A FRAMEWORK FOR PROFILING CRIME
REPORTED USING SOCIAL MEDIA – A CASE OF
TWITTER DATA IN KENYA

BY
CIDEON K. SUMBEIYWO

A Project Report Submitted to the School of Science
and Technology in Partial Fulfillment of the
Requirement for the Degree of Master of Science in
Information Systems and Technology

UNITED STATES INTERNATIONAL UNIVERSITY –
AFRICA

SUMMER 2018
STUDENT’S DECLARATION

I, the undersigned, declare that this is my original work and has not been submitted to any other college, institution or university other than the United States International University – Africa in Nairobi for academic credit.

Signed: ________________________       Date: ______________
        Gideon Kipkorir Sumbeiywo (ID No 620128)

This project has been presented for examination with my approval as the appointed supervisor.

Signed: ________________________       Date: ______________
        Dr. Leah Mutanu

Signed: ________________________       Date: ______________
        Dean, School of Science and Technology
COPYRIGHT

All rights reserved; No part of this work may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without express written authorization from the writer.

Gideon Kipkorir Sumbeiwo © 2018
CRIME PROFILING HELPS LAW ENFORCEMENT AGENCIES UNDERSTAND, TACKLE AND SOMETIMES PREDICT THE NEXT MOVE BY CRIMINALS. THIS CAN BE ACHIEVED BY MONITORING AND STUDYING PATTERNS AND TRENDS THAT HAVE OCCURRED IN THE PAST AND CONTINUE TO OCCUR IN THE PRESENT. SOCIAL MEDIA PLATFORMS SUCH AS FACEBOOK, GOOGLE PLUS, INSTAGRAM, REDDIT AND IN THIS CASE TWITTER, HAVE CREATED PLATFORMS WHERE PEOPLE SHARE VIEWS, OPINIONS AND EMOTIONS ALL THE WHILE INFLUENCING AND INFORMING OTHERS.

This research set out with four objectives that would enable it to be successful in coming up with a framework for profiling. The objectives were, Identifying and extracting crime data from social media, Performing Sentiment analysis on crime data, Designing a Framework for profiling of crime, using the Twitter social media platform and Testing of the framework.

This study set out to find crime related data from members of the public from Twitter and see if it can be used to profile crime. Collecting and extracting of tweets was done from the Kenyan Twitter population, cleaned and sentiment analysis carried out. The analysis of sentiments was done using a lexicon based algorithm within the SentiStrength open source tool.

Through the framework model created in the project, the preprocessed data was analyzed in regards to the environment and author variables. Then establishing how they can be used to monitor or profile crime. The findings from this analysis led to viewing of patterns and hotspots of crimes on real life digital maps.

The results provided insight into the dynamics around different authors who tweet about crime and what differentiates one from another. These results where then investigated and compared to derive an understanding of the dynamics. Finally the framework was tested via two types of tests, usability test and Pearson correlation coefficient.

The major study findings showed that it is possible to come up with a framework model to use in the tracking of crime in Kenya from Twitter. Usability and correlation coefficient tests proved the framework successful from user feedback and remodeling of the framework.

Major recommendations for future studies is to have a data collection method for a streaming API instead of a standard API. This API together with a tailor made application that collects real time data will ensure better collection of data. The Law Enforcement Agencies can then collect data continuously and store in their databases. There is also a need to have local language content within the dictionaries. This will better localize the sentiments of the tweets to include few Swahili, Sheng or tribal words.
ACKNOWLEDGMENT

I would like God for giving me the courage and strength to embark on and complete this project.

I would like to offer my sincere gratitude to my supervisor Dr. Leah Mutanu for expertise, intelligence, knowledge and persistence all through this project.

I would also like to thank my family and everyone else who offered their support while I was doing this project. Please accept my gratitude and sincere appreciation.

Thank you
DEDICATION

I would like to dedicate my Masters to my wife, daughter, parents, siblings and friends with a special gratitude to my wife, daughter and parents who continued to pray, encourage, support and continually push me in my pursuit for my masters.

I also dedicate this dissertation to the Masters in Information Systems and Technology USIU faculty especially Dr. G. Chege and Dr. L. Mutanu who has always sacrificed her busy schedule to offer advice and encouraged me to finish the project.
# TABLE OF CONTENTS

STUDENT’S DECLARATION........................................................................................................................................... ii
COPYRIGHT.................................................................................................................................................................... iii
ABSTRACT........................................................................................................................................................................ iv
ACKNOWLEDGEMENT......................................................................................................................................................... v
DEDICATION...................................................................................................................................................................... vi
LIST OF TABLES.............................................................................................................................................................. x
LIST OF FIGURES............................................................................................................................................................. xi

Chapter 1: Introduction ..................................................................................................................................................... 1
1.1 Background of the Problem ........................................................................................................................................... 1
1.2 Statement of the Problem .............................................................................................................................................. 2
1.3 Purpose of the Study ....................................................................................................................................................... 3
1.4 Research Objectives ....................................................................................................................................................... 3
1.5 Justification of the Study ............................................................................................................................................... 3
1.6 Scope of the Study ......................................................................................................................................................... 4
1.7 Definition of Terms ....................................................................................................................................................... 4
1.8 Chapter Summary ......................................................................................................................................................... 5

Chapter 2: Literature Review ........................................................................................................................................ 6
2.1 Introduction................................................................................................................................................................. 6
2.2 Theoretical Foundation ............................................................................................................................................... 6
  2.2.1 Routine Activity Theory ........................................................................................................................................... 6
  2.2.2 Crime Pattern Theory ............................................................................................................................................... 7
  2.2.3 Rational Choice Theory ........................................................................................................................................... 8
2.3 Identifying and Extracting Crime Data from Social Media .......................................................................................... 9
  2.3.1 Data Characteristics .................................................................................................................................................. 9
  2.3.2 Identifying and Extraction ....................................................................................................................................... 10
2.4 Sentiment analysis on crime data ............................................................................................................................. 13
  2.4.1 Sentiment Analysis and Local Languages ............................................................................................................ 18
2.5 Profiling of Crime, using the Twitter social media platform ...................................................................................... 19
  2.5.1 Profiling of crime in the U.S.A ............................................................................................................................ 19
  2.5.2 Profiling of crime in Kenya .................................................................................................................................... 20
  2.5.3 Using Twitter to profile crime ............................................................................................................................... 22
  2.5.4 Developing a Framework for Profiling Crime ........................................................................................................ 23
2.6 Testing a Framework .................................................................................................................................................... 26
2.7 Conceptual Framework ................................................................................................................................................ 27
<table>
<thead>
<tr>
<th>Chapter 3: Methodology</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>30</td>
</tr>
<tr>
<td>3.2 Research Design</td>
<td>30</td>
</tr>
<tr>
<td>3.3 Population and Sampling Design</td>
<td>31</td>
</tr>
<tr>
<td>3.3.1 Population</td>
<td>31</td>
</tr>
<tr>
<td>3.3.2 Sampling</td>
<td>31</td>
</tr>
<tr>
<td>3.4 Data Collection Methods</td>
<td>31</td>
</tr>
<tr>
<td>3.5 Research Procedures</td>
<td>32</td>
</tr>
<tr>
<td>3.6 Data Analysis Methods</td>
<td>34</td>
</tr>
<tr>
<td>3.7 Ethical Considerations</td>
<td>34</td>
</tr>
<tr>
<td>3.8 Chapter Summary</td>
<td>35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 4: Model</th>
<th>36</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction</td>
<td>36</td>
</tr>
<tr>
<td>4.2 Analysis</td>
<td>36</td>
</tr>
<tr>
<td>4.3 Modeling</td>
<td>37</td>
</tr>
<tr>
<td>4.4 Testing</td>
<td>38</td>
</tr>
<tr>
<td>4.5 Chapter summary</td>
<td>39</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 5: Results and Findings</th>
<th>41</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Introduction</td>
<td>41</td>
</tr>
<tr>
<td>5.2 Identifying and extracting crime data from social media</td>
<td>41</td>
</tr>
<tr>
<td>5.3 Sentiment analysis on crime data</td>
<td>44</td>
</tr>
<tr>
<td>5.4 Frame work for profiling of crime, using the Twitter social media platform</td>
<td>48</td>
</tr>
<tr>
<td>5.5 Testing of the framework</td>
<td>54</td>
</tr>
<tr>
<td>5.5.1 Usability Testing</td>
<td>54</td>
</tr>
<tr>
<td>5.5.2 Framework Testing</td>
<td>55</td>
</tr>
<tr>
<td>5.5.3 Collaborative Analysis</td>
<td>58</td>
</tr>
<tr>
<td>5.6 Chapter summary</td>
<td>59</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 6: Discussion, Conclusions and Recommendations</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1 Introduction</td>
<td>60</td>
</tr>
<tr>
<td>6.2 Summary</td>
<td>60</td>
</tr>
<tr>
<td>6.3 Discussions</td>
<td>61</td>
</tr>
<tr>
<td>6.3.1 Identifying and Extracting Of Data from Social Media</td>
<td>61</td>
</tr>
<tr>
<td>6.3.2 Performing Sentiment Analysis on Crime Data</td>
<td>62</td>
</tr>
<tr>
<td>6.3.3 Designing a Frame Work for Profiling of Crime, using the Twitter</td>
<td>62</td>
</tr>
</tbody>
</table>
Social Media Platform ................................................................. 62
6.3.4 Testing of the Framework .................................................. 63
6.4 Conclusions .............................................................................. 64
6.4.1 Identifying and Extracting of Data from Social Media ........ 64
6.4.2 Performing Sentiment Analysis on Crime Data .................. 64
6.4.3 Designing a Framework for Profiling of Crime, Using the Twitter Social Media Platform ...................................................... 65
6.4.4 Testing of the Framework .................................................. 65
6.5 Recommendations ..................................................................... 65
References ............................................................................................. 68
## LIST OF TABLES

Table 2-1: Example acronyms and meaning ................................................................. 15
Table 5-1: Crime search words used to extract crime data from Twitter ...................... 41
Table 5-2: Top 24 words with highest negative sentiment ........................................... 46
Table 5-3: Sentiment analysis results figures ................................................................. 47
Table 5-4: Frequency of crime mentions by County ...................................................... 49
Table 5-5: Frequency and Type of tweets in specific areas ........................................... 51
Table 5-6: Summary of crimes in Table 5.5 ................................................................. 51
Table 5-7: Top 20 authors with details .......................................................................... 53
Table 5-8: Percentage of crime related tweets over other tweets. Also activity levels of the authors .................................................................................................................. 54
Table 5-9: Correlation matrix ....................................................................................... 66
LIST OF FIGURES

Figure 2-1: Routine activity theory ................................................................. 7
Figure 2-2: Crime pattern theory ................................................................. 8
Figure 2-3: Big Data Challenges ................................................................. 9
Figure 2-4: Sentiment Analysis Architecture .............................................. 18
Figure 2-5: A proposed framework for global cyber-attacks ......................... 28
Figure 4-1: Research variables and their attributes ..................................... 38
Figure 5-1 Geographical location coverage of tweet collection ..................... 42
Figure 5-2: Twitter Account used to access data ........................................... 43
Figure 5-3: Twitter Archiver extracting Twitter data ...................................... 43
Figure 5-4: Search parameters ..................................................................... 44
Figure 5-5: Database of tweets ..................................................................... 44
Figure 5-6: Removing of duplicates ............................................................... 45
Figure 5-7: Sentiment analysis of all the tweets ............................................ 46
Figure 5-8: Graphs depicting sentiment analysis of all the tweets .................. 47
Figure 5-9: Map of Counties with mentions of crime data ......................... 48
Figure 5-10: Mapping of Crime location in different parts of Kenya ............. 49
Figure 5-11: Time period of the day authors’ tweet ................................... 52
Figure 5-12: Twitter users who follow the most users and Authors with the most followers ................................................................. 52
Figure 5-13: Coding of data ....................................................................... 56
Figure 5-14: Resultant framework model ..................................................... 58
Chapter 1: Introduction

1.1 Background of the Problem

A crime can be defined as a harmful act towards a state, society, community or individual that goes against an agreed upon law of the land. Crime has therefore existed as long as man has been civilized and lived according to a set of binding rules. To counter crime, man came up with laws and law enforcement to protect against those crimes. Therefore, law enforcement agencies (LEAs) exist in our world with a view to enforce agreed upon laws.

In order to fight crime effectively LEAs need to study criminal patterns and trends which include but not limited to, individuals and groups of criminals, types of crime, victims, analyzing and predicting crime patterns. These criminal activities can be studied and behaviours looked at. Furthermore all these activities and more can be referred to as criminal profiling.

Crime profiling helps LEAs understand, tackle and sometimes predicting the next move by criminals. This can be achieved by monitoring and studying patterns and trends that have occurred in the past and continue to occur in the present. With advancements in technology, the internet and specifically social media, humans and in turn criminal behavior has moved from not only the real world, but also the digital domain.

Social media platforms such as Facebook, Google Plus, Instagram, Reddit and in this case Twitter, have created platforms where people share views, opinions and emotions all the while influencing and informing others. Twitter has over 330 million users and therefore provides a large audience and in this studies case, a large volume of data to use (Minter, 2017).

Crime and criminal activity is a constant and continuous threat to society that occurs either premeditated or spontaneously. Law enforcement agencies have key roles in stopping or responding to perpetuated crime and the factors linked to it such as the victims and offenders. In order to monitor and respond to crime effectively, LEAs need mechanisms in place that notify them of criminal acts occurring in their area of jurisdiction.

In Kenya, traditional methods of reporting crime to the police include communicating via telephone or website and physically reporting to a police station. There is fear of visiting police
stations due to stigma including poor service or corruption. This therefore makes police records inaccurate as they lack the true picture on the ground.

Various social media platforms exist in the world but this study focused on Twitter because of the simplicity of collecting data from the platform. Twitter sends a 280 character or less phrase (tweet) to as many people as possible. The message is sent in a small enough text format which is easier to collect and analyze as opposed to other platforms that use large amounts of text, images, videos and web links. This feature therefore made Twitter ideal for data collection and analysis to be completed within the time allocated for the project.

Previous research studies mostly analyze Twitter data and the sentiment involved then report on it. Most sentiment analysis studies also classify sentiment into binary or 3-way negative, positive and neutral sentiments. This study will aim to more than just sentiment analysis. It will identify, extract, store, carry out sentiment analysis, create a model and test that model in relation to the data collected. Most studies on Twitter sentiment analysis only report on what percentage are positive, negative or neutral.

1.2 Statement of the Problem

The major problem with this subject area is the collection of crime data by LEAs. Data on crime is limited due do a number of reasons and without data, LEAs cannot effectively monitor, cover and profile criminal activities. Data collection on crime is limited to reports from members of the public.

