SOCIAL DATA MINING FOR CRIME INTELLIGENCE

H. ISAH

PHD

2017

Social Data Mining for Crime Intelligence: Contributions to Data Quality Assessment and Prediction Methods

Haruna ISAH

Submitted for the Degree of

Doctor of Philosophy

Department of Computer Science

University of Bradford

2017

Abstract

Haruna Isah

Social Data Mining for Crime Intelligence: Contributions to Social Data Quality Assessment and Prediction Methods

Keywords: Social networks analysis, data mining, social network data quality, digital crime intelligence

With the advancement of the Internet and related technologies, many traditional crimes have made the leap to digital environments. The successes of data mining in a wide variety of disciplines have given birth to crime analysis. Traditional crime analysis is mainly focused on understanding crime patterns, however, it is unsuitable for identifying and monitoring emerging crimes. The true nature of crime remains buried in unstructured content that represents the hidden story behind the data. User feedback leaves valuable traces that can be utilised to measure the quality of various aspects of products or services and can also be used to detect, infer, or predict crimes. Like any application of data mining, the data must be of a high quality standard in order to avoid erroneous conclusions.

This thesis presents a methodology and practical experiments towards discovering whether (i) user feedback can be harnessed and processed for crime intelligence, (ii) criminal associations, structures, and roles can be inferred among entities involved in a crime, and (iii) methods and standards can be developed for measuring, predicting, and comparing the quality level of social data instances and samples. It contributes to the theory, design and development of a novel framework for crime intelligence and algorithm for the estimation of social data quality by innovatively adapting the methods of monitoring water contaminants.

Several experiments were conducted and the results obtained revealed the significance of this study in mining social data for crime intelligence and in developing social data quality filters and decision support systems.

Declaration

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Haruna Isah

Acknowledgements

To God be all the glory. I couldn't have achieved such a great feat without the support of many people. First and foremost, I am forever indebted to my wife, my two kids, my parents, family and friends. They are my root and have made countless sacrifices to provide the optimal starting conditions and motivations for me to flourish. I wouldn't be in a position of enrolling and completing a PhD thesis without their endless support and love. They have been a welcome distraction.

I wish to express my unreserved heartfelt gratitude to my supervisory team which includes my principal supervisor Professor Daniel Neagu and my associate supervisor Dr Paul Trundle for their valuable support before my enrolment into the PhD programme, during funding applications and throughout the entire PhD journey. I couldn't have endured in completing this thesis without their support and understanding. Special thanks to the panel of examiners (Dr Ian Knopke, Dr Dhaval Thakker, and Professor Crina Oltean-Dumbrava) for their comments and feedback which has greatly enhanced the quality of this thesis.

Special thanks to the Commonwealth Scholarship Commission for providing me with the required studentship and grants. Am also indebted to the Federal University of Technology (FUT) Minna for supporting me throughout my studies. Same goes to the staff of the School of Electrical Engineering and Computer Science, other staff of the University of Bradford (especially Sue Baker, the International Student Adviser), Churches Together in Britain and Ireland (CBTI), WAW 2015 School organisers, and many other individuals and organisations who assisted me in achieving my dreams.

Finally, I would like to extend my thanks to the entire Tangale and my Church communities (especially my Pastors, Dr and Mrs Akpo Onduku, Dr and Mrs Amos Fatokun, Col and Mrs Solomon Inuwa), researchers, my guardians (Prof. Nebath N. Tanglang and Prof. Solomon L. Lamai), all my guarantors, and organisations that have provided me with different kind of grants, feedback on my research, publication reviews, and datasets used in my experiments.

Table of Contents

List of Figures

List of Tables

Glossary

Acronyms	Meaning
ACU	Association of Commonwealth Universities
AIC	Akaike Information Criterion
API	Application Programming Interface
BOW	Bag of Words
CSC	Commonwealth Scholarship Commission
DARPA	Defence Advanced Research Projects Agency
DTM	Document Term Matrix
FBI	Federal Bureau of Investigation
FDA	Food and Drug Administration
FOAF	Friend Of A Friend
GCHQ	Government Communications Headquarters
GraphML	Graph Markup Language
JSON	JavaScript Object Notation
INTERPOL	International Police Organization
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
MQDB	Medicines Quality Database
NAFDAC	National Agency for Food and Drug Administration and Control
NCSC	National Cyber Security Centre
PQM	Promoting the Quality of Medicines
SNA	Social Network Analysis
SVD	Singular Value Decomposition
TDM	Term Document Matrix
USP	United States Pharmacopeial Convention

VIF	Variable Inflation Factor
VSM	Vector Space Model
WHO	World Health Organization
WWW	World Wide Web
XML	eXtensible Markup Language

Chapter One: Introduction

The invention and openness of the Internet, the World Wide Web (WWW), and online social networks have changed our society both positively and negatively. In the positive dimension, for instance, information is now instant and more accessible, people can now share their opinions, experiences, and concerns about virtually everything online and global business transactions are now effectively and efficiently conducted through online electronic mediums. The explosion of user generated content on the Internet, the increase in computing power, and advances in artificial intelligence have turned social data mining into a thriving field and crucial commercial domain (Nathan and Richard, 2014). The negative dimensions, mainly fuelled as a result of internet's seamless accessibility (Scrivens et al., 2017), anonymity and pseudonymity include fraud, online abuse and anti-social behaviour, spread of misinformation and communication hijacking, spread of malicious URLs, artificial grassroots campaigns (Astroturfing), manipulation of search engines (Kyumin et al., 2014) and other uses of the Internet for exploitative and malicious purposes. Both dimensions can bring out the best or the worse in people, but most importantly, they create conditions for remarkable innovation, creativity and invention (Bartlett, 2015).

Crimes and fraud are increasingly becoming digital in nature, occurring in vast volumes and at real-time velocity (Mena, 2011). With the advent of e-commerce, digital currency, and online banking technologies, many traditional or physical-world crimes have taken a new turn into developing an online presence (Soska and Christin, 2015). Let's consider a typical crime such as the online advertisement and sale of illicit pharmaceuticals, which is of specific interest to this thesis. Imagine a scenario where a first time buyer searching for a medical product to buy online; with the growing number of Internet pharmacies, where can the buyer start? How can one verify whether a website or a product being

advertised or displayed online for sale is fake or genuine? How can one find out if the owner of an online marketplace is a benign or a deceptive user? Can we identify, characterise, and infer relationships among key actors and other entities involved in a digital crime? Counterfeit crime is a huge problem plaguing the pharmaceutical industry globally. According to the Federal Bureau of Investigation (FBI), the World Health Organization (WHO) estimates that at least 10 percent of medications sold around the world are counterfeits as such creating an illegal market worth more than $200 billion annually (Congdon, 2015). According to the INTERPOL, the increasing prevalence of counterfeit and illicit goods has been compounded by the rise in Internet trade, where medications can be bought easily, cheaply and without a prescription. Fig. 1.1 for example, is a one week (18-25th August 2017) Google Trends[1] chart for "counterfeit drugs" search, on exploring the chart, one will notice a peak of 100 search frequency on 22nd August 2017 at 11:00 AM, this indeed calls for attention as to the location and any other factor that might have led to the high interest in the term.

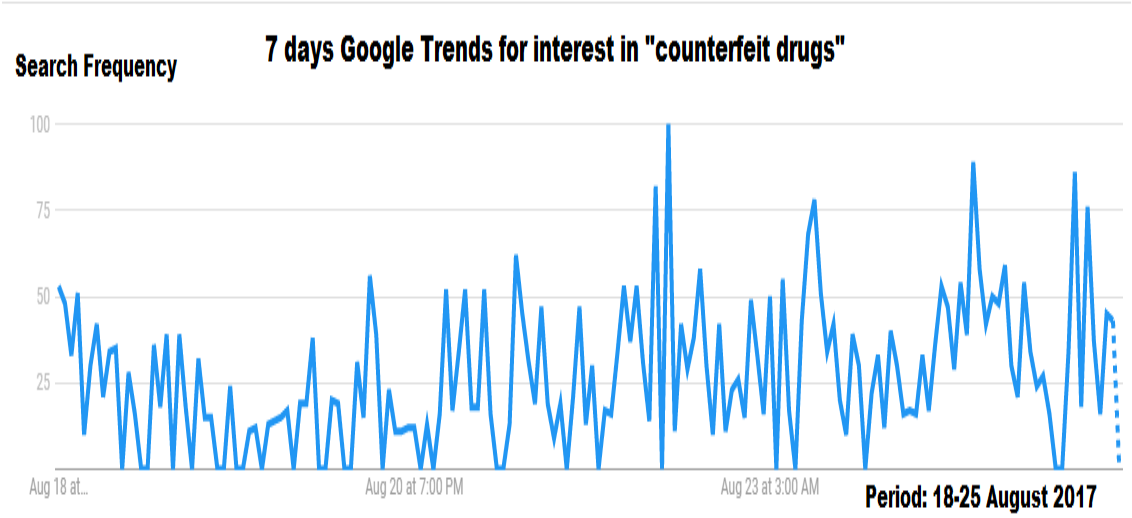


Fig. 1.1. One week Google Trends for "counterfeit drugs" search

Further exploration of the chart within this period by region revealed that the top countries where the search originated and the frequency are: the United Kingdom (100); United States (32); India (26), and Egypt (23). The rise in counterfeit drug

[1] https://trends.google.co.uk/trends/

search interest in the US and UK may be connected to recent seizure[2] and sentencing of those involved in the distribution of fake pills containing Fentanyl[3] which killed two people in Pennsylvania. Such news may generate discussions and will encourage many others to publicly share related experiences leading to social data that can be analysed in order to mitigate future crime occurrences. It is in that regard that term “crime intelligence” became relevant to this work; crime intelligence in the context of this research is the compilation, analysis, and/or dissemination of information in an effort to anticipate, prevent, or monitor criminal activity (UNODC, 2011).

Narrowing down to the pharmaceutical domain, the INTERPOL reported that perpetrators of pharmaceutical crime operate across national borders in range of activities that include the import, export, manufacture, sale, and distribution of counterfeit and illicit medicines and are attracted by huge profit potential. These criminals have also learnt to organise themselves into specialised chains of commands and responsibilities, sometimes in the deep Web, learning from, and using one another’s skills and capabilities (Ball, 2013). For example, Silk Road (Christin, 2013) and Evolution (Compton, 2015) operated internationally as a Tor hidden service and used Bitcoin as their exchange currency. The commonly traded products on these and many darknet markets include counterfeit DVDs, data, bank account and card details, fake subscriptions, counterfeit money, fake vouchers, cannabis, and prescription drugs. These therefore become necessary for organisations and agencies such as the World Health Organization (WHO), Partnership for Safe Medicines, Center for Safe Internet Pharmacies (CSIP), Permanent Forum on International Pharmaceutical Crime (PFIPC), Pharmaceutical Security Institute (PSI), Institute of research against counterfeiting medicines (IRACM), United Nations Office on Drugs and Crime (UNODC), and many similar organisations that are involved in the field to plan and coordinate operations in order to disrupt these transnational criminal networks. Despite all these efforts, pharmaceutical crime continues to

[2] http://www.safemedicines.org/2017/08/30000-counterfeit-pills-containing-fentanyl-seized-in-arizona.html

[3] http://www.safemedicines.org/2017/08/fake-pills-containing-fentanyl-kill-two-in-pa-seller-sentenced-to-18-years.html

thrive, hence the need for law enforcement and anti-counterfeiting agencies to do better product counterfeit prediction and prevention.

Counterfeit goods are products such as designer clothes, accessories, electricals or cosmetics that are fake but sold as authentic. This definition is according to the Action Fraud, the UK's national reporting centre for fraud and cybercrime. Counterfeits at one time were largely restricted to products such as watches, designer apparel, and movies; today counterfeiting is a major problem in pharmaceutical, automotive, and software industry (Berman, 2008). Pharmaceutical products which are produced and sold with the intent to deceptively represent its origin, authenticity or effectiveness are generally referred to as counterfeit medications (Wertheimer and Wang, 2012). According to the WHO, a counterfeit medicine is one which is deliberately and fraudulently mislabelled with respect to identity and/or source. The pharmaceutical counterfeit market is particularly intricate with its complexity enhanced by use of the Internet for online pharmacies. The effects of product counterfeiting include decline in long-term sales and damage in brand reputation faced by the affected firms, losses in employment, losses in income and sales tax revenues, expenses associated with increased trade deficits as well as the costs of detecting and controlling counterfeiting activity, and for medical and pharmaceutical products, effects include drug resistance, public health and safety issues which may lead to death (Dégardin et al., 2014).

Until recently, most product counterfeiting and other digital crime detection strategies are mostly reactive and incident specific solutions that are not suitable for spotting new and emerging crime techniques. These solutions in most cases may be the result of analytical test in the laboratory, reported cases or known threats, targeted government and industry market surveillance or law enforcement raids and seizures such as the operations of the Medicines Quality Monitoring (MQM) programs of the USP in collaboration with the Promoting the Quality of Medicines (PQM) program that led to the pharmaceutical quality data called Medicines Quality Database (MQDB) and Poor-Quality Medicines ALERT described in (Krech et al., 2014) as well as the operations of international crime

control agencies such as INTERPOL[4] and Partnership for Safe Medicines[5]. To be able to tackle future crime occurrence, law enforcement and anti-counterfeiting agencies need to do better product counterfeit prediction and prevention. These agencies should be proactive in preventing consumers from suffering sustained effects of counterfeit products and services; this is why developing models that can describe, measure, and assign probabilities of why certain products or services may be counterfeited in space and time and finding focal points for most effective use of these agencies resources are very important scientific and practical problems.

There are many other strategies that have been developed for mitigating product counterfeiting, these include developing early warning signals, budgeting to monitor, deter, and remove counterfeits, the use of demand-side strategies such as authentication technology, and the use of supply-side strategies such as website monitoring (Berman, 2008). The study in (Abbasi et al., 2010) proposed and applied statistical metrics for distinguishing between fake and genuine online pharmacy website. Automated crime analysis and prediction using data mining, network analysis, and machine learning techniques on historic crime data can complement the reactive and incident solutions.

Data mining involves the systematic analysis of large datasets in order to (dis)prove existing hypotheses or ideas or to discover new or previously unknown (but valuable) information using machine learning methods (McCue, 2014). Network analysis which involves the representation and analysis of relationships or information flow between individuals, groups, organisations or other entities (Samatova et al., 2013) finds application in the process of understanding and mapping of associations between entities involved in a crime, for example it attempts to explore where and when certain crimes occurred, the methods used, types of property, people, or vehicles involved. Machine learning is a branch of artificial intelligence that employ pattern recognition techniques in the analyses of vast amounts of data in order to predict some behaviour such as criminal intent and activity (Mena, 2011). The field of machine learning seeks to learn from

[4] https://www.interpol.int/Crime-areas/Pharmaceutical-crime/Operations
[5] http://www.safemedicines.org/

historical activities in order to predict future criminal behaviours; these can be, for example, burglaries, money laundering, or intrusion attacks. This study posits that the process of surveillance and regulation of the advertisement, import, export, and sales of counterfeit products and other related crimes can also be enhanced through the observation and analysis of publicly available counterfeit and other crime related data. This is the fundamental goal of this research, to harness, track, observe and measure relevant online reported views, experiences, and opinions of pharmaceutical and other related product users for predictive anti-counterfeiting.

An important task associated with the analysis of crime and intelligence information is the knowledge of data which is the lifeblood of the analytical community. Regardless of how perfect a dataset might seem to be, it almost always has some shortcomings (McCue, 2014). Generally, data can be said to have a quality problem if it doesn't mean what you think it is or does not do what should do. For instance, the dataset may contain many tables with unknown properties, it may also contain detriments such as typos, multiple formats, missing values, or it may lack metadata (Dasu and Johnson, 2003). The challenge is enormous in user generated or natural language text data especially those obtained from social media. The shortcomings in text data include irrelevant and gibberish post and comments by spammers, fake and bogus reviews, use colloquial forms of language, intentional misspellings (Agarwal and Yiliyasi, 2010). These issues or shortcomings in data can adversely affect the performance of machine learning algorithms (Sessions and Valtorta, 2006) and crime analysis results and include concepts such as reliability, validity, and timeliness. According to (Zafarani et al., 2014), distortion, outliers, missing values, and duplicates are quality issues that need to be looked into when preparing data for use in data mining algorithms. Recently, we are beginning to witness the emergence of malevolent bots and fake accounts increasingly being created to manipulate or influence internet conversations (Oentaryo et al., 2016)

and for promoting lies[6], misinformation and propaganda by governments and individuals.

Poor quality data can skew analytic results and raises serious concerns regarding the accuracy of crime related natural language processing results. Poor quality data can lead to incorrect search and analytic results, loss of revenue, and scepticism towards the data publisher; usually, different uses of the same data needs to be viewed from different quality aspects (Debattista, 2017). The authors of (Zafarani et al., 2014) illustrate these quality issues in an experiment that measures the average number of followers of users on Twitter. They claim that celebrity accounts with many followers in such experiments can be considered as outliers; such accounts can easily distort the average number of followers per individuals and other analysis based on the obtained result. Missing values in user profiling and prediction tasks include accounts with no personal information such as age, location, or hobbies while duplicates are multiple instances of social data instances (blog posts, tweets, or profiles) with the exact same feature values. These issues require special considerations and handling before carrying out analysis. For example, there is the need to identify or detect data related issues that can affect data mining results and there is also the need for methods that can be used to measure, rank, or predict the level and analytical effects of these issues and how they can be mitigated, there is also the need to ensure that user privacies are preserved, and finally, there is the need for a minimum expected quality requirement and established standards that all analytic data sources must meet.

Crime intelligence, therefore requires data to be of good quality level. This thesis is, therefore, centred on the following two major concerns, (1) the need for automated methods and tools for pharmaceutical and related crime prediction, and (2) the need to measure and ensure that the crime prediction data is of the suitable quality level. To be able to ensure that user privacy are preserved, the names of users and organisation are either randomly coded or anonymised throughout this research.

[6] https://www.theguardian.com/technology/2017/jun/19/social-media-proganda-manipulating-public-opinion-bots-accounts-facebook-twitter

1.1. Research Motivation

This research is funded by the Commonwealth Scholarship Commission (CSC) in the United Kingdom. The CSC believes that a PhD without any development impact is a complete disconnect, hence apart from the aspiration to discover and learn new things, one of the fundamental motivation for conducting this research is to contribute to the attainment of the Millennium Development Goals (MDGs) established following the Millennium Summit of the United Nations in 2000. Goal 6 of the MDGs is "to combat HIV/AIDS, malaria, and other diseases" while Goal 8 is "to develop a global partnership for development". Target E of Goal 8 is aimed at providing access to affordable essential drugs in developing countries in cooperation with pharmaceutical companies while Target F of Goal 8 is to make available the benefits of new technologies, especially information and communications in cooperation with the private sector. These goals and targets cannot be achieved until when pharmaceutical and related crimes are mitigated to the barest minimum or when only medications that meet required quality standards are supplied to those who need it.

The availability and ease of access to large user feedback data that has been left by other customers and the realisation in (McGlohon et al., 2010) and (Zhang et al., 2012) that the decision of whether or not to make an online purchase is often driven by this feedback were also a great motivation. Product users are becoming important stakeholders in product design and evaluation, for instance, concerns about adverse events associated with Essure (a popular and permanent form of birth control)[7] led to the emergence of Essure Problems Facebook group[8] which currently has over 34,000 members at the time of writing this thesis. The increasing number of complaints regarding Essure in that group made FDA held a public advisory committee meeting in 2015 to discuss the products and the possibility of further risk assessment. This is just one example of the power of social media communication. Exploring such data may lead to the discovery of valuable entities and associations.

[7] https://www.wired.com/2015/09/facebook-group-got-fda-reconsider-type-birth-control/
[8] https://www.facebook.com/groups/Essureproblems/

Again, the successes of social media sentiment analysis in various problem domains especially healthcare as well as the interest by the US Food and Drug Administration (FDA) and the Defence Advanced Research Projects Agency (DARPA)[9] on tracking social media data was another great incentive. The FDA considers the rise of social media communication on the Internet as a new opportunity to interface with the public with respect to emerging hazard situations involving its regulated products, the project goal was to detect adverse events and it aimed to achieve this through the following objectives (i) analyses of social media including user-generated content such as blogs, forums, message boards, wikis and podcasts to provide baselines on consumer sentiment prior to risk communication and to depict changes in social media buzz following the risks communications, and (ii) surveillance through social media listening for early detection of adverse events and food-borne illness. DARPA, however, considers social media as the next battleground, its interest is to learn how to mould social data to prevent adverse outcomes, it aims to achieve this through the following objectives (i) to detect, filter, classify, measure, track, and exploit the formation, development and spread of ideas and concepts (memes), as well as the purposeful or deceptive messaging and misinformation on social media, (ii) to identify participants and intent, and measure effects of persuasion campaigns, and (iii) to counter adversary messages and influence operations. We then posit that mining social data or user feedback which include ratings, natural language reviews, social media feedback, and other forms of user experiences expressed online can provide valuable knowledge that can enhance the process of crime surveillance.

Another major drive for this research is the availability and access to a huge quantity of social data. Data is evidence for machine learning forensics, the vast amounts of digital data being collected with every credit card swipe, phone calls, on social media networks, online chats, e-mail, browser and search history can be utilised to combat crime. (Mena, 2011). The explosive growth of social data from popular communication and networking platforms such as Facebook, Twitter, Instagram and blogging sites also played a major role in the design of

[9] https://www.darpa.mil/program/social-media-in-strategic-communication

this study. Interestingly, most of these platforms now offer Application Programming Interfaces (API) for data download. There are also many archived datasets from these platforms that can be accessed for research purposes. For example, the Northwestern University Centre for Ultra-scale Computing and Information Security provide Millions[10] of Amazon Reviews, Tweets, and Facebook comments for sentiment analysis research and the Tweet dataset described in (Kyumin et al., 2011) that was collected as a result of a long-term study of content polluters on Twitter. However, there is a need to verify that the datasets are indeed what they were claimed to be.

Finally, the success of any data mining application depends on the quality of the underlying data being analysed (Berndt et al., 2015), this is a fundamental motivation for further studies on data quality, which simply refers to the assessments of data against defined standards and its fitness for use (Herzog et al., 2007). Real-life data is often dirty, i.e. it can be inconsistent, inaccurate, incomplete, obsolete and duplicated (Fan, 2015). Unlike structured data sources where content is entered and managed by a selected few individuals, social data allows anyone to express an opinion on other users' contributions which often lead to the proliferation of low-quality content (Pelleg et al., 2016). These are the major motivations for this study.

1.2. Thesis Statement

The aim of this study is to apply knowledge discovery techniques for crime prediction and to develop novel methods and algorithms for social data quality measurement. The objective of this thesis is twofold. Firstly, it is concerned with the development of innovative social data mining techniques and frameworks in order to improve the understanding, intervention, and policy-making on digital crimes (with a specific interest in pharmaceutical crimes). Secondly, it seeks to contribute to the theory, design and development of models, algorithms, and management frameworks for the measurement and evaluation of social data quality.

[10] http://cucis.ece.northwestern.edu/projects/Social/sentiment_data.html

1.3. Scope

According to Her Majesty's Inspectorate of Constabulary[11], digital crimes are (i) Internet-facilitated crimes, where the internet and smart devices are used in planning and committing traditional crimes such as online abuse and terrorist communications (ii) Cyber-enabled crimes, which can be carried out either offline or online at an unprecedented scale and speed, examples include fraud, illegal drugs or firearms trade, and child sexual exploitation, and (iii) Cyber-dependent crimes, which include crimes on computer and information technology infrastructure such as the deliberate creation and spread of malware for financial gain, hacking and denial of service attacks. Both the Internet-facilitated and Cyber-dependent crimes have been the focus of research in both industry and academic Cyber Security departments over the years.

This study is specifically focused on knowledge discovery for understanding and mitigating cyber-enabled crimes which are largely unexplored, unlike cyber-dependent and Internet-facilitated crimes which are the main focus of most cyber security departments in research institutions. It also identified several issues that may affect data analysis results, the major focus here is on social data quality and the extent to which data are appropriate for a given information or analytic need.

1.4. Research Questions

Apart from the law enforcement and regulatory agencies crime incident databases, much of the information on criminals and crime victims are found in unstructured text in the grey literature (technical reports, white papers and preprints), newspapers, social media and blog threads. In most cases, these data are massively accessible and can be exploited for crime surveillance. It is in that regard that the novel research question (RQ1) was proposed as follows:

- **Research Question 1 (RQ1):** can user feedback be harnessed and processed for crime intelligence?

[11] https://www.justiceinspectorates.gov.uk/hmic/publications/real-lives-real-crimes-a-study-of-digital-crime-and-policing/

Many digital crimes require a high degree of organisation and involve a network of humans (such as manufacturers, users, and distributers) and systems (such as payment gateways, e-commerce, and delivery or shipping systems). In many cases, it is not straight forward to construct an association network among these entities, often because data does not exist or is incomplete. Means of uncovering hidden networks and key members in a network are required in order to be able to disrupt crime associations. This concern then leads to the novel research question (RQ2) as follows:

- **Research Question 2 (RQ2):** can criminal associations, structures, and roles be inferred among entities involved in a crime but not directly connected?

However, the data we are harnessing must be of the required quality standard in order to avoid misleading outcomes in achieving the goal of using data from social media for crime intelligence. This concern then becomes the consequence of the novel research question (RQ3) as follows:

- **Research Question 3 (RQ3):** can methods and standards be developed for measuring, predicting, and comparing the quality level of social data instances and samples?

Answers to these questions are vital to a range of applications including the understanding, intervention, and policy-making on crimes, as well as in the development of filters on social data platforms and in establishing social data quality standards.

1.5. Methodology

The strategy adopted in this thesis involves a mix of exploratory and empirical research. Specifically, the methodology follows a similar research process in (Leskovec, 2008) which involves observation, model design, and algorithm/system development. Throughout this research, social data which comprises of user review ratings, natural language text, and networks of relationships among entities will be harnessed and processed. Rating, text mining, and graph analysis techniques such as Bayesian estimation, topic mining, sentiment analysis, and network analysis have been utilised for crime

intelligence. Innovative methods of social data quality estimation, ranking, and prediction have been proposed and evaluated using real world datasets. The overall goal is to discover risk, segment and detect criminal behaviour and intent, measure the effect of poor quality social data in analytics and search, and finally to push towards the development of social data quality filters and standards.

1.6. Contributions and Publications

The contributions of this thesis to the discipline of computing, in the mitigation of digital crimes and towards social quality standards include:

- Novel observations and analysis of user feedback for digital crime intelligence.
- An innovative method for inferring entity relationships in crime data using k-partite network representation and projection.
- Gold standard data and benchmarks for pharmaceutical crime prediction.
- Novel characterisation of contaminants and their effects on social data quality.
- A framework for the estimation, ranking, comparison, and dynamic analysis of social data quality level.
- Identification of future research directions regarding social data quality standards.

This thesis is based on the summary of the following academic publications as well as other new unpublished studies.

- **Social Media Analysis for Product Safety Using Text Mining and Sentiment Analysis**. In this paper, the utterances of people on social media were reported to be a rich or imperfect source of information about their state, personalities, relationships, and concerns. The extraction of counterfeit related post and comments on social media was carried out and existing text mining techniques were utilised to process the data for predictive anti-counterfeiting. The experimental analysis output can serve as a gold standard training data for pharmaceutical predictive anti-counterfeiting. The paper presentation was made in August 2014 during the 14th UK Workshop on Computational Intelligence (UKCI) held at the

University of Bradford. The following are seminars and workshops related to this paper:

- **Social Media Analysis for Combating Counterfeit Products**. This presentation was part of the University of Bradford MPhil to PhD progression assessment and it took place in October 2014. The presentation demonstrated the application of knowledge discovery and data mining methods for enhancing the process of surveillance of the advertisement and sale of fake pharmaceutical and related products by exploring electronic information on counterfeit activities from diverse sources. The main lesson learnt was that developing models, algorithms, and systems that can describe, measure, and predict why certain products may be counterfeited in space and time are very important scientific and practical problems.
- **Sentiment Analysis of User-Generated Contents for Pharmaceutical Product Safety**. This presentation was made in July 2015 during the 3rd edition of the PhD-Summer Course on Machine Learning: A Computational Intelligence Approach (MLCI) held at the Department of Informatics, Bioengineering, Robotics and Systems Engineering, the University of Genoa in 2015. The main focus here was to understand the methods of text mining and sentiment analysis. The main lesson learnt was that when text mining and sentiment analysis techniques are combined in a project on social media data, the result is often a powerful descriptive and predictive tool.
- **Social Media Mining for Crime Intelligence**. This presentation was made in November 2016 as part of the School of Electrical Engineering and Computer Science research seminar and industry talk series. Again, it hypothesised that the utterances of people on social media can provide a clue to their state, personalities, relationships, and concerns. The presentation focused mainly on the techniques for mining social media text data and propose an experimental analysis with the goal of learning whether user reviews are predictive of pharmaceutical crime.

- **Bipartite Network Model for Inferring Hidden Ties in Crime Data**. This paper presents a novel approach for tackling situations where direct associations among network entities involved in a crime are of particular interest but lacking in the available data. It proposed an innovative method of inferring entity relationships by representing the data as a k-partite graph of common attributes and then projecting the graph to its unipartite components. The model was evaluated using two case studies and results obtained revealed hidden ties among criminals that were not obvious in the data. The paper presentation was made in August 2015 during the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) held at TELECOM ParisTech in Paris.
- **Measuring the Effect of Social Data Contaminants**: This contribution is currently being considered for publication in a journal, it featured research investigations towards measuring and standardising social data quality. The work reported that social data (a goldmine of consumer data) can be leveraged by brands in order to get insight into the tastes, preferences and purchase intent of their target market and by organisations for opinion detection and forecasts. The challenge, however, is that these data are being contaminated by bots, unsolicited posts, reviews or comments. The problem was casted to be analogous to evaluating the quality of water due to many characteristics similar to both social data and water. An innovative method for measuring the quality class of social data instances and for the dynamic assessment of social data quality over time was then proposed. An experimental analysis was carried out and results obtained using five different Tweet datasets revealed the significance of the study in standardising social data quality benchmarks. This study also opens new opportunities for developing quality filters in systems that support search operations as well as in the development of social data quality management standards similar to those of water quality management.

1.7. Outline

The remainder of this thesis is structured into six chapters. Each chapter covers one main topic of this thesis and is a self-contained unit that can be read

independently from the other chapters. This chapter introduced the problem context, it then states the contributions, methodology, and structure of the thesis.

- **Chapter Two. Social Data Content Mining for Crime Intelligence:** evaluates a text mining framework in a novel application that utilises user feedback to describe, measure, and predicts digital crimes.
- **Chapter Three**. **Social Data Network Mining for Crime Intelligence:** proposes and evaluates an innovative method for inferring entity relationships in crime data using k-partite network representation and projection.
- **Chapter Four. Social Data Quality Estimation:** proposes a novel framework for the characterisation of contaminants and their effects on social data quality, as well as the estimation, ranking, comparison, and dynamic analysis of social data instance quality level.
- **Chapter Five. Towards Social Data Quality Standard:** is a comparative analysis and a step towards social data quality standard. It also demonstrates the utility of our proposed algorithm for social data sample quality estimation in social search.
- **Chapter Six. Conclusions:** is a recapitulation of previous chapters and suggestion of directions for future research.

Chapter Two: Social Data Content Mining for Crime Intelligence

2.1. Introduction

The wide availability of products and services being offered through e-commerce platforms necessitates the need to accurately assess and evaluate the quality and genuineness of these offers; several research efforts have shown that consumer feedback is an important source of information that can be utilised to measure the quality of both the content and source of objects of interests such as products, restaurants, vendors, movies, music, information technology solutions, cloud services, apps, and Web pages (Ranchal et al., 2015). Because of their capability at propagating real-time information to a large audience (Li et al., 2014b), user interactions on social media platforms also leave valuable traces of feedback and opinions that can help in shaping the products and services being discussed (Wang et al., 2015). These feedback are in form of numerical ratings, review text, and reviews of review, it is an expression of the experiences of users and what they feel about the whole or aspect of a product or service (Moghaddam and Ester, 2011). Businesses can interact, solicit or extract relevant public posts and comments of their customers on their websites, on specialised websites, or on social media sites in order to understand their current and potential customers' outlook and business decisions. Review of reviews can either be in form of rating or review text. Yelp, Pharmacy Reviewer, and TripAdvisor are a typical example of specialised feedback sites for local businesses, online pharmacies, and travellers respectively. These sites are dedicated to gathering and aggregating ratings and reviews for product and service quality ranking. We will use the phrase “user feedback” to denote aspect ratings, review text, and reviews of review (which can either be as a rating or text).

The decision to purchase an online product or subscribe to a service is currently being influenced by the feedback given by other consumers (Kriti, 2017).

Companies use sentiment analysis to develop marketing strategies through the assessment and prediction of public attitudes toward their brand (Cambria et al., 2013). User feedback plays an important role in decision making (Mukherjee et al., 2013b), it is helping potential consumers to evaluate and compare products or services (Li et al., 2017). It also aid product and service producers in their design evaluations (Li et al., 2014a). The ratings and the text are two of the most important items in user feedback (Jindal and Liu, 2008) and comprise an important opinion data sources (McAuley et al., 2012), (Fei et al., 2013). The aggregation and mining of user feedback can provide early clues about product allergies, adverse events, rogue products or vendors, and valuable insights into a population (Paul and Dredze, 2011). Product manufacturers need to track patients' opinions and experiences regarding their products for decision making (Kaiser and Bodendorf, 2012). We can also get an idea of the user's concerns and the item's features from even a single review (Tan et al., 2016). While a numeric rating tells whether a user likes or dislikes an item, the associated review is capable of explaining the underlying reasons; we rely on review text to understand better how users rate items (Tan et al., 2016).

The main goal of this chapter is to answer RQ1 defined in section 1.4 by exploring the possibility of inferring the likelihood of criminal activities based on user feedback data. We will be adapting some fundamental mathematical and data mining techniques for the purpose of representing user feedback as well as mining topics, computing sentiment scores, and inferring word relationships. The rest of this chapter is organised as follows, section 2.2 is a concise review of related literature on user feedback data mining, section 2.3 introduces fundamental concepts, definitions, and state of the art models of rating and text data representational, section 2.4 showcased rating data aggregation models that and demonstrated their utility and limitation in estimating product and service quality, section 2.5 introduced topic modelling techniques and demonstrate their utility in uncovering hidden themes in natural language text, section 2.6 introduced sentiment analysis techniques and demonstrate their utility in quantifying opinion polarity, affect, emotion, mood, conditional expressions, sarcasm, and negation in natural language text, section 2.7 detailed our innovative product safety framework for harnessing and processing the views and

experiences of consumers of popular brands of drug and cosmetic products reported as status updates on social media platforms, section 2.8 presents results obtained and lessons learnt through experimental work aimed at measuring the true quality of regulated products using aspect rating data as well as on the use of textual social data for inferring rogue vendors of pharmaceutical and cosmetic products, section 2.9 highlighted issues that can affect the validity of data analysis results and recommendations on mitigating same, and finally section 2.10 is the chapter conclusion which provides an overview of all objectives met and a pointer to our future work.

2.2. Literature Review

This section presents an overview of the existing literature regarding the nature and state of the art techniques of representing and mining user feedback data. With the growing importance of user feedback data, the need for mining themes opinions, attitudes, and emotions from the data grow rapidly. This section is structured into paragraphs each highlighting our review of the nature of feedback data and the applications of Bayesian estimation, topic mining, sentiment analysis, and other mathematical techniques in mining feedback data.

According to (Moghaddam and Ester, 2011) and (McAuley and Leskovec, 2013a), feedback data are generally categorised under numerical ratings, review text and reviews of review which can either be in form of rating or review text. A rating is a set of a numerical or categorical scale designed to elicit subjectivity information about a quantitative or a qualitative attribute of an aspect (product, organisation, individual, or service). A review text is a subjective evaluation of a product or service by their customers using a natural language text. The opinions of consumers in forms of post or comment regarding certain products or services on social media platforms also constitute a review text. These descriptions agree with those of (Yang et al., 2016) which states that a rating is a numeric score whereas a review text is a comment with a more detailed description of sentiments. In general, a rating indicates whether a user likes or dislikes an item while the associated review text explains the reason behind the like or dislike (Tan et al., 2016).

Rating aggregation and ranking are the major challenges that naturally arise when rating data accumulate. These tasks are usually related and involve the aggregation and sorting of rating data such that the highest-rated item appears at the top while the lowest-rated item at the bottom. Aggregation is useful in ranking the quality of product vendors (McGlohon et al., 2010), in recommendation systems (Tan et al., 2016) and other applications that utilise preference space measures such as star ratings. Evan Miller have had a look into many practical rating aggregation algorithms from the most trivial ones that simply compute the difference between the positive and negative ratings to intermediate ones such as the lower bound of Wilson score confidence interval in (Miller, 2009), and then to the more advanced Bayesian methods in (Miller, 2012), (Masurel, 2013), (Bengfort, 2014), (Miller, 2014), and (Landy, 2015). Further analysis of other ranking functions in practice include the article in (Salihefendic, 2015a) that highlight how news items on Hacker News are ranked based on three criteria (ratings, gravity and time), the related articles in (Miller, 2015a), (Miller, 2015d) and (Salihefendic, 2015b) that explored how Reddit's story and comment rankings work, the article in (Miller, 2015c) that proposed a model for estimating Tweet's quality based on retweeting behaviour and time, and the article in (Miller, 2015b) which explores how the Splatoon ranking system works in theory.

To understand how users evaluate a product or a service, we must understand the hidden themes in their textual feedback. Relevant research tasks associated with textual reviews include review or opinion summarisation (Hu and Liu, 2004), (Nishikawa et al., 2010), recommender system development (Ganu et al., 2009), sentiment and emotion analysis (Turney, 2002), (Brody and Elhadad, 2010), (Liu, 2012), (Kiritchenko et al., 2014), (Liu, 2015), (Fersini et al., 2016), fake or deceptive reviewer detection (Lim et al., 2010), (Jindal et al., 2010), (Wang et al., 2012a), (Mukherjee et al., 2012), (Fei et al., 2013), (Mukherjee et al., 2013a), (Li et al., 2017), review quality analysis (Liu et al., 2007), (Lu et al., 2010), duplicate and untruthful review detection (Jindal and Liu, 2008), (Mukherjee et al., 2013b), (Li et al., 2014a), topic detection (Lin and He, 2009), (McAuley and Leskovec, 2013a), (Lim and Buntine, 2014), topic diversion (Wang et al., 2015), and user account activity classification (Li et al., 2014b). The proliferation of social media platforms has prompted a surge of interest in learning

from social data using topic mining and sentiment analysis techniques (Wang et al., 2012c). These techniques, especially topic modelling and sentiment analysis are useful in feedback data mining and are of special interest to our research.

Topic modelling has previously been used for a variety of applications, including ad-hoc information retrieval, geographical information retrieval, and the analysis of the development of ideas over time in the field of computational linguistics (Hornik and Grün, 2011). It provides us with a way to infer the latent thematic structure behind a large collection of text documents using a suite of algorithms called topic models (Blei, 2013). The output of topic models is useful in search and summarisation in large archives of texts as well as in pattern discovery in genetic data, images, and social networks (Blei, 2012). While most topic models are unsupervised, mostly used to construct features for classification, the model in (Mcauliffe and Blei, 2008) is however supervised and is suitable for classification task because there is a response variable associated with each document. There has been a significant progress in recent years on topic modelling technologies in text mining, information retrieval, natural language processing, and other related fields (Wang et al., 2013). Popular topic modelling methods include Probabilistic Latent Semantic Analysis (PLSA) proposed in (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) proposed in (Blei et al., 2003). We will be utilising the implementation of these models in our user feedback data explorations.

People are now highly opinionated, they hold and express their views about everything from international politics to pizza delivery; sentiment analysis often referred to as opinion mining, is the field of study that analyses people's views, evaluations, attitudes, and emotions expressed in written language (Nathan and Richard, 2014). Sentiment analysis aid consumers to know what other people think about a product or service and also guide producers to gauge public opinion with respect to their products or services. It is also useful to policy and decision makers to judge the impact of their policies on a given community (Liu, 2015). In (Wang et al., 2012b), (Gerber, 2014), and (Chen et al., 2015) tweets were utilised as additional features in criminal activity prediction, the authors reported that these models outperformed the versions of the models without the Twitter data. The authors also argued that these results are

indications that location, timing, and content of tweets are informative with regard to future events. In terms of its commercial potential, the study in (Kessler and Nicolov, 2015) highlight some typical sentiment analysis potentials in the automobile industry.

Narrowing down to our targeted domain of pharmaceutical and other related crime, the growing number of online users and interactions such as e-commerce and business transactions which provide fewer direct sensory cues, have less immediate gratification, entail more legal uncertainties, and present more opportunities for fraud and abuse (Liu et al., 2014). The issue of limited resources is constantly putting pressure on law enforcement agencies in fighting crime, the use of analytics to both respond to and ward off crime is now more critical than ever. According to IBM Crime Insight and Prevention solution white report[12], crime prediction and prevention analytics help agencies make the best use of available resources to monitor, measure and predict crimes and crime trends. The report added that the analysis of crime data (such as historical crime incidents, profiles, maps and typology) as well as enabling factors (such as weather) and trigger events (such as holidays or paydays) can provide insight that can help to track criminal activities, to predict the likelihood of incidents, to effectively deploy resources and to solve criminal cases faster.

Motivated by the state of the art and prospects of crime prediction as well as the availability and significance of user feedback data, we aimed to advance on these and try something useful, novel and challenging. In the next sections, we will provide details of the input representation, mathematical techniques, and tools for mining user feedback in order to infer product and service quality, uncover themes in review text, compute sentiment scores, infer crimes, and investigate data quality issues. We will be laying much emphasis on the domain of pharmaceutical crime which is of special interest to us.

2.3. Social Data Definition and Representation

Social network and other sites that allow user participation are becoming a useful source of information on consumer insights like demographics, preferences,

[12] ftp://ftp.software.ibm.com/software/data/sw-library/cognos/pdfs/whitepapers/wp_crime_prediction_and_prevention.pdf

sentiments, questions, and expectations. This same information is readable by humans, made for human interaction, and can be used by individuals or organisations for formulating strategies. The massive amount and speed of this information, however, exceeds human processing capacity. This is where computers and algorithms come in, we need a means of automatically sifting through this large amount of information. Currently, computers do not understand natural language communication unless it is transformed into its preferred language. Social data is social media that is readable by computers (Reddy, 2013).

Definition 2.1. Social Data: is defined as the expression of the computer-readable format of publicly shared content and metadata from social media platforms such as Wikipedia, Disqus, Twitter, Facebook, YouTube, and Tumblr (Reddy, 2013). In this study, social data denote a collection of user generated contents such as user rating and reviews, Tweets on Twitter, and updates on Facebook or Instagram.

In machine learning and data mining, effective data representations can lead to an improved performance and may unveil some hidden information in a dataset (Zhong, 2013). The performance of data mining and machine learning methods heavily relies on the choice of data representation (Faruqui, 2016). In order to learn and process valuable information, effective techniques are required for the extraction and representation of the data. Although beyond the scope of this work, there is also the need to evaluate the performance of different data representation on the same dataset and how each is affected by quality issues. Next, the structures and methods of representation of both rating and review text extracted from user feedback data will take the centre stage.

2.3.1. Representation of Rating Data

A rating is an intended interpretation of the satisfaction of users of a product or services in terms of numerical values (Moghaddam and Ester, 2011). Most review websites use ratings in the range from 1 to 5 stars (see TABLE 2.1). The star rating is one of the popular measures for product evaluation (Zhang et al., 2012) and (Ranchal et al., 2015). There are, however, many variants of rating scales, these include unary scale, 5-star scale in half star increments, 3-star rating

system, a scale from -3 to +3 or -5 to +5, up/down vote or thumb-up/down, and 100 points slider.

TABLE 2.1. Aspect star rating scale, description and value

Star Rating	Description	Value
☆☆☆☆☆	Excellent	5.0
☆☆☆☆	Above Average	4.0
☆☆☆	Average	3.0
☆☆	Below Average	2.0
☆	Poor	1.0

These rating scales are widely used in online marketplaces such as EBay, Netflix, Amazon, IMDb, Apple App store, Google Play store, as well as by news and social networks in order to gather user opinions of an entire aspect or aspect features. Other rating websites include Epinions, TripAdvisor, Skytrax, Yelp, and Pharmacyreviews. A sample of a five star rating scale from Pharmacy Review website used for rating five aspects of Online Pharmacies is illustrated in Fig.2.1.

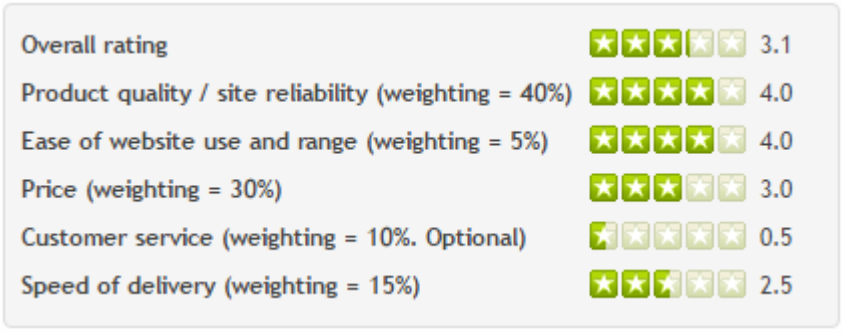


Fig. 2.1. Five star aspect rating scale from Pharmacy Reviewer

These aspects include the product quality or site reliability, the ease of website use, the product(s) price, the customer services, and the speed of delivery. Each of these aspects was assigned an importance weighting factor against which the overall rating is computed. The extraction of online rating data is often straight forward scraping or by using an application programming interface (API). The matrix in TABLE 2.2 is a simple representation of a rating data consisting of five users and items, it describes values giving for each user-item pair, i.e. a value (in

this case stars in the range 1 to 5) that represents the degree of preference of a particular user to a given item.

TABLE 2.2. A simple matrix representation of rating data

	Item 1	Item 2	Item 3	Items 4	Item 5
User 1	0	3	0	3	0
User 2	4	0	0	2	0
User 3	0	0	3	0	0
User 4	3	0	4	0	3
User 5	4	3	0	4	0

The zero's represent situations where the users have not rated the item. In practice, the matrix would be even sparser, with the typical user rating only a tiny fraction of all available items (Leskovec et al., 2014). A sparse matrix is a better and much more memory-efficient representation of the same information, it only retains the information about the non-zero values (see TABLE 2.3).

TABLE 2.3. Sparse matrix representation of rating data

User Index	Item Index	Rating
1	2	3
1	4	3
2	1	4
2	4	2
3	3	3
4	1	3
4	3	4
4	5	3
5	1	4
5	2	3
5	4	4

It is expressed as a coordinate list or vector of (i, j) pairs, where i is a vector of row indices and j is a vector of column indices. Therefore only the (i, j) pairs that have values are included. Given the vectors for both users and items, one commonly performed analysis is to estimate the degree to which a user would prefer an item by computing the distance between the user's and item's vectors. In situations when the data is very large leading and computationally costly, dimensionality reduction methods such as Singular Value Decomposition (SVD)

are often employed to reduce the matrix by approximation into a product of two long and thin matrices (Leskovec et al., 2014).

This study focuses on the aggregation of user feedback data in order to rank the quality of pharmaceutical vendors, products, and their aspects. An efficient R and Python sparse matrix implementations will be utilised to represent rating data.

2.3.2. Representation of Text Data

Most information that has value in law enforcement and intelligence analysis resides in the unstructured or narrative format as text data (McCue, 2014). In this study, text data refers to the aggregated natural language evaluation of a product or service in textual form. Major sources of text data include social media feedback and textual components of user reviews. The extraction of text data is usually carried out by scraping the textual content of social media content or user review documents or by using an application programming interface (API). A sample of review text from Pharmacy Reviewer website is shown in TABLE 2.4. Here we only showed the title and content of the text data, there will be more metadata in most cases.

Text mining, according to Lexalytics[13], is the process of collecting and processing high-quality information from unstructured text such as large streams of Twitter posts, a collection of scientific papers, or user reviews. Text mining holds considerable promise for public safety and security mining and analysis (McCue, 2014). The basic objective of text mining can be reduced to the extraction and discovery of relevant and valuable information from large volumes of natural language text data (Banchs, 2012). Majority of the algorithms that are used to catch the underlying information in a text data require the textual input to be in the form of vectors rather than strings of plain text. The goal of text processing is to convert the many forms in which words can occur in a text document into a more consistent representation (Croft et al., 2010). The raw review text data has to go through several steps of processing in order to have the right abstract representation that can be feed into algorithms. The type and nature of representation of a text document determine the kind of mining

[13] https://www.lexalytics.com

algorithms that can be applied to it. Depending on the intended application, there are multiple ways of representing text documents, these according to (Zhai, 2017) include String representation (for string processing applications), Word representation (for word relation analysis, topic analysis, sentiment analysis, Syntactic representation (for syntactic graph analysis), Entities and Relational representation (for knowledge graph and information network analysis), Logic and Predicates representation (for integrated analysis of scattered knowledge and logic inference).

TABLE 2.4. Samples of review text from Pharmacy Reviewer

Title	Review Text
Reliable, but beware	Reliable IOP. Average speed for shipping. Product is just fine. Beware though, you will be bombarded with annoying phone calls and emails from now to eternity. Don't give them a cell phone number or email address that you ever intend to use again. They're relentless when it comes to spamming their customers. I would never use them for that reason. There are other IOP's offering equal or better values that have respect for their shoppers.
Paid never got order	Paid them over a month ago, nothing has arrived. Blaming it one the "post office"
Low prices, great quality	Got my stuff within the promised period including bonus pills added for being a repeat customer. they saved me a lot of $$:) Buying online has been easier all of a sudden.

Word-based representation is the suitable approach for our study and in general for natural language processing. It involves some form of chopping up or tokenisation of the text data into smaller features or terms such as characters, words, n-grams, and skip grams. The tokens are then usually transformed to vectors and matrices for further analysis. A character is simply a unit of information in a text such as an alphabet or symbol. In English language, words are often separated from each other using whitespaces while n-grams or shingles are a set of n words or characters that occur in an ordered sequence in a text. A skip gram is a representation of a set of n-grams by a subset of it, for instance "review data" is a skip gram in "review text data". Regular expressions, often the first model of text processing tasks, are a sequence of string patterns that utilises whitespaces and other patterns to capture generalisations in the text (Banchs, 2012). Regular expressions are used to determine if a string, matches a pattern

and to split natural language sentences into individual tokens. Regular expressions find application in the extraction of tags, usernames, emails, and URL's from text data. We will describe text representation models by first presenting some fundamental observations, properties and regularities governing large volumes of text, which are generally referred to as corpus statistics.

These techniques and existing implementations of regular expressions and text representation models in R and Python will be utilised to extract and process tokens from textual user feedback data.

Properties of text data: despite its apparent randomness, a simple exploration of text data in any language will make evident that text sequence is usually characterised by a lot of regularities (Banchs, 2012). The following are fundamental properties that are implicit in languages:

- **Zipf's law:** the law states that, for any language, if we one plots the frequency of words versus their rank for a sufficiently large collection of textual data, one will observe a clear trend, which resembles a power law distribution. In simple terms, given a set of words or vocabulary of a language, the Zipf's law states that only very few words are responsible for the largest proportion of the entire text data while most of the words in the vocabulary seldom occur. This means that words in a language do not have the same probability of appearing in a given segment of text, some have a very large probability of occurrence while other have a very low probability.
    - Another interesting characteristic of natural language text, also due to Zipf's law, is that frequently used words in a language are, in most cases, the shortest ones.
- **Intermittency or burstiness**: this property illustrates the tendency of words within a text sequence to repeat themselves following some specific patterns that resemble bursts of occurrences (Banchs, 2012). This means that vocabulary words are not evenly distributed along text sequences.
- **Word co-occurrence**: accounts for the frequency of appearance of a pair of words alongside each other in the whole collection's vocabulary of textual data. Word co-occurrence can be computed by disregarding the

relative position of the two words in each word pair under consideration or by taking into consideration their order or specific locations within the given segment of the text. Particularly, co-occurrences are of great utility in specific problems such as lexical acquisition, semantic similarity estimation, collocation extraction, and in more general areas of study such as statistical semantics. Pointwise mutual information (PMI) is a very useful concept for assessing the nature of a given word co-occurrence. The mutual information of two events measures how much the joint probability of such events deviate from what would be expected if they were independent and is defined in (Banchs, 2012) as follows:

$$I(w_a, w_b) = \log\left(\frac{p(w_a, w_b)}{(p(w_a)p(w_b))}\right) \quad (2.1)$$

where $p(w_a, w_b)$ denotes the joint probability for events w_a and w_b which means the occurrence of words a and b in a text, and $p(w_a)$ and $p(w_b)$ denote the individual probabilities for the same events. If w_a and w_b were actually independent, their joint probability would be equal to the product of the individual probabilities and the value of $I(w_a, w_b)$ would reduce to zero; if the w_a and w_b are not independent, then either a positive affinity effect is observed in which the words tend to co-occur more often than expected or a negative affinity also called rejection effect is observed in which the two words tend to co-occur less often than expected in the case of statistical independence. In such cases, $I(w_a, w_b)$ values would be greater or less than zero, respectively.

Mathematical models of text representation: Here, we are going beyond just words to larger conventional units of text such as sentences, paragraphs and documents. Common mathematical word representation methods include n-gram, bag of words, vector space, and latent semantic analysis models.

- **N-gram model**: derived from an approximation of the probability of a sequence of words, which is based on a Markov property assumption which in the context of this work, means that a given word only depends on a fixed number of preceding words. Given a unit of text w which consist

of a sequence of words $w_1, w_2, \ldots, w_m$. The probability of such a sequence as defined in (Banchs, 2012) can be decomposed in the following product of probabilities using chain rule:

$$p(w) = p(w_1, w_2, \ldots, w_m) \quad (2.2a)$$

$$p(w) = p(w_1)p(w_2|w_1)p\big(w_3\big|w_1, w_2 \ldots p(w_m|w_1, w_2 \ldots w_{m-1})\big) \quad (2.2b)$$

According to the Markov property, n-gram models are defined by approximating the conditional probabilities in equation (2.2b) by conditional probabilities that only depend on the previous n-1 words in the sequence. In this way, n-gram models of order one (1-gram or unigram), order two (2-gram or bigram), order three (3-gram or trigram), and so on, were defined in (Banchs, 2012) as follows:

$$1-gram: p(w) \approx p(w_1)p(w_2) \ldots p(w_m) \quad (2.3)$$

$$2-gram: p(w) \approx p(w_2|w_1)p(w_3|w_2)p(w_4|w_3) \ldots p(w_m|w_{m-1}) \quad (2.4)$$

$$3-gram: p(w) \approx p(w_3|w_2, w_1)p(w_4|w_3, w_2) \ldots p(w_m|w_{m-1}, w_{m-2}) \quad (2.5)$$

$$n-gram: p(w) \approx \prod_i p(w_i|w_{i-1}, w_{i-2} \ldots w_{i-n+1}) \quad (2.6)$$

Different from the 1-gram case, word order is taken into account in the 2-gram, trigram, and the general n-gram model. The probability of a given word depends on the word immediately before in the 2-gram, it depends on the previous two words in the 3-gram, and so on.

