

Sentiment Analysis of Social Media Data using Supervised Learning and Ekman's Basic
Emotions Theory

By

Thabiso Diale

Reg. No: 13100023

Department of Computer Science and Information Systems

Faculty of Science

Botswana International University of Science and Technology

dialet@biust.ac.bw

+267-71353643

A Dissertation Submitted to the Faculty of Science in Partial Fulfilment of the
Requirements for the Award of the Degree of Master of Science in Information
Systems of BIUST

SUPERVISOR:

Dr. Hlomani Hlomani

Department of Computer Science and Information Systems

Faculty Of Sciences

Botswana International University of Science and Technology

Mobile: +267 75025393 Work: +267 4900117

hlomanihb@biust.ac.bw

Signature  _____

Date: 18-09-2018 _____

© Thabiso Diale, September 2018

Declaration and Copyright

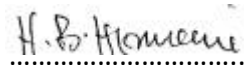
I, THABISO DIALE, declare that this thesis is my own original work and that it has not been presented and will not be presented to any other university for a similar or any other degree award.

Signature

This thesis is copyright material protected under the Berne Convention, the Copyright Act of 1999 and other international and national enactments, in that behalf, on intellectual property. It must not be reproduced by any means, in full or in part, except for short extracts in fair dealing; for researcher private study, critical scholarly review or discourse with an acknowledgement, without the written permission of the office of the Provost, on behalf of both the author and the BIUST.

Certification by Supervisor

The undersigned certifies that he/she has read and hereby recommends for acceptance by the College of Science a thesis titled: **Sentiment Analysis of Social Media Data using Supervised Learning and Ekman's Basic Emotions Theory**, in fulfilment of the requirements for the degree of Master of Science of BIUST.



.....

Dr.

(Supervisor)

Date: 17-09-2018.....

Acknowledgements and Dedication

My gratitude goes to God for making this thesis a success. I would like to thank my supervisor Dr Hlomani for guiding me during my studies and for motivating me when I felt discouraged. He was very patient with me all the time, may God bless him.

In addition, I would like to take this opportunity to show my appreciation to my colleagues for their support throughout the period of this research.

This research work is dedicated to my mother who has been the pillar of my strength. Without your support, this work would have not been possible. Finally, to my partner, you have been more understanding and very supportive. Thank you.

Table of Contents

Declaration and Copyright	i
Certification by Supervisor	ii
Acknowledgements and Dedication	iii
List of Tables.....	vii
List of Figures	viii
List of Abbreviations	ix
Abstract	x
CHAPTER 1: Introduction	1
1.1 Overview	1
1.2 Background	2
1.2.1 Sentiment Analysis.....	3
1.2.2 Crime Prediction and Classification	3
1.3 Problem Statement	4
1.4 Objectives.....	5
1.5 Research Questions.....	5
1.6 Significance of Study	5
1.7 Thesis Structure.....	7
CHAPTER 2: Literature Review	8
2.1 Introduction	8
2.2 Sentiment Analysis of Data	8
2.2.1 Social Media	10
2.3 Related Works on Sentiment Analysis and Crime Prediction	12
2.4 Adoption of Sentiment Analysis and Crime in Botswana	15
2.5 Ekman's six basic Emotions Theory.....	16
2.6 Machine Learning Techniques and Sentiment Classification.....	17
2.6.1 Naïve Bayes	19
2.6.2 Random Forest	21
2.6.3 J48 Classifier	22
2.6.4 k-Nearest Neighbour (K-NN)	23
2.7 Research Gap found in Literature	24
CHAPTER 3: Methodology.....	25
3.1 Introduction	25
3.2 Sentiment Analysis Workflow	25

3.2 Data Collection	26
3.2.1 Crime Dataset.....	27
3.2.2 Crime Statistics Report.....	27
3.3 Data Pre-processing	28
3.3.1 Data Cleansing.....	29
3.3.2 Dataset Translation	30
3.3.3 Replacing of abbreviations, emoticons, hashtags, repeating words	32
3.3.4 Word Tokenisation	32
3.3.5 Removal of Stop-words	32
3.3.6 Word Stemming	33
3.3.7 Feature Selection	34
3.4 Data Labelling.....	34
3.4.1 Applying Polarity Score	35
3.5 Building Classifiers.....	36
3.5.1 The Prediction Model.....	37
3.6 Testing the Classifiers.....	38
3.7 System Evaluation	38
3.8 Summary of Workflow	39
CHAPTER 4: Results and Discussion	40
4.1 Introduction	40
4.2 Evaluation Metrics for Sentiment Classification	40
4.2.1 Accuracy	41
4.2.2 Precision	41
4.2.3 Recall	42
4.2.4 F-Measure	42
4.2.5 Kappa Statistics	43
4.2.6 Mean Absolute Error (MAE).....	43
4.2.7 Summary of Evaluation Results.....	44
4.3 Discussion.....	45
4.3.1 Random Forest Classification	45
4.3.2 Naïve-Bayes Classification.....	45
4.3.3 KNN Classification	46
4.3.4 J48 Classification	46
4.4 Sentiment Results and Police Crime report (PCR)	47
4.4.1 Sentiment Data	47
4.4.2 Police Crime Report.....	49

4.4.3 Sentiment Analysis vs. Police Crime Report.....	50
4.5 Threats to Validity	51
4.6 Summary	51
CHAPTER 5: Conclusion and Future Work.....	52
5.1 Introduction	52
5.2 Outline of this research.....	52
5.2.1 Reflections on the Objectives and Research Questions.....	52
5.3 Conclusion.....	54
5.4 Future Work	55
References.....	57
Appendices.....	66

List of Tables

Table 2.1: The 5 V's of Big Data.....	11
Table 4.1: The accuracy of all four classifiers used for sentiment classification	41
Table 4.2: The precision for each classifier	42
Table 4.3: Recall percentage for each classifier	42
Table 4.4: The f-measure for each classifier	43
Table 4.5: The kappa statistics for each classifier	43
Table 4.6: The MAE for each classifier	44
Table 4.7: Summary of all the classifiers showing accuracy, precision, recall and f-measure, MAE and kappa statistics	45
Table 4.8: Confusion Matrix.....	48
Table 4.9: Crime Distribution by districts.....	50

List of Figures

Figure 2.1: Sentiment Analysis techniques	10
Figure 2.2: The Machine Learning Process.....	18
Figure 2.3: The process for supervised learning approach	19
Figure 2.4: A sample of a decision tree.....	22
Figure 2.5: K-NN example	23
Figure 3.1: The sentiment analysis approach	26
Figure 3.2: Steps followed for pre-processing of the data.....	29
Figure 3.3: Stemming process- words before and after stemming	33
Figure 3.4: The steps for setting up machine learning classifier.....	37
Figure 3.5: The prediction model.....	38
Figure 3.6: The overall architecture for our methodology	39
Figure 4.1: Percentage comparison of the sentiments.....	49

List of Abbreviations

API	Application Program Interface
ANEW	Affective Norms for English Words
ARFF	Attribute Related File Format
BOW	Bag of Words
DT	Decision Tree
ID3	Iterative Dichotomiser
GTD	Global Terrorism Database
KDD	Knowledge Discovery
KNN	K-Nearest Neighbour
MAE	Mean Absolute Error
ML	Machine Learning
NB	Naïve Bayes
NLP	Natural Language Processing
START	Study of Terrorism and Responses of Terrorism
SVM	Support Vector Machine
TD-IDF	Term Frequency and Inverse Document Frequency
PCR	Police Crime Report
URL	Unified Resource Locator
WEKA	Waikato Environment for Knowledge Analysis

Abstract

Social media has emerged as an effective source to investigate people's opinions in the context of a variety of topics and situations, including crime. Crime solving can be a difficult task hence requiring human intelligence together with experience. Crime data is usually big and full of noise; hence manual analysis of this data is tedious and sometimes impossible to do. Data mining techniques such as Sentiment Analysis can help in analysing such big data. In Sentiment Analysis, polarity in a text is identified using text processing and classification. For our study, we used textual sentiment analysis in retrieving the sentiment or emotion carried by a piece of text at sentence level. We used multi-classification of data by categorising data using the Ekman's basic emotions theory. This research work applied Sentiment Analysis of Facebook data in order to predict the public emotion regarding crime issues affecting their lives. A model was designed and trained using supervised learning approaches; Naïve Bayes, J48, K-NN and Random Forest. The model was tested on Facebook crime dataset. The results from the experiments showed that Random Forest outperformed all the other classifiers with an accuracy of 80%. Naïve Bayes followed with an accuracy of 71.29%, K-NN with 69.21% and J48, which was the least performing classifier, achieved 65.92%. The results of the predictive model were also used to demonstrate a correlation between the moods observed on social media and the Botswana Police Annual Report of 2016. Results show that sentiments have an impact on the outcome of the Police Report, hence showing a positive correlation. Results provide valued information that will assist the Botswana Police Service to put in strategy their actions with considerations to public sentiments and hence improve the process of handling crime related issues.

Keywords— Sentiment Analysis, Social Media, Supervised Learning, Naïve Bayes, Random Forest, J48, k-NN, Classification Algorithms, Crime Prediction, Ekman's Basic Emotions

CHAPTER 1: Introduction

1.1 Overview

In the past, traditional methods such as radios and newspapers were used to air public views. Sometimes it was difficult for the public to interact on these. Today, the availability of social media has given opportunities for people with similar interests to exchange opinions and views regarding any particular subjects. The availability of the Internet and rich resources such as review sites, online forums and web blogs, provide users with a chance to share their opinions [1]. Social media closes the gap between the real world and the imaginary world. However, due to the large amount of data posted on social media sites, it is a tedious process to collect such data manually, and it is often impossible for users to read all of the data.

To date, many techniques such as data mining and statistical analysis, social media mining, opinion mining or sentiment analysis have been put in place to solve this social media problem. Data mining and statistical analysis discover previously unknown patterns and information which brings a lot of benefits. Furthermore, sentiment analysis, which tries to understand the opinions of the writer, can be used, for example, to discover information that can help businesses to make better decisions based on the information that has been collected. Social media mining involves the representation and analysis of similar patterns of data from e.g, Twitter and Facebook. It encompasses tools to model and mine meaningful patterns from large scale social media data. Opinion mining or sentiment analysis is the process aiming to determine whether polarity of a textual corpus (document, sentence) tends to be positive, negative or neutral [2]. The data mining and sentiment analysis techniques work hand in hand to produce very useful information which can be applied to different domains. In business, sentiment analysis helps to analyse customer behaviour in order to improve their products, to get a chance for new business opportunities and better customer service [2]. In politics, sentiment analysis can help predict election outcomes and help in campaigns [3]. Similarly, in the movies business, customers can review or rate movies thus helping the organisation in bettering their decision making.

One of the areas of research that needs such data mining techniques is the security domain, mostly the crime section. According to Bolla [4], crime is an act that does not only harm an individual at hand, but it involves the nation as a whole. It is a forbidden act punishable by law. As society evolves, so does crime. In order for the justice system to keep up with the rapid pace of crime, it is important to constantly study crime and various methods to better handle the advancing situations of crime and the advancement of criminal methods [5]. There is also a need to implement crime prevention strategies to analyse crime data for better understanding of crime patterns and individuals engaging in crime, using data mining techniques [6]. In the past, the Botswana Police have tried to use social media such as Facebook to understand crime related issues. This was done by manually going through posts and pinpointing certain pointers they needed. This proved to be difficult because of the huge amount of data found in social media spaces. Hence why the need for automatic techniques to perform this task.

The aim of this study therefore is to apply sentiment analysis to empirically analyse crime data in order to predict the public's moods regarding crime in Gaborone (the capital city of Botswana). This will be achieved by taking advantage of the use of machine learning techniques for classification. This study also takes advantage of the abundance of information that is available in the public domain like social media platforms. We chose to focus this study on the crime domain because crime is one of the most concerning issues in Botswana, with highest crime rates in Gaborone region. Sentiment Analysis can be of importance in enhancing the Botswana Police Service in improving their services and identify new opportunities. This study is centred at sentence level as opposed to document level. When sentences are used as functional unit of analysis than the whole document, performance is improved.

1.2 Background

This section gives a brief account of history about the topic at discussion- Sentiment Analysis and Crime prediction. This sets the platform for further in-depth discussion of the Literature Review.

1.2.1 Sentiment Analysis

Sentiment analysis creates a fast growing research area, especially since the emergence of Web 2.0 and its related technologies. The goal of sentiment analysis is to decide the opinion of a writer. Recent studies have shown that studying or analysing people's feelings or opinions in social media platforms is useful for forecasting market behaviours, political views and product reviews [1]. The recent massive usage of micro-blogging services, particularly Twitter, has shifted attention to sentiment analysis of microblogging posts and tweets. Many research domains have studied sentiment analysis. For example, Chong [7] presented a preliminary experiment on tweets sentiment analysis. The experiment was aimed to extract sentiments based on subjects that are found in tweets. It detected the sentiments that referred to specific subject using Natural Language Processing (NLP) techniques. According to Cambria [8], sentiment analysis is a restricted NLP problem, where it is essential to understand the positive or negative sentiments in each sentence and the target entities. Depending on whether the target of a study is a document, text or sentences, different NLP and sentiment analysis tasks can be achieved. The first step to sentiment classification for sentence level, is to classify a sentence as subjective or objective, hence known as subjectivity classification. Subjective sentences are then classified as positive or negative, also known as sentence level classification. If you do not know the target of a sentence, then polarity identified for the sentence will not be beneficial.

1.2.2 Crime Prediction and Classification

Previous works have been done to analyse sentiments in the context of information security. These include some works done on detecting crime activities on the web, including prediction of crime and prediction of future crimes or attacks. A study by Azeez and Aravindhar [9] indicated that crime can be predicted by focusing on patterns and trends from various contributing factors. They explained that for one to have an effective crime prediction model or technique, it is important to consider various data inputs. In another research by McClendon and Meghanathan [10], they did a comparative study between violent crime patterns from communities using datasets provided by California and historical crime statistics. The Waikato Environment for Knowledge Analysis (WEKA) tool was used to do the analysis of the data using linear

regression, decision stump algorithms and additive regression. The results showed that linear regression performed best. Decision stump performed poorly which could be attributed to certain factors of unpredictability in the various crimes and associated features. Nouh et al. [11] proposed a design of a cybercrime intelligence framework to help identify profiles that post malicious Unified Resource Locator (URL) and looks at social relations between accounts. Yang et.al [12] also performed an empirical study to learn cybercrime communities in Twitter [12] [10].

1.3 Problem Statement

Classification algorithms mostly used for crime predictions are on basis of historical data, lacking real-time feasibility. This means that the initiation of an investigation is usually prompted by an external event (reported crime) or an internal event. This may leave other crimes unreported and hidden. Also crime data is usually big and manual interpretation of this data consumes a lot of time during analysis, making analysis almost impossible. Most of the studies on sentiment analysis classify sentiments using polarity as positive, neutral and negative, thus having limitations of not capturing all possible emotions. While this kind of analysis can be used for other application areas such as product reviews and movie reviews, it may not be suitable for crime data due to multiple emotions.

The problem this study addresses is the need for automatic classification based on multiple emotions. Therefore, in our study, we apply multi-class classification to sentences, something that previous approaches do not adapt more often. There is a lack of evidence in terms of literature that shows the adoption of sentiment analysis in the information and security sector of Botswana. This sector can implement the sentiment analysis techniques on social media to help with classification of their (large) data by using machine learning techniques. The work presented here was motivated by R. Iqbal et.al [13], who concluded in their study that there is need to apply other classification algorithms on crime data and to evaluate their prediction performance.

1.4 Objectives

To apply sentiment analysis of social media data using supervised learning approaches to predict people's emotions on crime, and establish how these relate to the annual crime statistics report.

Specific Objectives

To achieve the general objective, the following are the specific objectives that need to be achieved:

1. To collect data from social media
2. To manually perform pre-processing of data in order to remove noisy and irrelevant data
3. To train and test the data using supervised learning techniques; Naïve Bayes, Random Forest, K-NN and J48
4. To evaluate the supervised learning techniques based on accuracy, f-measure, Mean Absolute Error (MAE), precision, recall and Kappa Statistics
5. To compare the sentiment classification results with the annual Police Crime Report (PCR)

1.5 Research Questions

To achieve the above stated objectives, the formulated research questions have to be answered:

1. How can crime data be pre-processed prior to sentiment analysis?
2. Which supervised methods exist for sentiment classification of social media data and any data in general?
3. Which test and evaluation measures can be used for sentiment analysis of social media crime data?