They reporting via phone, sending a message via official LEA websites or physically reporting to a police officer or a police station. By delving into the social media domain, LEAs can connect directly to citizens who can give any crime related data at their own convenience, safely and in a timely manner.

Crime inherently is a negative topic. If sentiment analysis is carried out on crime alone, and sentiment analysis results were to be classified in a binary or 3-way classification of positive, negative or neural, the results will be negatively skewed and not give much more. There needs to be a step further in data mining and processing so as to add value to sentiment analysis on crime related research (Dhanhani, 2018).
Danthala (2015) continues to explain the need for more work done in the research apart from just stopping at sentiment analysis. The data collected and the analysis should go beyond just tabulating results but rather use them as a platform for further value addition into creating a framework that LEAs will find useful in their daily operations.

1.3 Purpose of the Study

Carrying out data mining research using the Twitter platform exposes this study to a large, widespread, current and simple format repository of data. This study will collect and extract tweets specific to crime in Kenya. These tweets will then be preprocessed into different datasets and sentiment analysis be carried out on them.

Analysis of those datasets will occur where the aim of profiling trends and patterns of crime in Kenya, profiling users who post tweets and classifying of victims. Once these profiles are done, the study will look at the viability of that data and decide if it contains concrete usable data. Finally it will this study will test the framework to confirm its viability and value.

Law enforcement agencies will benefit from the analyzed and processed Twitter data by obtaining a clear direction on the situation on the ground from the citizens point of view. Various types of policing strategies exist such as community-policing, evidence-based policing and intelligence-based policing. With collection of data from Twitter, the LEAs will bolster up other strategies with information before even going out into the community to collect evidence or intelligence.

1.4 Research Objectives

The research identified four objectives that would enable it to design a framework for profiling crime. These objectives are:

i. Identifying and extracting crime data from social media.

ii. Performing Sentiment analysis on crime data

iii. Designing a Framework for profiling of crime, using the Twitter social media platform.

iv. Testing of the framework.

1.5 Justification of the Study

Millions of people use Twitter every second to express themselves to the world (Minter, 2017). They post tweets on what they are experiencing, feeling, or what others have experienced. If a user has had a crime committed against them, or heard someone else tell them, they are likely to share on social media or tell someone who will share it.
Because LEAs are constantly striving to identify crime as early as possible, social media has a vast untapped wealth of data that can be mined to facilitate crime fighting. If this data can be extracted, analyzed, classified and presented in a form that can help in profiling and even predicting crime, LEAs can get ahead of the criminals and better carry out their mandate of protecting people and property.

There is anonymity in using third party entities who are independent and alert on criminal activities occurring. Victims, people close to them, or people who witnessed a crime can inform third party groups or activists who report it to the world via tweets. These groups and individuals have a lot of information and can be used as a key source of profiling data.

1.6 Scope of the Study

This project will concentrate on mining data related to crime from Twitter in Kenya. Both historical and present data will be sifted through to collect crime and crime related issues only. Analysis of trends and pattern of crime which include locations, times and any other dynamics the study might uncover will be performed. Other aspects of data to be collected will be the user accounts that posts on crime issues, profiling those accounts, if they are victims or third parties.

The data collection will span two to three months of tweet gathering on key crimes that LEAs usually deal with constantly. The crimes targeted will revolve around the most common crimes found in Kenya as well as the crimes that are a big concern for LEAs. This will call for understanding of the crime situation in the country.

1.7 Definition of Terms

A description of terms associated with this research and tweets is provided in this section.

**Emoticons**: These are facial expressions pictorially represented using letters and punctuation expressing the mood of the user (Danthala, 2015).

**Target**: Twitter users have profile names that can be referred to by using the “@” symbol, which then automatically alerts them e.g. @PeterKibet (Mukindia, 2017).

**Hashtags**: Hashtags are used to link a particular tweet to an ongoing trend or discussion, denoted by # (the topic) e.g. #Corruption (Danthala, 2015).
Natural language processing—(NLP) is a field of computer science, artificial intelligence and linguistics concerned with the interactions between computers and human (natural) languages. Specifically, it is the process of a computer extracting meaningful information from natural language input and/or producing natural language output (Bolla, 2014).

Opinion mining—opinion mining (sentiment mining, opinion/sentiment extraction) is the area of research that attempts to make automatic systems to determine human opinion from text written in natural language (Gerber M., 2014).

Scraping—collecting online data from social media and other Web sites in the form of unstructured text and also known as site scraping, web harvesting and web data extraction (Cochran, 1977).

Sentiment analysis—sentiment analysis refers to the application of natural language processing, computational linguistics and text analytics to identify and extract subjective information in source materials (Khan, 2009).

1.8 Chapter Summary

This chapter introduced the research topic by looking at the background of the problem. It introduced the components involved such as LEAs, victims, offenders, users and the Twitter platform. A justification on the choice of social media platform was also presented. The dynamics of the Twitter platform and how it works was also discussed.

The statement of the problem was examined, indicating the challenges with traditional methods of collecting data on crime. In the purpose of the study, the reasons why data from Twitter will be beneficial and the viability of Twitter data for this study were outlined.

A discussion on the research objectives showing how the research be conducted from identification, extraction, analysis, coming up with and finally testing the framework was given. A justification of the study showing the gaps that exist currently and how they will be addressed was presented. The scope of the focused on Tweets generated in Kenya. Finally a definition of terms was listed. The chapter following will provide a detailed literature review on working with Twitter data.
Chapter 2: Literature Review

2.1 Introduction

This chapter of the study looks at previous studies by different researchers done in profiling of crime. Various researches have also been done on similar topics related to social media data mining and sentiment analysis of that data. The review of existing literature is based on the objectives listed in chapter one, examines what currently exists in the body of knowledge and how this particular research may contribute to it.

This review will evaluate, summarize, clarify and describe literature in regards to this topic. It will help in giving the theoretical base and assisting in the nature of this research. In addition to this, the purpose is to highlight ideas, knowledge, strengths and weaknesses on this topic. This literature review has been defined by a guiding concept of theoretical foundation, the research objectives and finally the conceptual framework.

This literature review will use sources from various areas including but not limited to book, scholarly articles, government reports, websites and research projects. Various considerations will be given such as recent material that are not older than ten years except for special considerations. Other considerations will include the authors with supporting empirical evidence, unbiased views and the contribution to the subject.

2.2 Theoretical Foundation

This section of the study will look at related theories relevant to this study. The theoretical frameworks assisted in looking at the key variables in this topic and highlighted their differences and circumstances. This helped the study explain the challenges, nature and meaning associated with this topic.

2.2.1 Routine Activity Theory

One theory is the routine activity theory or approach. This theory looks at predatory crimes which are crimes that involve conscious intent and preparation to act on a vulnerable target. This theory needs three elements to exist: a suitable target, a likely offender, and the absence of a capable guardian (Clarke, 1998)

As seen in Figure 2-1, the likely offender already exist therefore the other elements are the ones that can change. The guardian is anyone whose proximity or presence would discourage the crime
from occurring. The guardian is not necessarily the presence of a police officer or law enforcement officer, but even the presence of a co-worker, neighbour, doorman or housewife will have an impact against crime (Clarke, 1998).

**Routine Activity Theory**

![Routine Activity Theory Diagram](image)

**Figure 2-1: Routine activity theory (Source: Anonymous (2018))**

The third element of target refers to either an object or a victim, due to the fact that a crime can be committed against an object when the owner is away such as theft of a laptop. A target’s risk level is the chance a crime can be perpetuated against it. This risk level can be looked at using the following four main elements; Value, Inertia, Visibility and Access (Clarke, 1998).

The value of the target may interest the offender for example it could be because of its monetary gain. The inertia refers to the weight of the item, for example small lighter goods are easier to steal than heavier goods. Visibility is simply the exposure of the target. Access is how easy it will be for the offender to commit the crime physically (Clarke, 1998).

**2.2.2 Crime Pattern Theory**

Crime pattern theory shown in Figure 2-2 looks at the interactions between people and their physical environment. It analyzes the movement of crime in relation to people and things in time and space. There are three main elements for this theory: edges, paths and nodes. An individual’s everyday activities are located around certain specific areas, for example, school, home, work, entertainment or shopping areas (Clarke, 1998).
These locations that an individual goes to and from are called “Nodes”. These nodes are activity spaces but they also have a perimeter surrounding them which is the local environment. Crimes therefore occur either within the node or the perimeter, activity or awareness space respectively (Clarke, 1998).

![Crime pattern theory](source.png)

**Figure 2-2: Crime pattern theory (Source: Kashyap (2011))**

The second elements are the edges, which are the boundaries between different nodes. This could mean neighbourhoods where different people from different neighbourhoods may meet. Insiders or outsiders from one neighbourhood carry out a crime in the next neighbourhood and retreat for safety in their own neighbourhood. The third element is the path. Paths are the routes that connect different nodes and may form part of the offender’s awareness space and in turn locations for committing crimes (Clarke, 1998).

### 2.2.3 Rational Choice Theory

This method designed by Cornish, Derek, Clarke & Ronald (1986) believes that the reason actor (that is man), weighs ends and means, benefits and costs to make a rational choice. To meet any number of needs such as status, sex, excitement and money, an offender will make rational calculations, decisions and choices in the committing of a crime. This theory is based on the offender’s individualism as they seek to maximize their goals through self-interest.

These three theories seem to overlap in regards to criminal offenders and crime in general. Both routine activity and crime pattern theories have variables of location, offender and victim, while all three discuss the offender. Each gives different attention to various areas, routine activity theory (larger society), crime pattern theory (the local area) and rational choice theory (the individual).
These theories assist law enforcement in investigating why crime exist in some areas as well as predicting future occurrences of crime. They at times can help policy makers in tackling crime such as paying attention to the daily rhythm of activities and the geographical distribution of crime. This can also be used to tackle crime rates by adjusting the management and design of cities, towns and business areas. Ultimately what these theories tell us is that locality and society can change crime opportunity while the individual offender reacts to these changes.

2.3 Identifying and Extracting Crime Data from Social Media

2.3.1 Data Characteristics

Before identifying and extracting crime data from social media, it is first important to understand big data and specifically social media data. Traditionally with the dawn of computers in business, organizations really did not perform major analysis and operations on that data. Big data became more present from the year 2005 onwards with a need to incorporate different types of data into analysis and processing real time (Gewirtz, 2018).

Big data refers to huge volumes of data that cannot be handled or mined using traditional methods and the capacity surpasses conventional traditional databases according to Gewirtz (2018). An example can be seen is the popular social media platform Facebook, whereby they created algorithms that link huge numbers of signals from friends and users to enable sharing of similar ideas, views and various other things.

Gewirtz (2018) explains that organizations use big data for numerous uses but the major value usually falls between analysis, monitoring, assessing and enabling new services and products based on the ones existing. Analysis on big data can reveal hidden issues for organizations such as customers or user interactions, geographical and social data issues in the market and industry.

According to Danthala (2015), the major challenges and characteristics of big data are Velocity, Volume and Variety shown in Figure 2-3. They are used to characterize different aspects of big data and are called the 3 V’s.

![Figure 2-3: Big Data Challenges (Source: Gewirtz (2018))](image-url)
The speed at which the data is modified, created, processed and retrieved is referred to as *Velocity*. The data has to be processed and with it being huge, the time for performing various operations as well as the processing itself is a major challenge and has to be considered. The structure of traditional databases does not allow efficient performing of operations at an acceptable speed, thus not satisfying business and customer needs (Gewirtz, 2018).

The amount of data to be processed is the *Volume* and it is the second challenge. Processing and uploading of data are done in many business areas, and done at very high rates in terms of zetabytes and terabytes. Danthala (2015) states that daily uploads are around 10 terabytes on various devices such as networking, telecommunications and RFID, 12 terabytes of twitter data and 25 terabytes of Facebook data.

The third challenge is *Variety*, whereby it is the chance of presenting various types of data in a distributed environment. The categories include unstructured, semi structured and structured. The process of performing operations and analysis increase in difficulty from structured to unstructured. Social media like Tweets and Facebook posts can give one an understanding of product preferences and use through sentiment analysis. These are the issues when processing information from different data sets (Gewirtz, 2018).

Muthanna (2014) includes a forth challenge as *Veracity*. This refers to the abnormality, biases and noise in data. The data that is mined and stored meaningful to the problem being analyzed. This means that Veracity can be approached as the uncertainty of data due to model approximations, incompleteness and inconsistency, latency and ambiguities in the process of analyzing data.

### 2.3.2 Identifying and Extraction

After looking at the various challenges of big data above, the issues around extraction of that big data and specifically social media data are now better understood. There exist numerous social media platforms in the world today, this research will look into a number of the existing platforms and reasons why Twitter was chosen above the rest.

In defining social media, Haenlein (2010) refers to it as mobile-based and web-based Internet applications that enable the access, creation and exchange of user-generated content whose access is restricted or universally accessible depending on the particular platform. Carr (2015) further defines social media as communication that is computer-mediated where users not only create
profiles to show who they are but also to create content on their own, view and interact with content of others users especially their friends.

Ma (2017) looks at the use motivations, intensity of use, and time spent daily on social media platforms such as main social media platforms Snapchat, Facebook, Instagram and Twitter. In their research, they go on to describe the different platforms starting with Facebook. According to their website, Facebook (2017) states that it is the most popular social networking site (SNS). It allows users to connect with family members, acquaintances and friends, giving people the chance to share and post content such as status updates and photos. It was found in 2004 and has over 2.2 billion monthly users and over 1 billion daily active users.

The second most popular social media platform according to Ma (2017) is Instagram with 800 million monthly users who are active. Instagram is a mobile application where users take pictures, edit them with filters and other add-ons, and then share them on Instagram as well as other platforms such as Twitter and Facebook.

Snapchat is yet another mobile social media application where users receive and send time sensitive videos and photos that disappear after 24 hrs. Snapchat like Instagram allows users to communicate to their followers as well as direct message specific users privately. Snapchat has over 187 million active users (Ma, 2017).

Founded in 2006, Twitter has been branded as a microblogging website where users interact live using 280 character tweets to their followers. Users normally can communicate using hashtags, mentions and replies. Twitter has over 321 million active users and unique monthly visits of 1 billion monthly to websites from embedded tweets (Ma, 2017).

Ahmed (2017) argues that the Twitter infrastructure is unique to all other social media platforms. The uniqueness of Twitter allows users to be followed by any person, Twitter user or not. Only until a user enables the “Protected Tweets” feature is when the tweets will only be available to the people that follow them privately, but a majority of users do not do this.