- **Bag of words model**: In the extreme case of the unigram, the resulting model is completely independent of the order of words and is known as the bag of words (BOW) model. Topic models are a general class of BOW (Banchs, 2012).
- **Vector Space Model (VSM):** a VSM treats a text document as a BOW and ignores the dependence between words (Turney and Pantel, 2010). VSM captures both syntactic and semantic concepts in a sentence (Maas and Ng, 2010). A text document is conceptually represented in a VSM by

a vector of keywords extracted from the document and their associated weights representing the importance of the keywords in the text document and within the whole document collection. As illustrated in (Zafarani et al., 2014), given a set of text documents D in which each document is a set of words. The goal of VSM is to convert these textual documents to feature vectors. According to (Zafarani et al., 2014), a document i can be represented with a vector d_i, as follows:

$$d_i = \left(w_1, w_{2,\ldots,} w_{N,i}\right) \tag{2.7}$$

where $w_{j,i}$ represents the weight for word j that occurs in document i while N is the number of words used for vectorisation. We compute $w_{j,i}$ in binary terms by setting it to a value of 1 when the word j exists in document i or 0 when it does not; it can also be computing in frequency terms by setting it to the number of times the word j is observed in document i. However, ordinary VSM for large text document is very costly in terms of memory, it often ranges from tens of thousands to millions of dimensions. A more generalised approach is to use the term frequency-inverse document frequency (TF-IDF) weighting scheme. The TF-IDF is a measure for evaluating word important in a text and is used to filter less important words such as stop words. In the TF-IDF scheme described in (Zafarani et al., 2014), $w_{j,i}$ is computed as follows:

$$w_{j,i} = tf_{j,i}\, X\, idf_j \tag{2.8}$$

where $tf_{j,i}$ is the frequency of word j in document i. idf_j is the inverse TF-IDF frequency of word j across all documents, and according to (Zafarani et al., 2014), is computed as follows:

$$idf_j = \log_2 \frac{|D|}{|\{document \in D | j \in document\}|} \tag{2.9}$$

which is the logarithm of the total number of documents divided by the number of documents that contain word j. TF-IDF assigns smaller weights

to stop words, such as “and, “the”, and “because” which are often common in all documents. TF-IDF assigns higher weights to words that are less frequent across documents but highly frequent within the document they are used. This guarantees that words with high TF-IDF values are important and can serve as representative examples of the documents they belong to.

- **Latent Semantic Analysis (LSA):** this is another very useful technique for text representation, it is a matrix factorisation method for generating low dimensional word representations of a text document. It uses SVD to perform dimensionality reduction on the TF-IDF vectors of the text document[14]. The goal of LSA is to represent documents and words or terms in a unified way for exposing document-document, document-word, and word-word similarities (Thomo, 2009). The document and word are expressed as vectors with elements corresponding to these concepts and each element in a vector give the degree of participation of the document or word in the corresponding concept. LSA finds application in text data classification, clustering, and search. Both VSM and LSA are simple models for computing a continuous degree of similarity between words in a text.

2.4. Rating Aggregation

Reputation systems are often used to promote online trust by identifying true reputation scores of entities such as users, products, or services based on the feedback of others (Liu et al., 2014). In e-commerce, reputation systems collect, distribute and aggregate feedback about the past behaviour of actors (buyers and sellers) involved in the system. Historical buyers usually share their experiences, normally in the form of a numerical rating reflecting their level of satisfaction. These ratings are aggregated for each actor as a trust mechanism to promote or demote a product or service, to help users decide whom to trust and to deter participation by those who are dishonest (Liu et al., 2014). The automated mining of ratings to produce a re-calculated quality ranking score on e-commerce sites such as EBay and Amazon are a typical example of a reputation system. Given

[14] http://mccormickml.com/2016/03/25/lsa-for-text-classification-tutorial/

a set of ratings of products or aspects of a product from a wide range of users, an interesting challenge is to measure the true quality of the product or aspect (McGlohon et al., 2010). These measurements include aggregating, sorting, ranking and displaying the highest-rated product at the top and lowest-rated product at the bottom of an e-commerce site or to categorise services as highly recommended or not recommended. This is similar to the Google PageRank algorithm for ranking Web pages and EigenTrust algorithm in peer-to-peer networks (Liu et al., 2014).

Rating Aggregation Models: the aggregation of rating scores is often straight forward (positive ratings – negative ratings) as in Fig. 2.2, however, such simple aggregation method is not enough and can lead to a wrong conclusion regarding the quality of products or services, especially when some products or services have only very few ratings.

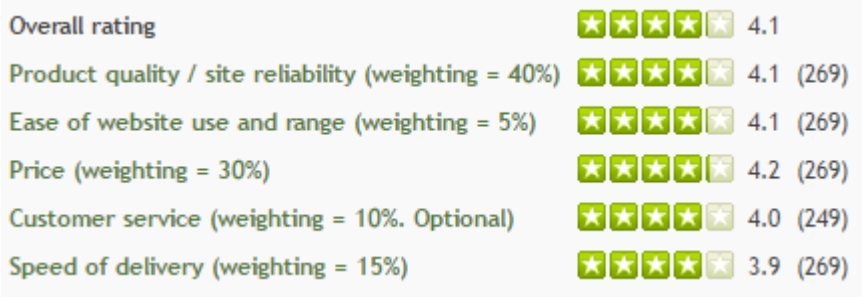


Fig. 2.2. Average Generic Doctor's ratings on Pharmacy Reviewer

Given a set of ratings R where each rating $r(o_i, a_j)$ is a numeric value representing author $a_j's$ opinion of object o_i, our overall goal which is the same with the study in (McGlohon et al., 2010) is to estimate and rank each object $o_i's$ quality q_i, relative to the other objects. Statistical models for estimating and ranking the quality q_i of user ratings are outlined in equations 2.10 to 2.14.

- **Average**: a frequently used baseline model is the average of all the ratings received from all authors. This according to (Ranchal et al., 2015) is given by equation 2.10 where the variables r_{i*} and r_{*j} refers to the set of all ratings for a given object o_i and the set of all ratings from a given author a_j.

$$q_i = \frac{1}{|r_{i*}|}\sum_{j \in r_{i*}} r_{ij} \tag{2.10}$$

- **Lower bound on normal confidence interval**: it is a model based on lower bound on normal confidence interval is used as better estimates in situations where ratings for certain products or objects are more consistent than those of others. An object rating is said to fall in a distribution around its true quality with some variance σ_i^2. The lower bound described in (Miller, 2009) is then used to calculate q_i (see equation 2.11). For a 95%, the constant $z_{\alpha/2}$=1.96.

$$q_i = \overline{r_i} - z_{\alpha/2}\frac{\sigma_i}{\sqrt{|r_{i*}|}} \tag{2.11}$$

For an object with only a few ratings, Wilson score (see equation (2.12)) which is based on lower bound on binomial confidence interval was recommended in (Miller, 2009) for the simplification of star ratings into positive or negative as well as to determine with confidence what percentage of users took some sort of action. Here $\hat{p}$ is the proportion of positive ratings for a given object o_i.

$$q_i = \hat{p} + \frac{z_{\propto/2}^2}{2|r_{i*}|} - z_{\alpha/2}\frac{\sqrt{r_{i*}\left[\hat{p}*(1-\hat{p})+z_{\propto/2}^2/4|r_{i*}|\right]/|r_{i*}|}}{\left(1+z_{\propto/2}^2/|r_{i*}|\right)} \tag{2.12}$$

The model in equation (2.13) assumed that some authors are more precise than others, therefore, each author has a variance σ_j^2. The model is also described as a stochastic function of the true quality q_i of an object o_i and the noise of an author. In equation (2.13), $\widehat{q_i}$ is a weighted average of ratings by all authors, where each author is weighted according to the noise among their ratings. It also follows from (Miller, 2009), that the noise can be computed as the sample variance around the quality scores.

$$\widehat{q_i} = \frac{\sum_{r_{i*}\frac{1}{\sigma_j^2}} * r_{ij}}{\sum_{r_{i*}\frac{1}{\sigma_j^2}}} \tag{2.13}$$

- **Bayesian models**: Bayesian statistics, which takes into account the fact that we lack enough data to make an estimation via the mean is becoming popular and was applied in (Miller, 2012), (Masurel, 2013), (Bengfort, 2014), and (Miller, 2014) in order to sort ratings with items that have an insufficient number of ratings. Bayesian rankings differ from other methods, it requires some assumptions about the system beforehand and closely reflects the opinion of a large number of users. We derived our motivation from (Chiang, 2012). Given a list of vendors indexed by i offering various medical products for sale and enlisted on Pharmacy Reviewer. For each vendor i, there are n_i reviews. Given that the simple naïve average of the vendor reviews is r_i and that the total number of all reviews for all vendor $N = \sum_i n_i$. Equation (2.14) described in (Chiang, 2012) is used to denote the overall average reviews for all vendors on Pharmacy Reviewer.

$$R = \frac{\sum_i n_i r_i}{N} \tag{2.14}$$

One important question however, is at what number of reviews N can one trust the average R as the measure of vendor quality on Pharmacy Review? Should a prospective customer trust and buy a medical product from a vendor with an average rating score of 4.8 out of 5 from only two reviews or the vendor with an average rating score of 4.2 out of 5 from 1000 reviews? Can we adjust the average rating based on review population? It also follows from (Chiang, 2012) that the Bayesian ranking (see equation (2.15)) is the adjusted score that lies between r_i and R. As n_i increases $\widehat{r_i}$ will approach r_i.

$$\widehat{r_i} = \frac{NR + n_i r_i}{N + n_i} \tag{2.15}$$

Example: Let's assume there are a total of 50 ratings for two online vendors Generic Doctor and Medstore Online on Pharmacy Reviewer with an overall average rating of 4.0. Given the rating distribution for the two vendors, we are interested in measuring the quality of each vendor from the user ratings based

on both the Average and Bayesian ranking models. We applied equation (2.14) and (2.15) to calculate and rank the ratings (see TABLE 2.5 for the scores and ranks).

$$R_{GD} = \frac{5.0 + 5.0 + 5.0}{3}$$

$$R_{GD} = 5.0,$$

$$\widehat{r_{iGD}} = \frac{50X4.0 + 3\dfrac{5.0 + 5.0 + 5.0}{3}}{50 + 3}$$

$$r_{iGD} = 4.057,$$

$$R_{MO} = \frac{3.0 + 3.5 + 4.5 + 4.0 + 5.0 + 5.0 + 5.0 + 4.5 + 5.0 + 4.0}{10}$$

$$R_{MO} = 4.35,$$

$$\widehat{r_{iMO}} = \frac{50X4.0 + 10\dfrac{3.0 + 3.5 + 4.5 + 4.0 + 5.0 + 5.0 + 5.0 + 4.5 + 5.0 + 4.0}{10}}{50 + 10}$$

$$r_{iMO} = 4.058.$$

The result shows that Medstore Online vendor will be ranked higher than Generic Doctor based on the adjusted Bayesian ranking even though it was rank the opposite way in terms of average ranking. The application of these models in aggregating online pharmacy vendors rating data will be demonstrated in subsequent sections using R and Python.

TABLE 2.5. Rating distribution for two vendors on Pharmacy Reviewer

Vendor	Rating: 5-Star	$\boldsymbol{R}$, rank	$\widehat{\boldsymbol{r_i}}$, rank
Generic Doctor (GD)	5.0, 5.0, 5.0	5.0, 1	4.057, 2
Medstore Online (MO)	3.0, 3.5, 4.5, 4.0, 5.0, 5.0, 5.0, 4.5, 5.0, 4.0	4.35, 2	4.058, 1

Despite the real life application of Bayesian ranking, it still has some limitations. For instance, most reviews are either very low (~1) or very high (~5). It is widely believed that reviews tend to have a bimodal (rather than Gaussian) distribution with centres around most negative and most positive reviews because in most cases it is those people who are extremely satisfied or dissatisfied that are more likely to care enough to submit ratings (Li et al., 2017).Rating systems also faced

gatekeeper and anonymity problem, the challenge of the best choice of scale and controls of review duplications and multiple accounts. The study in (Zhang et al., 2012) argued that the star rating average can be biased due to differences in individual grading standard (for instance, a lower score from a tough reviewer may still mean an object of good quality), again, the average score for an object with very few ratings is not statistically significant, and finally, many objects share similar scores, rendering such a rating system meaningless. Users' feedback usually contains rich textual reviews in addition to numerical ratings (Tan et al., 2016). The use of rating data for the ranking of product or service quality can be greatly improved by combining both the rating and the associated review text data. The next section will be focused on the mining themes and topics from textual user feedback.

2.5. Topic Mining

Rating data tells whether a user likes or dislikes a product or service, the textual review associated with the rating, however, is capable of explaining the users' concerns (Tan et al., 2016). Given a large collection of user feedback, an interesting research challenge is to figure out what the users are saying and why; these are embedded in the hidden themes and topics in the data. Topic mining provides a convenient way to perform unsupervised classification of text documents (Julia, 2017). According to David M. Blei, the seminal work on LSA in (Deerwester et al., 1990) launched topic modelling research. Topic models are mixed-membership models that extend and build on mathematical techniques and natural language processing methods such as the unigram and mixture of unigram models (Hornik and Grün, 2011). A well accepted practice is to explain the generation of each text document with a probabilistic topic model (Blei, 2013). Given a text document collection, the general assumption is that it was generated according to some model, the task or challenge, therefore, is to output the model that has generated the text. User feedback can be described as a sample of mixed topics, as such decoding the topics contained therein can be useful in many applications including criminal intelligence.

Topic Mining with LSA: basically, LSA finds a low dimensional representation of documents and words in a corpus, it maps words in documents into a concept space as illustrated in Fig. 2.3, the documents are first represented as BOW such

that word that usually appears together are represented as word patterns or concepts.

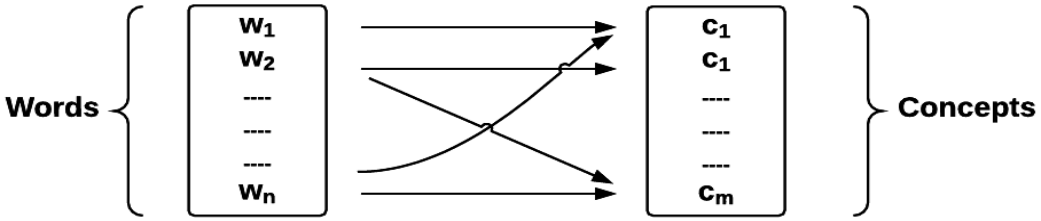


Fig. 2.3. Concept mapping of words in documents

Given an $m\ X\ n$ row-column matrix representation of documents denoted by A, LSA tries to reduce the number of rows (terms) while preserving the similarity structure among columns (documents) in the matrix. This task is called low-rank approximation and is achieved using methods of linear algebra such as SVD. Given that $B = A^T A$ is the document-document $(m\ X\ m)$ matrix while $C = AA^T$ is the term-term $(n\ X\ n)$ matrix with both B and C as square and symmetric, SVD uses matrices B and C to decompose A into three matrices represented by equation (2.16), where S is the matrix of the eigenvectors of B, U is the matrix of the eigenvectors of C, Σ is the diagonal matrix of the singular values which also refers to the square roots of the eigenvalues of B. According to (Thomo, 2009), LSA on A is represented as follows:

$$A = S\Sigma U^T \tag{2.16}$$

This low-rank approximation is then followed by a k-rank approximation that decomposes A to a k-dimensional space A_k as in equation (2.17) where small singular values are ignored, keeping only k entries (first k columns of S and U and first k columns and rows of Σ). It therefore, follows from (Thomo, 2009) that:

$$A_k = S_k \Sigma_k U_k^T \tag{2.17}$$

Apart from SVD, LSA also uses Non-negative Matrix Factorization (NMF) for the low-rank matrix approximation (Stevens et al., 2012). Some of the applications of LSA include search, text clustering and document classification.

Probabilistic Topic Mining: as explained in the tutorial in (Zhai, 2017), probabilistic topic mining tasks involve discovering k topics in a document and then filtering which topics are covered and to what extent in each documents (see Fig. 2.4).

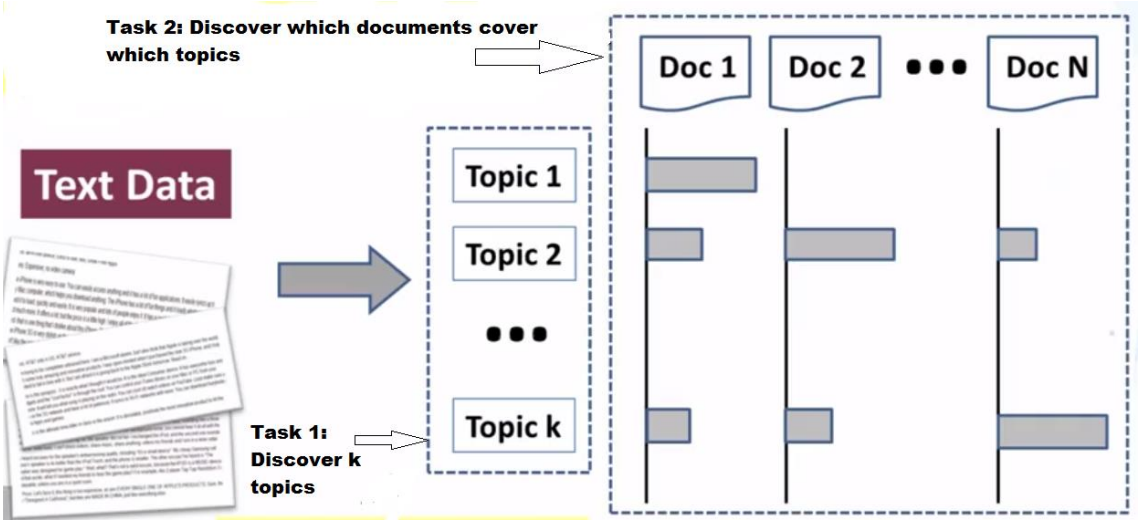


Fig. 2.4. Topic mining and analysis tasks (Zhai, 2017)

Given a collection of N text documents $C = \{d_1, d_2, \ldots, d_N\}$, one can use topic analysis to learn the different distributions of say k $topics$: $\{\theta_1, \theta_2, \ldots, \theta_k\}$ in the document collections. The coverage of topics in each document d_i is given by $\{\pi_{i1}, \pi_{i2}, \ldots, \pi_{ik}\}$, where $\sum_{j=1}^{k} \pi_{ij} = 1$, while π_{ij} is the probability of document d_i covering topic π_j. There are many strategies for defining $\theta_1, \theta_2, \ldots, \theta_k$, one basic approach is using terms as topics. The overall mining of the terms will then involve the parsing of all the text documents in C in order to obtain candidate terms followed by the design of a method for measuring topic coverage, which is how good each term is as a topic in a document. A simple approach of computing topic coverage is to simply count and normalise the occurrences of these terms so that the coverage of each topic in the document add to one, this according to (Zhai, 2017) is achieved using equation (2.18) for L documents.

$$\pi_{ij} = \frac{count(\theta_j, d_i)}{\sum_{L=1}^{k} count(\theta_L, d_i)} \tag{2.18}$$

The next and final step is to pick k terms with the highest scores. The main limitation of this approach is that there is only one term to describe a topic as such lacks the power to describe complicated topics. This limitation is addressed using probabilistic models that employ weighted multiple words as topics. According to (Zhai, 2017), the output in this case is a set of topics each with a word distribution as in equation (2.19), where $\forall_j \in [1,k]$, $\sum_{w \in V} p\left(w \middle| \theta_j\right) = 1$, while $\forall_i \in [1,N]$, $\sum_{j=1}^{k} \pi_{ij} = 1$.

$$\Lambda = (\{\theta_1, \theta_2, \dots, \theta_k\}, \{\pi_1, \pi_2, \dots, \pi_{1k}\}, \dots, \pi_{N1}, \pi_{N2}, \dots, \pi_{Nk}) \qquad (2.19)$$

The generative model in equation (2.20) is one method of solving this problem. A generative model $P(Data|Model, \Lambda)$, is a principle way of using statistical modelling to solve topic and other text mining problems. Given a natural language text data, the model utilises equation (2.20) described in (Zhai, 2017) to infer the most likely parameter values Λ^*.

$$\Lambda^* = argmax_{\Lambda} P(Data|Model, \Lambda) \qquad (2.20)$$

The simplest statistical language model is the unigram language model which generate text by generating each word independently. Basically, there are two problems to think about in terms of a statistical language model. First, given a model, how likely are we to observe a certain kind of data point? That is, we are interested in the sampling process. The other problem is called the estimation process, and it involves thinking of model parameters given some observed data. LDA and PLSA are some of the popular approaches to probabilistic topic modelling, in practical terms, both LDA and PLSA perform similarly for many tasks (Zhai, 2017). These two processes are accomplished differently in different topic modelling approaches and algorithms. The output of topic models can enable many applications such as clustering and can further associate topics with different contexts such as locations, authors, sources, and time (Blei, 2013). The non-textual components of user feedback data provide more context for analysing patterns in the topics. These components are usually called metadata and include among others, the time, locations, authors, or sources of the data. Such interesting patterns include topics in different locations, trending and fading

topics (Julia, 2017). The implementations of the topic modelling algorithms in R and Python will be utilised in order to discover hidden topics in textual user feedback.

2.6. Sentiment Analysis

The major elements of a sentiment analysis include, the sentiment or opinion expressed, the opinion target, the time of opinion expression, the opinion holder, the reason for opinion expression, the opinion qualifier, and the opinion types (Liu, 2012). Document, sentence, and aspect are the three commonly studied sentiment analysis levels or discourse granularity. The aspect or entity level analysis forms the core of sentiment analysis tasks as it aims to identify and measure the atomic unit of information contained in a sentiment expression. The classical subtasks of sentiment analysis include sentiment classification at different discourse (aspect sentence, and document) levels, opinion summarisation, opinion search and retrieval, emotion identification, aspect, and lexicon extraction (Liu, 2015).

Sentiment Analysis Tasks

The framework for representing a sentiment analysis problem at the aspect level, defined in (Zhang and Liu, 2014), is the quintuple in equation (2.21).

$$\left(e_i, a_{ij}, s_{ijkl}, h_k, t_l\right) \tag{2.21}$$

where e_i is the name of an entity, a_{ij} is an aspect of e_i, s_{ijkl} is the sentiment on aspect a_{ij} of entity e_i, h_k is the sentiment holder, and t_l is the time when the sentiment is expressed by h_k. The framework reduces to a quadruple when the entity is itself the aspect, hence e_i and a_{ij} together represent a single element called sentiment target. An entity e_i is represented as a finite set of aspects $\{a_{i1}, a_{i2}, \dots, a_{in}\}$, while a sentiment document d contains sentiment expressions on a set of entities $\{e_1, e_2, \dots, e_r\}$, a subset of their aspects from a set of opinion holders $\{h_1, h_2, \dots, h_p\}$, at some particular time point. The fundamental sentiment analysis tasks involves the identification and measurement of the type,

orientation, and intensity of a sentiment expression. In general, given a sentiment document d, we aim to discover the quintuples in equation (2.21) over d.

- **Entity Extraction:** an entity e represent a product, service, person, event, organisation, or topic that appears in an entity expression. It is represented as a pair e: (T, W), where T is its components or sub-components while W is set of attributes of T. The task here is to extract all entity expressions in d, and categorise or group synonymous or entity expressions into entity clusters (or categories) such that each entity expression cluster indicates a unique entity e_i.
- **Aspect Extraction:** The aspects of an entity e, are the components and attributes of e that appear in an aspect expression and can be explicit (usually nouns and noun phrases) or implicit (verbs, verb phrases, adjectives, and adverbs). The task here is to extract all aspect expressions of the entities, and categorise these aspect expressions into clusters such that each aspect expression cluster of entity e_i represents a unique aspect a_{ij}. Popular aspect extraction approaches, detailed in (Liu, 2012) and (Zhang and Liu, 2014), are based on language rules, sequence models, topic models, and supervised learning. The language rules method capture the various contextual properties of terms and their relations in a text. Aspect extraction can also be considered as a sequence labelling task which can be achieved using Hidden Markov Model, Conditional Random Fields, or any other sequence labelling model. Topic models such as the PLSA and LDA specifies a probabilistic procedure by which documents can be generated as such can be extended for sentiment aspect generation by regarding aspects as latent topics in sentiment documents.
- **Sentiment Holder Extraction**: the task here is to extract sentiment or opinion holders for sentiments or opinions from text or structured data and categorise them such that each cluster represents similar sentiment holders.

- **Time Extraction and Standardisation**: the task here is to extract the times when sentiments were given and standardise different time formats.
- **Sentiment Scoring or Estimation**: The task for a polarity estimation is to assign a numeric sentiment rating to an aspect a_{ij} and then determine whether a sentiment on the aspect is positive, negative or neutral based on a predefined decision boundary. This task is commonly approached either with linguistic or machine learning techniques. The linguistic constructs require domain lexicon while the machine learning methods are basically a mix of feature engineering and supervised learning on text data.

Recent studies on sentiment analysis indicated that it's beyond the measure of subjectivity, it is being extended to automatically quantify affect, emotion, mood, conditional expressions, sarcasm, sentiment composition, and negation in multimedia content; sentiment analysis also embeds the following emerging subtasks, mining of intentions, debates, and discussions, deceptive opinion detection, and review quality monitoring (Liu, 2015). Sentiment analysis can also be used to understand what words with emotional or opinion content are important for a particular text and how a narrative arc changes throughout its course; however, sentiment analysis of news and social media still remains challenging, this is because most previous approaches were proposed for product reviews (Julia, 2017). The next section will report our proposed framework detailed in Appendix III for ensuring product safety using sentiment analysis.

2.7. Product Safety Framework

Traditional crime data mining can only attempt to discover what is going on as far as digital crime is concern, but the whys and wherefores remain buried in an unstructured content that represents the hidden story behind the data; some of these unstructured data include large volumes of news stories, social media discussions, user reviews, notes, chats, web forms, emails, documents, voicemail, texts, blogs, forum discussions, regulatory filings, crime records, presentations, and invoices; automated frameworks are needed in order to

capture and analyse these and other natural language data for crime intelligence purposes; text mining, a specialised area of data mining that deals with text data, provides an automated solution for organising key concepts buried in these unstructured content in order to discover previously unknown patterns and concepts that may be useful to investigators; it involves the analyses of large amounts of the unstructured data and includes such tasks as taxonomy categorisation, concept clustering, topic analysis, entity extraction, sentiment analysis, and summarisation; text mining tools can also be used to convert unstructured content and parse it over to a structured format that is amenable to inductive and deductive learning (Mena, 2011).

The highly dynamic nature of online pharmacies with continuously appearing and disappearing entities and services are additional challenges that call for the development of robust and efficient trust and reputation mechanisms. People are now turning to the Internet to buy things because of convenience and cheaper options, however, in bilateral interactions involving risk such as purchasing medications online, a transaction shouldn't take place unless the party who moves first has a certain level of trust that the second party is willing to fulfil its obligations. Online trust and reputation systems are soft security mechanisms that allow users to provide credibility rating to a service provider; agents rely on reputation information to choose trade or coalition partners (Jøsang, 2008). Trust, in the context of this study, implies a decision by an agent to rely on someone or something. Trust and reputation systems capable of scoring or ranking pharmaceutical entities (such as brands, products, and websites) into predefined categories such as recommended or not recommended can provide evidence about whom patients and customers can trust to purchase pharmaceutical products as well as to share and accept information without additional verification.

The product safety framework in Fig. 2.5 is a novel application of text mining and sentiment analysis in harnessing and processing the views and experiences of consumers of popular brands of drug and cosmetic products reported on social media platforms. The application came about mainly due to the usefulness of social data analysis techniques, especially the successes of topic mining and sentiment analysis on social media data. The health sector

consists of professional health service providers, drug manufacturers, regulatory and enforcement agencies, etc. People who are seeking medical products, advice and care often find it difficult to obtain reliable information about the quality and competence of health service providers (Mosadeghrad, 2014). Feedback generation, aggregation and processing in the pharmaceutical marketplaces still remains a challenge, unlike those of traditional marketplaces such as EBay and Amazon which gather and rank buyers and sellers based on comments from each other after each transaction.

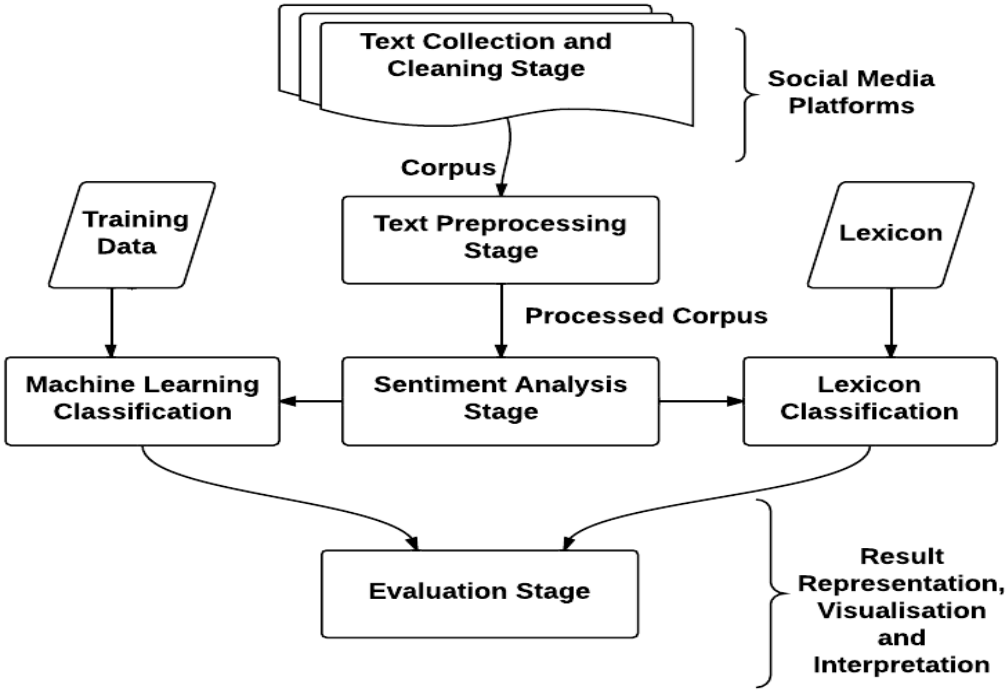


Fig. 2.5. Architecture of the proposed product safety framework

The framework is an innovative application of reputation computation pipeline and is in line with an effort to uncover patterns of criminal activities as well as the desire to predict when and where crimes are likely to occur in the future. It involves the extraction, conversion, and modelling of unstructured data in order to advance pharmaceutical crime investigation as well as to improve real-time crime detection systems. According to (Josang and Ismail, 2002), the two fundamental features that every reputation system should possess are: (i) reputation engine, which ranges from simple numerical adder to complex mathematical equations for calculating the value of reputation and (ii) propagation mechanism (centralised or distributed), which allow entities to obtain reputation

values. A measure of trust and reputation of pharmaceutical entities from user opinions using sentiment analysis and expert inputs can serve as the reputation score of a target entity and can be useful in safeguarding the public health. The proposed framework is composed of extractive, inductive, and deductive phases and its architecture are presented in Fig. 2.5. It comprises four stages: text collection and cleaning; pre-processing; sentiment analysis and finally evaluation.

Text Collection: at the text collection and cleaning stage, an API call for authentication and data extraction is invoked on Facebook Graph and Twitter APIs. The framework is designed to accommodate virtually all available social media APIs, but due to the dynamic nature of these APIs and for the purpose of this report, greater focus will be given to Twitter and Facebook. The Twitter API consists of the REpresentational State Transfer (REST) and Streaming APIs (Kevin, 2009). The REST API provides methods for authenticating applications, processing requests, handling imposed limits, etc. The Streaming API provides client applications with Twitter's global stream (public, user and site) of data. The Facebook Graph API provides the means of getting data into and out of the Facebook social graph. The framework employs both the REST and Streaming APIs for searching and fetching Tweets, while the Facebook Graph API is used for fetching pages, status updates and comments suggesting user experiences and views on drugs and cosmetic products; the collected text is noisy and in JavaScript Object Notation (JSON) format and methods for cleaning and parsing of the data to form a corpus, i.e. a collection of comments and Tweets, are incorporated for further processing.

Text Pre-processing: the corpus is then transformed into feature vectors; the BOW representation was used due to its simplicity and because preserving the order of the features in the corpus is not of particular interest in the application. A simple feature selection or pre-processing method to transform or tokenize the text stream to words was followed; these methods constitute a sequence of the following tasks; removing delimiters, converting all words to lower case, removing numbers and stop words, stemming words to their base and some application or domain specific feature transformations. The tokens are then represented as BOW sparse matrix using the $TF - IDF$ weighing scheme. Given a corpus C,

containing N documents defined as d_i, where $i = 1 \ldots N$, and Tweets tokenized as words or terms t. The $TF - IDF$ weighting scheme takes into account the relative importance of the word in the document described in (Manning et al., 2008) and assigns to term t_j a weight in document d_i given by:

$$TF - IDF(t_j,\ d_i) = TF(t_j,\ d_i) * IDF(t_j) \tag{2.22}$$

where $TF(t_j,\ d_i)$ denotes term frequency, the number of word occurrences in a document; $IDF(t_j) = log2\left(\frac{N}{df t_j}\right)$ denotes inverse document frequency, with $DF(t_j)$ representing the number of documents containing the word. Further sparseness is handled by selecting terms that appear in a minimum number of documents. The resulting bipartite representation is the input on which further tasks are performed.

Sentiment Scoring: this task was approached by employing both lexicon and machine learning methods. In the lexicon based approach, beside the corpus, a fundamental requirement is a pre-labelled word list or polarity lexicon. For an improved classification result as described in (Nathan and Richard, 2014), the framework merges two lexicons, application or domain specific preassembled lexicon and a generic English based lexicon developed and is being maintained by the authors of (Hu and Liu, 2004). Another requirement for the lexicon based classifier is a sentiment scoring function, of which there are several options, one of the most basic polarity computational scheme was described in (Gaston, 2012); all the words in the corpus or target collection are compared to the words in the lexicon, the overall sentiment score of the corpus or a subset will then be the difference between the numbers of positively and negatively assigned words. Therefore, the associated polarity score for each comment or Tweet in the corpus is given in (Hu and Liu, 2004) as:

$$Score = \sum_i^n pw\ - \sum_j^m nw \tag{2.23}$$

where pw and nw denote positive and negative words respectively. A comment or Tweet has an overall (ii) positive sentiment if $Score > 0$, (ii) neutral

sentiment if $Score = 0$, and (iii) negative sentiment if $Score < 0$. The total score for the corpus is then visualised and evaluated with simple descriptive statistics i.e. histogram and box plot. A more advanced scoring scheme include fuzzy reasoning, which is a computational intelligence technique that can be applied to improve the text classification and clustering tasks. In the machine learning based approach, beside the corpus, the fundamental requirement is a training dataset, already coded with sentiment classes. The classifier is trained or modeled with the labeled data such that new but similar documents are tested with the resulting model to have it predict the direction of the sentiment of the new documents. For the purpose of this report, Naive Bayes is used as a baseline classifier because of its efficiency as reported in (Manning et al., 2008). Feature words are assumed to be independent and then each occurrence was used to classify Tweets or comments into its appropriate sentiment class; this is called multinomial event model. It, therefore, follows from (Manning et al., 2008, Isah et al., 2014) that the classifier which utilises the maximum a posteriori decision rule can be represented as:

$$c_{map} = \underset{c \in C}{\operatorname{argmax}}\big(P(c|d)\big) = \underset{c \in C}{\operatorname{argmax}}\big(P(c)\prod_{1 \le k \le n_d} P(t_k|c)\big) \quad (2.24)$$

where t_k denotes the words in each Tweet or comment and C the set of classes used in the classification; again, $P(c|d)$ is the conditional probability of class c given document d, $P(c)$ is the prior probability of class c and the $P(t_k|c)$ conditional probability of word t_k given class c. It also follows from (Manning et al., 2008), that in order to estimate the prior parameters, equation (2.24) is then reduced to:

$$c_{map} = \underset{c \in C}{\operatorname{argmax}}\big(\log P(c) + \sum_{1 \le k \le n_d} \log P(t_k|c)\big) \quad (2.25)$$

To handle zero probabilities that may arise when a word does not occur in a particular class, $(TF - IDF)$ weighing or Laplace smoothing by adding 1 to each count is employed; with Laplace smoothing, $P(t|c)$ as described in (Manning et al., 2008) becomes:

$$P(t|c) = \frac{T_{ct}+1}{\sum_{t\prime\in V}(T_{ct\prime}+1)} + \frac{T_{ct}+1}{(\sum_{t\prime\in V} T_{ct\prime})+B\prime} \quad (2.26)$$

where B' refers to the number of terms contained in the vocabulary V. While the lexicon based sentiment classification result is evaluated with reference to a ground truth or by human judgment, contingency or truth tables are used to represent the output of the machine learning approach such that a baseline result can be used for performance comparison. The evaluation of the proposed framework will be carried out in the next section using datasets from Twitter and Facebook.

2.8. Experimental Work

Product manufacturers and service providers are beginning to utilise their customer feedback in order to track and manage satisfactions, complaints, and suggestions. These feedback as described in section 2.2 comprises of publicly available aspect ratings, review text, and reviews of review. Mining these feedback can provide a useful knowledge for understanding and predicting crimes. This experimental work will begin by first measuring the true quality of products using aspect rating data. It will then demonstrate how social media feedback can be useful in inferring rogue vendors of pharmaceutical and cosmetic products.

2.8.1. Datasets

Four datasets will be used throughout this chapter, the first dataset, denoted as D^1, is an archive of more than half a million rating and text reviews of fine foods from amazon spanning a period of more than 10 years, the second dataset, denoted as D^2, is a user feedback data regarding online pharmacy vendors and products, the third dataset, denoted as D^3, is a user feedback data that was collected on Facebook which comprises of the opinions of some cosmetic products and vendors, finally, the fourth dataset, denoted as D^4, is a user feedback data collected on Twitter which comprises of Tweets gathered based on pharma and medical related query combinations.

- $\boldsymbol{D^1}$: is a user feedback data on fine foods described in (McAuley and Leskovec, 2013b), it has the following variables, Id, unique identifier for the product (ProductId), unique identifier for the user (UserId),

ProfileName, number of users who found the review helpful (HelpfulnessNumerator), number of users who indicated whether they found the review helpful (HelpfulnessDenominator), rating between 1 and 5 (Score), timestamp for the review (Time), a brief summary of the review (Summary), and text of the review (Text).

- $\boldsymbol{D^2}$: is a user feedback of online pharmacy vendors and products extracted from Pharmacy Reviewer[15], it is a 5-star rating data having the following variables, the mean of the five aspect ratings (ave_rating), the measure of the site reliability with a weighting value of 40% (product_quality), ease of website use and range with a weighting value of 5% (website_usability), price satisfaction with a weighting value of 30% (price), customer service satisfaction with a weighting value of 10% and optional (cust_service), speed of delivery with a weighting value of 15% (delivery), the review text (review), number of comments for each review (comments_count), number of helpful vote for each review (helpful_vote), the online vendors name (pharmacy). Only reviews from three top rated vendors (Generic Doctor, Indian Pharma, and Meds-Easy) and one blacklisted vendor (Cheap Scrips) was extracted for comparative analysis. Undelivered orders was the main reason for blacklisting Cheap Scrips.
- $\boldsymbol{D^3}$: is a user feedback data we gathered from Facebook pages of 3 popular brands of drug and cosmetic related products (Avon, Dove and OralB) using the Facebook Graph API. For privacy issues, the brand names are randomly coded as Brand X, Brand Y and Brand Z. We initially retrieved the public contents of the targeted pages and then extracted user comments and opinions from popular posts suggesting product advertisement that (1) do not offer a prize in return (Brand Y and Z) and (2) offer a prize in return (Brand X).
- $\boldsymbol{D^4}$: is a user feedback data (about 11,431 Tweets) we gathered from Twitter API using combinations of the following query terms: “medicine”, “prescription”, “over the counter”, “side-effects”, “online pharmacy” and “antibiotics”.

[15] https://pharmacyreviewer.co/

2.8.2. Exploration and Aggregation of Rating Data

A prospective customer of a product will be able to read hundreds of relevant online reviews and determine which ones seem the most genuine and informative, but this is not easy especially when there are several thousand or even millions of reviews. This section demonstrates how data mining can be utilised in exploring and aggregating millions of customer reviews.

Rating Data Exploration: the dataset D^1 was used here in order to explore and understand the nature of rating data. The excellent work of Dr. Rob Castellano in (Castellano, 2016) demonstrates the exploration of user feedback. Looking at the distribution of ratings among all the reviews, we can see from the plot in Fig. 2.6 that the 5-star reviews constitute a large proportion of all reviews. There are 52268 (9%) reviews with one star rating, 29769 (5%) reviews with two star rating, 42640 (8%) reviews with three star rating, 80655 (14%) reviews with four star rating, and 363122 (64%) reviews with five star ratings. Looking at the distribution of ratings among all the reviews, we can see from the plot in Fig. 2.6 that the 5-star reviews constitute a large proportion of all reviews. There are 52268 (9%) reviews with one star rating, 29769 (5%) reviews with two star rating, 42640 (8%) reviews with three star rating, 80655 (14%) reviews with four star rating, and 363122 (64%) reviews with five star ratings. Looking at the distribution of ratings among all the reviews, we can see from the plot in Fig. 2.6 that the 5-star reviews constitute a large proportion of all reviews. There are 52268 (9%) reviews with one star rating, 29769 (5%) reviews with two star rating, 42640 (8%) reviews with three star rating, 80655 (14%) reviews with four star rating, and 363122 (64%) reviews with five star ratings. Looking at the distribution of ratings among all the reviews, we can see from the plot in Fig. 2.6 that the 5-star reviews constitute a large proportion of all reviews. There are 52268 (9%) reviews with one star rating, 29769 (5%) reviews with two star rating, 42640 (8%) reviews with three star rating, 80655 (14%) reviews with four star rating, and 363122 (64%) reviews with five star ratings Looking at the distribution of ratings among all the reviews, we can see from the plot in Fig. 2.6 that the 5-star reviews constitute a large proportion of all reviews. There are 52268 (9%) reviews with one star rating, 29769 (5%) reviews with two star rating, 42640 (8%) reviews with three star rating, 80655 (14%) reviews with four star rating, and 363122 (64%) reviews with

five star ratingsLooking at the distribution of ratings among all the reviews, we can see from the plot in Fig. 2.6 that the 5-star reviews constitute a large proportion of all reviews. There are 52268 (9%) reviews with one star rating, 29769 (5%) reviews with two star rating, 42640 (8%) reviews with three star rating, 80655 (14%) reviews with four star rating, and 363122 (64%) reviews with five star ratings.

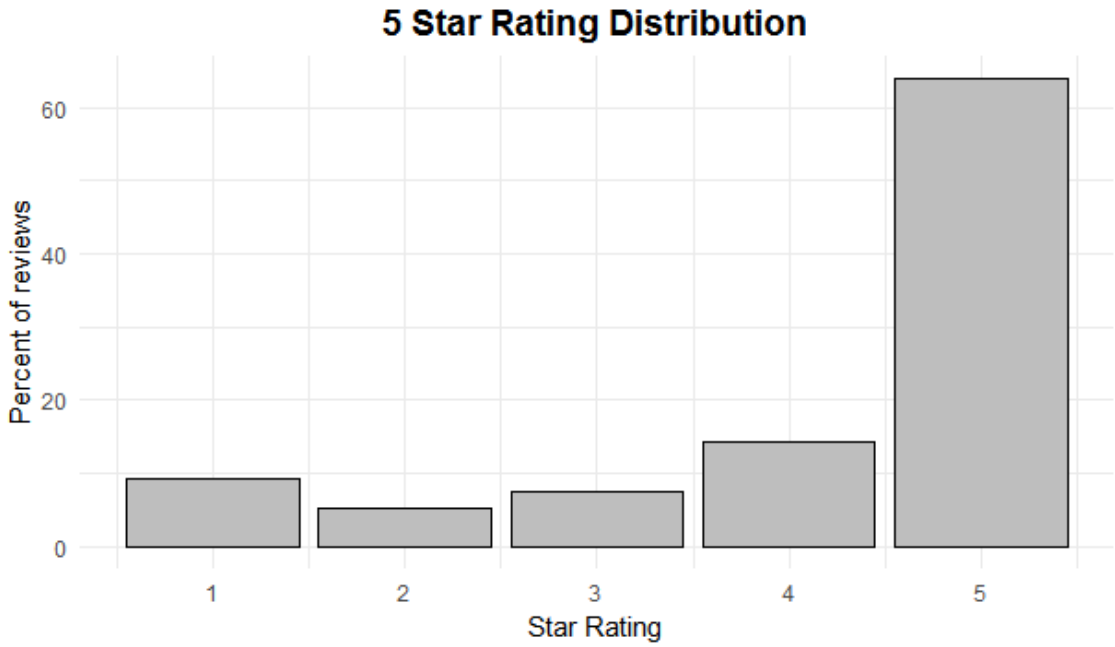


Fig. 2.6. Distribution of 5-star user ratings

TABLE 2.6 is a preview of two samples each from reviews with a score of five and one respectively. Some products or items have many reviews while others have only one review, the item with highest review count has 913 reviews. Some users have rated many items while others have rated only one item, the user with highest review count has rated 448 items. But how can these items be ranked based on their aggregate rating data?

Rating Data Aggregation: as pointed out in section 2.4, Bayesian estimation is superior to mean estimation method when it comes to the use of rating data to rank the quality of a product or service. Bayesian methods are widely applied in recommender services and other predictive algorithms that use star reviews and some other preference space measures, it is also used to estimate or refine star rating predictions. The excellent work on the computation of Bayesian mean of

movie rating in (Bengfort, 2014) will be adapted here. Again, dataset D^1 was utilised (because of its relatively large size) in order to understand rating data aggregation and to explore the top rated products in the dataset. A simple exploration of the quality ranking of randomly selected ten products using the average rating estimation method in equation (2.10) was performed.

TABLE 2.6. Fine food user review sample

Summary	Text	Score
Good Quality Dog Food	I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The product looks more like a stew than a processed meat and it smells better. My Labrador is finicky and she appreciates this product better than most.	5
Great taffy	Great taffy at a great price. There was a wide assortment of yummy taffy. Delivery was very quick. If your a taffy lover, this is a deal.	5
Bad	I fed this to my Golden Retriever and he hated it. He wouldn't eat it, and when he did, it gave him terrible diarrhea. We will not be buying this again. It's also super expensive.	1
Nasty No flavor	The candy is just red , No flavor . Just plan and chewy . I would never buy them again	1

As reported in (Bengfort, 2014), the results (see TABLE 2.7) quickly points to a problem with this method, high average scores were given to products with very few reviews (for instance, an average of 4.973451 for 113 review counts and an average of 4.906542 for 107 reviews).

TABLE 2.7. Product Quality Ranking: Average and Bayesian Methods

Product Id	Review Count	Average Ranking	Bayesian Ranking
B003B3OOPA	623	4.739968	4.629272
B001EO5Q64	567	4.746032	4.624797
B007R900WA	170	4.823529	4.465909
B000O5DI1E	107	4.906542	4.378981
B004EAGP74	389	4.802057	4.625285
B000GAT6NG	389	4.802057	4.625285
B001E8DHPW	389	4.802057	4.625285
B003QDRJXY	264	4.837121	4.584395
B000ED9L9E	113	4.973451	4.444785
B000NMJWZO	542	4.881919	4.744088

This means that using just a simple or naïve mean of rating data to measure product quality is not enough since the number of observations also contributes to the quality of a 5 star rating. This result calls our attention to a problem when there are few reviews for some products, hence the need for a rule that will allow us to rank products according to both the mean rating a vendor has received as well as the number of reviews the product has received. This notion is captured with Bayesian estimation, it allows us to forgo computing a direct value from a limited number of observations and instead creates a probability distribution that describes the observable space more completely. The probability distribution can then be utilised in order to come up with a better estimate for the aggregate rating computation. The probability distribution can also be used to assign confidences or bounds that can lead to better decision making. The application of the Bayesian ranking method in equation (2.15) resulted in an improved ranking (see TABLE 2.7). The Bayesian estimate embeds a prior and a likelihood (from observations) to the mean (Bengfort, 2014). So what lesson have we learnt?

Lessons and Limitations: feedback systems are valuable resources, as reported in both the introduction and literature review of this chapter, consumers make purchasing decisions based on the feedback of other consumers. However, these systems have been gamed by market forces, especially on newer products. Some manufacturers and marketers pay customers to write good reviews by giving reviewers their products and services for free or at a discount. A study of about 360,000 user ratings across 488 products on Amazon in (Kriti, 2017) discovered what is clearly an unrealistic proportion of 5-star ratings. Although not always the case, but according to Bin Liu[16], inflated or fake review is a common problem of review systems, it is also known by many names such as spam review, bogus review, deceptive review; the authors are often referred to as opinion spammers, review spammers, fake reviewers, or shill (stooge or plant), and there is a robust market that materially rewards them. The skewed distribution may be due to human use of rating systems or how the rating scale was implemented or placed on a website rather than being an indicative of fake reviews. Again not all reviews are equally valuable, there are clear and subtle differences between paid

[16] https://www.cs.uic.edu/~liub/FBS/fake-reviews.html

and organic reviews. Attempts are being made to reduce the impact of paid and bogus reviews, for instance, methods have been developed by Amazon to improve the relevancy of review rankings to ensure that reviews that could help shoppers the most are seen first; Professor Bin Liu and his team at the Computer Science department, the University of Illinois at Chicago have also been working for several years on a project titled “Opinion Spam Detection: Detecting Fake Reviews and Reviewers”. These efforts, however, are not yet to any great effect, for instance, despite being blacklisted in Pharmacy Reviewer, Cheap Scrips has its own review system (called Testimonials)[17] with the following title “REAL REVIEWS FROM REAL SHOPPERS” on its websites and at the time of writing this report there are currently a total of 45 reviews with an average of 5 out of 5 stars. The question now is who are the real shoppers? Those users who rated Cheap Scrips on Pharmacy Reviewer that resulting in it being blacklisted or those users who reviewed the vendor and gave it an excellent value of 5 on its website?

2.8.3. Textual Feedback Processing

Several properties and approaches for representing textual data were described in subsection 2.3.2, word-based representation was reported as the most preferred approach for natural language processing, hence will be utilised in this study. One of the most basic characteristics of a review is the number of words it contains. Fig. 2.7 is an illustration in (Castellano, 2016) of how word count varies with the rating. It was observed that 5-star reviews had the lowest median word count (53 words), while 3-star reviews had the largest median word count (71 words).

[17] http://www.cheapscrips.com/main/index.asp

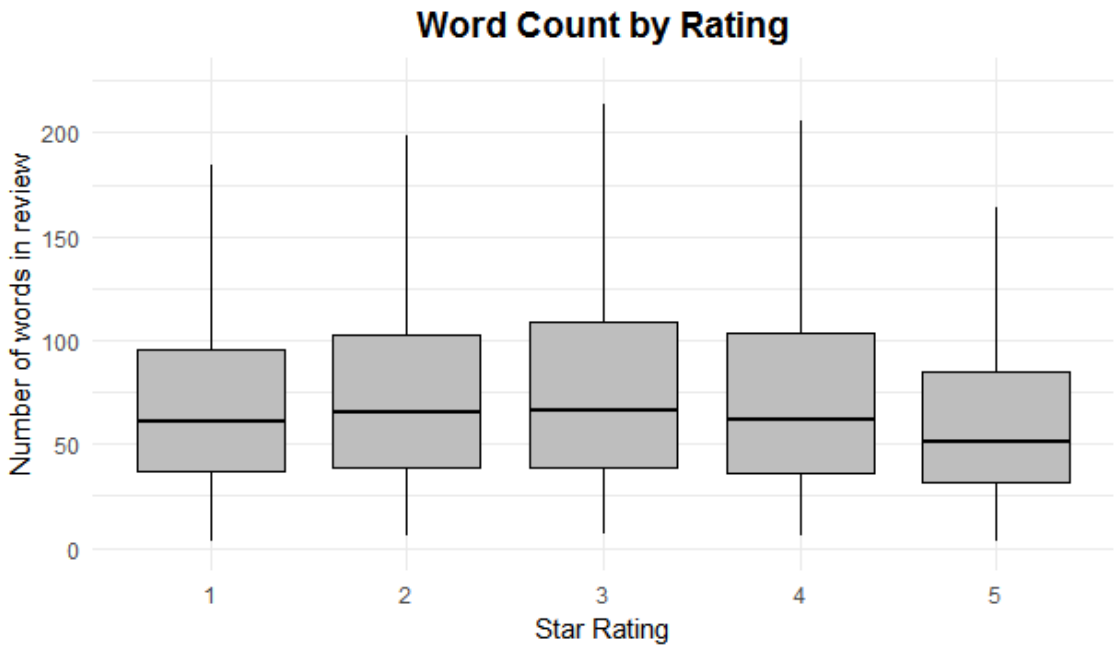


Fig. 2.7. Word count in user rating data by rating score

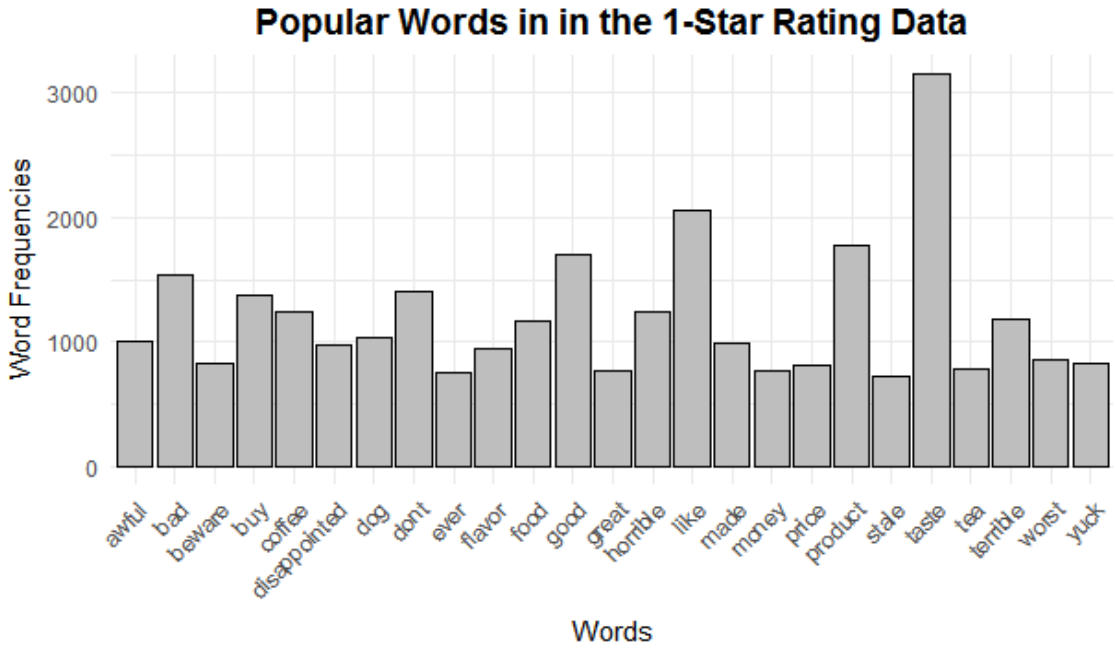


Fig. 2.8. Popular words in the 1-star rating data

A further exploration of the popular words in the summary or headline of each review revealed some important information regarding the concerns and experiences of the users. Some of the popular words in the 1-star review component (see Fig.2.8) that should be further investigated include awful, disappointed, beware, horrible, stale, terrible, worst, and yuck. In the 5-star

review component (see Fig. 2.9), popular words include awesome, best, delicious, excellent, favourite, healthy, love, perfect, tasty, wonderful, and yummy. There are however some words (good, great, like, and price) that appear in both components and may require further analysis such as negation or sarcasm detection.

