11788, 2021.

[117] A. Kannan, K. Kurach, S. Ravi, T. Kaufmann, A. Tomkins, B. Miklos, G. Corrado, L. Lukács, M. Ganea, P. Young, and V. Ramavajjala. Smart reply: Automated response suggestion for email. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, and R. Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 955–964. ACM, 2016. doi: 10.1145/2939672.2939801. URL `https://doi.org/10.1145/2939672.2939801`.

[118] P. Karisani and E. Agichtein. Did you really just have a heart attack?: Towards robust detection of personal health mentions in social media. In P. Champin, F. L. Gandon, M. Lalmas, and P. G. Ipeirotis, editors, *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 137–146. ACM, 2018. doi: 10.1145/3178876.3186055. URL `https://doi.org/10.1145/3178876.3186055`.

[119] P. Karisani, N. Karisani, and L. Xiong. Contextual multi-view query learning for short text classification in user-generated data. *arXiv preprint arXiv:2112.02611*, 2021. URL `https://arxiv.org/abs/2112.02611`.

[120] K. Kawintiranon and L. Singh. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.376. URL `https://aclanthology.org/2021.naacl-main.376`.

[121] P. Ke, H. Ji, S. Liu, X. Zhu, and M. Huang. SentiLARE: Sentiment-aware language representation learning with linguistic knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6975–6988, Online, 2020. Association for

Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.567. URL https://aclanthology.org/2020.emnlp-main.567.

[122] P. I. Khan, I. Razzak, A. Dengel, and S. Ahmed. Performance comparison of transformer-based models on twitter health mention classification. *IEEE Transactions on Computational Social Systems*, 2022.

[123] H. Khanpour and C. Caragea. Fine-grained emotion detection in health-related online posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1166, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1147. URL https://aclanthology.org/D18-1147.

[124] D. E. King and W. S. Pearson. Religious attendance and continuity of care. *The International Journal of Psychiatry in Medicine*, 33(4):377–389, 2003.

[125] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[126] A. Kniele and M. Beloucif. Uppsala University at SemEval-2023 task12: Zero-shot sentiment classification for Nigerian Pidgin tweets. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1491–1497, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.205. URL https://aclanthology.org/2023.semeval-1.205.

[127] A. Kothari, L. Foisey, L. Donelle, and M. A. Bauer. How do canadian public health agencies respond to the covid-19 emergency using social media: a protocol for a case study using content and sentiment analysis. *BMJ Open*, 11, 2021.

[128] P. Kralj Novak, J. Smailović, B. Sluban, and I. Mozetič. Sentiment of emojis. *PloS one*, 10(12):e0144296, 2015.

[129] V.-P. La, T.-H. Pham, M.-T. Ho, M.-H. Nguyen, K.-L. P. Nguyen, T.-T. Vuong, H. K. T. Nguyen, T. Tran, Q. V. Khuc, M.-T. Ho, and Q.-H. Vuong. Policy response, social media and science journalism for the sustainability of the public health system amid the covid-19 outbreak: The vietnam lessons. *Sustainability*, 2020.

[130] A. Lamb, M. J. Paul, and M. Dredze. Separating fact from fear: Tracking flu infections on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795, Atlanta, Georgia, 2013. Association for Computational Linguistics. URL `https://aclanthology.org/N13-1097`.

[131] V. Lampos, B. Zou, and I. J. Cox. Enhancing feature selection using word embeddings: The case of flu surveillance. In R. Barrett, R. Cummings, E. Agichtein, and E. Gabrilovich, editors, *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 695–704. ACM, 2017. doi: 10.1145/3038912.3052622. URL `https://doi.org/10.1145/3038912.3052622`.

[132] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[133] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[134] E. Laparra, S. Bethard, and T. A. Miller. Rethinking domain adaptation for machine learning over clinical language. *JAMIA open*, 3(2):146–150, 2020.

[135] B. X. Lee, F. Kjaerulf, S. Turner, L. Cohen, P. D. Donnelly, R. Muggah, R. Davis, A. Realini, B. Kieselbach, L. S. MacGregor, et al. Transforming

our world: implementing the 2030 agenda through sustainable development goal indicators. *Journal of public health policy*, 37(1):13–31, 2016.