The essence of Twitter is to send a 280 character or less tweet to as many people as possible, this key feature therefore makes Twitter ideal for research purposes and the collection of data. In addition, using Twitter Application Programming Interface (APIs), a researcher can collect almost
100% of the data in the Twitter universe both real time and historical for further uses (Ahmed, 2017).

When it comes to the extraction of data from Twitter a number of considerations need to be looked at. Firstly from the huge data sets in a cluster, the data has to be accessed and stored. There is a need for a standard computing platform that will scale down huge data into sizable data for computing from different data storage platforms (Danthala, 2015).

A second challenge with a large Twitter data set is the retrieval. Danthala (2015) explains that in certain scenarios this data is growing daily. It then becomes difficult to access large networks of data for specific actions. Thirdly, there needs to be a designed algorithm used to handle the problems raised by dynamic data characteristics and huge data volumes. Therefore, this research needs methods in performing operations on data sets found from social media.

There are a number of ways to identify and extract data with different researchers employing various techniques. Danthala (2015) research scope included fetching and analyzing users with the most retweeted statuses from a specific data set. That research used a Twitter API to collect tweets from that social networking platform.

Secondly, he used a distributed framework for big data processing called Apache Hadoop and MapReduce programming module, to reduce data in sizeable chunks from mapped frequent datasets. Finally, fetching of the Twitter IDs from the collected tweets, of those users with most retweeted statuses (Danthala, 2015).

Due to the short length of Twitter statuses, users use emoticons, acronyms, spelling mistakes and special characters to get a message across. Agarwal, Xie, Vovsha, Rambow and Passonneau (2011) acquired 11,875 tweets from a commercial source, which are available at a cost. This data was collected via a real-time steam, with no location, language or any other restriction put in the steaming process. This meant that even the language was not restricted and they had to translate some of the tweets into English before further processing.

They then clustered the tweets into various labels that were junk, neutral, negative or positive. They used human annotators that carry out Natural Language Processing (NLP) to understand human language. That research further eliminated tweets labelled as junk which the annotators
could not understand leaving an unbalanced sample of 8753 tweets. They further used stratified sampling to get 1709 tweets each from neutral, negative and positive classes (Agarwal et al., 2011).

Dhanhani (2018) extracted data from Twitter using a different method. He went for Logstash software which is an open source solution that can be a real time data collection tool with pipeline options. Logstash can unify, normalize, do logs and events collection, and sanitize the data into other formats.

The advantages of Logstash include the facilitation of data analysis and data ingestion via input and output plugins. Collection of different types of data such as some files or network streaming events is done by configuring inputs. The elimination of unnecessary data is done by configuring filters. Then the configuration allows the data output to be stored in other files (Danthala, 2015).

In the implementation of their research project, Kiruthika (2016) used the programming language R. It includes a “twitteR” package that connects via the Twitter API with Twitter credentials. He then performed preprocessing and text mining via the “SnowballC” and “tm” packages. Preprocessing is necessary to separate a number of different database fields such as text, id, favorite, replyToSID, favoriteCount, truncated, replyCount, replyTOSN and created in raw format.

R also has “DT”, “Shiny” and “markdown” packages that were used to create the graphical user interface. He also used a specially designed Twitter POS tagger for Twitter data that use specialized tags to annotate tweets, A (adjective), E (emotion), V (verb), @ (at-mention), N (noun) and U (URL). The problem with R is the researcher must be familiar with the programming language code.

2.4 Sentiment analysis on crime data

The huge amount of data extracted from Twitter is difficult to analyze without automation. To do this, various language processing techniques and models have to be employed. These include opinion mining and Natural Language Processing (NLP) tools that includes Sentiment Analysis (SA).

The mining of opinions, views, emotions and attitudes from speech, text, database sources and tweets through Natural Language Processing (NLP) is Sentiment analysis (SA). SA is whereby text is classified into categories such as “neutral”, “negative” or “positive”. SA can also be referred to as appraisal extraction, subjective analysis and opinion mining.
According to Sonawane (2016) businesses and marketers use sentiment analysis to understand the user’s requirements in relation to their services or products after analyzing the data. The core of SA are mainly sentiments, opinions, emotions, attitudes and appraisals. Other retrieval techniques such as Textual Information focus on analyzing or searching the present factual data. Although it is important to look at the factual contents, it is also necessary to look at the subjective characteristics as well.

Sentiment analysis is an area that offers opportunities for new application such as recommendation systems. These recommendation systems can look at goods or services reactions on social networking sites or blogs, sift through the negative and positive opinions, and then come up with a recommendation of public opinion. In SA the words belief, opinion, sentiment and view are used interchangeably, however there exists some differences between them.

Belief is defined as intellectual assent and acceptance, due to different opinions from different experts. Opinion is defined as a conclusion open to dispute. An opinion on one’s own feelings is a sentiment, and a view is a subjective opinion. An example of terminologies used in sentiment analysis are as follows,

\[ \text{SENTENCE} = \text{The story of the book was strong and interesting} \]
\[ \text{OPINION HOLDER} = \text{<sentence author>} \]
\[ \text{OBJECT} = \text{<book>} \]
\[ \text{FEATURE} = \text{<story>} \]
\[ \text{OPINION} = \text{<strong><interesting>} \]
\[ \text{POLARITY} = \text{<positive>} \]

An opinion can be represented mathematically as a quintuple \((t, h, so, f, o)\), where
\[ t = \text{time when opinion is stated;} \]
\[ h = \text{opinion holder, who is the entity or organization or person that expresses the opinion;} \]
\[ so = \text{polarity or orientation of the opinion;} \]
\[ o = \text{object;} \]
\[ f = \text{feature of the o – object.} \]

Sentiment Analysis includes various tasks such as extraction of sentiment, classification of sentiment, summarization of opinions, subjectivity classification or opinion spam detection to name a few. It aims to analyze sentiments of people, emotions, opinions, attitudes, etc. towards
elements such as services, products, organizations, individuals and topics, such as crime in this research study.

There has been a lot of work done in recent years in Sentiment Analysis on Twitter data by various researchers. Initial stages came up with binary classification that assigns reviews or opinions to negative or positive bipolar classes.

Due to Twitter restrictions of 280 characters per tweet, users use acronyms, emoticons and special characters among other techniques to get their message across. In their research, Agarwal et al. (2011) prepared two dictionaries for processing Twitter data, one for acronyms and another for emoticons.

They collected 170 emoticons used by users and defined their emotional state. For example “☹” was categorized as negative whereas “😊” was labelled as positive. Then all emoticons were categorized as follows: Extremely-negative, Negative, Neutral, Positive and Extremely-positive. Similarly, their research assembled an acronym dictionary made up of 5,184 acronyms from various different sources. For example “omg” is translated to “oh my god” and others as seen in Table 2-1.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Expansion in English</th>
</tr>
</thead>
<tbody>
<tr>
<td>gr8/t</td>
<td>Great</td>
</tr>
<tr>
<td>tbt</td>
<td>Throw back Thursday</td>
</tr>
<tr>
<td>fomo</td>
<td>Fear of missing out</td>
</tr>
<tr>
<td>bae</td>
<td>Before anyone else</td>
</tr>
</tbody>
</table>

Collected tweets were then processed as follows: a) using the emoticon dictionary, replacing sentiment polarity for corresponding emoticons, b) substituting all negations (e.g. cannot, no, never, not) with one tag, “NOT”, c) substitute Uniform Resource Locator (URLs) with tag “U”, d) substitute repeated characters by 3 characters, such as, “suuuuuuper to suuper” and e) substitute targets with “T” (e.g. @Tony) (Agarwal et al., 2011).

A technique carried out by Paroubek (2010), collected tweets using the Twitter API into a twitter corpus, they then annotated those tweets using emoticons. His proposed model was to classify tweets as negative, objective and positive. Using the multinomial Naïve Bayes method that uses elements such as N-gram and Parts Of Speech (POS) tags, they developed a sentiment classifier. He only used tweets with emoticons and therefore the exercise set was less efficient.
A two models approach including a Maximum Entropy model and a Naïve Bayes bigram model to classify tweets, was done by Movassate and Parikh (2009). They discovered the classifiers from Naïve Bayes were better than the Maximum Entropy one.

Go, Bhayani and Huang (2009) offered building models using Support Vector Machines (SVM), MaxEnt and Naïve Bayes which used distant supervision for sentiment analysis of Twitter data containing tweets with emoticons. This feature space comprised of POS, bigrams and unigrams. They concluded that unigrams were the most effective features and Server Vector Machines outperformed the rest of the models.

The method of automatic sentiment analysis made up of two phases was carried out by Feng and Barbosa (2010). The first phase included classifying tweets as subjective or objective, then further classifying the subjective tweets as negative or positive. The feature space comprised of exclamation marks, retweets, punctuation, hashtags and links in conjunction with POS and prior polarity of words features.

By using the Hoeffding tree, multinomial naïve Bayes and stochastic gradient descent (SGD) models, Frank and Bifet (2010), processed tweets provided by Firehouse API real-time Twitter data available publicly. They decided that when used with an appropriate learning rate, the stochastic gradient decent model was the best.

By developing a sentiment classification three way model into neutral, negative and positive classes. Agarwal et al. (2011) used models that were a tree kernel based model, a feature based model and a unigram model. The unigram model used over 9,000 features, the feature based model used 100 features and the tree kernel represented tweets as a tree. They concluded that features combining prior polarity of words and corresponding POS tags are key and play a huge role in classification. The tree kernel based model was the best performing model.

Davidov and Rappoport (2010) classified sentiment type using user defined hashtags in tweets, in different feature types such as patterns, n-grams, single words and punctuation. For sentiment classification they then combined them into a single feature vector. By using the K-Nearest Neighbor strategy for each example they constructed a feature vector to assign sentiment labels for each example in the training and test set.
Others such as Dai and Liang (2013) collected data using the Twitter API and created three labels of non-opinions, negative and positive. They filtered tweets containing opinions and implemented Unigram Naïve Bayes model and employed the simplifying independence of Naïve Bayes assumption. They used Chi square features extraction and Mutual Information to eliminate useless features. In the end, the placement of the tweet is predicted as either negative or positive.

For detecting polarity of tweets Garcia and Gamallo (2014) offered variations of Naïve Bayes classifiers. Two different classifiers were built, Binary (which made use of a polarity lexicon, classifying negative and positive. Ignoring neutral tweets) and Baseline (taught to classify tweets as neutral, positive and negative). Classifiers considered various features such as Multiword from different sources and Valence shifters, Lemmas (adverbs, nouns, adjectives and verbs) and Polarity Lexicons.

Turney (2002) carried out sentiment analysis using bag-of-words method whereby a document is represented as just a group of words and relationships between words was not considered. Sentiments of every word in the document was selected to determine the sentiment of the whole document and united with some aggregation functions.

In order to determine the emotional content of a word along different dimensions, Kamps et al. (2004) implemented the lexical database WordNet. They determined semantic polarity of adjectives and a distance metric on WordNet.

Combining various classification techniques and feature sets, Xia, Zong and Li (2011) came up with a framework for Sentiment Classification. They used three base classifiers (Support Vector Machines, Maximum Entropy and Naïve Bayes) and two feature sets types (Word relations and Part-of-speech information). For better accuracy and sentiment classification, they used ensemble approaches like Meta-classifier, fixed combination and weighted combination.

Luo, Osborne and Wang (2013) picked out the efficient techniques as well as challenges to mine opinions from Twitter data. Wildly varying language and spam makes opinion retrieval from Twitter a challenging mission. From the above literature we can tell that sentiment analysis architects follows a certain set of procedures and therefore it is possible to come up with an overall model for sentiment analysis as seen in Figure 2-4.
2.4.1 Sentiment Analysis on Local Languages

English text sentiment analysis has become an active and large research area with a number of commercial applications, but the language barrier limits the ability to evaluate the sentiment of many of the world’s population (Skiena, 2014).

For the English language, several well regarded sentiment lexicons exist, but for most of the world’s languages, this is not true. Skiena (2014), identified only 12 publicly available sentiment lexicons for 5 non-English languages (Italian, Chinese mandarin, Japanese, German and Arabic).

Sentiment analysis is an important area of NLP with a large and growing literature. Excellent surveys of the field include, establishing that rich online resources have greatly expanded opportunities for opinion mining and sentiment analysis. Godbole, Srinivasaiah and Skiena (2007) build up an English lexicon-based sentiment analysis system to evaluate the general reputation of entities.

Taboada, Brooke, Tofiloski, Voll and Stede (2011) present a more sophisticated model by considering patterns, including negation and repetition using adjusted weights. Liu (2010) introduces an efficient method, at the state of the art, for doing sentiment analysis and subjectivity in English. Researchers have investigated topic or domain dependent approaches to identify opinions.

Figure 2-4: Sentiment Analysis Architecture (Source Luce (2012))


Gînscă et al. (2011) work on better sentiment analysis system in Romanian. The ready availability of machine translation to and from English has prompted efforts to employ translation for sentiment analysis. They demonstrated that machine translation can perform quite well when extending the subjectivity analysis to multi-lingual environment, which makes it inspiring to replicate their work on lexicon-based sentiment analysis.

2.5 Profiling of Crime, using the Twitter social media platform.

This section will look at previous research studies carried out in profiling of crime in Kenya and how Twitter can be used for this.

2.5.1 Profiling of crime in the U.S.A

According to Young (2015), their research shows how data from technologies or social data like online searches, wearable technologies and social media can be used to predict crime. They carried out their research using big data infrastructure, data science and psychology science. This enabled them to come up with a framework to analyze social data and come up with predictions of likely crimes and their hotspots in and around University of California.

An existing framework was used by Lau, Kamal and Pramanik (2016) using structure analysis. This structure analysis was based on network mapping and centrality measures to detect crimes based on criminal network patterns. The study used 2004 to 2005 criminal data from the Sheriff’s department in Los Angeles County, USA. They looked at the attributes such as locations, weapons, organizations, types of crime then relating them to individual criminals.

In the research carried out by Gerber M. (2014), the research carried out crime profiling by investigating the use of spatiotemporally tagged tweets. The study used Twitter specific statistical
topic modeling and linguistic analysis to automatically identify discussion topics across major cities in the United States. By using a prediction crime model, these topics were incorporated to identify specific crime types. Compared to a standard kernel density estimation approach, Twitter data improved crime prediction.

2.5.2 Profiling of crime in Kenya

According to Mburu (2014) there is a growth of crime in major cities and across the world. Fatality and frequency of crime has gone up over the past few years posing a threat to socio-economic development. Criminal behavior variables are not very well understood and there is little research done on it in Kenya.