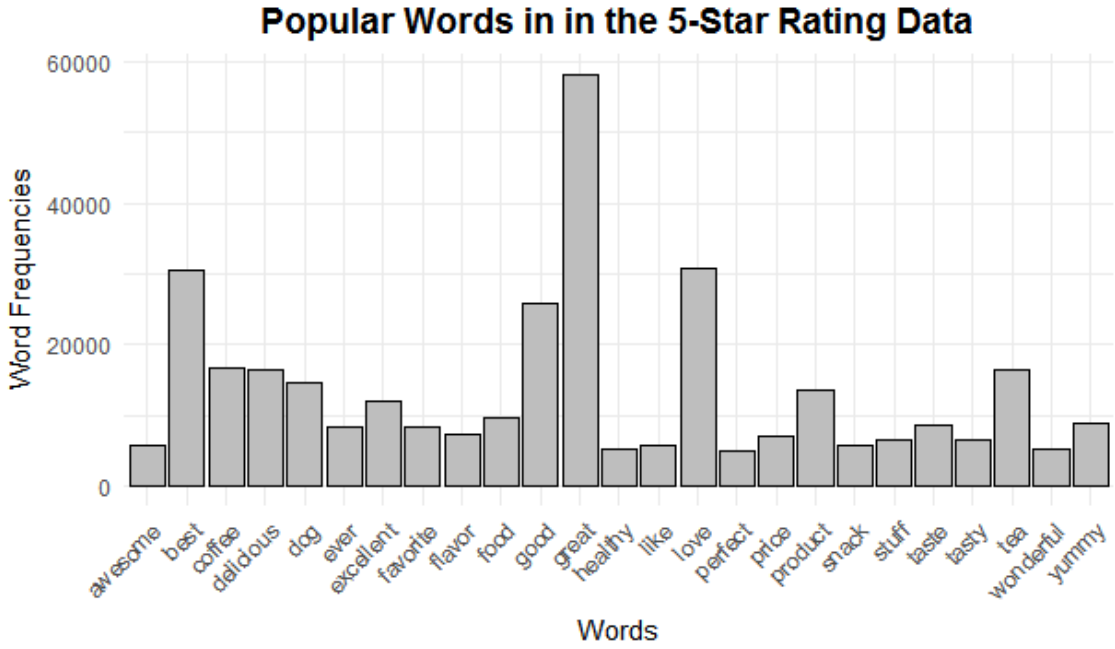


Fig. 2.9. Popular words in the 5-star rating data

Zipf's Law: Recall that in subsection 2.3.2, it was reported that Zipf's law is commonly observed in natural language corpus (text collection) and it states that the frequency that a word appears in a corpus is inversely proportional to its rank. It states in simple terms that if we plot the frequency of words versus their rank for a sufficiently large collection of textual data, we will observe a clear trend, which resembles a power law distribution. The law was examined using D^2, the Pharmacy Review dataset. From the resulting plot in log-log coordinates (see Fig. 2.10), a very long tails to the right was observed for all the four online pharmacy vendor reviews, most of the words occur rarely while fewer words occur frequently. It therefore, follows that the word distributions of most natural language corpus can be approximated with a Zipfian distribution or any of its close families such as the power law distributions.

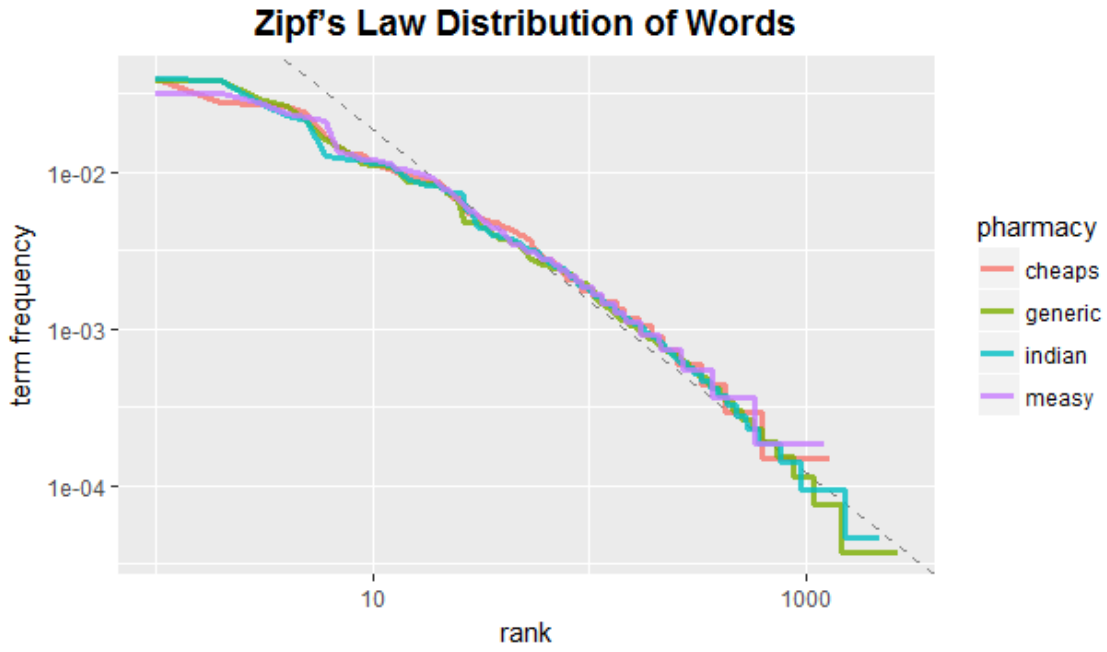


Fig.2.10. Zipf's law distribution for the four vendors

Word importance, co-occurrence and language models: as pointed out in subsection 2.3.2, a more generalised and common approach for text data representation is to use VSM which combines BOW or unigram model and TF-IDF weighting scheme**.** This is the simplified VSM representation of text data, the syntax and even the order of words is ignored. The TF-IDF weighing expressed mathematically in equation (2.8), is a useful measure for evaluating word importance in a text and is used in VSM to filter less important words such as stop words.

VSM of BOW or Unigrams and TF-IDF: an implementation of TF-IDF (bind_tf_idf) in the R tidytext package[18] was utilised in order to study the most important terms in the review text. There are many libraries and functions that implement BOW and TF-IDF, we opted for tidytext because of its flexibility to convert input data from and to other formats, and it also lends itself to quality visualisation. The resulting features (see Fig. 2.11) are the most important terms in each of the four vendor review documents in the dataset. Following a similar argument in (Julia, 2017), what measuring TF-IDF has done here is to show us that the reviewers used similar language, what distinguished one online pharmacy vendor from the others in the collection are the proper nouns (names

[18] https://cran.r-project.org/web/packages/tidytext/index.html

of people and places). This is obviously the goal of TF-IDF, it identifies words or terms that are important to a document within a collection of documents. Had it been that only term frequency (TF) was used, we would have gained an insight into how language is used in a natural language corpus but using TF-IDF enabled us to find words that characterise a specific document within document collections.

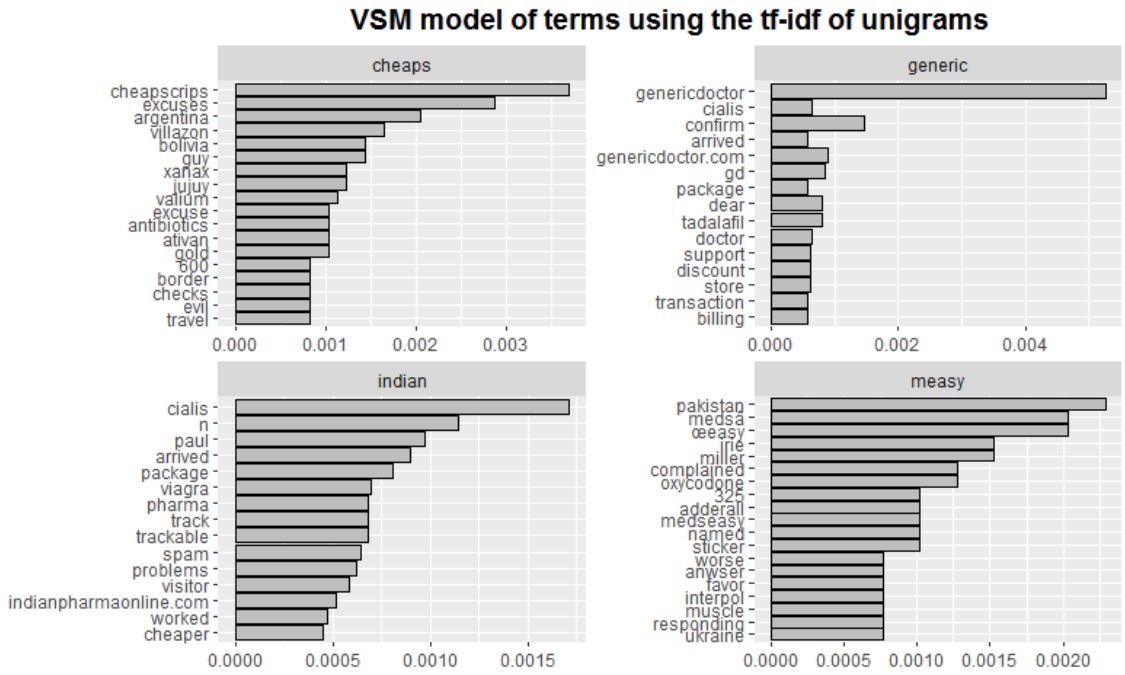


Fig. 2.11. VSM representation of reviews with TF-IDF of unigrams

VSM of Bigrams and TF-IDF: according to (Julia, 2017), many interesting text analyses, however, are based on the relationships between words, whether examining which words tend to follow others immediately, or which words tend to co-occur within the same documents. The n-gram model represented in equation (2.6) is a set of n words that co-occur in a text, it is often used as a feature to represent a text collection. We reported that the BOW (represented in equation (2.3)) is a special n-gram, also called 1-gram representation of a textual data in which the word or terms order was ignored. Popular n-gram representations include the 2-gram or bigram (represented in equation (2.4)), and the 3-gram or trigram (represented in equation (2.5)). It also follows from the same source that there are merits and demerits of examining the TF-IDF of pairs of consecutive terms rather than individual words, extracting higher n-grams such as bigram and

trigram rather than unigram features might capture further structure and provide context that makes natural language tokens more understandable. However, a typical two-term pair is rare than either of its component terms hence data representation with bigram are often very sparse. Bigrams and other higher n-grams are however superior to unigrams and more useful when processing a very large text corpus.

Again, an implementation of TF-IDF (bind_tf_idf) in the tidytext package in R was used. The interest here is to understand word relationships such as how often a word is followed by another, recall the discussion on Markov property in section 2.3 where it was stated that n-gram models are defined by approximating the conditional probabilities that only depend on the previous $n-1$ words in the sequence. The "n-grams" is an attribute in the function unnest_tokens() in the tidytext package and it takes n as the number of words (n=1 for unigram, n=2 for bigram, and so on) we wish to capture in each n-gram. Some interesting bigrams were observed in the results (see Fig. 2.12), for instance, "ship date" in Cheap Scrips (cheaps), "expiration dates" in Indian Pharma (indian), "repeat customer" in Generic Doctor (generic), and "stopped responding" in Meds-Easy (measy). These important word sequence would have been distorted in the unigram model.

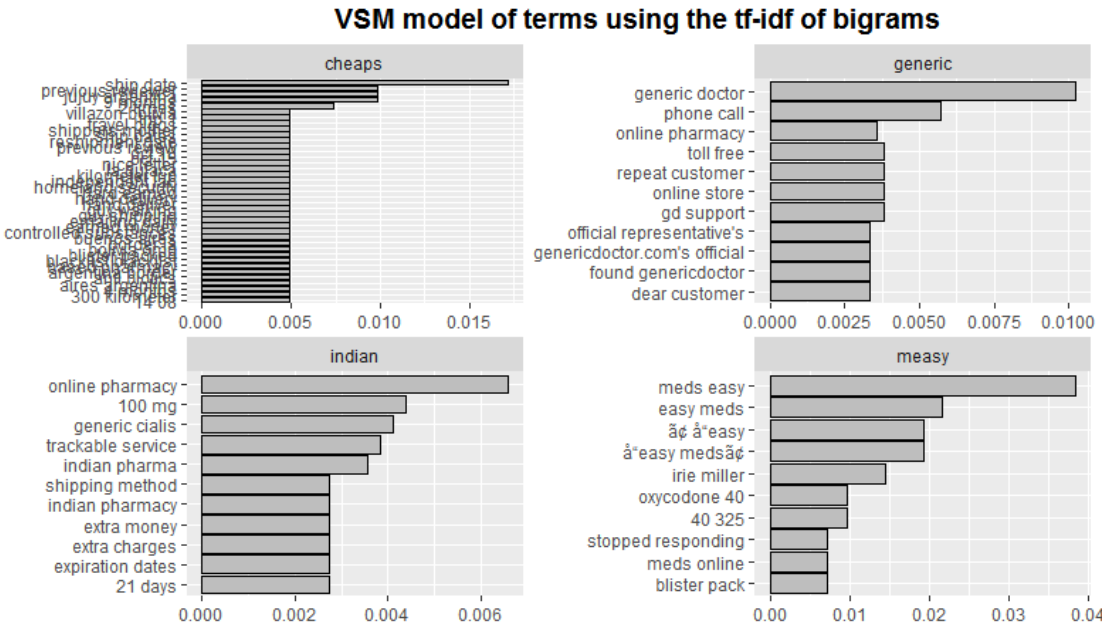


Fig. 2.12. VSM representation of reviews with TF-IDF of bigrams

As depicted in the result, tokenising by n-gram is a useful way to explore pairs of adjacent terms, it is also useful in describing terms that tend to co-occur within particular documents or vendor reviews even if they don't occur next to each other (Julia, 2017).

Word Associations: the interest here is in the frequency of association between words, this can be achieved by counting the number of times that crime related words appear within the same document and their most frequent correlated words. While n-gram is used to explore pairs of adjacent words, it cannot be used to extract words that tend to co-occur within particular documents if they do not occur next to each other. The widyr package[19] in the R library was utilised here in order to count correlating pairs of words in our vendors' review dataset. A function from the widyr library that lets us count common pairs of words co-appearing within the same document is the pairwise_count() function. However, in order to examine correlation among words, which indicates how often they appear together relative to how often they appear separately, a binary correlation measure called phi coefficient was used. The measure focused on how much more likely it is that either both words appear. The phi coefficient was applied in order to compare the correlation of the word "scam" and other words in each of the four subsets of our vendor review corpus. TABLE 2.8 is a comparison of the correlation result for Generic Doctor (generic) and Cheap Scrips (cheaps).

TABLE 2.8. Highly correlated words with scam in two vendor reviews

Generic Doctor	
Words	**Correlation**
review	0.45482819
experience	0.43301270
company	0.33245498
receive	0.32397580
week	0.32397580
confirm	0.32397580
business	0.31754265
people	0.31754265
card	0.30000000
money	0.30000000

Cheap Scrips	
Words	**Correlation**
ago	1.0000000
9	0.7071068
mail	0.7071068
waiting	0.6324555
refund	0.6324555
money	0.6324555
told	0.6324555
antibiotics	0.5000000
company	0.2500000
contact	0.2500000

[19] https://cran.r-project.org/web/packages/widyr/index.html

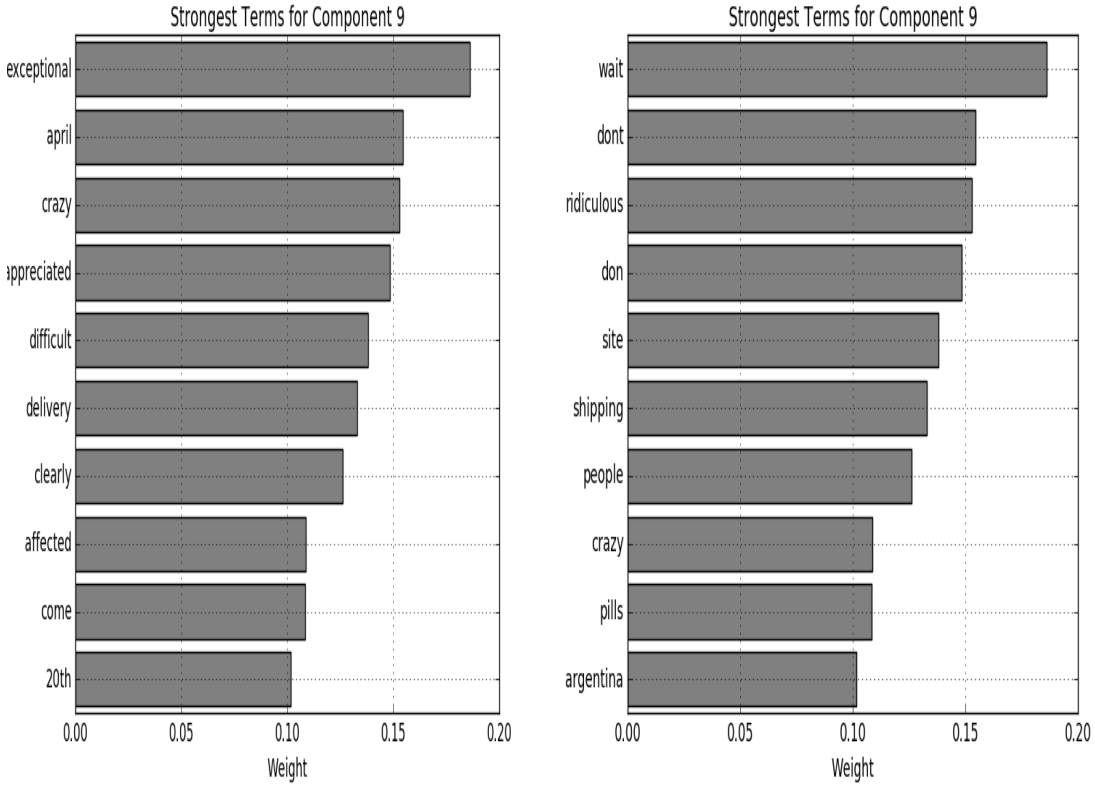
Recall that Generic Doctor is a top rated vendor while Cheap Scrips is a blacklisted vendor, therefore, higher correlation values were expected in the latter. This was the case, however, a detailed interpretation of this comparison may require a domain expert or knowledge.

2.8.4. Mining Topics from User Feedback

Here, we are not just interested in understanding the relationship between words in a text document, we are also interested in finding the underlying themes and topics in the textual component of the user feedback. Recall it was reported in section 2.5 that LSA, PLSA, and LDA are some of the popular approaches for discovering hidden topics in natural language data. The experimental work in (Stevens et al., 2012) also reported that LDA learns descriptive topics best while LSA is best at creating a compact semantic representation of documents and words in a corpus; the study also highlighted topic mining result evaluation as a fundamental issues and that it is primarily concerned with the degree to which the learned topics matched human judgments or enabled us to differentiate between themes and ideas embedded in a corpus. In order to evaluate unsupervised statistical method such as topic modelling, it is useful to try it on a very simple case where you know the right answer. In this case, it makes sense to reasonably assume that there will be some differences in the reviews of each of the vendors, especially between top rated and blacklisted vendors, hence the examination of each will be carried out in order to investigate how separate their word distributions (topics) are. This will also be an evaluation strategy for understanding whether the algorithm can correctly distinguish the topics in the four vendor review corpus. It also lets us double-check the usefulness of the method as well as to gain a sense of how and when it can go wrong. The experiments that follow utilised D^2, the Pharmacy Review dataset in order discover hidden themes and topics that differentiate between top rated vendor (Generic Doctor) and Blacklisted vendor (Cheap Scrips).

Topic Modelling Using LSA: as described in subsection 2.3.2 and section 2.5, LSA map words in documents into a concept space. LSA achieve this by first assuming that words have only one meaning, the text documents are then represented as BOW, and finally, the concepts are expressed as patterns of

words that usually appear together in the documents[20]. LSA takes a Term Document Matrix (TDM) or Document Term Matrix (DTM) as an input and performs SVD such that the resulting feature vector representation captures the core concept explained by the terms in the input. Scikit-learn, described in (Pedregosa et al., 2011) is a data analysis tool that implements both TfidfVectorizer and TruncatedSVD for performing LSA with the help of large scale eigenvalue solvers such as the ARPACK algorithm. This implementation of LSA was applied on each online pharmacy vendor review corpus. Fig.2.13 is a bar plot representation of one of the output components each for Generic Doctor and Cheap Scrips vendors, the components contain the top ten terms that represent inferred concept.



(a) Generic Doctor (b) Cheap Scrips

Fig. 2.13. Topic versus terms mapping using LSA

[20] https://technowiki.wordpress.com/2011/08/27/latent-semantic-analysis-lsa-tutorial/

It can be inferred from result a high level of satisfaction from the terms (exceptional and appreciated) in the Generic Doctor component and a possible dissatisfaction from the terms (ridiculous and wait) in the Cheap Scrips component. Despite some terms overlap (such as crazy), this actually makes sense, recall that Generic Doctor is among the top listed while Cheap Scrips is a blacklisted online vendor on Pharmacy Reviewer. Included also in this report is the top ten terms in the top ten components of both vendor reviews (see Appendix I and Appendix II, respectively). Comparing the ten components, one will notice terms such as scam, fake, and refund appearing in the Cheap Scrips review.

Topic Mining Using LDA: the LDA algorithm (Blei et al., 2003) implementations (Hornik and Grün, 2011) in R and Python Scikit-learn described in (Pedregosa et al., 2011) were applied in order to find natural groupings and themes in our online pharmacy vendor review corpus. For each combination, the LDA model computes the probability of that term being generated from that topic. Fig. 2.14 is the top terms within each review corpus using the R implementation. Besides estimating each topic as a mixture of words, the algorithms also models each document as a mixture of topics.

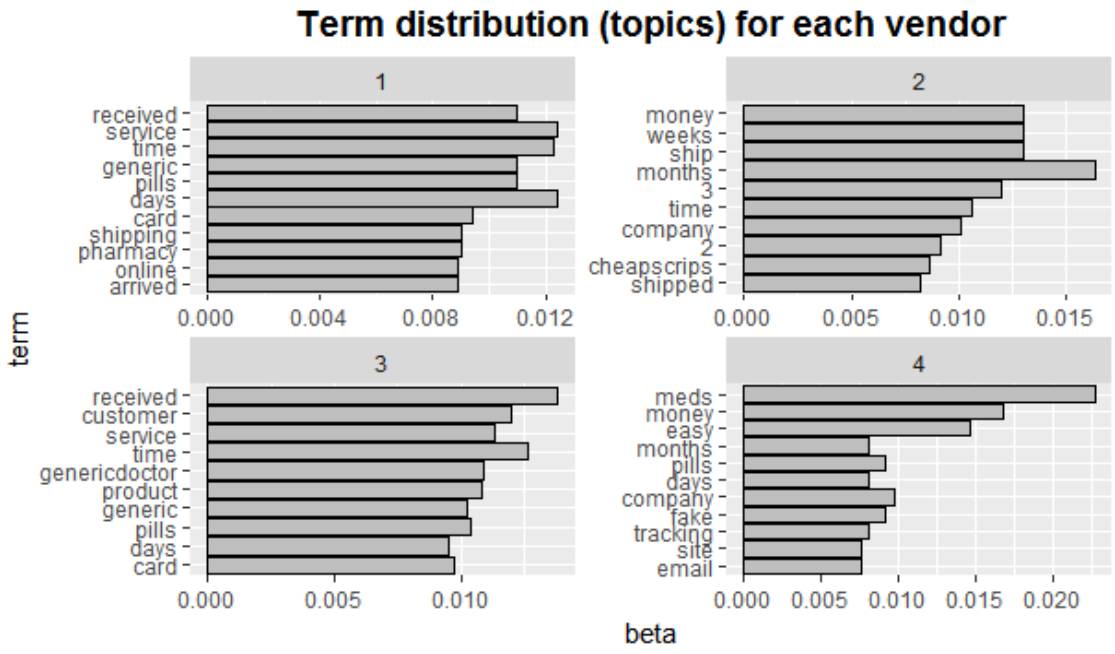


Fig. 2.14. Term distributions in each of the four vendors

This is called per-document-per-topic probabilities and can be estimated using a function called gamma. Each document in this analysis represented a single

vendor. Thus, one may want to know which topics are associated with each document. Recall that each of these values is an estimated proportion of words from that document that are generated from that topic. From the topic probabilities, it can be seen how well the unsupervised learning did at distinguishing the four vendor reviews. It is expected that reviews for each vendor would be found to be mostly (or entirely) generated from the corresponding topic. That is accomplished by trying to put the reviews back together in the correct vendors (see Fig. 2.15).

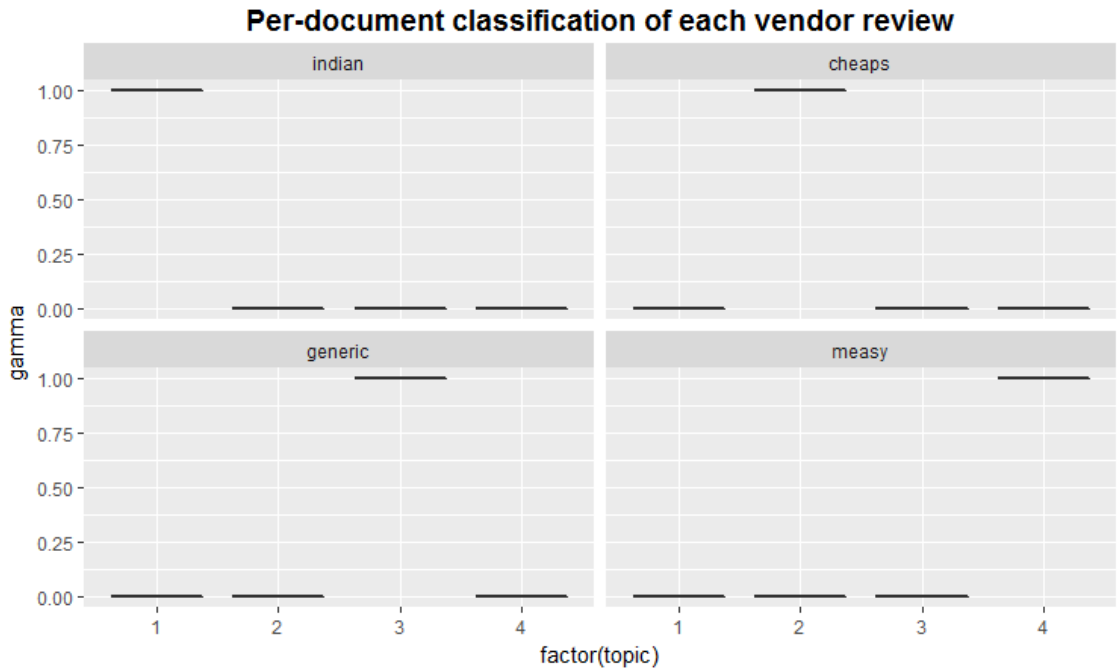


Fig.2.15. Per-document classification of each vendor review

As seen in the Fig. 2.15, almost all of the vendor reviews were uniquely identified as a single topic each and there are few or no any case where the topic most associated with a vendor belonged to another. Specifically, the following accuracies were obtained for each vendor: Indian Pharma (0.9999936), Cheap Scripts (0.9999766), Generic Doctor (0.9999947), and Meds Easy (0.9999737) which shows a great consensus. In the next subsection, our focus will go beyond investigating just word distribution in the different vendor reviews, to the study of sentiments expressed there in. The Python Scikit-learn implementation of LDA was also applied in order to compare the topics in the top and blacklisted vendors in the review corpus. The top ten word distributions of three topics each for Generic Doctor and Cheap Scrips is shown in TABLE 2.9.

TABLE 2.9. LDA inferred topics from top and blacklisted vendors

Topics	Generic Doctor	Cheap Scrips
Topic 1	received ordered order product did phone store card generic transaction	gold able used 09 just problem reviewer past editor check
Topic 2	complaints order card service reship used cancelled recommended weeks placed	got email com company send 600 scam order real weeks
Topic 3	great service prices genericdoctor customer product extremely products company online	ship orders order argentina just villazon jujuy bolivia place shipper

Again, despite natural language processing complexities, the result closely agrees with that of LSA and revealed some key differences in word usage. These differences, however, will be better explained using text mining techniques such as sentiment analysis that measures polarity or emotions in language.

2.8.5. Mining Sentiments from User Feedback

When users of certain products or services write a review, people often use their understanding of the emotional intent of words to infer whether a section or the entire review text is positive or negative or perhaps characterised by some other more nuanced emotion like joy, surprise or disgust. A common way to analyse the sentiment of a text document is to consider the text as a combination of its individual words or terms and the sentiment content of the whole text document as the sum of the sentiment content of the individual words or terms (Julia, 2017). The sentiment analysis techniques described in section 2.7 will be utilised here in order to measure the sentiment of keywords of interest expressed on some popular social media platforms.

Sentiment Analysis of Feedback Data from Facebook: this experiment is detailed in Appendix III and it demonstrates the usefulness of the proposed product safety framework in section 2.7 and Fig. 2.5. As described in subsection 2.8.1, the dataset is a user feedback data gathered from Facebook pages of 3 popular brands of drug and cosmetic related products (Avon, Dove and OralB) using the Facebook Graph API. See TABLE 2.10 for the comments summary. Over the entire corpus, tokenisation, stop word removal and conversion to lower

case functions was applied using the R tm[21] library in order to obtain only individual word features.

TABLE 2.10. Summary of extracted comments

Brand Name	Total no. of posts	Total no. of comments
Brand X	5000	654
Brand Y	665	1747
Brand Z	5000	957

The features were then represented as a BOW model with the $(TF - IDF)$ weighing scheme to generate a sparse matrix, limiting the output to a minimum of three characters word length. The sparse matrix was handled using an R function that removes sparse terms which have at least a 99 percentage of sparse elements. Frequent term analysis was then performed in order to generate popular words for creating preassembled lexicon, the method used in (Nathan and Richard, 2014) with the help of English dictionary and Thesaurus was adapted. The preassembled lexicon was then merged with social media slangs and the generic lexicon for sentiment classification in (Liu, 2012).

..

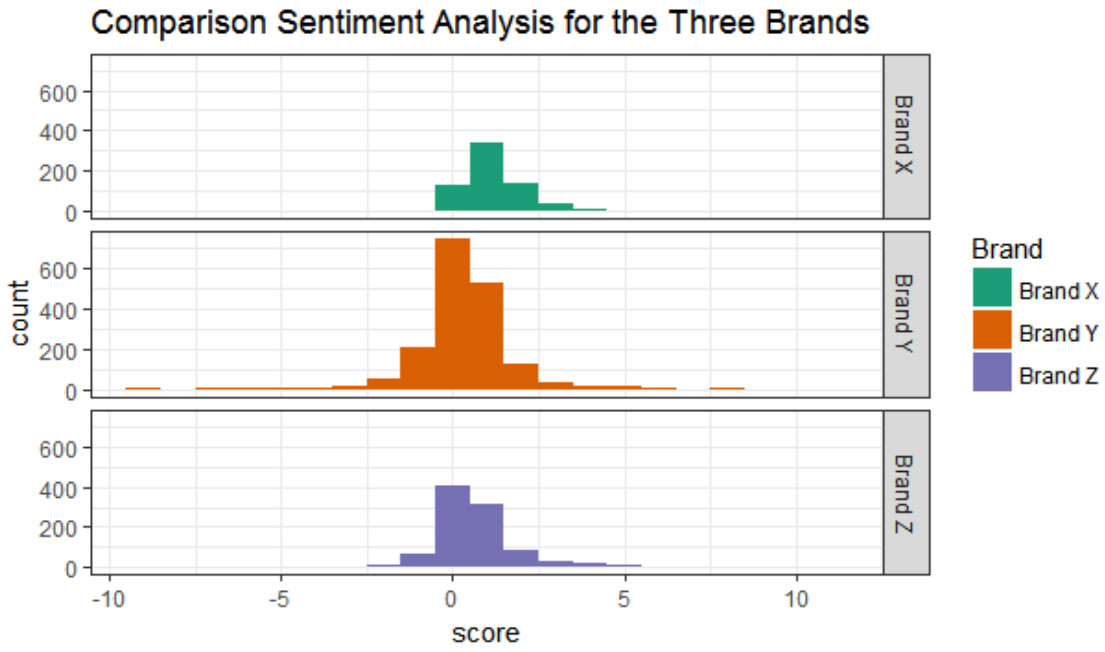


Fig. 2.16. Comparison sentiment analysis for the three brands

[21] https://cran.r-project.org/web/packages/tm/index.html

A comparison lexicon sentiment analysis was performed over the 3 different brands. The result obtained is presented in Fig 2.16. It is interesting to see the overall sentiment on all the three brands been positively skewed, with negative:neutral:positive ratios for Brand X, Brand Y and Brand Z approximately 1:42:175, 1:3:3 and 1:5:5 respectively; the high positive sentiment on Brand X confirms comments been prize driven. The distribution of these scores is presented in TABLE 2.11.

TABLE 2.11. Distribution of sentiment scores for the three brands

Brand	Sentiment Scores			
	Negative	**Neutral**	**Positive**	**Row Total**
Brand X	3	127	524	654
Brand Y	282	742	723	1747
Brand Z	85	408	464	957
Column Total	370	1277	1711	3358

Three separate datasets were also extracted in Brand Y each with the mention of the three generic products, Soap, Cream and Deodorants and then a comparison sentiment analysis was performed on the aggregated data; the distribution of these scores is presented in TABLE 2.12.

TABLE 2.12. Distribution of sentiment scores for the three products

Product	Sentiment Scores			
	Negative	**Neutral**	**Positive**	**Row Total**
Cream	5	11	11	27
Deodorant	21	23	46	90
Soap	11	26	56	93
Column Total	37	60	113	210

This is another interesting result with Soap been more positively skewed among the three products; the negative:neutral:positive ratio of Cream, Deodorant and Soap approximately been 1:2:2, 1:1:2 and 1:2:5 respectively. The result is presented in Fig. 2.17. The same problem of classifying the sentiment orientation of the entire corpus is also illustrated using a naïve Bayes sentiment classifier modeled with polarity and emotional lexicon.

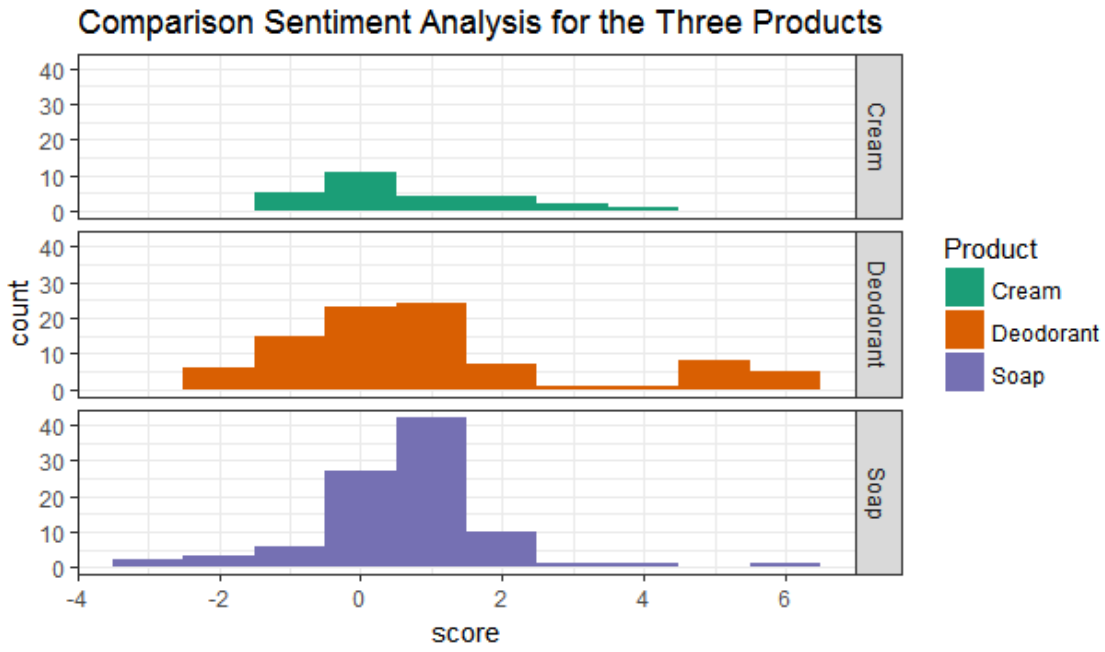


Fig. 2.17. Comparison sentiment analysis for the three products

The method classify_polarity classifies the comments as positive, neutral or negative. The classification results over the entire corpus for both methods are compared in TABLE 2.13. The negative scores for both methods agree closely while there is a sharp variation in both the neutral and positive scores.

TABLE 2.13. Lexicon versus machine learning sentiment scores

Method	Sentiment Scores			
	Negative	**Neutral**	**Positive**	**Total**
Lexicon	539	1436	1383	3358
Naïve Bayes	554	368	2436	3358

Sentiment Analysis of Feedback Data from Twitter: to demonstrate how the framework can be applied to model a machine learning classifier for predicting the sentiment of a given text, 11, 431 Tweets was gathered from Twitter API using combinations of the following query terms: “medicine”, “prescription”, “over the counter”, “side-effects”, “online pharmacy” and “antibiotics”. The Tweets were first cleaned then converted to a single corpus, with each Tweet represented as a single document. Thus there are 11,431 documents making up the corpus. The corpus was then represented as a sparse matrix for frequent term analysis and

lexicon generation as described in Appendix III. A lexicon sentiment analysis was then performed over the entire corpus such that the sentiment scores can be generated as described in TABLE 2.14; for simplicity, we categorised scores greater than zero as positives and scores lower than zero as negatives. The dataset was split into 75% i.e. 8573 labeled Tweets as a training set and 25% i.e. 2858 labeled Tweets as test set so that the classifier can be evaluated on data it had not seen previously. Naive Bayes algorithm was applied here as a baseline classifier.

TABLE 2.14. Random preview of polarity scores

Score	Tweet
-1	To everybody that knows me I think it's time for me to sign paperwork to take me off my antibiotics which will allow me to die in my home
2	Pretty sure we headed a hospital visit off today, fingers crossed these antibiotics help ASAP.
0	Took my antibiotics without eating and i was almost sick in work
-2	Its ridiculas how many times i get ill, iv either got a cold or on antibiotics , my emune system is rubbish

TABLE 2.15 is the representation of the result as a confusion matrix. A Naïve Bayes classifier was used to obtain a baseline result for assessing other classifiers.

TABLE 2.15. Naïve Bayes classifier result

Predicted	Actual		
	Negative	**Positive**	**Row Total**
Negative	531 0.639	180 0.089	711
Positive	300 0.361	1847 0.911	2147
Column Total	831 0.291	2027 0.709	2858

Looking at the table, it can be seen that 300 out of 831 positive messages (36 percent) were incorrectly classified as negative, while 180 of 2027 negative messages (8.9 percent) were incorrectly classified as positive; a total accuracy of about 83%. This performance will be used as a baseline for assessing other

classifiers. Again as detailed in Appendix III, the brand and product comparison results signify the usefulness of text mining and sentiment analysis on social media data while the use of machine learning classifiers for predicting the sentiment orientation provide a useful tool for users, product manufacturers, regulatory and enforcement agencies to monitor brand or product sentiment trends in order to act in the event of sudden or significant rise in negative sentiments.

Sentiment Analysis of Pharmacy Reviewer Data by Emotion: instead of sentiment scoring by polarity (positive, neutral, and negative) as in the previous example, the Generic Doctor and Cheap Scrips reviews were subjected to sentiment scoring by emotion using a naïve Bayes emotional sentiment classifier described in (Jurka, 2012). The method classify_emotion was utilised to classify the reviews into the following seven emotional categories joy, sadness, fear, surprise, anger, disgust, and unknown. The emotional classification of Generic Doctor review data (see Fig. 2.18) justified its top rated status, having approximately 62% joy, 6% sadness, 5% fear, 5% surprise, 2% anger, and 0% disgust, and 20% unknown.

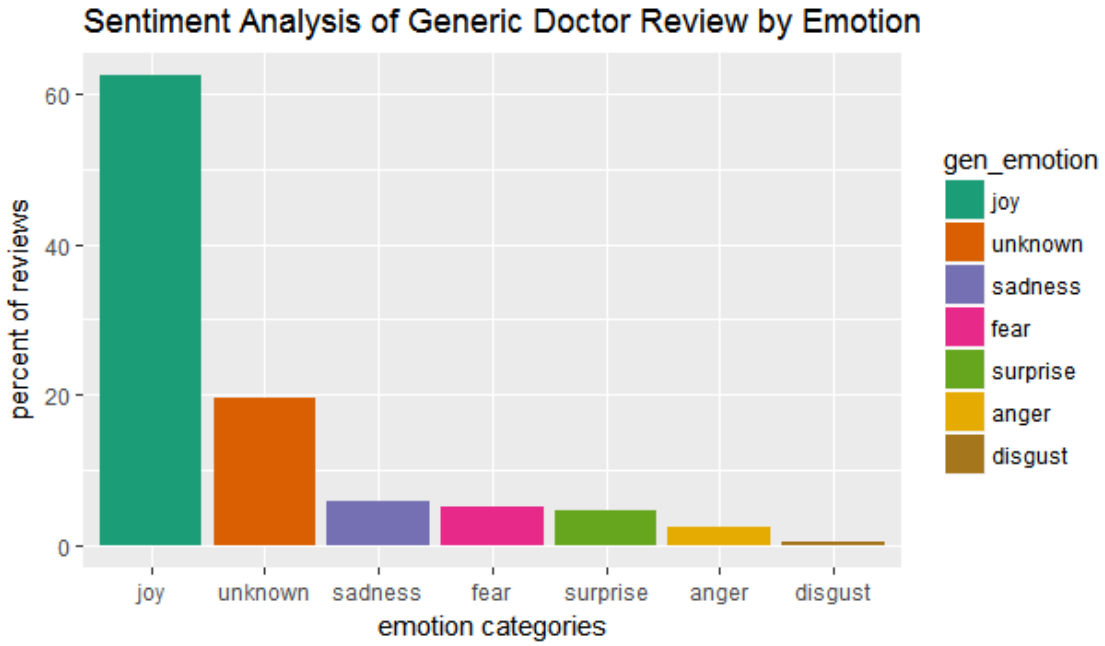


Fig. 2.18. Sentiment analysis of Generic Doctor review by emotion

That of Cheap Scrips (see Fig. 2.19) also justified its blacklisted status, having approximately 38% joy, 15% sadness, 4% fear, 4% surprise, 6% anger, and 3% disgust, and 30% unknown.

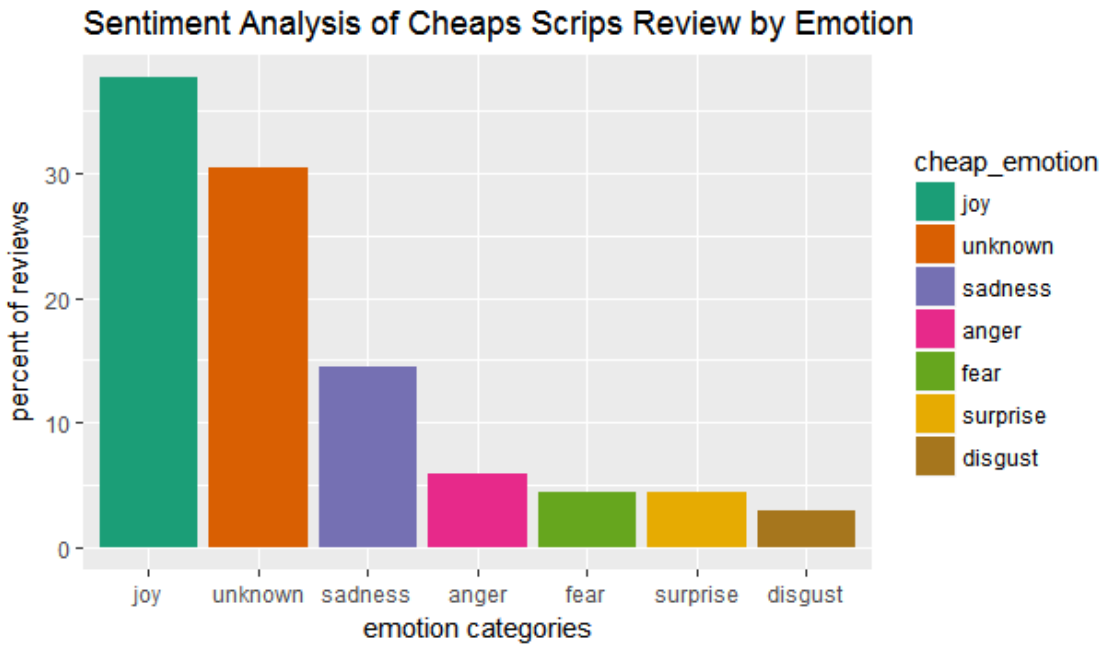


Fig. 2.19. Sentiment analysis of Cheap Scrips review by emotion

2.9. Issues in Social Data Content Mining

There are so many quantitative and qualitative issues that can affect the validity of data analysis results. For example, what quantity or data is sufficient mining and decision support? What level or data quality is acceptable? Data collected from real world sources are rarely ideal. Reputation systems have widely become victims to the unfair rating problem, where advisors (i.e., buyers providing feedbacks) provide misleading opinions about sellers, to alter their trust scores (Liu et al., 2014). Again, natural language text often contains some ambiguities and noise from a variety of sources such as misspellings, auto corrections, abbreviations, rendering errors, imperfect pre-processing steps, missed labelled classes, idiosyncratic language and slangs usage; in most cases, the more informal the communication, the more likely the noise (Berndt et al., 2015). Special fonts and symbols used on social media platforms such as emoji are a bit different on different devices and platforms and can cause serious data quality issues (Miller et al., 2016). Other sources of noise, according to (Prusa, 2015) include redundant data or instances that cannot be reliably distinguished as

belonging to a class. In some cases, bot accounts are created on social media to be posting multiple updates using identical text and/or emoji in order to distort search and analytic results as well as certain event outcomes.

The performance of data analytic applications depends to a great extent on the quality of the input data, predictive features, and outcome labels. Low quality models will result if the input data used for building predictive models are inaccurately labelled, flawed, or contains noise in the form of unwanted terms or phrases. Erroneously labelling documents could cause the machine learning algorithm to learn inappropriate language patterns and associate them with an incorrect target (Berndt et al., 2015). Noise could have a severe impact on the performance of classification algorithms (Abbasi et al., 2008). The impact of different levels of noise on sentiment analysis especially on social data has largely been ignored on the grounds that many of the sentiment learners are designed to be resilient to noise (Prusa, 2015). The study in (Miller et al., 2016) discovered that emoji font diversity can cause miscommunication and errors in sentiment analysis results.

Three text mining approaches described in (Berndt et al., 2015) include basic keyword or regular expression matching (useful for finding very specific patterns in data when known a priori), natural language processing (use of syntactic and semantic rules as well as domain information to construct controlled vocabularies and ontologies that underpin text processing and handling of negation, temporal relationships, hypothetical situations, and experiences), and the use of machine-learning algorithms (data-driven approaches that utilise statistical methods to group documents based on similarity scores. Among these three approaches, the machine learning algorithms are the most affected by data quality issues because they learn by examples.

Evaluation of the effects of noise on the performance of text mining techniques could determine if additional actions need to be taken to address the issue. Emoji is gaining popularity in social media communications and offers new possibilities for digital expression, it is therefore very important to study and understand its associated quality issues. Several experiments were conducted in (Berndt et al., 2015) in order to assess the robustness of some machine learning

algorithms applied in text mining in terms of the task complexity, the input data quality, and the target quality in the medical domain. An interesting result obtained suggest that the performance of the model degrades as more data instances with mislabelled targets are added. Inflated reviews and fake customer feedback is a common problem of review and feedback systems, hence the need to develop techniques that can detect and separate paid and bogus reviews from organic reviews. The issue of data quality will be fully addressed in Chapter Four.

2.10. Conclusions

This chapter focused on mining social data content for crime intelligence, it began by narrating how traditional crimes have taken a new turn with the advent of the Internet. The study focused on Cyber-enabled crime intelligence as it is largely unexplored. The targeted application area is the advertisement and sale of illegal pharmaceutical products. The wide availability of products and services being offered through e-commerce platforms necessitates the need to accurately assess and evaluate the quality and genuineness of both the source or vendor and the product or service. Recent studies revealed that consumer feedback (rating, review text, review of review, social media comments, and posts, are important source of information that is becoming key influencers of behaviours and policy decisions and can be explored using machine learning and data mining techniques in order to measure the perceived quality of both vendors and products or services. Current data mining and machine learning approaches to cyber-enabled crime intelligence are mainly based on the analysis of limited trail or log data such as mobile and online communications, travel and financial transactions of offenders and victims accounts. There are many evolving groups and pages on social media that are dedicated to providing feedback on their experiences of using some essential goods such as food, medicines, and cosmetics. These constituted a great source of data and motivation for this research.

A novel research question (RQ1) was proposed as follows, can user feedback be harnessed and processed for crime intelligence? This question was tackled by harnessing, aggregating, exploring, and modelling user feedback on certain products and brands available online in order to understand, learn, and predict hidden topics and sentiments. This was followed by the use of

mathematical techniques to measure the similarity between user feedback and crime related queries with the goal of inferring the likelihood of criminal activities. Novel experiments were carried out using state of the art methods of data representation, feedback aggregation, topic modelling, and sentiment analysis. First, the experiment reported how to best measure and rank the quality of online vendors of pharmaceutical products from customer ratings based on Bayesian ranking models. It then proposed and evaluated a novel framework for harnessing and processing the publicly reported views and experiences of consumers of popular brands of drug and cosmetic products. LSA was applied in order to measure the similarity between customer complaints crime and related keywords, while LDA was utilised in order to discover hidden topics in the user feedback. The results obtained for different vendor reviews revealed some hidden and interesting distributions of words that can be further utilised to infer themes and topics that are crime related. The sentiment analysis results also revealed some interesting results that can be useful in measuring significant changes or trends in the perceived quality of a vendor.

Finally, the study reported several issues that may affect the validity of the obtained results and how poor data quality can skew or introduce bias into the results. Experimental results of this nature can only be useful when the data being processed are of high quality. Issues of data quality will be tackled in Chapter Four and Chapter Five of this thesis.

Chapter Three: Social Network Mining for Crime Intelligence

3.1. Introduction

Many real life phenomena can be represented and explained with network graphs and matrices, this is especially the case in tackling organised pharmaceutical and insurance crimes as well as in domains where links and interconnections are of critical importance. Social networks represent relationships (connections and interactions) among social entities. According to (Tsvetovat and Kouznetsov, 2011), the science of Social Network Analysis (SNA) boils down to one central concept which is our relationships (such as friendship, influence, affection, trust, dislike, or conflict). These relationships can be binary (exist or not) or valued (quantified), directed or undirected, symmetric (both vertices must agree) or asymmetric (initiated by a single vertex), unimodal (vertices of the same type) or multimodal (actors of different types), etc. While link analysis is used to describe the mixing of different types or modes of vertices and edges in the same network, here for simplicity, we will use SNA to refer to both unimodal and multimodal networks.

Intelligence agencies have long used crime data analysis to understand the past crimes, recent advances in the field of data science and network analysis have introduced a new paradigm which aims to predict future crime called predictive policing (Tayebi and Glässer, 2016). The analysis of social networks can help us understand important people, roles, subgroups, and processes in a network (Golbeck, 2013). In the fight against crime, knowledge about the structure, flow of communications, and organisation of criminals and their associates is of fundamental importance to criminal intelligence agencies. However, the process relies on the ability to obtain and use reliable data (UNODC, 2011). Using social network analysis to analyse crime data obtained from various sources including phone traffic records, social networks, surveillance data, interview data, police operational and intelligence data could

be useful when deciding when to initiate destabilising actions or where to allocate resources to prevent future crime occurrence (Ferrara et al., 2014).

Predictive policing is aimed at using historic data to identify and predict future offenders, victims, criminal collaborations, and crime locations. For instance, co-offending network analysis contributes to crime intelligence by detecting hidden links and predicting potential links among offenders. Recent successes in crime intelligence include co-offending network disruption (finding a set of network actors whose removal will maximally destabilise the network) and organised crime detection i.e., understanding and destabilising the structure and sustainability strategies of organised crime groups (Tayebi and Glässer, 2016). Criminal intelligence process relies on the ability to obtain and use data. Network connections can also arise from information in sources such as text, databases, sensor networks, communication systems, simulated data, surveys, company data, covert or dark network, and newswire (Campbell et al., 2013). These connections and all the processes associated with it have a central role in crime intelligence. As detailed in Appendix IV, there are, however, situations in many criminal datasets where there is no direct connection among the criminals. Such is the case in (Pfeffer and Carley, 2012) involving the extraction of organisational structure of covert network from textual data obtained from news archive, also in (Compton, 2015) which is an online black market (Evolution, that became extinct in 2015) comprising of list of underground vendors with their associated products, and finally the medical quality database described (Krech et al., 2014) comprising of a list of manufacturers (including rogues) with their associated products. Developing methods and tools that can reveal or infer hidden relationship among criminals will not just be interesting but necessary for predictive policing. Many life questions can be answered by studying links in data, for example, who are the criminals? What is their modus operandi? Any community structure among them?

A graph representation of crime data is useful in understanding criminal network structure and in identifying clusters, cliques, and key players in the network (Ferrara et al., 2014). Creating networks based on item co-occurrence in the articles includes folding two-mode networks of articles and items to one-mode networks of items (Pfeffer and Carley, 2012). The analysis of underground forums

can also provide key information about who controls a criminal network or sells an illegal good or service, and the size and scope of the cybercrime underworld (Afroz et al., 2014). Vulnerability in organised crime groups that can be exploited by researchers and law enforcement agencies include how trust is earned among peers and the way they get their money or the e-currency they use (Greenstadt, 2015). This study is particularly focused on understanding and characterising hidden criminal collaborations from data, the goal is to develop methods for network information extraction and hidden links identification in crime data. It will also investigate ties and community structure in crime data in order to understand the operations of both traditional and cyber criminals, as well as to predict the existence of organised criminal networks. Two major interesting research directions include inferring relationship based on common attributes (co-occurrence) and other metadata among entities involved in a crime but not directly connected, the identification of structures and key individuals or influential groups directing the overall operations of a criminal network. These directions then led to the novel research question RQ2 which aims to investigate whether criminal associations, structures, and roles can be inferred among entities involved in a crime but not directly connected. The strategy adapted here to address RQ2 are (i) modelling a bipartite network in order to infer relationships between actors and resources involved in crime and (ii) analysing vertices and community structures of the resulting network.

This chapter is organised as follows: section 3.2 is a review of social network analysis (SNA) literature as it applies to organised crime detection and analysis, section 3.3 is a discussion on fundamental background of network analysis, section 3.4 describes the proposed model for inferring hidden ties in data; section 3.5 details the experimental work and case studies for evaluating the proposed model, section 3.6 is a discussion on issues relating to data quality and its effect on analysis result, and section 3.7 concludes the chapter and provide some recommendation for future work.

3.2. Literature Review

The Internet, WWW, and social networks have become enablers of crime coordination across dispersed areas. Despite the current legislations and efforts by many agencies and stake holders, very little is yet known about the structure,

sustainability strategy, communication flow and how trust is assured among criminal groups (Broadhurst et al., 2014). However, the study in (Yip et al., 2012) has shown that the least common denominator of organised crime (including cybercrime) is a relationship. Conventional and cyber criminals are more likely to participate in loosely associated illicit networks or inferred links than formal organisations or explicit social links (Sarvari et al., 2014). This relationship or association can be represented as a network in which the vertices are the criminals while the edges are the criminal interactions.

According to the reports in (Motoyama et al., 2011), (Yip et al., 2012), (Afroz et al., 2014) and (Compton, 2015), criminals often utilise underground forums (chatrooms and private messaging services) in the darknet to exchange information on abusive tactics and engage in the sale of illegal goods and services. The nature of organisation among these criminals depends on their targeted activity, these can be (i) purely aimed at online targets such as swarms (consisting of ephemeral clusters of individuals with no leadership, e.g. the Anonymous) or hubs (organised with a clear command structure and a focal point of core criminals around which peripheral associates gather, e.g. the LulzSec) (ii) based on the use of online tools to enable conventional crimes, these comprised clustered hybrids (articulated around a small group of individuals and focused on specific activities or methods e.g. carding networks) or extended hybrids (less centralised consisting of many associates and subgroups), and (iii) a combination of both online and offline targets such as hierarchies or aggregates depending on their degree of cohesion and organisation (Broadhurst et al., 2014).

Criminal data often contain a variety of network entities such as persons, products, organisations, locations, http links, vehicles, weapons, properties, and bank accounts. Learning relationships and associations between these entities is a critical part of uncovering criminal activities and fighting crimes. SNA was described in (Tsvetovat and Kouznetsov, 2011) as a study of human relationships by means of graph theory. SNA has found application over the years in evidence mapping, understanding relationship patterns, and in the identification of key members in a criminal group (Sarvari et al., 2014). SNA is enabling investigators to be able to detect patterns within and across product lines as a potential developing crime ring, as such saving companies and organisations from losses

and further development of the crime ring. Detailed in (Kirchner and Gade, 2011), (Kriegler, 2014), (Broadhurst et al., 2014), and (Fortunato, 2010), the commonly used SNA metrics used by these investigators include (i) vertex, centrality, and network density measures for the identification of pivotal vertices and potential fraud hotspots, sub-structures, and structural holes, (ii) clustering coefficient measures for network classification and path prediction, (iii) link analysis metrics for generating investigative leads by mapping criminal actors and the links between them as well as for uncovering missing information that may be hidden in a criminal network, and (iv) community detection, density, cohesion, and stability for understanding the group structure and organisation in criminal networks.