[136] K. Lee, A. Agrawal, and A. Choudhary. Forecasting influenza levels using real-time social media streams. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 409–414. IEEE, 2017.

[137] H. Lent, E. Bugliarello, M. de Lhoneux, C. Qiu, and A. Søgaard. On language models for creoles. *arXiv preprint arXiv:2109.06074*, 2021.

[138] R. Lerrigo, J. T. Coffey, J. L. Kravitz, P. Jadhav, A. Nikfarjam, N. H. Shah, D. Jurafsky, and S. R. Sinha. The emotional toll of inflammatory bowel disease: Using machine learning to analyze online community forum discourse. *Crohn's & Colitis 360*, 2019.

[139] C. Li, L. J. Chen, X. Chen, M. Zhang, C. P. Pang, and H. Chen. Retrospective analysis of the possibility of predicting the covid-19 outbreak from internet searches and social media data, china, 2020. *Eurosurveillance*, 25, 2020.

[140] L. Li, Q. Zhang, X. Wang, J. J. Zhang, T. Wang, T.-L. Gao, W. Duan, K. K. fai Tsoi, and F. yue Wang. Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo. *IEEE Transactions on Computational Social Systems*, 7:556–562, 2020.

[141] X. Li, L. Bing, W. Zhang, and W. Lam. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5505. URL `https://aclanthology.org/D19-5505`.

[142] Y. Liao, X. Jiang, and Q. Liu. Probabilistically masked language model capable of autoregressive generation in arbitrary word order. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 263–274, Online, 2020. Association for Computational Linguist-

ics. doi: 10.18653/v1/2020.acl-main.24. URL `https://aclanthology.org/2020.acl-main.24`.

[143] L. T. S. Lim, Z. J. G. Regencia, J. D. Cruz, F. D. V. Ho, M. S. Rodolfo, J. T. Ly-Uson, and E. S. Baja. Assessing the effect of the covid-19 pandemic, shift to online learning, and social media use on the mental health of college students in the philippines: A mixed-method study protocol. *PLoS ONE*, 17, 2022.

[144] S. Lindblad, G.-B. Wärvik, I. Berndtsson, E. Jødal, A. Lindqvist, G. M. Dahlberg, D. Papadopoulos, C. Runesdotter, K. Samuelsson, J. Udd, and M. W. Johansson. School lockdown? comparative analyses of responses to the covid-19 pandemic in european countries. *European Educational Research Journal*, 20:564 – 583, 2021.

[145] H. Liu, F. Morstatter, J. Tang, and R. Zafarani. The good, the bad, and the ugly: uncovering novel research opportunities in social media mining. *International Journal of Data Science and Analytics*, 1(3):137–143, 2016.

[146] Y. Liu and Y. fang Brook Wu. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *AAAI Conference on Artificial Intelligence*, 2018.

[147] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL `https://arxiv.org/abs/1907.11692`.

[148] Y. Liu, C. Whitfield, T. Zhang, A. Hauser, T. Reynolds, and M. Anwar. Monitoring covid-19 pandemic through the lens of social media using natural language processing and machine learning. *Health Information Science and Systems*, 9(1):1–16, 2021.

[149] Z. Liu, Y. Lin, and M. Sun. *Representation learning for natural language processing*. Springer Nature, 2020.

[150] M. O. Lwin, J. Lu, A. Sheldenkar, P. J. Schulz, W. Shin, R. Gupta, and Y. Yang. Global sentiments surrounding the covid-19 pandemic on twitter: analysis of twitter trends. *JMIR public health and surveillance*, 6(2):e19447, 2020.

[151] R. K. Mahabadi, S. Ruder, M. Dehghani, and J. Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. *arXiv preprint arXiv:2106.04489*, 2021.

[152] D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, et al. Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3):93–118, 2018.

[153] R. Mao, C. Lin, and F. Guerin. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898, 2019.

[154] J. Mcauliffe and D. Blei. Supervised topic models. *Advances in neural information processing systems*, 20, 2007.