Masese (2007) explains that Nairobi is the commercial, political and administrative capital of Kenya, having a population of over 3 million residents. Additionally satellite towns such as Kikuyu, Mavoko, Thika and Ruiru supply the capital with over 1.5 million people as workforce that comes to Nairobi on a daily basis.

There are also neighbouring urban, peri-urban and rural areas of Thika, Kiambu, Machakos and Kajiado districts that receive services such as specialized healthcare, good schools and roads. Mushrooming of unplanned and unauthorized informal settlements and growth of slums such as Mukuru Kwa Njenga, Kibra and Mathare, are prevalent within and in the periphery of city of Nairobi. These settlements hold over 60% of Nairobi’s population (Masese, 2007).

Masese (2007) continues that the issue of crime in Nairobi is not limited to informal settlements but also closely associated with them. Reasons span from a sense of helplessness and hopelessness due to high levels of unemployment, inadequate accessibility and weak social control institutions. All these factors contribute to creation of crime avenues. Nearby richer neighborhoods also make robbery and theft attractive for youth. Therefore crime comes out of social and behavioral influences in social, cultural, economic and political contexts.

Nairobi, similarly to other urban areas of the world, has become a criminal magnet for individuals seeking to commit offences. Among other serious crimes, sexual and physical assault, armed robbery, burglaries, housebreaking, murder, carjacking and mugging are all common. This is a constant concern for residents who see it as living in a spiral of insecurity and violence. (Masese, 2007)
Law enforcement agencies are in charge of protecting the public from criminals and a tool for prevention as well as managing crime is profiling. Douglas, Ressler, Burgess and Hartman (1986) writes that criminal profiling is a valued means by which the field of investigations can be narrowed and has been successfully used by law enforcement in several areas. Profiling does not give the specific identity of the criminal but rather shows the type of criminal, victim most likely to be associated with the crime as well as characteristics and trends of the crime itself.

Vorpagel (1982) describes criminal profiling as an educated attempt to get specific details about a certain type of suspect, gather a collection of leads and get a biographical picture of tendencies, trends and behavioral patterns. Holmes (1989) continues that narrowing down the field of investigation is the goal of profiling. It identifies consistencies in the method of committing crime linked to an offender’s personality and behavior at the crime scene.

Different authors classify criminal profiling in a number of ways, according to Ross (2010). These include geographic profiling, victimology, investigative psychology, criminal investigative analysis and behavioral evidence analysis.

- Geographic profiling can either look at an offender’s previous crimes to predict their location or locations that are frequented by crime instances. Profilers tend to focus on the likely location the offender frequents for entertainment and recreation, home or work address or travel route. This can also include looking at patterns in the timing to come up with profiles of likely offender’s.

- Victimology is where by the victims characteristics are studied and analyzed. The victim’s characteristics can give themselves to inferences on the offender’s modus operandi, signature behaviors and motive. The profiler will carry out a risk assessment of the victim, this includes the amount of risk their lifestyle places them in, the amount of risk the offender was willing to take in the crime and the amount of risk the victim was in at the time of the crime.

- Investigative psychology uses psychologists as profilers who use peer-reviewed research to determine facts about an offender using true psychological techniques. Also there is criminal investigative analysis also known as crime scene analysis established by the U.S.A Federal Bureau of Investigation (F.B.I). Patterns of criminal behavior and groups based on
those behavior are identified by profilers. Finally behavioral evidence analysis is performed where character traits and behavioural patterns are used in building a criminal profile.

In profiling the crime puzzle, Mburu (2014) explains that in the presence of crime generators, clusters of crime emanate as well as criminals who successfully commit crimes, feel encouraged to carry out more in the same surroundings. The offender will have an area or zone they consider safe to commit crimes. This safety area will then start expanding over time and welcome serial offending and repeat victimization. LEAs focus and attention should be on discoveries and managing those clusters via anticipatory strategies towards the ultimately goal of crime prevention.

2.5.3 Using Twitter to profile crime

There have been various social media big data studies, Braga and Bond (2008) conducted one utilizing emerging computational criminological methods. This study was exploratory in nature and collected, transformed, analyzed and linked ‘big social data’. This was to look at the crime pattern estimation problem with an aim of building big data statistical models.

Other studies such as Tumasjan, Sprenger, and Sandner (2010) looked at the German general election and measured Twitter sentiment in relation to the candidates and concluded that this data was as accurate as the voting patterns.

In another study Asur and Huberman (2010) looked at movies sentiments and posts from Twitter, correlated the frequency and concluded that this prediction method was more accurate than the Hollywood stock market. Sakaki, Okazaki and Matsuo (2010) compared estimates between conventional geological sensors methods and Twitter data on the epicentres of earthquakes and found the latter more accurate.

Such studies show that social media or Twitter can generate relevant data from naturally occurring events, and use that to augment and complement conventional data to study offline phenomena occurrences.

Ann ecological analysis of crime in London experiment carried out by Matthew L. Williams (2016) using data from Twitter to test the hypothesis that actual police crime rates are associated with crime related tweets. Computational methods handling new forms of online social data, allow criminologists to gain meaningful insights into modern social processes at exceptional speed and scale. The challenge however comes in marshaling these new forms of data.
The Williams et al. (2016) study made an assumption that within an offline phenomenon, each Twitter user is a sensor. In the manner of Raudenbush and Sampson (1999), the study considered these nodes or sensors as part of a wider sensor-net for systematic social observation covering ecological zones. Abbott (1997) refers to these sensors that observe natural phenomena as the feel, sights and sounds of the streets.

According to Wilson and Kelling (1982), the case of ‘broken windows’ includes minor public incivilities such as litter, graffiti and drinking in the street. These serve as signs of unwillingness of residents to intervene in a crime, call the police or confront strangers, signs that enable potential criminals.

The information on physical and social disorder can be published by sensors in four ways: as perpetrators, first hand witnesses, second-hand observers (e.g. spread of rumours or media reports) and as victims. Williams, Burnap and Sloan (2016) looked at these four publishing modes of Twitter as signatures of crime and disorder that have various characteristics. Data from these sensors includes spatial and temporal information. Some are based on variation in perceptions of disorder while others are activated such as published tweets.

Sensors are not always online as they be offline such as sleeping or working. Their data may be noisier from curated data as interpretation of certain contexts may be difficult such as rumors and sarcasm. Nevertheless the number of Twitter users referred to as sensors by Williams et al. (2016) is impressive. With over 321 million Twitter users and over 500 million tweets globally, 2.2 million of those users in Kenya.

### 2.5.4 Developing a Framework for Profiling Crime

This section looks at previous researches on testing frameworks that have previously been created in the past by a variety of researchers.

Tobin and Christpher (2007) came up with a solution framework for crime related weblog that includes visualizations, contents analysis and links. The main framework pillars are: 1. specific community topics extraction, 2. stipulating the relationship between social network bloggers, 3. Sentimental and content analysis, 4. Visualization of different abstraction levels. The framework also included some searching techniques, ranking algorithms with HITS algorithms, top N documents from seed set and, independent and dependent neighbourhood graphs.
They did not use NLP because use of the sometimes improper language structure in blogs, although ignoring this will increase false positive results and reduce the accuracy. Due to operating in blogs it did not verify results, no elimination techniques of false positives and there were no recommendations on implementation technologies for this framework (Tobin and Christopher, 2007).

A method was created by Paxson and Zhang (2011) to identify both verified and automated Twitter accounts based on tweeting behavior. The study found out that automated tweets have a structured pattern while organic ones are randomly distributed. In terms of verified accounts, the study showed that celebrities are the majority as well as poplar companies.

Unexpectedly, verified users have more automated accounts as opposed to non-verified accounts, at 6.9% and 16% respectively. Source of tweets were API’s for automated while organic ones mostly used their phones. The study however did not include number of followers, age of accounts and sentiment or content analysis for improved accuracy (Paxson and Zhang, 2011).

Ratkiewicz, Meiss, Goncalves, Conover, Flammini and Menczer (2011) framework uses graph analysis to analyze the behavior of users in Twitter feeds. This framework, based on topologies of the Twitter communication network is supposed to identify non-organic and organic tweets. It is represented by memes and events, which represent information and actors respectively. Weights are allocated on basis of certain units and edges in the network. They however did not use sentiment analysis with the graph analysis, and did not include user evaluation and classification of the tweet accuracy, which increases false negative and false positive.

A study was proposed by Li, Lei, Khadiwala and Chang (2012) of a system made up of offline and online processes. The offline process retrieves data from an API crawler then extracts the environment related tweet according to the classifiers, then carries out indexing and storage of the data. The online process contains clustering and ranking which is intended to search online for analytics and detect events.

Tweets are grouped according to their timeframe and geo location by a clustering model. Tweets were extracted using various techniques such as linear regression, ranking hashtags, user profile analysis and number of retweets. However it did not show how to correlate a hashtag to an event.
or distinguish its relevance, and only used enable geo location options on the user side (Li et al., 2012).

In looking at non-textual data of real life events such as the football world cup. Chierichetti, Kleinberg and Kumar (2014) looked at user interactions and influencers of the event interact, as well as extraction of raw tweets and retweets of these events using linear classifier methodology.

This Chierichetti et al. (2014) study found out that users are less social on the platform during the event but activity increased immensely after an event occurred. However the accuracy of the results were in doubt due to identifying credible sources of the event. Furthermore the methodology does not identify the good or bad impact of the event to society.

Gerber M., (2014) carried out a study using Twitter statistical and linguistic analysis in crime related topics, using crimes in Chicago city, USA as a sample, specifically narrowing in on crime types and geo location records. The study looked at geo location tagged tweets in historical data, comparing the Twitter feature analytic and traditional linear regression. He found out the Twitter data improved the results but location of crime predictions were not given to assist LEAs. Also he did not identify criminals based on network analysis of the accounts, look at trending topics for better credibility and accuracy.

A study similar to the one above was done by Sarvari, Abozinadah, Mbaziira and Mccoy (2014) but in addition used eigenvector and page rank algorithms in social network analysis. The data used was leaked from the Nigerian advanced fee fraud scammer data theft service. They then got Facebook profiles of related criminals and their friends coming up with social network analysis that includes ranks of centrality for key members. Although the study did not validate email address communications for identification and strengthening of the relationships.

Using a subset of the AIntP-3 data model from the North Atlantic Treaty Organization (NATO) model, Kaati, Rezine and Berzinji (2012) proposed graph algorithms to identify key network nodes in a terrorist network. They found out that the financial manager is the most important player but did not show how to detect the terrorist network in the diffused social network. Also they did not show how to map actors or nodes to entities only explaining the application of graph analysis of the NATO model using data from the news and not social media categories extraction.
Detection of crime patterns was done in a study by Bolla (2014) using social media in two domain; intensity and sentiment analysis on the tweets, and with respect to certain cities, geo based analysis of the tweets. They looked at subjective analysis that is opinion and not fact based, to do this they used a number of sentiment analysis tools. Using cut based and Naïve Bayes classification having both multiclass and binary classification.

Bolla (2014) continues to explain that binary is made up of negative and positive while multiclass is made up of: strong negative, negative, neutral, positive and strong positive. The study also used Recursive Neural Tensor Network sentiment analysis and normalizing emotional rating of dictionary based English words.

For the 2011 Irish general elections Smeaton and Bermiongham (2011) collected data using hashtags of names and abbreviations of selected key terms. They compared traditional volume based analysis and sentiment analysis. They found volume based successful but experienced failure with sentiment analysis. The failure was not addressed but they had not distinguished between peoples reaction to the news and people preferences. Also the credibility of the tweets and errors within them.

Ramteke, Shah, Godhia and Shaikh (2016) also analyzed election results from tweets collected via Twitter API and then used hashtag clustering to manually label the data. They classified the emotions polarity of sentences using Valence Aware Dictionary and Sentiment Reasoner (VADER). They then used various sentiment analysis techniques multinomial Naïve Bayes, Support vector machines (SVM) lib linear, SVM rbf kernel and SVM linear kernel, concluding the latter was the most accurate. The studies sentiment analysis was good but failed in automating the process requiring extraction of tweets from JSON to CSV format for analysis. They also did not compare old and new data.

2.6 Testing a Framework

In regards to literature on testing a framework, studies within experiment hypothesis testing, traditionally researcher randomly select the sample from the population according to Eisenhardt (1989). The goal of the sampling process in this type of study is to obtain accurate statistical evidence within the population on the distribution of variables.

According to Eisenhardt (1989) multiple data collection methods are combined by theory-building researchers. While archival sources, observations and interviews being particularly common,
inductive researchers are not confined to these choices. Different investigators employ only some of these data collection methods.

Various test exist such as usability tests explained by Satam, Taslim, Adnan and Manaf (2016) as having five common characteristics. These characteristics look at (i) participants viewing the framework, (ii) goals of the framework, (iii) real users of the framework, (iv) problems and recommended changes and (v) participants doing real tasks. An innovative experimental design and research methodology that can accommodate these characteristics may provide shared understanding between practitioner and academician.

Various researchers have used the purpose of the Pearson's correlation statistic is to test the degree of association between numerical variables. For example Kang and Kang (2017) use it in their work when predicting crime occurrence from multi-modal data. Ellaway and Shareck, (2011), used Pearson correlation to show how the attributes "objective" and "perceived crime" measures were not strongly correlated with Pearson correlation coefficients ranging between 0.20 among women and 0.25 among men.

2.7 Conceptual Framework

In order to explain a phenomenon and come up with a conceptual framework, synthesis of literature including observations and views of previous researches was necessary. This framework will identify and try to come up with an understanding of how various variables in this study connect.

McGaghie, Bordage and Shea (2001) describes that for the presentation of a particular research question that drives the study being reported based on the problem statement, the conceptual framework sets the stage. The problem statement of a thesis gives the issues and context that made the researcher conduct the study. This study will draw from the theoretical framework given in section 2.1 on the makeup of profiling crime.

The three theories of rational activity, crime pattern and rationale choice theories seem to overlap in regards to criminal offenders and crime in general. The three theories have similarities that revolve around certain variables. In rational activity it is the victim, offender and crime. In crime pattern theory it is edges, paths and nodes, and for rational choice theory it is the offender and their choices.
In looking for a unifying conceptual framework encompassing the three theories, this study has come up with a number of independent variables and a dependent variable. The dependent variable is crime, this requires input from other variable to exist. A framework in Figure 2.5 by Kshetri (2005) is based on cyber-attack patterns globally. Even though this framework deals with another issues, some aspects of it can be borrowed for this study. The building blocks connecting of sources, targets and mechanisms within it, can be applied to this studies research.