1.6 Significance of Study

Although sentiment analysis of crime from social media to some extent has been studied, there are still gaps identified, therefore there is still a need to carry out research

to extend to these. The following are some highlights that speak to the significance of this research.

1. Sentiment analysis is still relatively new in the context of crime and security; therefore there is need to do it using different techniques and different datasets in order to compare results with other previous studies within the same domain and also in different domains. This contributes baselines and test cases to the body of knowledge of sentiment analysis, in particular and to the application areas (security in this particular case). In the case of this study, the datasets that will be used are completely new and different from what has been done in the past literature. Through sentiment analysis and crime correlations, it is possible to mine crime connections/links which can provide leads towards critical points in the crime chains.
2. This study is the first of its kind in Botswana, there is no past literature that has been done on sentiment analysis in the Police sector. The results for this study could be beneficial to the sector and the nation at large, thus can improve decision making, adoption of new strategies and methods.
3. The results will be useful to researchers in the field of Information Systems who are willing to pursue research on sentiment analysis. Other researchers could reproduce our experimental set-up to carry out their research. They may also have access to our crime dataset and use it for data analysis. This platform forms a basis for further technical and analytical research on crime trends providing opportunities in other fields/businesses (security system resourcing, like optimisation tools for dynamic resource allocations).
4. The results of this study could also help raise people's awareness on crime hence helping the Botswana Police to carry out criminal investigations, crime prevention initiatives and crime detection efforts.

Overall, this study does not only speak to the societal issues but also contributes to the scientific body of knowledge as enumerated above.

1.7 Thesis Structure

The thesis is organized into a number of chapters. The first chapter gave an overview of the thesis and introduced the problem, the objectives and research questions. In chapter two, we present existing literature, a background and gaps in the literature. Chapter 3 discussed the methodology, tools and technologies for achieving each objective for this study. In chapter 4, we discuss the experiments and the results evaluation. Chapter 5 is the conclusion derived from this study and direction for future work.

CHAPTER 2: Literature Review

2.1 Introduction

In this chapter we review the previous researches conducted in the area of the current study. We review sentiment analysis and its approaches. We also discussed previous works done on sentiment analysis and social media, giving examples of different domains (education, movie reviews, product reviews). This chapter further conferred sentiment analysis in the area of crime prediction and gaps that were found in the past literature.

2.2 Sentiment Analysis of Data

Sentiment analysis is a study that analyses people's opinions, or emotions towards e.g. services, products, issues, events, topics and individuals [14]. According to Kontopoulos [2], sentiment analysis is a process one's polarity over a piece of text is determined to be either negative, neutral or positive. Liu [14] explains sentiment analysis as a process categorising textual information to determine emotions, feelings and attitude over a subject. Another term for sentiment analysis, is opinion mining. An opinion is a judgment, viewpoint, or a statement that is not conclusive [15]. It may deal with subjective matters in which there is no conclusive finding. An example of a sentiment could be: "I love movies" or "This movie is the best movie I have ever seen". Sentiment analysis is a data mining technique that uses machine learning techniques to evaluate text [16]. It has the capability to triangulate methods such as quantitative and qualitative, by real time data collection and analysis. It provides a simple, fast and less expensive alternative to traditional qualitative market research techniques like observations. Manual analysis of data is time consuming not only because of time spent on analysing the actual data but also on the time spent on cleaning, organising the data. Sentiment analysis applications that already exist have already shown that the knowledge extracted from this type of analysis is useful to many life sectors.

Sentiment analysis falls within the Big Data paradigm. Big data is used for any collection of complex and large data sets that are difficult to process using traditional data processing applications [17]. Determining sentiments from this big data is demanding, and doing it manually is practically impossible [18]. Different methods have been formulated to execute automatic polarity classification of textual data. Despite this,

human analysis is essential in performing sentiment analysis as automated systems are unable to make decisions or classify sentiments correctly on their own [19]. Till today, different sentiment analysis techniques are being implemented with different aspects of evaluation for big data opinion analysis. Sentiment analysis is associated with two sub-tasks which are sentiment detection and sentiment classification. Sentiment detection is categorizing units of text into documents that express opinions or personal perspectives versus those that express objective (factual) accounts of matters [20]. Sentiment classification is concerned with finding whether a unit of text has positive or negative opinion of a certain entity or occurrence. This current study will take the latter route of determining the polarity of text.

There are multiple methods for measuring text sentiments, including lexical-based (also known as knowledge) approaches and machine learning methods. Figure 2.1 shows these approaches and their methods. Lexicon approaches are based on opinion words, i.e. words that are mostly used to express positive or negative sentiments. These opinion words are contained in a dictionary called opinion lexicon. Lexical methods will not be discussed deeply in this study because this study is focused mainly on machine learning approaches. As depicted by Figure 2.1, machine learning approaches are divided into supervised and unsupervised methods. These will be discussed in Section 2.6 of this chapter. Despite the wide use and popularity of some methods, it is not clear which method is better suited to identify the polarity (positive, negative or neutral) in a post [21]. This could be because current literature does not provide a method of comparison among existing methods.

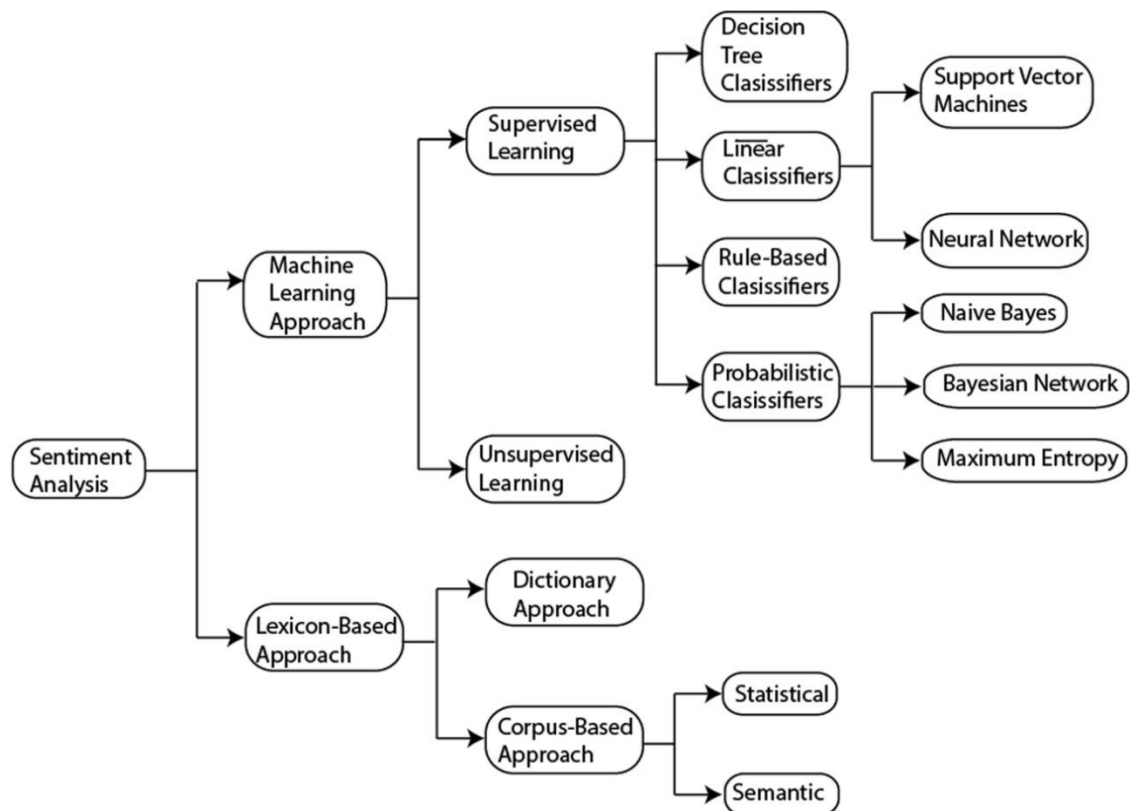


Figure 2.1: Sentiment Analysis techniques [22]

2.2.1 Social Media

Web 2.0 gives users the choice to interact amongst one another through social media in a virtual community. Social media has made life easier these days and business people are taking this route. Social media has many benefits compared to the traditional media that has been used in the past due to a number of reasons. It provides real time information, unprompted, no need to recall past behaviour and it is cost effective. This resulted in social networking sites, web applications and blogs. With the presence of online expressions such as reviews, ratings, online opinions are a benefit for businesses looking to market their products and identify new opportunities. As good as social media looks, it has its challenges: these can be classified as the 5V's of big data, as summarised in Table 2.1. The biggest challenge of social media is that of noise. Social media has data that is unstructured, informal and mostly difficult to understand. As far, there have been many techniques to try and remove this noise in order to make sense out of this data. Businesses are looking to automate and adopt the use of sentiment analysis by automating their processes of purifying noise [19]. Microblogs permit users to

interchange content such as sentences, individual images or video links. Researchers are also focusing on developing automatic tools for opinion mining. Several tools are already available in the market that helps companies to extract information from the internet. Some of these tools include SenticNet, Converseon, Factiva, Sentiment140 and SocialMention [16]. Researchers are also taking the route of sentiment analysis in micro-blogs to carry out and analyse sentiments such as in product reviews [22].

Table 2.1: The 5 V's of Big Data

The 5 V's	
Volume	Amount of data generated every second. The volume of data is rapidly growing making it difficult for analysis
Velocity	The speed at which data travels and generated. E.g. social media data can go viral in a blink of an eye
Variety	Data comes in different forms i.e. structured and unstructured. Today we are able to analyse all kinds of data being videos, images, text and so on
Veracity	Since there are many different forms of big data, data can be untrustworthy, making the quality of data uncontrollable. Big data technology therefore allows us to deal with this
Value	The most important. This is turning data into value so as to better understand and target customers and optimise business processes

In recent years, due to the vast amount of information available in social platforms, many researches focus on applying sentiment analysis to these data [23], [24] However, most all these works deal with English texts and retrieve them from Twitter, since it is easier to extract data from. Twitter is the most popular microblog service on the internet allowing posts to be publicly available to everyone. “With several hundred million users, its user-base is huge and the content produced currently amounts to 58 million tweets per day on average” [25]. Classifying texts according to polarity depends on the domain that they are used in. In addition to this, accuracy of sentiment analysis

can be calculated in many ways but the most common is scoring accuracy in comparison to a human. Any natural language processing that scores accuracy of 80% is considered great [26]. Sentiment analysis of Twitter data is used in various areas of research, e.g. in the stock market [27], [28] and public health [29]. An example of the use of the public mood could be seen during the Olympic games in London in 2012 [30]. Daley Thompson, a famous athlete, generated a light show that was exhibited on the London eye. A sentiment analyser analysed the tweets as positive and negative about the Olympic Games and the London eye according to the percentage of positive tweets.

Several types of information are of interest in the financial or business world. For example, investors may be concerned about financial news that are related to their investment. Companies may want to know about their competitors, customers, suppliers; customers may also need to know about other customer reviews. Sentiment analysis applications have surfaced that may be of help to different sectors such as financial news, products reviews, political elections and healthcare [14]. Sentiment and emotion analysis has gained eminence in government and multi-national companies to take advantage of the public mood to win elections. Political campaigns are placing more emphasis on social media tools as a low cost platform for connecting with voters and promoting engagement among users in their political base. The successful use of social media in the US presidential campaign of Barrack Obama in 2008, and the Arab Spring in 2011 has established Twitter, Facebook and other social media as integral parts of the political campaign toolbox [31]. One of the most famous studies in this field of market prediction through sentiment analysis is by Bollen et al. [27]. The study discovered a match in the Twitter feeds as causality in Granger's sense in the stock market. About 2.7 million people posted, and 10 million tweets were mined and sentiments extracted from this.

2.3 Related Works on Sentiment Analysis and Crime Prediction

This section highlights what has been done in previous studies in the area of crime prediction using sentiment analysis. Determining polarity of sentiments in crime is a difficult task. Usually polarity is assigned as positive, negative and neutral. This type of polarity is mostly used and is suitable with product reviews, movie reviews and not with crime reviews because of the latter's many attributes. Words can have different

meaning depending on the context that they are used in and polarity of words may differ from one domain to the next. For example, the word '*unpredictable*' can mean positive in a movie domain but negative in a crime domain. Sentiment classifiers trained in one domain may not perform well in other domains [32]. Learning based classifiers might not be as suitable for analysing newly created data, therefore, using a learning-based classifier is favourable if the use case and the domain of analysis is known beforehand [33]. More generally, as Turney [34] observed, sentiment analysis tasks become more difficult as the topic becomes more abstract. Several studies have been carried out on text mining, crime analysis and prediction using machine learning techniques and lexical methods. Crime analysis includes exploring and detecting crimes and their relationships with criminals. A study by Bolla [4], was carried out to perform sentiment analysis of online social networks to help detect crime patterns. Twitter was used to extract tweets from cities considered to be the most dangerous or safest in the United States. 100,000 crime related tweets were collected over 20 days. With this data, Forbes reports were used to compare with the tweets to check for any similar relationships. ANEW (Affective Norms for English Words) approach, deep learning and recursive model were used to do sentiment analysis. ANEW approach is developed to provide a set of emotional ratings for large number of words in English. In this study, ANEW approach was used to cluster the tweets according to a category of five polarities (very positive, positive, neutral, negative and very negative). Recursive deep models were used for sentiment analysis to overcome limitations of a dictionary-based approach. The results of this study show that indeed there is a correlation between online data with the crime statistics from the police.

Wijeratne et al. [35] analysed social media footprints of street gangs. The paper examines messages posted by gangs over a geographical area using Twitter as a source of data. A framework was developed to capture tweets posted by these gangs. Results discovered that posts published by gang members reveal the types of sentiment and what the platform can automatically detect. Early work by D'ecary et al. [36], used a small number of keywords in identifying gang related posts to determine what they share online. This study also revealed the significant increase in internet use by street gang members in recent years. Hernandez et al. [37] used linear regression to perform sentiment analysis method on Twitter content to predict future attacks on the web.

Support Vector Machines (SVM) was used to learn patterns in training data to predict labels from unseen samples. The datasets used were from the ISIS attacks and the Israeli websites in retaliation to the Gaza-Israel conflict. Results showed a high correlation between the sentiment of Twitter users and actual attacks on the web. A number of similar machine learning algorithms were applied on different data sets in order to achieve different objectives or functions.

For example Saeed et al. [38] used decision tree and Naïve Bayes (NB) classifiers on dataset of criminal activity to predict attributes and event outcomes. In this case, Naive Bayes was more reliable and more accurate on crime analysis compared to decision tree. Similarly in another study by Shojaee et al. [39], they conducted an experiment to obtain better supervised learning algorithms to predict crime status and identify crime patterns. The algorithms used were Naïve Bayes, neural networks, decision tree (J48), SVM and k-nearest neighbour (K-NN). Naïve Bayes did well with an error rate of 0.898%, followed by K-NN with (0.895%), neural networks with (0.892%), then decision tree (J48) achieving (0.727%) and SVM at (0.678%) [39]. Tolan [40] came up with a study based on predicting terrorist groups responsible for attacks in Egypt from 1970 until 2013. Experiments were done on WEKA for classification. Five techniques were compared; Naïve Bayes, SVM, K-NN, Tree Induction (C4.5) and Iterative Dichotomiser (ID3). This study used data from Global Terrorism Database (GTD) from National Consortium for the study of terrorism and Responses of Terrorism (START). SVM was more accurate than Naïve Bayes and K-NN in mode amputation approach, while ID3 had lowest classification accuracy, although it performs well in other measures and in a Litwise deletion approach. K-NN outperformed the other classifiers in its accuracy but the overall performance of SVM is acceptable than other classifiers.

Almanie [41] carried out experiments using Naïve Bayes and decision tree to predict crime types. The study focused on finding spatial (area) and temporal (time) criminal hotspots. Two crime datasets for Denver from year 2010-2015 and Los Angeles in year 2014 were compared through statistical analysis. This study predicts the likelihood of a crime happening in a particular location at a specific time. Apriori algorithm was used to produce frequent patterns of criminal hotspots. Decision tree and NB classifier

predicted crime within a place at a specific time. Naïve Bayes did better performing at 51% accuracy in Denver and 54% prediction accuracy in Los Angeles. Decision tree did not do well reporting at 42% in Denver and 43% in Los Angeles.