The study in (Yip et al., 2012) analysed anonymised private messaging records of carding forums' using degree distribution, assortativity, rich club phenomenon, transitivity, small world phenomena, connectivity and cohesive subgroup metrics with the goal of uncovering the underlying structural and behavioural properties of cybercriminals. The study in (Motoyama et al., 2011) explored the relationships (buddy, private message, and thread networks) and social dynamics of six underground forums with the aim of understanding how the users interact, gain and sustain trust or reputation, and how they impact e-crime market efficiencies. Doppelgänger Finder was used in (Afroz et al., 2014) to evaluates duplicate identities identification and to discover and group unknown identities in carding forums when ground truth data is unavailable. The study in (Greenstadt, 2015) utilises six centrality measures to characterise cybercrime forum reports, the results obtained showed that criminal groups are organised into two distinct communities, these are gangs (limited in size with one central leader who decides for the group) and mobs (thousands of members divided into multiple sub-groups and often share relatively equal centrality rankings). An analysis of the products that were traded in the defunct online black market called Evolution was carried out in (Compton, 2015), first a vendor co-occurrence relationships defined based on the number of vendors who sell both incident products were built to visualise the Evolution product network, then a market basket analysis was carried out in order to quantify the results suggested by the visual analysis.

An analysis of the community structure of Nigerian scammers on Craigslist were carried out in (Park et al., 2014) with the aim of understanding scammers location based on IP and shipping addresses, understanding their modus operandi based on their level of automation and the sophistication of the tools they used, and finally learning features that can be used to distinguish a scam email from a legitimate email, the results obtained shows that the scammers are organised into tightly and loosely connected groups. Another related analysis of Nigerian scammers in (Sarvari et al., 2014) was aimed at understanding the criminal that has a central role in the network, detection of subgroups and communities in the network, brokers of collaboration and information in the network, the ranking of criminals according to their importance and influence in the network. The results obtained revealed that only ten groups are responsible for about fifty percent of the scam attempts we receive.

An analysis of cybercriminal ecosystem on Twitter was also carried out in (Yang et al., 2012) in order to investigate the inner and outer social relationships among user accounts and results obtained indicated that the criminal hubs are more inclined to follow criminal accounts (which tend to be socially connected) thereby forming a small-world. A network analysis of South African arms deal and corruption carried out in (Kriegler, 2014) profiles the dynamics of criminal networks and the strategic relationships that make their network resilient, results from the analysis reveals that an actor can have many separate relationships in the network through which resources such as information, money, and influence are shared with other actors. The removal of this actor would disconnect a large part of the network and resources flow across the network. The study in (Paulo et al., 2013) combine techniques from logic programming, viral marketing, and community detection to associate arrested individuals with a criminal gang by assigning such individuals a degree of membership that represents the confidence that the individual is in a given gang. The different results obtained from these studies would allow authorities to better utilise their resources and devise more effective intervention and disruption strategies in the future.

In terms of procedures and methods for building networks from data, different authors adopt different approaches, for instances in (Sarvari et al., 2014) a large scale criminal network was constructed from a smaller set of leaked data

which contains email addresses that were used to further scrape the network from Facebook profiles. The study in (UNODC, 2011) identified three main types of data sources for criminal intelligence, these sources include open (publicly available such as the grey literature), closed (with limited access and availability such as criminal records, vehicle registration, and weapons licensing data) and classified (highly restricted such as data from covert operations). The study then outlined the procedure for the network formation and analysis to include understanding client needs, obtaining and use of relevant data, data quality evaluation, data collation strategy, data integration and analysis, knowledge dissemination, and finally, re-evaluation. The recommendation for building a criminal network in (Kirchner and Gade, 2011) however, is for the analyst to first use expert knowledge to create broad subsets of the network, the set of vertices that are flagged are then selected and then finally the network of all possible links only pertaining to these vertices are reproduced.

One of the most important tasks in crime prediction is the analysis of the collaborations patterns or relationships between offenders (Tayebi, 2015). There are situations where we are interested in analysing certain networks but lack data or metadata regarding the relationship among the objects of interest. For instance, the study in (Chen et al., 2016) aimed at understanding the social structure of hidden and hard to reach populations such as drug users and sex workers addressed the issue of data unavailability due to privacy and non-disclosures by reconstructing the underlying network structure through which the subjects are recruited. The study in (Tang et al., 2011) and (Zhuang et al., 2012) looked into the factors that imply and how to automatically learn or infer the type of social relationships in a network in situations where users do not take time to label them, i.e. given a data on the behaviour, history, and interactions among users, the task is to estimate how likely these users are family members. The study in (Tang et al., 2012) however, developed a framework by incorporating social theories into a machine learning model for classifying social relationship types by learning across heterogeneous networks, the authors reported a F1-score of 90% for inferring manager-subordinate relationships in an enterprise email network. The authors of (Sharma et al., 2014) explored the extent to which social ties between a pair of users on Twitter can be inferred based on the

common activities of the users such as the number of common celebrity profiles they are following. The work in (Xiao et al., 2014) proposed and evaluated with good performance, a method for estimating the similarity, hence potential social tie between users according to their physical location histories represented by Global Positioning System (GPS) trajectories with a semantic location history (SLH) instead of using social structures.

This chapter is focused on criminal network formation from data where the relationship between actors are hidden but can be inferred. It proposes to use the concept of co-occurrence in historic data which involve the use of bipartite network model, a special type of network representation belonging to a k-partite group where the vertices are divided into two sets, and only connections between two vertices in different sets are allowed (Latapy et al., 2008). As detailed in the next immediate section, given a bipartite network with a direct relationship between objects and their properties, the goal here is to extract co-occurrence relationships from the network, this is similar to the work in (Kötter et al., 2015) whose goal however, was to extract taxonomy trees that described the is-a relation among the object properties in a network. The k-partite network representation naturally suits criminal datasets that may lack direct connections among criminals and allows for inferring of hidden ties as well as in understanding the relationships between actors on one side and features or traits on the other side in the network (Chessa et al., 2014).

3.3. Network Analysis Fundamentals

The focus in the previous chapter was centered on the analysis of rating and text data. This chapter will, however, look into the mathematical representation and analysis of network data. Unlike text which is unstructured, networks or graphs represent structured data. A network (also called graph) is a theoretical construct composed of points called vertices that are connected by lines called edges (Samatova et al., 2013). Representing data in form of networks can increase the explanatory power of models derived from the data. Knowing the vertices and edges with their associated features is all that is needed to analyse a social network (Golbeck, 2013). SNA essentially takes a network with vertices and edges and finds distinguished properties of the network through formal analysis

(Tayebi and Glässer, 2016). This study takes the position of a network rather than graph theorist by referring to the relational data between accounts as network and then using established graph theories and equations to describe and explain the network. Although there are many types of networks as described in (Zafarani et al., 2014), the emphasis here will be mainly on undirected weighted unipartite and bipartite networks. The section begins by describing how network data can be represented, it will then explore some metrics that are used to explain network structures, groups, operations, and processes.

3.3.1. Network Representation

Mathematically, a network or graph G is denoted as pair $G(V, E)$, where V represents the set of vertices or actors in the network while E represents the set of edges that connect vertices in the network. It therefore, follows from (Zafarani et al., 2014) that the mathematical representation for a set of vertices is given by:

$$V(G) = \{v_1, v_2, \dots, v_n\} \tag{3.1}$$

where V is the set of vertices and $v_i, 1 \leq i \leq n$, is a single vertex while $|V| = n$ is a measure the size of the network. Two vertices v_1 and v_2 in a network are adjacent when v_1 and v_2 are connected via an edge, i.e. v_1 is adjacent to $v_2 \equiv e(v_1, v_2) \in E$. In a similar fashion, edges in a network are represented in (Zafarani et al., 2014) by:

$$E(G) = \{e_1, e_2, \dots, e_m\} \tag{3.2}$$

where E is the set of edges and $e_i, 1 \leq i \leq m$, is a single vertex while $|E| = m$ is a measure the size of the edges set. In a directed graph (usually represented using arrows, an edge $e(v_i, v_j)$ is represented using an arrow that starts at v_i and ends at v_j. Edges that start and end at the same vertex are called loops or self-links and are represented as $e(v_i, v_i)$. Two edges $e_1(a, b)$ and $e_2(c, d)$ are incident when they share one endpoint (connected via a vertex), i.e. $e_1(a, b)$ is incident to $e_2(c, d) \equiv (a = c) \vee (a = d) \vee (b = c) \vee (b = d)$.

There are a variety of text-based and visual methods for representing network data. The text-based methods which include adjacency or edge list, adjacency matrix, incidence matrix, eXtensible Markup Language (XML), and JavaScript Object Notation (JSON) are usually the input formats for most of the visual techniques and are more suitable for representing large networks (Golbeck, 2013). These representations are also very useful because they are able to store the vertex and edge sets in a way that (i) does not lose information, (ii) can be easily manipulated by computers, and (iii) can have mathematical methods applied easily (Zafarani et al., 2014). Given that the number of vertices in a network G is $n = |V(G)|$ while the number of edges in the same network is $m = |E(G)|$, let's say that vertices are given by $V(G) = \{v_1, v_2, \dots, v_n\}$ while the edges are given by $E(G) = \{e_1, e_2, \dots, e_n\}$, the following definition then follows from (Samatova et al., 2013):

Adjacency and edge List: in an adjacency list, every vertex is linked with a list of all the vertices that are connected to it while in an edge list each element is an edge and is represented as (v_i, v_j), denoting that vertex v_i is connected to vertex v_j (Zafarani et al., 2014). The adjacency list representation of G is a list of length n such that the i^{th} element of the list is a list that contains one element for each vertex adjacent to v_i. An adjacency list requires about $n + 2m$ bytes for an undirected graph and about $n + m$ bytes for a directed graph to be represented. One can easily find vertices that are adjacent to the i^{th} vertex using algorithms for finding a path between two vertices that can run in $O(n + m)$ time. Adjacency list and edge list representations takes significantly less space for large and sparse or low-density networks hence useful for persistent storage and internal data representation. The disadvantage of this representation however, is that it only allows iterating over the edges, but not fast search or traversal through the network (Tsvetovat and Kouznetsov, 2011).

Adjacency Matrix: the adjacency matrix, also known as a sociomatrix representation of G is a $n\,X\,n$ matrix. A value of 1 in the adjacency matrix indicates a connection between vertices v_i and v_j, and a 0 denotes no

connection between the two vertices. This means if $a_{r,c}$ is the value in the matrix at row r and column c, then $a_{r,c} = 1$ if v_r is adjacent to v_c; otherwise, $a_{r,c} = 0$. In a general sense however, any real number can be used to show the strength of relationships between two vertices. In the adjacency matrix representation, diagonal entries represent self-links or loops (Zafarani et al., 2014). With adjacency matrix representation, one can easily answer whether v_r and v_c are adjacent or in general whether there is an edge between two vertices. About n^2 bytes is required to represent an adjacency matrix. Despite its popularity, the major downside of adjacency matrix is that $a_{r,c} = 0$ cells take the same amount of memory as the other cells, the ratio of $a_{r,c} = 1$ cells to $a_{r,c} = 0$ is called density and most online social networks have density of 0.1% or less and the larger a social network is, the lower its density will be (Tsvetovat and Kouznetsov, 2011). This is called sparse matrix because it populated primarily with zeros is often the case in in most social networks because of the relatively small number of interactions among the actors. According to (Zafarani et al., 2014), an adjacency matrix is commonly formalised as:

$$A_{i,j} = \begin{cases} 1 & \text{if } \mathrm{v}_i \text{ is connected to } \mathrm{v}_j, \\ & 0 \text{ otherwise.} \end{cases} \tag{3.3}$$

Incidence Matrix: the incidence matrix representation of G is a $n\ X\ m$ matrix. If $a_{r,c}$ is the value in the matrix at row r and column c then:

- If G is undirected, $a_{r,c} = 1$ if v_r is the head or the tail of e_c; otherwise, $a_{r,c} = 0$.
- If G is directed, $a_{r,c} = -1$ if v_r is the tail of e_c; $a_{r,c} = 1$ if v_r is the head of e_c, $a_{r,c} = 0$ if e_c is a loop; otherwise, $a_{r,c} = 0$.

With incidence matrix representation, one can easily find which vertices make up an edge.

Other representations: other commonly used network representation methods enumerated in (Golbeck, 2013) include JSON and XML, using standards such as the Graph Markup Language (GraphML) and Friend Of A Friend (FOAF).

3.3.2. Network Structures and Properties

When trying to understand networks, one often want to identify important vertices or get a sense of how interconnected a network is compared to other networks. Beyond vertices and edges, there are some fundamental structures and properties that are important for describing and understanding networks, these include descriptions of vertices, their connections, and their roles in the network (Tsvetovat and Kouznetsov, 2011). What follows next are the fundamental definitions of vertex, edge, and network specific measures that find application in our experimental work.

Vertex and Edge Measures: fundamental vertex and edge specific measures include: degree, degree centrality, closeness centrality, betweenness centrality, eigenvalue centrality, PageRank and local clustering coefficient.

- **Neighbourhood**: for any vertex v_i, in an undirected network, the set of vertices it is connected to via an edge is called its neighbourhood and is represented as $N(v_i)$. However, in directed networks, the vertex v_i has incoming neighbours $N_{in}(v_i)$ which are vertices that connect to v_i and outgoing neighbours $N_{out}(v_i)$ which are vertices that v_i connects to.
- **Degree**: the degree of a vertex v_i in an undirected network, denoted as d_{v_i} is the number of edges connected to the vertex. In the case of directed network, vertices have in-degrees (number of edges pointing toward the vertex) denoted as $d^{in}{}_{v_i}$ and out-degrees (number of edges pointing away from the vertex) denoted as $d^{out}{}_{v_i}$. Some interesting properties of vertex degree include the following:
    - In an undirected network, there is an even number of vertices having odd degree.
    - The summation of all vertex degrees in an undirected network is equal to twice the number of edges, i.e. $\sum_i d_{v_i} = 2|E|$.

- In a directed network, the summation of in-degrees is equal to the summation of out-degrees, i.e. $\sum_i {d^{in}}_{v_i} = \sum_j {d^{out}}_{v_j}$.

- **Centrality:** it defines how important a vertex is within a network. There are many metrics for measuring vertex centrality in a network, each metric taking a different view on what a central vertex is. Commonly used centrality measures include the following:
  - **Degree Centrality**: it ranks vertices with more connections higher in terms of centrality. The degree centrality C_d for vertex v_i in an undirected network is given in (Zafarani et al., 2014) as follows:

$$C_d(v_i) = d_{v_i} \tag{3.4}$$

In a directed network, degree centrality can be measured using (i) in-degrees (${d^{in}}_{v_i}$) which measures prominence or prestige, (ii) out-degrees (${d^{out}}_{v_i}$) which measures gregariousness, or (iii) the combination of both which means ignoring edge directions and equivalent to measuring the degree centrality of an undirected network. These equations also defined in (Zafarani et al., 2014) are given in equations (3.5), (3.6), and (3.7) respectively.

$$C_d(v_i) = {d^{in}}_{v_i} \tag{3.5}$$

$$C_d(v_i) = {d^{out}}_{v_i} \tag{3.6}$$

$$C_d(v_i) = {d^{in}}_{v_i} + {d^{out}}_{v_i} \tag{3.7}$$

It also follows that the degree centrality measure requires normalisation to be able to be utilised for comparing centrality values across different networks. The normalisation can be carried by using any of the following methods (i) maximum possible degree, (ii) maximum degree, or (iii) degree sum. These equations also defined in (Zafarani et al., 2014) are given in equations (3.8), (3.9), and (3.10) respectively.

$$C_d^{norm}(v_i) = \frac{d_{v_i}}{n-1} \tag{3.8}$$

$$C_d^{max}(v_i) = \frac{d_{v_i}}{max_j d_{v_j}} \tag{3.9}$$

$$C_d^{sum}(v_i) = \frac{d_{v_i}}{\sum_j d_{v_j}} = \frac{d_{v_i}}{2|E|} = \frac{d_{v_i}}{2m} \tag{3.10}$$

- **Eigenvector Centrality**: the eigenvector centrality tries to generalise degree centrality by incorporating the importance of vertex neighbours in undirected networks or incoming vertex neighbours in directed networks. The eigenvector centrality ($c_e v_i$) of a vertex v_i defined in (Zafarani et al., 2014) is computed using equation (3.11) as a function of the vertex neighbours' centralities. Adjacency matrix A of the network is utilised in the equation in order to keep track of the vertex neighbours.

$$c_e(v_i) = \frac{1}{\lambda} \sum_{j=1}^{n} A_{j,i} \, (c_e v_j) \tag{3.11}$$

  where λ is some fixed constant. Equation (3.11) can be rewriting as $\lambda C_e = A^T C_e$, where $C_e = (C_e(v_1), C_e(v_2), \dots, C_e(v_n))^T$ is the eigenvector of adjacency matrix A^T and λ is the corresponding eigenvalue. One limitation of eigenvalue centrality is that a vertex despite having many edges connected to it can have its eigenvector centrality value becomes zero. This is the case in directed networks since centrality is only passed when we have (outgoing) edges.

- **Katz Centrality**: in order to overcome the limitations of eigenvector centrality in directed networks, Katz centrality adds a bias term β to the eigenvector centrality value. The Katz centrality value is defined in (Zafarani et al., 2014) and given by:

$$C_{Katz}(v_i) = \alpha \sum_{j=1}^{n} A_{j,i} \, C_{Katz}(v_j) + \beta \tag{3.12}$$

where α is a constant for controlling the effects of the. Similar to eigenvector centrality, Katz centrality also encounters some challenges associated with directed networks. One such challenge is that once a vertex becomes an authority (having high centrality), it then behave in an undesirable manner by passing all its centrality along all of its out links.

- **PageRank**: in order to overcome the limitations of Katz centrality, PageRank centrality divides the value of passed centrality by the number of outgoing links (out degree) from that vertex such that each connected neighbour gets a fraction of the source vertex centrality. PageRank is also defined in (Zafarani et al., 2014) as:

$$C_p(v_i) = \alpha \sum_{j=1}^{n} A_{j,i} \frac{C_p(v_i)}{d_{v_j}^{out}} + \beta \tag{3.13}$$

where $d_{v_j}^{out}$ is non-zero.

- **Betweenness Centrality**: the goal of betweenness centrality it to determine how important vertices are in connecting other vertices, for a vertex v_i, the betweenness centrality is measured by computing the number of shortest paths between other vertices that pass through v_i. Equation (3.14) defined in (Zafarani et al., 2014) is used to measure v_i's betweenness centrality or the central role it plays in connecting any pair of vertices s and t.

$$C_b(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}} \tag{3.14}$$

where σ_{st}, also known as information pathways, is the number of shortest paths from vertex s to t, and $\sigma_{st}(v_i)$ is the number of shortest paths from s to t that pass through v_i. Similar to degree centrality, betweenness centrality needs to normalised before it can be used to compare vertices across networks. This is achieved by

computing the maximum value it takes i.e. $\forall(s,t), s \neq t \neq v_i$, $\frac{\sigma_{st}(v_i)}{\sigma_{st}} = 1$.

- **Closeness Centrality**: the intuition of centrality here is that the more central vertices are, the more quickly they can reach other vertices. This also means that the central vertices have a smaller average shortest path length to other vertices. Closeness centrality is defined in (Zafarani et al., 2014) as:

$$C_c(v_i) = \frac{1}{\bar{l}_{v_i}} \tag{3.15}$$

where $\bar{l}_{v_i} = \frac{1}{n-1}\sum_{v_j \neq v_i} l_{i,j}$ is vertex v_i's average shortest path length to other vertices.

- **Local Clustering Coefficient:** estimates how strongly neighbours of a vertex v_i (vertices adjacent to v_i) are themselves connected. This is also a measure of transitivity (linking behaviour) at the vertex level. Local clustering coefficient is defined in (Zafarani et al., 2014) as:

$$C(v_i) = \frac{Number\ of\ Pairs\ of\ neighbours\ of\ v_i\ that\ are\ connected}{Numbers\ of\ pairs\ of\ neighbours\ of\ v_i} \tag{3.16}$$

Network Measures: measures that can be used to describe the structure of the entire network include: density, degree distribution, connectivity and centralisation. These are defined in (Zafarani et al., 2014) as follows:

- **Degree Distribution**: as the name suggests, degree distribution simply describes the distribution of all vertex degrees in a network. The degree distribution $P(d)$, gives the probability that a randomly selected vertex v_i has degree d_{v_i}. Because $P(d)$ is a probability distribution $\sum_{d=0}^{\infty} P(d) = 1$. In a network with n vertices, $P(d)$ is defined as:

$$P(d) = \frac{n_d}{n} \tag{3.17}$$

where n_d is the number of vertices with degree d. Usually, $P(d)$ is plotted as a histogram in which the x-axis represents the degree (d) while the y-axis represents either the number of vertices having that degree (n_d) or the fraction of vertices having that degree $P(d)$.

- **Connectivity**: a network is connected if there exists a path between any pair of its vertices. A vertex v_1 is connected to a vertex v_j (or v_j is reachable from v_i) if it is adjacent to it or there exists a path from v_i to v_j. A directed network is weakly connected if there exists a path between any pair of vertices, without following the edge directions (i.e., directed edges are replaced with undirected edges). The network is however, strongly connected if there exists a directed path (following edge directions) between any pair of vertices.
- **Components**: a component in an undirected network is a subgraph such that, there exists a path between every pair of vertices inside the component. A component in a directed network is strongly connected if, for every pair of vertices v_i and v_j, there exists a directed path from v_i to v_j and one from v_j to v_i. The component is weakly connected if replacing the directed edges with undirected edges results in a connected component.
- **Centralisation**: using the distribution of vertex centrality measures to understand a network as a whole.
- **Shortest Path**: in a connected network with multiple paths existing between any pair of vertices, the path that has the shortest length $\left(l_{i,j}\right)$ is called the shortest path.
- **Diameter**: the diameter of a connected network is defined as the length of the longest shortest path between any pair of vertices in the network, i.e.

$$diameter = \max_{(v,v_j)\in VXV} l_{i,j} \tag{3.18}$$

- **Transitivity and Clustering Coefficient**: transitivity measures linking behavior or how edges are formed in a social network. A transitive behavior

in a network requires at least three edges which along with the participating vertices create a triangle. Higher transitivity in a network results in a denser network, which in turn is closer to a complete graph. Therefore, we can determine how close graphs are to the complete graph by measuring their transitivity using the global clustering coefficient. Clustering coefficient is the probability that any two randomly chosen neighbours of a vertex v_i in a network are connected themselves. Transitivity is observed when triangles are formed and is analysed using clustering coefficient given by:

$$C = \frac{(Number\ of\ triangles)X3}{\text{Number of connected triples of vertices}} \tag{3.19}$$

where a triple is an ordered set of three vertices, connected by two (i.e., open triple) or three (closed triple) edges.

- **Reciprocity**: counts the number of reciprocal pairs in a network. When all edges are reciprocal, any directed network can have a maximum of $|E|/_2$ pairs, hence its reciprocity can be computed using the adjacency matrix of the network $\boldsymbol{A}$ using equation (3.18).

$$R = \frac{\sum_{i,j,i<j} A_{i,j}A_{j,i}}{|E|/_2} \tag{3.20}$$

- **Density**: a measure of how many edges are in a given set compared to the maximum possible number of edges in the network is or the ratio of the number of edges $|E|$ that a network has over the maximum it can have $\binom{|V|}{2}$:

$$\gamma = \frac{|E|}{\binom{|V|}{2}} \tag{3.21}$$

Once a network has been constructed and measurements have been calculated, the resulting dataset can be used for many applications.

3.3.3. Network Communities

A useful way to understand a large network is to analyse some sections or subgraphs of the network often referred to as egocentric networks, this allows us to identify common social roles and structures. Network communities are groups

of vertices that are similar to each other based on certain criteria, each vertex ends up in a cluster whose vertices are most similar to it irrespective of whether they are connected by an edge or not (Fortunato, 2010). In a network whose vertices can be assigned positions and embedded in an n dimensional Euclidean space, the similarity or dissimilarity between the vertices can be computed using any norm (l_m) such as: the Euclidean distance ($l^2\ norm$) and Taxi cab or Manhattan distance $(l_1 - norm)$, another important class of measures of vertex similarity is based on properties of random walks on graphs (Pons and Latapy, 2006).

A community can be real-world or virtual and explicit (such as LinkedIn, Yahoo, or Facebook groups) or implicit (such as individuals with the same taste for certain movies on a movie rental site or online vendors selling similar products), it describes as a group of individuals with shared interests or characteristics often interacting with one another. In the context of this research, the study of groups or communities in social networks is essential in understanding organised crimes which are mainly observable in a group setting. Among the many tasks geared towards understanding social network groups include community detection and evaluation as well as the temporal analysis of communities (Zafarani et al., 2014).

Community Detection: is aimed at finding implicit communities, for instance, we can use vertices to represent products or Webpages and then use edges between the vertices based on some similarity metrics to detect a group of individuals at the same location, with the same gender, or who bought the same products. Community detection algorithms try to find dense subgraphs in networks through the optimisation of some criteria and the use of heuristics. Given a network $G(V, E)$, the task of community detection is to find a set of communities $\{C_i\}_{i=1}^{n}$ in G such that $\cup_{i=1}^{n} C_i \subseteq V$. There are many community detection approaches and algorithms, the authors of (Zafarani et al., 2014) categorised these approaches as either (i) membership-based, where the interest is on grouping vertices with respect to some attributes or measures such as similarity, degree, or reachability or (ii) group-based, where the interest is on finding communities that are modular, balanced, dense, robust, or hierarchical.

A membership-based community detection algorithm should assign members with similar characteristics to the same community. The algorithm search for subgraphs in a network and only subgraphs that have vertices with specific characteristics (such as vertex similarity, degree or familiarity, and reachability) are considered as communities. Cliques are the most common searched subgraph in networks based on vertex degrees. A clique requires that all objects of a subgraph are connected to each other while a $k-$clique is a complete subset of size k of a network (Sarvari et al., 2014). Clique percolation method (Palla et al., 2005) is a well-known algorithm for searching cliques in a network. In terms of vertex reachability, one seek subgraphs where vertices are reachable from other vertices via a path. Finally, in terms of vertex similarity, one seek to measure the similarities among vertices and assigned the most similar vertices to be in the same community. The concept of structural equivalence where similarity is inferred from the adjacency relationships between vertices is used for networks that cannot be embedded in space. Structural equivalence methods are usually utilised to measure vertex similarities while clustering algorithms are commonly used to find communities (Zafarani et al., 2014).

It also follows from (Zafarani et al., 2014) that the goal of group-based community detection is to seek communities that are modular, balanced, dense, robust, or hierarchical. In balanced group-based community detection, a more balanced and natural partitioning of the network into communities is obtained by utilising variants of minimum cut methods that define an objective function for minimising (or maximising) during the cut-finding procedure. In robust group-based community detection, the goal is to find subgraphs robust enough such that removing some edges or vertices does not disconnect the k-connected subgraph. In modularity group-based community detection, the goal is to measure the distant between real world communities from randomly generated communities in a network. Modularity maximisation tries to maximize this distance. Consider a partitioning of the network G into k partitions $P = P_1, P_2, \ldots P_k$. For a partition P_x, this distance is defined in (Zafarani et al., 2014) as:

$$\sum_{v_1,v_2 \in P_x} A_{i,j} - \frac{d_{v_i} d_{v_j}}{2m} \quad (3.22)$$

It also follows from (Zafarani et al., 2014) that the distance is generalised as in equation (3.23) for partitioning P with k partitions.

$$\sum_{x=1}^{k} \sum_{v_1,v_2 \in P_x} A_{i,j} - \frac{d_{v_i} d_{v_j}}{2m} \quad (3.23)$$

And finally, the normalised version of this distance also defined in (Zafarani et al., 2014) also defined in equation (3.24) as modularity Q.

$$Q = \frac{1}{2m} \langle \sum_{x=1}^{k} \sum_{v_1,v_2 \in P_x} A_{i,j} - \frac{d_{v_i} d_{v_j}}{2m} \rangle \quad (3.24)$$

In modularity group-based community detection, the goal is to seek dense communities which have sufficiently frequent interactions. A network $G = (V, E)$ is dense if $|E| \geq \gamma = \frac{|E|}{\binom{|V|}{2}}$, where γ is the network density (Zafarani et al., 2014). Typical examples of dense communities are cliques, clubs, and clans. Finally, in hierarchical group-based community detection, the goal is to seek communities at a single level but in form of hierarchies with sub or super communities. Hierarchical clustering (agglomerative or divisive) algorithms are commonly used methods for community detection in networks while dendrogram is used to visualise how communities are merged or split using the algorithm. The Girvan-Newman algorithm (Girvan and Newman, 2002) is commonly used to find communities using divisive hierarchical clustering.

There are two major approaches to evaluate how accurately a network community detection task has been performed (Zafarani et al., 2014). The first approach is when the ground truth is available or when there is at least a partial knowledge of what communities should look like, in this case, any of the following evaluation measures are used: precision and recall, F-measure, purity, and normalised mutual information (NMI). Precision (P), which measures the fraction of vertex pairs that the community detection algorithm has correctly assigned to the same community and Recall (R), which measures the fraction of vertex pairs that the algorithm assigned to the same community of all the pairs that should

have been in the same community is defined in equation (3.25) and equation (3.26) in (Zafarani et al., 2014) respectively.

$$P = \frac{TP}{TP+FP} \tag{3.25}$$

$$R = \frac{TP}{TP+FN} \tag{3.26}$$

The terms in these equations are defined thus (i)True Positive (TP): which means similar vertices are assigned to the same community, (ii) True Negative (TN): which means dissimilar vertices are assigned to different communities, (iii) False Negative (FN): which means similar vertices are assigned to different communities, (iv) False Positive (FP): which means dissimilar vertices are assigned to the same community. According to (Zafarani et al., 2014), F-Measure uses harmonic mean to consolidate precision and recall into one measure and is given in equation (3.27).

$$F = 2\frac{PR}{P+R} \tag{3.27}$$

Purity is the fraction of vertices that have labels equal to their community's majority label. First, we assume that the majority of a community represents the community, we then use the label of the majority against the label of each vertex in the community to evaluate the algorithm. Purity is given in (Zafarani et al., 2014) as:

$$purity = \frac{1}{N}\sum_{i=1}^{k} \max_{j} \left|C_i \cap L_j\right| \tag{3.28}$$

where k is the number of communities, N is the total number of vertices, L_j is the set of instances with label j in all communities, and C_i is the set of members in community i. Mutual information (MI) measures the amount of information that two random variables share. MI can be used to measure the information that a cluster or community carries regarding the ground truth. According to (Zafarani et al., 2014), Normalised mutual information (NMI) is given in equation (3.29).

$$NMI = \sum_{h\in H}\sum_{l\in L} n_{h,l} \log \frac{n * n_{h,l}}{\sqrt{\left(\sum_{h\in H} n_h \log\frac{n_h}{n}\right)\left(\sum_{l\in L} n_l \log\frac{n_l}{n}\right)}} \quad (3.29)$$

where L and H are labels and found communities, n_h and n_l are the number of data points in community h and with label l, respectively while $n_{h.l}$ is the number of vertices in community h and with label l and finally, n is the number of vertices. It then follows that an NMI value close to one indicates high similarity between the communities detected and labels. A value close to zero indicates a long distance between them. These methods finds application in our experimental work.

3.4. Inferring Hidden Ties in Crime Data

The main challenge of data extraction for network analysis lies in the choice of network elements (vertices, relationships or edges, and attributes) that can best answer the targeted research questions. Depending on the problem at hand, common choices for these network elements include individuals, groups, organisations, bank accounts, products, URL's, resources, affiliates, and Internet infrastructures for vertices; the common choices for edges include friendship, ownership, distribution, advertisement (the edges can be binary, weighted, directed, undirected, multipartite or multiplex relationship); and finally the choices for attributes are usually additional information such as location, colour, and time. The extraction of these network elements is straight forward in structured data with known connections such as police records or criminal mobile phone records.. This study focused on situations where the direct extraction of these network elements is not feasible. Common approaches include the extraction and linking of network elements from a semi or unstructured data such as text documents and narratives (Ball, 2013), in this case the network elements may constitute the union of all statements per text document, vertices can be concepts or ideational kernels represented by one or more words, while edges are the links between two or more concepts.

3.4.1. Network of Items Co-occurrences

Recall that the goal of this chapter is to infer network elements from crime data in order to unmask social structures that can be disrupted by law enforcement

agencies. Of special interest to this research is organised crime and gang structure which was described in (Ball, 2013) as typically fluid and informal but far from random. A situations in a criminal dataset where there are no data on direct connection among criminals was modelled in a previous work in Appendix IV, the model utilised the concept of co-occurrence in (Leydesdorff and Vaughan, 2006) and (Buzydlowski, 2015) which is simply the counting of paired data within a collection unit (such as an online pharmacy shopping cart with purchased medications, locations, vendors, and buyers as variables in the data). Once the items and collection units are defined, the co-occurrence counts which indicate an association between items can then be derived. This is usually represented by a co-occurrence matrix, with the items of interest occupying the rows and columns while their intersection indicates the co-occurrence. The association between items is made stronger with each additional pairing until non-trivial patterns emerge. The co-occurrence matrix will then finally be transformed into a k-partite network depending on the vertex type. Bipartite or 2-partite network will be used to demonstrate how ties can be inferred from crime data. The following definitions of some basic SNA terms were adapted (Latapy et al., 2008), (Opsahl, 2013), and (Zafarani et al., 2014).

Definition 3.1. A **bipartite network** $G_B = (A, B, E)$ is a network formed by two non-overlapping sets of vertices A and B and by a set of edges E, such that every edge joins a vertex in A with a vertex in B. The number of vertices in set A is denoted by n_a while the number of vertices in set B is denoted by n_b. If V represent the total vertices in the network then $V = A \cup B$, $A \cap B = \emptyset$, and $E \subseteq A\ X\ B$. Bipartite networks are extremely useful in representing the membership or affiliation of one vertex v_i to a group (see Fig. 3.1), for example actors-movies, vendor-products, and researcher-papers affiliations.

Definition 3.2. The **incidence matrix** of a bipartite network $G_B = (A, B, E)$ is $n_a\ X\ n_b$ matrix of elements $A_{x,y}$ defined in (Latapy et al., 2008) as follows:

$$A_{x,y} = \begin{cases} 1 & if\ vertices\ v_x\ and\ v_y\ are\ linked \\ 0 & otherwise \end{cases} \tag{3.30}$$

For unweighted and undirected bipartite network or

$$A_{x,y} = \begin{cases} w_{x,y}\ if\ vertices\ v_x\ and\ v_y\ are\ linked\ with\ weight\ w_{x,y} \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad otherwise \end{cases} \quad (3.31)$$

For weighted and undirected bipartite network.

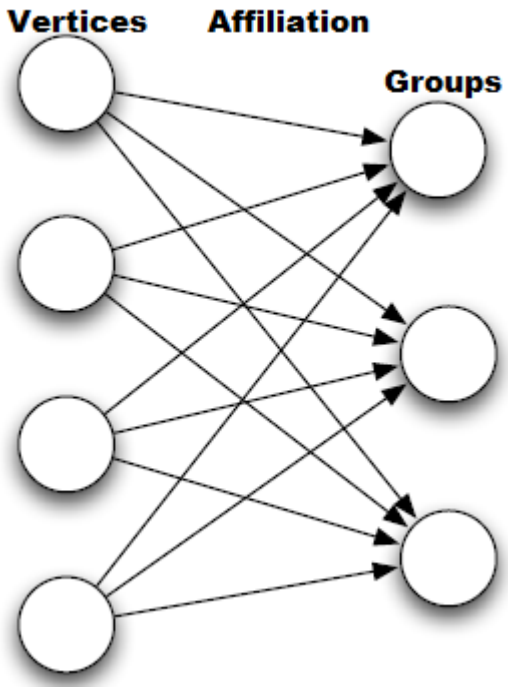


Fig.3.1. A typical bipartite network representation

Definition 3.3. Bipartite network projection is a process of transforming a bipartite network into its unipartite components so that existing methods and algorithms can be applied in the analysis of the network. Fig.3.2 is an illustration in (Kitsak and Krioukov, 2011) of bipartite network projection for nine vertices $(1,2,\dots,9)$ belonging to four different groups (A,B,C,D). The resulting networks in Fig. 3.2(a) and Fig.3.2(b) are unipartite components of the original network.

3.4.2. Bipartite Network Projection

Given a data with labelled instances of crime incidents, the goal here is to construct bipartite network based on co-occurrence concepts and then project the network to its unipartite networks for further analysis. As detailed in Appendix IV, let $A = \{a_1, a_2, \dots, a_i\}$, represent the vertex sets of vendors and $B = \{b_1, b_2, \dots, b_i\}$, represent the vertex sets of products, let's assume only vendors

and products are associated by an edge and that there is no between vendor-vendor or product-product association.

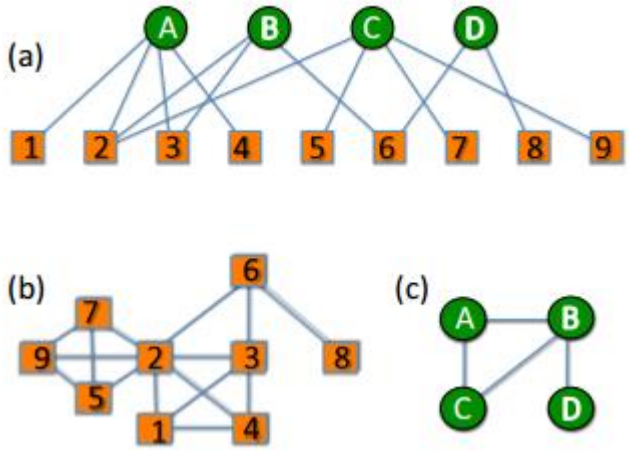


Fig.3.2. Bipartite network (a) and its unipartite projections (b) and (c)

The set of edges or relationships $E \subseteq A\ X\ B$, are the initiated actions such as the manufacture or sale of a product by a vendor while the edge weights represent the total co-occurrences of similar actions. The pairs a_x, b_y denote a vendor a_x, who is associated to a product b_y. The sets A, B, and E can be represented as a bipartite graph $G = (A \mid B, E)$, where A and B are called the partite sets of the network vertices that are connected by an edge iff $(a_x, b_y) \in E$, with $1 \leq x \leq i$ and $1 \leq y \leq j$ where i is the number of unique vendors and j is the number of unique products in the network. The cardinality or the number of edges in the bipartite graph is represented by $n = |E|$. The pairs a_x, b_y can also be represented as b_y, a_x denoting a product b_y associated by a vendor a_x, hence the vendor-product network can be represented by a weighted undirected bipartite network M_{ixj}. In order to infer ties among vendors in the network, the bipartite network M_{ixj} is transformed to its unipartite components A and B or vendor-vendor and product-product network respectively. These networks are obtained in a process called bipartite projection described in (Latapy et al., 2008).

3.5. Experimental Work

This section reports two case studies that were carried out in order to evaluate how ties among criminals can be inferred using bipartite network modelling. The first case study analysed traditional counterfeiting crime (manufacturer-manufacturer) network, and the other analysed Darknet (vendor-vendor) network.

3.5.1. Datasets

We utilised two datasets in this chapter, the first, denoted as D^5 is the MQDB and Poor-Quality Medicines Alert data[22] described in (Krech et al., 2014) while the second dataset denoted as D^6 is the archive of Evolution, an online black market described in (Compton, 2015) that operated on the Tor network from January 2014.

- $\boldsymbol{D^5}$**:** contains results of the collaborative post-marketing surveillance activities from 2003 to 2017 of the Promoting the Quality of Medicines (PQM) program to monitor the quality of medicines in the current 16 participating countries in three continents which include Africa (Ethiopia, Ghana, Kenya, Mozambique, and Nigeria), Asia (Cambodia, Lao PDR, Philippines, Thailand, and Viet Nam), and South America (Bolivia, Colombia, Ecuador, Guatemala, Guyana, and Peru). The MQDB database is made available by the United States Pharmacopeia Convention (USP) and contains 15913 cases (at the time of writing this report) of more than 30 therapeutic indications, information of whether the medicine is counterfeit or not, and many other metadata.
- $\boldsymbol{D^6}$: leaked few weeks after it disappeared on March 18, 2015, and contains nearly 100 GB of daily crawls of the site dating back to its inception.

3.5.2. Rogue Manufacturer-Manufacturer Network

According to the INTERPOL, pharmaceutical crime involves the manufacture, trade and distribution of fake, stolen or illicit medicines and medical devices, it also constitutes the counterfeiting and falsification of medical products, their packaging and the associated documentation as well as theft, fraud, illicit

[22] http://apps.usp.org/app/worldwide/medQualityDatabase

diversion, smuggling, trafficking, illegal trade of medical products and the money laundering associated with it. Pharmaceutical crime data may be composed of a variety of entities such as people, organisations, brands, locations, storefronts, websites, bank accounts, and product delivery agencies. These entities may form informal networks composed of: (i) thousands of storefronts in various locations (ii) affiliate websites run by associates', (iii) many botnet spamming partners who are paid to advertise illicit online pharmacy networks, (iv) covert systems for processing online orders, and (v) regular mail or courier services distributors. Interactions in any informal network formed among these entities will be very difficult to track, this also allows the key actors to evade detection for long periods of time. Once these criminal groups are identified and their habits are known, law enforcement authorities may begin to assess crime trends in order to forecast and hamper the development of perceived future criminal activities (UNODC, 2011). This is the motivation for this experiment, this study believe that the network analysis of pharmaceutical crime data can be useful in modelling indirect relationships among important entities involved (for example the criminals which include manufacturers, advertisers, and distributors, the products they sell, the banks that process their credit and debit card transactions or the delivery services used for shipping the products). The tasks in this case study include (i) data extraction, (ii) network representation, (iii) vertices and network level analysis and (iv) group analysis.

Data Extraction: using the year of sampling criteria, all the data instances of D^5 as described in subsection 3.5.1 was extracted. At the time of writing this report, the database contains about 15,913 instances of medicines collected and tested from 16 countries. The subset of the records with confirmed counterfeiting incidents was also extracted by following the custom data extraction guideline. The dataset was then filtered by removing duplicate records and all the rows that contain Missing, Unknown and N/A instances. The following variables were considered most relevant to the task at hand: Year, Manufacturer, Product, Country, Province, Dosage, Date of Sample Collection, and Test Result.

Network Representation: the co-occurrence concept, as described in subsection 3.4.1 was then applied in order to construct an undirected, weighted bipartite network (see Fig.3.3) between manufacturers (red vertices) with fake

incidences and their associated products (green vertices) and called it rogue manufacturer-product network. In accordance with equation (3.31), the edge weight represents the co-occurrence frequency of the manufacturer-product instances. One fundamental assumption made here is that all manufacturers with at least one product counterfeiting are rogues.

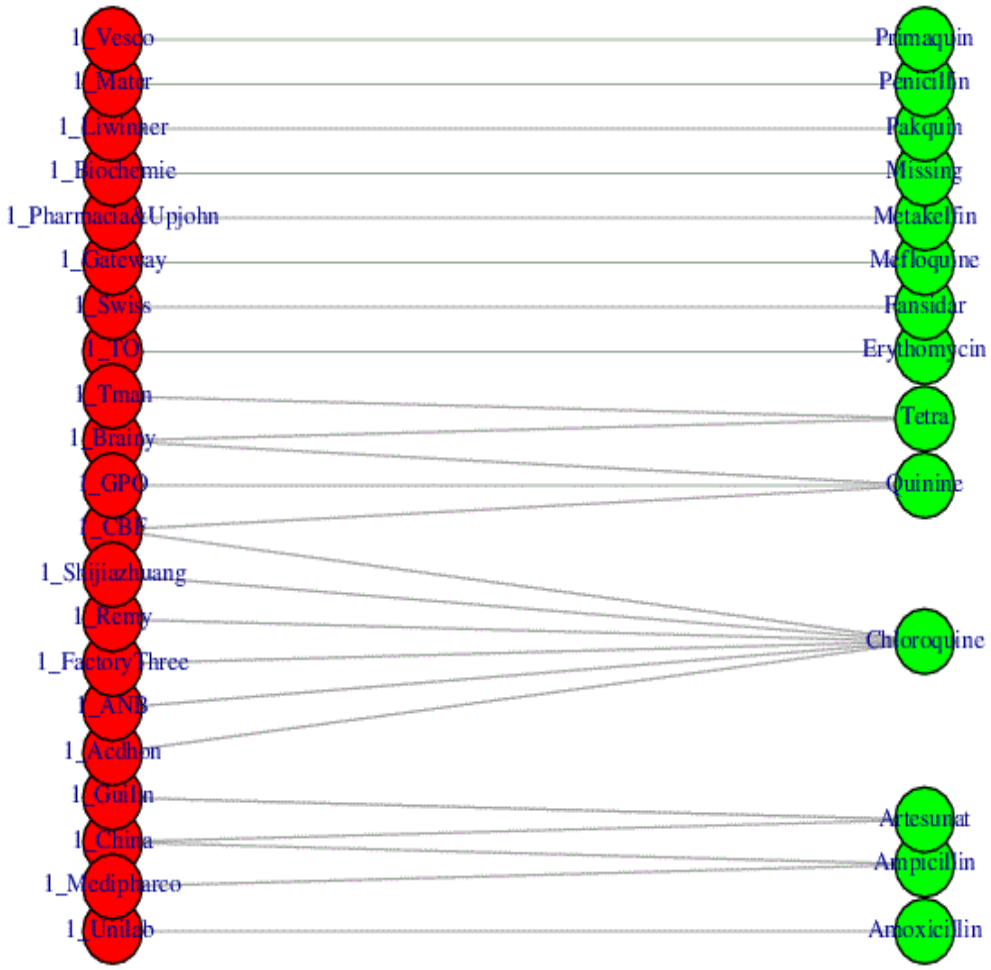


Fig.3.3. Rogue manufacturer-product bipartite network

The unipartite manufacturer-manufacturer network that resulted from projecting Fig.3.3 is shown in Fig.3.4. The resulting network between the manufacturers now becomes the network of interest and can be easily subjected to further analysis.

Manufacturer-Manufacturer Network Analysis: the aggregate metrics of the largest connected component of the network in Fig.3.4 will be reported here. SNA metrics for understanding network structures and properties introduced in subsection 3.3.2 was applied here. Results obtained include: the number of

unique vertices = 9, number of unique edges = 19, geodesic distance (diameter) = 3, average geodesic distance = 1.4321, and network density = 0.5278.

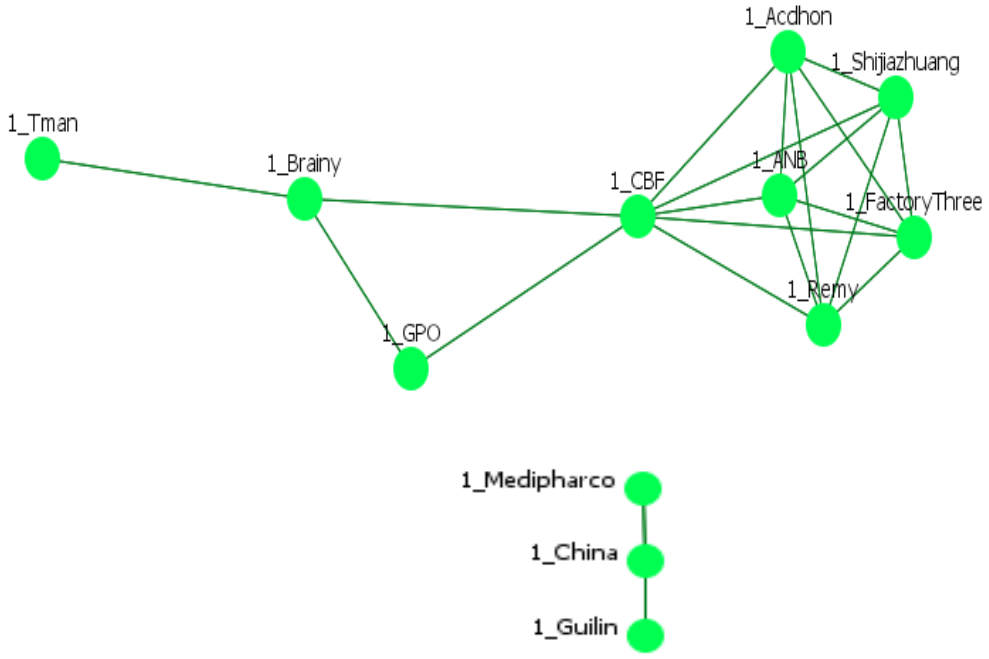


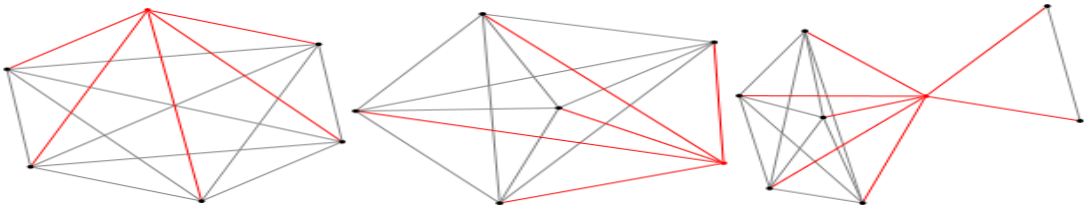
Fig.3.4. Rogue manufacturer-manufacturer network

The vertex-specific network metric results (see TABLE 3.1.) for the larger component was obtained by applying equations (3.4), (3.11), (3.14), and (3.15) for a degree, eigenvector, betweenness, and closeness centralities respectively.

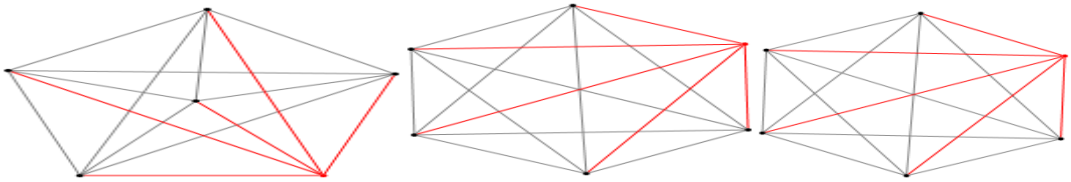
TABLE 3.1. Rogue network vertex-specific metrics

Vertices	Degree	Eigenvector	Betweenness	Closeness
1_FactoryThree	5	0.149	0	0.083
1_ANB	5	0.149	0	0.083
1_Shijiazhuang	5	0.149	0	0.083
1_Remy	5	0.149	0	0.083
1_CBF	7	0.163	15	0.111
1_Acdhon	5	0.149	0	0.083
1_GPO	2	0.040	0	0.071
1_Brainy	3	0.041	7	0.077
1_Tman	1	0.008	0	0.050

Following a similar reasoning in (Nadji et al., 2013), the important vertex in the rogue network based on the results in TABLE 3.1 is 1_CBF, this is because it has the highest degree and centralities. A vertex with the most neighbours (degrees) is often said to be a key member with influence in its local neighbourhood.



(a) 1_Acdhon (b) 1_ANB (c) 1_CBF



(d) 1_FactoryThree (e) 1_Remy (f) 1_Shijiazhuang

Fig.3.5. Subgraphs of the rogue network

Manufacturer-Manufacturer Network Group Analysis: the metrics for understanding groups in SNA introduced in subsection 3.3.3 was also applied, first, the 1.5 degrees egocentric networks of each vertex was extracted and the subgraphs with more than three edges is shown in Fig. 3.5. The star topology of the egocentric network of the vertex 1_CBF indicates its relative importance as a switch or hub in the rogue network. Four different community detection algorithms: Girvan Newman, Clauset Newman Moore, Wakita Tsurumi and Walktrap described in (Fortunato, 2010) was also applied in order to study the natural clusters in the rogue network. TABLE 3.2 is the summary of the community detection results for communities with a minimum of three vertices.

TABLE 3.2. Rogue network communities

Algorithms			
Vertices in each cluster	Girvan Newman	Clauset Newman Moore	Wakita Tsurumi
Cluster 1	1_FactoryThree 1_ANB 1_Shijiazhuang 1_Remy 1_CBF 1_Acdhon	1_FactoryThree 1_ANB 1_Shijiazhuang 1_Remy 1_Acdhon	1_FactoryThree 1_ANB 1_Shijiazhuang 1_Remy 1_Acdhon
Cluster 2	1_GPO 1_Brainy 1_Tman	1_GPO 1_Brainy 1_Tman 1_CBF	1_GPO 1_Brainy 1_Tman 1_CBF
Cluster 3	1_China 1_Guilin 1_Medipharco	1_China 1_Guilin 1_Medipharco	1_China 1_Guilin 1_Medipharco

The result for the Walktrap method is presented in Fig.3.6. One of the possible take away lessons from this study is that when working with massive crime data with location and time attributes, these grouping might signal an element of organisation among the criminals. These naturally occurring clusters may be based on patterns of social ties rather than formal group memberships. These analyses may aid security agencies to further investigate whether the clusters are actually involved in an organised network, the evolution and duration of the network, what sustains it, and how it can be disrupted.

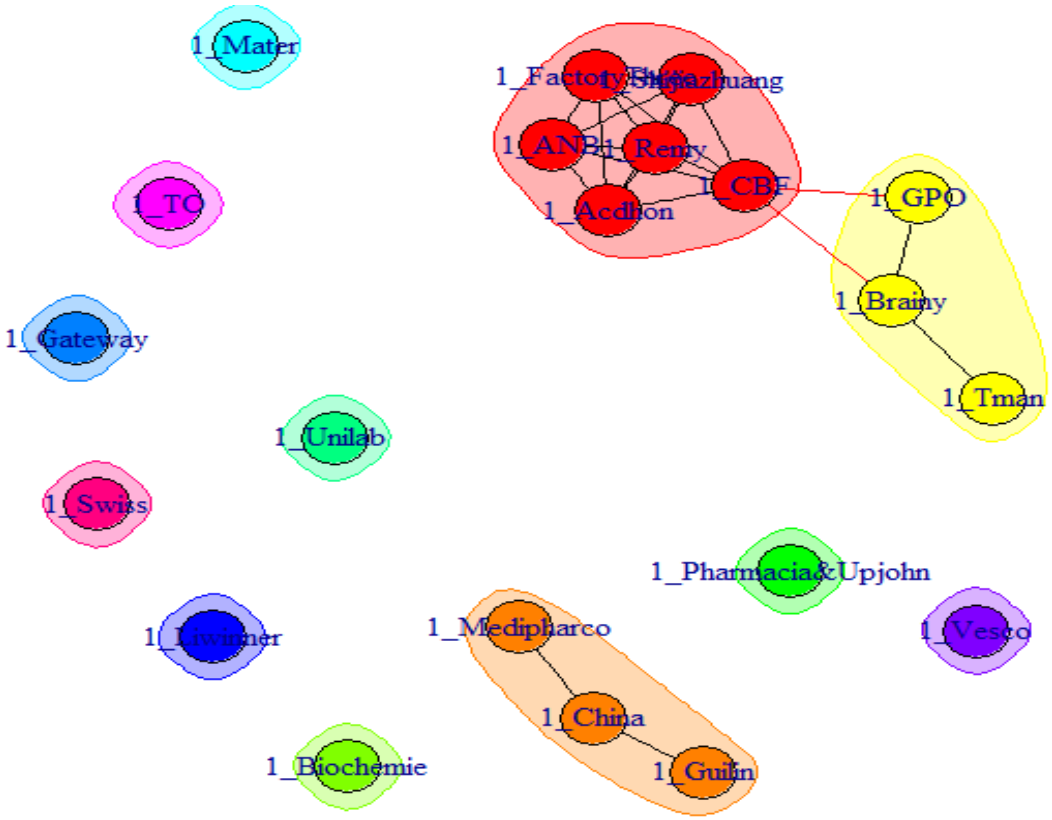


Fig.3.6. Rogue network communities using Walktrap algorithm

Vertices with a central position in their clusters, i.e. sharing a large number of edges with the other group partners, may have an important function of control and stability within the group while vertices lying at the boundaries between modules may play an important role in mediation and lead the relationships and exchanges between different communities. It is interesting to note that most of these incidents were recorded in one country. The rogue network was further subjected to a more strict community detection methods so as to detect cliques. TABLE 3.3 is the clique community detection result.