**Diagram/schematic of theory**

![Diagram](image)

**Figure 2-5: A proposed framework for global cyber-attacks. (Source: Kshetri (2005))**

### 2.8 Chapter Summary

This chapter looked at literature review in regards to the research objectives and subject area. It began with looking at the theoretical foundation by bringing out the major theories that exist, the degree to which these theories have been investigated, the relationships between them, and in the end developing a combined view on how the different theories can be looked at. This can now help the research in coming up with a more concrete and sound framework based on existing tried and tested crime theories.

This chapter then delved into the specific research objectives digging deeply into each aspect of the objectives. In the first objective, identifying and extracting crime data from social media was examined. Social media, big data and the characteristics surrounding it, were discussed highlighting how to identify and extract crime data from.

This review then looked into the second objective, sentiment analysis on crime data, where past studies on sentiment analysis as well as studies on crime data, including various approaches and
techniques were discussed. Finally this review looked at the last objective, profiling of crime in Kenya, using the Twitter social media platform. Previous studies that looked at crime in Kenya, using Twitter to profile crime and related frameworks were examined.
Chapter 3: Methodology

3.1 Introduction

This chapter will look at the methodology used in carrying out the research for this study. The research is intending to come up with a framework to profile crime in Kenya using Twitter. This therefore means that the ideal data to gather and analyze is from Twitter itself. The selection of tools needed to gather this data had to revolve around social media data gathering instruments.

In this study, the key to gathering data was identifying, extracting and eventually analyzing that data in various ways, including carrying out sentiment analysis on the data. Calculating and selecting the population size to use and sampling techniques was necessary given the vastness of social media platform.

3.2 Research Design

Crime and criminal activity is a constant and continuous threat to society that occurs either in a premeditated manner or spontaneously. Law enforcement agencies have key roles in stopping or responding to perpetuated crime and the factors linked to it such as the victims and offenders. In order to monitor and respond to crime effectively, LEA need mechanisms in place that notify them of criminal acts occurring in their area of jurisdiction.

Traditional methods of reporting crime include calling the police hotline which in Kenya is 999, 112 or 911, sending a message via official LEA websites or physically reporting to a police officer or a police station. These methods although accessible are not effectively utilized. According to the National Crime Research Centre (NCRC) chief executive Gerald Wandera, approximately 70% of crimes that occur in Kenya are not reported by citizens because of fear of visiting police stations (Mukindia 2017).

Twitter is a highly popular online social networking platform that allows users to post a message referred to as a tweet, composed of a maximum of 280 character and having it instantly accessible to over 320 million Twitter users’ worldwide as well as non-registered users who have access to the internet. Kenya has 2.2 million monthly users with 1 million of them using Twitter daily (Minter 2017).

With so many people using Twitter to tweet on various topics including crime related activities, makes this platform a good source of information for this subject matter. Twitter Inc. Company
allows developers to connect to the Twitter system via an Application Programming Interface (API) allowing the researcher to search on previously posted tweets by its users. These tweets can then be extracted, stored and then analyzed by the researcher.

Once the data is extracted, the researcher can then carry out sentiment analysis by using computational linguistics, text analysis and natural language processing to identify and extract subjective information from the tweets. This will help determine the overall contextual polarity and the author’s attitude within the text.

Upon completion of the sentiment analysis processes, the researcher can then look at categorizing the tweets and building a framework working with the variables stated in the conceptual framework. This includes both dependent variables of location and offender, and independent variable of victim.

3.3 Population and Sampling Design

3.3.1 Population

According to Minter (2017), Kenya has 2.2 million monthly Twitter users with 1 million of them using Twitter daily. An analysis done by Sochin Ltd by the BusinessWatchTeam (2017), a strategic communications consulting firm which looked at 728,074 Twitter posts from Kenya and their specific geo-tagging, found results of 85 percent referenced Nairobi, 6.1 percent Mombasa, 3.8 percent Nakuru, 2.5 percent Kisumu and Uasin Gishu. This study showed that Twitter users are mostly concentrated in Nairobi.

3.3.2 Sampling

For population size since it is not possible to separate twitter users who will tweet about crime from the rest of the tweets, therefore the entire total Kenyan Twitter population figure of 2.2 million users was used. This involved looking at all tweets within Kenya that users had sent, then sampling the data I needed using specific keywords.

3.4 Data Collection Methods

This section will describe the methods, instruments and tools used to collect the data. There exist a wide range of data collection methods for various research objectives. These can be specific to qualitative techniques, quantitative techniques or both. It is worthy to note that although qualitative
and quantitative data collection methods are different, each has its own value. StatisticalTrainingUnit (2018).

Various techniques exist such as interviews, questionnaires, surveys, observations, documents, records and focus groups amongst others. Interviews are whereby a researcher asks a specific set of questions to a subject. This can be in the form of telephone call, face to face or Computer Assisted Personal Interviewing. This was impractical considering the size of the sample.

Another technique is questionnaires which make use of rating scales and checklists. A rating scale is useful for monitoring behavior on a continuum. A checklist is a list of characteristics, behaviours or other entities the study is looking for. The survey participant or researcher simply checks if each item on the list is observed, true, present or vice versa. Due to the type of data needed for this study, a questionnaire could not sufficiently cover the sample size needed.

Document review is whereby existing documents are looked at for relevance to the study. Focus groups is a group interview of a small group of people who have similar characteristics. The group is guided by a facilitator on specific set of topics and they participate in discussions. Both these methods are not idea for this study as the Twitter platform and amount of data needed will not fit into these data collection methods

Majority of the data required for this study was to come from Twitter. The data had to be identified and extracted. This will involve coming up with a list of keywords to use as a guideline in collecting data. These keywords will be used in conjunction with an Application Programming Interface (API) to scrape Twitter for specific data.

3.5 Research Procedures

As stated in the previous section, the first step was to identify and then extract the data from Twitter. Using the twelve prevalent crimes the research had identified, a list of keywords to use when searching for data on Twitter was developed. The key word “Crime” was added to allow collection of more related data. The list of crimes targeted are described in List 3-1;

<table>
<thead>
<tr>
<th>No.</th>
<th>Targeted crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Terrorism / terrorist / terrorism / terrorist</td>
</tr>
<tr>
<td>2.</td>
<td>Carjacking</td>
</tr>
<tr>
<td>3.</td>
<td>Muggings</td>
</tr>
<tr>
<td>4.</td>
<td>Rape</td>
</tr>
</tbody>
</table>
5. Murder  
6. Gun related crime  
7. Drugs / drug dealing / drug dealer / drug user / drug peddler  
8. Pickpocket  
9. Thief / thieves / theft  
10. Robbing / robbery  
11. Burglar / burglary  
12. Vandal / vandalism  
13. Kidnap / kidnapping  
14. General crime data 

List 3-1: Crimes targeted from Twitter data (Source: Author) 

Once the keywords were identified, a Twitter search which was conducted. This involved using an API and the Twitter data set to extract the data. To gain access an independent Twitter account was set up for this research with the user as “Project_254” and API application named “Tazama Uhalifu”

The API took used were Tweeter Archiver and PlusOne Social. Twitter Archiver allows searching and saving tweets into Google Spreadsheets from keywords. The tweets are then saved automatically into spreadsheets. With Twitter Archiver you can search a specific multiple keywords geographical area within the search parameters.

PlusOne Social is similar to Twitter Archiver where keywords are used to search Twitter via geographic location, but instead stores its data into a Microsoft Access database. Once the search is initiated, the tweets and all the key attributes are stored in an MS-Access database.

These tools however allowed the collection of tweets within the past 7 days. This is the standard API available to all users that also has restrictions on rate of tweets and requests to the system per user. This study however required more data to address this challenge the researcher contacted Twitter Inc. after which access was grant to historical data. This study therefore proceeded with collecting historical data.

After the identification and extraction, the data was cleaned by removing duplicates then an overall sentiment analysis was carried out. This involved both with all the data and with stop words removed. The data was then categorized in terms of location, time and authors.
The location data plotted on maps and the author’s data further analyzed to show trends and patterns

3.6 Data Analysis Methods

To achieve the second research objective Sentiment analysis was used to determine an author’s attitude to a document’s overall context polarity or subjectivity on a particular topic. The study carried out sentiment analysis from the tweets extracted using the SentiStrength tool which is a lexicon-based classifier algorithm.

For the analysis of data, I used a number of quantitative techniques and primarily carried out sentiment analysis. The main sentiment analysis algorithm I used was a lexicon-based classifier called SentiStrength. SentiStrength comes included in Mozdeh, a free Microsoft Windows program for sentiment, content, gender, time series and keyword analysis of social media text.

SentiStrength’s output is made up of two integers, a positive and negative sentiment scored from 1 to 5, where 1 signifies no sentiment and 5 for strong sentiment for both. A neutral text would be shown as 1, 1 while a text score of 3, 5 would have a moderate positive sentiment and strong negative sentiment. The goal of using two scales of negativity and positivity is to detect the sentiment expressed rather than its overall polarity.

Although Mozdeh can collect tweets as well, it was still restricted by Twitter standard API. Mozdeh also has an option where extracted tweets from third party software can be imported in any number of database formats. The data was imported in a tab-delimited format data for analysis.

The data was analyzed in a number of ways, the first analysis method revolved around the environment. In regards to the environment I carried out analysis on the location of the reported crimes and mapping them graphically on a map. The frequency of those locations mapped was also included in the analysis.

3.7 Ethical Considerations

A researcher should follow research ethics with a code of moral guideline on how to conduct a study in a morally acceptable way. These guidelines seek to prevent researchers from engaging in scientific misconduct such as not maintaining the privacy and confidentiality of the participants. The researcher did not show any bias in this study.
Other ethical issues for this study were considered. All data scraped from Twitter was not displayed apart from the key words relating to crime issues. The authors of the tweets were not named and only identified via a unique number. In regards to the usability tests, the participants were informed their inputs will be treated with the highest confidence and no names or names of their respective department will be mentioned.

3.8 Chapter Summary

This chapter explained the process of identifying, gathering, analyzing, and processing information necessary for this research study. The chapter goes through why it was necessary to carry out the research to come up with a new framework for profiling crime. Specific strategies utilized were mentioned in regards to studying the underlining research objectives.

The dynamics of the Twitter platform and specifically the situation in Kenya was discussed. This assisted in identifying a population sample size and sampling technique to collect the data.

This chapter also looked at the data collection method of APIs and the tools that can be used in conjunction with them. An overview of data analysis method were introduced where various analysis methods will be used as well as sentiment analysis techniques.
Chapter 4: Model

4.1 Introduction

This Chapter will look at the process of building a crime profiling model from Twitter data. The model will be a representation of the research objective and a study of the properties of the system. This will involve guiding and describing the process of translating research into practice, explaining and understanding the influences to implementation outcomes (Nilsen, 2015).

The literature review carried out in Chapter 2, aimed to analyze and describe different existing frameworks, models and theories. The model and modelling process is closely related to those theories. Although the model is drawn from the theories, in this Chapter it is refined with a view of focusing on profiling of crime.

The proposed model is used to guide for the research to achieve the objectives outcome. This outcome or framework will be defining a specific viewpoint that this study will use in analyzing and interpreting the data gathered and understanding the variables. These concepts will act in building knowledge by challenging or validating theoretical assumptions (Nilsen, 2015).

The process of conceptual modeling is then used to create a technique or process that can bring about a better understanding of this system being modeled. To be effective or efficient the conceptual model, this study will need a criteria for comparison with the real world data. This model will take all the variables into account and try look at the relationships and functionality and eventually explaining the data (Nilsen, 2015).

4.2 Analysis

This study is addressing the issue of having a framework to profile crime. Crime does not exist in a vacuum and in order to exist, certain components or variables must come together. Crime data is usually limited to certain sources and is not usually reflective of the real world situation. By having social media platforms such as Twitter, the platform allows users to give out crime information and this study aims to build a framework and model from that data.

The objective of this study is to profile crime in Kenya from Twitter data. This involves having a plan on specific data variables and analyzing the data. Before any analysis of the data was to commence, this study compared the variables and their attributes against the data attributes that can be collected.
This study aims at looking within crime related Twitter data and to find out what variables and components can be extracted in relation to crime. Extracting these variables, then coming up with a model to solve the research objective and test the validity of that model and if possible predict crime information.

Using the conceptual framework from Chapter 2 section 2.7, this study looks at the dependent variable of crime, together with the relationships of two independent variables, environment and author. From reviewing literature on various theories, this study arrived at a particular theoretical framework for use in creating a model.

To make the data from the variables relate to the study after extraction, this study will look at the data on individual attributes to understand the variables. For example, environment variable data is hard to specify, but by collecting data on the location and time will help build information on the environment variable.

4.3 Modeling

This topic deals with the actual modeling of the variables that will form the blueprints for the crime profiling framework. A review of existing literature exposed various variables that can be obtained for crime mapping. These include environment and author. For purposes of this research, data was collected from social media.

The variables that can be obtained from twitter include environmental related variables such as location, time and the area characteristics. The author related variables such as victim type, frequency and level of influence. These variables were used to inform our model. The independent variables of environment and author directly impact on the dependent variable, crime reports. These variables collectively build on the researches objectives and are expounded on in Figure 4.1

The ideal model will be drawn up and needs to be adapted to the specifics of the Twitter data environment. This will enable for a better effective synergy between the data and framework. This study will take data from the variables of environment and author, based on their attributes. These attributes will help the research build on the study and test the model.

The determinant variables hypothesized from Fig. 4-1 were seen as ideal in influencing the outcome of crime reports as each variable has its corresponding attributes. Therefore these
attributes eventually impact on the dependent variable of crime reports. The research study simplified and refined all the factors available to create this model that would cater for specific measurable and collectable data. This research proposed model on the profiling of reporting on crime is presented in Fig. 4-1.

**Figure 4-1: Research variables and their attributes (Source: Author)**

In this study, data collection needed to correspond with the attributes. Therefore the collection of data had to be pinned on specific attributes. Firstly the Twitter data will have to be extracted from the sample size which is the entire Kenyan Twitter population size. The data will be classified based different attributes and analyzed.

Twitter data has different data fields but the data extracted is in streams and need to be stored in a database. Different extraction tools are used in conjunction with the Twitter API has to sort out the data attributes and store. Once data is in database form is when it can be analyzed easier and the researcher can use necessary attributes.

### 4.4 Testing

Once the study carries out this research, it is necessary to test the outcomes of the results. This study will carry out three test, a usability test, correlation coefficient test and collaborative analysis. The main outcomes for the usability test is to look at (i) participants viewing the framework, (ii) goals of the framework, (iii) real users of the framework, and (iv) problems and recommended changes (Satam et al. 2016).
The usability test in this study involved giving six government crime analysts a demonstration of the framework and having them interact with it. The analysts then had a look at the model and were then given a questionnaire with three questions to fill in answers. The following questions were asked:

1. Would you a crime analyst use this framework, and why?
2. What aspects of the framework did you like best?
3. What improvements can the designer of the framework make?