2.4 Adoption of Sentiment Analysis and Crime in Botswana

Crime in Botswana has grown over the years, making it a serious problem affecting the entire country. According to Crime Statistics of 2015, crime has risen compared to the previous year of 2014 [42]. Gaborone region was reported to have the highest number of crime cases, making it the highest crime prone area. Currently in Botswana, the police rely on the public (historical data) to report cases, and to come up with these crime reports. These traditional methods are not able to capture all factors relating to crime, which means some of them (crimes) might not be captured at all. Therefore there is imminent need to use automatic data analysis techniques to carry such tasks. According to literature, nothing has been done on sentiment analysis in any kind of domain in the context of Botswana. Such analysis techniques exist which could be applied on many fields such as information and security, education, healthcare and politics. In education for example, these could be used for classifying student performance levels, cluster similar students based on educational background, age, or sex. Other models such as decision trees and Bayesian can be applied on the prediction of student courses and understanding of student life in general. In politics, it could be used to predict elections and help candidates to carry out their campaigns. For healthcare, these techniques could predict diseases and help in diagnosis.

For the information and security domain, sentiment analysis techniques could impact on future decision making processes and help in coming up with more effective strategies and policies. They could also help detect criminal hotspots, culprits and the mostly affected areas of crime. People go on social media to discuss different topics of interest, and such data could be useful to the Botswana Police service in order to predict for example, future crimes. Even though the Police have not engaged sentiment analysis techniques yet, the Police are gearing towards the use of social media. Currently, the Botswana Police is working in hand with IT professionals to go on social media to monitor crimes. According to a report by [43], the Botswana Police Service has warned

the public about crimes committed on social media platforms such as Facebook and Twitter. Apparently there are those people who misuse social media unlawfully. Such crimes include blackmail, extort, use of insulting language and posting of pornographic content. The police said that these crimes are punishable by law. The police also use social media data:

1. for interventions
2. to see most frequently talked about topics, to help them carry out investigations
3. Social media has the power to decide court rulings

2.5 Ekman's six basic Emotions Theory

Paul Ekman and others have developed theories on how to categorise some emotions [44]. These emotions are said to have ties to universal facial expressions and physiological processes such as increased heart rate and perspiration. Ekman's emotions are one of the most agreed upon universal emotions [44]. These represent facial expressions of emotion being *happiness, sadness, surprise, anger, fear, disgust*. By using these emotions, it gives sentiment analysis for understanding consumer feelings concerning a certain product. Fine-grained emotion labels, in contrast to positive or negative labels, would increase the effectiveness of sentiment classifiers [45]. Ekman's theory has been applied in many works in order to classify emotions. Aman et al. [46] organized a corpus annotated with rich emotion information. They collected blog posts directly from the Web and prepared a list of seed words for six basic emotion categories proposed by Ekman. In their study, they added *mixed emotion* and *no emotion*, resulting in eight categories to which a sentence could be assigned. While sentiment analysis usually focuses on documents, this work focuses on the sentence-level analysis. The annotators used kappa to measure the emotions in terms of high, medium and low. For classification experiments, SVM and Naïve Bayes were used. SVM did better by accuracy of 73.89%, and Naïve Bayes with 10 fold cross validation gave an accuracy of 65.6%. In another study by Chaffar and Inkpen [47], the authors adopted a supervised approach to classifying emotions using Ekman's six emotions. They used a mixed emotion dataset that consist of blogs, fairy tales and news headlines. SVM, J48 and Naïve Bayes classifiers

were used to perform these experiments. SVM performed well with an accuracy of 71.69% better than the others.

Sentiment analysis was used to analyse students in relation to education services, policies and issues. Five emotion categories for Ekman's were utilised (anger, fear, joy, sadness, and surprise), in which joy and surprise emotions were assigned to positive class while anger, fear, and sadness were assigned members of negative class. Negative emotion, disgust, was removed because the emotion is similar to anger and leads to making sentiment classes biased [45]. The study adopted the Ekman's theory to label sentiments and have a deeper understanding of the public's emotions. A study by Do et al. [48], investigated public responses expressed on Twitter during the 2015 Middle East Respiratory Syndrome (MERS) outbreak in South Korea. To conduct the analysis, they developed an emotion classification scheme that is particularly attuned to Korean and Twitter domain. They classified tweets into seven categories that include neutral (no emotion) and Ekman's six basic emotions: happiness, sadness, anger, disgust, fear, and surprise.

2.6 Machine Learning Techniques and Sentiment Classification

Machine learning is a subfield in computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence [21]. Its focus is on prediction that relies on known properties learned from the training data, and has been used in many domains of research. Sentiment analysis is a part of NLP. Using machine learning (in a correct way) can help you boost the performance of NLP systems or sentiment analysis tools. Regardless of the platform being used, machine learning follows a simple process as shown in Figure 2.2. The goal is to come up with a model as it is the main thing that applications can submit requests to in order to gain insight on new data. The process starts with a question; what are you trying to learn from your machine learning experiment [49]. Once you have prepared data, you can now apply machine learning algorithm to produce a model.

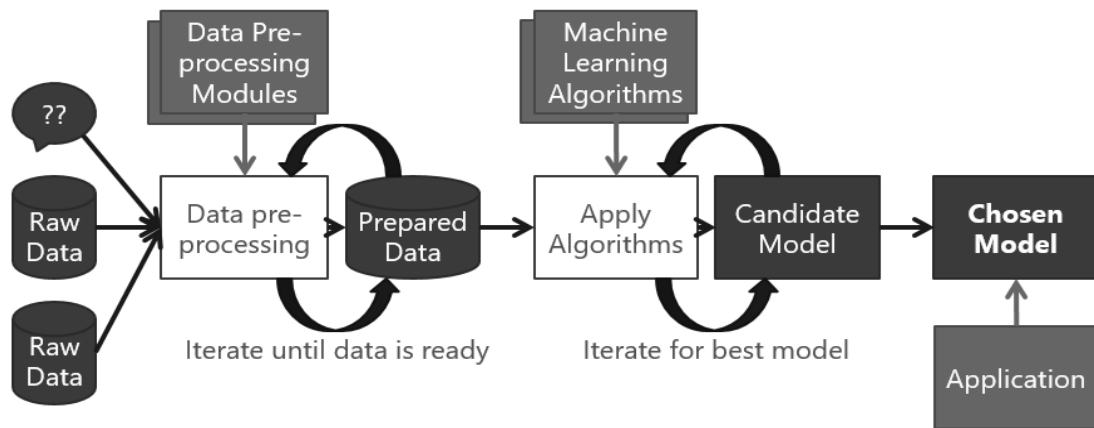


Figure 2.2: The Machine Learning Process [50]

An example of the use of machine learning techniques is presented by Pang et al.[51]. The study compares different techniques for classifying movie reviews, with SVM performing best at an accuracy of 82.9%. Another study by Boiy et al. [18] used classifiers Naïve Bayes, Maximum Entropy and SVM to classify movie reviews and document blogs based on unigrams, bigrams and adjectives. Naïve Bayes proved to be the fastest. In education domain, Troussas et al. [52] found Naïve Bayes to be the best technique, while Song et al. [53] found SVM to be the best when using the above three techniques. This shows that even for the same domain, different machine learning techniques give different results, prompting a need for testing several techniques. Machine learning methods for classifying data to their sentiment polarities do well in specific domains. Machine Learning techniques are divided into supervised and unsupervised techniques. In general, supervised techniques need prior training. This takes great amount of time and training datasets are not always available in all domains. Unsupervised approaches do not need to have any labelled data for training. Our study adopts the use of supervised methods. This is because they have been considered as the best techniques for text classification.

Supervised binary classification classifies text into three classes of positive, neutral and negative. According to this approach, a sentiment classifier is trained to distinguish positive, negative or neutral sentiment in textual corpora. Supervised methods must be manually trained an assigned polarity values, and must learn features corresponding to the value. Then the model is able to predict the sentiment of new, previously unseen

instances. To determine accuracy of a model, there has to be a test set which tests if a model produces expected results. The steps involved in supervised learning are shown in Figure 2.3. The input for the process is training data. Data is trained and labelled, then fed to machine for learning. Then new data is presented and fed to the prediction model so that it can be evaluated. The predictive model is supposed to output the expected labels which the model was built on. Sentiment analysis is domain-specific when using supervised learning approaches. As a result, when data is transferred to another domain, a sentiment classifier often performs badly [54]. Supervised methods include Naïve Bayes, Decision trees (Random Forest and J48), K-NN and Support Vector Machines. Some of these algorithms are adopted for this current study and are explained in the next section.

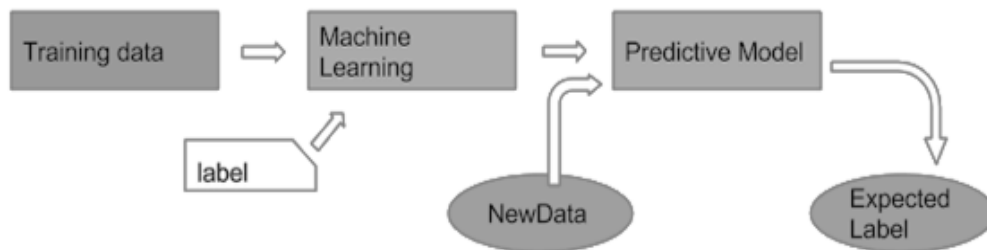


Figure 2.3: The process for supervised learning approach [55]

2.6.1 Naïve Bayes

NB classifier is based on the Bayes theorem, and this technique is linear, meaning that all data samples can be separated by a line where each vector represents a sample. One important component of this classifier is its independence assumption among the categories meaning that if assigned a negative, positive or neutral it assumes the words are conditionally independent of each other [18]. It is called “naive” because it assumes independence between text features. It is used to predict things based on its prior knowledge and assumptions. Naïve Bayes has been used to carry out sentiment analysis in order to observe social media and how the latter relates to conventional media. It is different from other classifiers like K-NN and SVM such that it builds upon probabilities (of a feature belonging to a certain category) whereas the other approaches interpret the document feature-matrix spatially [55]. Posterior probability, class conditional probability and prior probability play an important part in determining the results of a particular instance [56].

As reviews are being classified, the only features that any classifier considers during the training process are the words. Therefore as the naïve rule states, the words are independent of each other. The formula for NB is as adapted from literature [57] is described below:

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)} \quad (1)$$

- Where $P(H|E)$ is the probability posterior of P given the evidence, evidence being what we have seen or the dataset. The probability of outcome knowing the evidence [58]
- $P(H)$ is the prior probability, $P(E|H)$ is the likelihood of evidence if H is true
- $P(E)$ is the prior evidence if it is true

For the case of this current study, suppose that we have comments that are classified as either happiness or sadness. Then we are also given that in the previous comments that Happiness comments are more than Sadness comments, and then the new comments are not known. The formula [57] to see the likelihood of new comments could be written as :

$$P(S|H) = \frac{P(S)*P(H|S)}{P(H)} \quad (2)$$

Where S is for 'Sadness' and H is for 'Happiness. Then we can write our prior probability as:

$$\begin{aligned} \text{Prior Probability of Happiness comment } P(H) &= \frac{\text{no of Happiness comments}}{\text{Total no of comments}} \\ \text{Prior Probability of Sadness comment } P(S) &= \frac{\text{no of Sadness comments}}{\text{Total no of comments}} \end{aligned}$$

This classifier has disadvantages such that if a category variable has category that are not observed in the training data set, the model will assign a zero probability and would

not be able to make predictions [59]. This is known as zero frequency. To overcome this, we can use smoothing techniques such as Laplace estimation.

2.6.2 Random Forest

Random Forest and J48 classifiers fall under the decision tree family. Before we discuss these classifiers, it is important to know what a decision tree is. It is a structure that includes a root node, branches, and leaf nodes [60]. It breaks down a dataset into smaller subsets. Decision trees can be built using many different methods such as ID3 (Interactive, Dichotomiser). Classification and regression trees (CART) was later developed to handle binary trees. According to Hongbo [61] and Kalpana [62] the strengths of Decision Tree (DT) are that it assigns a class label to an unseen record, as well as explains why the decision is made in an easy-to-understand classification rule. DT classifies unseen records efficiently, and it can handle both categorical and continuous attributes. Decision trees are well fitted for classifying crimes and criminal objects [63].

Figure 2.4 shows a simple decision tree of crime types. Rectangular nodes are used for testing nodes while the oval shapes are for the results. From the example, if one has a criminal record, the type of crime is defined. If it is rape, then the person faces jail, hence if pickpocketing then they are given a fine. If one does not have a criminal record, then no jail or fine. Now, going back to Random Forest, these are an ensemble learning method for classification that operate by constructing a multitude of decision trees at training time and outputting the class by individual trees [64]. The higher the number of trees in the forest, the higher the accuracy. It is good with handling missing values. It can be used both for classification and the regression problems. Random Forest has been used in many applications such as e-commerce, banking, medicine and the stock market.

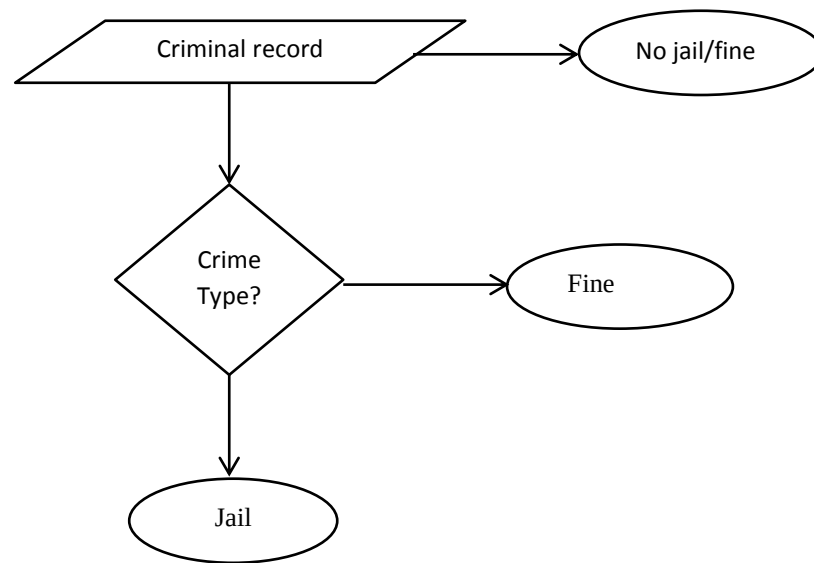


Figure 2.4: A sample of a decision tree

Random forest is one of the most accurate learning algorithms available. It has advantages that the training data are less sensitive to outlier data. Parameters can be set easily and eliminates the need for pruning the trees which improves accuracy. For many datasets, it produces a highly accurate classifier. It can run efficiently on large datasets. Although it is a good classifier, it also has a negative side; it easily over-fits its class. To reduce the number of over-fits, one has to reduce the number of trees as well as decrease the vague links that are present.

2.6.3 J48 Classifier

J48 or C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan [63]. It is a program that creates a decision tree based on a set of labelled input data. It also deals with missing values, numeric attributes and noisy data. Although it can deal with noisy data, its performance usually becomes slow for large and noisy data. In order to handle continuous attributes, it creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it. To construct a tree, top down approach is applied until a stopping criterion is reached and different methods such as Gain-in entropy is used for making nodes [39]. J48 has been used in many classification studies. For example, this algorithm has been used by Gholap [65] to predict soil fertility. It gave the highest accuracy of 91.90% amongst the

other decision tree classifiers being NB-Tree and SimpleCart. These other classifiers gave 85.51% and 91.75% respectively. A disadvantage to using this classifier is that Tree depth is linked to tree size, and to the number of examples. So the size increases linearly with the examples.

2.6.4 k-Nearest Neighbour (K-NN)

K-NN is the simplest of all the Machine Learning algorithms which can be used for prediction [66]. It can be expressed easily without the training model (so no learning required). It is sometimes known as the 'lazy algorithm'. It defers the decision to generalize beyond the training examples until a new query is met. K-NN can be used for both classification and regression. It makes predictions using the training dataset directly. Given a query instance x to be classified, q to be classified:

Let $x_1, x_2, \dots, x_k$ denote the k instances from training examples that are nearest to x_q , then return the class that represents the maximum of the k instances [67].

