



Strathmore
UNIVERSITY

Strathmore University
SU+ @ Strathmore
University Library

Electronic Theses and Dissertations

2018

Open source intelligence gathering for hate speech in Kenya

Banchale G. Adhi
Faculty of Information Technology (FIT)
Strathmore University

Follow this and additional works at: <https://su-plus.strathmore.edu/handle/11071/5980>

Recommended Citation

Adhi, B. G. (2018). *Open source intelligence gathering for hate speech in Kenya*

(Thesis). Strathmore University. Retrieved from <https://su-plus.strathmore.edu/handle/11071/5980>



STRATHMORE UNIVERSITY
Faculty of Information Technology
Master of Science in Information Systems Security

Open Source Intelligence Gathering for Hate Speech In Kenya

BY

Banchale A. Gufu

093926

SUPERVISOR: DR JOSEPH SEVILLA

APRIL 2018

DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

© No part of this thesis may be reproduced without the permission of the author and Strathmore University.

Name: Banchale A. Gufu

Signature:

Date :



This research proposal of Banchale A. Gufu was reviewed and approved by the following:

Dr. Joseph Sevilla,

Senior Lecturer, Faculty of Information Technology,

Strathmore University

Dr. Joseph Orero,

Dean, Faculty of Information Technology,

Strathmore University

Professor Ruth Kiraka,

Dean, School of Graduate Studies,

Strathmore University

ABSTRACT

The Internet has been celebrated for its ability to erode barriers between nations. Social media is a powerful medium that can unite, inform, and move people. One post can start a chain of events that changes the world. It gives users fast access to and sharing of information and facilitates ease of communication. However, the Internet allows for a lot of negativity as well. There has been an increase in hate speech activities on social media in the Kenyan cyber space.

The National Cohesion and Integration Commission (NCIC) was established to facilitate and promote equality of opportunity, good relations, harmony and peaceful co-existence between persons of the different ethnic and racial communities of Kenya, and to advise the Government on all aspects thereof (Act No, 12, 2008). In particular, the NCIC Act of 2008 is mandated to curb hate speech.

This research studied existing hate speech detection tools in use by NCIC, then identified gaps and challenges faced. A technical solution (tool for analysing hate speech) was proposed that can be implemented by the NCIC and the government to respond to hate-speech cases perpetrated through social media platforms. The developed tool tracked challenges and gaps in the existing tools currently in use by NCIC for hate speech monitoring, detection and analysis. Due to the differences in Application Programming Interface (API) implementation on the variety of social media platforms used in Kenya, the scope of this research is limited to Twitter.

This research employed the use of predictive analytics for text classification using Naïve Bayes. A tool that uses the predictive model in assistance to detection of hate-speech online was developed to conceptualize the solutions discussed in this research.

Keywords: social media, hate speech, National Cohesion and Integration Commission (NCIC), Communication Authority of Kenya (CA), Sentiments Analysis



TABLE OF CONTENTS

DECLARATION.....	ii
ABSTRACT.....	iii
LIST OF ABBREVIATIONS	viii
DEFINITION OF TERMS.....	ix
ACKNOWLEDGEMENT.....	xi
DEDICATION.....	xii
CHAPTER ONE: INTRODUCTION.....	13
1.1 Background	13
1.2 Social Media	15
1.3 Open Source Intelligence.....	16
1.4 Statement of the Problem.....	17
1.5 Research Objectives.....	18
1.6 Research Questions.....	18
1.7 Relevance of the Research.....	19
1.8 Scope and Limitations	19
CHAPTER TWO: LITERATURE REVIEW.....	20
2.1 Introduction.....	20
2.2 Sentiment Analysis Algorithms	20
2.2.1 Lexicon Algorithms	20
2.2.2 Machine Learning Algorithms (MLA)	21
2.3 Understanding Social Media in Kenyan Perspective.....	23
2.4 Challenges Regulating Social Media	25
2.5 Developing a Speech Analysis Tool	25
2.6 Existing Tools in Hate Speech Detection	27
2.6.1 Perspective API.....	27
2.6.2 Spice Hate Speech Detection.....	28
2.6.3 Hate Speech Blocker.....	28
2.6.4 Umati Online Monitoring Project in Kenya	28
2.6.5 Uchaguzi Online Monitoring Project in Kenya.....	29
2.6 Conclusions.....	30
CHAPTER THREE: RESEARCH METHODOLOGY	32
3.1 Challenge Identification	32
3.2 Research Design	32
3.3 Data Collection Methods	33
3.3.1 Observation	33
3.3.2 Questionnaires	33
3.4 Data Classification and Analysis	34
3.5 Implementation of the System	34
3.6 Overview of Agile Software Development Methodology.....	35
3.7 Agile Software Development Life Cycle (SDLC)	35
3.7.1 Concept.....	36
3.7.2 Inception	36