The second test this study will carry out is Pearson correlation coefficient. This test will take the variables in the framework model. From the variables, the study will look at the attributes that make up the variables and were extracted from the data. These variables will be correlated against the thirteen crime topics to find out the significance of those variable attributes. This test allows the study to identify the key variables and attributes in the framework model.

The third test will be a collaborative analysis between the results and findings from this study and existing National Police records on the ground. This will enable this study to confirm or refute the model as being reflective of the real situation reported to the Police.

**4.5 Chapter Summary**

This Chapter looked at model and modeling process planned for the proposed framework. The Chapter explained the evolution from real life theories to the conceptual framework and finally the model. This process allows translating research from objectives into the eventual outcome and why it is important to build on previous existing knowledge.

The Chapter also talks about using Twitter and extracting the variable specific data with it. To be able to build a realistic and testable model, it was necessary to study Twitter database field and come up with a plan of extracting them. The Chapter also talks on being able to analyze the variables and attributes that will be extracted.

This Chapter continued to look at the actual modeling, picking the dependent and independent variables along with attributes. Without clearly identifying the variables, the eventual model will not be practical. The Chapter also discussed the sample size and the process that will be used to collect the data.
Finally this Chapter looked at the testing process what types of tests, in this case was usability, correlation coefficient and collaborative comparison analysis. Outlining the desired outcomes of those tests. This Chapter also listed the procedures to be followed on the eventual tests of the model. These included creating a simple questionnaire to be distributed to real world crime analysts in the field as well as factor analysis on different datasets.
Chapter 5: Results and Findings

5.1 Introduction

This Chapter is where findings of the research are reported based on the methodology that was applied to gather data. The results in this section are a true statement of the findings of the research arranged in a logical sequence without interpretation or bias. The findings will be used later in Chapter six to explore whether the research objectives were attained. (Annesley, 2010)

The results and findings in this Chapter will be structured according to the research objectives making sure it flows from beginning to end. In some cases, results of tables will be presented with a small description such as comparisons, but the reasons why certain differences or correlations occur will be discussed in Chapter 6.

As mentioned above, this Chapter is structured following the research objectives. This will enable an understanding of each objectives findings as they all touch on different areas. A combination of descriptions, tables and charts will be used to display each of the results. The results from the dataset and analysis is huge and therefore the findings displayed here are only analyzed by key results.

5.2 Identifying and extracting crime data from social media.

In line with the first objective it was necessary to identify and extract crime data from Twitter. As previously mentioned, crime data on Twitter was not enough and therefore the entire Kenyan Twitter population was used. This consequently entailed separating tweets related to crime from the rest. This was done by searching using the keywords shown in Table 5-1.

<table>
<thead>
<tr>
<th>No.</th>
<th>Search word</th>
<th>Targeted crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Terroris</td>
<td>Terrorism / terrorist / terrorism / terrorist</td>
</tr>
<tr>
<td>2.</td>
<td>Carjack</td>
<td>Carjacking</td>
</tr>
<tr>
<td>3.</td>
<td>Mug</td>
<td>Muggings</td>
</tr>
<tr>
<td>4.</td>
<td>Rape</td>
<td>Rape</td>
</tr>
<tr>
<td>5.</td>
<td>Murder</td>
<td>Murder</td>
</tr>
<tr>
<td>6.</td>
<td>Gun</td>
<td>Gun related crime</td>
</tr>
<tr>
<td>7.</td>
<td>Drug</td>
<td>Drugs / drug dealing / drug dealer / drug user / drug peddler</td>
</tr>
<tr>
<td>8.</td>
<td>Pickpocket</td>
<td>Pickpocket</td>
</tr>
<tr>
<td>9.</td>
<td>Thie</td>
<td>Thief / thieves / theft</td>
</tr>
<tr>
<td>10.</td>
<td>Robbery</td>
<td>Robbing / robbery</td>
</tr>
<tr>
<td>11.</td>
<td>Burglar</td>
<td>Burglar / burglary</td>
</tr>
</tbody>
</table>
If the keyword search was done in Twitter it would return results from the entire world, it was therefore necessary to come up with geographical location parameters to restrict the search to within a certain area. This involved coming up with points on the map that will have a radius large enough to cover the areas of interest.

To determine the actual ground area, a tool called “Gmaps-radius” was used. It that allows a user to draw a specific circular radius on Google Maps. In Figure 5-1 is a geographic representation of the area which Twitter data was collected from.

![Geographical location coverage of tweet collection.](image)

Once the keywords and geographical range had been identified, the study needed a way to search Twitter. This involved coming up with an Application Programming Interface (API) and Twitter application within the API that will grant me access. The study set up an independent Twitter account named “Project254” and API application named “TazaUhalifu” as seen in Figure 5-2.

![Twitter Profile](image)
Once the API was ready for use, a tool that could connect via the API and search Twitter was required. The tools were identified, Tweeter Archiver and PlusOne Social.

Twitter Archiver allows searching and saving tweets into Google Spreadsheets from keywords as seen in Figure 5-3. The tweets are then saved automatically into spreadsheets. With Twitter Archiver you can search a specific multiple keywords geographical area within the search parameters.
PlusOne Social is similar to Twitter Archiver where keywords are used to search Twitter via geographic location, as shown in Figure 5-4, and stores data into a Microsoft Access database.

Once the search is initiated, the tweets and all the key attributes are stored in an MS-Access database as shown in Figure 5-5. After assessing both tools, it was clear that Twitter Archiver extraction capabilities were limited, as it was extracting very few tweets by sampling. PlusOne social on the other hand extracted all the tweets based on the parameters.

5.3 Sentiment analysis on crime data

Once the data was identified and extracted successfully, the second objective involved carrying out sentiment analysis on the data. Due to multiple collections of tweets, some of the data had duplicate tweets. Therefore duplicates had to be removed within the database as seen in Figure 5-6.
As mentioned in Chapter 3, for the sentiment analysis algorithm, a lexicon-based classifier called SentiStrength was used. This algorithm is included into a piece of software called Mozdeh that was used for this study. A screen shot showing the process of sentiment analysis within Mozdeh and results of tweets with corresponding sentiments is given in Figure 5-7.
Figure 5-7: Sentiment analysis of all the tweets

A general representation of the total number of words and their frequencies with the stop words such as, the, to, of, a, etc. removed, the dataset, the total number of words reduced as seen in Table 5-2. Table 5-2 shows a summary of the top twenty four words from the highest negative sentiment to the lowest. The words in Table 5-2 appear the most frequent in the tweets with the highest negative sentiment.

Table 5-2: Top 24 words with highest negative sentiment

<table>
<thead>
<tr>
<th>Word</th>
<th>VocabCount</th>
<th>SubjectiveCount</th>
<th>Av Pos Subj</th>
<th>Av Neg Subj</th>
</tr>
</thead>
<tbody>
<tr>
<td>abuse</td>
<td>133</td>
<td>131</td>
<td>1.191</td>
<td>4.145</td>
</tr>
<tr>
<td>murder</td>
<td>683</td>
<td>672</td>
<td>1.185</td>
<td>4.088</td>
</tr>
<tr>
<td>rape</td>
<td>490</td>
<td>468</td>
<td>1.246</td>
<td>4.049</td>
</tr>
<tr>
<td>attempted</td>
<td>128</td>
<td>127</td>
<td>1.157</td>
<td>4.008</td>
</tr>
<tr>
<td>moi</td>
<td>212</td>
<td>193</td>
<td>1.119</td>
<td>3.938</td>
</tr>
<tr>
<td>death</td>
<td>107</td>
<td>106</td>
<td>1.009</td>
<td>3.887</td>
</tr>
<tr>
<td>school</td>
<td>125</td>
<td>109</td>
<td>1.119</td>
<td>3.872</td>
</tr>
<tr>
<td>keb</td>
<td>143</td>
<td>131</td>
<td>1.16</td>
<td>3.87</td>
</tr>
<tr>
<td>girl</td>
<td>262</td>
<td>223</td>
<td>1.099</td>
<td>3.825</td>
</tr>
<tr>
<td>charge</td>
<td>156</td>
<td>154</td>
<td>1.143</td>
<td>3.805</td>
</tr>
<tr>
<td>charged</td>
<td>183</td>
<td>180</td>
<td>1.028</td>
<td>3.778</td>
</tr>
<tr>
<td>case</td>
<td>138</td>
<td>125</td>
<td>1.208</td>
<td>3.68</td>
</tr>
<tr>
<td>officer</td>
<td>133</td>
<td>121</td>
<td>1.05</td>
<td>3.628</td>
</tr>
<tr>
<td>day</td>
<td>107</td>
<td>91</td>
<td>1.429</td>
<td>3.396</td>
</tr>
<tr>
<td>nairobi</td>
<td>217</td>
<td>192</td>
<td>1.219</td>
<td>3.385</td>
</tr>
<tr>
<td>police</td>
<td>253</td>
<td>208</td>
<td>1.125</td>
<td>3.332</td>
</tr>
</tbody>
</table>
Table 5-3 shows an overall sentiment set of results. The scale used is 1 to 5 for both positive and negative sentiment, where 1 signifies no sentiment and 5 strong sentiment for both. Positive sentiment means having a positive thing to say about crime e.g. “Crime has reduced in my area”. Negative sentiment is the opposite i.e. “Crime is too much in Nairobi” The overall total negative sentiment of 78.42% against an overall total positive sentiment of 23.92%, with the no sentiment results removed. This shows that all the tweets collected are heavily leaning towards negativity.

**Table 5-3: Sentiment analysis results figures**

<table>
<thead>
<tr>
<th>Score</th>
<th>Pos</th>
<th>Neg</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>16.61%</td>
<td>28.87%</td>
</tr>
<tr>
<td>3</td>
<td>6.71%</td>
<td>13.36%</td>
</tr>
<tr>
<td>4</td>
<td>1.10%</td>
<td>34.19%</td>
</tr>
<tr>
<td>5</td>
<td>0.00%</td>
<td>2.00%</td>
</tr>
</tbody>
</table>

The summary of figures in Table 5.3 is illustrated graphically below in Figure 5-8. Figure 5-8 shows two graphs, each depicting sentiment strength of positive against negative. Again it clearly shows heavy leaning of the results towards negative sentiment. These results show that the content of the tweets will be reflective to the real world crime situation, that is, naturally inherently negative. Therefore there was no need to separate the tweets as the majority are negative.

![Figure 5-8: Graphs depicting sentiment analysis of all the tweets](image-url)
5.4 Framework for profiling of crime, using the Twitter social media platform.

From Chapter 4, this research came up with a proposed framework model to profile crime using Twitter data. This framework uses three key variables and their relationships. These three variables are the environment, the tweet author and the crime itself.

The environment has various attributes that associated to it, key being location and the characteristics of the location. This study therefore had to extract details pertaining to the environmental attributes from the tweets stored. For location data analysis the study required to have certain search lists identified to run through the tweet database in order to separate the data needed for analysis.

The location search lists or dictionaries are used to assist in analyzing crime occurrences in different locations. The locations dictionaries start widely from the counties and regions, narrowing down through cities, towns and finally roads, streets and regions.

Using online geographic information software (GIS) analyzed data is mapped according to location frequency. The locations mentioned in relation to crime are in twenty seven counties and are shown in Figure 5-9 with the figures in Table 5-3. Nairobi has the highest mentions with 129, Mombasa second with 49 and Lamu third with 11. Figure 5-9 shows 27 of the 47 counties with crime mentioned in tweets for the dataset. This means that for the period, 20 counties in Kenya did not have any tweets about crime in their area.

Figure 5-9: Map of Counties with mentions of crime data
The results in Table 5-4 show that the vast majority of the tweets about crime are in Nairobi County. Meanwhile the least mentioned counties include Turkana, Kwale, Isiolo, Homa Bay and Garissa. This also shows that the further you go from Nairobi County, the fewer the mentions on crime occur. The counties that border other countries have very few or no mentions at all.

The distribution of crime on a map is presented in Figure 5-10 for different parts of Kenya. In Fig. 5-10.a) crime is shown concentrated within Nairobi city and only a few mentions in the satellite towns.

a) Nairobi and the environs.
Within Mombasa City, Fig. 5-10.b) shows crime spread between Diani / Ukunda, Likoni and Mtwapa.
b) Mombasa area.

Fig. 5-10. c) shows the mentioning of crime in parts of Rift Valley and Western Kenya. The majority of areas identified were major town such as Kisumu, Nakuru and Naivasha.

c) Rift Valley and Western parts of Kenya.

**Figure 5-10**: Mapping of Crime location in different parts of Kenya.

Further analysis of major neighbourhoods, areas, roads and streets within the dataset returned detailed results. Table 5.5 shows, frequency and type of tweets in specific areas. From the table drugs were the most common type of crime discussed on twitter. Further, tweets on drug related crimes are more prevalent in Likoni as compared to Ukunda and Likoni. Tweets on gun related crimes were not as common as drug related crime.

The results in Table 5-5 also show that murder is mainly in two cities. In Nairobi County, murder is mostly in Thika Highway, Kibera, Limuru and Mavoko while in Mombasa County it is
mentioned in Diani. During the period of collection, it is clear that a boda boda strike (motorcycle taxis) strike mentioned around Kenyatta Avenue, CBD, Ngong road and Haile Selassie avenue.

Table 5-5: Frequency and Type of tweets in specific areas.