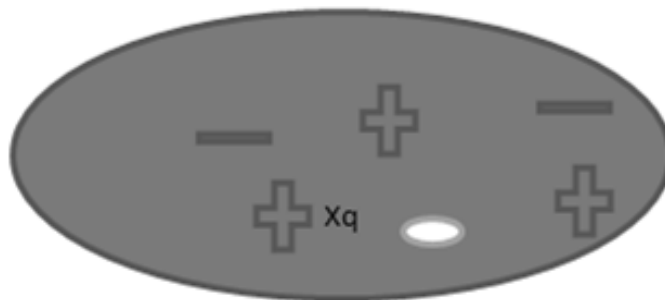


Figure 2.5: K-NN example

According to Figure 2.5, if $K = 5$, then query instance x_q will be classified as positive since three of its nearest neighbours are classified as positive. This algorithm has drawbacks such that it assigns equal weight to each attribute, which may cause confusion when there are many irrelevant attributes in the data [40]. K-NN works well with a small number of input variables, but struggles when the number of inputs is very large. Blachnick and Strzempa [68] showed that a properly configured K-NN may be as highly effective, if not more than other classification algorithms.

2.7 Research Gap found in Literature

This chapter analysed the literature on sentiment analysis and machine learning in the context of crime and crime prediction. After carefully going through literature, these are the gaps we found;

1. Clustering techniques are mostly used for crime mining and prediction. However, these methods are limited in the use of significant amounts of data due to high computational intensity [69]. Furthermore, they are inadequate to capture complex interactions. For our study, we used the supervised learning techniques since they are suitable for classifying textual data.
2. Most studies relied on historical data, lacking real time feasibility. With a few studies on crime prediction and sentiment analysis, these studies do not include understanding moods of the public concerning crime, and how these moods may affect the crime going up or down, say for a particular city.
3. Again we extended the previous commonly used polarities (positive, negative, neutral) by using the Ekman's theory to have an in-depth understanding of these sentiments, which this has not been the case in previous literature.

Our study therefore uses sentiment analysis of social media (Facebook) and machine learning classification to detect how social media sentiments can have an impact in crime rising or going down. The next chapter presents the methodology adapted towards development of the model.

CHAPTER 3: Methodology

3.1 Introduction

The large amount of data concerning crime is considered a treasure of knowledge; therefore unfolding it can bring benefits. Data mining techniques are used to analyse large amounts of data, in many areas including in prediction of crime. Sentiment analysis constitutes a branch of data mining that categorizes text data into positive, negative and neutral sentiments. This study takes advantage of sentiment analysis by using it on the application of crime in order to understand sentiments of the public. This study also aims at building a model to understand people's opinions on crime related issues. This will further help the Botswana Police sector in incorporating these sentiments in producing their annual crime report and helping them in making decisions.

Our approach is experimental in nature as the datasets will be analysed, tested and evaluated. The research design for this study is exploratory. The primary reason for exploratory research is in shedding light in a condition and identify specific objectives. Exploratory research is ideal for use if someone wants to clearly understand a state of affairs or make a decision. Both qualitative and quantitative techniques are ideal for exploratory research and both techniques were used for this current study. Qualitative data is unstructured data that one can extract some insights and understanding from. It relies much on interpretation, and once you have begun to collect the data, you will begin to review it and use initial findings to shape how one collects and interpret data in the future. It is useful when the researcher wants to know the feelings and attitudes attached to data, for example, in the case of our study where we want to know the public's sentiments. Quantitative techniques are the use of statistical analysis of data in order to measure it. In this study, we used statistics for our data presentation and interpretation in order to quantify attitudes, opinions and behaviours.

3.2 Sentiment Analysis Workflow

Our sentiment analysis model followed the workflow presented in Figure 3.1. It is made up of these stages: data collection, text pre-processing, data labelling, building and testing of classifiers, system evaluation and the presentation of results. These steps are

discussed in detail in sections that follow. For each of these steps, we described the objectives and how they were achieved.

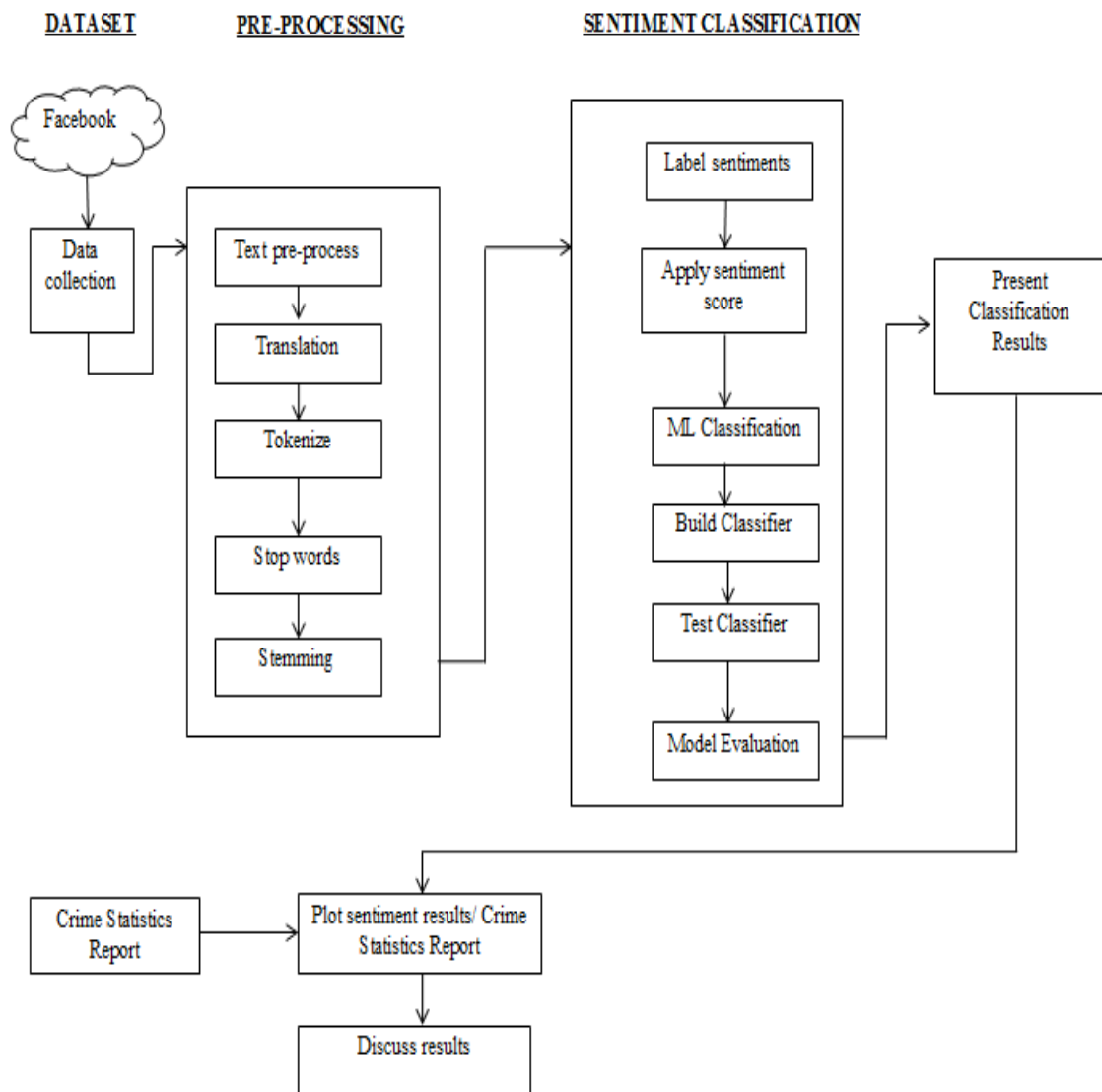


Figure 3.1: The sentiment analysis approach

3.2 Data Collection

This step responds to the 1st objective of our study- To collect social media data. To achieve this objective, we needed to collect data from two entities- crime data from Facebook and the Police Crime Statistics report of year 2016. These datasets are described in the sections below.

3.2.1 Crime Dataset

For our project, we extracted data using a tool called Netlytic [70]. It is a social network analyser tool that can automatically summarise and discover social networks from online conversations [70]. Netlytic has a number of capabilities being that; it is able to capture, analyse data from different social media podiums e.g. Facebook, Twitter. Netlytic allows access to its data through its API, allowing users to capture as much records as possible. We collected data from four different groups on Facebook. These are Fight Crime in Gaborone, Women and Men against Children Abuse, Botswana Online Magazine and the Voice online newspaper. The reason for choosing these groups was because they have a platform where people discuss about crime, making it the perfect platform for our study. Facebook was chosen as our data collection point because it covers a larger audience (in Botswana) than other social media like Twitter. Facebook allows access to its Application Program Interface (APIs). Although we had a system in place to collect data, it was difficult to choose the relevant comments and status updates. This was because these groups do not entirely report crimes in Gaborone only, so we had to selectively choose only data that is relevant to our area of research. We ended up manually selecting and separating subjective comments and status updates from objective ones (for Gaborone region only). This process of collection was iterative, thus taking a lot of time; hence going back to collecting if what we had collected was not enough.

We collected data of about 1,000 comments and status updates over a period of 12 months, from January to December 2016. We combined the data from these groups into one file (dataset). We refer to this file as “crime dataset”. Our crime dataset was saved to an Attribute Related File Format (ARFF).

3.2.2 Crime Statistics Report

Every year the Botswana Police releases an annual crime report stating the percentages of crimes reported in Botswana and the areas which those crimes happened. For our study, we used the crime statistical report of year 2016 for Gaborone region. The crime report was presented in text, tables and graphs. From this report, we only used a section of the tables where the crime statistics for Gaborone region were presented. This data was used to get statistics of crimes committed in Gaborone and to help determine which area in Gaborone has the highest percentage of crime.

3.3 Data Pre-processing

The second objective of this study was: to perform pre-processing to remove noisy data. Data pre-processing is often neglected but it is an important phase in the sentiment analysis process [47]. It is a first stage in the process of knowledge discovery of the KDD process, it reduces the complexity of the data and has a big impact on the accuracy and efficiency of results [71]. Using mining techniques and sentiment analysis on noisy data is indeed a big task in the sentiment analysis research [3]. Sentiment analysis may be improved by pre-processing the data. The aim in pre-processing is help present data in an organized manner hence increasing the machine interpretation on the text [7]. Analysing data that has not been carefully screened can produce misleading results, and can cause confusion in analysis. Experimental results that have been reported show that if pre-processing is performed well, using methods such as filtering and data transformation can improve classifiers performance [72]. Data pre-processing techniques performed for our study include: data cleansing, tokenization, stemming, feature selection, removal of stop-words and document translation for our crime dataset. These processes are shown in Figure 3.2. Some of these processes are done manually, e.g. document translation, removal of abbreviations and replacing of repeating words. Further tuning of the data was done automatically before modelling the data. These include removing of stop-words; stemming, word tokenisation and feature selection. Each step is explained in the next section. Since the data we collected was unstructured and noisy, this process of data pre-processing was time consuming and tedious.

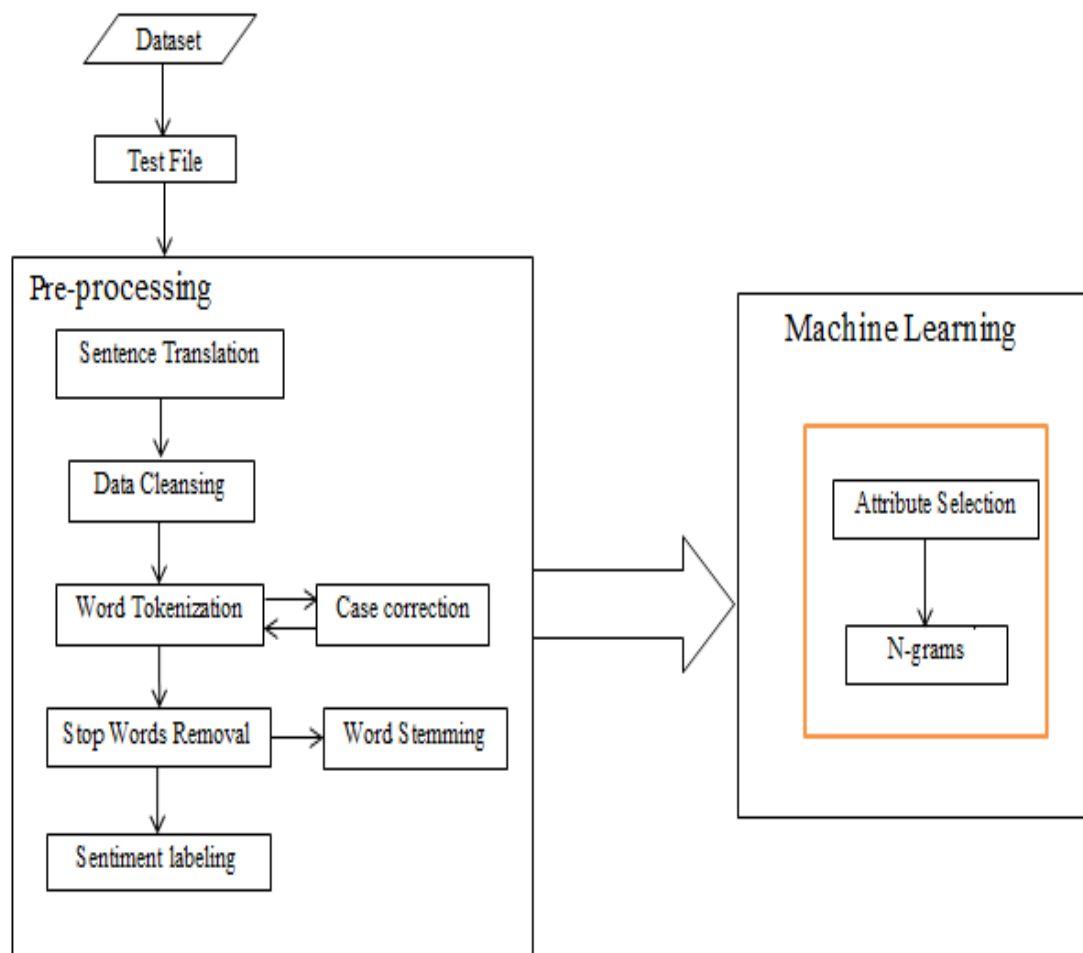


Figure 3.2: Steps followed for pre-processing of the data

3.3.1 Data Cleansing

Real-world data tend to be incomplete, noisy and most of the time inconsistent. Data cleansing or cleaning is needed to curb this situation by filling in missing values, smoothing out noise while identifying outliers and correcting inconsistencies in the data. After collecting our data, we did what is known as subjectivity analysis. The aim here was to decide if a sentence is subjective or not. By subjective we mean whether it carries human beings feelings or not. Subjectivity analysis is defined as distinguishing factual sentences from those used to present opinions [73]. Such sentences which do not carry any emotions are known to be objective. In the case of this study, these sentences were discarded. This process was done manually by reviewing sentences one by one, then discarding all sentences that were not relevant for this study. Still on this, we used some keywords to search for the subjective sentences that we wanted to keep.

We looked for keywords that commonly appear in crime related studies, these are some of the keywords we found, hence used for our search (but not limited to): *danger, guns, attack, survive, report, dangerous, police, criminals, kill* etc. After collection, our data contained a lot of noise and irrelevant fields.

To keep our data content, consistent and unique, all unnecessary fields like the links, source, type, publication date, author, author_id, title, like_count, comment_on were removed; we were only left with description of comments and status updates. After this step, we were left with 540 status updates and comments to be used for classification. This is because we discarded some sentences which were not relevant to our study. Filling in missing values was done manually. This process was time consuming because we had to go through our dataset line by line checking for missing information.