3.7.3	<i>Construction/ Iteration</i>	37
3.7.4	<i>Release/Transition</i>	38
3.7.5	<i>Production</i>	39
3.8	Validation	40
CHAPTER FOUR: SYSTEM DESIGN AND ARCHITECTURE		41
4.1	Introduction	41
4.2	System Design	41
4.2.1	<i>Functional Requirements</i>	41
4.2.2	<i>Non-Functional Requirements</i>	42
4.2.3	<i>Software Requirements</i>	42
4.3	System Architecture	44
4.4	Use Case Diagram	46
4.4.1	<i>Detailed Use Case Descriptions</i>	47
4.5	Sequence Diagram	48
4.6	System Analysis	49
CHAPTER FIVE: SYSTEM IMPLEMENTATION, TESTING AND RESULTS		54
5.1	Introduction	54
5.2	System Implementation	54
5.2.1	<i>Main Dialog Interface</i>	54
5.2.2	<i>Twitter API Configuration</i>	55
5.2.3	<i>Collection of Tweets</i>	56
5.2.4	<i>Cleaning of Tweets</i>	58
5.2.5	<i>Preprocessing of Tweets</i>	60
5.2.6	<i>Training Data</i>	61
5.3	Testing	64
5.4	Validation	66
5.4.1	<i>Manual Validation</i>	66
5.4.2	<i>Automated Validation</i>	67
CHAPTER SIX: DISCUSSION OF RESULTS		69
6.1	Introduction	69
6.2	Discussion	69
CHAPTER SEVEN: CONCLUSIONS AND RECOMMENDATIONS		71
7.1	Conclusions	71
7.2	Recommendations	72
7.3	Future Work	72
REFERENCES		74
APPENDIX A: Interview Guide		78
APPENDIX B: Python Program		82

Table of Figures

Figure 3.1: The Stages of the Agile Software Development Life Cycle	36
Figure 3.2: Understanding the Agile Software Development Lifecycle and Process Workflow .	40
Figure 4.1: The Database Schema	44
Figure 4.2: System Design Architecture.....	46
Figure 4.3: Use Case Diagram.....	47
Figure 4.4: Sequence Diagram.....	49
Figure 4.5: Data Flow Diagram	50
Figure 4.6: OSINT Gathering for Hate Speech	51
Figure 4.7: Configure Twitter API Keys	51
Figure 4.8: Tweets Collection for Sentimental Analysis	52
Figure 4.9: Cleaning Tweets	53
Figure 4.10: Pre-processor Algorithm Implementation in Python.....	53
Figure 5.1: Main Dialog Interface	55
Figure 5.2: Twitter API Configuration	56
Figure 5.3: Tweets Collection for Sentimental Analysis	57
Figure 5.4: Collected Tweets the Output of Figure 5.3	57
Figure 5.5: Raw Tweets Saved in Table Raw_Tweets	58
Figure 5.6: Cleaning Tweets	59
Figure 5.7: Cleaned Tweets Saved in Table Cleaned_Tweets.....	59
Figure 5.8: Tweets Classified as Neutral	60
Figure 5.9: New Tagged Tweets Stored in Analysed_Tweets Table.....	61
Figure 5.10: Training Dataset	63
Figure 5.11: Contents of Kikuyu-Mugiki Dataset	63
Figure 5.12: Testing Process.....	65
Figure 5.13: Python Script Named Manual_Validation.py	66
Figure 5.14: Predictions of Python Script in Figure 5.13	67

Table of Tables

Table 4.1: User Keywords	47
Table 5.1: Summary of the Automatic Testing Results.....	68

LIST OF ABBREVIATIONS

NCIC – National Cohesion and Integration Commission.

MAU – Monthly Active User.

DAU – Daily Active User.

YoY – year on Year.

CA – Communications Authority of Kenya.

ISP – Internet Service Providers.

SA – Sentimental Analysis.

MLA – Machine Learning Algorithm.

RF – Random Forest algorithm.

SVM – Support Vector Machine.

NB – Naïve Bayes

FAQs – Frequently Asked Questions.

SDLC – Software Development Life Cycle.

ERD – Entity Relationship Diagram.

CSV – Comma Separated Value.

API – Application Programming Interface.

DFD – Data Flow Diagram.

NLTK – Natural Language ToolKit.

GIF – Graphics Interchange Format.

URL – Uniform Resource Locator

DEFINITION OF TERMS

Naïve Bayes – is a Natural Language Processing (NLP) Application Programming Interface (API) that processes text-based data for classification as per the intent of the developer.

Twitter – An online social networking and micro blogging service that enables users to send and read short 140-character messages (Twitter, 2017).

Internet Service Provider – is a company that provides subscribers with internet access.

Social media – Websites and applications like Facebook, Twitter and Instagram that enable users create and share content or to participate in social networking.

Twitter REST API – is an Application Programming Interface (API) based on Representational State Transfer (REST) which allows other developers create programs that have access to data on Twitter.

Twitter Streaming API – is an Application Programming Interface (API) provided by Twitter which provides very high throughput in comparison to Twitter REST API of real-time access to subsets of both public and protected Twitter data. It contains public statuses of all users filtered using User Id, Keywords or by random sampling.

CSV format – is a comma separated values file used to exchange large files of data like in databases. Data is saved in a table structured format.

Sentimental Analysis – is the thorough research by computational study of how opinions and perspectives can be related to one's emotion and attitude show in natural language in respect to an event (Deng and Wiebe, 2015).

Daily Active Users – the number of users accessing a platform per day.

Emoji – A small image or icon use to express emotions electronically.

Python – Python is a general-purpose programming language that is simple, powerful high-level and very dynamic, with its syntax allowing programmers express concepts in fewer lines in comparison to other programing languages like C. its design focusing on code readability.



ACKNOWLEDGEMENT

I would like to acknowledge God for His grace, strength and good health as I undertook this research. My sincere gratitude to the members of the Faculty of Information Technology staff including: my supervisor, Joseph Sevilla for his continued commitment to guide and support this research from its inception to completion.