[155] A. McCosker, P. Kamstra, T. De Cotta, J. Farmer, F. Shaw, Z. Teh, and A. Soltani Panah. Social media for social good? a thematic, spatial and visual analysis of humanitarian action on instagram. *Information, Communication & Society*, 24(13):1870–1890, 2021.

[156] L. McDonald, B. Malcolm, S. Ramagopalan, and H. Syrad. Real-world data and the patient perspective: the promise of social media? *BMC medicine*, 17(1):1–5, 2019.

[157] S. F. McGough, J. S. Brownstein, J. B. Hawkins, and M. Santillana. Forecasting zika incidence in the 2016 latin america outbreak combining traditional

disease surveillance with search, social media, and news report data. *PLoS Neglected Tropical Diseases*, 11, 2017.

[158] J. Mendelsohn, C. Budak, and D. Jurgens. Modeling framing in immigration discourse on social media. *arXiv preprint arXiv:2104.06443*, 2021.

[159] O. Metwally, S. Blumberg, U. Ladabaum, S. R. Sinha, et al. Using social media to characterize public sentiment toward medical interventions commonly used for cancer screening: an observational study. *Journal of medical Internet research*, 19(6):e7485, 2017.

[160] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[161] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. Rosenquist. Understanding the demographics of twitter users. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, 2011.

[162] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, LA, USA, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1001. URL `https://aclanthology.org/S18-1001`.

[163] S. M. Mohammad. Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 323–379. Elsevier, 2021.

[164] S. M. Mohammad and P. D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465, 2013.

[165] S. Molaei, M. Khansari, H. Veisi, and M. Salehi. Predicting the spread of influenza epidemics by analyzing twitter messages. *Health and Technology*, pages 1–16, 2019.

[166] B. L. Monroe, M. P. Colaresi, and K. M. Quinn. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2008.

[167] F. Morstatter, J. Pfeffer, H. Liu, and K. Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 400–408, 2013.

[168] S. Muralidhara and M. Paul. healthy selfies: Exploration of health topics on instagram. *JMIR Public Heal. Surveill*, 4(2):10150,.

[169] S. Muralidhara, M. J. Paul, et al. # healthy selfies: exploration of health topics on instagram. *JMIR public health and surveillance*, 4(2):e10150, 2018.

[170] T. Nakatani. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. In *Proc. Interspeech 2019*, 2019.

[171] P. Nandwani and R. Verma. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):81, 2021.

[172] U. Naseem, J. Kim, M. Khushi, and A. G. Dunn. Identification of disease or symptom terms in reddit to improve health mention classification. In *Proceedings of the ACM Web Conference 2022*, pages 2573–2581, 2022.

[173] U. Naseem, B. C. Lee, M. Khushi, J. Kim, and A. G. Dunn. Benchmarking for public health surveillance tasks on social media with a domain-specific pretrained language model. *arXiv preprint arXiv:2204.04521*, 2022.

[174] I. Nazaruddin, S. B. Rezki, and Y. Rahmanda. Love of money, gender, religiosity: The impact on ethical perceptions of future professional accountants. *Business & Economic Horizons*, 14(2), 2018.

[175] D. Q. Nguyen, T. Vu, and A. Tuan Nguyen. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online, Oct. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.2. URL `https://aclanthology.org/2020.emnlp-demos.2`.

[176] Q. C. Nguyen, Y. Huang, A. Kumar, H. Duan, J. M. Keralis, P. Dwivedi, H.-W. Meng, K. D. Brunisholz, J. Jay, M. Javanmardi, and T. Tasdizen. Using 164 million google street view images to derive built environment predictors of covid-19 cases. *International Journal of Environmental Research and Public Health*, 17, 2020.

[177] A. Nikfarjam, J. D. Ransohoff, A. Callahan, E. Jones, B. Loew, B. Y. Kwong, K. Y. Sarin, N. H. Shah, et al. Early detection of adverse drug reactions in social health networks: a natural language processing pipeline for signal detection. *JMIR public health and surveillance*, 5(2):e11264, 2019.