3.3.2 Dataset Translation

The collected Facebook data was written in both non-English (Setswana) and English languages. For consistency purposes we had to use one language so our analysis tool could not get confused in the process. “Past research on sentiment analysis have been on English language, so sentiment analysis systems in other languages tend to be less accurate” [74]. In comparison with English, it is challenging to execute sentiment analysis on other languages [74], [75], [76]. The experiments with sentiment analysis systems show that in languages where a strong sentiment analysis system does not exist, translating text to English from that language is a viable option. Some researchers argue that translation may have an effect on data by losing the meaning [77], [78]. Even though these authors think so, it is evident that indeed translation is necessary in some cases. In a study by Kaur and Kumar [18], they translated their dataset written in Hindi to English for consistency. In another study by Wang et al. [79], they translated their dataset from Chinese to English. In all these above mentioned studies, the datasets were still satisfactory even after translation. We came up with a few strategies for carrying out the translation process. We engaged a focus group, made consultations within the group and we applied the use of back translation.

Engaging a focus group

We formed a focus group of 8 participants. Focus groups are designed to take insights on a defined area of interest in a tolerant and non-threatening environment. The reason for engaging a focus group was to avoid biasness by the main researcher during translation. The focus group centred on:

- Translating dataset from Setswana to English language
- Data cleansing
- Correcting grammar and spelling
- Labelling data according to 6 polarity categories of emotions (this will be discussed under section 3.4)

The participants in the focus group had experience with the culture and were fluent with the both languages, hence why we chose them for this task. There is no corpus for Setswana language for natural language processing. Again, there are no available applications that translate documents and sentences from Setswana to English. Instead, the ones that are available only translates between both languages (Setswana to English and vice versa), word by word. Therefore translating word by word was not sufficient in our case because some words are independent hence this may distort meaning. We used Setswana Oxford Living Dictionaries for reference and to interpret words that were difficult to translate. The quality of translation was influenced by factors such as: the memoirs of the researcher and our translators, the translator's knowledge of the language and the culture and the researcher's and translator's fluency in the both languages. Secondly to avoid our dataset from losing its meaning, consultations were made within the group, and issues were discussed and agreed upon. Thirdly, back translation as one of the most common techniques used in cross-cultural research was adopted. It involved looking for equivalents through the translation of items from the source language to the target language. This was done in order to cross check for any mistakes that might occur during the translation process. For many researchers, combining the above techniques is seen as the best and most efficient way to deal with translation related problems [77]. Hence the adoption of these methods in our study.

3.3.3 Replacing of abbreviations, emoticons, hashtags, repeating words

This process was done by the focus group, by manually going through the data and removing or replacing all unnecessary data such as hashtags, and any other data that may distort meaning. Abbreviations (e.g., OMG, WTF, and BRB) and emoticons were translated. A hashtag like *#crimemustfall* was replaced with 'crime must fall'. One of the primary challenges with analysis of casual text, for example, was dealing with the problem of rambling spelling mistakes. Spelling was therefore corrected, including grammar. Repeating words were removed, for example 'so so *saaad*' was replaced with 'so *sad*'.

3.3.4 Word Tokenisation

Word Tokenisation is the process of breaking down or splitting words referred to as tokens [26]. Tokenizing the data makes it easy to separate and sieve only words that enhance worth to sentiment score. This is achieved by using delimiters such as comma, stop and space. Again, we did what we call case correction. All uppercase sentences were corrected to lowercase, this was done for consistency, uniformity and in order to avoid confusion during the machine learning classification of our data.

3.3.5 Removal of Stop-words

These are words that do not play a role in deciding sentiments in a sentence. For this project, we populated words from The Humboldt Digital Library [80]. These are words like:

{is, are, the, us, been, at, was, it, her, him, because, an, done, due, are, etc.}.

It is very important to perform this task on our dataset because it has a number of advantages being:

1. It increases classifier performance thus decreasing the run time during classification
2. Storage space of vectors are decreased

A code snippet was done using Python library to achieve this task. The code snippet is under Appendices section of this thesis.

3.3.6 Word Stemming

The goal of stemming is to reduce words to their base or original form. Stemmers remove the 'ing' and the 'ed' and 'en' form from words. For example, from our crime dataset, as shown by Figure 3.3, words like *kill*, *killing*, *killed*, *kills* were reduced to their root word- *kill*. When the feature vectors are built for these words, they will be treated as separate words, hence calculated separately. If stemming is applied, only the word *kill* will be added in the dictionary of the classifier, with word frequency 4. This therefore reduces the storage space of the vector. Stemming helped increase our classifiers accuracy for our crime dataset. There are several stemming techniques that are available, but for this project we used the Porters stemmer.

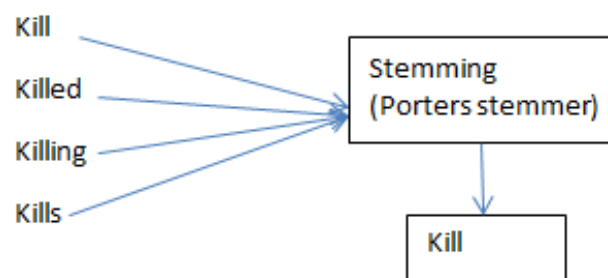


Figure 3.3: Stemming process- words before and after stemming

The reason for choosing this type of stemmer is because it is the least aggressive algorithm over all stemmers, with the specifics of each algorithm being fairly lengthy and technical. In stemming, a word is reduced to its stem making an assumption that each one is semantically connected. There are two important points to consider while using a stemmer:

1. Words that do not mean the same thing are separated
2. Morphology forms of a word are considered to share same meaning, hence are denoted to the same stem

3.3.7 Feature Selection

This is the process of selecting words or features (as used in text classification) that are more applicable to machine learning problem and help in model building. Only those features that help in determining sentiments are considered. The selection of features for our multi classification model is vital as it acts as an input for the classifier and it determines the sentiment of a sentence. Feature selection is imperative because it can improve accuracy in results, and it can also reduce training time since unwanted data will be less. All features selected must be in order and should be applied one after another.

➤ N-grams

N-gram feature selection is a series of words given length n , where n items are read as each word for a given instance. If $n=1$, then it is unigram, $n=2$ is bigram, followed by $n=3$ for trigrams [56]. For our current project, n -grams with different sizes were applied to our dataset.

For example, in the case of our crime dataset, if a comment or status update is written as: *"I hate thieves wish they die"*.

The unigrams, bigrams and trigrams of this post are:

Unigrams: {'I', 'hate', 'thieves', 'wish', 'they', 'die'}

Bigrams: {'I hate', 'hate thieves', 'thieves wish', 'wish they', 'they die',}

Trigrams: {'I hate thieves', 'thieves wish they', 'they die'}

In this case, the use of n -grams can help detect the sentiment of a sentence. For example, in the above sentence, 'I hate' can be reviewed to be a negative comment. For our dataset, bigrams seemed to be effective when compared to unigrams and trigrams. This step raised accuracy of the classifiers used in our project. Many experiments in past literature have proved that N -grams improve the superiority of feature sets [81].

3.4 Data Labelling

This is a step where the data is trained and made ready for the classifier. The data is the fuel for the classifier; it will be fed to the algorithm for learning purpose. Data was labelled according to Ekman's emotions theory; anger, fear, happiness, sadness, disgust

and surprise. These emotions capture all possible emotions that could be available. To extend the normal polarity (of positive, negative, neutral) used in most previous studies; we assigned our data as follows: *happiness* was assigned to positive class; *surprise* was assigned to neutral class while *anger*, *fear*, *sadness* and *disgust* were assigned to the negative class. For this task, we engaged a focus group of 8 postgraduate students (the same group that did the dataset translation). The reason for choosing a focus group was so that we capture average opinion and sentiment of people, and also to avoid unfairness by the main researcher. Again, by using a focus group, the noise and inaccuracies in labelling can be minimized. The individuals received no training, though they were given some guidelines to look for and samples of sentences to illustrate the kind of annotations required. The labelled data was prepared over a timeframe of 2 weeks. It was difficult to label this dataset because some of the comments were unclear as to which polarity they fall in. In this case we ended up labelling them as neutral.

3.4.1 Applying Polarity Score

For sentiment detection of crime news, we followed the dictionary based approach that uses Bag of Words (BOW) technique for text mining. This method is based on J. Bean for his implementation on Twitter sentiment for airline companies [82]. BOW is a technique used to convert text documents into a vector describing its features. After matching the different categories, we identified which bag a certain block of data (test data) belongs to. To build the polarity dictionary, we compiled six words from the Ekman's basic emotions theory (disgust, happiness, surprise, sadness, fear, anger). These were further categorised into 3 groups; positive (happiness), neutral (surprise) and negative (sadness, fear, anger, disgust).

The dictionaries for the positive, neutral and negative words were then created. Thereafter, vectors were formed to count each word. For this method, we used a statistical method known as td-idf (term frequency and inverse document frequency). We then applied an algorithm to compute the sentiment score of the document. The sentiment polarity score was assigned as follows: positive words were assigned as (1); neutral assigned as (0) and negative assigned (-1). Here, we are considering an assumption that if the score of the document is 0, then the label assigned will be neutral,

if 1, it will assign positive, hence if -1 then negative. The following are steps followed to consider polarity in a sentence:

1. Positive, negative and neutral polarity of each word is calculated.
2. Add the positive polarities of words in the comment gives the positive polarity of the whole comment.
3. In the same way, the negative polarities of all the words gives the negative polarity of the comment.
4. Adding neutral polarities of all the words in the comment gives the neutral polarity of the comment.
5. The comment is then given positive or negative or neutral, depending on which polarity is greater than the other.

3.5 Building Classifiers

This section covered the third objective which was: to train and test the data using supervised learning techniques, Naïve Bayes, Random Forest, K-NN and J48. We needed to train our data before feeding it to the classifiers, so that whenever we input new data, results can be easily predicted. Our approach adopted the supervised method for building the classifiers. In this approach, the training dataset is made up of data occurrences with the labels specified for them. At the start, the classifiers learn the data in a supervised manner. A model is built from the training data set which value of the class label is known. Once the model is built, it is then used to label the sentences on the test set automatically as opposed to human labelling which was done on training data. For text classification, mostly NB, SVM and Random Forest have been used and perform well in text classification. For our method, we used Naïve Bayes, K-NN, J48 and Random Forest to build our model because they are good when used in classification. NB is simpler and fast to forecast the class of the test data set. It also performs well in a multi-class prediction hence its use for this study. Naïve Bayes has got limitations though, e.g. it does not consider the sequence of words and requires higher computational effort. The reason for choosing Random Forest is because it is considered best for sentiment analysis and very simple to understand [64], [83]. J48 algorithm was chosen because it has been considered as the most efficient machine learning algorithm for prediction of crime data [63]. K-NN was used because it is easy to use and implement. One weakness

of K-NN is its large storage requirements. However, these limitations would not affect the use of these chosen classifiers in this study because it is being used only to augment the classification of data only.

The classifiers were implemented and tested using the Weka tool. The steps for building up the machine learning classifiers are shown in Figure 3.4. Once the model is built with training data, the model is applied to test data to predict the target values. Our train data consists of crime data, which will be trained with labels. The test data which contains new crime data (with no labels), consists of data which the model has not seen before, hence used to evaluate the model.

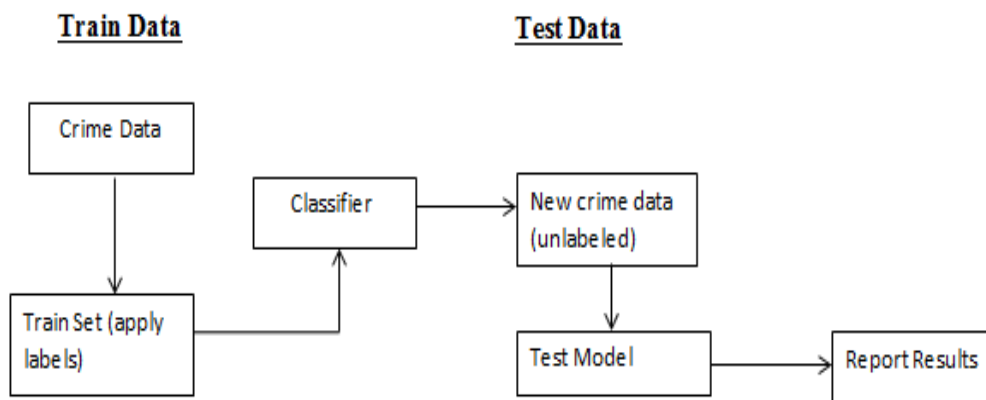


Figure 3.4: The steps for setting up machine learning classifier

3.5.1 The Prediction Model

This part plays a very important role in sentiment classification. The prediction model in Figure 3.5 labels the data and trains the classifiers. With the trained model, the test set predicts the instance or sentiments and set the label to the uncategorized data. An assumption was made that happiness belong to the positive class, surprise belongs to the neutral class and sadness, anger, fear, disgust make the negative class. Our aim here is to predict all the above classes. So the model was trained using 3 classes of positive, negative and neutral. During prediction, when a test instance is applied, the classifier outputs any of the three labels [0, 1, -1]. The instances which are classified as negative are labelled as -1, positive given 1 and neutral given 0.

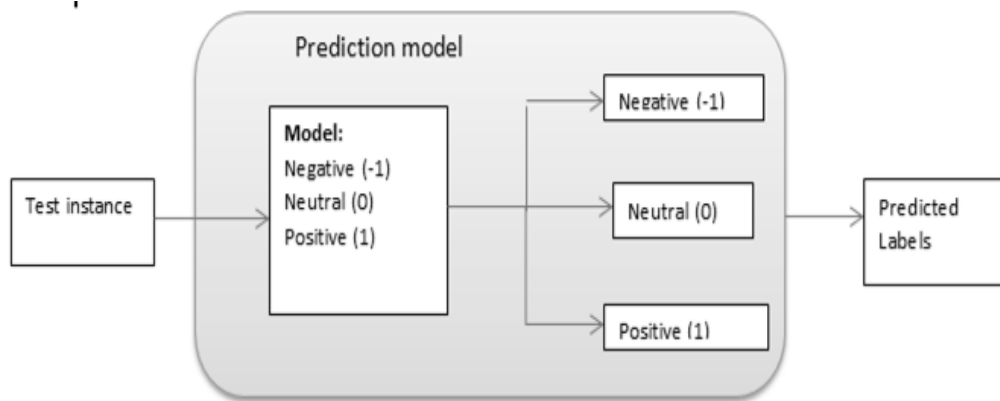


Figure 3.5: The prediction model

3.6 Testing the Classifiers

We used the train and test options in Weka for testing our crime dataset. The data was divided into training (given 70%) and test set (30%). This splitting was done such that the sentiment label spreading in the actual train-set was properly maintained. For example, if the train set has labels (happiness, sadness, anger, fear), the test set should have exactly the same labels. There is no way the test set could introduce a new label which was not mentioned in the train set. This process of testing is iterative, we had to loop around testing the four chosen algorithms until we have exhausted all of them.

3.7 System Evaluation

This section is in line with the 4th objective which was: to evaluate the supervised learning techniques based on accuracy, f-measure, Mean Absolute Error (MAE), recall, Kappa Statistics and precision. Evaluation is testing for accuracy of algorithms or model used for carrying a particular study. The accuracy of a sentiment analysis system is how well it is in agreement with human verdict. More sophisticated measures have been implemented but evaluation of sentiment analysis systems still remains a complex matter [19]. We evaluated all four classifiers (NB, J48, Random Forest, K-NN). The results for these classifiers were discussed in Chapter 4- Results and Discussion, which follows next. The classification results were then compared with the Police Report. This answers the last objective which was: to compare the sentiment classification results with the annual report (PCR). This was also covered in Chapter 4- Results and Discussion.

3.8 Summary of Workflow

In this section, we described the overall workflow of the whole sentiment analysis approach and the comparison with the annual PCR. This was done to relate the sentiments observed on Facebook with the PCR.