DEDICATION

I want to thank God for granting me with the ability to complete this project. To my dearest husband and children, thank you for giving me time to be able to work on this to completion.



CHAPTER ONE: INTRODUCTION

1.1 BACKGROUND

There has been tremendous increase of cyber bullying and online hate speech in Kenya over the past few years. Traditional hate speech has always existed. It has always been produced to support the status of one's own group and to discriminate against the others, but social media has now made it more visible than before. Expressions and beliefs based on emotions are emphasised and they are also circulated online, as it provides an inexpensive communication medium that allows anyone to quickly reach millions of users worldwide. The Internet allows people to protect themselves behind the screen and interact with each other in a more anonymous environment. According to Hitz, 37% of the world's population is active on social media and there are nearly 2.8 billion active social media users across the globe, and it is expected to rise to almost 3 billion users by 2020 (Hitz, 2017).

Hate Speech on Social Media fueled the 2007 post-election violence witnessed in the country. It was one of the known contributors for ethnic strife and tension. After the 2007-2008 post-election violence, the Government of Kenya enacted the National Cohesion and Integration Act (Act No, 12, 2008) to promote national cohesion and integration. The Act consequently instituted the National Cohesion and Integration Commission (NCIC) to oversee and monitor content in media such as radio, television and mobile phones in a bid to govern hate speech (National Council for Law Reporting, 2008). Hateful comments against an individual solely do not qualify as hate speech; this is because hateful comments can only be considered as hate speech if they target the individual as part of a group Sambuli, N., Morara, F., & Mahihu, C. (2013). The movement of hate speech mongers towards the digital cyber space needs to be addressed by the government before it escalates further as it is experienced regularly in Kenyan cyber space.

The National Cohesion and Integration Commission (NCIC) was established to facilitate and promote equality of opportunity, good relations, harmony and peaceful co-existence between persons of the different ethnic and racial communities of Kenya, and to advise the Government on all aspects thereof (Act No, 12, 2008). Under Sections 13 and 62 of the National Cohesion and Integration Act (NCI Act of 2008), the NCIC is mandated to curb hate speech, a role that strives towards national cohesion and integration.

Kenya endured perhaps the worst hate speech experienced ever in the just concluded 2017 general elections. People from different ethnic groups propagated hate speech against their opponents. In some cases, hate speech has resulted into violence. On the other hand, the NCIC has been facing challenges in detecting hate speech and ethnic or racial contempt that can result to violence. Ahead of the August 2017 election, NCIC and Communications Authority of Kenya cracked down on 176 social media accounts for propagating hate speech and 33 were under prosecution (“State cracks down on 176 social media accounts over hate speech”, 2017)

According to the State of Blogging & Social Media in Kenya report (*The State of Blogging & Social Media in Kenya 2015 Report*, 2015), there are 4.3 million Kenyan users on the Facebook platform. In the recent report published by Kenyan technology writer Kemibaro, M. (2015), Kenya has a confirmed 700,000+ monthly active users (MAUs) on Twitter, of 1.4 million to 2.1 million users in total, with 80% of the users accessing the service daily. In terms of Daily Active Users (DAUs), the number is approximately 570,000+. Twitter growth has been doubling in Kenya year on year (YoY). This means that around this time next year there will be around 1.4 million users. Twitter MAUs in Kenya are anywhere between 2.8 million to 4.2 million users in total, factoring in those who do not login.

As people spend more time online these days and most of our daily routines getting digitized, from shopping, entertainment to banking. They also express their opinion about their experiences and views online. It becomes a difficult duty to those tasked to monitor these communication channels and analyse the intentions and impact of views being expressed. Automated techniques to assist analysts tasked with the monitoring are used to collect and determine the semantic orientation (positive or negative) of the collected text data.

1.2 SOCIAL MEDIA

Although social media is often associated with large network services such as Facebook and Twitter, the term social media refers to a larger family of service platforms. These services can be clustered into six groups (Kaplan and Haenlein, 2010) which are: collaborative projects (e.g., Wikipedia), blogs, including microblogs (e.g., Twitter), content communities (e.g., YouTube), social networking sites (e.g., Facebook, LinkedIn), virtual game worlds (e.g., World of War Craft) and virtual social worlds (e.g., Second Life). Internet accessibility in Kenya has continued to penetrate and grow in numbers over the past 15 years and this has led to a growing significance of ethnic hatred and incitement in social media giants like Facebook and Twitter among others (Mutahi & Kimari, 2017).

In comparison to other countries in the East Africa bloc, Kenya's Internet usage and penetration is higher, and this has mainly been facilitated by the introduction of smartphones and cheap mobile data bundles provided by Internet Service Providers (ISP) and mobile networks. According to the Communications Authority (CA) of Kenya Sector Statistics report Q1 (2017), the number of broadband subscriptions was recorded at 17.6 million, a raise from 15.4 million posted in the preceding quarter representing a growth of 14.3 per cent. This translated to broadband

penetration level of 38.8 per cent as at the end of quarter under review up from last quarter's 34.2 per cent.