[178] S. I. Nikolenko, S. Koltcov, and O. Koltsova. Topic modelling for qualitative studies. *Journal of Information Science*, 43(1):88–102, 2017.

[179] A. L. Nobles, E. C. Leas, S. Noar, M. Dredze, C. A. Latkin, S. A. Strathdee, and J. W. Ayers. Automated image analysis of instagram posts: Implications for risk perception and communication in public health using a case study of# hiv. *PloS one*, 15(5):e0231155, 2020.

[180] I. K. Nti, A. F. Adekoya, and B. A. Weyori. Predicting stock market price movement using sentiment analysis: Evidence from ghana. *Applied Computer Systems*, 25:33 – 42, 2020.

[181] S. Nyawa, D. Tchuente, and S. Fosso-Wamba. Covid-19 vaccine hesitancy: a social media analysis using deep learning. *Annals of Operations Research*, pages 1 – 39, 2022.

[182] F. Ofli, F. Alam, and M. Imran. Analysis of social media data using multimodal deep learning for disaster response. *arXiv preprint arXiv:2004.11838*, 2020.

[183] B. Ofoghi, M. Mann, and K. Verspoor. Towards early discovery of salient health threats: A social media emotion classification technique. In *biocomputing 2016: proceedings of the Pacific symposium*, pages 504–515. World Scientific, 2016.

[184] W. H. Organization et al. *The world health report 2007: a safer future: global public health security in the 21st century*. World Health Organization, 2007.

[185] V. Osadchiy, T. Jiang, J. N. Mills, and S. V. Eleswarapu. Low testosterone on social media: Application of natural language processing to understand patients' perceptions of hypogonadism and its treatment. *Journal of Medical Internet Research*, 22, 2020.

[186] S. Pandrekar, X. Chen, G. Gopalkrishna, A. Srivastava, M. Saltz, J. Saltz, and F. Wang. Social media based analysis of opioid epidemic using reddit. In *AMIA Annual Symposium Proceedings*, volume 2018, page 867. American Medical Informatics Association, 2018.

[187] A. Park and M. Conway. Tracking health related discussions on reddit for public health applications. In *AMIA annual symposium proceedings*, volume 2017, page 1362. American Medical Informatics Association, 2017.

[188] P. Patel and K. Mistry. A review: Text classification on social media data. *IOSR Journal of Computer Engineering*, 17(1):80–84, 2015.

[189] D. U. Patton, W. R. Frey, K. A. McGregor, F.-T. Lee, K. McKeown, and E. Moss. Contextual analysis of social media: The promise and challenge of eliciting context in social media posts with natural language processing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 337–342, 2020.

[190] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12: 2825–2830, 2011.

[191] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001): 2001, 2001.

[192] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL `https://aclanthology.org/N18-1202`.

[193] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[194] J. Phang, T. Févry, and S. R. Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*, 2018. URL `https://arxiv.org/abs/1811.01088`.

[195] R. Plutchik. A general psychoevolutionary theory of emotion. 1980.

[196] R. Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier, 1980.

[197] J. Poushter, C. Bishop, and H. Chwe. Social media use continues to rise in developing countries but plateaus across developed ones. *Pew research center*, 22:2–19, 2018.

[198] S. Pradeepa, K. R. Manjula, S. Vimal, M. S. Khan, N. K. Chilamkurti, and A. K. Luhach. Drfs: Detecting risk factor of stroke disease from social media using machine learning techniques. *Neural Processing Letters*, pages 1–19, 2020.

[199] D. Pruss. Zika discourse in the americas: A multilingual topic analysis of twitter. *PLoS One*, 14(5):0216922,.

[200] Z. Qin and Y. Song. The sacred power of beauty: Examining the perceptual effect of buddhist symbols on happiness and life satisfaction in china. *International Journal of Environmental Research and Public Health*, 17(7):2551, 2020.

[201] D. Qiu, Y. Yu, and L. Chen. Emotion analysis of covid-19 vaccines based on a fuzzy convolutional neural network. *Cognitive Computation*, pages 1 – 15, 2022.