TABLE 3.3. Network clique communities

Communities	Rogue Cliques
3-clique community	21
4-clique community	15
5-clique community	6
6-clique community	1

There were a total of 43 clique communities starting from a 3-clique community in the rogue network. The size of the largest clique community is 6 with a single clique. Clique communities in the rogue manufacturer-manufacturer network may also lead to the detection of a closely linked organised criminal network.

3.5.3. Darknet Vendor-Vendor Network

The Internet is now a catalyst for illicit online pharmacies, marketing falsified medicines to the public. According to INTERPOL's report "Pharmaceutical Crime on the Darknet" there are several online market places on the hidden part of the internet (Darknet) offering prescription medicines together with cannabis for sale to the public, these market places such as Silk Road, Agora and Evolution provide online access anonymity via anonymising software such as P2P or Tor and payment anonymity via crypto-currencies such as Bitcoin to criminals with shipments being sent across the world between source, transit and destination countries. According to (McCoy et al., 2012b) and (McCoy et al., 2012a), large-scale abusive advertising is a profit-driven endeavour, abuse-advertised goods and services such as spam-advertised Viagra, search-advertised counterfeit software and malware-advertised fake anti-virus has been dominated by an affiliate business model comprised of independent advertisers acting as free agents acquiring traffic via spam or search, and in turn paid on a commission basis by their sponsors who handle the back end, customer service and payment processing. Counterfeit pharmaceutical affiliate business models such as GlavMed, SpamIt and RX-Promotion involve a range of sponsors providing drugstore storefronts, drug fulfilment, shipping, payment processing customer service and independent advertisers or affiliates who are paid a commission for promoting the program using botnets to send spam email or manipulating search engine results (McCoy et al., 2012b). The next experiment will look into how networks can be inferred from darknet transactions based on the proposed

bipartite network model for inferring ties in crime data. It will also showcase the analysis of structures and key persons in the inferred networks.

Data Extraction: The leaked dataset D^6 described in subsection 3.5.1 was downloaded and a sample extracted, the data consists of the product that each vendor sells each day categorised under top-level categories such as Drugs, Digital Goods, Fraud Related, and Weed. These are further subdivided into product-specific pages and each page contains several listings by various vendors as illustrated in TABLE 3.4 also available in (Compton, 2015). All records for Analgesics category was extracted and more than 5000 records with the variables "Vendor", "Products", and "Date" was obtained.

TABLE 3.4. Sample of darknet market data

Vendor	Products
MrHolland	Cocaine, Cannabis, Stimulants, Hash
Packstation24	Accounts, Benzos, IDs & Passports, SIM Cards, Fraud
Spinifex	Benzos, Cannabis, Cocaine, Stimulants, Prescription, Sildenafil Citrate
OzVendor	Software, Erotica, Dumps, E-Books, Fraud
OzzyDealsDirect	Cannabis, Seeds, MDMA, Weed
TatyThai	Accounts, Documents & Data, IDs & Passports, PayPal, CC & CVV
PEA_King	Mescaline, Stimulants, Meth, Psychedelics
PROAMFETAMINE	MDMA, Speed, Stimulants, Ecstasy, Pills
ParrotFish	Weight Loss, Stimulants, Prescription, Ecstasy

Data Representation: an undirected, weighted bipartite network between vendors and their associated products was constructed and called a Darknet vendor-product bipartite network. For illustration purpose, Fig. 3.7 shows the resulting network for the nine instances in TABLE 3.4 with the vendors in bold blue while the products are in black colour. The edge weight (not shown) represent the co-occurrence frequency of vendor-product instances.

Vendor-Vendor Analysis: the network was then transformed to its unipartite components, the vendor-vend43or network obtained consists of 102 vertices, 952 edges and 4 connected components and is shown in Fig.3.7. As in the previous experiment using MQDB dataset, the largest connected component of the network is shown in Fig.3.8.

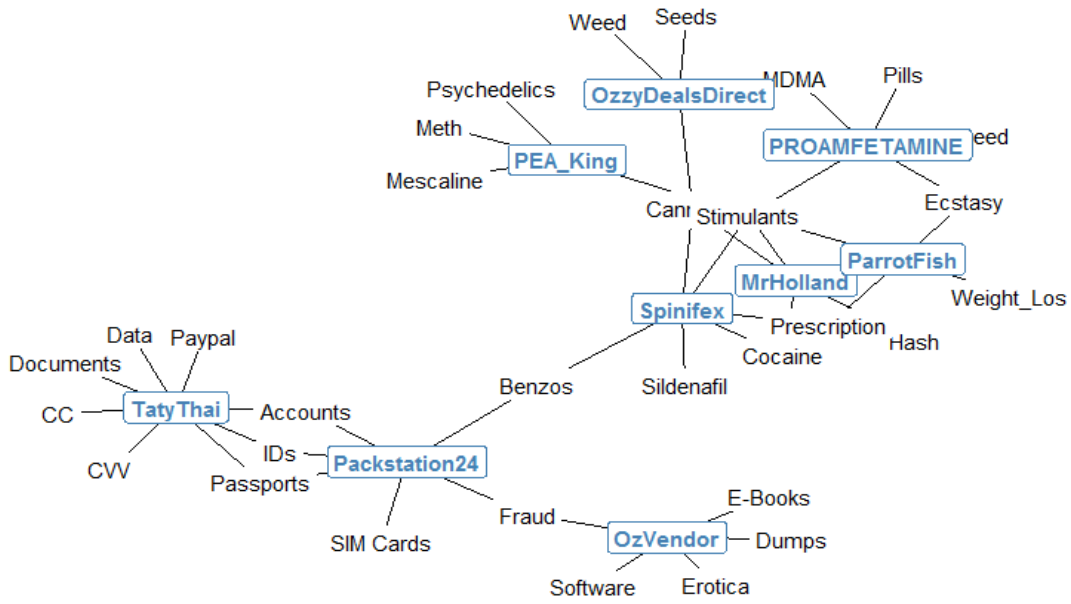


Fig. 3.7. Darknet vendor-product bipartite network for nine vendors

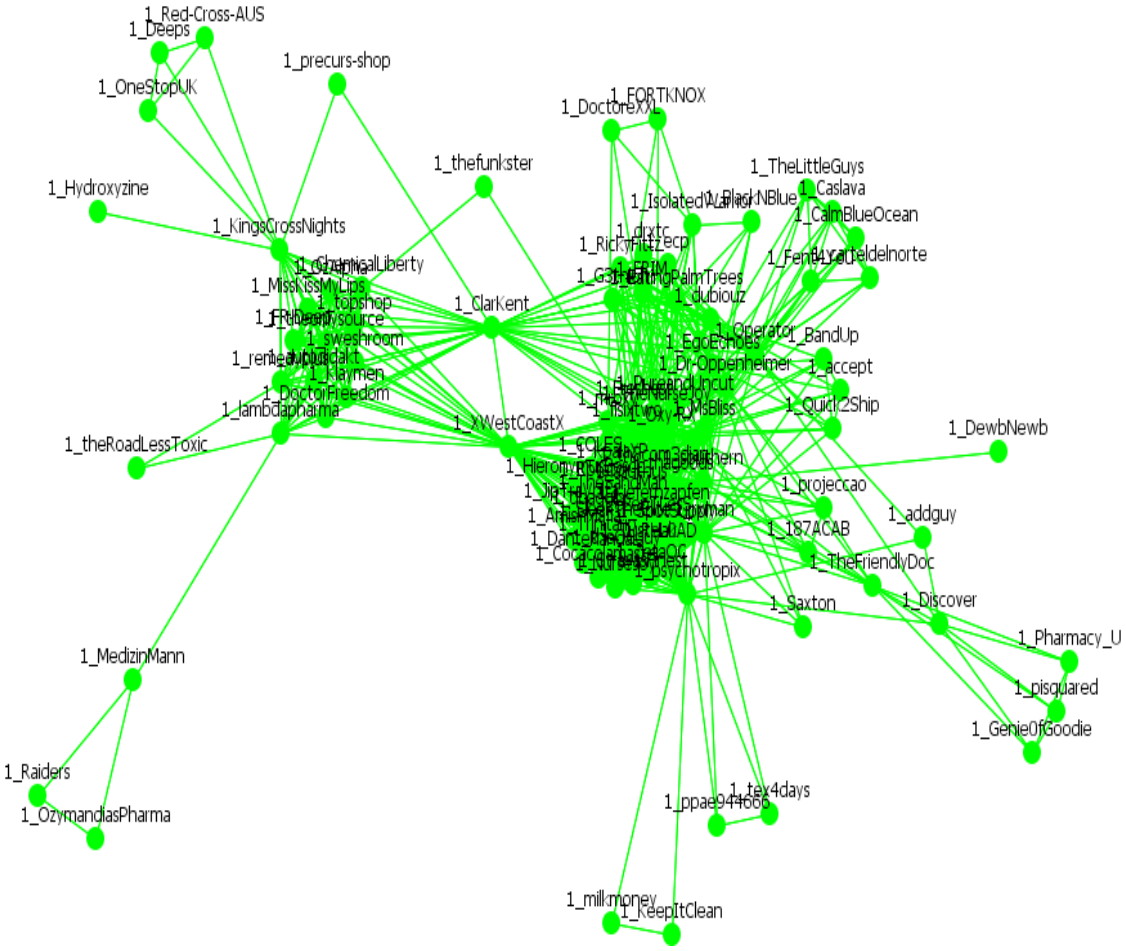


Fig.3.8. Darknet vendor-vendor network

The aggregate metrics of this connected component include: number of unique vertices = 95, number of unique edges = 947, geodesic distance (diameter) = 6, average geodesic distance = 2.1682, and network density = 0.2121. The important vendor in the Darknet network is 1_TheNurseJoy, it has the highest degree and centralities. Four clusters each of vertices 35, 26, 24, and 10 respectively were obtained when Clauset Newman Moore community detection algorithm described in (Fortunato, 2010) was applied. There were, however, 7 and 24 communities found with Wakita Tsurumi and Girvan Newman algorithms respectively. The size of the largest clique community is 28 with a single clique. Fig. 3.9 is the plot of the resulting communities obtained using Louvain method described in (Blondel et al., 2008). The result clearly shows a separation of the vertices into two communities (green and pink colours), these might be a network of vendors working together or just happenstance. The result can serve as a lead and it is meant to help security agencies to efficiently allocate resource.

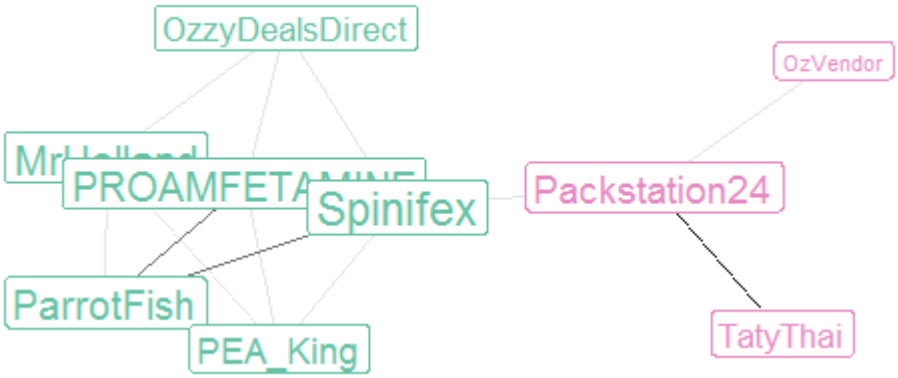


Fig. 3.9. Darknet network communities for the nine vendors

3.6. Issues with Social Network Data Mining

Despite its potentials, network data also has pitfalls. In recent times, network data seems to be easily and quickly accumulated and mined from various API's. However, the ease with which network data can be obtained does not necessarily correlate with quality inferences (Danneman and Heimann, 2014). Again these API's consistently undergo changes with respect to the accessibility of the data and also in the way they work. The quality of network data obtained from social media platforms need to be checked before any analysis, this is because many of the online accounts that form the network are either missing some vital

information, are anonymous, or even created by minors, as such there is always a possibility for skewness in the analysis results (Ravindran and Garg, 2015).

There are ongoing efforts by researchers to develop methods for checking network data quality, for instance, the thesis in (Debattista, 2017) developed a framework to analyse a variety of quality problems in linked datasets and enable consumers to make educated decisions when searching for datasets that are “fit for use”. It is our desire to improve on these efforts in order to advance research on social data quality analysis. As stated in the previous chapter, the difference between the existing research and ours is the novelty and the methods of estimation of quality which will be introduced in chapter four.

3.7. Conclusions

Many types of crimes require a high degree of organisation and specialisation. Again, many criminal networks show various levels of organisational structures, they may operate as swarms, hubs, hierarchies, aggregates, or hybrids according to their activity, the degree of cohesion, or organisation. The organisation of crime may also occur in the darknet (with a restricted network where individuals interact within online discussion forums and chat rooms) for illegal trade, communication, and peer-to-peer file sharing. Identifying criminal groups within a network, mapping their relationship to one another, and understanding their command structure can be essential in strategic intelligent decision makings. SNA is being used to automatically identify criminal clusters that may not have been obvious in a crime dataset. These naturally occurring clusters are based on patterns of social ties rather than formal group memberships. Identifying clusters and their boundaries in a criminal network allow for a classification of vertices according to their structural position in the groups. These results may be useful for operational strategies, for instance, vertices with a central position in their clusters may have an important function of control within the group while vertices lying at the boundaries between clusters may play an important role between two communities. However, situations often arise in a criminal dataset where we lack direct connection among criminals. This chapter focused on answering RQ2 which is aimed at learning hidden relationships between suspicious contents and actors in crime data. This aim was achieved by utilising the concept of item co-occurrence and bipartite network projections. The model was evaluated using two

case studies and the results obtained were very significant and can reveal some hidden ties among criminals that were not immediately obvious in the data. Beyond this dataset, SNA can also be used to reveal relationships between user accounts sending pharmaceutical spam and the spam URL's. While the objectives of this chapter were achieved by answering RQ2, future study should consider working closely with intelligence agencies on real life scenarios. Further development of network data quality methods for tackling quality issues in a huge dataset of relationships is also highly recommended.

Chapter Four: Social Data Quality Estimation

4.1. Introduction

Social data from user generated contents on social media platforms are becoming rich information sources for both individual and organisational needs. For example, question and answer portals provide a valuable knowledge repository that can serve as a gold mine for information retrieval (Bian et al., 2008) and a viable alternative to Web search (Bian et al., 2009). Customer reviews are actively being utilised by other prospective consumers to judge the quality of products or services before purchase and for knowledge discovery (Ott et al., 2011). With millions of active users around the world posting Tweets and updates on other social media platforms, real time search and tracking of events and news before they are reported in the mainstream media becomes easy (Benevenuto et al., 2010). Again, user generated content on social media is gaining attention as a source of descriptive annotations for digital objects such as photos or videos (Momeni et al., 2013). However, the presence of noise in social data has led to a growing concern for measuring content quality on social information networks (Agichtein et al., 2008). The noise, which is also called contaminants in this study is due to the "fluid" nature of social data sources which creates an opportunity for suspicious and automated accounts to introduce polluted contents (Benevenuto et al., 2009). For instance, Twitter's popularity is making it a fertile breeding ground for spammers to proliferate, often masking themselves as legitimate users latching trending topics, generating traffic and revenue (Fabrício et al., 2010), and compromising users' security and privacy (Yardi et al., 2009). Automated malevolent bots are now being designed and deployed to spread spam messages (Freitas et al., 2016), commercial advertisements, and radical opinions (Clark et al., 2016).

The quality of social data is still a major concern for users as well as in data analytics, this is because distinguishing between good and bad instances of

social data is difficult due to issues such as the veracity of contents and accounts, the large volume of content being created per time, the velocity with which social data are being created, and best practice abuse due to difficulties in content moderation. Unlike conventional data sources such as surveys and operational databases, social data sources are inherently open to users who can create accounts, network with one another, generate contents, interact, share and consume information (Chai, 2011). Such openness and interaction make the basis of social data richness and popularity. However, maintaining this openness is important for the continued growth and impact of social data sources. The challenge is that as the availability and openness increases, the task of identifying high-quality social data becomes increasingly important (Agichtein et al., 2008).

Efforts are, however, being made by researchers to automate the measurement of social data quality, notable ones include the utilisation of clues from collaboration and edit history in measuring the quality of wiki articles in (Lim et al., 2006). One of the fundamental challenges of tackling social data quality that still remains unsolved is the difficulty of defining the concept of quality. We all have our own view of what quality really means, the same content appearing in two different communities may be perceived at different quality levels. Most approaches for detecting and filtering spam in social data are currently focused on developing automated techniques for profiling and classifying suspicious accounts from legitimate ones. Despite these efforts, automated accounts are still evolving and developing strategies to emulate organic account activities. Again, while these accounts can be identified and suspended, the same users can create new accounts and resurface again (Clark et al., 2016); some 'normal' users are also in the habit of spamming trendy discussions to (self-)promote their own agenda. It is, therefore, an interesting challenge to characterise social data quality and then develop customised techniques for filtering noise in social data. Research efforts are still needed in the area of identification and measurement of noise in social data as well as in evaluating whether or not a given instance or sample of social data is good or bad for a given information need. These concerns led to the novel research question (RQ3) as follows:

- **Research Question 3 (RQ3):** can methods and standards be developed for measuring, predicting, and comparing the quality level of social data instances and samples?

This question is further broken down into two parts. The first part, which is the focus of this chapter, looks into methods for measuring and predicting the quality level social data instances. The second part, which will be detailed in the next chapter, focuses developing benchmarks for social data quality comparison. This chapter proposes the adoption of the methods for water quality monitoring to characterise and develop methods for measuring the quality of social data. It is targeted at identifying and penalising noise or spam content itself in social data rather than accounts. The justification of the water versus social data analogy is due to the following three main reasons (i) the science of water quality evaluation is very well established, (ii) water bodies can suitably be compared with archived social data while flowing river can be likened to social data streams, and (iii) in our opinion, the indicators and parameters for measuring water quality show good adaptability to measure social data quality.

The rest of this chapter is organised as follows, section 4.2 is a concise review of related literature on social data content quality assessment, section 4.3 highlights the analytical effects of noise in social data, section 4.4 introduces fundamental concepts, definitions, and proposed algorithms for estimating the level, effects, and dynamics of noise in social data, section 4.5 presents results obtained and lessons learnt through experimental work using datasets from Twitter, and finally section 4.6 is the chapter conclusion which provides an overview of all objectives met and a pointer to our future work.

4.2. Literature Review

In today's data driven world where individuals and organisations are looking to use crowd feedback for analysing operations and finding ways to improve, bad data can lead to poor analysis and bad business decisions. The issue of poor data quality is even more challenging in social data due to the loose user control nature. For instance, many wiki sites allow multiple content edits and contributions often without requiring the editors and contributors to create user profiles (Lim et al., 2006). Another major challenge with social media is the

presence of noise such as spam, malware, unwanted or duplicate updates, misinformation, and anything that is not relevant or valuable. This noise misleads, exploit and manipulate social media discourse sometimes resulting in several levels of societal harm, a typical example is the artificial inflation of support for a candidate in an election which can endanger democracy by influencing the election outcomes (Ferrara et al., 2016).

Social data quality has been defined differently by different authors, for example, the ratio of the number of out of vocabulary (OOV) words to the total number of words was used in (Duan et al., 2010) to measure the language quality of Tweets. The authors also refer to the number of words as a measure of the information richness of a Tweet, as such a long sentence is apt to contain more information than a short one. In terms of accounts, the authors proposed four metrics to measure the influence of users' authorities on Tweets, these include the number of followers a user has, the number of times a user is referred to in Tweets, the number of lists a user appears in, and the popularity score of users based on retweet relations and PageRank algorithms. The authors also hypothesised that (i) users who have more followers and have been mentioned in more Tweets, listed in more lists and retweeted by more important users are thought to be more authoritative, and (ii) a Tweet is more likely to be an informative Tweet rather than pointless babble if it is posted or retweeted by authoritative users. In question and answer portals where quality refers to interesting, well formulated and factually accurate content, explicit feedback from users was used as a quality indicator in (Agichtein et al., 2008), (Bian et al., 2009), and (Pelleg et al., 2016). The quality of comments on images and videos on Flickr and YouTube was described in (Momeni et al., 2013) as the comment usefulness from the user's perspective, i.e. a comment is useful if it provides descriptive information about an object beyond the usually very short title accompanying it. In the case of microblogs, the criteria used for judging the quality or interestingness of Tweets in (Choi et al., 2012) and (Webberley et al., 2016) was the count of content influenced retweets while the best criteria in (Vosecky et al., 2012) include well-formedness, factuality, and navigational features.

The methodological approach to social data quality measurement in (Agichtein et al., 2008) is a binary decision tree model developed to automatically

learn and separate high-quality content from the rest using features derived from human-labelled quality judgments. A logistic regression model was used in a semi-supervised learning approach in (Bian et al., 2009) and in a supervised context in (Pelleg et al., 2016) to learn the content quality of a question and answer forums. A rule learner, fuzzy logic classifier, and Support Vector Machines were used in (Chai, 2011) to measure and evaluate the quality of Web forum posts using the forums content, usage, reputation, temporal and structural features. The authors of (Choi et al., 2012) proposed a retweet-based quality model that find and promote informative Tweets for improved microblog search. Support Vector Machines was utilised in (Vosecky et al., 2012) to filter and rank Tweets based on a quality score in order to improve Twitter's basic search functionality. A Bayesian network classifier was used in (Webberley et al., 2016) for the collective assessment of Tweets interestingness.

Twitter initially presents Tweets in chronological order, it later introduced a new ranking strategy that considers the popularity of Tweets in terms of a number of retweets. The new ranking still allows a lot of pointless Tweets to flood the relevant Tweets. In (Duan et al., 2010), a new Tweet ranking method was proposed based on learning to rank strategy which uses the content relevance of a Tweet, the account authority, and Tweet-specific features such as whether a URL link is included in the Tweet. The authors reported that the best feature conjunctions include whether a Tweet contains URL or not, length of Tweet and account authority. Twitter users previously follow their interests and rationale when deciding whether a retrieved post is of interest to them or not. In their pursuit to build search algorithms that are capable of identifying interesting and relevant social post for a given topic, the authors of (Ke et al., 2012) studied whether there exist additional characteristics of social data that are more predictive of its relevance and interestingness than the measure of the keyword similarity with the query. More recently, Twitter itself announced a quality filter (Leong, 2016) for penalising lower-quality contents.

This study, however, is beyond Twitter and differs from the above listed studies in scope, features and estimation of quality. It extends the social data content quality analysis method to the next level by adapting the methods of monitoring of contaminants in water to monitor social data content quality.

4.3. Water vs Social Data Quality Analogy

Here, some fundamental quality characteristics and concepts similar to both social data and water will be outlined, this will be followed by the formulation of the social data quality challenge in analogy with current methods of measuring water quality. So, what connects water and social data? Let's begin by describing water, which is a universal solvent generally regarded in its pure form as a transparent, tasteless, odourless liquid with a freezing point of 0°C and a boiling point of 100°C or in its less impure state such as rain, oceans, lakes, rivers; water has a tendency to pick up "foreign" materials through which it flows (Spellman, 2013). In a presentation[23] on how dirty data and non-consumer posts distort the insights brands gain from social media, Networked Insights stated that social media, because of its "fluid" nature and openness, also harbours "foreign" contents created by malevolent bots or computer algorithms that masquerade as humans.

In the same manner that water is used for drinking, agricultural, and industrial processes, social data are fast becoming the hottest commodity in market research (Clark et al., 2016). Social data have become a source of interest in public health and economic research crisis management, on-the-fly campaign assessments, and sentiment analysis[24]. However, both water and social data are fundamentally vulnerable to contaminants (impurities) that affect their quality. Threats to water quality are manifold, from industrial wastes and radiation to natural phenomena such as dissolved bacteria, algae, and solids (Postolache et al., 2012) while threats to social data quality in the context of this study include deceptive and abusive links, tags, and accounts, duplicate instances, and non-informative terms. Therefore, as much as it is of paramount importance to keep water quality at its highest level possible, it is also important to measure and assess social data quality before taking measurements and decisions.

In order to measure and categorise water into appropriate quality scores, it is either respondent (the water users) are asked to assign a quality score within a given range or by taking measurements of indicators of water samples at

[23] https://www.slideshare.net/networkedinsights/networked-insights-how-dirty-is-big-data-2
[24]https://www.millwardbrown.com/Insights/Point-of-View/Social_Measurement/

different experimental conditions. Both user feedback and experimental values are then transformed to standard water quality units. There are many metrics for such transformations, one popular metric[25] that combine all the above characteristics in order to rank the quality of different water samples is called water quality index (WQI) represented in the ranges: 0-24 (very bad), 25-49 (bad), 50-69 (fair), 70-89 (good), and 90-100 (excellent). The WQI score and its variants are used to compare data from several sites as well as to look at trends over time on a single site (Wilde et al., 1998).

TABLE 4.1. Basics of social data and water quality analogy

	Social Data Quality (Proposed)	Water Quality (Survey, 2015)
Goals	To measure and identify whether instances or samples of social data are meeting designated uses To identify common social data contaminant and their sources To analyse social data quality trends over time To measure the analytical effects of contaminated instances	To measure and identify whether waters are meeting designated uses To identify water contaminant and sources To analyse water quality trends over time To screen for water quality impairment
Input	Social data sample: posts, updates, comments/replies, ads	Water sample: lakes, rivers, reservoirs, groundwater
Output	Social data sample quality score, status, index, or rank	Water sample quality score, status, index, or rank
Contaminants (Quality Indicators)	Deceptive links and hashtags; characters, words, tags, links, and symbols created to amplify the visibility of misleading information, etc.	Sediments, solids, wastes, microorganisms, toxic substances, disinfectants, inorganic and organic chemicals, etc.
Metric	Social Data Quality Index ($SDQI$)	Water Quality Index (WQI)
Application	To study the extent to which contaminants can skew data analytics result. To compare the quality of different social data samples.	To evaluate or compare the suitability of water for different applications. To select appropriate water treatment technique.

[25] http://www.keithalcock.com/creeks/awqi.shtml

Constraints: while there are no established standards for social data quality, crowdsourcing (social data users and experts) is being utilised to label social data instances for a specific application into different categories of interest. Two other issues that may account for differences in social data vs water quality characterisation were also noted, first, unlike water, social data sample is a collection of many instances or datum each of which may have different contaminant; and second, different social data sources have different quality indicators or features, for example, the quality features for Twitter and Instagram data sample may include the presence or counts of irrelevant, deceptive, or misleading links and hashtags while that of ads on listing Websites may include illicit or prohibited ads. These issues, however, do not invalidate our analogy but calls for specific approaches. TABLE 4.1 describes the analogy of the fundamental concepts of social data and water quality. The next section will highlight the analytical effects of contaminants in social data.

4.4. Fundamental Definitions and Metrics

Before getting into social data quality, this study first provides the definition of some fundamental concepts and metrics of data quality as follows:

Definition 4:1. Quality: according to the online Oxford Dictionary, quality is the standard of something measured against other things of a similar kind or the degree of excellence of something. The International Organization for Standardization (ISO) however defined quality as the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs.

Discovering whether data are of acceptable quality is a very difficult task (Herzog et al., 2007). The authors of (Pipino et al., 2002) reported that data and information are often used synonymously, but in practice, the two concepts are often differentiated by describing information as processed data; they then utilised the quality dimensions in (Kahn et al., 2002) and proposed subjective and objective metrics for improving organisational data quality assessment and improvement. In this work, we will be using the term data to refer to both processed and unprocessed data. These data quality dimensions defined in (Kahn et al., 2002) include the following:

- **Accessibility**: defined as the extent to which data is available, or easily and quickly retrievable.
- **Appropriate Amount**: defined as the extent to which the volume of data is appropriate for the task at hand.
- **Believability**: defined as the extent to which data is regarded as true and credible.
- **Concise Representation**: defined as the extent to which data is compactly represented.
- **Consistent Representation**: defined as the extent to which data presented in the same format.
- **Ease of Manipulation**: defined as the extent to which data is easy to manipulate and apply to different tasks.
- **Free of Error**: defined as the extent to which data is correct and reliable.
- **Interpretability**: defined as the extent to which data is in appropriate languages, symbols, units, and the definitions are clear.
- **Objectivity**: defined as the extent to which data is unbiased, unprejudiced, and impartial
- **Relevancy**: defined as the extent to which data is applicable and helpful for the task at hand.
- **Reputation**: defined as the extent to which data is highly regarded in terms of source or content.
- **Security**: defined as the extent to which access to data is restricted to maintain its security.
- **Timeliness**: defined as the extent to which data is sufficiently up-to-date for the task at hand
- **Understandability**: defined as the extent to which data is easily comprehended
- **Value Added**: defined as the extent to which data is beneficial and provides advantage from its use.

However, according to the authors of (Herzog et al., 2007), the seven most commonly cited data quality dimensions are relevance, accuracy, timeliness, accessibility and clarity of results, comparability, coherence, and completeness.

They also described high quality data and how they can be obtained. According to them, high quality data are those that meet the following criteria (i) fit for use in their intended operational, decision-making and other roles and (ii) conformance to standards; these can be achieved by prevention (filtering bad data at the point of entry), detection (proactively looking for bad data that had already entered), and finally, repair (fixing bad data). Big data, according (Cai and Zhu, 2015), faces the following quality challenges: (i) increased difficulty in data integration due to the diversity of data sources, types, and complexity in structures, (ii) difficulty in judgement within a reasonable amount of time due to its tremendous volume, (iii) difficulty in real time processing and evaluation because its timeliness is very short and (iv) absence of unified and approved standards. The authors then formulated a dynamic quality assessment process with a feedback mechanism for big data. There are many interesting data quality definitions in the literature, the following are fundamental to this study:

- **Quality of Forum Post:** (Weimer et al., 2007) and (Nayer et al., 2008) defined several quality dimensions (surface, lexical, syntactic, forum-specific, originality, relevance, posting-component, and similarity features) for modelling and predicting how an online forum community perceives post quality using supervised classification techniques that learn from human ratings to categorised posts into predefined quality classes. According to the authors, the rating represents a seed value for each post and can be leveraged in filtering the forum content.
- **Quality of Product Review:** (Liu et al., 2007) defined the quality of product review based on four categories (best, good, fair, and bad) which represent different values of the reviews to users' purchase decision. Given a training dataset $D = \{x_i, y_i\}_1^n$, the authors then developed a model that can minimise the error in prediction of y given x in order to address the problem of detecting low-quality products reviews. The terms $x_i \in X$ and $y_i = \{high\ quality, low\ quality\}$ represents a product review and a label. The study in (Lu et al., 2010) also defined several combination of features including textual, author, and social network information for assessing review quality.

- **Quality of Question and Answer:** (Agichtein et al., 2008) defined several features (content, usage, and user relationship) for ranking the quality of questions and answers in a forum and modelled a binary classifier that learnt to automatically separate high quality content from the rest. The study in (Bian et al., 2009) defined the following concepts regarding the content of question and answer forums, first, the quality of a question measured as a score between 0 and 1 indicating a question's effectiveness at attracting high-quality answers, second, the quality of an answer also measured as a score between 0 and 1 indicating the responsiveness, accuracy, and comprehensiveness of an answer to a question, third, the answer reputation measured as a score between 0 and 1 indicating the expected quality of a user's answers, and finally, question reputation measured as a score between 0 and 1, indicating the expected quality of the user's questions. The Thesis in (Chai, 2011) also defined several quality dimensions (content, usage, reputation, temporal and structural) for assessing the quality of user generated contents.
- **Quality of Tweet:** (Choi et al., 2012) and (Miller, 2015c) defined the quality of a Tweet as the percent of readers who will retweet it and it takes into account the passage of time, as well as retweeting behaviour outside the author's immediate network.

Most of the above definition of social data quality focused on the reputation of the source and user feedback such as the number of retweets, likes and comments. However, because of the "fluid" nature of control and moderation of social media discussions, this study argued that for a complete quality evaluation, both the fitness for use and conformance to standards should also be taken into consideration. In addition to the motivation for adapting water quality evaluation methods in section 4.1, this is also one of the reasons why this research is focused on measuring those objects that appear in social data contrary to established standards and best practice policies rather than measuring social data quality directly from user feedback. The level of these "foreign" objects either in social data instance or sample will then be used to infer social data quality.

The definition of some fundamental concepts of social data quality will be outlined here, this will be followed by the formulation of the social data quality

challenge in analogy with the current methods of measuring water quality. To effectively explore the water and social data parallel, this study proposes that a social data lake collected at a finite time is analogous to a water lake, while real-time streaming social data is analogous to flowing water in a river with many sources and tributaries. A typical example of a water lake is a swimming pool, while an example of social data lake is the “Disasters on social media” data sample in CrowdFlower[26] data library. The scope of this current work is centred on social data lake, however, suggestions on how the methods can be extended to real-time social data applications have been provided.

Definition 4.2. Social Data Contaminant: defined as objects such as duplicate instances, non-informative terms, repeated links, tags, and other entities created and embedded in social data in order to amplify the visibility of misleading information. For example, the images in Fig. 4.1, Fig. 4.2 and Fig. 4.3 are Tweets with contaminant features. These data instances can degrade analytic, search and other quality expectations of social data as such should be penalised. Social data instances containing contaminants are also referred in this study as contaminated. These social data contaminants are analogous to water contaminants which are objects in the water that can degrade water quality, such include organic and inorganic chemicals, sediments, wastes, microorganisms, toxic substances, and disinfectants. The presence of contaminants and water characteristics are used to indicate the quality of water (Postolache et al., 2012). This concept is, therefore, extended to social data.

Fig. 4.1. Tweet with replicate or repeated hashtags

Fig. 4.2. Tweet with multiple unrelated hashtags

Fig. 4.3. Tweet with multiple username mentions

[26] https://www.crowdflower.com/data-for-everyone/

TABLE 4.2 is a list and description of variable notations that finds application in this chapter.

TABLE 4.2. List and description of variable notations

Notation	Description
$\boldsymbol{S}$	a sample of social data lake
$\boldsymbol{S_u}$	an unlabelled subset of S
$\boldsymbol{S_l}$	a labelled subset of S
$\boldsymbol{S_{l-}}$	the contaminated subset of S_l, also the target class
$\boldsymbol{S_{l+}}$	a non-contaminated subset of S_l
$\boldsymbol{v}$	number of instances v in the labelled set S_l
$\boldsymbol{d} = \{\boldsymbol{d_1}. \boldsymbol{d_2}, \dots \boldsymbol{d_n}\}$	the data frame of the n data instances in S
$\boldsymbol{d_u} = \{\boldsymbol{d_{u1}}, \boldsymbol{d_{u2}}, \dots, \boldsymbol{d_{um}}\}$	the data frame of the m data instances in S_u
$\boldsymbol{d_{l-}} = \{\boldsymbol{d_{l1}^-}, \boldsymbol{d_{l2}^-}, \dots, \boldsymbol{d_{li}^-}\}$	the data frame of the i data instances in S_{l-}
$\boldsymbol{d_{l+}} = \{\boldsymbol{d_{l1}^+}, \boldsymbol{d_{l2}^+}, \dots, \boldsymbol{d_{lj}^+}\}$	the data frame of the j data instances in S_{l+}
$\boldsymbol{x} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_h}\}$	predictor feature vectors in h dimensional space
$\boldsymbol{d_u^i}$	an i[th] instance of unlabelled social datum in S_u
$\boldsymbol{f(z)}$	logistic or sigmoid function
$\boldsymbol{P(d_u^i \in S_{l-} \mid x^o) = p(x^o)}$	probability of d_u^i as a member of S_{l-} given $\boldsymbol{x^o}$
$\boldsymbol{M_{1,2,\dots m}}$, **M**	*m* fitted models before model selection, best model respectively
$\boldsymbol{\theta = (\alpha, \beta_1, \beta_2, \dots, \beta_h)}$	coefficients of a logistic regression model
$\boldsymbol{L(\theta)}$	likelihood function
$\boldsymbol{p_i}$	estimated probability of the class membership of d_u^i
$\boldsymbol{t}$	threshold value or class membership decision boundary
$\boldsymbol{Q}$	the quality level of d_u

Definition 4.3. Individual Quality (Q_i): the quality of an individual instance of a social data such as a Tweet is defined as the probability measure (with value between 0 and 1) of the degree of similarity or closeness of that instance to a contaminated social data instance. The closer (Q_i) is to 1, the low quality the instance is. Individual quality measurement is useful in detecting contaminated instances and in making sure that social updates follow best practices.

Definition 4.4. Relative Quality (Q_r)**:** the relative quality of a social data instance is a measure of users' perception, it takes into account the number of reactions such as likes, comments/replies, and retweet/share. Relative quality measurement captures how users feel about a social data instances. It can be viewed as a measure of satisfaction.

Definition 4.5. Absolute Quality (Q_a): the absolute quality of a social data instance is the sum of the individual quality (Q_i) and relative quality (Q_r) of a social data instance and is given by:

$$Q_a = Q_i + Q_r \tag{4.1}$$

Definition 4.6. Sample Social Data Quality (Q_s): the quality of a social data sample is defined as the degree of excellence of social data with respect to the weight of contaminated instances contained in the sample. This definition is also analogous to water quality which refers to those characteristics that make water appealing, useful, and suitable for a particular use (Spellman, 2013). The higher the concentration of contaminated instances in both social data and water, the lower the quality of the sample and vice versa. The social data quality Q_s for n instances of social data in a sample is an aggregate or summation of the absolute quality class of each instance (Q_a). Given that $Q_{a\_} = \sum_1^u S_{l\_}$ and $Q_{a_+} = \sum_j^u S_{l_+}$ represent contaminated to the non-contaminated social data instances respectively, then Q_s can be expressed as the following proportion or ratio:

$$Q_s = Q_{a\_} : Q_{a_+} \tag{4.2}$$

As such the higher the value of $Q_{a\_}$ the poor the quality of the sample Q_s.

4.5. Effect of Contaminants in Social Data

Most machine learning research assumes that the data feeding algorithms is of high quality, i.e. accurate, complete and timely (Sessions and Valtorta, 2006). However, data quality problems can lead to unrealistic or noticeably strange answers in statistical analysis and estimation (Herzog et al., 2007). The effect of bad data in analytics will be illustrated in the next section. There are simply too many updates that many social media users don't care about yet these updates keep streaming constantly through their timeline. Studies have found that social data is full of noise because its contents are written and posted without much revision. For example, a common complaint about using Twitter data is that 99% of it is useless (Liu, 2013). Due to the quick and short messaging nature of Twitter, people use acronyms, make spelling mistakes, use emoticons and other

characters that express special meanings (Agarwal et al., 2011). Currently, social data is contaminated with noise and abuse which include duplicate contents, pointless babbles, gibberish, information hijacking. According to the studies (Ott et al., 2011) and (Ott et al., 2012), there has been a growing concern about fictitious reviews that have been deliberately written to look and sound authentic and to deceive the reader. Deceptive, low quality and unreliable updates often are a problem when an online communication platform becomes popular. The rise in social media popularity has also been paralleled by the rise of unwanted and disruptive entities in the networks (Chai, 2011). This section looked into the effect of contaminated social data instances in two case studies.

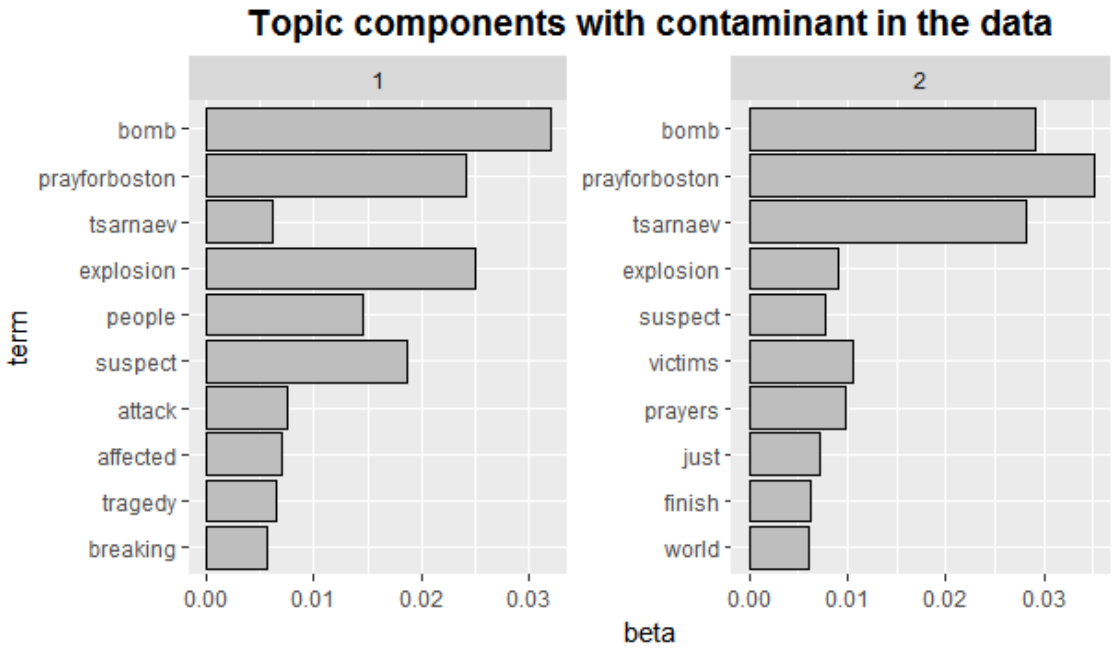
4.5.1. Effect of Social Data Contaminants in Topic Modelling

Contaminated social data can degrade the quality and credibility of online discussions and search systems (which are a crucial part of applications in corporations, government, and many other domains). The presence of contaminants above a certain threshold in social data can skew or generate misleading search and content analysis results (Clark et al., 2016). For example, Fig. 4.4 is an illustration of the effect of non-informative or unwanted instances on topic models of disaster Tweets detailed in (Gupta et al., 2014). Fig. 4.4 (a) and (b) shows the top ten words in two different topic distributions of the unfiltered dataset (including non-informative Tweets) while Fig. 4.4 (c) and (d) shows the variants of the same topic distribution of the filtered dataset (after removing the non-informative instances). One will notice some differences in the beta values in of the terms in the filtered and unfiltered component of the top ten word distributions. Some words such as “breaking and tragedy” in topic component 1 has completely disappeared outside the distribution. Again, duplicate information consumes resources and provides little or no new information to the user (Croft et al., 2010).

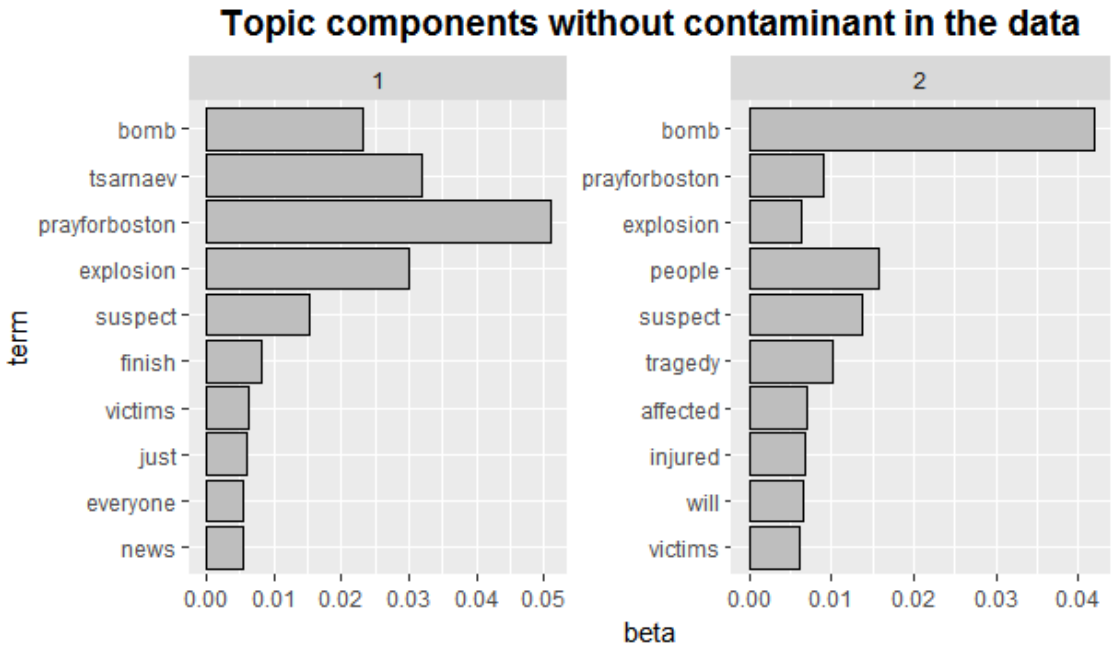
4.5.2. Effect of Social Data Contaminants in Sentiment Analysis

Again, the effect of contaminated social data in sentiment analysis result was demonstrated using the same datasets used in subsection 4.5.2. Fig. 4.5(a) shows the sentiment polarity scores of the unfiltered dataset (including non-informative Tweets) while Fig. 4.5(b) shows the variants of the same sentiment polarity scores of the filtered dataset (after removing the non-informative

instances). One will easily notice the slight change in the sentiment polarity values. The unfiltered set contains 40% negative, 14% neutral, and 46% positive while the filtered sets contain 41% negative, 15% neutral, and 44% positive. This slight variation may be costly for certain applications.

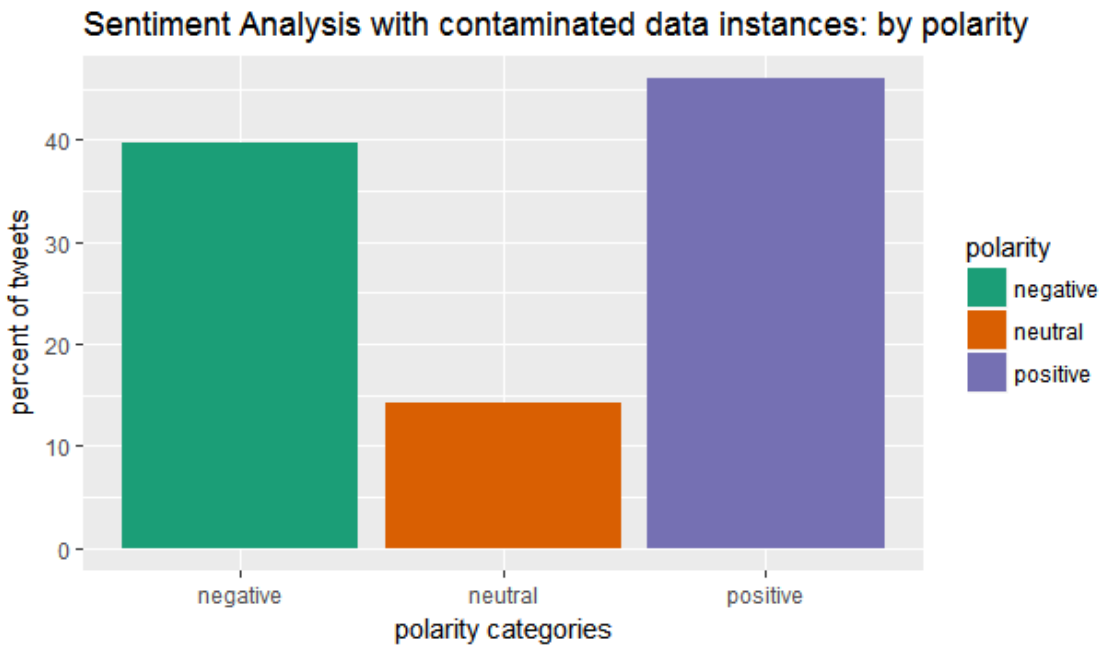


(a) Topic 1: unfiltered dataset (b) Topic 2: unfiltered dataset

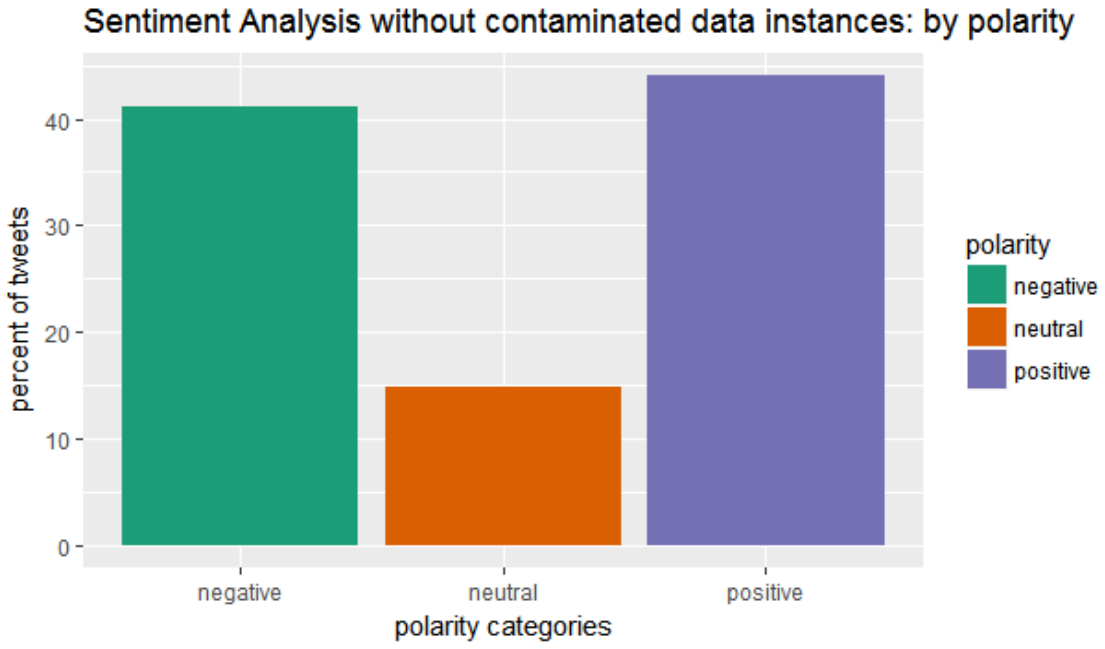


(c) Topic 1: filtered dataset (d) Topic 2: filtered dataset

Fig. 4.4. The effect of non-credible data instances on topic models



(a) Data include contaminated instances



(b) Contaminated instances filtered

Fig. 4.5. The effect of non-credible data instances on sentiment analysis

Unless checked, contaminated social data can de-value real time search services (Benevenuto et al., 2010), as such, social data sources must discourage abuse while encouraging participation (Chai, 2011). Therefore, in order to mitigate errors and bias in data analytics results due to quality issues as well as to extend

the social data content quality analysis method to the next level, this study adapt the methods of monitoring water quality (or contaminants in water) to monitor social data content quality. Details of the innovative approach will be outlined in the next section.

4.6. Estimating the Level of Contaminants in Social Data

In this study, the challenge of estimating the level of contaminants in social data is casted as a supervised learning in which it is assumed that relevant sample data is available or can be extracted and a reasonable portion of the data has been labelled with the relevant quality classes. It, therefore, follows that $d = \{d_1. d_2, \dots d_n\}$ can also be represented as $d = \{d_u, d_{l-}, d_{l+}\}$, such that $n = m + i + j$ (see TABLE 4.2 for full description of notations). Again, all the data instances v in the labelled set S_l can also be represented as $d_l = \{d_{l-}, d_{l+}\}$, such that $v = i + j$.

4.6.1. Problem Definition

Given a social data lake S, let's assume its portion denoted as S_l is labelled with quality class levels while the rest of the portion of S denoted as S_u is unlabelled (see TABLE 4.2 for full description of notations). For binary class learning, the labelled set S_l is further categorised as either contaminated (S_{l-}) or non-contaminated (S_{l+}). All the data instances n in S are denoted as $d = \{d_1. d_2, \dots d_n\}$, while the data instances m in S_u are denoted as $d_u = \{d_{u1}, d_{u2}, \dots, d_{um}\}$. The data instances v of the two subsets in S_l are represented as $d_{l-} = \{d_{l1}^-, d_{l2}^-, \dots, d_{li}^-\}$ for the i data instances in S_{l-}, and are represented as $d_{l+} = \{d_{l1}^+, d_{l2}^+, \dots, d_{lj}^+\}$ for the j data instances in S_{l+}, where $d_{l1}^-, d_{l2}^-, \dots, d_{li}^-$ are the contaminated data instances while $d_{l1}^+, d_{l2}^+, \dots, d_{lj}^+$ are the non-contaminated data instances.

4.6.2. Algorithm for Social Data Quality Estimation

Given a sample of social data lake denoted as $S = S_u \cup S_l$, where $S_u \subset S$ is unlabelled, while $S_l \subset S$ is labelled as either contaminated (S_{l-}) or non-

contaminated (S_{l+}), as such, $S_l = S_{l-} \cup S_{l+}$. The Venn diagram in Fig. 4.6 is a representation of the set membership of all the instances of S. The social data quality estimation challenge is formerly defined as the probability estimation of the target class membership (S_{l-}) of the m data instances $d_u = \{d_{u1}, d_{u2}, \ldots, d_{um}\}$ in S_u over the feature set $x = \{x_1, x_2, \ldots, x_h\}$, reversed engineered from the labelled set S_l, where h is the number of features extracted. First, S_l is reversed engineered to identify the optimal feature combination x^o from the feature set x that can effectively discriminate between $S_{l-}$and S_{l+}. A model M is then developed for estimating the probability of the class membership of the m data instances d_u using x^o. The resulting probability values are then transformed to a quality value (Q) which is used to rank and infer the quality category of different social data samples for different information needs.

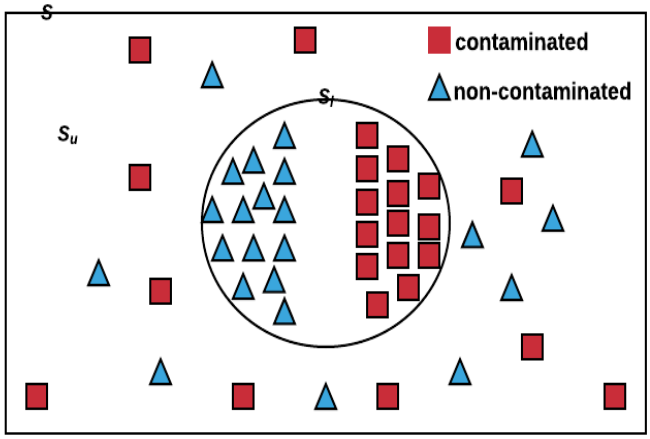


Fig 4.6. Venn diagram representation of the dataset

This quality class estimation can be achieved using any classifier that can return class membership probability values. In this study, logistic regression described in equation (4.3) was utilised.

$$P\left(d_u^i \in S_{l-} | x^o\right) = 1/\left(1 + e^{-(\alpha + \sum_k \beta_k x_a)}\right) \quad (4.3)$$

where α and β are regression coefficients that will be computed from S_l, and $a = 1, 2, \ldots, h$. The choice of logistic regression as a modelling technique is because (i) it suits our goal of predicting class probabilities rather than just

quantitative variables, (ii) of the ease of assessing the predictor's importance, and (iii) the conditional distribution $P(d \in S_{l-}|x^o)$ is a Bernoulli distribution with the binary outcome: S_{l-} or S_{l+}. The estimated probabilities denoted as p_i can then be used to rank the quality level (Q) of the social data. Making predictions with a logistic regression model is as simple as plugging in numbers into the logistic regression equation and calculating the result. There are also several metrics for accessing the goodness of fit of the model, commonly used metrics include the Akaike Information Criterion (AIC), McFadden's pseudo-R squared ($R^2_{McFadden}$), and the area under the curve (AUC) of the receiving operating characteristic (ROC) graph.