1.3 OPEN SOURCE INTELLIGENCE

Open-source intelligence (OSINT) is the use of public data to derive some actionable information in order to address a specific question or for use in decision making. Lawfully, any person could obtain the public data by request or observation, as well as other unclassified data that has limited public distribution or access. The latter is referred to as "grey literature" and includes non-proprietary information from companies and other organizations. These methods do not utilize any data which is covert or proprietary (U.S.A Joint Military Intelligence Training Center, October 1996 Open Source Intelligence Professional Handbook)

Internet users can get the desired information they need by just querying a search engine like Google. For OSINT, it is more of understanding where this data is coming from, efficient ways to collect and analyse it. Not all parts of the Internet are indexed by the search engines for instance social media websites. Public source is an umbrella term comprising of other related public data sources such as academia publications, media sources, website (web) content and open government documents. During the course of this research, OSINT refers only to the part which only uses internet (web) as its medium. This type of OSINT is also referred to as WEBINT (web intelligence) though it might be confusing and ambiguous since Internet and Website are not entirely the same thing. What is implied by WEBINT could have been correct few decades ago when the web was the primary technology on the Internet. Today the Internet is a coagulation of different technologies and protocols (Chauhan & Panda, 2015).

1.4 STATEMENT OF THE PROBLEM

Hate mongers have been known to move from the cyber space to the actual physical world to promote, fund and finance violent crimes. For example, this has been witnessed in the just concluded 2017 General Election in Kenya. Hate messages disseminated online are increasingly common, largely attributed to issues of anonymity, itinerancy, permanency and cross-jurisdiction of online content (United Nations Educational, Scientific and Cultural Organisation, 2015). While NCIC and other law enforcement firms both private and governmental are operational, hate speech propagation in social media is wide spread, with most going undetected. Currently the NCIC uses semi-automated processes with minimal automation to identify, monitor and analyse hate speech perpetrated via social media in Kenyan's cyber space. The semi-automated method is overwhelming, time consuming and is prone to human errors during interpretation.

Most of the hate speech suspects have been tried in our courts of law and found to be innocent. Some of the suspects have repeatedly committed the same offenses without prosecution, despite produced evidence. For a successful prosecution there is a need for using methods and tools that provide evidence that is admissible in a court of law following the principle rules of digital forensics like collection and preservation of evidence in its raw format.

The quality of the algorithms employed in these monitoring tools determines the quality of accuracy expected in real world cases. Unfortunately, most of the currently used tools are generic in nature. In Kenya, social media users mix a bit of vernacular languages with Swahili and English. If a tool is only trained to analyse English content without focusing on the local slang, it might miss to capture active hate mongers and run the risk of having a lot of false positives as well.

1.5 RESEARCH OBJECTIVES

The main purpose of the study was to design, develop and implement a tool that automatically monitors online hate speech by displaying feeds of propagated hate speech and as it happens which requires less human intervention for NCIC.

The specific objectives will be to:

1. To identify the challenges faced by National Cohesion and Integration Commission (NCIC) while detecting, monitoring and analysing hate speech in Kenyan cyber space.
2. To review existing tools that are used in hate speech detection and to identify gaps that exists in these tools.
3. To design, develop, test and implement a tool for detecting, monitoring hate speech in Kenya with digital forensics principles in design.
4. To validate that the automated tool can effectively identify hate speech in Kenyan context.

1.6 RESEARCH QUESTIONS

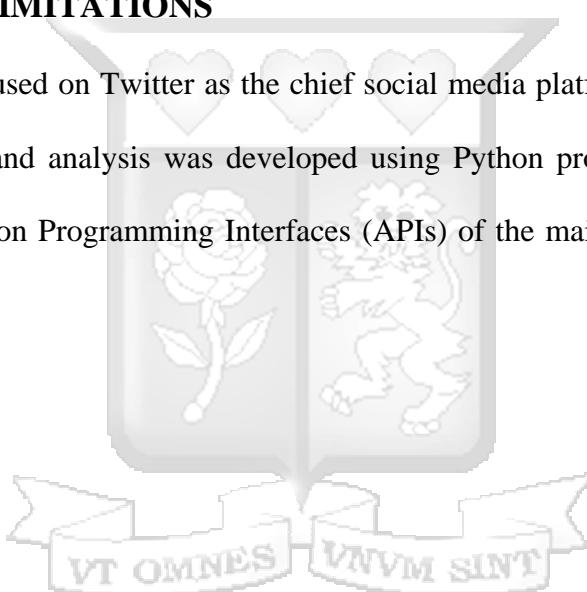
1. What are the challenges of the National Cohesion and Integration Commission (NCIC) in the identification and prosecution of hate speech?
2. What are the tools and systems that are available that address the gathering of information about hate speech?
3. How do you design, develop, test and implement a tool that can detect hate speech in Kenya and provide evidence for prosecution if needed?
4. Is the tool effective to automatically identify hate speech?

1.7 RELEVANCE OF THE RESEARCH

The NCIC has a mandate to curb hate speech, therefore there is a need to sanitise social media in Kenya. By developing a monitoring tool that will automatically detect, monitor and analyse hate speech in Kenyan cyber space, most of NCIC's challenges will be mitigated. An automated tool will reduce man interaction with the tool thus minimizing human errors and enhancing efficiency.

1.8 SCOPE AND LIMITATIONS

This research focused on Twitter as the chief social media platform study case. The tool for automatic detection and analysis was developed using Python programming language and integrated with Application Programming Interfaces (APIs) of the main social media platforms used.



CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

This chapter explores existing literature on hate speech and the techniques employed in automated text analysis. It describes existing tools and depicts gaps, especially in Kenyan environment. The chapter also reviews relevant literature to further comprehend the concept and investigate the research problem.

2.2 Sentiment Analysis Algorithms

Sentiment Analysis (SA) is the thorough research by computational study of how opinions and perspectives can be related to one's emotion, and attitude show in natural language in respect to an event (Deng and Wiebe, 2015). Recent events show that the sentiment analysis has reached up to great achievement which can surpass the positive versus negative and deal with whole arena of behavior and emotions for different communities and topics. With the use of different algorithms, good amount of research has been carried out for prediction of social opinions in the sentimental analysis field. Sentimental detection and analysis uses two main types of algorithms: Lexicon and Machine learning based algorithms.

2.2.1 Lexicon Algorithms

Lexical analysis utilises lexicon of pre-tagged words, a dictionary of words precompiled before their use, for a specific purpose. Using a hand-tagged adjective lexicon, Hatzivassiloglou, V., & Wiebe, J. M. (2000) reported success rate of 80% on single phrase of text, demonstrating that the subjectivity of an evaluative sentence could be determined. Comparison to its later advancements is as stated below.

- a) Using hand-tagged adjectives lexicon methodology to test a database of movie reviews by Kennedy and Inkpen (2002) with a result rating at 62%.
- b) Turney (2002) used Internet search engine to check the polarity of words, based on the AltaVista search engine queries: target word + 'good' and target word + 'bad', to work on the earlier work of Kennedy and Inkpen (2002), thus increasing the success rate to 65%. This comparison was done by Thakkar H. and Patel D. (2015).

With more developments and researches done, the accuracy rate remained in the range of 62% to 64%, till a research by Turner, P.D. (2003) on Measuring Praise and Criticism, in which effects of adjective orientation and gradeability on sentence subjectivity by Hatzivassiloglou, V., Wiebe, J. (2000) was employed. It yielded an accuracy of 82%, by evaluating the semantic gap between the words and simply subtracting the positive from the negative.

2.2.2 Machine Learning Algorithms (MLA)

Based on Dezyre (2016), machine learning algorithms are classified into supervised, unsupervised or reinforcement algorithms. Supervised MLA search for patterns within the value assigned to each data point by making predictions on given set of samples. They perform classification on the target corpus, after being trained with training data. Unsupervised MLA algorithms have no assigned value to the data points but organise the data into a group of clusters describing its structure thus make it organised and simple for analysis. Reinforcement MLA algorithms, based on each data point, choose an action then evaluate the result. Overtime, this algorithm changes its strategy in order to learn better and achieve results that are more accurate.

In accordance to Bai, Nie, & Paradis, (2004) and Pei & Wu, (2014), several machine learning algorithms are used for sentimental detection and analysis namely: N-gram, Random Forest, Naïve Bayes and Support Vector Machine (SVM). N-gram as SVMs have a solid theoretical foundation in statistical analysis making them suitable for language tooling when used for text classification, topic detection and information retrieval. An n-gram language tool can therefore be applied to text classification in a similar manner to a Naïve Bayes model. (Peng, 2003).

Random Forest (RF) Algorithm trains multiple decision trees, with each tree trained using a random subset of the vector features. To minimise cost function that evaluates the performance of the trees, the values of the intermediate nodes are kept up-to-date. The decisions of each tree are combined using a voting algorithm that gives the result. The sequence of features and the value of the feature generate the path to a leaf that represents the decision. Liombart, R. Ò., & Duran, C. J. (2017).

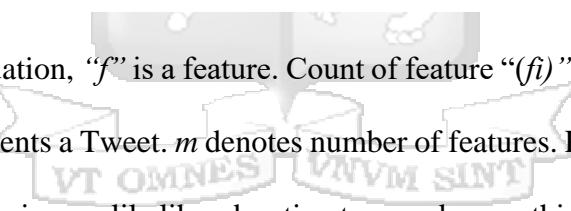
Naïve Bayes and SVM are the most commonly used algorithms. Naïve Bayes is based on the Bayes rule in probability theory and statistics, which states that the probability of an event, based on prior knowledge of conditions that might be related to the event. A research done on documents retrieved from online sites reports an overall Naïve Bayes correctness of 75.6%, in cross validation experiments, on a dataset that consists of 100 documents for each of 12 categories (Yahyaoui, M. 2001).

The SVM being one of the widely used supervised machine learning algorithms for textual polarity detection, considers that each set of features represents a position inside a hyperspace and the SVM tries to divide it using hyperplane maximising the distance between this hyperplane and each vector, minimising the objective function. This space division is hard to accomplish, and

sometimes impossible, for the SVM can use a margin that allows to misclassify some examples but increases the overall performance Liombart, R. O., & Duran, C. J. (2017). The accuracy rate of a SVM is approximately 82%. Despite having a solid theoretical foundation, SVMs are more accurate than most algorithms in performing classifications (Joachims, T. 1998). It supports comparison of given data to a list of words then its classification of that data to the rightful category or class (Vishal & Sonawane, 2016). Let d be the Tweet and c^* be a class that is assigned to d . The equation is;