[202] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

[203] A. S. Raamkumar, S. G. Tan, and H. L. Wee. Measuring the outreach efforts of public health authorities and the public response on facebook during the covid-19 pandemic in early 2020: Cross-country comparison. *Journal of Medical Internet Research*, 22, 2020.

[204] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[205] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[206] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551, 2020.

[207] S. Reimer-Kirkham. Nursing research on religion and spirituality through a social justice lens. *Advances in Nursing Science*, 37(3):249–257, 2014.

[208] S. Rosenthal, N. Farra, and P. Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.00741*, 2019.

[209] M. Roser, H. Ritchie, E. Ortiz-Ospina, and J. Hasell. Coronavirus pandemic (covid-19). *Our world in data*, 2020.

[210] S. Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017. URL `https://arxiv.org/abs/1706.05098`.

[211] S. Ruder. *Neural transfer learning for natural language processing*. PhD thesis, NUI Galway, 2019.

[212] M. Saeidi, G. Bouchard, M. Liakata, and S. Riedel. Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *International Conference on Computational Linguistics*, 2016.

[213] M. Santillana, A. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Computational Biology*, 11, 2015.

[214] R. Sawhney, H. Joshi, S. Gandhi, and R. R. Shah. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7685–7697, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.619. URL `https://aclanthology.org/2020.emnlp-main.619`.

[215] R. Sawhney, H. Joshi, L. Flek, and R. R. Shah. PHASE: Learning emotional phase-aware representations for suicide ideation detection on social media.

In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2415–2428. Association for Computational Linguistics, April 2021. doi: 10.18653/v1/2021.eacl-main.205. URL `https://aclanthology.org/2021.eacl-main.205`.

[216] L. Sax, H. Zimmerman, J. Blaney, B. Toven-Lindsey, and K. Lehman. Diversifying undergraduate computer science: The role of department chairs in promoting gender and racial diversity. *J. Women Minor. Sci. Eng*, 23(2): 101–119,.

[217] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[218] C. Scheele, M. Yu, and Q. Huang. Geographic context-aware text mining: enhance social media message classification for situational awareness by integrating spatial and temporal features. *International Journal of Digital Earth*, 14:1721 – 1743, 2021.

[219] S. Schneider, A. Baevski, R. Collobert, and M. Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[220] S. Seibel and N. Veilleux. Factors influencing women entering the software development field through coding bootcamps vs. computer science bachelor's degrees *.

[221] E. Seltzer, E. Horst-Martz, M. Lu, and R. M. Merchant. Public sentiment and discourse about zika virus on instagram. *Public health*, 150:170–175, 2017.

[222] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[223] A. Seyeditabari, N. Tabari, and W. Zadrozny. Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*, 2018.

[224] R. R. Sharma. Role of spiritual intervention in executive burnout. *Asia Pacific Business Review*, 2(2):25–36, 2006.

[225] S. Sidana, S. Mishra, S. Amer-Yahia, M. Clausel, and M.-R. Amini. Health monitoring on social media over time. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 849–852, 2016.

[226] K. Singh, K. Allen, R. Scheckler, and L. Darlington. Women in computer-related majors: A critical synthesis of research and theory from 1994 to 2005. *Review of Educational Research*, 77(4):500–533,.

[227] G. E. Smith, A. J. Elliot, I. Lake, O. Edeghere, R. Morbey, M. Catchpole, D. L. Heymann, J. Hawker, S. Ibbotson, B. McCloskey, et al. Syndromic surveillance: two decades experience of sustainable systems–its people not just data! *Epidemiology & Infection*, 147, 2019.

[228] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*, 2020. URL `https://arxiv.org/abs/2004.09297`.

[229] T. Sosea and C. Caragea. CancerEmo: A dataset for fine-grained emotion detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8892–8904, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.715. URL `https://aclanthology.org/2020.emnlp-main.715`.

[230] J. Staiano and M. Guerini. Depeche mood: a lexicon for emotion analysis from crowd annotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–433, Baltimore, Maryland, 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2070. URL `https://aclanthology.org/P14-2070`.