The entire process of estimating the quality class of social data is illustrated in $Algorithm\ 1$. The process begin with a reasonable size of labelled social data and then extract as many statistical features from the content as possible. Redundant features and those with high degree of multicollinearity are then removed with the method $SelectFeatures()$ which returns the best predictor variable subset (x^o) by employing the variable inflation factor (VIF_k) on all the extracted features. The selected features are then used to model logistic regression estimators $M_{1,2,\dots,k}$ for different variable combinations $1, 2, \dots, k$. The method $SelectModel()$ returns the best model M by taking as attributes, the returned models $M_{1,2,\dots,k}$, the model selection metrics AIC, $R^2_{McFadden}$ and AUC. The best model is then used in the method $predict()$ to estimate the probability of class membership of all the instances of S_u. The probability estimates are then sorted and returned for final categorisation based on a suitable decision boundary b. In analogy with water quality estimation, $Algorithm\ 1$ is called social data quality index ($SDQI$).

$Algorithm\ 1$: Estimating the quality class of social data (SDQI)

Input: v instances of $S_l = S_{l-} \cup S_{l+}$, and m instances of S_u

Output: m instances of S_u with associated quality labels Q

Feature Engineering

$\boldsymbol{for}$ *each instance* i *in* $v, \in S_l$ $\boldsymbol{do}$

$ExtractFeatures(S_l,\ x = x_1,\ x_2, \dots, x_n)$

$SelectFeatures(x, VIF_k)$

$Return\ x^o$

$\boldsymbol{end\ for}$

Fitting Data

$M_{1,2,\dots,k} \leftarrow P(d \in S_{l-}|x^o) = 1/\left(1 + e^{-(\alpha+\sum_k \beta_k x_a)}\right),\ k = 1, 2, \dots, n$

Model Selection

$\boldsymbol{for}$ *each* $M_{1,2,\dots,k}$ $\boldsymbol{do}$

$SelectModel(M_{1,2,\dots,k}, AIC,\ R^2_{McFadden},\ AUC)$

$Return\ M$

$\boldsymbol{end\ for}$

Quality Class Membership Estimation

$\boldsymbol{for}$ *each instance* i *in* $m, \in S_u$ $\boldsymbol{do}$

$p_i \leftarrow predict(S_u, M)\ ,\ p_i$

$Sort(p_i)$

Return p_i

Decision Boundary: Binary

$\boldsymbol{for}$ *each* p_i

$$Q = \begin{cases} i = contaminated, & p \geq b \\ i = noncontaminated, & otherwise \end{cases}$$

$Return\ Q$

$\boldsymbol{end\ for}$

The problem of data quality, in general, is highly application-specific and domain dependent (Dasu, 2013), the proposed quality estimation method can be customised appropriately. Next, a dynamical system model for understanding the evolution of contaminated instances will be developed in analogy with the model of chemical pollution in a lake detailed in (Cornette et al., 2013).

4.7. Dynamical System Model for Quality Assessment

Consider a body of water (lake) with many tributaries flowing through and out of it, let's assume the lake at a time, t is of high quality to support swimming and animal life. Because the water body has many incoming (including contaminant sources from industrial or agricultural wastes) and outgoing tributaries, the water quality in the lake is expected to keep changing dynamically relative to the released chemical waste at any given time.

Change in chemical waste in water per day: The daily change in the quality of the water body was dynamically modelled in (Cornette et al., 2013) as *daily change in chemical waste per day = amount of waste added per day – amount of waste removed per day* using the relation in equation (4.4), i.e.

$$W_{t+1} - W_t = R - \frac{F}{V} W_t \tag{4.4}$$

where V is the volume of the lake, F the daily flow of water through the lake, R the daily release of chemicals into the lake, W_t the chemical waste at time t, W_{t+1} the chemical waste at time $t + 1$. The quantity $\frac{F}{V} W_t$ in equation (4.4) refers to the amount of chemicals removed from the lake down an exit river per day. The chemical concentration C_t was represented as a multiple of the total chemical waste W_t as in equation (4.5).

$$C_t = \frac{W_t}{V} \tag{4.5}$$

Change in social data quality per day: This study posits that the above model of chemical waste per day can be adapted to social data. This is because many of the assumptions for water quality dynamics holds for social data. First, the quality (value/informativeness) of social posts on a given topic streaming on social media within a time window of interest is dynamic, despite their effort in

filtering contaminated instances, the larger and more popular these social data platforms become the more attractive they become to contaminants and contaminated instances. This dynamical model of change in a chemical waste of water per day is hereby extended to an analogous model of change in social data contaminant per day on a social media platform. The proposed model is given as *daily change in social data quality per day = total number of contaminated social data added per day – total number of contaminated social data removed per day* using the relation in equation (4.6), directly adapted from equation (4.4), i.e.

$$X_{t+1} - X_t = S_{l\_} - \ S_{l\_r} \tag{4.6}$$

where $S_{l\_}$ is the contaminated portion of labelled social data S_l, S_{l_r} the portion of $S_{l\_}$ being removed, X_t the total contaminated social data instances at time t, X_{t+1} total contaminated social data instances at time $t + 1$. The model can be utilised to compute the number and distribution of contaminated instances in social data sample over time. It can also be applied in understanding what characteristic of social data attracts contaminated instances.

4.8. Experimental Work

In this section, we will evaluate our proposed methods with Twitter datasets. We will first demonstrate $Algorithm\ 1$ which describe how contaminated instances of social data can be estimated using supervised learning approach. We will then apply the algorithm on existing and extracted datasets in order to compare the concentration of contaminated instances in the different samples.

4.8.1. Datasets

A total of 172,772 Tweets were collected using the Twitter streaming API during the 2016 Brit Awards held on 24 February 2016 using #BRITs2016 as a query term. The Tweets were collected for few hours as the event was happening. This dataset is denoted as D^7 and each Tweet is taken as a single text document.

Pre-processing: exact duplicates in D^7 was removed and a corpus was created comprising of only Tweets (36048 instances) written in the English language. For the purpose of this evaluation, the definition of contaminants is limited to

Tweet features such as deceptive hashtags, links, username mentions, and word counts. Tweets such as those in Fig.4.1, Fig.4.2, and Fig.4.3 are obviously low in information and can be easily discriminated based on their statistical features. Despite being easy to visually identify, these instances still overcome platform and application specific noise filters and are able to penetrate into the user feeds.

Feature Creation and Extraction: in this work, Tweet features reported in (Choi et al., 2012), (Vosecky et al., 2012), and (Webberley et al., 2016) that have proved successful in the literature of Twitter data mining were considered. The initial feature sets x_T, include: Average Characters per Tweet (C_T), Average Word per Tweet (W_T), Word Difference per Tweet (WD_T), Average Unique Word per Tweet (UW_T), Average Hashtag per Tweet (H_T), Average Mention per Tweet (M_T), and Average URL per Tweet (U_T), User Listed Count (ULC_T), User Friends Count (UFC_T), User Statuses Count (USC_T), User Followers Count (UFC_T), Favourites Count (FC_T), User Language (UL_T), User Verification Status (UVS_T), User Geolocation Status (UGS_T), and User Client Sources (UCS_T).

TABLE 4.3. Confusion matrix for the quality class prediction

Test instances	Predicted: No	Predicted: Yes	Total
Actual: No	**2653**	10	2663
Actual: Yes	8	8142	8150
Total	2661	8152	**10813**

A small proportion of the entire #BRITs2016 datasets was first labelled into two categories (contaminated or non-contaminated). The labelling was done by predicting the class labels of more than five million instances of Twitter data obtained from (Kyumin et al., 2011) categorised by source (content pollutants and non-pollutants) the labelled instances was then utilised as initial training sets and then automatically (with some manual checks) extended the labels to the entire dataset using supervised learning approach (see TABLE 4.3 for details of the labelled Tweets). Following the word representation of the corpus using the VSM

with TF-IDF, a visual description of the importance of each of the predictor variables in the feature set x_T in discriminating the target class is shown in Fig. 4.7. From these plots, one can make a rough estimate of the contributions of each predictor variable in discriminating the target variable. A regression model is said to provide a better fit to the given data if it shows an improvement over a model with fewer predictor variables (Mathew, 2015).

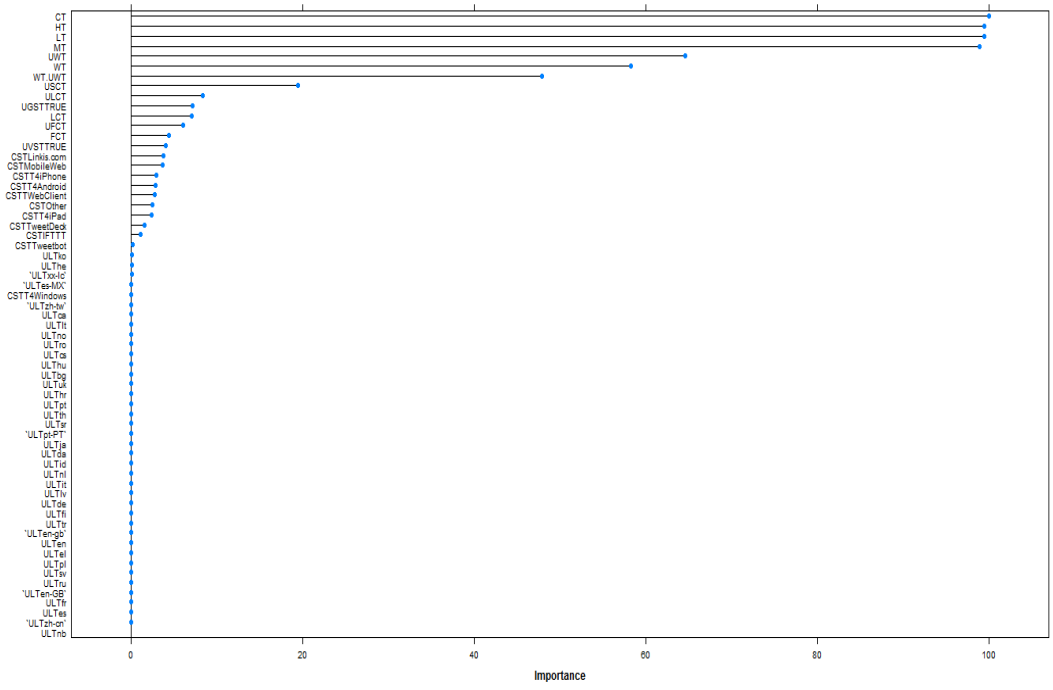


Fig. 4.7. Variable importance

The following feature selection metrics VIF_k, AIC, and $R^2_{McFadden}$ were applied in order to select best feature sets ${x^O}_T$. The selected features were then utilised to train the logistic regression model described in equation (4.3). Finally, the model is applied to estimate the probability that an unlabelled social data instance is a member of the contaminated class. The processed data (36048 instances) was divided into training (25235 = 70%) and test (10813 = 30%) sets. TABLE 4.3 is the confusion matrix for the quality class prediction over the test set. The accuracy (AUC of the estimator is obtained from the area of the ROC curve in Fig. 4.8 as 0.9998218. Apart from the analysis of the statistical properties of the Tweet corpus, we could also explore other parameters such as the devices (client sources) used for posting the

Tweets, the time of posting, and the characteristic of accounts posting the Tweets. The analysis of the accounts is very important in identifying key actors and groups that are targeting and contaminating social data.

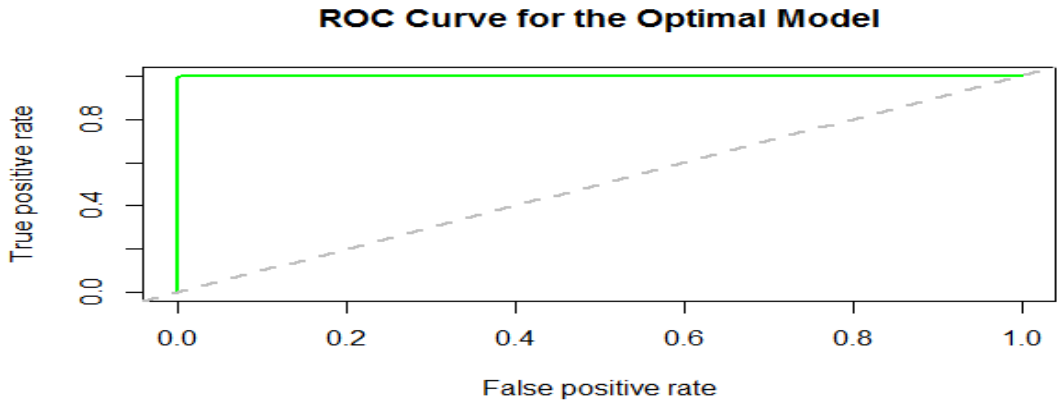
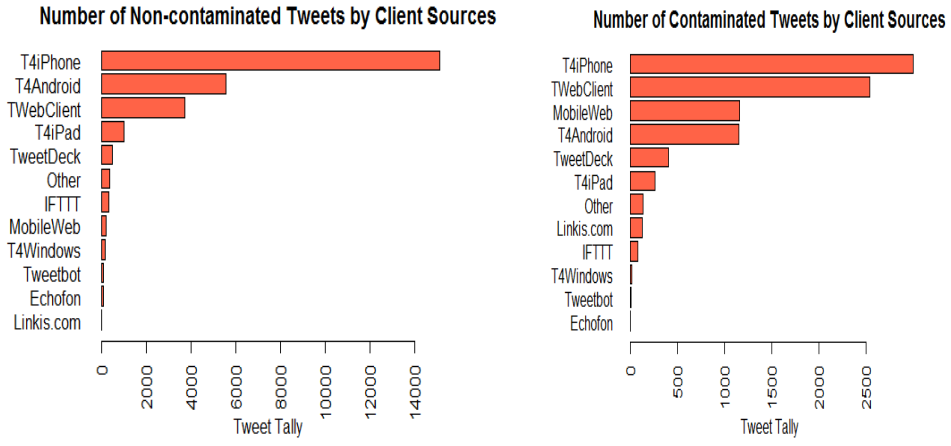


Fig. 4.8. ROC Curve for the optimal model

For instance, the contamination concentration of the Tweet client sources (see Fig. 4.9) demonstrate some interesting results with higher contamination concentration from Twitter Web Client and Linkis.com (a free link customisation service for social promotion). It was also discovered that many of the accounts posting these contaminated instances were later suspended by Twitter.



(a) Non-contaminated Tweets (b) Contaminated Tweets

Fig. 4.9. Barplot of client sources

4.8.2. Dynamical Modelling

Recall that the proposed dynamical model in equation (4.6) can be used to compute the number and distribution of contaminated instances in social data sample over time. The model utilises the output of the quality estimator in $Algorithm\ 1$ to measure abnormal peaks in the distribution of contaminated instances in a social data sample over time. The processed dataset D^7 was again represented into hourly time component for quality trend analysis (monitoring of contaminated to non-contaminated proportions of over a given time period). For a sample spanning several weeks, the measurement can be dividing into days or weeks.

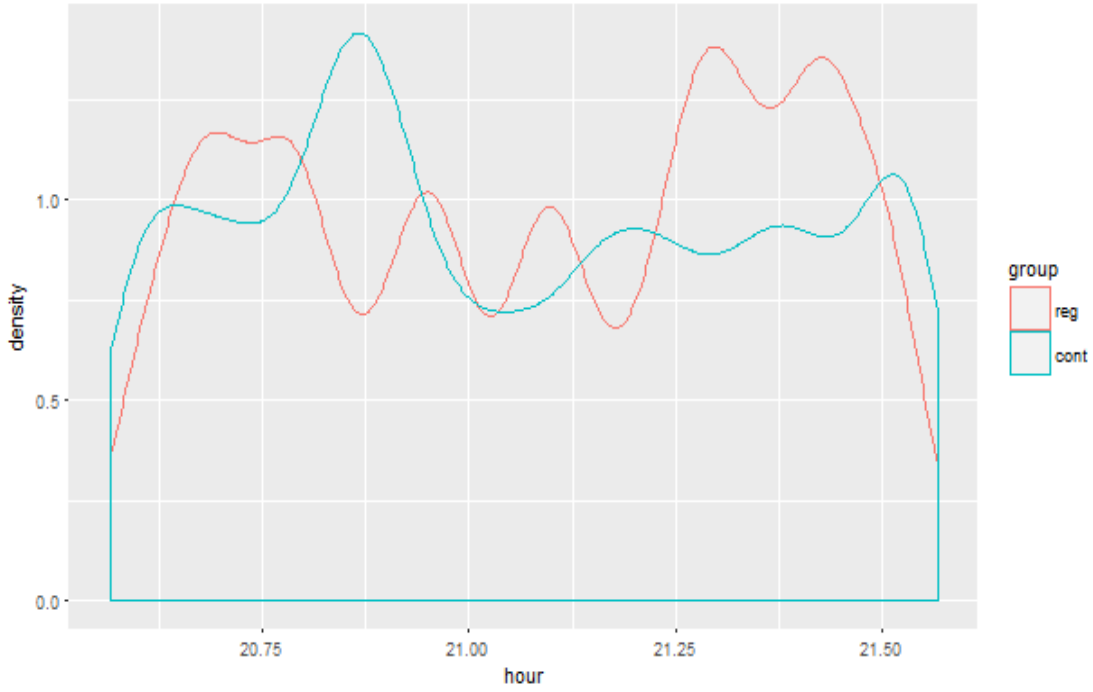


Fig. 4.10. Tweets quality trend analysis using kernel density plot

The kernel density plot in Fig. 4.10 shows the evolution of both contaminated (cont) and non-contaminated (reg) Tweets within the Brits2016 event window (the live shows commenced at 20:00 and ended at 23:00 UK time). One can easily deduce from the graph that there were only a few contaminated instances at the commencement of the show, the contaminated instances then increased to its maximum peak between 20:00 and 21:00 and then normalised to a constant value. It is logical to infer that the sharp fall in the number of the contaminated instances is due to the default Twitter noise filters

which may probably have some lag time. This underscores the need for an advanced and real-time noise filter.

4.9. Conclusions

One major challenge with social data is the presence of noise (objects such as duplicates, non-informative terms, repeated tags or links created and embedded in a social update in order to amplify the visibility of misleading information) in the mainstream subjects. Contamination or noise in real world data are errors that can significantly impact the analysis and conclusions drawn from the data. In structured data, these errors are often called data glitches and include typos, missing values, outliers, duplicate entries, and distorted values. Signatures are commonly used to detect and tag glitches in structured data. The quality of the data is often quantified by the glitch prevalence using an objective measurement such as glitch scores and indices. In terms of unstructured data such as web documents, quality is viewed as a continuous spectrum consisting of high quality and low quality documents at each end with most of the pages somewhere in between the two extremes. In this study, social data contaminant was defined as objects such as duplicate instances, non-informative terms, repeated links, tags, and other entities created and embedded in social data in order to amplify the visibility of misleading information. Social data instances containing contaminants are referred to as contaminated and can degrade search and other quality expectations of social data analytic results.

The bulk of research effort on mitigating contaminants and contaminated social data focused mainly on the detection of suspicious accounts. A major challenge, however, regarding account-based approaches is that automated accounts are still evolving and developing strategies to emulate organic account activities. This underscores the importance of developing methods that can measure and filter low quality instances of social data based on the content features. The problem was considered to be analogous to evaluating the quality of water due to many characteristics similar to both social data and water. This study proposed an innovative method for measuring the quality class of social data instances and for the dynamic assessment of social data quality. This study focused more on social data lake and further study may look into streaming social data and online algorithm.

The results obtained are useful in developing filters on social data platforms and decision support systems employing machine learning features on what dataset can be used for given decision-making tasks. For instance, in order to standardise sentiment analysis result, specific data quality criteria can be drafted so that published analytical results can be effectively compared. The result can also be a guideline for individuals and organisations offering open datasets for market analysis.

Chapter Five: Towards Social Data Quality Standard

5.1. Introduction

Given a sample of social data, this study tried to understand the analytical effect and prediction of contaminated instances in the previous chapter. It also addressed critical data quality issues related to fitness for use, user satisfaction, and conformance to best practice policies by adapting the well-established methods of water quality monitoring. With regards to fitness for use, instances of social data that are completely irrelevant were filtered during the data collection process, user satisfaction were measured directly from user reactions such as likes, comments, and shares or retweets in the case of data from Twitter, finally, regarding conformance which is usually platform specific, social data from Twitter was utilised Tweet features (similar to water quality indicators) that adulterate a Tweet to become a contaminated instance were defined. These features were then integrated into a model for estimating the contaminated likelihood of unlabelled Tweets. Using raw and filtered version of Tweets in an experiment, some bias in topic modelling and sentiment analysis results introduced by contaminated instances were observed. An algorithm for predicting and ranking social data instances based on their level of contaminants and a dynamical system model for measuring trends in social data quality were developed and evaluated to be very useful. Again, in social retrieval applications where users often search or analyse updates on events of interest as it happens, a filtering technique that can complement the automated account classification effort is needed.

This chapter, is a further step in the same direction, here, the focus is on answering the second part of RQ3, which is focused on developing benchmarks that can be used to compare the suitability of social data samples for different information needs. The proposed algorithm (see section 4.6) for estimating the quality level of social data called $Algorithm1$ will be utilised for comparing

several samples of Tweet datasets. Apart from its commercial application, an answer or answers to this part of RQ3 can also be a further step in developing a formal background for social data quality monitoring. Other interesting research application of this and previous chapter contributions include real time evaluation of the effect of contaminated social data in search and analytic results.

The rest of this chapter is organised as follows, section 5.2 introduced fundamental concepts and definitions of social data quality benchmarking, section 5.3 is an analogy between the well-established water quality standard and the proposed social data quality standard (a case study of Twitter data), section 5.4 presents social data comparison results obtained and lessons learnt through experimental work using datasets from Twitter, and finally section 5.5 demonstrate the utility of the contributions in this and previous chapters in social search, and finally, section 5.6 is the chapter conclusion which provides an overview of all that objectives that were met and a pointer to future work.

5.2. Definitions

Currently, organisations are being heavily impacted by bad data as a result of high degree of inaccurate information, which among other reasons, is due to the outdated data management strategies that are in place today (Quality, 2015). As reported in the previous chapter, measuring the quality level of an instance or sample of social data as well as the comparative evaluation of various social data samples is a challenging and difficult task because of the subjective nature of natural language text. What looks good to a particular user in terms of quality may be perceived differently or bad by another user. However, there should be an established benchmark or standard (as in the case of water quality) that describe the desired quality of social data or the level of contaminated instances that can be tolerated in a sample for a given data application. So, what are standards?

Definition 5.1. Standard: defined according to online Oxford dictionary as (i) a level of quality or attainment, (ii) a required or agreed level of quality or attainment, and (iii) something used as a measure, norm, or model in comparative evaluations.

Definition 5.2. Social Data Quality Standard: defined as a benchmark against which the quality of social data samples can be compared. This is also analogous

to water quality standard which is a benchmark against which monitoring data are compared to assess the health of waters (Spellman, 2013).

5.3. Social Data vs Water Quality Standard Analogy

Water sources are limited, hence the need to keep its quality at the highest level possible (Postolache et al., 2012). The major task of water quality assessment involves the sampling and analysis of water constituents and conditions. This is achieved by conducting regular testing of water quality indicators and comparing the results to some form of established standards. Organisations have become obsessed with data and the insight that it can provide, as a result the quality of data is growing in importance. Without good quality data, it becomes difficult to gain the desired level of insight. Data quality should take a central role, should be managed and monitored, and should be part of core business practices of an organisation (Quality, 2015). The next two subsections will look into the parallel between water and social quality.

5.3.1. Water Quality Standards

There are three main dimensions of water quality characterisation. First, the physical characteristics of water such as turbidity, colour, taste, odour, and temperature which are apparent to the senses of smell, taste, sight, and touch. Second, the chemical characteristics of water which refers to measures of dissolved impurities such as total dissolved solids (TDS), alkalinity, hardness, fluoride, metals, organics, and nutrients (Spellman, 2013). Finally, the biological characteristics of water which refers to the presence or otherwise of waterborne pathogens that cause waterborne diseases such as bacteria, virus, and protozoa. There exist different water quality standards for different locations and applications, for example in (Postolache et al., 2012), the acceptable ranges of water contaminant indicators established for human consumption are Turbidity (1 to 5 NTU), salinity (500 TDS), and pH (6.5 to 9.0).

5.3.2. Social Data Quality Standards

In the same manner, this study proposes the analogous indicators of physical characteristics for social data to include (near)-duplicates and gibberish (instances with repeated characters, words, tags, or other entities) that can be spotted in a sample; while that of chemical characteristics for social data include

hijacked or irrelevant instances and instances with deceptive entities such as tags, links, and user mentions; and finally, the analogous biological characteristics for social data include instances containing phishing links or contents. The next section will look into a case study of Twitter data quality standard.

5.4. Quality Comparison of Data Samples

As stated in TABLE 4.1, one major application of the proposed social data quality estimation result is that the labelled dataset now serves as a gold standard for Tweet quality modelling and comparative analysis of related data samples. The goal in this section is to measure and compare the proportion of contaminated data instances in four Twitter datasets. This is similar to comparing the quality of different water samples for similar or different applications.

5.4.1. Dataset

Four different Twitter datasets were harnessed for quality comparison. The first sample (Star Wars), denoted as D^8 was collected on May 4, 2016, using the two hashtags #MayThe4thBeWithYou and #StarWarsDay used by Star War fans for celebrating the Star Wars Day. The other three datasets (Disaster, denoted as D^9 represent Tweets on Disasters on social media, GOP Debate, denoted as D^{10} represent Tweets on the First GOP debate, and Deflategate, denoted D^{11} represent Tweets on New England Patriots Deflategate, D^9, D^{10} , and D^{11} were obtained from Crowdflower[27], a collection of cleaned, annotated and enriched open social data.

5.4.2. Social Data Quality Comparison

Algorithm 1 was applied on each sample (see TABLE 5.1 for a detailed description and results obtained). A higher proportion of contaminated instances was expected in the Star Wars dataset in relation to the other 3 datasets (as it was not cleaned). From results obtained, as expected, the Star Wars contains the highest proportion of contaminated instances (almost half of the dataset). It is however interesting to observe that the Disaster dataset

[27] https://www.crowdflower.com/data-for-everyone/

was the least affected. Although specific to Twitter data, these results are extensible and can be useful in developing filters on social data platforms as well as in making choices of what dataset can be used for a given decision-making task.

TABLE 5.1. Tweet samples with class percentage estimates

Class	Star Wars	Disaster	GOP Debate	Deflategate
contaminated	5247 (49.97%)	798 (7.34%)	3589 (25.87%)	4380 (37.07%)
non-contaminated	5253 (50.03%)	10078 (92.66%)	10282 (74.13%)	7434 (62.93%)
Total	10500	10876	13871	11814

5.5. Quality Ranking in Social Search

Drinking water suppliers now provide search tools for consumers to always check and monitor the water quality in their areas. For example, the Thames Water Utilities Limited[28] provided a Web-based tool for its customers to always monitor their drinking water characteristics to check hardness, lead pipes, and smell. In the same manner, information retrieval systems must always ensure that relevant documents appear near the top of a search result. A typical social search or information retrieval process (see Fig. 5.1) involves a user (u) who types or pronounces a query (q) to a search system and the system retrieves documents (d) from an indexed collection by first ranking the documents according to some metrics such as the documents relevance score to the query term and then returning the top ranked documents to the user. Social information consumers usually visit social media platforms for news on real time events, these users often are faced with the frustration of retrieving spam and non-informative information.

[28] https://my.thameswater.co.uk/dynamic/cps/rde/xchg/corp/hs.xsl/899.htm

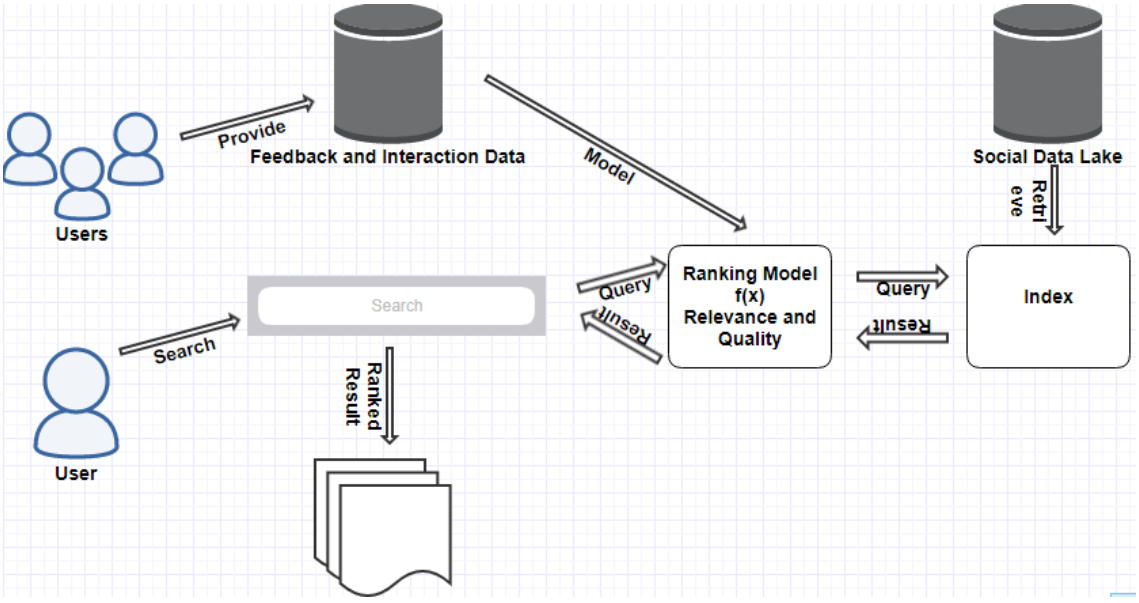


Fig. 5.1. Social search architecture with quality ranking method

The focus of this section is to contribute to the improvement of user experience in social information search by offering $Algorithm\ 1$ as an advanced support system or optional feature for further ranking of social data based on the level of contaminants of each instance in a social stream.

TABLE 5.2. Sample Tweets and query from #Brits2016 dataset

Query (q)	Adele #brits2016
Doc1	#brits2016 pls no #adele pls no #adele pls no #adele pls no #adele pls no #adele pls no #adele pls no #adele pls no #adele **8:47 PM**
Doc2	And @Adele takes home her 2nd BRIT Award for the night! Best Single for “Hello” #BRITs #BRITs2016 **8:48 PM**
Doc3	just caught a quick glimpse of the #BRITs2016 and WOW, my singing babe @Adele won award after award...well done! LOVE HER!!! roll on March! **9:11 PM**
Doc4	@ITV @brits @TheO2 @ITV @Adele @antanddec @justinbieber @bauermedia @thisisglobal #RihannaBRITs #ScottieMcClue #DinkyDoo #BRITs2016 #Adele **9:33 PM**

Consider the sample of four Tweets labelled as Doc1, Doc2, Doc3, Doc4 in TABLE 5.2. If the Tweets were posted by a single user and were to appear in the timeline of the user or followers of the user, the default option may be inverse

chronological order, i.e. Doc1:Doc2:Doc3:Doc4. Assuming a user interested in Tweets mentioning Adele during the Brit Awards show typed the query “Adele #brits2016” in the Twitter search box. Let’s say Twitter uses the popular vector space model for its relevance ranking, the retrieved result (see TABLE 5.3) will be in the order Doc4:Doc3:Doc2:Doc1.

TABLE 5.3. Relevance versus quality based ranking models

documents	relevance estimate	contaminated estimate
Doc1	0.04995497	0.9963914 (contaminated)
Doc2	0.2687147	0.1217354 (non-contaminated)
Doc3	0.2928053	0.02172186 (non-contaminated)
Doc4	0.7033982	0.9999925 (contaminated)

Imagine a trending topic attracting millions of Tweets streaming at real time, search result based on the relevance retrieval may degrade the quality of user expectation. Doc1 and Doc4 are obviously non-informative as far as Brits Award event is concerned. Incorporating our method as an optional filtering strategy, the retrieved result will be in the following order Doc2:Doc3:Doc1:Doc4. This analysis can be easily extrapolated into other social data by just defining the relevant features that can be used as input to the algorithm. The analysis is also relevant in the Pharmaceutical domain especially in aiding users searching for online medications.

5.6. Conclusions

This chapter reports the contribution of this study in developing standards that can be used to compare the suitability of social data samples for different information needs. It specifically focused on Twitter data and addressed critical quality issues related to fitness for use, user satisfaction, and conformance to best practice policies. The chapter began by observing those salient features of social data that allows into the public stream, contaminated instances that contravene Twitter best practice policies. These features were compared with the three main dimensions of water quality. The method that was developed in Chapter Four for estimating the probability that a social data instance is

contaminated was extended here to compare several samples of Twitter datasets with known quality level.

The results obtained met all expectations and are useful in developing filters on social data platforms and in making choices of what dataset can be used for a given decision-making task. Experimental results obtained using five different Tweet datasets revealed the significance of the study in standardising social data quality benchmarks. It also demonstrated the use of the proposed method in enhancing the level of user satisfaction in social search. This study recommends a dynamic analysis of the effect of contaminated instances on different data representations and knowledge discovery techniques, extending this work to data streams and other social data platforms, and developing industry wide data quality standards and decision support systems.

Chapter Six: Conclusions

People are now turning to the Internet to buy things because of convenience and cheaper options. The growing popularity of the Internet offers new opportunities for cybercriminals to conduct their nefarious activities. With the advent of e-commerce, digital currency, and online banking technologies, many traditional or physical-world crimes have taken a new turn into developing an online presence. Many digital crimes require a high degree of organisation and involve a network of humans (such as manufacturers, users, and distributers) and systems (such as payment gateways, e-commerce, and delivery or shipping systems). The wide availability of products and services being offered through e-commerce platforms necessitates the need to accurately assess and evaluate the quality and genuineness of the offers and the offerors. Again, the highly dynamic nature of online pharmacies with entities continuously appearing and disappearing and services are additional challenges that call for the development of robust and efficient trust and reputation mechanisms. Furthermore, a means of uncovering hidden networks and key members in a network is required in order to be able to disrupt crime associations.

Traditional crime data mining based on crime incident reports can only attempt to discover what is going on as far as crime is concerned, but the true nature remain buried in unstructured content that represents the hidden story behind the data. User interactions on social media and other online platforms leave valuable traces of feedback and opinions that can help in shaping the products and services being discussed. This feedback is also an important source of information that can be utilised to measure the quality of both the content and sources of products and services. It can also be used to detect, infer, or predict crimes. However, these data must be of the required quality standard in order to avoid misleading outcomes. There are many reputation systems that enable users of products and services to provide feedback, in some cases, positive opinions often mean profits and fame for businesses and individuals. This, therefore, became a very strong incentive for people to manipulate the system by posting fake opinions or false reviews in order to promote or discredit certain

products. The effectiveness of crime intelligence is highly dependent on the quality of the data, hence the need to measure and ensure that social data for crime prediction are of the suitable quality level. This underscores the importance of developing methods that can measure and filter low quality instances of social data based on the content features. In an attempt to contribute towards mitigating issues of digital crime and poor data quality, the following three novel research questions RQ1, RQ2, and RQ3 were respectively proposed thus, can user feedback be harnessed and processed for crime intelligence? Can criminal associations, structures, and roles be inferred among entities involved in a crime but not directly connected? And can methods and standards be developed for measuring, predicting, and comparing the quality level of social data instances and samples? This study applied state of the art knowledge discovery techniques for crime prediction and developed novel methods and algorithms for data quality analysis and prediction. Answers to these questions are vital to a range of applications including the understanding, intervention, and policy-making on crimes, as well as in the development of filters on social data platforms and in establishing social data quality standards.

RQ1 was addressed in Chapter One by adapting some fundamental mathematical and data mining techniques for the purpose of representing user feedback as well as mining topics, computing sentiment scores, and inferring word relationships. More precisely, a novel product safety framework (reputation system) was proposed for crime intelligence using social data; it utilised Bayesian estimation, topic mining, and sentiment analysis techniques in order to rank the quality of pharmaceutical vendors, products, and their aspects. The framework is an innovative reputation computation pipeline and is in line with an effort to uncover patterns of criminal activities as well as the desire to predict when and where crimes are likely to occur in the future. Word association was also employed to measure the association between crime related words as well as the words in the reviews of top rated and blacklisted vendors. In the word correlation experiments, high correlation values were obtained for the blacklisted vendors; this confirmed the hypothesis on the use of word association techniques for crime intelligence. The topic modelling result also revealed some interesting results that can be useful in crime intelligence. For instance, the high level of satisfaction from

the terms in the reviews of the top rated vendors and a possible dissatisfaction from the terms in the reviews of the blacklisted vendors can be utilised to infer failed deliveries, poor quality products, etc. The brand and product polarity (positive, neutral, and negative) sentiment comparison results signify the usefulness of text mining and sentiment analysis on social media data while the use of machine learning classifiers for predicting the sentiment orientation provide a useful tool for users, product manufacturers, regulatory and enforcement agencies to monitor brand or product sentiment trends in order to act in the event of sudden or significant rise in negative sentiments. Again, the emotional (sadness, anger, fear, surprise, disgust, and joy) sentiment classification results also justified the rating status of both top rated and blacklisted vendors.

RQ2, which focused on typical situations in criminal datasets where there is no information on the direct connection among criminals, was addressed in Chapter Three through an innovative method for inferring entity relationships by employing the concept of co-occurrence. A novel approach was proposed for representing historic crime data as a k-partite graph of common attributes and then projecting the graph to its unipartite components. Two case studies were conducted; the first analysed traditional counterfeiting crime (manufacturer-manufacturer) networks and this study hypothesised that the network analysis of pharmaceutical crime data can be useful in modelling indirect relationships among important entities involved (for example the criminals which include manufacturers, advertisers, and distributors, the products they sell, the banks that process their credit and debit card transactions or the delivery services used for shipping the products). The second case study analysed Darknet (vendor-vendor) networks and it investigates how networks can be inferred from illegal transactions in the darknet. Results obtained in both case studies may be useful for operational strategies, for instance, vertices with a central position in their clusters may have an important function of control within a group while vertices lying at the boundaries between clusters may play an important role between two groups. Once these criminal groups are identified and their habits are known, law enforcement authorities may begin to assess crime trends in order to forecast and hamper the development of perceived future criminal activities. These

analyses may also aid security agencies to further investigate whether the clusters are actually involved in an organised network, the evolution and duration of the network, what sustains it, and how it can be disrupted. Beyond these analyses, SNA can also be used to reveal relationships between user accounts sending pharmaceutical spam and the spam URL's.

RQ3 was actually a consequence of RQ1 and RQ2, both Chapter Two and Chapter Three recommended a further work in data quality, especially social data. It was discovered that the quality of social data is still a major concern for users as well as in data analytics, this is because distinguishing between good and bad instances of social data is difficult due to issues such as the veracity of contents and accounts, the nature and openness of its providers, the concept of anonymity, the large volume of content being created per time, the velocity with which social data are being created, and best practice abuse due to difficulties in content moderation. RQ3 was addressed in Chapter Four and Chapter Five. Chapter Four looked into methods for measuring and predicting the quality level social data instances. It considered the problem to be analogous to evaluating the quality of water due to many characteristics similar to both social data and water. It then proposed an innovative method for estimating the quality class of social data instances and for the dynamic assessment of social data quality. The justification of the water versus social data analogy is due to the following: firstly, the science of water quality evaluation is very well established; secondly water bodies can suitably be compared with archived social data while a flowing river can be likened to social data streams, and finally the indicators and parameters for measuring water quality show good adaptability to measure social data quality. Experimental results obtained using Tweet data uncovers some bias in topic mining and sentiment analysis due to contaminated instances and revealed the significance of the study in developing filters on social data platforms and in the development of decision support systems employing machine learning features on what dataset can be used for given decision-making tasks. Chapter Five focused on developing benchmarks for social data quality comparison. For instance, in order to standardise sentiment analysis result, specific data quality criteria can be drafted so that published analytical results can be effectively compared. The result can also be a guideline for individuals and organisations

offering open datasets for market analysis. The chapter also demonstrated the use of the proposed method in enhancing the level of user satisfaction in social search.

In terms of crime intelligence using social data, this study recommends that further research in this line should consider (i) working closely with social data security agencies in order to be able to correlate the result obtained with real crime incidents, (ii) using other sources of datasets such as search logs and the many evolving groups and pages on social media that are dedicated to providing feedback on their experiences of using certain goods or services (iii) comparing the results with mainstream media report archives such as the newspaper section of LexisNexis[29] library. Sentiment-based identification of suspicious actors and contents is becoming very interesting but requires an interdisciplinary approach. Major social media platforms such as Facebook and Twitter are beginning to develop research departments at the intersection of computer science, statistic, and social science for aggregating crowd wisdom. A recent study in (Scrivens et al., 2017) reported that sentiment analysis can enhance the process of identifying users of interest to counter-extremism agencies.

Regarding social data quality, as time passes, social communication will keep changing in content and structure, data quality, therefore, becomes dynamic. The challenges of dynamic data quality promise to provide research opportunities for a long time to come (Labouseur and Matheus, 2017). This study recommends the following (i) a dynamic analysis of the effect of contaminated instances on different data representations and knowledge discovery techniques, (ii) extending this work to data streams and other social data platforms, and (iii) developing industry wide data quality standards and decision support systems.

[29] https://www.lexisnexis.com/ap/academic/form_news_wires.asp

References

Abbasi, A., Chen, H. & Salem, A. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Trans. Inf. Syst.,* 26**,** 1-34.

Abbasi, A., Zhang, Z., Zimbra, D., Chen, H. & Nunamaker, J. F. 2010. Detecting fake websites: the contribution of statistical learning theory. *MIS Q.,* 34**,** 435-461.

Afroz, S., Islam, A. C., Stolerman, A., Greenstadt, R. & McCoy, D. Doppelganger Finder: Taking Stylometry to the Underground. 2014 IEEE Symposium on Security and Privacy, 18-21 May 2014 2014. 212-226.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O. & Passonneau, R. Sentiment analysis of twitter data. Proceedings of the workshop on languages in social media, 2011. Association for Computational Linguistics, 30-38.

Agarwal, N. & Yiliyasi, Y. Information quality challenges in social media. International Conference on Information Quality (ICIQ), 2010. 234-248.

Agichtein, E., Castillo, C., Donato, D., Gionis, A. & Mishne, G. 2008. Finding high-quality content in social media. *Proceedings of the 2008 International Conference on Web Search and Data Mining.* Palo Alto, California, USA: ACM.

Ball, P. 2013. *Unmasking organised crime networks with data* [Online]. @BBC_Future. Available: http://www.bbc.com/future/story/20130709-unmask-crime-networks-with-data [Accessed 2017].

Banchs, R. E. 2012. *Text mining with MATLAB*®, Springer Science & Business Media.

Bartlett, J. 2015. *What 'dark net' drug buyers say about their dealers - Telegraph* [Online]. Telegraph. Available: http://www.telegraph.co.uk/technology/internet/11466413/What-dark-net-drug-buyers-say-about-their-dealers.html [Accessed 2017].

Benevenuto, F., Magno, G., Rodrigues, T. & Almeida, V. Detecting spammers on twitter. Collaboration, electronic messaging, anti-abuse and spam conference (CEAS), 2010. 12.

Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J. & Gonçalves, M. Detecting spammers and content promoters in online video social networks. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009. ACM, 620-627.

Bengfort, B. 2014. *Computing a Bayesian Estimate of Star Rating Means* [Online]. District Data Labs. Available: https://districtdatalabs.silvrback.com/computing-a-bayesian-estimate-of-star-rating-means [Accessed 2017].

Berman, B. 2008. Strategies to detect and reduce counterfeiting activity. *Business Horizons,* 51**,** 191-199.

Berndt, D. J., McCart, J. A., Finch, D. K. & Luther, S. L. 2015. A Case Study of Data Quality in Text Mining Clinical Progress Notes. *ACM Trans. Manage. Inf. Syst.,* 6**,** 1-21.

Bian, J., Liu, Y., Agichtein, E. & Zha, H. Finding the right facts in the crowd: factoid question answering over social media. Proceedings of the 17th international conference on World Wide Web, 2008. ACM, 467-476.
Bian, J., Liu, Y., Zhou, D., Agichtein, E. & Zha, H. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. Proceedings of the 18th international conference on World wide web, 2009. ACM, 51-60.
Blei, D. 2013. Probabilistic Topic Models: Origins and Challenges. *2013 Topic Modeling Workshop at NIPS*.
Blei, D. M. 2012. Probabilistic topic models. *Communications of the ACM,* 55**,** 77-84.
Blei, D. M., Ng, A. Y. & Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research,* 3**,** 993-1022.
Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment,* 2008**,** P10008.
Broadhurst, R., Grabosky, P., Alazab, M., Bouhours, B. & Chon, S. 2014. An analysis of the nature of groups engaged in cyber crime.
Brody, S. & Elhadad, N. An unsupervised aspect-sentiment model for online reviews. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010. Association for Computational Linguistics, 804-812.
Buzydlowski, J. W. 2015. Co-occurrence analysis as a framework for data mining. *Journal of Technology Research* 6.
Cai, L. & Zhu, Y. 2015. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal,* 14.
Cambria, E., Schuller, B., Xia, Y. & Havasi, C. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems,* 28**,** 15-21.
Campbell, W. M., Dagli, C. K. & Weinstein, C. J. 2013. Social network analysis with content and graphs. *Lincoln Laboratory Journal,* 20**,** 61-81.
Castellano, R. 2016. *Exploratory visualization of Amazon fine food reviews - NYC Data Science Academy Blog* [Online]. Available: http://blog.nycdatascience.com/student-works/amazon-fine-foods-visualization/ [Accessed 2017].
Chai, K. E. K. 2011. *A machine learning-based approach for automated quality assessment of user generated content in web forums.* PhD, Citeseer.
Chen, L., Crawford, F. W. & Karbasi, A. 2016. Seeing the unseen network: inferring hidden social ties from respondent-driven sampling. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence.* Phoenix, Arizona: AAAI Press.
Chen, X., Cho, Y. & Jang, S. Y. Crime prediction using Twitter sentiment and weather. 2015 Systems and Information Engineering Design Symposium, 24-24 April 2015 2015. 63-68.
Chessa, A., Crimaldi, I., Riccaboni, M. & Trapin, L. 2014. Cluster analysis of weighted bipartite networks: a new copula-based approach. *PloS one,* 9**,** e109507.
Chiang, M. 2012. *Networked Life: 20 Questions and Answers*, Cambridge University Press.

Choi, J., Croft, W. B. & Kim, J. Y. Quality models for microblog retrieval. Proceedings of the 21st ACM international conference on Information and knowledge management, 2012. ACM, 1834-1838.
Christin, N. Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace. Proceedings of the 22nd international conference on World Wide Web, 2013. ACM, 213-224.
Clark, E. M., Williams, J. R., Jones, C. A., Galbraith, R. A., Danforth, C. M. & Dodds, P. S. 2016. Sifting robotic from organic text: a natural language approach for detecting automation on Twitter. *Journal of Computational Science,* 16**,** 1-7.
Compton, R. 2015. *Darknet Market Basket Analysis* [Online]. Available: http://ryancompton.net/2015/03/24/darknet-market-basket-analysis/ [Accessed 2017].
Congdon, K. 2015. *FBI Insights On Counterfeit Drug Prevention* [Online]. PharmaOnline. Available: https://www.pharmaceuticalonline.com/doc/fbi-insights-on-counterfeit-drug-prevention-0001 [Accessed].
Cornette, J. L., Ackerman, R. A. & Nykamp, D. Q. 2013. *A model of chemical pollution in a lake* [Online]. Math Insight. Available: http://mathinsight.org/chemical_pollution_lake_model [Accessed 2017].
Croft, B., Metzler, D. & Strohman, T. 2010. *Search Engines: Information Retrieval in Practice*, Pearson.
Danneman, N. & Heimann, R. 2014. *Social media mining with R*, Packt Publishing Ltd.
Dasu, T. 2013. Data glitches: Monsters in your data. *Handbook of Data Quality.* Springer.
Dasu, T. & Johnson, T. 2003. *Exploratory data mining and data cleaning*, John Wiley & Sons.
Debattista, J. 2017. *Scalable Quality Assessment of Linked Data.* Universitäts- und Landesbibliothek Bonn.
Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science,* 41**,** 391-407.
Dégardin, K., Roggo, Y. & Margot, P. 2014. Understanding and fighting the medicine counterfeit market. *Journal of Pharmaceutical and Biomedical Analysis,* 87**,** 167-175.
Duan, Y., Jiang, L., Qin, T., Zhou, M. & Shum, H.-Y. 2010. An empirical study on learning to rank of tweets. *Proceedings of the 23rd International Conference on Computational Linguistics.* Beijing, China: Association for Computational Linguistics.
Fabrício, B., Gabriel, M., Tiago, R. & Virgílio, A. Detecting spammers on twitter. Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS}, 2010.
Fan, W. 2015. Data Quality: From Theory to Practice. *SIGMOD Rec.,* 44**,** 7-18.
Faruqui, M. 2016. *Diverse Context for Learning Word Representations.* University of Trento.
Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M. & Ghosh, R. 2013. Exploiting Burstiness in Reviews for Review Spammer Detection. *ICWSM,* 13**,** 175-184.

Ferrara, E., De Meo, P., Catanese, S. & Fiumara, G. 2014. Visualizing criminal networks reconstructed from mobile phone records. *arXiv preprint arXiv:1407.2837*.
Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. 2016. The Rise of Social Bots. *Communications of the ACM, Vol. 59 No. 7.*
Fersini, E., Pozzi, F. & Messina, E. 2016. Approval network: a novel approach for sentiment analysis in social networks. *World Wide Web***,** 1-24.
Fortunato, S. 2010. Community detection in graphs. *Physics reports,* 486**,** 75-174.
Freitas, C., Benevenuto, F., Veloso, A. & Ghosh, S. 2016. An empirical study of socialbot infiltration strategies in the Twitter social network. *Social Network Analysis and Mining,* 6**,** 1-16.
Ganu, G., Elhadad, N. & Marian, A. Beyond the Stars: Improving Rating Predictions using Review Text Content. WebDB, 2009. Citeseer, 1-6.
Gaston, S. 2012. *Mining Twitter with R* [Online]. Available: https://sites.google.com/site/miningtwitter/home [Accessed 2017].
Gerber, M. S. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems,* 61**,** 115-125.
Girvan, M. & Newman, M. E. 2002. Community structure in social and biological networks. *Proceedings of the national academy of sciences,* 99**,** 7821-7826.
Golbeck, J. 2013. *Analyzing the social web*, Newnes.
Greenstadt, R. 2015. *The Upworthy Don: Formulas That Drive Google, Klout, Facebook Help Researchers Understand Organized Cybercrime - DrexelNow* [Online]. Available: http://drexel.edu/now/archive/2015/April/organized-cybercrime/ [Accessed 2017].
Gupta, A., Kumaraguru, P., Castillo, C. & Meier, P. Tweetcred: Real-time credibility assessment of content on twitter. International Conference on Social Informatics, 2014. Springer, 228-243.
Herzog, T. N., Scheuren, F. J. & Winkler, W. E. 2007. What is Data Quality and Why Should We Care? *In:* HERZOG, T. N., SCHEUREN, F. J. & WINKLER, W. E. (eds.) *Data Quality and Record Linkage Techniques.* New York, NY: Springer New York.
Hofmann, T. Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999. ACM, 50-57.
Hornik, K. & Grün, B. 2011. topicmodels: An R package for fitting topic models. *Journal of Statistical Software,* 40**,** 1-30.
Hu, M. & Liu, B. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004. ACM, 168-177.
Isah, H., Trundle, P. & Neagu, D. Social media analysis for product safety using text mining and sentiment analysis. 2014 14th UK Workshop on Computational Intelligence (UKCI), 8-10 Sept. 2014 2014. 1-7.
Jindal, N. & Liu, B. Opinion spam and analysis. Proceedings of the 2008 International Conference on Web Search and Data Mining, 2008. ACM, 219-230.
Jindal, N., Liu, B. & Lim, E.-P. Finding unusual review patterns using unexpected rules. Proceedings of the 19th ACM international

conference on Information and knowledge management, 2010. ACM, 1549-1552.
Jøsang, A. 2008. Online reputation systems for the health sector. *Electronic Journal of Health Informatics,* 3**,** 8.
Josang, A. & Ismail, R. The beta reputation system. Proceedings of the 15th bled electronic commerce conference, 2002. 2502-2511.
Julia, S. D., Robinson 2017. *Text Mining with R*

A Tidy Approach, O'Reilly.
Jurka, T. P. 2012. sentiment: Tools for Sentiment Analysis. R package version 0.2.
Kahn, B. K., Strong, D. M. & Wang, R. Y. 2002. Information quality benchmarks: product and service performance. *Communications of the ACM,* 45**,** 184-192.
Kaiser, C. & Bodendorf, F. Mining Patient Experiences on Web 2.0 - A Case Study in the Pharmaceutical Industry. 2012 Annual SRII Global Conference, 24-27 July 2012 2012. 139-145.
Ke, T., Fabian, A., Claudia, H. & Geert-jan, H. 2012. What makes a tweet relevant for a topic? *2nd Workshop on Making Sense of Microposts.*
Kessler, J. S. & Nicolov, N. 2015. The JDPA Sentiment Corpus for the Automotive Domain. *The Handbook of Linguistic Annotation*.
Kevin, M. 2009. *Twitter API: Up and Running*

Learn How to Build Applications with the Twitter API, O'Reilly Media.
Kirchner, C. & Gade, J. 2011. Implementing social network analysis for fraud prevention. *CGI Group Ind*.
Kiritchenko, S., Zhu, X. & Mohammad, S. M. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research,* 50**,** 723-762.
Kitsak, M. & Krioukov, D. 2011. Hidden variables in bipartite networks. *Physical Review E,* 84**,** 026114.
Kötter, T., Günnemann, S., Faloutsos, C. & Berthold, M. R. Extracting taxonomies from bipartite graphs. Proceedings of the 24th International Conference on World Wide Web, 2015. ACM, 51-52.
Krech, L. A., El-Hadri, L., Evans, L., Fouche, T., Hajjou, M., Lukulay, P., Phanouvong, S., Pribluda, V. & Roth, L. 2014. The Medicines Quality Database: a free public resource. *Bulletin of the World Health Organization,* 92**,** 2-2A.
Kriegler, A. 2014. Using social network analysis to profile organised crime.
Kriti, A. R., Needleman 2017. Can You Trust Reviews on Amazon? *Webpage*.
Kyumin, L., Brian, D. E. & James, C. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. International AAAI Conference on Web and Social Media (ICWSM'11), 2011.
Kyumin, L., James, C. & Calton, P. 2014. Social Media Threats and Countermeasures.
Labouseur, A. G. & Matheus, C. C. 2017. An Introduction to Dynamic Data Quality Challenges. *J. Data and Information Quality,* 8**,** 1-3.
Landy, J. 2015. *How not to sort by average rating revisited* [Online]. Available: http://efavdb.com/ranking-revisited/ [Accessed 2017].
Latapy, M., Magnien, C. & Vecchio, N. D. 2008. Basic notions for the analysis of large two-mode networks. *Social Networks,* 30**,** 31-48.