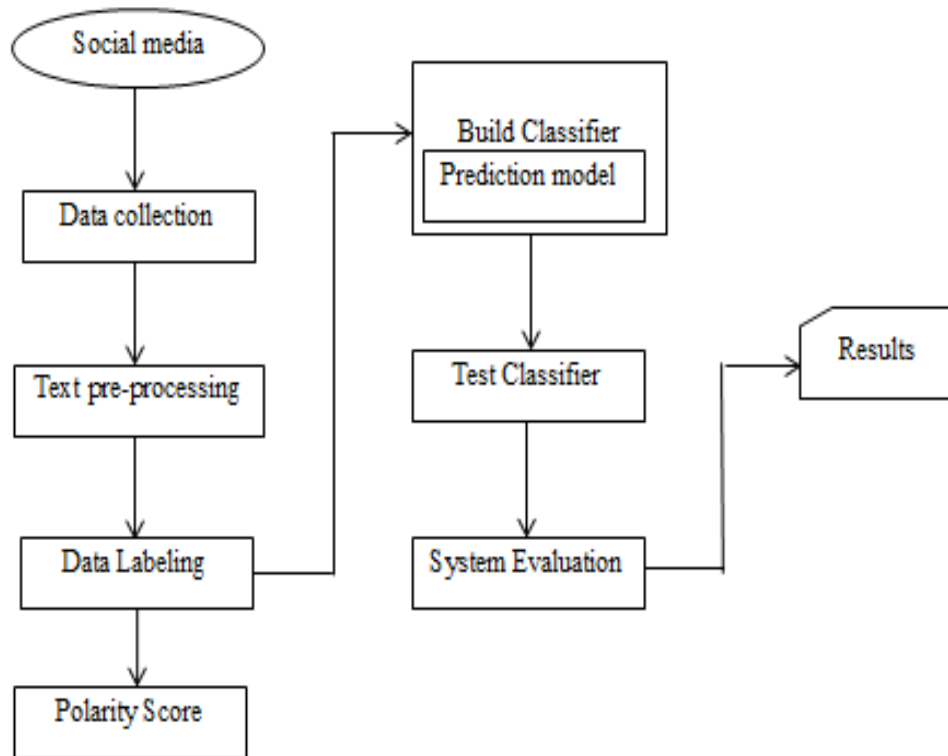


Figure 3.6: The overall architecture for our methodology

The process is presented in Figure 3.6 and is summarised by these points:

1. Collection of crime data from Facebook
2. Pre-processing of the data to remove noise and to select features
3. Labelling of the data and applying polarity score to data
4. Train and testing of the model on crime dataset
5. Evaluate the model
6. Comparison of sentiment classification with the Police Crime Statistics Report
7. Discussion of the results

CHAPTER 4: Results and Discussion

4.1 Introduction

In this chapter we started by presenting and discussing the results of the prediction model for crime news sentiment analysis component, and also compared our results with previous literature. About 540 comments and status updates were used for classification by Naïve Bayes, J48, K-NN and Random Forest. We established the accuracy and performance of each classifier in order to select the one with the highest accuracy and compare our results with the previous studies. During the experiments, we trained and tested our model. Training data is data which the model is built on, and test data verifies the performance of the model. The experiments were run on Waikato Environment for Knowledge Analysis (WEKA) version 3.8.0 tool. WEKA is a framework for model training, classification and prediction. It emanated through the perceived need for a unified workbench that would allow researchers easy access to techniques in machine learning [84]. In the second part of this chapter, we compared the results we got from our prediction model with the Police Crime Report of 2016, in order to see if there is a correlation between the two.

4.2 Evaluation Metrics for Sentiment Classification

Our study adopts the use of supervised learning approach. They are considered as white box classification models- meaning that they provide explanation for the classification, and can be used directly for decision making (they teach the system what to look for). We used NB, Random Forest, J48 (known as C4.5 in Weka) and K-NN (known as IBK in Weka) to train our model. The reason for using these classifiers is that they are good for text classification. Once the model was trained, we needed to measure its performance. We evaluated the accuracy, referring to the number of correctly classified instances over incorrectly classified instances; the precision, recall and F-measure or f1-score. We also evaluated the classifiers for the error rates using Kappa Statistics and the Mean Absolute Error (MAE). The algorithms were also used to find sentiments on Facebook data and to find a correlation between the Police crime report and the people's sentiments.

4.2.1 Accuracy

Our classifiers were first evaluated for accuracy; which is the percentage of correctly classified instances against all predicted instances. The formula for accuracy is depicted by:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Where TP is the true positive, TN is the true negatives, FP is the false positive and FN is the false negatives.

In our study, accuracy indicated the number of correct sentiments from all the sentiments applied to the dataset. An accuracy of 100% means that all the sentiments are predicted correctly [85]. Table 4.1 illustrates the accuracy scores for each classifier. Random Forest has the highest accuracy of 80%, followed by Naïve Bayes with 71.29%, and K-NN with 69.23%. J48 had the lowest accuracy of 65.92%.

Table 4.1: The accuracy of all four classifiers used for sentiment classification

Classifier	Accuracy
Random Forest	80.00%
K-NN	69.23%
Naïve Bayes	71.29%
J48	65.92%

4.2.2 Precision

We used precision to further understand the performance of the classifiers. Precision, also known as Positive Predictive Values (PPV), is the portion of data which is classified correctly. It calculates the portion of true positives over all positive predicted instances. The higher the fraction, the more precise the classifier is [85]. The formula is denoted by:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

Where TP= true positive, FP= false positive

According to Table 4.2, the precision for Random Forest was 0.800; K-NN was 0.691, Naïve Bayes 0.722 and J48 was 0.677. Random Forest had a higher precision more than others, and J48 had the lowest precision.

Table 4.2: The precision for each classifier

Classifier	Precision
Random Forest	0.800
K-NN	0.691
Naïve Bayes	0.722
J48	0.677

4.2.3 Recall

Recall is the true positive rate, which is the fraction of items that are actually positive and were predicted positive. The formula for recall was shown by:

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

Where TP = True Positive, while FN = False Negative.

Recall is the ratio of the number of positive examples classified over all the positive examples. As presented in Table 4.3, Random Forest had a recall rate of 0.800, which is the highest amongst all the other classifiers. K-NN had a recall of 0.692; Naïve Bayes had 0.713 and J48 with 0.659. The lowest recall was recorded by J48.

Table 4.3: Recall percentage for each classifier

Classifier	Recall
Random Forest	0.800
K-NN	0.692
Naïve Bayes	0.713
J48	0.659

4.2.4 F-Measure

F-measure or F1-score combines recall and precision to provide for a single measurement. The formula for F-measure is:

$$F - Measure = \frac{2 \times precision \times recall}{precision + recall} \quad (6)$$

According to Table 4.4, Random Forest had the highest f-measure of 0.796, followed by Naïve Bayes with 0.713 and K-NN with 0.673. J48 had the lowest f-measure of 0.654.

Table 4.4: The f-measure for each classifier

Classifier	F-Measure
Random Forest	0.796
K-NN	0.673
Naïve Bayes	0.713
J48	0.654

4.2.5 Kappa Statistics

Kappa Statistics measures the agreement between items, i.e. between the classifications and true classes. The formula is adopted from [86], which is denoted as:

$$K = \frac{P(A) - P(E)}{1 - P(E)} \quad (7)$$

Where P (A) = percentage agreement and P (E) is the chance agreement. Kappa of 0 is doing better than chance hence kappa of 1 is a perfect agreement [87] [85]. Table 4.5 shows that Random Forest had the highest kappa of 0.654, followed by Naïve Bayes by 0.583 and J48 by 0.583. K-NN had the lowest agreement of 0.409.

Table 4.5: The kappa statistics for each classifier

Classifier	Kappa Statistics
Random Forest	0.654
K-NN	0.409
J48	0.583
Naïve Bayes	0.583

4.2.6 Mean Absolute Error (MAE)

MAE is used to measure how close forecasts or predictions are to the eventual outcomes [85]. MAE is adopted from [86], and the formula denoted by:

$$MAE = \frac{1}{N} \sum_{i=1}^n |fi - yi| = \frac{1}{n} \sum_{i=1}^n |e_i| \quad (8)$$

The MAE is an average of the absolute errors: $e_i = |fi - yi|$, where fi is the prediction and yi is the true value. From table 4.6, it shows that Random Forest had MAE of 0.027, K-NN with 0.031, NB with 0.122 and J48 with 0.159. According to literature, any MAE between 0 and 1 is accepted and the higher you go the worst it gets [85]. Lower values indicate that the forecasted values are better predictions than just using the last known target value [85]. From Table 4.6 it shows that Random Forest had a good MAE as compared to other classifiers.

Table 4.6: The MAE for each classifier

Classifier	Mean Absolute Error
Random Forest	0.027
K-NN	0.031
J48	0.150
Naïve Bayes	0.122

4.2.7 Summary of Evaluation Results

Table 4.7 summarises the accuracy, precision, recall, F-measure, kappa statistics and the MAE for all the four classifiers we used for our experiments. Overall, we observed that Random Forest had the highest accuracy of 80%, followed by Naïve Bayes with 71.29%, k-NN with 69.23% and J48 scored 65.92% respectively. Altogether Random Forest got the highest precision of 0.800, recall of 0.800, f-measure of 0.796, and kappa agreement of 0.654 and the mean error margins of 0.027. K-NN had an accuracy of 69.23% lower than Naïve Bayes at 71.29%, but the mean error for K-NN was better at 0.031 than that of NB with 0.150. We saw this again with K-NN; it had a higher accuracy than J48 at 65.92%, but the kappa agreement for J48 was much better at 0.583 than K-NN at 0.409.

Table 4.7: Summary of all the classifiers showing accuracy, precision, recall and f-measure, MAE and kappa statistics

Classifier	Accuracy	Precision	Recall	F-Measure	Kappa Statistics	Mean Absolute Error
Random Forest	80.00%	0.800	0.800	0.796	0.654	0.027
K-NN	69.23%	0.691	0.692	0.673	0.409	0.031
Naïve Bayes	71.29%	0.722	0.713	0.713	0.583	0.150
J48	65.92%	0.677	0.659	0.654	0.583	0.122

4.3 Discussion

In this section, we discuss our findings and further compared them with the related works done in the area of crime prediction. Weka was used for the implementation of all the four classifiers- Random Forest, Naïve Bayes, K-NN and J48. We tested our data using the train and test data.

4.3.1 Random Forest Classification

According to the experiments performed on the crime dataset, Random Forest performed with an accuracy of 80% better as compared to other classifiers. The results achieved by Random Forest are similar to a study by Gupte et al. [64], who got an accuracy of 80% for stock prediction using sentiment analysis. In view of sentiment analysis, Random Forest classifier has an upper hand with high accuracy and performance. This makes the classifier best fit for sentiment analysis. In another similar study to classify Twitter emotional content, Random Forest achieved an accuracy of 82.72% [83]. This margin is almost in the same range with our accuracy margin for Random Forest. In a research to predict credit risk, Random Forest did better than J48 [88]. This good performance is prevalent even on our study.

4.3.2 Naïve-Bayes Classification

In this study, Naïve Bayes achieved an accuracy of 71.29%. This classifier did better than J48 and K-NN. This is because Naïve Bayes is considered the simplest probability based

classifier. In a study by Iqbal et.al [13] to predict crime category, Naïve Bayes got an accuracy of 70.80%. This is nearer to the accuracy we got after applying Naïve Bayes to our dataset. Again, in another study, we see NB achieving an accuracy of 71.33% better than J48 and K-NN [89]. Although in most of the studies of classification, Naïve Bayes has been doing better than decision tree classifiers [39]. On contrast, with the above study, we saw Random Forest performing better than Naïve Bayes in our study.

4.3.3 KNN Classification

From our classification results, K-NN performed with an accuracy of 69.23%. K-NN outperformed J48 which is the lowest performing. Similar to our results, in a study by [90], K-NN outperformed J48. The only difference in their study is that K-NN also outperformed NB. This is not the case in our study. However, there is no solid reason as to why K-NN performs better than J48 and NB on this task. K-NN works by assigning equal weight to each attribute; this can cause confusion when there are many irrelevant attributes in the data. Again, since K-NN uses number of nearest neighbour k as one of the parameter in classifying an object, this value has the possibility to affect its performance. That is why it has been the worst performing classifier in the past literature [91].

4.3.4 J48 Classification

From our results, J48 got an accuracy of 65.92%. It was the lowest performing classifier in this current study. J48 is good with handling missing values attributes, but our dataset did not have a lot of missing values, hence why the low accuracy. It is not unusual that this classifier performs lower than others. In a study by Martínez-Cámara et al. [92], J48 got an accuracy of 68.20% lower than other classifiers- SVM, Naïve Bayes and K-NN, for classifying Spanish text. In the same study, NB, K-NN and SVM did better than J48. This is also the case in the current study.

To summarise the classification results, based on the benchmarks used to measure the algorithms in this study, all the 4 classifiers achieved better results than a baseline of 50% random accuracy as expected. The classifiers achieved different accuracies with a margin ranging from 65.23% to 80%. Random Forest did better than all the other classifiers, with the lowest error margin. Studies show that the performance of the Random Forest is always rising [64]. They are known to be one of the most efficient

classifiers, hence why their good performance on our dataset. The sentiment analysis model results show that the model can go up to 80% of accuracy. With regards to Naive Bayes being outperformed by Random Forest, this could be attributed to that NB builds words upon probabilities. Each word is assumed to be independent. This can have a negative impact on the results because some words maybe falsely classified. Although K-NN is categorised as the lazy classifier, we saw it doing better than J48. The reason for poor results by J48 could be because of the feature selection methods we used in our datasets. Feature selection methods have an influence on the accuracy of classifiers. The additional features of J48 are accounting for missing values, but for our dataset, we did not have a lot of missing values. J48 is suitable for handling very complex classification problems and can be used for making critical decision. We believe our dataset was not so complex, hence why poor performance by J48.

4.4 Sentiment Results and Police Crime report (PCR)

We analysed social media data from Facebook and compared it with the crime report of year 2016. The purpose was to see if sentiments have a relation with crime statistics. Below we discuss how we achieved this task by explaining our prediction model (sentiment analysis component) and the actual PCR. The PCR gives us the trustworthiness of our model and verifies our prediction.

4.4.1 Sentiment Data

During this phase, we trained our data with different instances. Whenever a test instance was applied, the classifier outputs any of these labels; negative (anger, sadness, disgust, fear) positive (happiness) and neutral (surprise). We used the confusion matrix to further understand our sentiments. A confusion matrix contains information about actual and predicted classifications for any given classification algorithm. From the above experiments done on Section 4.2, Random Forest came out as the best model to predict sentiments; therefore we used its confusion matrix to see how best it can predict values as shown in Table 4.8.

Table 4.8: Confusion Matrix

a	b	c	d	e	f	←Classified as
84	7	0	1	3	4	a= happiness
7	106	0	2	0	2	b=sadness
8	0	51	2	0	4	c= fear
9	8	6	72	5	3	d= disgust
0	6	4	3	62	6	e= surprise
5	7	3	0	5	57	f= anger

From Table 4.8, 'a' class (happiness) had 99 instances in total. 84 instances were correctly classified into class 'a'. 7 instances were incorrectly classified under class 'b', 1 incorrectly classified under class 'd', 3 incorrectly classified as 'f'. For class 'b' (sadness), there were 117 instances in which 106 were correctly classified. 7 instances were incorrectly classified under 'a', 2 incorrectly classified under class 'd' and 2 incorrectly classified under class 'f'. We can observe that even with the remaining classes (c, d, e, f), the Random Forest is doing well in terms of classifying our data correctly. From Table 4.8, we wanted to see the percentage for each class to determine which class has the highest sentiments.

We calculated the percentages of the confusion matrix for each sentiment label. We achieved this by dividing the total number of correctly classified instances for class 'a' with the total number of all instances. We then multiplied this number by 100 to come up with a percentage. We repeated the same steps for classes' b, c, d, e, and f. We then populated a Column chart using Microsoft Excel 2010 application, as shown in Figure 4.1. From this chart, sadness seems to have the highest percentage of correctly classified instances by 90.6%, followed by happiness with 84.4%, and fear with 78.5%. Surprise and anger were correctly predicted by 76.5% and 74.0% respectively. The sentiment with the lowest correctly classified was disgust with 69.9%.