<table>
<thead>
<tr>
<th>Area</th>
<th>Freq.</th>
<th>Crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likoni</td>
<td>15</td>
<td>Drug (abuse, trafficking)</td>
</tr>
<tr>
<td>Thika</td>
<td>10</td>
<td>General</td>
</tr>
<tr>
<td>Diani</td>
<td>8</td>
<td>Murder</td>
</tr>
<tr>
<td>Kikuyu</td>
<td>7</td>
<td>General (Thug killed)</td>
</tr>
<tr>
<td>Ukunda</td>
<td>6</td>
<td>Drug</td>
</tr>
<tr>
<td>Kisii - migori road</td>
<td>4</td>
<td>Drug (Bhang)</td>
</tr>
<tr>
<td>Thika Highway</td>
<td>4</td>
<td>Murder</td>
</tr>
<tr>
<td>Kibera</td>
<td>3</td>
<td>Murder</td>
</tr>
<tr>
<td>Kenyatta Avenue</td>
<td>3</td>
<td>Boda boda strike</td>
</tr>
<tr>
<td>CBD</td>
<td>3</td>
<td>Boda boda strike</td>
</tr>
<tr>
<td>Nairobi-Nakuru highway</td>
<td>2</td>
<td>Carjacking</td>
</tr>
<tr>
<td>Mumias</td>
<td>2</td>
<td>General Crime</td>
</tr>
<tr>
<td>Litein</td>
<td>2</td>
<td>Shooting</td>
</tr>
<tr>
<td>Limuru</td>
<td>2</td>
<td>Murder</td>
</tr>
<tr>
<td>Kitengela</td>
<td>2</td>
<td>General crime</td>
</tr>
<tr>
<td>Mtswapa</td>
<td>2</td>
<td>Drug</td>
</tr>
<tr>
<td>Eldama Ravine</td>
<td>2</td>
<td>Drug (illicit brew)</td>
</tr>
<tr>
<td>Siaya</td>
<td>2</td>
<td>General Crime</td>
</tr>
<tr>
<td>Landis road</td>
<td>2</td>
<td>Robbed, Gun</td>
</tr>
<tr>
<td>Building</td>
<td>2</td>
<td>Drug</td>
</tr>
<tr>
<td>Syokimau</td>
<td>2</td>
<td>Police sergent arrested</td>
</tr>
<tr>
<td>Dandora</td>
<td>2</td>
<td>Crime general</td>
</tr>
<tr>
<td>Naivasha</td>
<td>1</td>
<td>Gun crime</td>
</tr>
<tr>
<td>Mavoko</td>
<td>1</td>
<td>Murder</td>
</tr>
<tr>
<td>Eldoret – Kisumu road</td>
<td>1</td>
<td>General crime</td>
</tr>
<tr>
<td>Ngong road</td>
<td>1</td>
<td>Boda boda strike</td>
</tr>
<tr>
<td>Haile Sellasie</td>
<td>1</td>
<td>Boda boda strike</td>
</tr>
<tr>
<td>Mtswapa</td>
<td>1</td>
<td>Drug (Heroin)</td>
</tr>
</tbody>
</table>

A summary of crimes by frequency across all counties is shown in Table 5-6. The results show that drugs and murder were the most common types of crime. Carjacking and gun related crimes were the least common.

Table 5-6: Summary of crimes in Table 5-5.

<table>
<thead>
<tr>
<th>Related crime</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>31</td>
</tr>
<tr>
<td>General Crime</td>
<td>25</td>
</tr>
<tr>
<td>Murder</td>
<td>18</td>
</tr>
<tr>
<td>Boda Boda Strike</td>
<td>8</td>
</tr>
<tr>
<td>Gun crime</td>
<td>5</td>
</tr>
<tr>
<td>Carjacking</td>
<td>2</td>
</tr>
<tr>
<td>Police related</td>
<td>2</td>
</tr>
</tbody>
</table>
In addition to analysis on location and crime, this study carried out analysis on time periods such as hours of the day for tweets. Below in Figure 5-11 is a summary of that analysis. The results did not show any significant data as the time of tweeting is distributed evenly.

![Figure 5-11: Time period of the day authors’ tweet](image)

Another type of analysis covered in this study is looking at the authors and data relating to them. Authors are the creators of tweets, sometimes referred to as tweeters. Twitter users can follow authors and read everything an author posts. The more followers an author has, the more people will read their tweets and ultimately the more influence they have. Figure 5-12 shows a general summary of author information.

![Figure 5-12: Twitter users who follow the most users and Authors with the most followers](image)

Figure 5-12 shows the users with the most followers. The top four followed users are media houses or their products such as ntvkenya, citizentvkenya, KTNKenya and dailynation. Notably individual users are also most followed such as KoinangeJeff, C_NyaKundiH, bonifacemwangi, RobertAlai, xtiandela and MohaJichoPevu. These individuals are either media personalities or activists.
Table 5-7 is a summary of the top 10 authors of crime related data for the period of this study with various information on each. The authors are listed with their user system id, tweets and retweets of crime data, total tweets ever made, count of followers, if the account is location enabled, the location of the account was set up, the language they use and the date the account was created.

From the results users’ tweets and number of followers are not related. This can be seen with user 1 who tweeted 771 time on crime but only has 1743 followers. On the other hand, user number 5 tweeted 344 times on crime but has 1,846,754 followers.

Table 5-7: Top 20 authors with details

<table>
<thead>
<tr>
<th>No.</th>
<th>User System ID</th>
<th>Crime Tweets and Retweets</th>
<th>Followers Count</th>
<th>Geo Enabled</th>
<th>Location</th>
<th>Language</th>
<th>Account Created</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2853324261</td>
<td>771</td>
<td>1743</td>
<td>False</td>
<td>Nairobi, Kenya</td>
<td>en</td>
<td>31-10-14</td>
</tr>
<tr>
<td>2</td>
<td>4202526155</td>
<td>504</td>
<td>3072</td>
<td>False</td>
<td>Nairobi, Kenya</td>
<td>en</td>
<td>16-11-15</td>
</tr>
<tr>
<td>3</td>
<td>201220566</td>
<td>480</td>
<td>9915</td>
<td>False</td>
<td>Nairobi</td>
<td>en</td>
<td>11-10-10</td>
</tr>
<tr>
<td>4</td>
<td>2178508505</td>
<td>367</td>
<td>3324</td>
<td>False</td>
<td></td>
<td>en</td>
<td>12-11-13</td>
</tr>
<tr>
<td>5</td>
<td>70394965</td>
<td>344</td>
<td>1846754</td>
<td>False</td>
<td>Nairobi, Kenya</td>
<td>en</td>
<td>31-08-09</td>
</tr>
<tr>
<td>6</td>
<td>4761718329</td>
<td>308</td>
<td>5198</td>
<td>False</td>
<td>Nairobi, Kenya</td>
<td>en</td>
<td>10-01-16</td>
</tr>
<tr>
<td>7</td>
<td>343326011</td>
<td>292</td>
<td>740052</td>
<td>True</td>
<td>Nairobi, Kenya</td>
<td>en</td>
<td>27-07-11</td>
</tr>
<tr>
<td>8</td>
<td>25985333</td>
<td>269</td>
<td>1876780</td>
<td>True</td>
<td>Nairobi</td>
<td>en</td>
<td>23-03-09</td>
</tr>
<tr>
<td>9</td>
<td>9245516928489</td>
<td>263</td>
<td>136</td>
<td>False</td>
<td></td>
<td>en</td>
<td>29-10-17</td>
</tr>
<tr>
<td>10</td>
<td>181289232</td>
<td>195</td>
<td>11808</td>
<td>True</td>
<td>Nairobi - Kenya</td>
<td>en</td>
<td>21-08-10</td>
</tr>
</tbody>
</table>

Table 5-8 is an analysis of the top 10 users and the percentage of crime related tweets they send in relation to all the other tweets. Also added is the activity level, i.e. the total number of tweets they have ever sent since creation of their account.

Table 5-8 may be used to find users who specialize in tweeting about crime. For example user number has the most tweets on crime but compared to the total amount of tweets, crime takes up only 6.56%. On the other hand, user number 6 has 63.90% of all their tweets on crime. This table shows that some users concentrate more on crime tweets.
Table 5-8: Percentage of crime related tweets over other tweets. Also activity levels of the authors.

<table>
<thead>
<tr>
<th>No.</th>
<th>UserSystemID</th>
<th>Crime Tweets and Retweets</th>
<th>Tweets for the study period</th>
<th>Percentage of crime related tweets</th>
<th>Total Tweets since acc creation</th>
<th>Account Created</th>
<th>Avg tweets per year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2853324261</td>
<td>771</td>
<td>11756</td>
<td>6.56</td>
<td>211098</td>
<td>31-10-14</td>
<td>52775</td>
</tr>
<tr>
<td>2.</td>
<td>4202526155</td>
<td>504</td>
<td>1908</td>
<td>26.42</td>
<td>46971</td>
<td>16-11-15</td>
<td>15657</td>
</tr>
<tr>
<td>3.</td>
<td>201220566</td>
<td>480</td>
<td>3752</td>
<td>12.79</td>
<td>76962</td>
<td>11-10-10</td>
<td>9620</td>
</tr>
<tr>
<td>4.</td>
<td>2178508505</td>
<td>367</td>
<td>6570</td>
<td>5.59</td>
<td>420743</td>
<td>12-11-13</td>
<td>84149</td>
</tr>
<tr>
<td>5.</td>
<td>70394965</td>
<td>344</td>
<td>4978</td>
<td>6.91</td>
<td>389235</td>
<td>31-08-09</td>
<td>43248</td>
</tr>
<tr>
<td>6.</td>
<td>476178329</td>
<td>308</td>
<td>482</td>
<td>63.90</td>
<td>41283</td>
<td>10-01-16</td>
<td>20642</td>
</tr>
<tr>
<td>7.</td>
<td>343326011</td>
<td>292</td>
<td>1620</td>
<td>18.02</td>
<td>124319</td>
<td>27-07-11</td>
<td>17760</td>
</tr>
<tr>
<td>8.</td>
<td>25985333</td>
<td>269</td>
<td>3916</td>
<td>6.87</td>
<td>358308</td>
<td>23-03-09</td>
<td>39812</td>
</tr>
<tr>
<td>9.</td>
<td>92455169284209</td>
<td>263</td>
<td>4310</td>
<td>6.10</td>
<td>7734</td>
<td>29-10-17</td>
<td>7734</td>
</tr>
<tr>
<td>10.</td>
<td>181289232</td>
<td>195</td>
<td>2000</td>
<td>9.75</td>
<td>258154</td>
<td>21-08-10</td>
<td>32269</td>
</tr>
</tbody>
</table>

5.5 Testing of the framework.

As explained in Chapter four, the testing of the framework will be done using a usability test, correlation coefficient and collaborative comparison analysis. The first test as explained in Chapter 4 was a questionnaire made up of three questions relating to the framework given to six crime analysts. The questions and responses were as follows:

5.5.1 Usability Testing

i. Using the Framework

Overwhelmingly all crime analysts responded positively to the framework and agreed they would use it in their roles. They pointed out what they liked most from the framework is the ability to categorize different crimes and look at data from each. The analysts expounded that they would find it easier to follow crime trends per specific crime.
The analysts also pointed out the ability to record locations of crimes and track hotspots or patterns of crime would be beneficial. By having locations where crime is mentioned will allow crime analysts compare data from various other sources. The analysts were pleased with having an extra source of crime information.

ii. Framework Strong Points

The analysts mentioned some aspects of the framework were limited such as the geo tagging of the tweet authors lacking. This they found would have further given them more data to work with. Although the geo tagging itself does not indicate the location where the crime occurred, they still mentioned it.

Another aspect of the framework they did not like was the various steps and stages that the data must physically be processed through. They opined that they would very much prefer a highly simplified process where they just click and results are displayed.

iii. Framework Improvements

The analysts requested that although they understand the framework was a research study, any implementation in future should be simplified and very user friendly. As it is now, not many users are able to get the results on their own without assistance from the researcher.

Another improvement to the system according to the analysts is a real time monitoring or reporting system. This is having the framework and model automated in such a way that when crimes data is entered in Twitter, it is summarized and displayed. This can involve dedicated monitoring systems where the API runs on continuously capturing data and processing it.

5.5.2 Framework Testing

The second test is a Pearson correlation coefficient carried out on the framework to find out the relationships between the variables of environment and author, and crime reports. This was done by looking at observed variables and attributes within the datasets. This study collected six attributes related to the author and environment.
This study then coded the data as seen in Table 5.8 according to the various attributes.

To carry out the statistical calculations, IBM SPSS Statistics software was used. The data was checked for a patterned relationship amongst the variables in the Correlation matrix in Table 5-9. The correlation technique used was Pearson correlation coefficient.

From the correlation matrix in Table 5-9, the study found that for the environment variable, the strongest determinants of most crimes are, rural or urban area, and the income level of the area. These seem to have the highest positive correlation across most of the crimes. This therefore shows that those two attributes are key for the variable environment and in turn key for the model.

Furthermore, for the author variables, the attributes of gender and identity be it an individual or an organization, have the strongest determinants for the outcome of most of the crime data. These attributes are key to the author variable. In addition, the correlations are not above .90 and therefore indicate no problem with multicollinearity. This is further confirmed by the determinant being above the rule of thumb of .000001.

Table 5-9: Correlation matrix

<table>
<thead>
<tr>
<th>Correlations</th>
<th>The place mentioned</th>
<th>Rural or urban area</th>
<th>High, Medium or Low Income</th>
<th>individual / organization</th>
<th>Gender of author</th>
<th>Number of followers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terrorism1</td>
<td>Pearson Correlation</td>
<td>-.013</td>
<td>-.113**</td>
<td>-.100**</td>
<td>-.025</td>
<td>-.075**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.425</td>
<td>.000</td>
<td>.000</td>
<td>.141</td>
<td>.000</td>
<td>.018</td>
</tr>
<tr>
<td>Category</td>
<td>Pearson Correlation</td>
<td>Sig. (2-tailed)</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>---------------------</td>
<td>-----------------</td>
<td>-----</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carjacking2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.24</td>
<td>0.049</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.056</td>
<td>0.034</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.044</td>
<td>-0.003</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muggings3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.004</td>
<td>-0.001</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.011</td>
<td>-0.004</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.008</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.859</td>
<td>0.859</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rape4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.036</td>
<td>0.563</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.686</td>
<td>0.094</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.193</td>
<td>-0.106</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Murder5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.034</td>
<td>-0.109</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.138</td>
<td>0.029</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.091</td>
<td>-0.010</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gun_crime6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.004</td>
<td>-0.103</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.101</td>
<td>-0.033</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.082</td>
<td>-0.011</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drugs7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.037</td>
<td>-0.124</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.154</td>
<td>0.053</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.074</td>
<td>-0.019</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pickpocket8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.028</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.011</td>
<td>-0.010</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theft9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.085</td>
<td>0.037</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.007</td>
<td>0.001</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.012</td>
<td>-0.010</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Robbery10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.013</td>
<td>-0.063</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.071</td>
<td>0.066</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.007</td>
<td>0.035</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vandalism11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.009</td>
<td>-0.007</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>0.002</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.019</td>
<td>-0.006</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kidnapping12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.008</td>
<td>-0.115</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.152</td>
<td>-0.018</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.084</td>
<td>-0.048</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.004</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General_Crime13</td>
<td>Pearson Correlation</td>
<td>Sig. (2-tailed)</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>---------------------</td>
<td>----------------</td>
<td>-----</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-.059</td>
<td>.000</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-.169</td>
<td>.000</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-.212</td>
<td>.000</td>
<td>3591</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-.063</td>
<td>.000</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.123</td>
<td>.000</td>
<td>3592</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-.066</td>
<td>.000</td>
<td>3591</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

The analysis of these variables shows that both the place mentioned and the number of followers are not significant determinants to the respective attributes, and as such can be removed from the proposed framework model. Figure 5-14 shows the new framework model without the weak correlation attributes.