Leong, E. 2016. *New Ways to Control Your Experience on Twitter* [Online]. Available: https://blog.twitter.com/2016/new-ways-to-control-your-experience-on-twitter [Accessed].
Leskovec, J. 2008. *Dynamics of large networks.* Carnegie Mellon University.
Leskovec, J., Rajaraman, A. & Ullman, J. D. 2014. *Mining of Massive Datasets*, Cambridge University Press.
Leydesdorff, L. & Vaughan, L. 2006. Co-occurrence matrices and their applications in information science: extending ACA to the web environment. *Journal of the Association for Information Science and Technology,* 57**,** 1616-1628.
Li, H., Chen, Z., Liu, B., Wei, X. & Shao, J. Spotting fake reviews via collective positive-unlabeled learning. Data Mining (ICDM), 2014 IEEE International Conference on, 2014a. IEEE, 899-904.
Li, H., Fei, G., Wang, S., Liu, B., Shao, W., Mukherjee, A. & Shao, J. 2017. Bimodal Distribution and Co-Bursting in Review Spam Detection.
Li, H., Mukherjee, A., Liu, B., Kornfield, R. & Emery, S. Detecting campaign promoters on twitter using markov random fields. Data Mining (ICDM), 2014 IEEE International Conference on, 2014b. IEEE, 290-299.
Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B. & Lauw, H. W. Detecting product review spammers using rating behaviors. Proceedings of the 19th ACM international conference on Information and knowledge management, 2010. ACM, 939-948.
Lim, E. P., Vuong, B.-Q., Lauw, H. W. & Sun, A. 2006. Measuring qualities of articles contributed by online communities.
Lim, K. W. & Buntine, W. Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, 2014. ACM, 1319-1328.
Lin, C. & He, Y. Joint sentiment/topic model for sentiment analysis. Proceedings of the 18th ACM conference on Information and knowledge management, 2009. ACM, 375-384.
Liu, B. 2012. *Sentiment analysis and opinion mining*.
Liu, B. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*, Cambridge University Press.
Liu, H. Some computational challenges in mining social media. 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), 25-28 Aug. 2013 2013. xxxvii-xxxvii.
Liu, J., Cao, Y., Lin, C.-Y., Huang, Y. & Zhou, M. Low-Quality Product Review Detection in Opinion Summarization. EMNLP-CoNLL, 2007. 334-342.
Liu, X., Datta, A. & Lim, E.-P. 2014. *Computational trust models and machine learning*, CRC Press.
Lu, Y., Tsaparas, P., Ntoulas, A. & Polanyi, L. Exploiting social context for review quality prediction. Proceedings of the 19th international conference on World wide web, 2010. ACM, 691-700.
Maas, A. L. & Ng, A. Y. A probabilistic model for semantic word vectors. NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2010. 1-8.
Manning, C. D., Raghavan, P. & Schütze, H. 2008. *Introduction to Information Retrieval*, Cambridge University Press.

Masurel, P. 2013. *Of Bayesian average and star ratings* [Online]. Available: http://fulmicoton.com/posts/bayesian_rating/ [Accessed 2017].
Mathew, A. 2015. Logistic Regression in R – Part Two. *Mathew Analytics.*
McAuley, J. & Leskovec, J. Hidden factors and hidden topics: understanding rating dimensions with review text. Proceedings of the 7th ACM conference on Recommender systems, 2013a. ACM, 165-172.
McAuley, J., Leskovec, J. & Jurafsky, D. Learning attitudes and attributes from multi-aspect reviews. Data Mining (ICDM), 2012 IEEE 12th International Conference on, 2012. IEEE, 1020-1025.
McAuley, J. J. & Leskovec, J. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. Proceedings of the 22nd international conference on World Wide Web, 2013b. ACM, 897-908.
Mcauliffe, J. D. & Blei, D. M. Supervised topic models. Advances in neural information processing systems, 2008. 121-128.
McCoy, D., Dharmdasani, H., Kreibich, C., Voelker, G. M. & Savage, S. Priceless: The role of payments in abuse-advertised goods. Proceedings of the 2012 ACM conference on Computer and communications security, 2012a. ACM, 845-856.
McCoy, D., Pitsillidis, A., Jordan, G., Weaver, N., Kreibich, C., Krebs, B., Voelker, G. M., Savage, S. & Levchenko, K. Pharmaleaks: Understanding the business of online pharmaceutical affiliate programs. Proceedings of the 21st USENIX conference on Security symposium, 2012b. USENIX Association, 1-1.
McCue, C. 2014. *Data mining and predictive analysis: Intelligence gathering and crime analysis*, Butterworth-Heinemann.
McGlohon, M., Glance, N. S. & Reiter, Z. Star Quality: Aggregating Reviews to Rank Products and Merchants. ICWSM, 2010.
Mena, J. 2011. *Machine learning forensics for law enforcement, security, and intelligence*, CRC Press.
Miller, E. 2009. *How Not To Sort By Average Rating* [Online]. Available: http://www.evanmiller.org/how-not-to-sort-by-average-rating.html [Accessed 2017].
Miller, E. 2012. *Bayesian Average Ratings* [Online]. Available: http://www.evanmiller.org/bayesian-average-ratings.html [Accessed].
Miller, E. 2014. *Ranking Items With Star Ratings: An Approximate Bayesian Approach* [Online]. Available: http://www.evanmiller.org/ranking-items-with-star-ratings.html [Accessed 2017].
Miller, E. 2015a. *Deriving the Reddit Formula* [Online]. Available: http://www.evanmiller.org/deriving-the-reddit-formula.html [Accessed 2017].
Miller, E. 2015b. *Evaluating Splatoon's Ranking System* [Online]. Available: http://www.evanmiller.org/evaluating-splatoons-ranking-system.html [Accessed 2017].
Miller, E. 2015c. *Inferring Tweet Quality From Retweets* [Online]. Available: http://www.evanmiller.org/inferring-tweet-quality-from-retweets.html [Accessed 2017].
Miller, E. 2015d. *Ranking News Items With Upvotes* [Online]. Available: http://www.evanmiller.org/ranking-news-items-with-upvotes.html [Accessed 2017].

Miller, H., Thebault-Spieker, J., Chang, S., Johnson, I., Terveen, L. & Hecht, B. 2016. Blissfully happy" or "ready to fight": Varying Interpretations of Emoji. *Proceedings of ICWSM,* 2016.

Moghaddam, S. & Ester, M. 2011. ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.* Beijing, China: ACM.

Momeni, E., Cardie, C. & Ott, M. Properties, Prediction, and Prevalence of Useful User-Generated Comments for Descriptive Annotation of Social Media Objects. ICWSM, 2013.

Mosadeghrad, A. M. 2014. Factors influencing healthcare service quality. *International journal of health policy and management,* 3**,** 77.

Motoyama, M., McCoy, D., Levchenko, K., Savage, S. & Voelker, G. M. An analysis of underground forums. Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, 2011. ACM, 71-80.

Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M. & Ghosh, R. Spotting opinion spammers using behavioral footprints. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013a. ACM, 632-640.

Mukherjee, A., Liu, B. & Glance, N. Spotting fake reviewer groups in consumer reviews. Proceedings of the 21st international conference on World Wide Web, 2012. ACM, 191-200.

Mukherjee, A., Venkataraman, V., Liu, B. & Glance, N. S. What yelp fake review filter might be doing? ICWSM, 2013b.

Nadji, Y., Antonakakis, M., Perdisci, R. & Lee, W. 2013. Connected Colors: Unveiling the Structure of Criminal Networks. *In:* STOLFO, S. J., STAVROU, A. & WRIGHT, C. V. (eds.) *Research in Attacks, Intrusions, and Defenses: 16th International Symposium, RAID 2013, Rodney Bay, St. Lucia, October 23-25, 2013. Proceedings.* Berlin, Heidelberg: Springer Berlin Heidelberg.

Nathan, D. & Richard, H. 2014. *Social Media Mining with R*, Packt Publishing

Nayer, W., Motaz, E.-S., Heba, A. & Waleed, A. Automatic Scoring of Online Discussion Posts. 2nd ACM workshop on Information credibility on the web (WICOW '08), 2008 Napa Valley, California.

Nishikawa, H., Hasegawa, T., Matsuo, Y. & Kikui, G. Optimizing informativeness and readability for sentiment summarization. Proceedings of the ACL 2010 Conference Short Papers, 2010. Association for Computational Linguistics, 325-330.

Oentaryo, R. J., Murdopo, A., Prasetyo, P. K. & Lim, E.-P. 2016. On Profiling Bots in Social Media. *In:* SPIRO, E. & AHN, Y.-Y. (eds.) *Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part I.* Cham: Springer International Publishing.

Opsahl, T. 2013. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks,* 35**,** 159-167.

Ott, M., Cardie, C. & Hancock, J. Estimating the prevalence of deception in online review communities. Proceedings of the 21st international conference on World Wide Web, 2012. ACM, 201-210.

Ott, M., Choi, Y., Cardie, C. & Hancock, J. T. Finding deceptive opinion spam by any stretch of the imagination. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 2011. Association for Computational Linguistics, 309-319.

Palla, G., Derényi, I., Farkas, I. & Vicsek, T. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *arXiv preprint physics/0506133*.

Park, Y., Jones, J., McCoy, D., Shi, E. & Jakobsson, M. 2014. Scambaiter: Understanding targeted nigerian scams on craigslist. *system,* 1**,** 2.

Paul, M. J. & Dredze, M. 2011. You are what you Tweet: Analyzing Twitter for public health. *Icwsm,* 20**,** 265-272.

Paulo, D., Fischl, B., Markow, T., Martin, M. & Shakarian, P. 2013. Social network intelligence analysis to combat street gang violence. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* Niagara, Ontario, Canada: ACM.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. & Dubourg, V. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research,* 12**,** 2825-2830.

Pelleg, D., Rokhlenko, O., Szpektor, I., Agichtein, E. & Guy, I. When the crowd is not enough: Improving user experience with social media through automatic quality analysis. Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, 2016. ACM, 1080-1090.

Pfeffer, J. & Carley, K. M. 2012. Rapid modeling and analyzing networks extracted from pre-structured news articles. *Computational and Mathematical Organization Theory,* 18**,** 280-299.

Pipino, L. L., Lee, Y. W. & Wang, R. Y. 2002. Data quality assessment. *Commun. ACM,* 45**,** 211-218.

Pons, P. & Latapy, M. 2006. Computing communities in large networks using random walks. *J. Graph Algorithms Appl.,* 10**,** 191-218.

Postolache, O., Girão, P. S. & Pereira, J. M. D. 2012. Water quality monitoring and associated distributed measurement systems: An Overview. *Water Quality Monitoring and Assessment.* InTech.

Prusa, J. D. 2015. *An evaluation of machine learning algorithms for tweet sentiment analysis*, Florida Atlantic University.

Quality, E. D. 2015. The data quality benchmark report. *Experian Data Quality***,** 1-10.

Ranchal, R., Mohindra, A., Zhou, N., Kapoor, S. & Bhargava, B. Hierarchical Aggregation of Consumer Ratings for Service Ecosystem. 2015 IEEE International Conference on Web Services, June 27 2015-July 2 2015 2015. 575-582.

Ravindran, S. K. & Garg, V. 2015. *Mastering social media mining with R*, Packt Publishing Ltd.

Reddy, E. E. 2013. Social Data vs Social Media.

Salihefendic, A. 2015a. *How Hacker News ranking algorithm works – Hacking and Gonzo – Medium* [Online]. @Medium. Available:

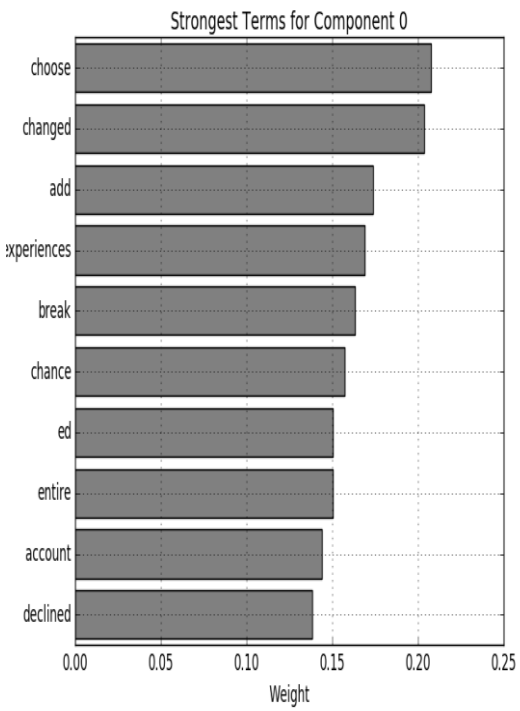
https://medium.com/hacking-and-gonzo/how-hacker-news-ranking-algorithm-works-1d9b0cf2c08d [Accessed 2017].
Salihefendic, A. 2015b. *How Reddit ranking algorithms work – Hacking and Gonzo – Medium* [Online]. @Medium. Available: https://medium.com/hacking-and-gonzo/how-reddit-ranking-algorithms-work-ef111e33d0d9 [Accessed 2017].
Samatova, N. F., Hendrix, W., Jenkins, J., Padmanabhan, K. & Chakraborty, A. 2013. *Practical graph mining with R*, CRC Press.
Sarvari, H., Abozinadah, E., Mbaziira, A. & Mccoy, D. Constructing and Analyzing Criminal Networks. 2014 IEEE Security and Privacy Workshops, 17-18 May 2014 2014. 84-91.
Scrivens, R., Davies, G. & Frank, R. 2017. Searching for signs of extremism on the web: an introduction to Sentiment-based Identification of Radical Authors. *Behavioral Sciences of Terrorism and Political Aggression***,** 1-21.
Sessions, V. & Valtorta, M. 2006. The Effects of Data Quality on Machine Learning Algorithms. *ICIQ,* 6**,** 485-498.
Sharma, U., Suman, A. & Shannigrahi, S. 2014. Inferring social ties from common activities in twitter. *Proceedings of the 25th ACM conference on Hypertext and social media.* Santiago, Chile: ACM.
Soska, K. & Christin, N. Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem. USENIX Security Symposium, 2015. 33-48.
Spellman, F. R. 2013. *Handbook of Water and Wastewater Treatment Plant Operations*, CRC Press
Stevens, K., Kegelmeyer, P., Andrzejewski, D. & Buttler, D. Exploring topic coherence over many models and many topics. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012. Association for Computational Linguistics, 952-961.
Survey, U. S. G. 2015. *National Field Manual for the Collection of Water Quality Data: U.S. Geological Survey Techniques of Water-Resources Investigations*, book 9, chaps. A1-A10, available online at http://pubs.water.usgs.gov/twri9A.
Tan, Y., Zhang, M., Liu, Y. & Ma, S. Rating-Boosted Latent Topics: Understanding Users and Items with Ratings and Reviews. IJCAI, 2016. 2640-2646.
Tang, J., Lou, T. & Kleinberg, J. 2012. Inferring social ties across heterogenous networks. *Proceedings of the fifth ACM international conference on Web search and data mining.* Seattle, Washington, USA: ACM.
Tang, W., Zhuang, H. & Tang, J. 2011. Learning to Infer Social Ties in Large Networks. *In:* GUNOPULOS, D., HOFMANN, T., MALERBA, D. & VAZIRGIANNIS, M. (eds.) *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III.* Berlin, Heidelberg: Springer Berlin Heidelberg.
Tayebi, M. A. 2015. *Predictive Models for Public Safety Using Social Network Analysis.* Applied Sciences: School of Computing Science.
Tayebi, M. A. & Glässer, U. 2016. Social Network Analysis in Predictive Policing. *Social Network Analysis in Predictive Policing.* Springer.

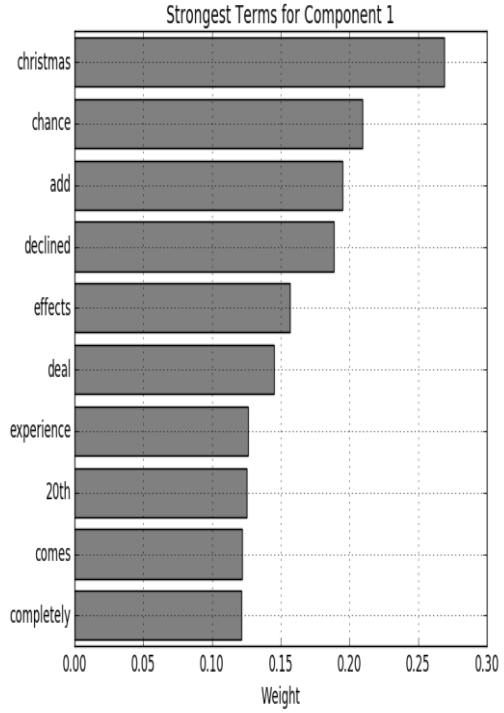
Thomo, A. 2009. *Latent Semantic Analysis Tutorial* [Online]. Available: http://alexthomo.blogspot.com/2009/03/latent-semantic-analysis-tutorial.html [Accessed June 2017].
Tsvetovat, M. & Kouznetsov, A. 2011. *Social Network Analysis for Startups: Finding connections on the social web*, " O'Reilly Media, Inc.".
Turney, P. D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th annual meeting on association for computational linguistics, 2002. Association for Computational Linguistics, 417-424.
Turney, P. D. & Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research,* 37**,** 141-188.
UNODC 2011. *Criminal Intelligence: Manual for Analysts*, United Nations Publications.
Vosecky, J., Leung, K. W.-T. & Ng, W. Searching for quality microblog posts: Filtering and ranking based on content analysis and implicit links. International Conference on Database Systems for Advanced Applications, 2012. Springer, 397-413.
Wang, G., Xie, S., Liu, B. & Yu, P. S. 2012a. Identify online store review spammers via social review graph. *ACM Transactions on Intelligent Systems and Technology (TIST),* 3**,** 61.
Wang, J., Yu, C. T., Yu, P. S., Liu, B. & Meng, W. 2015. Diversionary comments under blog posts. *ACM Transactions on the Web (TWEB),* 9**,** 18.
Wang, Q., Xu, J., Li, H. & Craswell, N. 2013. Regularized Latent Semantic Indexing: A New Approach to Large-Scale Topic Modeling. *ACM Trans. Inf. Syst.,* 31**,** 1-44.
Wang, X., Brown, D. E. & Gerber, M. S. Spatio-temporal modeling of criminal incidents using geographic, demographic, and twitter-derived information. 2012 IEEE International Conference on Intelligence and Security Informatics, 11-14 June 2012 2012b. 36-41.
Wang, X., Gerber, M. S. & Brown, D. E. Automatic crime prediction using events extracted from twitter posts. International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, 2012c. Springer, 231-238.
Webberley, W. M., Allen, S. M. & Whitaker, R. M. 2016. Retweeting beyond expectation: Inferring interestingness in Twitter. *Computer Communications,* 73**,** 229-235.
Weimer, M., Gurevych, I. & Mühlhäuser, M. 2007. Automatically assessing the post quality in online discussions on software. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions.* Prague, Czech Republic: Association for Computational Linguistics.
Wertheimer, A. I. & Wang, P. G. 2012. *Counterfeit medicines volume I: Policy, economics and countermeasures*, ILM Publications.
Wilde, F., Radtke, D., Gibs, J. & Iwatsubo, R. 1998. National field manual for the collection of water-quality data: US Geological Survey Techniques of Water-Resources Investigations, book 9, chap. *A-6, variously paged*.
Xiao, X., Zheng, Y., Luo, Q. & Xie, X. 2014. Inferring social ties between users with human location history. *Journal of Ambient Intelligence and Humanized Computing,* 5**,** 3-19.

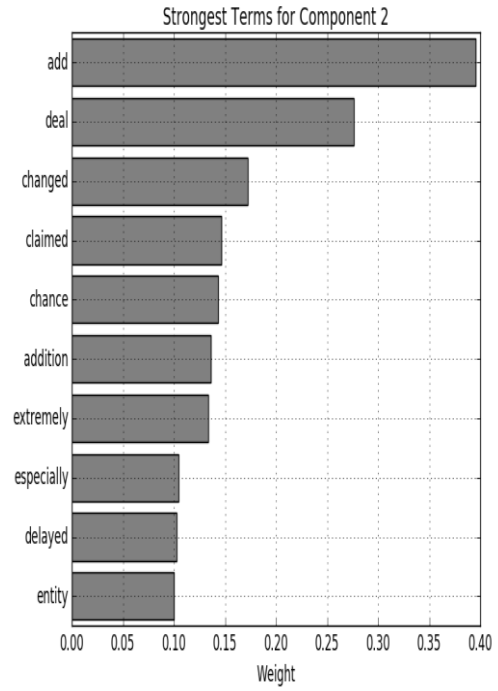
Yang, C., Harkreader, R., Zhang, J., Shin, S. & Gu, G. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. Proceedings of the 21st international conference on World Wide Web, 2012. ACM, 71-80.
Yang, X., Yang, G. & Wu, J. 2016. Integrating rich and heterogeneous information to design a ranking system for multiple products. *Decision Support Systems,* 84**,** 117-133.
Yardi, S., Romero, D. & Schoenebeck, G. 2009. Detecting spam in a twitter network. *First Monday,* 15.
Yip, M., Shadbolt, N. & Webber, C. Structural analysis of online criminal social networks. 2012 IEEE International Conference on Intelligence and Security Informatics, 11-14 June 2012 2012. 60-65.
Zafarani, R., Abbasi, M. A. & Liu, H. 2014. *Social media mining: an introduction*, Cambridge University Press.
Zhai, C. 2017. *Text Mining and Analytics* | *Coursera* [Online]. Coursera. Available: https://www.coursera.org/learn/text-mining [Accessed].
Zhang, K., Cheng, Y., Liao, W.-k. & Choudhary, A. 2012. Mining millions of reviews: a technique to rank products based on importance of reviews. *Proceedings of the 13th International Conference on Electronic Commerce.* Liverpool, United Kingdom: ACM.
Zhang, L. & Liu, B. 2014. Aspect and Entity Extraction for Opinion Mining. *In:* CHU, W. W. (ed.) *Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenge and Opportunities.* Berlin, Heidelberg: Springer Berlin Heidelberg.
Zhong, C. 2013. *Improvements on Graph-based Clustering Methods*.
Zhuang, H., Tang, J., Tang, W., Lou, T., Chin, A. & Wang, X. 2012. Actively learning to infer social ties. *Data Min. Knowl. Discov.,* 25**,** 270-297.

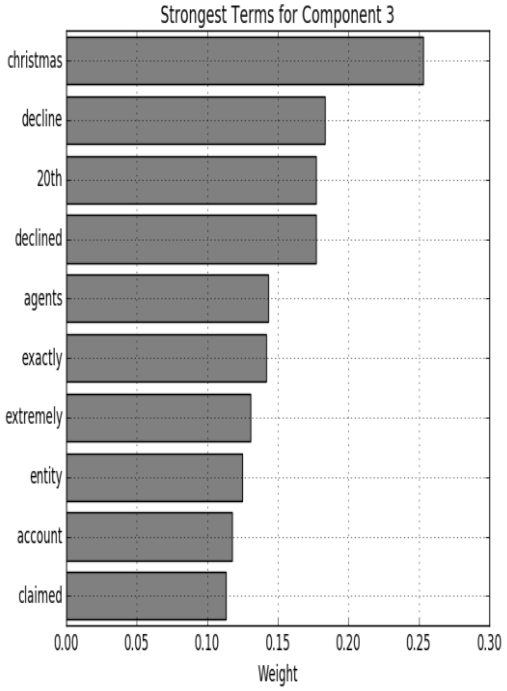
Appendices

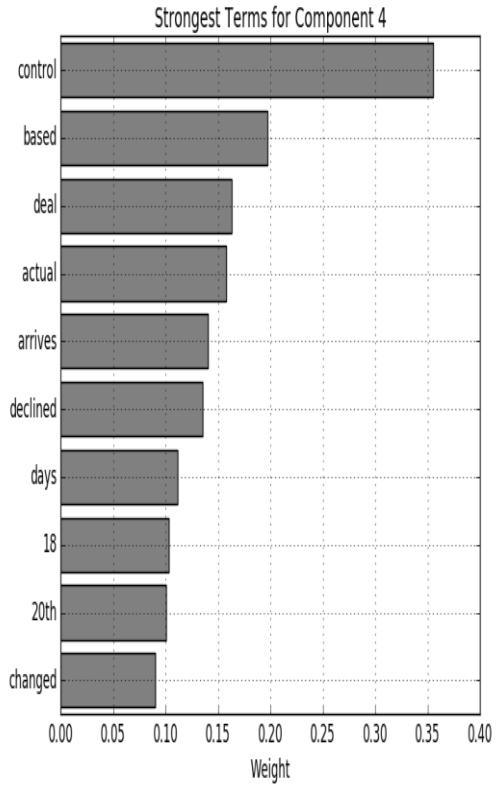
Appendix I: The Top Ten Terms and Components in Generic Doctor Review Corpus

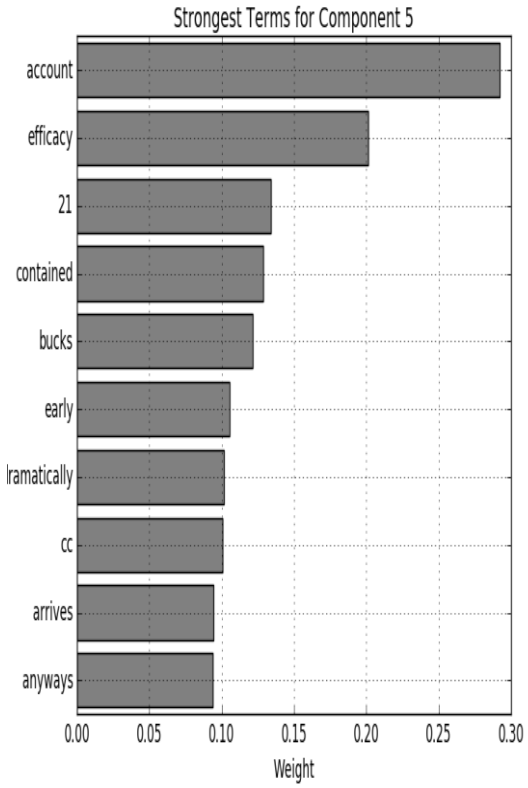


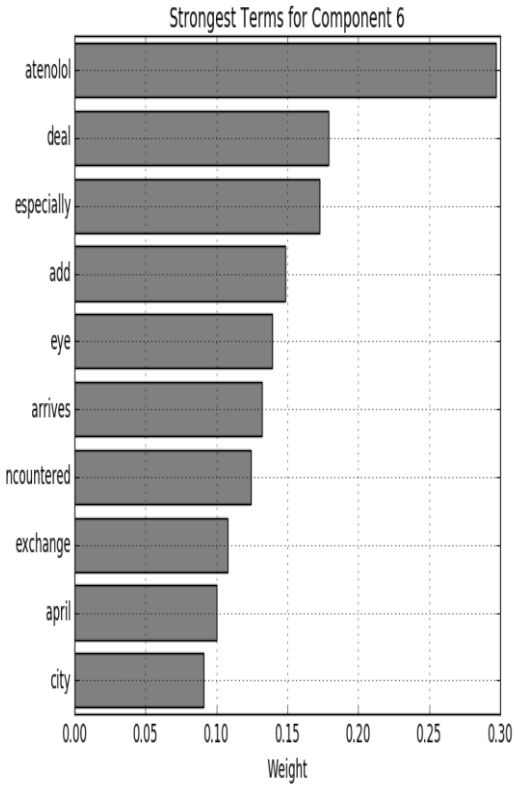


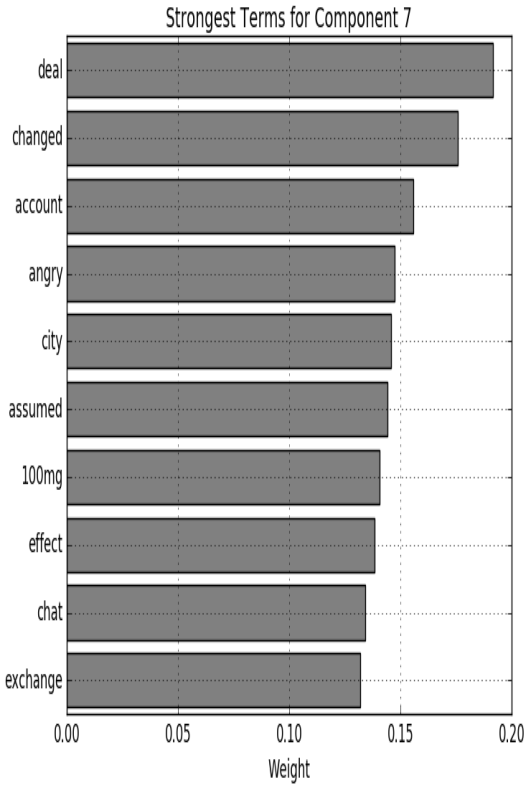


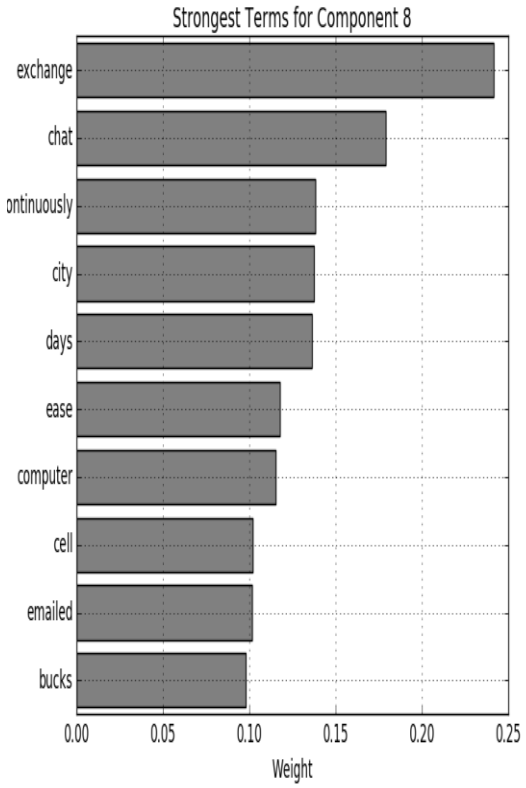


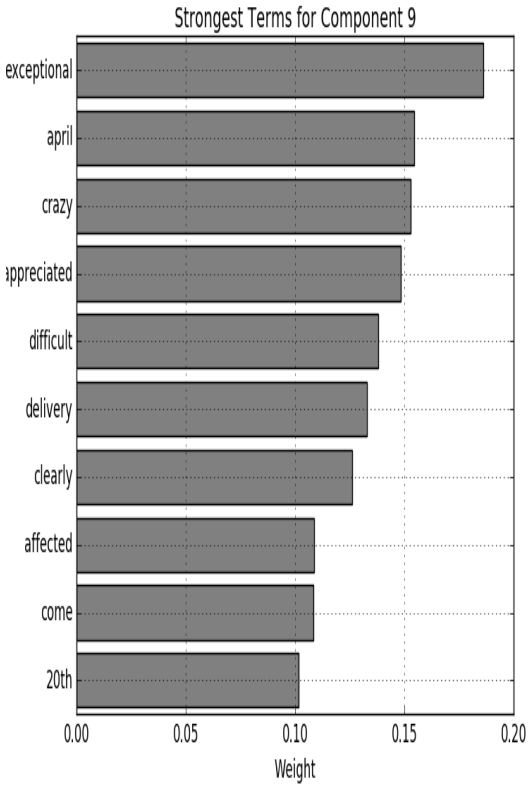

Strongest Terms for Component 4
control
based
deal
actual
arrives
declined
days
18
20th
changed
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Weight


Strongest Terms for Component 5
account
efficacy
21
contained
bucks
early
ramatically
cc
arrives
anyways
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Weight

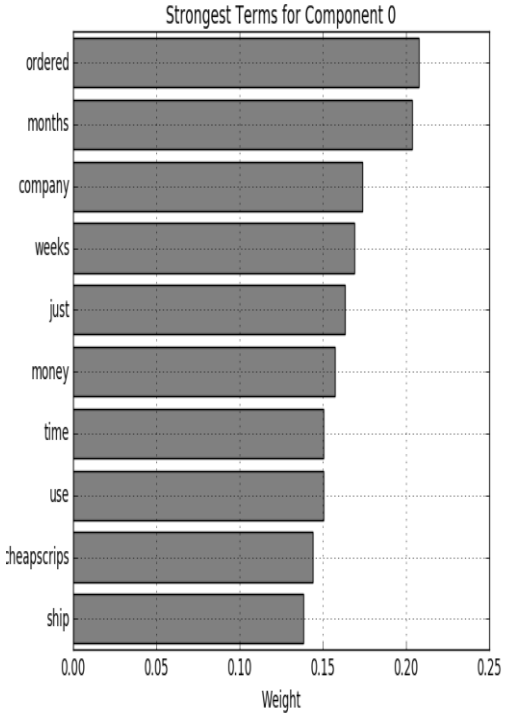

Strongest Terms for Component 6
atenolol
deal
especially
add
eye
arrives
ncountered
exchange
april
city
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Weight

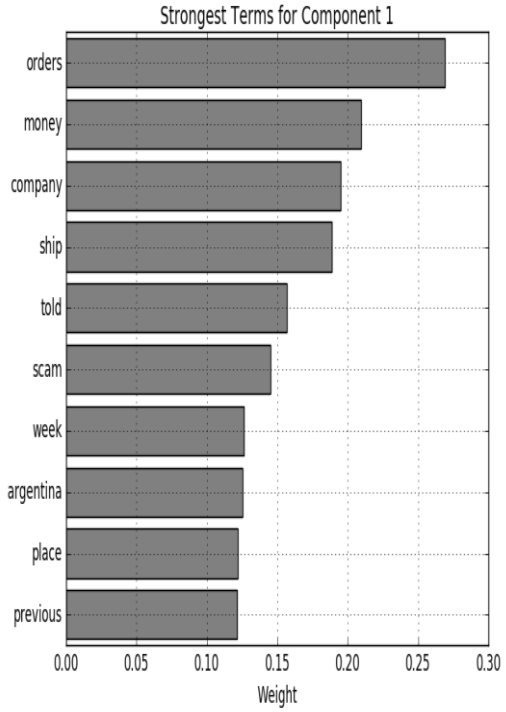

Strongest Terms for Component 7
deal
changed
account
angry
city
assumed
100mg
effect
chat
exchange
0.00
0.05
0.10
0.15
0.20
Weight

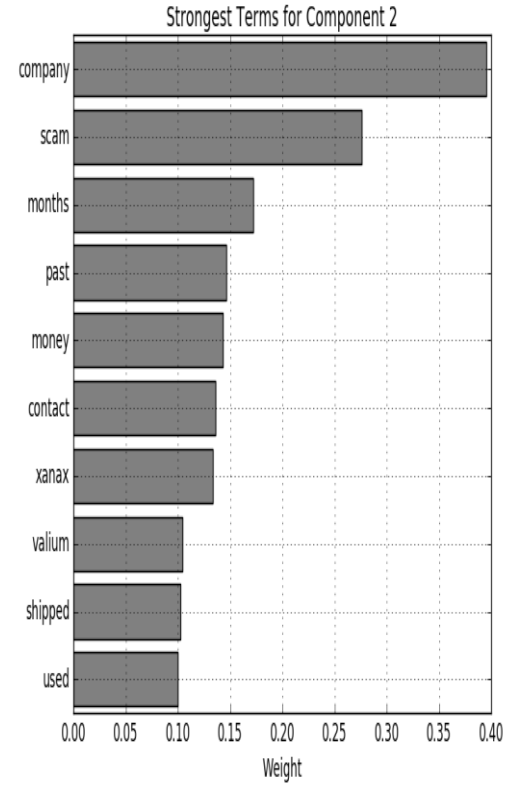

Strongest Terms for Component 8
exchange
chat
ontinuously
city
days
ease
computer
cell
emailed
bucks
0.00
0.05
0.10
0.15
0.20
0.25
Weight

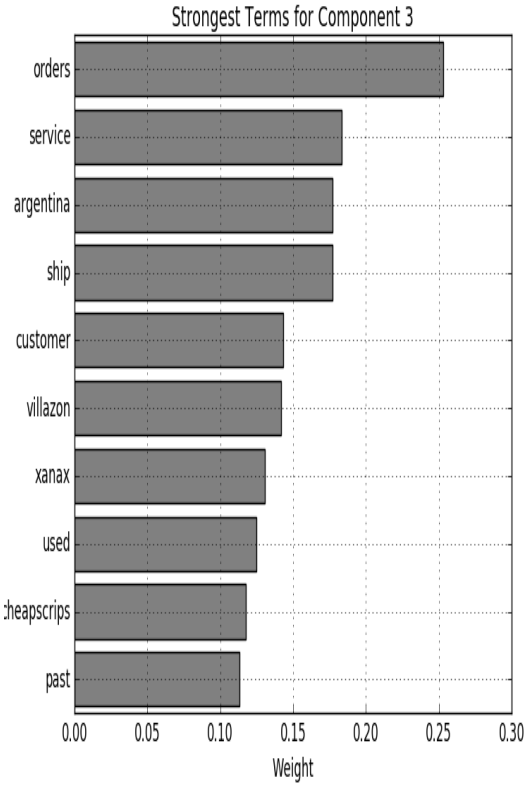

Strongest Terms for Component 9
exceptional
april
crazy
appreciated
difficult
delivery
clearly
affected
come
20th
0.00
0.05
0.10
0.15
0.20
Weight

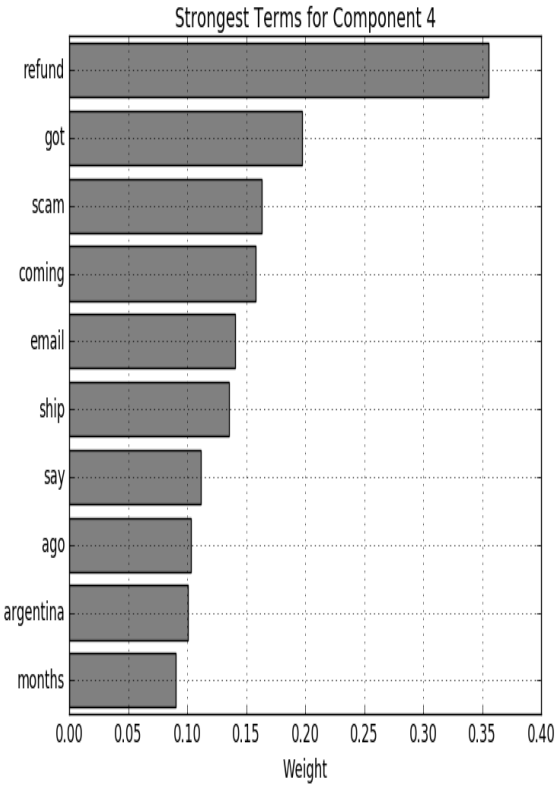
Appendix II: The Top Ten Terms and Components in Cheap Scrips Review Corpus

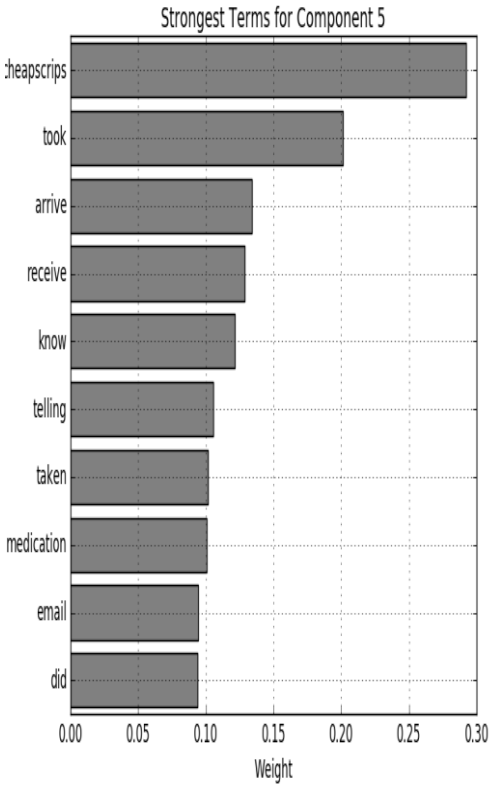


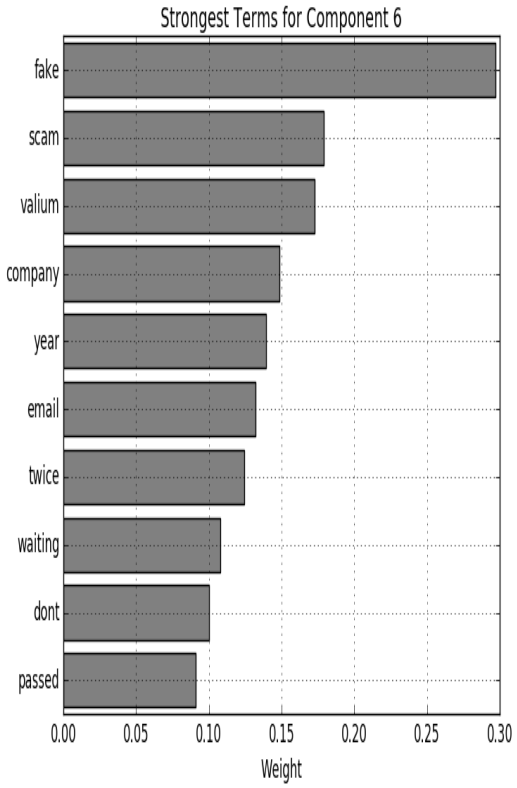


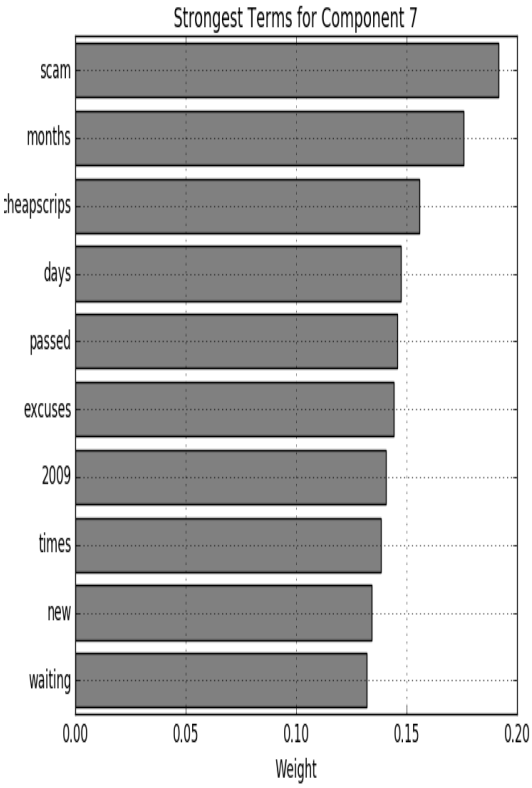


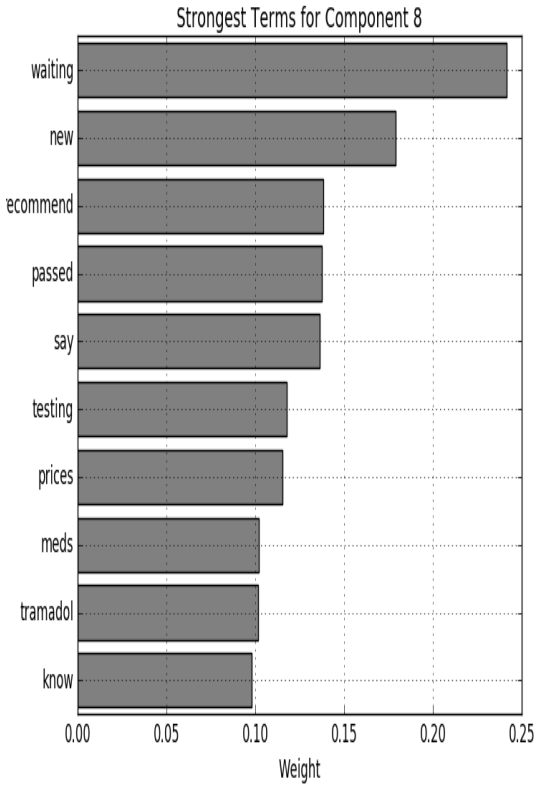


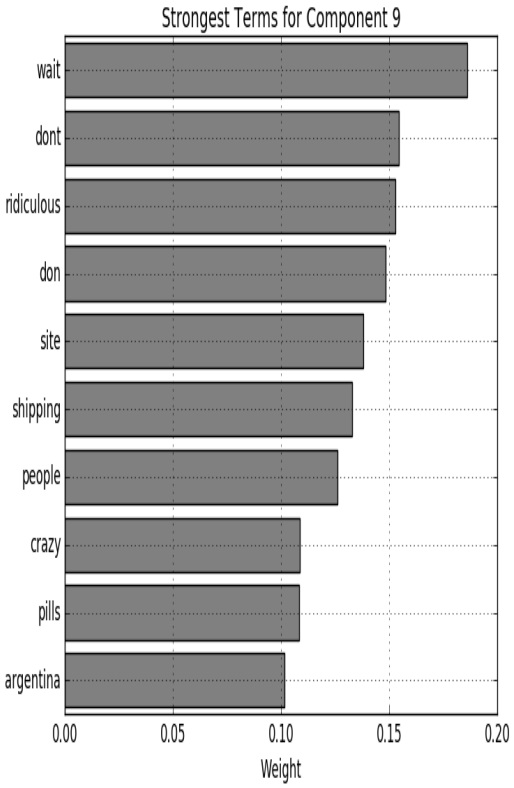

Strongest Terms for Component 4
refund
got
scam
coming
email
ship
say
ago
argentina
months
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Weight


Strongest Terms for Component 5
:heapscrips
took
arrive
receive
know
telling
taken
medication
email
did
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Weight


Strongest Terms for Component 6
fake
scam
valium
company
year
email
twice
waiting
dont
passed
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Weight


Strongest Terms for Component 7
scam
months
:heapscrips
days
passed
excuses
2009
times
new
waiting
0.00
0.05
0.10
0.15
0.20
Weight


Strongest Terms for Component 8
waiting
new
ecommend
passed
say
testing
prices
meds
tramadol
know
0.00
0.05
0.10
0.15
0.20
0.25
Weight


Strongest Terms for Component 9
wait
dont
ridiculous
don
site
shipping
people
crazy
pills
argentina
0.00
0.05
0.10
0.15
0.20
Weight

Appendix III: UKCI2014

Social Media Analysis for Product Safety using Text Mining and Sentiment Analysis

Haruna Isah, Paul Trundle, Daniel Neagu
Artificial Intelligence Research (AIRe) Group
School of Electrical Engineering and Computer Science
University of Bradford
Bradford, UK
H.Isah@student.bradford.ac.uk, P.R.Trundle@bradford.ac.uk, D.Neagu@bradford.ac.uk

***Abstract*— The growing incidents of counterfeiting and associated economic and health consequences necessitate the development of active surveillance systems capable of producing timely and reliable information for all stake holders in the anti-counterfeiting fight. User generated content from social media platforms can provide early clues about product allergies, adverse events and product counterfeiting. This paper reports a work in progress with contributions including: the development of a framework for gathering and analyzing the views and experiences of users of drug and cosmetic products using machine learning, text mining and sentiment analysis; the application of the proposed framework on Facebook comments and data from Twitter for brand analysis, and the description of how to develop a product safety lexicon and training data for modeling a machine learning classifier for drug and cosmetic product sentiment prediction. The initial brand and product comparison results signify the usefulness of text mining and sentiment analysis on social media data while the use of machine learning classifier for predicting the sentiment orientation provides a useful tool for users, product manufacturers, regulatory and enforcement agencies to monitor brand or product sentiment trends in order to act in the event of sudden or significant rise in negative sentiment.**

Keywords*— *text mining; sentiment analysis; product safety; social media; machine learning

I. INTRODUCTION

The global scourge of counterfeit drug and cosmetic products poses a great threat to public safety. One strategy for combating counterfeit products is through the effective communication and tracking of early warning signals of product allergies, side or adverse effects, drug resistance and disease outbreaks [*1*]. Social media is now a platform for sharing virtually all kinds of information; studies on Twitter data have demonstrated that aggregating millions of messages can provide valuable insights into a population [*2*], hence, pharmaceutical manufacturers need to track patients' opinions on their products for decision making [*3*].

To address the menace of drug and cosmetic products counterfeiting, a surveillance system capable of harnessing and tracking online reported views and experiences of users of these products is needed. We propose novel research questions that include the following:

1. What is the public sentiment about a given brand (s) of drug and cosmetic product?

2. By monitoring conversation about a given brand (s) of drug and cosmetic products over a sample user population, can we be able to predict evolving adverse effects and the possibility of product counterfeiting?

We approach these problems by applying text mining and sentiment analysis techniques on drug and cosmetic product related text data. The following are the contributions of this work:

1. We propose a product safety framework for harnessing and processing the views and experiences of users of popular brands of drug and cosmetic products reported on social media platforms using text mining and sentiment analysis.

2. We utilised the framework to gather and analyse views and experiences of users of cosmetic and drug products in form of tweets and Facebook comments using lexicon and machine learning approaches.

3. We demonstrated how to develop custom lexicon and training data and also modeled a Naive Bayes classifier for sentiment prediction of views and experiences of users of popular brands of drug and cosmetic products.

This work is organised thus: Section II provides a review of related work; Section III details the structure and flow of the proposed framework; Section IV demonstrates the application of the framework using two case studies. The final section describes conclusions and future work.

II. RELATED WORK

Text mining is a specialised domain that applies data mining techniques over text; some of the early attempts of exploratory data analysis over text are recorded in [*4*]. Sentiment analysis aims to identify and extract opinions, moods and attitudes of individuals and communities; the authors of [*5*] provided a technical survey and early work on sentiment analysis. When text mining and sentiment analysis techniques are combined in a project on social media data, the result is often a powerful descriptive or predictive tool; in [*6*], text mining was successful applied to extract Facebook posts for sentiment classification during the Arab Spring event.

Recent applications of sentiment analysis relevant to our study include crime surveillance; in [*7*], social media monitoring systems were reviewed and five essential elements that every monitoring system should include were suggested, while in [*8*], a framework was proposed to study the reactions, sentiments, and communication of civilians in response to terrorist attacks. The authors of [*9*] proposed two computational methods for estimating social media sentiment and reported better performance in comparison to standard techniques.

Sentiment analysis of social media data has also been applied for tracking disease outbreaks; the authors of [*10*] described a method for extracting tweets for early warning and outbreak detection during a Swine Flu pandemic demonstrating a strong contribution for alerting relevant stake holders for prompt action.

Twitter data allows for geographical and spatial analysis and in [*11*] a framework was developed for visualising public sentiment variations using tweets gathered based on counties and countries in the whole of UK during the event that took place following the birth of prince George in 2013; this application is particularly useful in tracking the upwards or downwards trend of product allergies in a given region over a given period of time.

The authors of [*12*] applied text mining techniques to investigate consumer attitudes towards global brands, they reported that Twitter can be used as a reliable method in analyzing attitudes towards global brands.

Pharmacovigilance is of particular interest to us, it involves the monitoring of adverse effects of pharmaceutical products. As reported in [*13*], Internet users can provide early clues about adverse drug events through their Internet surfing log data. A similar work is found in [*14*], where the visualization of CHFpatients.com forum chat sentiments was used to measure the effectiveness of a drug through quantifying its side effects, particularly for the benefit of the forum members and their physicians. We aim to apply the same approach to drug and cosmetic product users, with the intended beneficiaries been the product users, manufacturers, regulatory and enforcement agencies.

Sentiment analysis can be approached as one or combination of supervised, semi-supervised and unsupervised classification tasks; lexicon and machine learning are the popular approaches; the lexicon based approach as detailed in [*5*], [*15*] and [*16*] uses dictionaries of words annotated with their semantic orientations; the learning based approach also described in [*5*] and [*17*] require creating a model by training a classifier with labeled examples; and finally the utilization of the combination of both approaches as described in [*18*]. A recent review on these techniques was presented in [*17*]; two performance issues discovered with regard lexicon approaches are: 1) how to deal with context dependent words and 2) how to address multiple entities with varying orientations within a single sentence; one of the suggested approach for tackling these issues is the use of holistic lexicon described in [*15*] which involve exploiting external evidences and linguistic conventions of natural language expressions. The machine learning approach is reported to outperform the lexicon approach, yet suffers a general drawback of labeling large training data.

Three machine learning based classifiers, Naive-Bayes, Maximum Entropy and Support Vector Machines and a hybrid technique, label propagation, were investigated in [*17*] and a number of issues that need to be addressed before using these techniques for sentiment classification were outlined. The reported drawback of the Naive-Bayes classifier is the assumption that features are independent of each other; Maximum Entropy suffers from over-fitting in the event of sparse data. It is, however, reported that its performance can be improved by introducing a priori for each feature; Support Vector Machines was reported to outperform the other techniques, its major drawback however is the difficulty in identifying the important words that influenced the classification process due to its black box nature. The review outlined a semi-supervised method called label propagation; the technique improves the accuracy of the classification process by using the Twitter follow graph.

Driven by these motivations, especially the successes of text mining and sentiment analysis on social media data for surveillance [*7*], [*8*], [*9*] and [*10*], and brand reputation monitoring [*12*] applications and also in sentiment classification framework design in [*8*], [*11*] and [*17*], we proposed a novel framework for harnessing and processing the views and experiences of consumers of popular brands of drug and cosmetic products reported as status updates, tweets or comments on social media platforms. We also aim to adapt a combination of these approaches and computational intelligent techniques such as fuzzy sentiment scoring, in the framework for performance comparison with the traditional lexicon and other methods. A detailed description of the proposed framework is provided in the next section.

III. The Proposed Framework

The architecture of the proposed framework for harnessing and processing the views and experiences of customers of popular brands of drug and cosmetic products is presented in Fig. 3.1

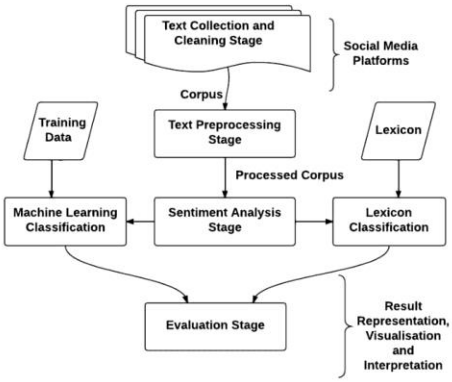


Fig. 3.1. Architecture of the proposed framework

The framework comprises four stages: text collection and cleaning; pre-processing; sentiment analysis and finally evaluation. The following is a brief description of these components:

A. Text Collection and Cleaning Stage

Most organisations and businesses including social media platforms now create Application Programming Interfaces (APIs) to share data. At the text collection and cleaning stage, an API call for authentication and data extraction is invoked on Facebook Graph and Twitter APIs. The framework is designed to accommodate virtually all available social media APIs, but due to the dynamic nature of these APIs and for the purpose of this report, greater focus will be given to Twitter and Facebook. The Twitter API consists of the REpresentational State Transfer (REST) and Streaming APIs [*19*]. The REST API provides methods for authenticating applications, processing requests, handling imposed limits, etc. The Streaming API provides client applications with Twitter's global stream (public, user and site) of data. The Facebook Graph API provides the means of getting data into and out of the Facebook social graph. The framework employs both the REST and Streaming APIs for searching and fetching tweets, while the Facebook Graph API is used for fetching pages, status updates and comments suggesting user experiences and views on drugs and cosmetic products. Fig. 3.2 describes the data collection workflow.

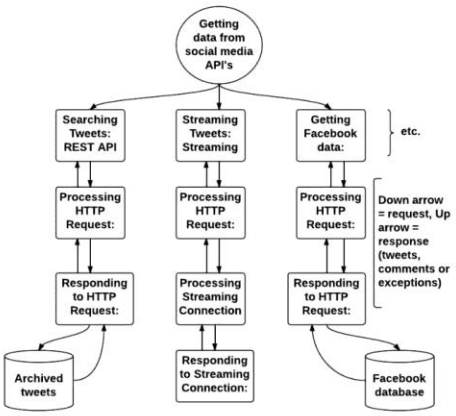


Fig. 3.2. Data Collection workflow

The collected text is noisy and in JavaScript Object Notation (JSON) format and methods for cleaning and parsing of the data to form a corpus, i.e. a collection of comments and tweets, are incorporated for further processing.

B. Pre-processing Stage

At this stage, the corpus is transformed into feature vectors; for our purpose of conducting this work, we adapted the bag of words representation described in [*20*] and [*21*] due to its simplicity and because preserving the order of the features in the corpus is not of particular interest in the application. We adapted a simple feature selection or pre-processing method to transform or tokenize the text stream to words; these methods constitute a sequence of the following tasks; removing delimiters, converting all words to lower case, removing numbers and stop words [*21*], stemming words to their base and some application or domain specific feature transformations. The tokens are then represented as bag-of-words sparse matrix using the term frequency-inverse document frequency ($tf - idf$) weighing scheme described in [*20*] and [*21*].