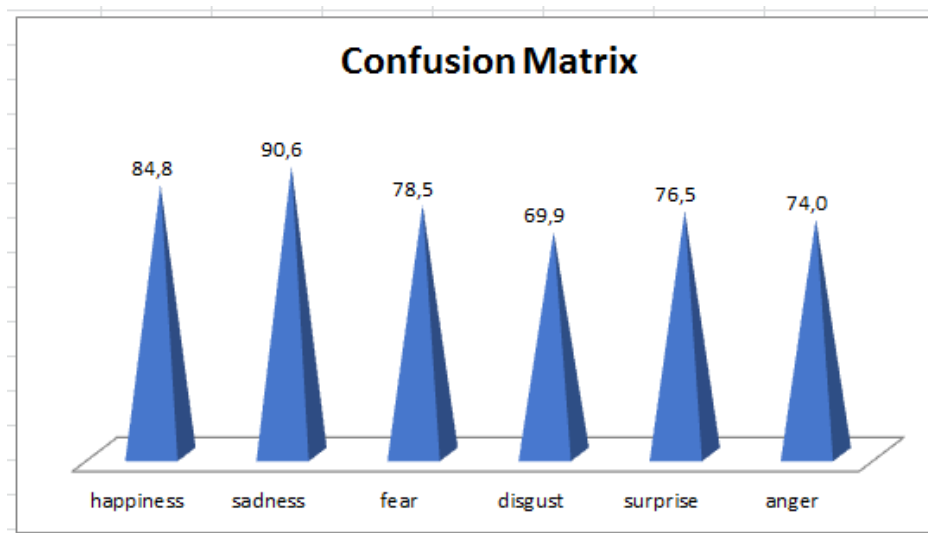


Figure 4.1: Percentage comparison of the sentiments

From Figure 4.1, we can conclude that if most of the sentiments are classified as negative; it means for that period most people were very uneasy and the crime statistics for the year 2016 are likely to go up. However, if sentiments are mostly classified as positive, this means crime statistics are likely to go down. Hence if neutral; no change is expected in crime rising or going down. This means that people's moods were stable. Our prediction model showed that most comments and status updates were negative. This implied that since the comments were negative, the results in the crime statistics are likely to go up.

4.4.2 Police Crime Report

The data presented in Table 4.9 was adopted from the Botswana Police Services Annual Crime Statistics Report of year 2016 [93]. The data used to compile this report was from January to December 2016, but as for the data from 2015, it was only included in the report for comparison purposes. This is the same period as the Facebook data was collected (from January to December 2016). Table 4.9 presents crime distributions by the Police districts in Botswana. Our study was mainly concerned with the district of Gaborone, which covers Police district No.3 (Gaborone Central and East) and District no 13 (Mogoditshane and Gaborone West) as highlighted. From this report, it showed that there has been an increase in crime by 1.6% in 2016 for Gaborone Central and East as compared to year 2015. For Mogoditshane district, crime went down by -1.2% from

2015. To determine if these statistics have gone up or down, we added the two numbers to calculate the average. So, we added 1.6% and -1.2%, which gives us 0.4%. Now, we conclude that crime for the Gaborone region has gone up by 0.4% in year 2016. This increase in crime could be attributed to that there are more economic activities in the city hence the number of crime incidents tends to be higher.

Table 4.9: Crime Distribution by districts [93]

Police Districts	Year 2015	Year 2016	% change
No.1- Kutlwano (Francistown South)	16932	18105	6.9
No.2- Serowe	9890	10298	4.1
No. 3 – Gaborone (Central & East)	23234	23603	1.6
No.4- Lobatse	6331	6359	0.4
No.5- Maun	7829	8205	4.8
No.6- Gantsi	2298	2211	-3.8
No.7- Kasane	2900	3257	12.3
No.8- Letlhakane	3897	3708	-4.8
No.9- Tsabong	2293	2489	8.5
No.10- Selibe Phikwe	7109	7179	0.9
No.11- Molepolole	10914	11929	9.3
No.12- Mochudi	4568	4610	0.9
No.13- Mogoditshane & Gabs/West	28381	28039	-1.2
NO. 14-Kanye	8515	8133	-4.5
No.15- Francistown (Francistown North)	18439	23737	28.7
No.16- Mahalapye	6520	7351	12.7

4.4.3 Sentiment Analysis vs. Police Crime Report

To assess the prediction power of our model, the PCR (discussed in section 4.4.2) and the sentiment analysis component results (discussed in section 4.4.1) were compared. It was difficult to compare the data because we did not have a lot of features to use for comparison. Instead we opted to use the confusion matrix results for the sentiments classification. From the matrix, it showed that *sadness* category had the highest percentage, meaning that the public's sentiments were classified as negative. We then checked with the PCR to relate these two. We observed that the crime statistics for Gaborone district went up. This could be that people were not happy with the services they were given by the police or they felt vulnerable to being attacked at any time, hence feeling unsafe. It could also mean that people were victims of crime, hence why a higher number in crime statistics. In conclusion we can say that people's sentiments

from social media have an impact in the outcome of the crime statistics being high or low. It is evident that our model is reliable and good with prediction. Law enforcing agencies can take great advantage by using our model for prediction.

4.5 Threats to Validity

Throughout this research we experienced some limitations which could be threats to our results. The limitations are discussed below:

The Police depend on the public to report cases, but not all of these cases are reported to them, therefore some cases may remain unknown. This could mean that there are some cases which are not captured on the Police files. This therefore can account for variation in the results. Unfortunately this phenomenon was beyond our control. Another contributing threat to validity to our research results was that some people may report false information on social media. These were used in coming up with the results and might not reflect the exact true results. Again, not everyone in Botswana is active on Facebook, so we may only find the ones who are more active and more likely to expose their thoughts and feelings. Most comments show negativity but it does not mean there are no positive aspects concerning crime issues.

4.6 Summary

This chapter discussed the evaluation metrics used to measure the 4 classifiers we used on our dataset. Results showed that Random Forest outperformed all the classifiers. We used the confusion matrix for Random Forest to determine which sentiments have the highest percentage of correctly classified instances, and to which sentiment label they fall in. After comparing the matrix results with the PCR, we conclude that sentiments have the power to determine if crime statistics will go up or down

CHAPTER 5: Conclusion and Future Work

5.1 Introduction

This section discusses the contributions of this research work, emphasising on the research questions to answer the objectives; and recommendations for future work.

5.2 Outline of this research

Our study was motivated by a research gap indicated by [13] which stated that there is a need for further research to apply other classification algorithms on crime data. The main objective of this study was to apply sentiment analysis of social media data by using supervised learning methods in order to predict the moods of the public about crime, using Ekman's six basic emotions theory. Our aim was to build a model to predict news polarity which may have a correlation to the Police statistics report. The predictive model classifies sentiments as happiness, sadness, fear, anger, disgust and surprise. The main aim in analysing sentiments is to understand why people are posting. Moods observed from social media may be used to correlate with annual statistical police report. By analysing these sentiments from conversations about crime, Police can come to the right conclusions about conversations. A positive review in social media may mean that the public is satisfied with the services that the Police forces are offering them including punishing law breakers and culprits. A negative review may mean the public is not satisfied. To achieve our main objective, there are some steps which we followed: data collection, pre-processing of data, labelling of data, building and testing of classifiers. The model which we built was then evaluated using some performance measures.

The main contribution of this research work is the crime dataset for Gaborone region which this thesis has shown its usefulness. Again a workflow for the sentiment analysis process has been demonstrated which shows the integration between qualitative research and the computational techniques.

5.2.1 Reflections on the Objectives and Research Questions

In order to achieve this research work, we came up with objectives and research questions that were up for investigation during the inception of this research. The main

aim was to apply sentiment analysis of social media using supervised learning approaches to predict the moods of the public, and to correlate these moods with the Police Statistics Report of 2016. The objectives and their matching questions are explained below:

1. To collect data from social media

This objective has been addressed during our research undertaking. The crime dataset used in this study was collected from Facebook. We used a software tool called Netlytic to help us to collect our data. The data was collected from January 2016 to December 2016. This data included comments and status updates on any issues regarding crime. We collected this data from different Facebook groups- Fight Crime in Gaborone, The Voice Online Newspaper, Women and Men Against Children Abuse and Botswana Online Magazine.

2. To manually perform pre-processing of data in order to remove noisy and irrelevant data.

Research Question 1- How can crime data be pre-processed prior to sentiment analysis?

This objective and research question has been addressed. The text data we collected was unstructured, so we pre-processed the data using different techniques such as stemming, stop-words removal, and case correction. We started by cleaning the data by removing the links, source, type, publication date which did not have any meaning to our data. We removed all comments and status updates which were objective. All unnecessary data such as abbreviations, hashtags, and emoticons were removed from our dataset. For automatic pre-processing of data, Weka was used to perform feature selection.

3. To train and test the data using supervised learning techniques; Naïve Bayes, Random Forest, K-NN and J48

Research Question 2- Which supervised methods exist for sentiment classification of social media data and any data in general?

This objective and research question has been addressed. We used Weka software to train our data and build the sentiment analysis model. Sentiment classification techniques such as Naïve Bayes, Random Forest, K-NN and J48 were used to train this model. For testing the model, we used Weka tool. We used train and test options for splitting our data, giving the train data 70% and test data 30%.

4. To evaluate the supervised learning techniques.

Research Question 3- Which sentiment classification schemes can be used for evaluation of social media crime data?

This objective and research question has been addressed. Evaluation is measuring (using statistics) the performance of the classifiers. We evaluated the classifiers; Naïve Bayes, Random Forest, K-NN and J48 using accuracy, f-measure, precision, Mean Absolute Error, recall, and Kappa Statistics, using Weka. We then compared the results of the classifiers performance.

5. To compare the sentiment classification results with the annual Police Crime Report

We computed the confusion matrix statistics and the column chart for sentiment classification using Microsoft Office Excel. We then compared the results with the PCR. A relationship between the social media sentiments and the PCR was seen. This last objective was also achieved.

5.3 Conclusion

Through social media, people obtain information about a decision by communicating with one another; hence social media has become an important aspect of decision making. This current work used Facebook comments and status updates to analyse peoples sentiments regarding crime. Data was then pre-processed, but took long to do because of the complexity in the data. The comments and status updates were used to predict the sentiments regarding crime. Sentiments from social media have the power to predict outcomes. Previous works on sentiment analysis used polarity as positive, negative and neutral. This study extended this polarity by including the work of Ekman's basic emotions, by categorising crime data into 6 categories of; happiness, fear, anger,

sadness, surprise and disgust. We call this kind of data classification multi-classification. These emotions were further categorised as: positive (happiness), neutral (surprise) and negative (sadness, fear, anger, disgust). We believe that by using Ekman's [44] work, the sentences were categorised with all the possible emotions that could be found. Also, previous works on crime prediction and mining used mostly unsupervised methods such as regression and clustering methods, which proved to be inadequate. Our study used supervised learning approaches. We managed to use these supervised methods to build a model to predict emotions. The classifiers did a good job when classifying data. When given any data the model will be able to assign polarity to further predict emotions.

Our study made a conclusion that there is a correlation between the Facebook emotions with the actual Police crime report. This kind of study can help law enforcement sectors in decision making, improving their services and sensitising the public on crime.

5.4 Future Work

This section provides a beam of light in direction of possible future avenues of research.

Our study only uses one platform (Facebook) as a data source. This could be a limitation because not all platforms are included in making conclusions. Using different social media outlets like MySpace, YouTube and Twitter would help improve accuracy. For building the predictive model, only one dataset was used. We believe that in the future, using a number of datasets could help when it comes to the testing and comparing of different classifiers. We are also considering applying different feature selection methods to the different datasets and compare their accuracies. This will give an overall conclusion on different kinds of data hence biasness decreased. Also, our crime dataset could be increased so as to improve accuracy. Most classifiers give better results when the dataset is big. In this study we only collected and used text on social media. To extend this, we will analyse crime generated content other than text, we will use for example, images and videos. These also carry very rich information that could be useful for crime analysis and prediction.

We saw a correlation of sentiments from Facebook with the PCR. For future research, there is need for prediction in crime patterns and trends, by concentrating patterns

from numerous aspects. As an extension, this study will investigate how to spread over sentiment analysis to Facebook data, by observing how public opinion changes and how a particular issue/topic advances over a certain period. For example, during a national crisis like a public strike, the Police might get interested about topics discussed during that time. Again, our study collected data for a year and we had to wait for a Police yearly report in order to verify our results. Instead of waiting this long for a report, data can be analysed monthly and compared with the monthly crime statistics report. This way it would be helpful to pin-point important issues and keep track of any changes in the data.

References

- [1] Z. Kechaou, M. Ammar, and M. Alimi A, "Improving e-learning with sentiment analysis of users opinions. Global Engineering Education Conference (EDUCON)-- Learning Environments and," in *IEEE Global Engineering Education Conference (EDUCON) – "Learning Environments and Ecosystems in Engineering Education,"* 2011, pp. 1032–1038.
- [2] E. Kontopoulos, C. Berberidis, T. Dergiades, and N. Bassiliades, "Ontology-based sentiment analysis of twitter posts," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4065–4074, 2013.
- [3] K. Ahmed, N. El Tazi, and A. H. Hossny, "Sentiment Analysis over Social Networks: An Overview," *2015 IEEE Int. Conf. Syst. Man, Cybern.*, no. April 2016, pp. 2174–2179, 2015.
- [4] R. A. Bolla, "Crime pattern detection using online social media," M.S Thesis, Missouri University of Science and Technology, 2014.
- [5] C. S. Sadhana and B. K. Sangareddy, "Survey on Predicting Crime Using Twitter Sentiment and Weather Data," in *National Conference on Emerging Trends and Advances in Information Technology*, 2016.
- [6] O. G. N and J. Alhasan, "Analytical Study of Some Selected Classification Algorithms in WEKA Using Real Crime Data," vol. 4, no. 12, pp. 44–48, 2015.
- [7] W. Y. Chong, B. Selvaretnam, and L.-K. Soon, "Natural Language Processing for Sentiment Analysis: An Exploratory Analysis on Tweets," *2014 4th Int. Conf. Artif. Intell. with Appl. Eng. Technol.*, pp. 212–217, 2014.
- [8] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 15–21, 2013.
- [9] J. Azeez and D. J. Aravindhar, "Hybrid approach to crime prediction using deep learning," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2015, pp. 1701–1710.
- [10] L. McClendon and N. Meghanathan, "Using Machine Learning Algorithms to

- Analyse Crime Data,” *Mach. Learn. Appl. An Int. J.*, vol. 2, no. 1, pp. 1–12, 2015.
- [11] M. Nouh, J. R. C. Nurse, and M. Goldsmith, “Towards designing a multipurpose cybercrime intelligence framework,” in *Proceedings - 2016 European Intelligence and Security Informatics Conference, EISIC 2016*, 2017, pp. 60–67.
 - [12] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, “Analyzing spammers’ social networks for fun and profit: a case study of cyber criminal ecosystem on twitter,” *WWW ’12 Proc. 21st Int. Conf. World Wide Web*, pp. 71–80, 2012.
 - [13] R. Iqbal, M. Azrifah, A. Mustapha, P. Panahy, and N. Khanahmadliravi, “An Experimental Study of Classification Algorithms for Crime Prediction,” *Indian J. Sci. Technol.*, vol. 6, no. 3, pp. 4219–4225, 2016.
 - [14] B. Liu, “Sentiment Analysis and Subjectivity,” *Handb. Nat. Lang. Process.*, no. 1, pp. 1–38, 2010.
 - [15] S. Baliyan, “The effectiveness of information and communication technology curriculum: a case of private senior secondary schools in Botswana,” *Int. J. Sci. Res.*, vol. 5, no. September, pp. 197–206, 2012.
 - [16] M. Rambocas and J. Gama, “Marketing Research : The Role of Sentiment Analysis,” Trinidad and Tobago, 2013.
 - [17] L. Barbosa and J. Feng, “Robust Sentiment Detection on Twitter from Biased and Noisy Data,” *Coling*, no. August, pp. 36–44, 2010.
 - [18] H. J. Kaur and R. Kumar, “Sentiment analysis from social media in crisis situations,” in *International Conference on Computing, Communication & Automation*, 2015, pp. 251–256.
 - [19] M. Shinde-pawar, “Formation of Smart Sentiment Analysis Technique for Big Data,” *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 2, no. 12, pp. 7481–7488, 2014.
 - [20] B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis,” *Found. Trends Inf. Retr.*, vol. 1, no. 2, pp. 91–231, 2006.