$$C^* = \arg \max_c P_{NB}(c | d)$$

$$P_{NB}(c | d) = \frac{(P(c)) \sum_{i=1}^m p(f_i | c)^{n_i(d)}}{P(d)}$$

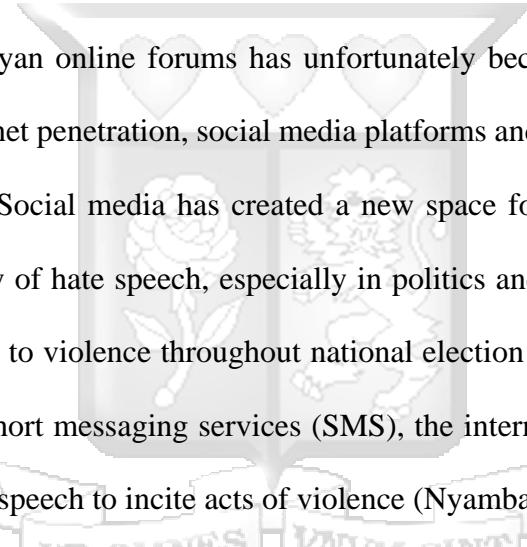


From the above equation, “ f ” is a feature. Count of feature “(f_i)” is denoted with $n_i(d)$ and is present in d which represents a Tweet. m denotes number of features. Parameters $P(c)$ and $P(f/c)$ are computed through maximum likelihood estimates, and smoothing is utilised for unseen features. Python NLTK library is used to both train and classify Tweets using Naïve Bayes Machine Learning technique (Satapathy, Govardhan, Raju, & Mandal, 2014).

2.3 Understanding Social Media in Kenyan Perspective

Social media in Kenya is viewed as a platform where people can connect with friends, share views and thoughts, criticise others and as well follow the affluent in the society. It is efficient for all classes of people either social or antisocial. All they have to do is be online either via hand held

gadgets, personal computers or in cyber cafes to either post, view, comment or share posts. Social medial though mostly used by the youth, it is an open field exploited by all age groups. Fair prices for internet bundles by Internet Service Providers, most homesteads in urban centers having WIFI access and cheap cyber café services have heightened social media access. Some parents from financially stable families have bought their children smartphones, giving them access to social media and internet in general. (“How app assists parents manage child's phone”, 2016). The strength of social media is that its information is largely public unless its owner has set the privacy setting to private.



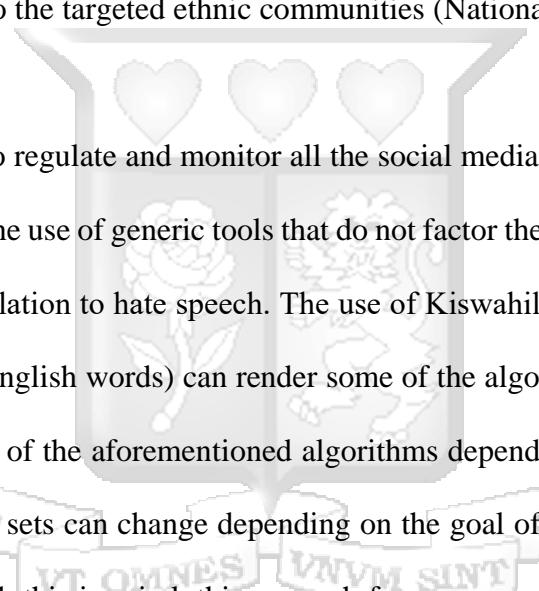
Hate speech in Kenyan online forums has unfortunately become a common occurrence with the growth of the internet penetration, social media platforms and cheaper mobile computing devices in the recent past. Social media has created a new space for the dissemination of hate speech. Kenya has a history of hate speech, especially in politics and this was done through the use of incitements and calls to violence throughout national election campaign period and as the conflict unfolded. Media, short messaging services (SMS), the internet and mobile phones were used as transmitters of hate speech to incite acts of violence (Nyambane, 2012).

According to the Kenyan constitution (The Constitution of Kenya, 2010), article 33 provides that every person has the right to freedom of expression, which includes freedom to seek, receive or impart information or ideas, freedom of artistic creativity and academic freedom of expression. The constitution however goes ahead to dictate that this freedom of expression does not extend to propaganda for war, incitement to violence, hate speech or advocacy for hatred that constitutes ethnic incitement, vilification of others or incitement to cause harm.

2.4 Challenges Regulating Social Media

Kenya has over forty-two ethnic tribes, each with its own unique way of communicating.

Almost all ethnic communities in Kenya have some kind of stereotypes about others, either positive or negative (National Cohesion and Integration Commission, 2013). Most negative statements depict feelings of contempt and general hate towards targeted communities resulting in heightened friction and animosity among various ethnic communities. The negative statements are often expressed in coded language well known to the members of the community who use it and may or may not be known to the targeted ethnic communities (National Cohesion and Integration Commission, 2013).



It is therefore hard to regulate and monitor all the social media platforms communication. This problem is fueled by the use of generic tools that do not factor the contextual problems faced in Kenyan cyberspace, in relation to hate speech. The use of Kiswahili and ‘Sheng’ a local slang (mixture of Kiswahili and English words) can render some of the algorithms discussed in section 2.2, ineffective. The quality of the aforementioned algorithms depends on the quality of training data used. The training data sets can change depending on the goal of an analysis and the nature of data being collected. With this in mind, this research focuses on using Twitter as the selected social media platform for the study.

2.5 Developing a Speech Analysis Tool

Speech analysis is also referred to as opinion mining or sentiment analysis which is a field of natural language processing (NLP) that tries to identify and extract this kind of subjective information from text data (Liu, 2012). There are mainly two approaches of determining the semantic orientation or what is also known as polarity in NLP, rule based and machine learning.