[231] J. Streich, J. Romero, J. G. F. M. Gazolla, D. Kainer, A. Cliff, E. T. Prates, J. B. Brown, S. Khoury, G. A. Tuskan, M. Garvin, et al. Can exascale computing and explainable artificial intelligence applied to plant biology deliver on the united nations sustainable development goals? *Current opinion in biotechnology*, 61:217–225, 2020.

[232] A. Subbaswamy and S. Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2): 345–352, 2020.

[233] S. tajner. Automatic text simplification for social good: Progress and challenges. In *Findings*, 2021.

[234] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565, Baltimore, Maryland, 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1146. URL `https://aclanthology.org/P14-1146`.

[235] L. Tang, B. Bie, and D. Zhi. Tweeting about measles during stages of an outbreak: A semantic network approach to the framing of an emerging infectious disease. *American journal of infection control*, 46(12):1375–1380, 2018.

[236] J. Taylor and C. Pagliari. Mining social media data: How are research sponsors and researchers addressing the ethical challenges? *Research Ethics*, 14 (2):1–39, 2018.

[237] W. L. Taylor. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.

[238] A. C. Teirlinck, E. K. Broberg, A. S. Berg, H. Campbell, R. M. Reeves, A. Carnahan, B. Lina, G. Pakarna, H. Bøås, H. Nohynek, H.-D. Emborg, H. Nair, J. Reiche, J. Oliva, J. O'Gorman, J. Paget, K. Szymański, K. Danis,

M. Socan, M. Gijón, M. Rapp, M. Havlíčková, R. Trebbien, R. Guiomar, S. Hirve, S. Buda, S. van der Werf, A. Meijer, and T. K. Fischer. Recommendations for respiratory syncytial virus surveillance at the national level. *The European Respiratory Journal*, 58, 2021.

[239] A. B. Temitope. In defense of nigerian pidgin. *Journal of Languages and Culture*, 4(5):90–98, 2013.

[240] H. Tian, C. Gao, X. Xiao, H. Liu, B. He, H. Wu, H. Wang, and F. Wu. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4067–4076, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.374. URL `https://aclanthology.org/2020.acl-main.374`.

[241] N. Tomašev, J. Cornebise, F. Hutter, S. Mohamed, A. Picciariello, B. Connelly, D. Belgrave, D. Ezer, F. C. v. d. Haert, F. Mugisha, et al. Ai for social good: unlocking the opportunity for positive impact. *Nature Communications*, 11(1):1–6, 2020.

[242] T. Uwalaka. Nairaland and the reconstruction of the public sphere in nigeria. In *Refereed Proceedings of the Australian and New Zealand Communication Association Conference: Rethinking Communication, Space and Identity, Queenstown, NZ, http://www. anzca. net/conferences/past-conferences/, ANZCA*, 2015.

[243] A. Valiavska and S. Smith-Frigerio. Politics over public health: Analysis of twitter and reddit posts concerning the role of politics in the public health response to covid-19. *Health Communication*, pages 1–10, 2022.

[244] M. Valizadeh, P. Ranjbar-Noiey, C. Caragea, and N. Parde. Identifying medical self-disclosure in online communities. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computa-*

tional Linguistics: Human Language Technologies, pages 4398–4408, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.347. URL https://aclanthology.org/2021.naacl-main.347.

[245] A. J. van Rensburg, M. Poggenpoel, C. P. Szabo, and C. Myburgh. Referral and collaboration between south african psychiatrists and religious or spiritual advisers: Views from some psychiatrists. *South African Journal of Psychiatry*, 20(2):40–45, 2014.

[246] A. Vandormael, M. Adam, M. Greuel, and T. W. Bärnighausen. An entertainment-education approach to prevent covid-19 spread: study protocol for a multi-site randomized controlled trial. *Trials*, 21, 2020.

[247] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[248] M. Viviani, C. Crocamo, M. Mazzola, F. Bartoli, G. Carrà, and G. Pasi. Assessing vulnerability to psychological distress during the covid-19 pandemic through the analysis of microblogging content. *Future Generation Computer Systems*, 125:446–459, 2021.