We define our corpus as C, containing N documents defined as d_i, where $i = 1 \dots N$, and tweets tokenized as words or terms t. The $tf - idf$ weighting scheme takes into account the relative importance of the word in the document and assigns to term t_j a weight in document d_i given by:

$$tf - idf(t_j,\ d_i) = tf(t_j,\ d_i) * idf(t_j) \quad (1)$$

where:

$tf(t_j,\ d_i)$ denotes term frequency, the number of word occurrences in a document;

$idf(t_j) = log2\left(\frac{N}{df t_j}\right)$ denotes inverse document frequency, with $df(t_j)$ representing the number of documents containing the word. Fig 3.3 describes the pre-processing workflow.

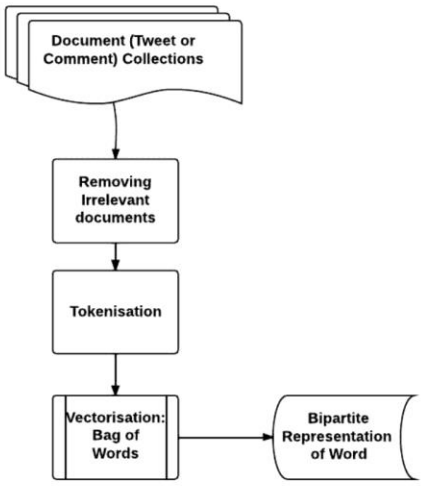


Fig. 3.3. Preprocessing workflow

Further sparseness is handled by selecting terms that appear in a minimum number of documents. The resulting bipartite representation is the input on which further tasks are performed.

C. Sentiment Analysis Stage

This stage handles the polarity measurement, sentiment classification and clusterisation of the entire corpus and for some targeted entities. We approach these tasks by employing both lexicon and learning methods.

1. Lexicon-based sentiment classification

In the lexicon based approach, beside the corpus, a fundamental requirement is a pre-labelled word list or polarity lexicon. For an improved classification result as described in [*22*], the framework merges two lexicons, application or domain specific preassembled lexicon and a generic English based lexicon developed and being maintained by the authors of [*23*]. As described in Fig. 3.4, another requirement for the lexicon based classifier is a sentiment scoring function, of which there are several options, one of the most basic polarity computational schemes is described in [*24*]; all the words in the corpus or target collection are compared to the words in the lexicon, the overall sentiment score of the corpus or a subset will then be the difference between the numbers of positively and negatively assigned words. Therefore, the associated polarity score for each comment or tweet in the corpus is given by:

$$Score = \sum_{i}^{n} pw - \sum_{j}^{m} nw \qquad (2)$$

where pw and nw denote positive and negative words respectively;

A comment or tweet has an overall positive sentiment if Score > 0,

A comment or tweet has an overall neutral sentiment if Score = 0,

A comment or tweet has an overall negative sentiment if Score < 0,

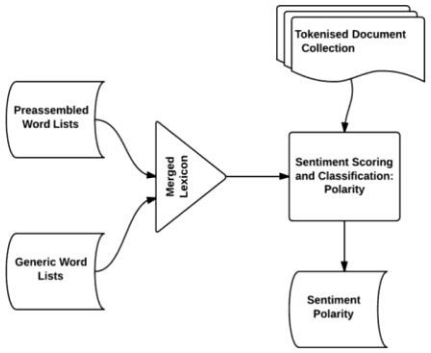


Fig. 3.4. Lexicon based sentiment classification workflow

The total score for the corpus is visualized and evaluated with simple descriptive statistics i.e. histogram and box plot. A more advanced scoring scheme include fuzzy reasoning, which is a computational intelligence technique that can be applied to improve the text classification and clustering tasks; the technique was used in [*25*] to generate intuitive fuzzy numbers for 150 words formed from a feature word list for sentiment classification of hotel management reviews, with higher accuracy and recalling rate.

2. Machine learning-based sentiment classification

In the machine learning based approach, beside the corpus, the fundamental requirement is a training dataset, already coded with sentiment classes. As described in Fig. 3.5, the classifier is trained or modeled with the labeled data such that new but similar documents are tested with the resulting model to have it predict the direction of the sentiment of the new documents. For the purpose of this report, Naive Bayes is used as a baseline classifier because of its efficiency as reported in [*21*]. We assume the feature words are independent and then use each occurrence to classify tweets or comments into its appropriate sentiment class; this is called multinomial event model.

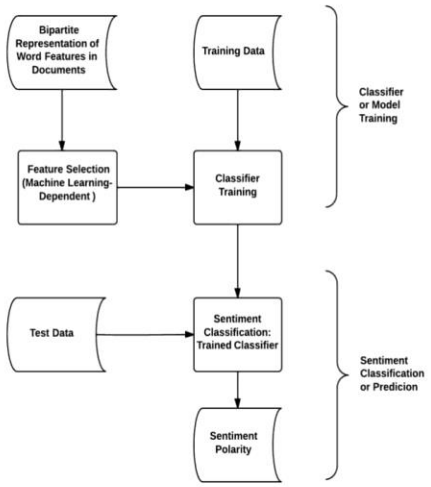


Fig. 3.5. Machine learning sentiment classification workflow

It follows from [*21*] that our classifier which utilises the maximum a posteriori decision rule can be represented as:

$$c_{map} = \underset{c \in C}{\operatorname{argmax}}\big(P(c|d)\big)$$

$$= \underset{c \in C}{\operatorname{argmax}}\left(P(c) \prod_{1 \le k \le n_d} P(t_k|c)\right) \qquad (3)$$

where t_k denotes the words in each tweet or comment and C the set of classes used in the classification; again, $P(c|d)$ is the

conditional probability of class c given document d, $P(c)$ the prior probability of class c and the $P(t_k|c)$ conditional probability of word t_k given class c. To estimate the prior parameters, equation (3) is then reduced to:

$$c_{map} = \underset{c \in C}{\operatorname{argmax}} \left(\log P(c) + \sum_{1 \le k \le n_d} \log P(t_k|c) \right) \quad (4)$$

To handle zero probabilities that may arise when a word does not occur in a particular class, $(tf - idf)$ weighing or Laplace smoothing by adding 1 to each count is employed; with Laplace smoothing, $P(t|c)$ becomes:

$$P(t|c) = \frac{T_{ct}+1}{\sum_{t \in V}(T_{ct\prime}+1)} + \frac{T_{ct}+1}{(\sum_{t \in V} T_{ct\prime})+B\prime} \quad (5)$$

where B' refers to the number of terms contained in the vocabulary V.

D. Evaluation Stage

The lexicon based sentiment classification result is represented as a histogram of polarity measures and the result is evaluated with reference to a ground truth or by human judgment. We use contingency tables or truth table to represent the output of the classifier and a baseline result for performance comparison.

IV. Case Studies

A. Text Mining and Sentiment Analysis of Facebook Comments

To demonstrate the usefulness of the proposed framework, we collected user comments and opinions from Facebook pages of 3 popular brands of drug and cosmetic related products (Avon, Dove and OralB) using the Facebook Graph API. For privacy issues, the brand names will be randomly coded as Brand X, Brand Y and Brand Z. We initially retrieved the public contents of the targeted pages and then extracted user comments and opinions from popular posts suggesting product advertisement that: 1) do not offer a prize in return i.e. Brand Y and Z and 2) offer a prize in return i.e. Brand X. TABLE 4.1 is a summary of the collected comments.

TABLE 4.1. Summary of Collected Comments

Brand Name	Total number of posts retrieved	Total number of comments extracted
Brand X	5000	654
Brand Y	665	1747
Brand Z	5000	957

Over the entire corpus, we apply tokenization, stop word removal and conversion to lower case functions to obtain only individual word features. The features were then represented as bag of word model with the $(tf - idf)$ weighing scheme to generate a sparse matrix, limiting the output to a minimum of three characters word length. The sparse matrix was handled using a function that remove sparse terms which have at least a 99 percentage of sparse elements. We then perform frequent term analysis to generate popular words for creating our preassembled lexicon, we adapted the method used in [22] by using the English dictionary and Thesaurus. The preassembled lexicon is then merged with social media slangs and generic lexicon for sentiment classification. A comparison lexicon sentiment analysis was performed over the 3 different brands. The result is presented in Fig 4.1. It is interesting to see the overall sentiment on all the three brands been positively skewed, with negative:neutral:positive ratios for Brand X, Brand Y and Brand Z approximately 1:42:175, 1:3:3 and 1:5:5 respectively; the high positive sentiment on Brand X confirms comments been prize driven.

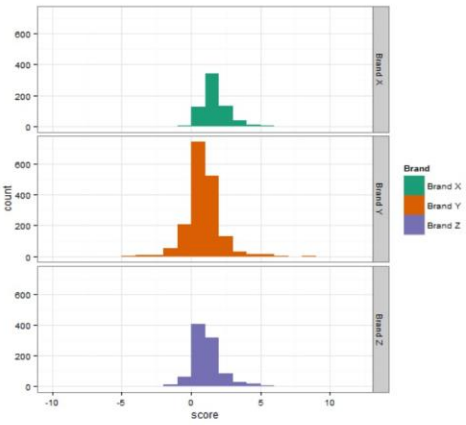


Fig. 4.1. Comparison sentiment analysis for Brand X, Brand Y and Brand Z

The distribution of these scores is presented in TABLE 4.2.

TABLE 4.2. Distribution of Sentiment Scores for the 3 Brands

Brand	Sentiment Scores			
	Negative	Neutral	Positive	Row Total
Brand X	3	127	524	654
Brand Y	282	742	723	1747
Brand Z	85	408	464	957
Column Total	370	1277	1711	3358

We also extracted three separate datasets in Brand Y each with the mention of the three generic products, soap, cream and deodorants and then performed comparison sentiment analysis on the aggregated data; the distribution of these scores is presented in TABLE 4.3. This is another interesting result with Soap been more positively skewed among the three products; the negative:neutral:positive ratio of Cream, Deodorant and Soap approximately been 1:2:2, 1:1:2 and 1:2:5 respectively.

TABLE 4.3. DISTRIBUTION OF SENTIMENT SCORES FOR THE 3 PRODUCTS

Product	Sentiment Scores			
	Negative	Neutral	Positive	Row Total
Cream	5	11	11	27
Deodorant	21	23	46	90
Soap	11	26	56	93
Column Total	37	60	113	210

The result is presented in Fig. 4.2.

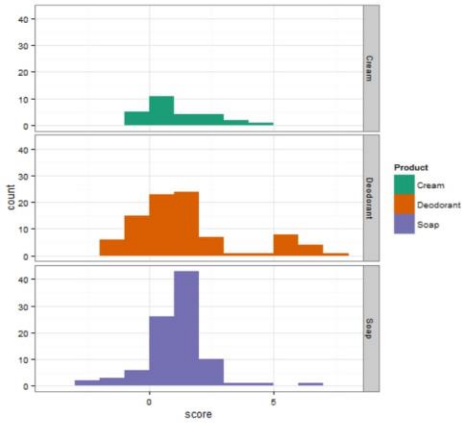


Fig. 4.2. Comparison sentiment analysis for soap, cream and deodorant

We illustrate the same problem of classifying the sentiment orientation of the entire corpus with a naïve Bayes sentiment classifier described in [26], modeled with polarity and emotional lexicon. The method classify_polarity classifies the comments as positive, neutral or negative. The classification results over the entire corpus for both methods are compared in TABLE 4.4.

TABLE 4.4. COMPARISON OF LEXICON AND MACHINE LEARNING SENTIMENT SCORES OVER THE ENTIRE COMMENT CORPUS

Method	Sentiment Scores			
	Negative	Neutral	Positive	Total
Lexicon	539	1436	1383	3358
Naïve Bayes	554	368	2436	3358

The negative scores for both methods agree closely while there is a sharp variation in both the neutral and positive scores. Evaluation of what method provided accurate result depends on several factors and is beyond the scope of this work.

B. Text Mining Sentiment Analysis of Twitter data

To demonstrate how the framework can be applied to model a machine learning classifier for predicting the sentiment of a given text, we collected about 11,431 tweets with the following keywords: medicine; prescription; over the counter; side-effects; online pharmacy; and antibiotics. The tweets were first cleaned then converted to a single corpus, with each tweet represented as a single document. Thus we have 11,431 documents making up the corpus. The corpus is then represented as sparse matrix for frequent term analysis and lexicon generation as described in part A.

We also performed lexicon sentiment analysis over the entire corpus as in part A, such that the generated sentiment scores containing 11,431 tweets classified with corresponding scores as described in TABLE 4.5 will serve as an initial dataset; for simplicity, we categorized scores greater than zero as positives and scores lower than zero as negatives.

TABLE 4.5. RANDOM PREVIEW OF POLARITY SCORES

Score	Tweet
-1	To everybody that knows me I think it's time for me to sign paperwork to take me off my antibiotics which will allow me to die in my home
2	Pretty sure we headed a hospital visit off today, fingers crossed these antibiotics help ASAP.
0	Took my antibiotics without eating and i was almost sick in work
-2	Its ridiculas how many times i get ill, iv either got a cold on antibiotics , my emune system is rubbish

The dataset was split into 75% i.e. 8573 labeled tweets as a training set and 25% i.e. 2858 labeled tweets as test set so that the classifier can be evaluated on data it had not seen previously. We apply the Naive Bayes algorithms as our baseline classifier. TABLE 4.6 is the representation of the result as a confusion matrix.

TABLE 4.6. NAÏVE BAYES CLASSIFIER RESULT

Predicted	Actual		
	Negative	Positive	Row Total
Negative	531 0.639	180 0.089	711
Positive	300 0.361	1847 0.911	2147
Column Total	831 0.291	2027 0.709	2858

Looking at the table, we can see that 300 out of 831 positive messages (36 percent) were incorrectly classified as negative, while 180 of 2027 negative messages (8.9 percent) were incorrectly classified as positive; a total accuracy of about 83%. This performance will be used as a baseline for assessing other classifiers.

CONCLUSION

We have demonstrated how machine learning techniques can be used to infer sentiments over social media data suggesting the views and experiences of drug and cosmetic product users. First a framework is developed for harnessing and tracking these views and experiences using text mining and sentiment analysis, we then conducted two case studies using the framework for comparison of sentiment analysis over three cosmetic brands coded as: Brand X, Brand Y and Brand Z and also over three products: Soap, Cream and Deodorant. A Naïve Bayes classifier was used to obtain a baseline result for assessing other classifiers. This paper reports a work in progress and the initial brand and product comparison results signify the usefulness of text mining and sentiment analysis on social media data while the use of machine learning classifiers for predicting the sentiment orientation provide a useful tool for users, product manufacturers, regulatory and enforcement agencies to monitor brand or product sentiment trends in order to act in the event of sudden or significant rise in negative sentiments.

Future work will consider comment spamming, comparing different machine learning sentiment classification performances, temporal analysis for detecting up or down trend of sentiment of a particular brand or product as well as clustering tweet and user sentiments by location.

ACKNOWLEDGMENT

Special thanks to the Commonwealth Scholarship Commission through the Association of Commonwealth Universities (ACU) for providing studentship to the main author.

REFERENCES

[1] K. Dégardina, Y. Roggoaand and P. Margot, "Understanding and fighting the medicine counterfeit market," Journal of Pharmaceutical and Biomedical Analysis, vol. 87, pp. 167-175, January 2014.

[2] M.J. Paul and M. Dredze, "You are what you Tweet: Analyzing Twitter for Public Health," in 5th International AAAI Conference on Weblogs and Social Media (ICWSM 2011), Barcelona, 2011.

[3] C. Kaiser and F. Bodendorf, "Mining Patient Experiences on Web 2.0 - A Case Study in the Pharmaceutical Industry," in SRII Global Conference (SRII), California, 2012, pp. 139-145.

[4] A. M. Hearst, "Untangling text data mining," in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Maryland, 1999, pp. 3-10.

[5] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, January 2008.

[6] J. Akaichi, Z. Dhouioui, and M.J. Lopez-Huertas Perez, "Text mining facebook status updates for sentiment classification," in System Theory, Control and Computing (ICSTCC), 2013 17th International Conference , Sinaia, 2013, pp. 640-645.

[7] M.D. Sykora, T.W. Jackson, A. O'Brien, and S. Elayan, "National Security and Social Media Monitoring: A Presentation of the EMOTIVE and Related Systems," in 2013 European Intelligence and Security Informatics Conference (EISIC), Uppsala, 2013, pp. 172-175.

[8] M. Marc and C. Lee. Vincent, "A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter," Information Systems Frontiers, vol. 13, no. 1, pp. 45-59, March 2011.

[9] K. Glass and R. Colbaugh, "Estimating the sentiment of social media content for security informatics applications," in IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, 2011, pp. 65-70.

[10] E. de Quincey and P. Kostkova, "Early Warning and Outbreak Detection Using Social Networking Websites: The Potential of Twitter," in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Istanbul, Turkey: Springer Berlin Heidelberg, 2010, ch. 3, pp. 21-24.

[11] V. D. Nguyen, B. Varghese, and A. Barker, "The royal birth of 2013: Analysing and visualising public sentiment in the UK using Twitter," in IEEE International Conference on Big Data, California, 2013, pp. 46-54.

[12] M. Mostafa Mohamed, "More than words: Social networks' text mining for consumer brand sentiments," Expert Systems with Applications, vol. 40, no. 10, pp. 4241-4251, August 2013.

[13] R. W. White, N. P. Tatonetti, N. H. Shah, R. B. Altman, and E. Horvitz, "Web-scale pharmacovigilance: listening to signals from the crowd," J Am Med Inform Assoc, March 2013.

[14] B. Chee, K.G. Karahalios, and B. Schatz, "Social Visualization of Health Messages," in 42nd Hawaii International Conference on System Sciences, HICSS '09. , Big Island, 2009, pp. 1-10.

[15] X. Ding, B. Liu and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining, New York, 2008, pp. 231-240.

[16] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," Journal of Computational Linguistics, vol. 37, no. 2, pp. 267-307, June 2011.

[17] S. Bhuta, A. Doshi, U. Doshi, and M. Narvekar, "A review of techniques for sentiment analysis Of Twitter data," in International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, 2014, pp. 583-591.

[18] F. Balage, P. Pedro and T. A. S. Pardo, "NILC_USP: A Hybrid System for Sentiment Analysis in Twitter Messages," in Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Georgia, 2013, pp. 568--572.

[19] M. Kevin Makice, Twitter API: Up and Running: O'Reilly Media, 2009.

[20] M. Shafiei, S. Wang, R. Zhang, E. Milios, B. Tang, J. Tougas, R. Spiteri, "Document Representation and Dimension Reduction for Text Clustering," in IEEE 23rd International Conference on Data Engineering Workshop, Istanbul, 2007, pp. 770-779.

[21] C. D. Manning, P. Raghavan, and H. Schtze, Introduction to Information Retrieval: Cambridge University Press, 2008.

[22] R. Heimann and N. Danneman, Social Media Mining with R.: PACKT Publishing, 2014.

[23] M. Hu and B. Liu, "Mining and summarizing customer reviews," in KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, 2004, pp. 168-177.

[24] S. Gaston. (2014, June) Mining Twitter with R. [Online]. https://sites.google.com/site/miningtwitter/home

[25] X. Feng Li and D. Li, "Sentiment Orientation Classification of Webpage Online Commentary Based on Intuitionistic Fuzzy Reasoning," Applied Mechanics and Materials, vol. 347 - 350, pp. 2369-2374, August 2013.

[26] P. Jurka Timothy, Tools for Sentiment Analysis, 2012.

Appendix IV: ASONAM2015

Bipartite Network Model for Inferring Hidden Ties in Crime Data

Haruna Isah, Daniel Neagu, Paul Trundle
Artificial Intelligence Research Group
Department of Computing, University of Bradford
Bradford, UK
(H.Isah, D.Neagu, P.R.Trundle)@bradford.ac.uk

***Abstract*— Certain crimes are hardly committed by individuals but carefully organised by group of associates and affiliates loosely connected to each other with a single or small group of individuals coordinating the overall actions. A common starting point in understanding the structural organisation of criminal groups is to identify the criminals and their associates. Situations arise in many criminal datasets where there is no direct connection among the criminals. In this paper, we investigate ties and community structure in crime data in order to understand the operations of both traditional and cyber criminals, as well as to predict the existence of organised criminal networks. Our contributions are twofold: we propose a bipartite network model for inferring hidden ties between actors who initiated an illegal interaction and objects affected by the interaction, we then validate the method in two case studies on pharmaceutical crime and underground forum data using standard network algorithms for structural and community analysis. The vertex level metrics and community analysis results obtained indicate the significance of our work in understanding the operations and structure of organised criminal networks which were not immediately obvious in the data. Identifying these groups and mapping their relationship to one another is essential in making more effective disruption strategies in the future.**

Keywords—network analysis, bipartite network; organised criminal network; underground forum

I. INTRODUCTION

The Internet and related technologies lend themselves perfectly to crime coordination across dispersed areas [1]. The least common denominator of organised crime is human relationships, social networking is inevitable among criminal groups responsible for the provision of illicit goods and services [2]. Despite the ongoing efforts by governments, law enforcement agencies, academic researchers, and the security sector, little is yet known about the preferred structures, longevity, and how trust is assured among criminal groups [1]. Available empirical data suggest that conventional and cyber criminals are more likely to be involved in loosely associated illicit networks rather than formal organisations [1] [3]. Organised criminal groups often involve multiple offenders connected through various relationships [4]. These relationship can be represented as a network where the nodes are the criminals while the edges are the criminal interactions. Social network analysis, defined as a theoretical and methodological paradigm for sophisticated examination of complex social structures in [2], has a long history of application to evidence mapping in both fraud and criminal conspiracy cases [5], it is useful in understanding the patterns of relationships among criminal groups and in identifying key members in the group [6]. Node centrality and network density measures in social network analysis are useful in identifying pivotal nodes and potential fraud hotspots, sub-structures, structural holes and clustering coefficient measures are used for network classification and path prediction [7]. Link analysis allows for mixing of different node and edge types in the same network and is useful in generating investigative leads and for uncovering missing information that may be hidden in a criminal network [4]. Groups, also called communities or clusters in a network, can be considered as fairly independent compartments with high concentrations of edges within groups of vertices and low concentrations between these groups in the network [8]. Group detection is a useful method for understanding the structure and organisation of criminals in a network.

Criminal intelligence process relies on the ability to obtain and use data. Three main sources of data identified in [9] include: open data such as newsletters, closed data in form of structured databases and classified data often collected through covert means. A common starting point in understanding criminal groups is to identify the criminal's associates i.e. identifying relationships between individuals and their roles in the criminal activities [9]. These relationship are usually obtained from email and phone communication logs [10], underground forums [2] [11] [12] [13], scraped using set of seeds or leaked data [6] [14], money trails [15] [16], crime records [17], by extracting and associating entities in the grey literature [5] [18] or by combinations of these sources.

Situations may arise in a criminal dataset where there is no direct connection among the criminals, such is the case in [18] which involves the extraction of organisational structure of covert network from textual data obtained from public news, the archive of Evolution, an online black market in [13] that recently disappeared comprising of list of underground vendors with their associated products, and the crime report in [19] comprising of list of rogue manufacturers with their associated products. In order to address the issue of lack of direct connection among criminals and the need to understand their organisational structure, we propose the following research questions: (i) can we infer relationship based on common attributes and other metadata among entities involved in crime but not directly connected? (ii) is there an individual or groups directing the overall operations in a criminal network? We address these problems by (i) modelling a bipartite network in

order to infer relationships between actors and resources involved in crime and (ii) analysing nodes and community structures of the resulting network.

This work is organised as follows: Section II provides a concise review of related work; Section III describes the research methodology; Section IV provides two case studies for evaluating our model and the final section describes conclusions and future work.

II. Related Work

Criminal network data may contain variety of entities such as persons, organisations, locations, URL's, vehicles, weapons, properties, bank accounts, etc. Learning associations between these entities is a critical part of uncovering criminal activities and fighting crimes [4]. Criminal groups show various levels of organisational structure. This organisation according to [1], depends on whether: (i) their activity is purely aimed at online targets such as swarms consisting of ephemeral clusters of individuals with no leadership as in the case of the Anonymous or hubs which are organised with a clear command structure and a focal point of core criminals around which peripheral associates gather as in the case of LulzSec, (ii) their activity uses online tools to enable conventional crimes such as clustered hybrids which are articulated around a small group of individuals and focused around specific activities or methods as in the case of carding networks or extended hybrids which are less centralised consisting of many associates and subgroups, and (iii) they combine online and offline targets such as hierarchies or aggregates according to their degree of cohesion and organisation. Social network analysis as a tool for understanding organised criminal groups involve the detection of structural changes in social networks with node and group level measures. The node level metrics include: degree and centrality measures while the group level metrics include: density, cohesion, group stability, etc. Group detection tasks in criminal network analysis involve detection of meaningful clusters of criminal actors, their interaction, the interaction of their subgroups and cliques in criminal data. Increased work in network analysis of criminal groups is reported recently in [2] [4] [5] [6] [10] [11] [12] [17] [18] [20] [21].

Criminals utilise underground forums in form of chatrooms and private messaging services to exchange information on abusive tactics and engage in the sale of illegal goods and services [2] [11] [12] [13]. Anonymised carding forums' private messaging records were modelled as graphs in [2] with the aim of uncovering the underlying structural and behavioural properties of cybercriminals, measures investigated include: degree distribution, assortativity, rich club phenomenon, transitivity and small world phenomena, connectivity and cohesive subgroups. Another analysis of underground forums aimed at understanding the social dynamics of six underground forums and how they impact e-crime market efficiencies was carried out in [11]. A recent study of underground forum interaction in [12] which uses six different centrality measures to produce a visual representation of cybercrime forum reports that criminal groups are organised into two distinct communities that resemble (i) gangs, which are limited in size with one central leader who makes all the decisions for the group and (ii) mobs, with hundreds or thousands of members that share relatively equal centrality rankings divided into multiple sub-groups. Vulnerability in organised crime groups that can be exploited by law enforcement agencies include how group members earn trust among pears and the way they get their money or the e-currency they use [12]. Market basket analysis of products traded in Evolution, an online black market operating on Tor network that recently became extinct was carried out in [13]. Results obtained from these studies would allow authorities to better utilise their resources and devise more effective disruption strategies in the future.

In [4], a link analysis technique that employs shortest path algorithms and priority first search to identify the strongest associations between entities in a criminal network was proposed and evaluated using Phoenix Police Department crime reports. The network of the 19 hijackers surrounding the tragic events of September 11th, 2001 were mapped through public data in [5], the result obtained revealed that the hijackers did not work alone but had accomplices who were on the planes, yet they were the conduits for money, skills and knowledge needed to execute the operation. Meta matrix model of concepts extracted from public news was used in [18] to detect and analyse the structure of a covert network.

Analysis of the community structure of Nigerian scammers were carried out in [6] and [10]. The study in [6] shows that these scammers are organised into tightly and loosely connected groups while the findings in [10] revealed that only ten groups are responsible for about 50% of the scam attempts we receive. In [17], n-clique and k-core algorithms compared fairly well with other propriety criminal group detection models. An analysis of cybercriminal ecosystem on Twitter in [20] reports that while network criminal hubs are more inclined to follow criminal accounts, the criminal accounts tend to be socially connected, forming a small-world. Network analysis of South African arms deal and corruption carried out in [21] reveals that a single actor can have many separate relationships through which resources (information, money, influence, etc.) are shared with other actors. If this actor were to be removed, a large part of the network would be disconnected from the rest and significant resources would be unable to reach large parts of the network.

Procedures for implementing social network analysis for organised crime prevention were described in [6] [7] [9]. The central theme in [6] is focused on constructing large scale social graph from a smaller set of leaked data and linking of the leaked data (set of email addresses) to Facebook profiles to scrape large scale social graphs. The main focus in [7] is centred around fraud analytics and the laid down procedures include: (i) building the network, (ii) graph sampling to select set of flagged nodes, (iii) exploring, observing and measuring fundamental network metrics, and (iv) applying mitigation measures based on the inference from the measured metrics. The main focus in [9] is operational, the procedure include understanding client needs, obtaining and use of relevant data, data quality evaluation, data collation strategy, data integration

and analysis, knowledge dissemination, and finally, re-evaluation.

Although our approach also utilises social network measures for understanding criminal network grouping, its novelty can be traced to the use of bipartite network model, a special type of graph representation where vertices are divided into two sets A and B, and only connections between two vertices in different sets are allowed [22]. The bipartite network representation naturally suits criminal datasets that may lack direct connection between criminals and allows for inferring of hidden ties among each of the two sets A and B as we can see in the next sections.

III. Methodology

The research tasks are categorised as follows: (i) criminal network extraction, (ii) network representation, (iii) measuring network metrics, and (iv) analysis of group dynamics.

A. Criminal Network Extraction

One of the main challenges of data extraction for network analysis lies in the choice of vertices, relationships, and attributes that can best answer the targeted research questions. Candidate choices for each of the three major network elements are: vertices (individuals, groups, organisations, bank accounts, products, URL's, resources, affiliates, Internet infrastructures), edges (friendship, ownership, distributor/advertiser and can be binary, weighted, directed, undirected, multipartite or multiplex relationship) and attributes (location, time etc.).

Network elements extraction is straight forward in structured data such as police records where vertices have been structured in tables and the edges are derived either as binary or weighted relationship between single or multimode vertices. In a semi or unstructured data such as text documents, network constitute the union of all statements per text document, vertices are concepts or ideational kernels represented by one or more words, while edges are the links between two or more concepts [18]. In the latter case, data collection is more of an approximation via text or natural language network analysis.

B. Network Representation: The Bipartite Network Model

Let $A = \{a_1, a_2, \dots, a_i\}$, represent the vertex sets of actors such as individuals, group of individuals or organisations capable of initiating an action over certain resources and $B = \{b_1, b_2, \dots, b_i\}$, represent the vertex sets of resources such as products, then an actor is uniquely connected to resources and no connection exist between actors and actors or resources and resources. The set of edges or relationships $E \subseteq A\ X\ B$, are the initiated actions while the edge weights represent the total co-occurrences of similar instance of connections. The pairs a_x, b_y denote an actor a_x, who is associated to a resources b_y. The sets A, B, and E can be represented as a bipartite graph $G = (A \mid B, E)$, where A and B are called the partite sets of the graph vertices that are connected by an edge iff $(a_x, b_y) \in E$, with $1 \leq x \leq i$ and $1 \leq y \leq j$ where i is the number of unique actors and j is the number of unique resources in the network. The cardinality or the number of edges in the bipartite graph is represented by $n = |E|$. The pairs a_x, b_y can also be represented as b_y, a_x denoting a resources b_y associated by an actor a_x, hence the actor-resources network can be represented by a weighted undirected bipartite network $\boldsymbol{M_{ixj}}$. In order to infer ties among actors in the network, we transform the bipartite network $\boldsymbol{M_{ixj}}$ to its unipartite components $\boldsymbol{A}$ and $\boldsymbol{B}$ or actor-actor and resource-resource network respectively. The unipartite networks are obtained in a process called bipartite projection described in [22]. The bipartite to unipartite transformation process is illustrated below. Fig.1. is an actor-resource bipartite network representation where the two sets of vertices are differentiated by red (resources) and green (actors) colours. The unipartite components are obtained by selecting one of the sets of vertices and linking two vertices from that set if they were connected to the same vertices of the other sets.

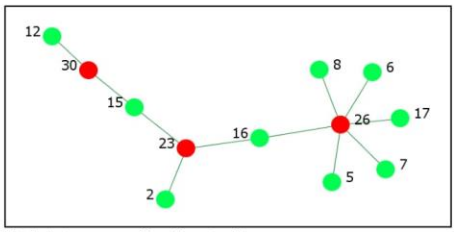


Fig.1. Actor-resource bipartite network

The A-projection of our actor-resource bipartite network $G = (A \mid B, E)$, shown in Fig.2. is the actor-actor network $G_A = (A, E_A)$ in which two vertices of A are linked together if they have at least one neighbour in common.

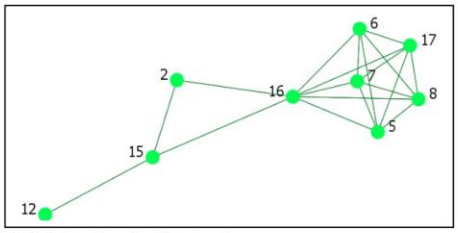


Fig.2. Actor-actor unipartite network

The B-projection of our actor-resource bipartite network $G = (A \mid B, E)$, shown in Fig.3. is the resources-resources network $G_B = (B, E_B)$ in which two vertices of B are linked together if they have at least one neighbour in common.

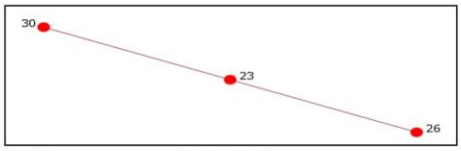


Fig.3. Resource-resource unipartite network

C. Network Metrics

When trying to understand networks, we often want to identify important vertices, locate subgroups, or get a sense of how interconnected a network is compared to other networks. Vertex and edge specific measures include: degree, degree centrality, closeness centrality, betweenness centrality, eigenvalue centrality, PageRank and local clustering coefficient. Measures that can be used to describe the structure of the entire network include: density, degree distribution, connectivity and centralisation.

If the actor-actor network matrix $\boldsymbol{A}$ is defined by:

$$a_{ij} = \begin{cases} 1 \; if \; an \; edge \; \text{exists} \; from \; vertex \; \text{j} \; to \; vertex \; \text{i} \\ 0 \; \text{otherwise} \end{cases},$$

then it follows from [6] that the degree of a vertex in the network defined as the number of edges connected to the vertex or the cardinality of the vertex neighborhood is given by:

$$d_i = \sum_j a_{ij} \tag{1}$$

the degree centrality is defined as:

$$D_i = \frac{d_i}{N-1} = \frac{\sum_j a_{ij}}{N-1} \tag{2}$$

where d_i is the degree of the vertices and $N-1$ is a normalization factor (N is the number of vertices in the network) and $0 \leq D_i \leq 1$. Closeness centrality of a vertex, defined as the distance of a vertex from other vertices or the sum of shortest paths between a vertex and all other vertices in a network is given by:

$$C_i = (L_i)^{-1} = \frac{N-1}{\sum_j d_{ij}} \tag{3}$$

where d_{ij} is the distance between vertices i and j. L_i is the normalized distance of a vertex from other vertices in a network. Betweenness centrality of a vertex, defined as the number of shortest paths in a network which passes through that vertex is given by:

$$B_i = \frac{\sum_{j<k} \frac{n_{jk}(i)}{n_{jk}}}{(N-1)(N-2)} \tag{4}$$

where n_{jk} is the number of shortest paths between j and k and $n_{jk}\ (i)$ the number of such paths which pass through vertex i. $(N-1)(N-2)$ is the normalization factor. Eigenvector centrality of a vertex determines to what extent a vertex is connected to other well-connected vertices and is given by:

$$x_i = \frac{1}{\lambda} \sum_{j \in M(i)} x_j = \frac{1}{\lambda} \sum_j a_{ij} x_j \tag{5}$$

where $M(i)$ is the set of neighbors of i and λ is a constant. Clustering coefficient of a vertex is the probability that any two randomly chosen neighbours of that vertex in a network are connected themselves, hence a measure of the density of a 1.5-degree egocentric network. Density of a network, defined as a measure of how many edges are in a given set compared to the maximum possible number of edges in the network is:

$$density = 2 * \frac{|E|}{\left(|V| * (|V|-1)\right)} \tag{6}$$

By counting how many vertices have each degree, a degree distribution is formed. Degree distribution deg(d) is defined as the fraction of vertices in a graph with degree d. Connectivity, also known as cohesion, is a count of the minimum number of vertices that would have to be removed before a network becomes disconnected. Centralisation uses the distribution of a centrality measure to understand the network as a whole. Once a network has been constructed and measurements have been calculated, the resulting dataset can be used for many applications.

D. Community Structure

A useful way to understand a large network is to analyse some sections or subgraphs of the network referred to as egocentric networks. Subgraph allow us to identify common social roles and structures. Community detection in networks is a typical clustering problem [8] [23] and is aimed at identifying modules by using the information encoded in the network topology. Communities are assumed as groups of vertices that are similar to each other. This assumption allows for computing the similarity between each pair of the vertices with respect to some local or global reference property such that each vertex ends up in a cluster whose vertices are most similar to it, irrespective of whether they are connected by an edge or not [8]. It follows from [23], that in a network whose vertices can be assigned positions and embedded in an n dimensional Euclidean space, the similarity or dissimilarity between the vertices can be computed using any norm L_m such as: the Euclidean distance (L_2-norm), Manhattan distance (L_1-norm), or the L_∞-norm. The concept of structural equivalence where similarity is inferred from the adjacency relationships between vertices is used for networks that cannot be embedded in space. Another important class of measures of vertex similarity is based on properties of random walks on graphs [8]. Such measure is the basis of the Walktrap algorithms in [23]. Another property of particular interest is whether or not all vertices in a group are connected to one another, when this happens, it is called a clique. A clique requires that all objects of a subgraph are connected to each other. A k-clique is a complete subset of size k of a graph [6].

IV. Experimental Work

We conducted two case studies: one on traditional counterfeiting crime (rogue manufacturer-manufacturer network), and the other on cybercrime (Darknet vendor-vendor

network) data, in order to evaluate how we can infer ties among criminals using bipartite network modelling.

A. Rogue Manufacturer-Manufacturer Network

Pharmaceutical crime involves the manufacture, trade and distribution of fake, stolen or illicit medicines and medical devices, it also constitute the counterfeiting and falsification of medical products, their packaging and the associated documentation as well as theft, fraud, illicit diversion, smuggling, trafficking, illegal trade of medical products and the money laundering associated with it [3].

Pharmaceutical crime data may be composed of a variety of entities such as: people, organisations, brands, locations, storefronts, websites, bank accounts, and product delivery agencies [4]. These entities may form networks composed of: (i) thousands of storefronts in various locations (ii) affiliate websites run by associates', (iii) many botnet spamming partners who are paid to advertise illicit online pharmacy networks, (iv) covert systems for processing online orders, and (v) regular mail or courier services distributors, thereby making it difficult to track at the same time allowing the key actors to evade detection for long periods of time [3] [4]. Once these criminal groups are identified and their habits known, law enforcement authorities may begin to assess current trends in crime in order to forecast and hamper the development of perceived future criminal activities [9].

Network analysis of archived pharmaceutical crime data can be useful in modelling indirect relationships among important entities involved in pharmaceutical product counterfeiting. These entities can be criminals (manufacturers, advertisers, and distributors), the products they sell, banks that process their credit and debit card transactions or the delivery services used by these criminals. The method can also be used to reveal relationships between user accounts sending pharmaceutical spam and the spam URL's. The case study tasks include: (i) data extraction, (ii) network representation, (iii) vertices and network level analysis and (iv) group analysis.

1) Rogue Data Extraction

Using the year of sampling criteria, we extracted all the data from the Medicines Quality Database (MQDB), a public and freely accessible online tool that tracks medicines tested for quality in selected countries in Africa, Latin America and south-eastern Asia [19]. Currently, the database contains about 13,319 instances of medicines collected and tested from 12 countries. We extracted subset of the records with confirmed counterfeiting incidents. We then filtered duplicate data and removed all rows containing Missing, Unknown and N/A records in the Manufacturers column. We considered the following variables most relevant for the task at hand: Year, Manufacturer, Product, Country, Province, Dosage, Date of Sample Collection, and Test Result.

2) Rogue Network Representation

We constructed an undirected, weighted bipartite network, shown in Fig.4. between manufacturers (red vertices) with fake incidences and their associated products (green vertices) and called it rogue manufacturer-product network. The edge weight represent the co-occurrence frequency of manufacturer-product instances. The assumption we made here is that all manufacturers with atleast one product counterfeiting are rogues.

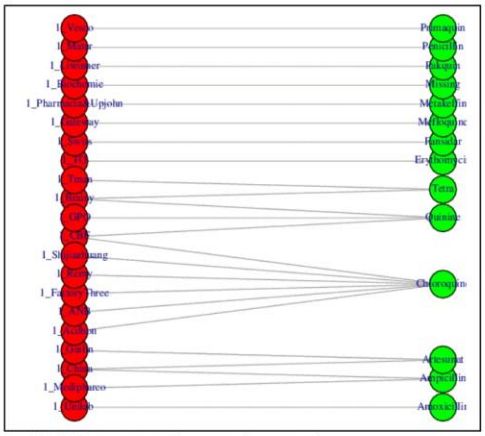


Fig.4. The rogue manufacturer-product network

The unipartite manufacturer-manufacturer network that resulted from projecting the network in Fig.4. is shown in Fig.5.

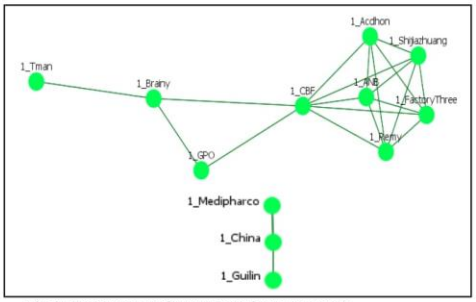


Fig.5. The rogue manufacturer-manufacturer network

3) Rogue Manufacturer-Manufacturer Network Analysis

We first report the aggregate metrics of the largest connected component of the network in Fig.5. These include: number of unique vertices = 9, number of unique edges = 19, geodesic distance (diameter) = 3, average geodesic distance = 1.4321, and network density = 0.5278. The vertex-specific network metrics for the larger component obtained by applying equation (1), equation (3), equation (4), and equation (5) for degree, betweenness, closeness and eigenvector centralities respectively are presented in TABLE I.

From the results in TABLE I, the important vertex in the rogue network is 1_CBF, it has the highest degree and centralities. A vertex with the most neighbours (degrees) can be said to be a key member with influence in its local neighborhood.

TABLE I. ROGUE NETWORK VERTEX-SPECIFIC METRICS

Vertices	*Degree*	*Betweenness*	*Closeness*	*Eigenvector*
1_FactoryThree	5	0	0.083	0.149
1_ANB	5	0	0.083	0.149
1_Shijiazhuang	5	0	0.083	0.149
1_Remy	5	0	0.083	0.149
1_CBF	7	15	0.111	0.163
1_Acdhon	5	0	0.083	0.149
1_GPO	2	0	0.071	0.040
1_Brainy	3	7	0.077	0.041
1_Tman	1	0	0.050	0.008

4) Rogue Manufacturer-Manufacturer Network Group Analysis

We extracted 1.5 degrees egocentric networks of each vertex and reported subgraphs with more than three edges in Fig. 6. The star topology of egocentric network of the vertex 1_CBF indicates its relative importance as a switch or hub in the rogue network.

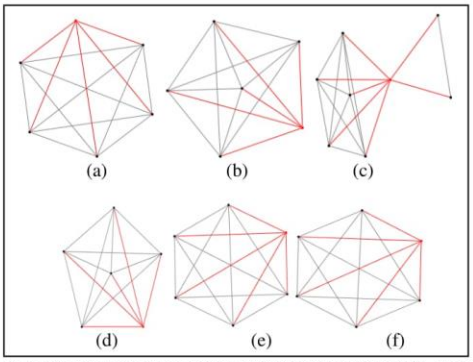


Fig.6. Subgraphs of the rogue network for the following vertices: (a) 1_Acdhon (b) 1_ANB (c) 1_CBF (d) 1_FactoryThree (e) 1_Remy (f) 1_Shijiazhuang

We applied four different community detection algorithms: Girvan Newman, Clauset Newman Moore, Wakita Tsurumi and Walktrap described in [8] in order to study the natural clusters in the rogue network. TABLE II is the summary of the community detection results for communities with minimum of three vertices. The result for the Walktrap method is presented in Fig.7.

When working with massive crime data with location and time attributes, these grouping might signal an element of organisation among the criminals. These naturally occurring clusters are based on patterns of social ties rather than formal group memberships. Vertices with a central position in their clusters, i. e. sharing a large number of edges with the other group partners, may have an important function of control and stability within the group while vertices lying at the boundaries between modules may play an important role of mediation and lead the relationships and exchanges between different communities.

TABLE II. ROGUE NETWORK COMMUNITIES

Vertices in each cluster	*Algorithms*		
	Girvan Newman	*Clauset Newman Moore*	*Wakita Tsurumi*
Cluster 1	1_FactoryThree 1_ANB 1_Shijiazhuang 1_Remy 1_CBF 1_Acdhon	1_FactoryThree 1_ANB 1_Shijiazhuang 1_Remy 1_Acdhon	1_FactoryThree 1_ANB 1_Shijiazhuang 1_Remy 1_Acdhon
Cluster 2	1_GPO 1_Brainy 1_Tman	1_GPO 1_Brainy 1_Tman 1_CBF	1_GPO 1_Brainy 1_Tman 1_CBF
Cluster 3	1_China 1_Guilin 1_Medipharco	1_China 1_Guilin 1_Medipharco	1_China 1_Guilin 1_Medipharco

It is interesting to note that most of these incidents were recorded in one country.

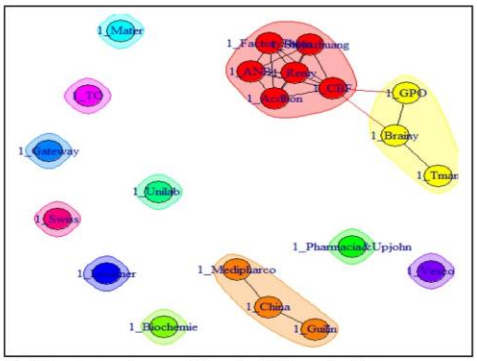


Fig.7. Rogue network communities using Walktrap algorithm

We further subject the rogue network to a more strict community detection methods so as to detect cliques. TABLE III is the clique community detection result.

TABLE III. NETWORK CLIQUE COMMUNITIES

Communities	*Rogue Cliques*
3-clique community	21
4-clique community	15
5-clique community	6
6-clique community	1

There were a total of 43 clique communities starting from a 3-clique community in the rogue network. The size of the largest clique community is 6 with a single clique. Clique communities

in the rogue manufacturer-manufacturer network may lead to the detection of organised criminal network.

B. Darknet Vendor-Vendor Network

The Internet is now a catalyst for illicit online pharmacies, marketing falsified medicines to the public. There are several online market places on the hidden part of the internet (Darknets) offering prescription medicines together with cannabis for sale to the public, these market places such as Silk Road, Agora and Evolution provide online access anonymity via anonymising software such as P2P or Tor and payment anonymity via crypto-currencies such as Bitcoin to criminals with shipments being sent across the world between source, transit and destination countries [24]. Large-scale abusive advertising is a profit-driven endeavor, abuse-advertised goods and services such as spam-advertised Viagra, search-advertised counterfeit software and malware-advertised fake anti-virus have been dominated by an affiliate business model comprised of independent advertisers acting as free agents acquiring traffic via spam or search, and in turn paid on a commission basis by their sponsors who handle the back end, customer service and payment processing [15] [16]. Counterfeit pharmaceutical affiliate business models such as GlavMed, SpamIt and RX-Promotion involve a range of sponsors providing drugstore storefronts, drug fulfillment, shipping, payment processing, customer service and independent advertisers or affiliates who are paid a commission for promoting the program using botnets to send spam email or manipulating search engine results [15].

TABLE IV. VENDOR-PRODUCT LISTING ON EVOLUTION DARKNET DATA

Vendor	***Products***
MrHolland	Cocaine, Cannabis, Stimulants, Hash
Packstation24	Accounts, Benzos, IDs & Passports, SIM Cards, Fraud
Spinifex	Benzos, Cannabis, Cocaine, Stimulants, Prescription, Sildenafil Citrate
OzVendor	Software, Erotica, Dumps, E-Books, Fraud
OzzyDealsDirect	Cannabis, Seeds, MDMA, Weed
TatyThai	Accounts, Documents & Data, IDs & Passports, PayPal, CC & CVV
PEA_King	Mescaline, Stimulants, Meth, Psychedelics
PROAMFETAMINE	MDMA, Speed, Stimulants, Ecstasy, Pills
ParrotFish	Weight Loss, Stimulants, Prescription, Ecstasy
Pharmacy_U	Analgesics, Drugs

1) Darknet Data Extraction

We were opportune to have access to the archive of Evolution, an online black market operating on Tor network few weeks after it disappeared on March 18, 2015 via [13]. The extracted data consists of what product each vendor sells on different dates and contains few top-level categories such as Drugs, Digital Goods, Fraud Related, etc. which are subdivided into product-specific pages and each page contains several listings by various vendors as illustrated in TABLE IV. We extracted all records for Analgesics category and obtained 5002 records with the variables "Vendor", "Products", and "Date".

2) Darknet Data Representation

We constructed an undirected, weighted bipartite network, shown between vendors and their associated products and called it Darknet vendor-product network. The edge weight represent the co-occurrence frequency of vendor-product instances. We then transformed the network to its unipartite components, the vendor-vendor network obtained consists of 102 vertices, 952 edges and 4 connected components as shown in Fig.7.

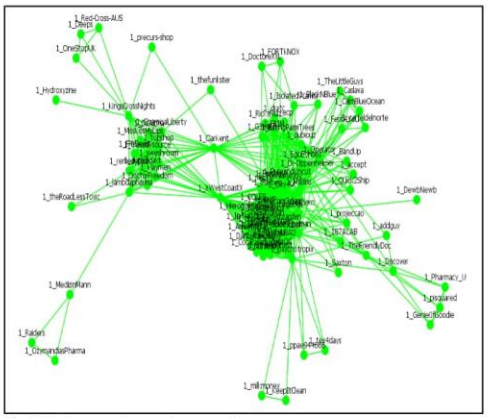


Fig.7. Darknet vendor-vendor network

3) Darknet Vendor-Vendor Network Analysis

We first report the aggregate metrics of the largest connected component of the network in Fig.7. These include: number of unique vertices = 95, number of unique edges = 947, geodesic distance (diameter) = 6, average geodesic distance = 2.1682, and network density = 0.2121. The important vendor in the Darknet network is 1_TheNurseJoy, it has the highest degree and centralities.

4) Darknet Vendor-Vendor Network Group Analysis

We obtained four clusters each of vertices 35, 26, 24, and 10 respectively when we applied Clauset Newman Moore community detection algorithms in [8]. There were however, 7 and 24 communities found with Wakita Tsurumi and Girvan Newman algorithms respectively. The size of the largest clique community is 28 with a single clique.

V. CONCLUSION

Many types of crimes require a high degree of organisation and specialisation. Criminal networks show various levels of organisational structures, they may operate as: swarms, hubs, hierarchies, aggregates, or hybrids according to their activity, degree of cohesion and organisation. The organisation of crime may also occur in the darknet where individuals interact within online discussion forums and chat rooms. Identifying criminal groups within a network and mapping their relationship to one another can be essential to making intelligent strategic decisions. Social network analysis is being used to automatically identify criminal clusters that may not have been obvious in a crime dataset. These naturally

occurring clusters are based on patterns of social ties rather than formal group memberships. Identifying these clusters and their boundaries allows for a classification of vertices according to their structural position in the groups. Vertices with a central position in their clusters may have an important function of control within the group while vertices lying at the boundaries between clusters may play an important role between different communities. Situations often arise in a criminal dataset where we lack direct connection among criminals. In this work, we model such data as a bipartite network in order to infer relationships between actors and resources based on their common attributes. We evaluated the model using two case studies and the results were very significant and can reveal some hidden ties among criminals that were not immediately obvious in the data.

We plan to undergo further evaluation of the model in a large scale case study and in collaboration with law enforcement agents. Weighted projection of the bipartite graph will be considered in the future.

ACKNOWLEDGMENT

Special thanks to the Commonwealth Scholarship Commission through the Association of Commonwealth Universities (ACU) for providing studentship to Haruna Isah.

REFERENCES

[1] B. Roderic, G. Peter, A. Mamoun and C. Steve, "Organizations and Cyber crime: An Analysis of the Nature of Groups engaged in Cyber Crime," *International Journal of Cyber Criminology* , vol. 8, no. 1, p. 1-20, 2014.

[2] M. Yip, N. Shadbolt and C. Webber, "Structural Analysis of Online Criminal Social Networks," in *IEEE International Conference on Intelligence and Security Informatics (ISI)*, Arlington, 2012.

[3] INTERPOL, "An analysis of the involvement of organized criminal groups in pharmaceutical crime since 2008," 2014.

[4] J. X. Jennifer and C. Hsinchun, "Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks," *Decision Support Systems,* p. 473 – 487, 2004.

[5] K. Valdis, "Mapping Networks of Terrorist Cells," *CONNECTIONS,* vol. 24, no. 3, pp. 43-52, 2002.

[6] S. Hamed, A. Ehab, M. Alex and M. Damon, "Constructing and Analyzing Criminal Networks," in *IEEE Security and Privacy Workshops*, San Jose, 2014.

[7] 2. C. G. Inc, "Implementing social network analysis for fraud prevention," 2011.

[8] F. Santo, "Community detection in graphs," *Physics Reports,* vol. 486, no. 3–5, pp. 75-174, 2010.

[9] UNODC, "Criminal Intelligence Manual for Analysts," United Nations Office , Vienna, 2011.

[10] P. Youngsam, J. Jackie, M. Damon, S. Elaine and J. Markus, "Scambaiter: Understanding Targeted Nigerian Scams," in *Network and Distributed System Security Symposium (NDSSS)*, San Diego, 2014.

[11] M. Marti, M. Damon, L. Kirill, S. Stefan and M. V. Geoffrey, "An analysis of underground forums," in *ACM SIGCOMM conference on Internet measurement conference (IMC '11)*, Berlin, 2011.

[12] D. University, "The Upworthy Don: Formulas That Drive Google, Klout, Facebook Help Drexel Researchers Understand Organized Cybercrime," 3 April 2015. [Online]. Available: http://drexel.edu/now/archive/2015/April/organized-cybercrime/. [Accessed 15 April 2015].

[13] C. Ryan, "Darknet Market Basket Analysis," 24 March 2015. [Online]. Available: http://ryancompton.net/2015/03/24/darknet-market-basket-analysis/. [Accessed 23 April 2015].

[14] L. Kirill, P. Andreas, C. Neha, E. Brandon, F. Márk, G. Chris, H. Tristan, K. Chris, K. Christian, L. He, M. Damon, W. Nicholas, P. Vern, M. V. Geoffrey and S. Stefan, "Click Trajectories: End-to-End Analysis of the Spam Value Chain," in *IEEE Symposium on Security and Privacy (SP '11)*, Oakland, 2011.

[15] M. Damon, P. Andreas, J. Grant, W. Nicholas, K. Christian, K. Brian, M. V. Geoffrey, S. Stefan and L. Kirill, "PharmaLeaks: understanding the business of online pharmaceutical affiliate programs," in *21st USENIX conference on Security symposium*, Berkeley, 2012.

[16] M. Damon, D. Hitesh, K. Christian, M. V. Geoffrey and S. Stefan, "Priceless: the role of payments in abuse-advertised goods," in *ACM conference on Computer and communications security (CCS '12)*, New York, 2012.

[17] O. Fatih, G. Murat, E. Zeki and O. Yakup, "Detecting criminal networks: SNA models are compared to proprietary models," in *IEEE Int'l Conf. on Intelligence and Security Informatics*, Arlington, 2012 .

[18] D. Jana and M. C. Kathleen, "Using Network Text Analysis to Detect the Organizational Structure of Covert Networks," in *Proceedings of the North American Association for Computational Social and Organizational Science Conference (NAACSOS)* , Pittsburgh, 2004.

[19] A. K. Laura, E.-H. Latifa, E. Lawrence, F. Tom, H. Mustapha, L. Patrick, P. Souly, P. Victor and R. Lukas, "The Medicines Quality Database: a free public resource," *Bulletin of the World Health Organization,* vol. 92, no. 1, p. 2–2A, 2014.

[20] Y. Chao, H. Robert, Z. Jialong, S. Seungwon and G. Guofei, "Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter," in *Proceedings of the 21st international conference on World Wide Web*, Lyon, 2012.

[21] K. Anine, "Using social network analysis to profile organised crime," Institute for Security Studies, 2014.

[22] L. Matthieu, M. Clémence and D. V. Nathalie, "Basic notions for the analysis of large two-mode networks," *Social Networks,* vol. 30, no. 1, p. 31–48, 2008.

[23] P. Pascal and L. Matthieu, "Computing communities in large networks using random walks," in *Computer and Information Sciences - ISCIS 2005*, Springer Berlin Heidelberg, 2005, pp. 284-293.

[24] INTERPOL, "Pharmaceutical Crime on the Darknet," 2015.