Rule based approach counts the number of negative and positive words in a text data. Prior to doing this, a lexicon, which is a resource where all the words have been classified as either positive or negative has been processed from a viably available pool of related data (Boržì, Faro, Pavone & Sansone, 2015). The use of lexicons can help determine what emotions does a word express, intensity of a word and general inquiry. For instance, “I hate Luos” does not express the same intensity as “I really hate Luos so much” despite the subject remaining the same.

Machine leaning approach involves the use of Naïve Bayes classification, Support Vector Machines and maximum entropy classifier. The key important part of this approach is the training data used and the features the researcher chooses. Training data is usually a human-annotated corpora and a feature can be a bag of words model. A bag of words model is collection of individual words in the training and testing text data sets. The corpus is used to train a classifier which operates by calculating a conditional probability of a text instance to belong to a specific class (e.g. negative or positive) given a set of attributes or features. The classifier is trained by generating all pairs of words (bi-grams) in all training data as a feature. A feature vector for a text can later be generated during analysis indicating the presence or absence of these bi-grams (Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. 2018).

Whichever approach is chosen for analysis, understanding the data format to be subjected to opinion mining is very important. Since each social media platform presents its data in a unique way, it is challenging to create a single tool that can analyse all existing social media platforms. However, regardless of which platform one chooses for analysis purpose, all the data collected for training, testing and analysis or lexicon must be cleaned into one format (Peersman, G., 2014). This is usually achieved by using regular expressions, or regex, which is a text pattern used to manipulate text data.

Publicly available lexicon data for tool development that a researcher can start with a General Inquirer Lexicon developed and maintained by Harvard University and the MPQA Lexicon developed and maintained by the university of Pittsburg. The Linguistic Inquiry, Word Count, Lexicon and Sent WordNet are other publicly available lexicon data for tool development (Potts, C., 2011). For Twitter sentiment analysis and opinion mining, a general training corpus developed and maintained by Niek Sander is usually advised for a start before making any other fine tuning. This training data is publicly available and contains at least 5000 classified Tweets at the time of this research. However, Twitter does not permit the distribution of raw Twitter data hence this training data only contains a Tweet ID and a label (positive, negative or neutral). Researchers have to get Niek training data set using Twitter API (Hasan, A., Moin, S., Karim, A., & Shamshirband, S., 2018).

2.6 Existing Tools in Hate Speech Detection

2.6.1 Perspective API

Perspective is an API developed by Jigsaw using Machine learning algorithms to estimate the perceived impact a comment might have. While using the toxic detector running on Google's servers, developers can identify bullying, harassment and abuse on social media, or more efficiently to filter invective from the comments on a news website. It weighs the toxicity of a comment. Currently, online communities and publishers, such as The New York Times, The Guardian, The Economist, and Wikipedia have partnered with Jigsaw to implement Perspective API to help them in their online conversations (Lichterman, J., 2017).

2.6.2 Spice Hate Speech Detection

There also exists Spice Hate Speech Detection tool. According to Kinnunen, Spice Hate Speech Detection tool was built for hate speech detection in order to assist officials track and detect hate speech in social media (Kinnunen, 2017). The tool monitors public communications between the candidates in social media then flag posts with hate speech as per the European Commission and Ethical Journalism Network. The tool uses several extraction methods such as: Bag-of-features, Word embedding and Machine learning. Support Vector Machine (SVM) and Random Forest (RF) Machine learning algorithms.

2.6.3 Hate Speech Blocker

Hate Speech Blocker seeks to resolve online hate speech without blurring the line between freedom of speech and censorship. It is a chrome plugin which once installed, will always notify the user of hateful words while he/she types. Usually, online users tend to vent their anger on unsuspecting victims even anonymously on social media thus making it quite hard to track. On the contrary, Hate Speech Blocker suggests to the writer that the words he/she is using could be construed as hateful. This can only happen if one has this Chrome plugin. It analyses text when one is typing, and flags out hateful words. It constantly checks the words against Hatebase, a nonprofit online service that collects data about hate and derogatory terms in various parts of the world and has the ability to reference the terms across different countries (Akl, A., 2016).

2.6.4 Umati Online Monitoring Project in Kenya

Umati Online Monitoring project's aim is to provide continuous monitoring of online media. Such projects are rare, but they have the potential to serve as early warning systems or

enable a reaction to incidents as they occur. The best-known project of this nature is the Umati project in Kenya (Ihub, 2013).

Umati Kenya is an online monitoring project that collects and analyses hate speech and or dangerous statements from Kenyan online. This gives the response team early warning thus taking measures as or before the incidents occur.

Real-Time Monitoring and Mapping by Umati was launched in October 2012, six months before the Kenya general elections (held on March 4, 2013).

Several challenges emerged as a result of the semi-automated with minimal automation data collection process. These included the possibility for correct misses/false alarms, as detailed by the signal detection theory. The Umati tool often displayed fatigue and varying levels of productivity as a result of task dullness.