[249] S. Volkova, E. Ayton, K. Porterfield, and C. Corley. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PLoS ONE*, 12, 2017.

[250] T. Vos, S. S. Lim, C. Abbafati, K. M. Abbas, M. Abbasi, M. Abbasifard, M. Abbasi-Kangevari, H. Abbastabar, F. Abd-Allah, A. Abdelalim, et al.

Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The Lancet*, 396(10258):1204–1222, 2020.

[251] L. Waguespack, J. Babb, and D. Yates. Triangulating coding bootcamps in is education: Bootleg education or disruptive innovation?

[252] D. Wang, P. Liu, M. Zhong, J. Fu, X. Qiu, and X. Huang. Exploring domain shift in extractive text summarization. *arXiv preprint arXiv:1908.11664*, 2019.

[253] R.-Q. Wang, Y. Hu, Z. Zhou, and K. Yang. Tracking flooding phase transitions and establishing a passive hotline with ai-enabled social media data. *IEEE Access*, 8:103395–103404, 2020.

[254] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth. Harnessing twitter" big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592. IEEE, 2012.

[255] W. Wang, I. Hernandez, D. A. Newman, J. He, and J. Bian. Twitter analysis: Studying us weekly trends in work stress and emotion. *Applied Psychology*, 65(2):355–378, 2016.

[256] X. Wang and E. W. Lee. Negative emotions shape the diffusion of cancer tweets: toward an integrated social network–text analytics approach. *Internet Research*, 31(2):401–418, 2021.

[257] X. Wang and Y. Wang. Sentence-level resampling for named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2151–2165, 2022.

[258] E. Wilder, L. Tabak, R. Pettigrew, and F. Collins. Biomedical research: Strength from diversity. *Science*, 342(6160):798,.

[259] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL `https://aclanthology.org/2020.emnlp-demos.6`.

[260] H. Woo, H. S. Cho, E. Shim, J.-K. Lee, K. Lee, G. Song, and Y. Cho. Identification of keywords from twitter and web blog posts to detect influenza epidemics in korea. *Disaster Medicine and Public Health Preparedness*, 12: 352 – 359, 2017.

[261] I. Wood and S. Ruder. Emoji as emotion tags for tweets. In *Proceedings of the Emotion and Sentiment Analysis Workshop LREC2016, Portorož, Slovenia*, pages 76–79, 2016.

[262] J. Wortmann. Religious coping. In *Encyclopedia of behavioral medicine*, pages 1873–1875. Springer, 2020.

[263] L. Wu and H. Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018.

[264] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. URL `https://arxiv.org/abs/1609.08144`.

[265] Q. Yang, H. Alamro, S. Albaradei, A. Salhi, X. Lv, C. Ma, M. Alshehri, I. Jaber, F. Tifratene, W. Wang, et al. Senwave: monitoring the global

sentiments under the covid-19 pandemic. *arXiv preprint arXiv:2006.10842*, 2020. URL `https://arxiv.org/abs/2006.10842`.

[266] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html`.

[267] T. Yigitcanlar, N. Kankanamge, A. Preston, P. S. Gill, M. Rezayee, M. Ostadnia, B. Xia, and G. Ioppolo. How can social media analytics assist authorities in pandemic-related policy decisions? insights from australian states and territories. *Health Information Science and Systems*, 8(1):1–21, 2020.

[268] S. D. Young, N. Mercer, R. E. Weiss, E. A. Torrone, and S. O. Aral. Using social media as a tool to predict syphilis. *Preventive medicine*, 109:58–61, 2018.

[269] R. Zafarani, M. A. Abbasi, and H. Liu. *Social media mining: an introduction*. Cambridge University Press, 2014.

[270] Y. Zhang and Q. Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[271] S. Zheng, K. Han, M. B. Rosson, and J. M. Carroll. The role of social media in moocs: How to use social media to enhance student retention. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pages 419–428, 2016.

[272] J. Zhou, J. Tian, R. Wang, Y. Wu, W. Xiao, and L. He. SentiX: A sentiment-aware pre-trained model for cross-domain sentiment analysis. In

Proceedings of the 28th International Conference on Computational Linguistics, pages 568–579, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.49. URL https://aclanthology.org/2020.coling-main.49.