**Figure 5-14: Resultant framework model**

### 5.5.3 Collaborative Analysis

The third test was a comparison with real world Police reports to collaborate the results of the study. The findings were able to show that the highest frequency reports of crime tweets was in Nairobi County with 159, Mombasa County with 41, Lamu with 11, Nakuru 10, Laikipia 8 and Kiambu 7. When compared with National Police Service (NPS) Crime report of 2016 in NPS (2016), the highest was Nairobi County, Mombasa, Kiambu and Nakuru. This goes to show that the Twitter data is close to the real life data on the ground.

Further collaborative data was found in the study when a comparison was made between County crime figures according to prevalence. The findings of this study showed that out of the top 27 Counties with crime reports in twitter, 18 of them were recorded with high crime rates according the National Police annual crime report 2016 (NPS, 2016).
5.6 Chapter summary

This Chapter looked at the results and finding of the research study. After a brief introduction of the results were presented in order of the research objects. First objective was identifying and extracting crime data. The process of how the data was collected is explained. The use of key words and geographic specific boundaries is outlined. Also explained is the creation of the Twitter profile and Twitter API used to connect to the platform and extract data. Then an insight into the choice of software to use as the primary extraction tool was given.

The second objective finding fell under sentiment analysis of the data. An introduction of the Mozdeh tool was shown and how SentiStrength works, its sentiment analysis algorithm works. The sentiment analysis results of all the words in the dataset are shown and then a summary of the top twenty words mentioned, excluding stop words. Further analysis of key words relating to crime are shown in a table form. The overall sentiment is analyzed both via figures and plotting of graphs.

The third objectives findings in regards to the framework are shown. The results structured under the different variables in the framework model. The first variable of environment results are displayed on frequency of crimes by County, towns then areas and roads. Corresponding maps highlight the locations using geographic information system software. Analysis of time periods and author results are also shown in this Chapter.

Finally is the testing objective where by the tests created in Chapter 4 are carried out. The first test is a usability test done with crime analysts. A questionnaire with three key questions was given and answers summarized and displayed in the Chapter. The second test was a statistical correlation coefficient done where to compare specific variables in the study. The third test was comparing real Police records to the study findings.
6.1 Introduction

This Chapter describes and interprets the significance of the study’s findings in relation to the research problem being examined. Taking the findings into consideration, this Chapter will explain any new insights on the problem. This Chapter will refer and connect to the objectives outlined in the introduction in Chapter 1.

This Chapter will start with giving a summary of important elements including the purpose of the study and specific objectives, research methodology and findings. This Chapter will then focus on the major findings organized according under the specific objectives. The results will also be interpreted and compared to relevant previous studies. This Chapter will also look at the major conclusions of the research from the research findings. Finally this Chapter will provide recommendations for future studies in this area.

6.2 Summary

This research study set out to come up with a framework to profile crime in Kenya using Twitter. Twitter is a social microblogging platform medium, where users post what they are doing or thinking about at a particular time. This study’s findings support the notion that Twitter can be used in providing data and specifically crime related data from members of the public.

By coming up with a framework to profile crime, the study aims at providing a process by which LEAs can use to improve their crime fighting activities. This framework has proved successful allowing LEAs to collect crime related data from Twitter and adding value to existing methods of data collection already employed by them.

This study had four specific objectives that it set out to achieve. The objectives were as follows:

i. Identifying and extracting crime data from social media.
ii. Performing Sentiment analysis on crime data
iii. Designing a Framework for profiling of crime, using the Twitter social media platform.
iv. Testing of the framework.

To carry out this study a number of tools and techniques were used. The first step was to find a way to identify and extract the data. This data comes from twitter users who are worldwide. Therefore there was a need to find a sample size, which was narrowed down to Kenya. Data extraction however showed that there are not many crime related tweets and therefore the
researcher chose to use all the tweets nationwide. The study therefore had to target the entire Kenyan Twitter population as its sample size and sift out the data not needed.

Once the data was identified and extracted, it had to be cleaned and sentiment analysis carried out. Sentiment analysis was necessary to ensure that the tweets collected were predominantly negative, to separate them from the positive ones. This study needed to look at only the negative tweets as those would have the necessary information to achieve the objectives.

The analysis of sentiments was done using a lexicon based algorithm within the SentiStrength free tool. The overall sentiment was analyzed as it was necessary to target specific types of tweets for further analysis. Once the acceptable dataset was identified, further analysis was done to analyze the relevant crime tweets.

Through the model created in the project, the preprocessed data could be analyzed in regards to the research variables. The analysis involved looking at the variables of environment and author and establishing how they can be used to monitor or profile crime. The findings from this analysis led to various results such as being able to transfer to maps via GIS. This then allowed viewing of patterns and hotspots of crimes from the data.

Other results from the analysis came in regards to the author variable. The results provided insight into the dynamics around different authors who tweet about crime and what differentiates one from another. These results where then investigated and compared to derive an understanding of the dynamics. Finally the framework was tested via two types of tests, usability test and correlation coefficient.

6.3 Discussions

6.3.1 Identifying and Extracting of Data from Social Media

According to Llewellyn, Grover, Alex, Oberlander, and Tobin (2013), extraction of data from Twitter is done via the streaming API that gives access of up to one percent of the produced data. They continue, the identification of Twitter data is done via random sample or query using user IDs, search words, location bounding boxed, hashtags or phrases. This study used search words and locational parameters as well as both streaming and historical tweets to increase the size of the dataset.
The data was successfully extracted and stored in a Microsoft Access Database, duplicates were removed and the data was exported as a Tab delimited format for further analysis. Tab delimited format was easier to work with due to the size compressed size and easy uploading for further analysis. Sentiment analysis was then carried out using the Mozdeh tool that uses lexicon based calculations.

The extraction of historical data was an added advantage as more data could be extracted as opposed to only using the limited streaming data. The data extracted was sufficient for some sort of results and findings but a larger dataset would have increased the details of the findings. The challenge was the tool only accessed approximately ten days data and therefore collection should have been carried out for six months continuously.

6.3.2 Performing Sentiment Analysis on Crime Data

Similarly to the SentiStrength algorithm, Smyth, Cheng and Russell (2016) came up with a lexicon-based method that combined crime data in a kernel density algorithm for the cities in China. They captured predictive hidden variables and calculated user ranking for the concept of user credibility to test crime and predict trends.

From the sentiment analysis it was observed that 78.42% of the crime related tweets had a negative sentiment. The results concur with Troussas, (2013) who used three different classifiers to classify Facebook crime messages. Results from their study found the Rocchio classifier had 73% negative sentiments and the Naïve Bayes classifier had 77% negative sentiment, both outperforming the Perceptron classifier.

Sentiment analysis on the dataset was successful as the overall result showed the vast majority of the tweets were of negative. This therefore enabled the study to take the entire dataset as a whole when carrying out further analysis. If the overall negative sentiment would have been much lower, then it would have been necessary to separate the data into two. This however was not the case in this study.

6.3.3 Designing a Frame Work for Profiling of Crime, Using the Twitter Social Media Platform.

A similar framework for detecting cyber bullying cases in schools was developed by Duwairi, Marji, Sha’ban and Rushaidat (2014) to analyze tweets in Arabizi and Arabic that was able to handle negation, emoticons and repetitions. The highest accuracy in detecting cyber bullying cases
they achieved was 76.8% when using Naïve Bayes classifier. These comparisons with other studies showed that the 78.42% was adequate enough to continue using the dataset as a whole for analysis.

Further findings in terms of type of crimes show that from tweets, the highest mentions are related to drugs, general word “crime”, murder, strikes, gun crime and carjacking. Again if compared to NPS crime report of 2016 in NPS (2016), vehicle and other thefts led, followed by theft by servant, dangerous drugs, stealing, criminal damage, economic crimes and homicides. This study’s result include most the same crimes toping the results.

The research then analyzed and gathered results on authors of tweets in line with the model variables. The results were able to show the top ranking authors in terms of followers and therefore the reach they extend. A comparison of the top 20 followed users from the dataset and the top 20 Twitter users in Kenya according to Twitter Counter Inc., 13 of the 20 are the same in TwitterCounterAPI (2018).

Analysis of the authors in findings showed the top 20 tweeters of crime data in Kenya. These top 20 authors were further categorized in terms of the amount of tweets they send on crime compared to tweets on any other topics. The authors who tweet mostly about crime were easily identifiable. This allowed the pinpointing of key crime tweeters who can be tracked for future data collection.

6.3.4 Testing of the Framework.

In testing the framework, the first test was a usability test. As mentioned in the literature various test exist such as usability tests can be used to evaluate the developed framework, (Satam et. al., 2016). This was a questionnaire made up of three questions relating to the framework given to six crime analysts. The questions and responses were recorded and summarized under three main section that include using the framework, the frameworks strong points and improvements to the framework.

The second test was conducted using a Pearson correlation coefficient carried out on the framework to find out the relationships between the variables of environment and author, and crime. Shareck and Ellaway, (2011), used Pearson correlation to show how the attributes are correlated when profiling crime in the neighborhood. This was done by looking at observed variables and attributes within the datasets. This study collected six attributes, three each from the
variables author and environment. These variables correlations were analyzed and the attributes that did not significantly affect the dependent variable were removed from the model.

The third test was a comparison with real world Police reports to collaborate the results of the study. The findings were able to show that the highest frequency reports of crime tweets was in Nairobi County with 159, Mombasa County with 41, Lamu with 11, Nakuru 10, Laikipia 8 and Kiambu 7. When compared with National Police Service (NPS) Crime report of 2016 in NPS (2016), the highest was Nairobi County, Mombasa, Kiambu and Nakuru. This goes to show that the Twitter data is close to the real life data on the ground.

6.4 Conclusions

6.4.1 Identifying and Extracting of Data from Social Media

From the research we note the significance of extracting data from twitter. This research used the tweeter API and MS Access to extract, store and filter data for purposes of analysis. The limitations of collecting with the API is how far the historical tweets can be collected from, approximately two weeks per search. This made it hard to achieve this objective as a lot of Tweeter data is required for analysis of this issue.

Despite this challenge, the tools were found to be able to capture and present the data in a format suitable for analysis. The first research objective, Identifying and Extracting of Data from Social Media, was therefore successfully achieved.

6.4.2 Performing Sentiment Analysis on Crime Data

This research used a lexicon based sentiment analysis tool to carry out its analysis. This is a dictionary based approach and ensured the use of recognizable and standardized English words in the analysis. The limitations included being limited to the dictionaries within the software and not having dictionaries of other languages spoken in Kenya.

Despite this limitation, majority of the tweets in Kenya are in English save for a few words from other local languages mixed in. This therefore was able to ensure the overall sentiment of the dataset be analyzed and reflect the true picture. The second objective, Performing Sentiment Analysis on Crime Data, was achieved.
6.4.3 Designing a Frame Work for Profiling of Crime, Using the Twitter Social Media Platform

For the third objective, designing a framework had to be approached from two angles. The study was to come up with a framework that covers both crime issues and twitter data. The framework had to include relevant issues to crime profiling as well as identifying variables that can be accessed from Twitter.

Limitations of building the model for this framework involved finding the right variables that can be extracted from Twitter. The attributes within Twitter must be usable and add to the significance of the framework. With the right tools and literature review, this study was able to successfully achieve the third objective, Designing a Frame Work for Profiling of Crime, Using the Twitter Social Media Platform.

6.4.4 Testing of the Framework

The final objective involved evaluation of the framework after all analysis and processes had been carried out. It is necessary to carry out tests on a framework to ensure that it works in a real world environment. For this study, the challenge was pinpointing the most effective tests to carry out.

The study therefore carried out two test, a usability test and a Pearson coefficient correlation. The usability test looked at responses from eventual users of the system. The correlation analysis test looked at observed variables and attributes within the datasets. The results from these two types of tests managed to achieve objective four, Testing of the Framework.

6.5 Recommendations

The biggest challenge for this research was the collection of data. Future studies in this area should look into a collection method for a streaming API instead of a standard API. Although historical data was adequate for this study, in future having an API that collect live data would be better. With a streaming API, data can be collected instantly and easily. This API together with a tailor made application that collects real time data will ensure better collection of data. The LEAs can then collect data continuously and store in their databases. The application connecting to the API can also be customized to fit particular specifications.

A standard API also has restrictions on rate of 1 percent of the tweets and requests to the system per user, and only go back 7 days. To be able to collect more tweets, it was necessary to get in
touch with Twitter Inc. via email, explain the research and requested for more access to historical tweets via the API. After a number of emails Twitter Inc. agreed to grant the API access to more historical data. To effectively collected tweets, the best option is having an enterprise API and collecting streaming data with little or no limitations.

Sentiment analysis results followed similar studies but further analysis requires the creation of local data dictionaries. Although the data dictionaries are standard and internationally accepted languages and emoji’s, there is need to have local language content with the dictionaries. This will better localize the sentiments of the tweets to include few Swahili, Sheng or tribal words.

Coming up with models or algorithms to easily obtain semantic classification or clustering of Kiswahili words from corpora. With a more comprehensive feature set and a thorough consideration of linguistic aspects unique to Kiswahili. This approach can be improved and refined to be a useful tool for automatic semantic feature extraction. The same approach can be used to obtain functional similarities of different word categories

The semantic classification obtained can be used to augment a lexicon or dictionary with semantic tags (codes). These semantic tags can also be used to improve the performance of various natural language processing tasks such as word sense disambiguation, for which semantic information is a prerequisite, part-of-speech tagging, syntactic parsing and information extraction.

Another challenge in the Twitter options was geographical location activation. For a user to have their location known requires permission from the user and geo-activation on the device. Since a majority of user do not activate this feature, in future it may be important to look into other ways of locating the users or devices such as tracking tweets and locations mentioned within those tweets.

Through the research findings, it was clear that the authors were key to content generation and data. The author variable therefore is the most important entity in this study. In future to effectively cover crime issues, tracking of authors and creating a database of those authors to log all the content created and come up with trends and patterns of those authors.

Evaluation of frameworks can be done with periodic comparisons of dataset. Datasets can be collected after certain periods and factor analysis carried out on the data. More correlation analysis on the relationships between variables should be done.
Although the data may seem late as something has already occurred, with the streaming API it can be collected as it happens notifying LEAs what is happening on the ground. This can then help the LEAs to deploy necessary resources after collaborating with information on the ground. This can therefore be a way to assist logistics and resource management.
References


Turney, P. D. (2002). humbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *40th annual meeting on association for computational linguistics* (pp. 417-424). Stroudsburg, PA, USA: Association for Computational Linguistics.