- [21] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods," in *Proceedings of the first ACM conference on Online social networks - COSN '13*, 2013, pp. 27–38.
- [22] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences," *Proceedings of the 2003 conference on Empirical methods in natural language processing*. pp. 129–136, 2003.
- [23] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," *Processing*, vol. 150, no. 12, pp. 1–6, 2009.
- [24] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in *In Proceedings of the Seventh Conference on International Language Resources and Evaluation*, 2010, pp. 1320–1326.
- [25] G. Tobias and R. Johansson, "Sentiment Analysis of Microblogs," M.S. Thesis, University of Gothenburg, 2013.
- [26] R. Singh and R. Kaur, "Sentiment Analysis on Social Media and Online Review," *Int. J. Comput. Appl.*, vol. 121, no. 20, pp. 44–49, 2015.
- [27] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.
- [28] J. Bollen, A. Pepe, and H. Mao, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," 2009, pp. 450–453.
- [29] M. J. Paul and M. Dredze, "You are what you Tweet: Analyzing Twitter for public health," *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. pp. 265–272, 2011.
- [30] J. Ortigosa-Hernández, J. D. Rodríguez, L. Alzate, M. Lucania, I. Inza, and J. A. Lozano, "Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers," *Neurocomputing*, vol. 92, pp. 98–115, 2012.
- [31] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welp, "Predicting elections with

- Twitter: What 140 characters reveal about political sentiment,” *Proc. Fourth Int. AAAI Conf. Weblogs Soc. Media*, pp. 178–185, 2010.
- [32] A. Aue and M. Gamon, “Customizing Sentiment Classifiers to New Domains: A Case Study,” *Proc. Recent Adv. Nat. Lang. Process.*, vol. 3, no. 3, pp. 16–18, 2005.
 - [33] D. Szerszen, “Sentiment Classification in Social Media An Analysis of Methods and the Impact of,” M.S. Thesis, KTH Royal Institute of Technology, 2016.
 - [34] P. D. Turney, “Thumbs up or thumbs down? Semantic Orientation applied to Unsupervised Classification of Reviews,” *Proc. 40th Annu. Meet. Assoc. Comput. Linguist.*, no. July, pp. 417–424, 2002.
 - [35] J. D. S Wijeratne Derek Doran, Amit Sheth, “Analyzing the Social Media Footprint of Street Gangs,” 2015.
 - [36] D. Décary-hétu and C. Morselli, “Gang Presence in Social Network Sites,” vol. 5, no. December, pp. 876–890, 2011.
 - [37] A. Hernández, V. Sanchez, G. Sánchez, and H. Pérez, “Security Attack Prediction Based on User Sentiment Analysis of Twitter Data,” pp. 1–8.
 - [38] U. Saeed, M. Sarim, A. Usmani, A. Mukhtar, A. B. Shaikh, and S. K. Raffat, “Application of Machine learning Algorithms in Crime Classification and Classification Rule Mining,” vol. 4, no. 3, pp. 106–114, 2015.
 - [39] I. Journal, D. C. Technology, S. Shojaee, U. Putra, A. Mustapha, U. Tun, H. Onn, F. Sidi, U. Putra, M. A. Jabar, and U. Putra, “A Study on Classification Learning Algorithms to Predict Crime Status,” vol. 7, no. May, 2013.
 - [40] G. M. Tolan and O. S. Soliman, “An Experimental Study of Classification Algorithms for Terrorism Prediction,” vol. 1, no. 2, 2015.
 - [41] T. Almanie, R. Mirza, and E. Lor, “Crime Prediction Based on Crime Types and using Spatial and Temporal Criminal Hotspots,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 4, pp. 1–19, 2015.
 - [42] Statistics Botswana, “Crime Statistics Report,” Gaborone, 2016.

- [43] pana-press, "Botswana Social media offenders irk Botswana Police," 2014. [Online]. Available: <http://www.panapress.com/Botswana--Social-media-offenders-irk-Botswana-Police--12-923298-42-lang2-index.html>. [Accessed: 01-Jul-2017].
- [44] P. Ekman, "An-Argument-For-Basic-Emotions.pdf," *Cogn. Emot.*, vol. 6, no. 3–4, pp. 169–200, 1992.
- [45] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the 2008 ACM symposium on Applied computing - SAC '08*, 2008, p. 1556.
- [46] S. Aman and S. Szpakowicz, "Identifying Expressions of Emotion in Text," *Text, Speech and Dialogue*, vol. 4629, pp. 196–205, 2007.
- [47] S. Chaffar and D. Inkpen, "Using a heterogeneous dataset for emotion analysis in text," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011, vol. 6657 LNAI, pp. 62–67.
- [48] H. J. Do, C. Lim, Y. J. Kim, and H. Choi, "Analyzing Emotions in Twitter During a Crisis : A Case Study of the 2015 Middle East Respiratory Syndrome Outbreak in Korea," 2016.
- [49] M. Kearn, "Machine Learning is for Muggles too!," *Microsoft Developer*, 2016. [Online]. Available: <https://blogs.msdn.microsoft.com/martinkearn/2016/03/01/machine-learning-is-for-muggles-too/>.
- [50] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014.
- [51] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2002, pp. 79–86.
- [52] C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno, and J. Caro, "Sentiment analysis

- of Facebook statuses using Naive Bayes Classifier for language learning,” *IISA 2013 - 4th Int. Conf. Information, Intell. Syst. Appl.*, no. July, pp. 198–205, 2013.
- [53] D. Song, H. Lin, and Z. Yang, “Opinion mining in e-learning system,” in *Proceedings - 2007 IFIP International Conference on Network and Parallel Computing Workshops, NPC 2007*, 2007, pp. 788–792.
- [54] G. Li, “Application of a clustering method on sentiment analysis,” *Journl Inf. Sci.*, vol. 38, no. 2, pp. 127–139, 2012.
- [55] J. Bukovina, “Journal of Behavioral and Experimental Finance Social media big data and capital markets — An overview,” *J. Behav. Exp. Financ.*, vol. 11, pp. 18–26, 2016.
- [56] S. Charan and R. Gangireddy, “Supervised Learning for Multi-Domain Text Classification,” San Jose State University, 2016.
- [57] H. V. Thakkar, “Twitter Sentiment Analysis using Hybrid Naive Bayes,” M.Tech.Dissertation, Sardar Vallabhbhai National Institute of Technology, 2013.
- [58] F. Derroncourt, “A simple explanation of Naive Bayes Classification.” pp. 1–6, 2015.
- [59] Y. W. Syaifudin and D. Puspitasari, “Twitter Data Mining for Sentiment Analysis on Peoples Feedback Against Government Public Policy,” *Int. J. Sci. Technol.*, vol. 3, no. 1, pp. 110–122, 2017.
- [60] A. Nichat A and J. Raut A, “Analysis of Student Performance Using Data Mining Technique,” *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 5, no. 2, pp. 1302–1309, 2017.
- [61] D. Hongbo, *Data Mining Techniques and Applications- Introduction*, no. 4. Cengage Learning, 2010.
- [62] K. Rangra and K. L. Bansal, “Comparative Study of Data Mining Tools,” *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 4, no. 6, pp. 2277–128, 2014.

- [63] E. Ahishakiye, D. Taremwa, and E. Omulo, "Crime Prediction Using Decision Tree (J48) Classification Algorithm," *Int. J. Comput. Inf. Technol.*, vol. 06, no. 03, 2017.
- [64] A. Gupte, S. Joshi, P. Gadgul, and A. Kadam, "Comparative Study of Classification Algorithms used in Sentiment Analysis," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 5, pp. 6261–6264, 2014.
- [65] J. Gholap, "Performance Tuning of J48 Algorithm for Prediction of Soil Fertility," 2012.
- [66] Y. Li and B.-L. Lu, "Feature selection based on loss-margin of nearest neighbor classification," *Pattern Recognit.*, vol. 42, no. 9, pp. 1914–1921, 2009.
- [67] S. Deokar, "Weighted k nearest neighbor." p. 40, 2009.
- [68] M. Kordos, M. Blachnik, and D. Strzempa, "Do we need whatever more than k-NN?," in *International Conference on Artificial Intelligence and Software Computing*, 2010, vol. 6113, no. PART 1, pp. 414–421.
- [69] R. Bosch, "Sentiment Analysis: Incremental learning to build domain models," M.S. Thesis, Universitat Pompeu Fabra, 2013.
- [70] Y. Salman, "Twitter network analysis and visualisation II: NodeXL," 2014. [Online]. Available: <https://mashe.hawksey.info/2011/09/twitter-network-analysis-and-visualisation-ii-nodexl/>.
- [71] S. S. Baskar, L. Arockiam, and S. Charles, "A Systematic Approach on Data Pre-processing In Data Mining," *COMPUSOFT, A n Int. J. Adv. Comput. Technol.*, vol. 2, no. Xi, pp. 335–339, 2013.
- [72] E. Haddi, "Sentiment Analysis; Text Pre-processing, Reader Views and Cross Domains," Ph.D. Disertation, Brunel University, 2015.
- [73] S. Ramaswamy, "Visualization of the Sentiment of the Tweets.," M.S. Thesis, Graduate Faculty of North Carolina State University, 2011.
- [74] S. M. Mohammad, "Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text," in *Emotion Measurement*, 2016, pp. 201–237.

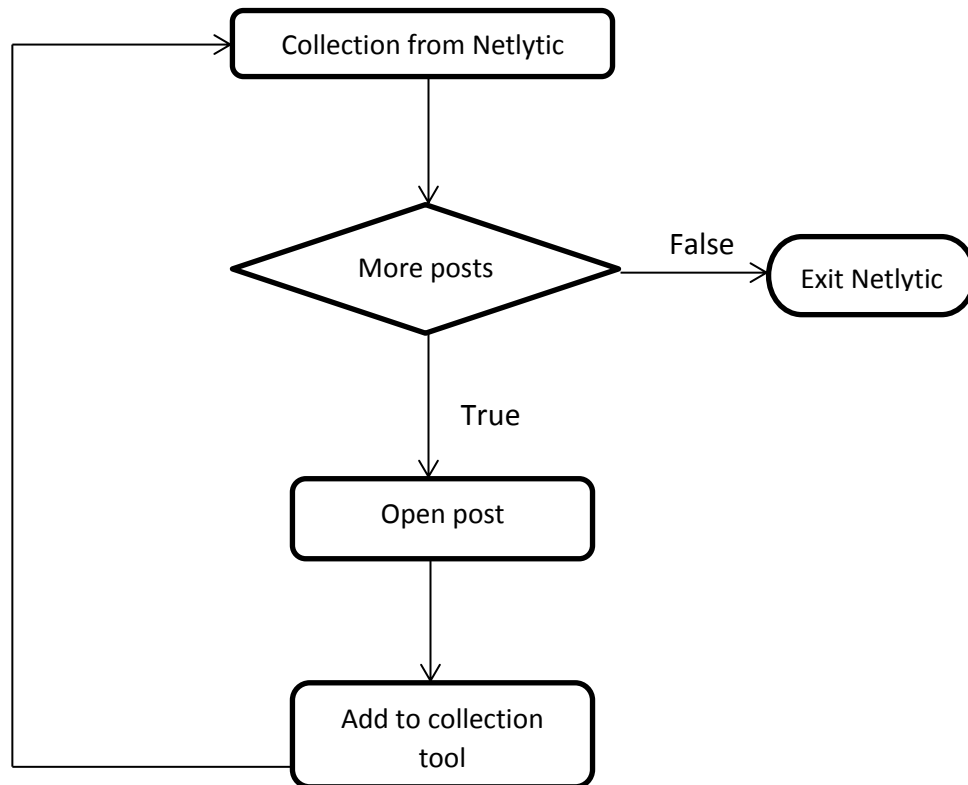
- [75] Z. Wang, V. Joo, C. Tong, and D. Chan, "Issues of Social Data Analytics with a New Method for Sentiment Analysis of Social Media Data," *2014 IEEE 6th Int. Conf. Cloud Comput. Technol. Sci.*, pp. 899–904, 2014.
- [76] X. Sun, C. Li, and F. Ren, "Sentiment analysis for Chinese microblog based on deep neural networks with convolutional extension features," *Neurocomputing*, vol. 210, pp. 227–236, 2016.
- [77] M. Birbili, "Translating from One Language to Another," *Soc. Res. Updat.*, no. 31, pp. 1–6, 2000.
- [78] F. van Nes, T. Abma, H. Jonsson, and D. Deeg, "Language differences in qualitative research: Is meaning lost in translation?," *Eur. J. Ageing*, vol. 7, no. 4, pp. 313–316, 2010.
- [79] Z. Wang, "Segment-based Fine-grained Emotion Detection for Chinese Text," no. October, pp. 52–60, 2014.
- [80] K. Einert, "Humboldt Digital Library – avhumboldt," *Humboldt Digital*, 2009. .
- [81] A. Gajjala, "Multi Faceted Text Classification using Supervised Machine Learning Models," MS.Thesis, San Jose State University, 2016.
- [82] Vicky Qian, "Step-by-Step Twitter Sentiment Analysis: Visualizing United Airlines' PR Crisis | iPullRank." 2017.
- [83] S. Lanka, "Analyzing Emotions in Twitter During a Crisis: A Case Study of the 2015 Middle East Respiratory Syndrome Outbreak in Korea," pp. 182–188, 2012.
- [84] S. S. Aksenova, "Machine Learning with WEKA," California, 2004.
- [85] C. M. Bishop, *Pattern Recognition and Machine Learning*, vol. 53, no. 9. New York: Springer, 2013.
- [86] E. Bhuvaneswari and V. R. S. Dhulipala, "The Study and Analysis of Classification Algorithm for Animal Kingdom Dataset," vol. 2, no. 1, pp. 6–13, 2013.
- [87] S. Mangasuli, B and P. S. Bakare, "Prediction of Campus Placement Using Data

- Mining Algorithm-Fuzzy logic and K nearest neighbor,” *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 5, no. 6, pp. 309–312, 2016.
- [88] L. Devasena, “Comparative Analysis of Random Forest, REP Tree and J48 Classifiers for Credit Risk Prediction,” in *International Journal of Computer Applications*, 2014, pp. 975–8887.
 - [89] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, and A. Perera, “Opinion Mining and Sentiment Analysis on a Twitter Data Stream,” in *International Conference on Advances in ICT for Emerging Regions (ICTer 2012)*, 2012, pp. 182–188.
 - [90] P. Horton and K. Nakai, “Better prediction of protein cellular localization sites with the k nearest neighbors classifier,” *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, vol. 5, pp. 147–152, 1997.
 - [91] A. Ashari, I. Paryudi, and A. Tjoa, “Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 11, pp. 33–39, 2013.
 - [92] E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López, “Opinion classification techniques applied to a Spanish corpus,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6716 LNCS, pp. 169–176, 2011.
 - [93] Botswana Police Service, “Annual Report,” 2016.

Appendices

APPENDIX A: The Data Collection Process

The following is a flowchart we followed when collecting our data from Facebook. We used Netlytic software tool.



APPENDIX B: The Focus Group

The translation and data labeling was carried out by a focus group of 8 postgraduate students. The most common purpose of a focus group is to provide an in-depth exploration of a topic about which little is known. Four primary stages of the focus group discussion were followed:

1. Introduction– The main researcher provides an overview of the goals of the discussion and introductions are made
2. Rapport Building Stage – Easily answered questions about the topic are asked to encourage participants to begin talking and sharing
3. In-depth discussion – The main researcher starts the discussion with group, encouraging conversation that reveals participants' feelings and thoughts
4. Closure– The main researcher summarizes the impressions or conclusions gathered and participants clarify, confirm, or elaborate on the information.

After discussion amongst the group these steps are followed:

1. The group brainstorms a sentence and agrees on how to translate.
2. If the group disagrees, the main researcher chips in and makes the final decision.
3. Also if a sentence is not clear, the main researcher clarifies it.

For labeling the data, these were steps followed:

1. The main researcher explains the labeling classes as happiness, sadness, anger, disgust, fear and surprise.
2. The group brainstorms and agrees on the labels.
3. Each sentence is given a label and if there is disagreement, the main researcher decides on the label.

APPENDIX C: A code snippet to remove stop-words

The following code was done using Python language to remove stop-words from the crime dataset.

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize, sent_tokenize
f = open("target.txt", "r")
s = f.read()
wrte = open("preprocessed.txt", "w")
t = sent_tokenize(s)
swords = set(stopwords.words('english'))
tokenized_sents = [word_tokenize(i) for i in t]
for i in tokenized_sents:
    fsent = [w for w in i if not w in swords]
    fsent = []
    for w in i:
        if w not in swords:
            fsent.append(w)
    r = " ".join(str(x) for x in fsent)
    print(r)
    wrte.write(r + "\n")

wrte.close()
f.close()
```