[273] B. Zou, V. Lampos, R. Gorton, and I. J. Cox. On infectious intestinal disease surveillance using social media content. In *Proceedings of the 6th International Conference on Digital Health Conference*, pages 157–161, 2016.

[274] B. Zou, V. Lampos, and I. J. Cox. Multi-task learning improves disease models from web search. In P. Champin, F. L. Gandon, M. Lalmas, and P. G. Ipeirotis, editors, *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 87–96. ACM, 2018. doi: 10.1145/3178876.3186050. URL https://doi.org/10.1145/3178876.3186050.

[275] O. Șerban, N. Thapen, B. Maginnis, C. Hankin, and V. Foot. Real-time processing of social media with sentinel: A syndromic surveillance system incorporating deep learning for health classification. *Information Processing & Management*, 56(3):1166–1184, 2019.

Appendix A

A.1 List of phrases

Table A.1: List of phrases used for offline activities

Activity	**Phrase**
Choir	I am doing choir
Choir	I am singing in a choir
Choir	I'm doing a choir
Choir	I'm singing in the choir
Meditation	I am doing meditation
Meditation	I'm meditating
Meditation	I'm doing meditation
Meditation	I am meditating
Meditation	I do meditation
Reflecting on nature	I am reflecting on nature
Reflecting on nature	I reflect on nature
Reflecting on nature	I ponder on nature
Prayer	I am praying
Prayer	I'm praying
Prayer	I'm doing prayer
Prayer	I am doing prayer
Prayer	I am praying
Prayer	I'm doing a prayer
Prayer	I am having a prayer
Prayer	I am in prayer
Prayer	I am holding a prayer
Prayer	I pray
Corporate worship	I am doing corporate worship
Corporate worship	I'm doing corporate worship
Corporate worship	I do corporate worship
Corporate worship	I am performing corporate worship
Corporate worship	I'm performing corporate worship
Corporate worship	I am participating in corporate worship
Corporate worship	I'm participating in corporate worship
Corporate worship	I am attending corporate worship
Corporate worship	I'm attending corporate worship
Corporate worship	I am at a corporate worship
Corporate worship	I'm at a corporate worship
Yoga	I am doing yoga
Yoga	I am practising yoga
Yoga	I'm doing yoga
Yoga	I do yoga
Yoga	I am participating in yoga

Table A.2: List of phrases used for online activities

Activity	**Phrase**
Choir	I am doing choir via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Choir	I am singing in a choir via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Choir	I'm doing a choir via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Choir	I'm singing in the choir via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Meditation	I am doing meditation via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Meditation	I'm meditating via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Meditation	I'm doing meditation via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Meditation	I am meditating via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Meditation	I do meditation via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Reflecting on nature	I am reflecting on nature via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Reflecting on nature	I reflect on nature via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Reflecting on nature	I ponder on nature via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Prayer	I am praying via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Prayer	I'm praying via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Prayer	I'm doing prayer via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Prayer	I am doing prayer via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Prayer	I am praying via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Prayer	I'm doing a prayer via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Prayer	I am having a prayer via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Prayer	I am in prayer via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Prayer	I am holding a prayer via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Prayer	I pray via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Corporate worship	I am doing corporate worship via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Corporate worship	I'm doing corporate worship via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Corporate worship	I do corporate worship via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Corporate worship	I am performing corporate worship via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Corporate worship	I'm performing corporate worship via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Corporate worship	I am participating in corporate worship via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Corporate worship	I'm participating in corporate worship via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Corporate worship	I am attending corporate worship via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Corporate worship	I'm attending corporate worship via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Corporate worship	I am at a corporate worship via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Corporate worship	I'm at a corporate worship via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Yoga	I am doing yoga via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Yoga	I am practising yoga via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Yoga	I'm doing yoga via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Yoga	I do yoga via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]
Yoga	I am participating in yoga via [Zoom *or* Microsoft Teams *or* Skype *or* Google Meet]