2.6.5 Uchaguzi Online Monitoring Project in Kenya

Ushahidi developed Uchaguzi-Kenya to enable citizens report problems occurring during Kenya's 2010 constitutional referendum and 2013 general election. Uchaguzi's main goal was to act as an early warning system and prevent the escalation of incidents. Other deployments have also taken place in Tanzania, Uganda, and Zambia in 2010 and 2011 Omenya, R. (2013). Uchaguzi monitored threats such as dangerous speech, rumors, and mobilization toward violence, alongside other issues related to security, polling station management, and vote counting and reporting Chan, J. (2012).

Required intelligence was available in open source with an estimate of 90% thus making its' imperative intelligence analysts become adept at mining open sources (Skyrme, 2007).

Recorded Future can help reduce research time, identify new sources, build timelines, chart networks, perform link analysis, and more with our open source threat intelligence.

Existing tools currently retrieve evidence by linking an individual or company with different social media accounts, locations, friends, age of friendships and IP details. These tools have the capabilities to analyse and monitor individual movements and behavior, but they do not have capabilities to perform sentiments analysis on hate speech (Gross, 2013).

2.6 Conclusions

Although the enhanced digital revolution in communication platforms and social interaction media has brought great entrepreneurial opportunities, hate speech has extensively taken root. This escalation of toxic behavior is harmful to people limitlessly leaving no room for safe havens. Hate speech detection and monitoring tools are rare and scarce, with those available neither adequately effective nor efficient for our local context use.

Based on the local hate speech detection tools, their use, result analysis, detection and monitoring is extensive, mostly in distinguishing between hate speech, profanity, and other texts. It was identified that these tools are not specific for automated hate speech detection as Umati Online Monitoring tool. Though Umati is semi-automated with minimal automated data collection process, this research focused on addressing these gaps by developing a fully automated tool both in detection and in data collection using naïve bayes algorithm to detect, analyse and monitor hate speech. The tools' development focuses on Twitter as the social media case study platform.

The rest of this research paper is organised as follows. Chapter 3 is the detailed research methodology encompassing challenge identification, research design, data collection methods, classification and analysis and system implementation. It also includes the Software Development Life Cycle of Agile Software. Chapter 4 clearly presents the system design and architecture of the

automated hate speech detection, monitoring and analyses tool. Its implementation is discussed in detail in chapter 5. Chapter 6 is result discussion fully addressing the study objectives and results of the study. The research culminates in chapter 7 with a conclusion, recommendation and future work.



CHAPTER THREE: RESEARCH METHODOLOGY

This chapter explains the methodology that the researcher adopted in development of the open source analysis tool for hate speech in Kenya. It expounds on the development tools and environment that will be implemented.

3.1 Challenge Identification

In order to identify the challenges of National Cohesion and Integration Commission (NCIC) in the identification and processing of hate speech, the researcher conducted interviews with NCIC's staff to find out the challenges faced in monitoring hate speech on social media. Real time data was used to monitor, observe and analyse the timeliness and efficiency of the current methods of hate speech detection tools used by NCIC.

To identify existing literature on tools, algorithms and challenges related to hate speech in Kenya, desktop research, questionnaires, and one on one interviews were conducted with five NCIC social media monitoring analysts, who monitor online social media, for information gathering. In addition, Frequently Asked Questions (FAQs) and focus groups were used.

3.2 Research Design

A research design describes how a research study is carried out giving clear blueprint of the entire phase to the very completion of the study. It includes the operationalising variables, selection of the point of interest in the study, data collection for hypotheses testing and result analysis (Thayer, 1993). In this research, research objectives were identified, with the design dealing with data collection, cleaning and categorising of the training data. A Naïve Bayes- based model (Wang, 2012) would be built for data classification as positive, negative and neutral.

Efficacy of the tool focuses on hate speech automated detection, analysing and monitoring facilitating the operation of NCIC in hate speech mitigation in Kenya.

3.3 Data Collection Methods

Several techniques were employed for data collection. Techniques employed are determined by the research type (Kothari, 2004). Data collection helps the researcher in drawing more directional conclusions and assist in decision making. In this research, data collection methods helped determine the machine learning techniques to employ. The Twitter search API (Bifet, 2010) was used to collect sample training corpus to train the Naïve Bayes model used for data classification. The target of the survey carried out were hate speech monitoring analysts of NCIC. In this study, a mixed method research (combination of qualitative and quantitative) was practiced, i.e. interviews, questionnaires and language biasedness based on cultural, habits and customs.

3.3.1 Observation

Observational technique was employed in determining the social and organisation requirements to use. Once collected data was classified it was analysed to see if the classifier correctly classified data into the three classes, namely negative, positive and neutral class.

3.3.2 Questionnaires

To incorporate different opinions from the interviewees, questionnaires render a more cognitive environment for acquiring good information. The questions were structured and required the analyst to give their answer in writing. It employed a mixture of open ended and closed ended

questions. Five analysts of NCIC filled the questionnaires. Questions like ‘Do you use automated tools to automatically detect, analyses and monitor hate speech on social media and have you identified any gaps in the tools you use for hate speech detection, analysis and monitoring?’ were asked. This is as shown in Appendix A.

3.4 Data Classification and Analysis

The features extracted in the training corpus along with a custom corpus that captures commonly used ‘sheng’ words were used to train the classifier. Steps taken to building the classifier model involved building a vocabulary (wordlist in training data sets or corpus). Representation of each used Tweet in the training data with presence or absence of the test Tweets for accuracy testing. All of these training steps used NTLK library. The raw Twitter data collected was classified and analysed using this custom trained Naïve Bayes model (Wang, 2012). The classified data along with the result were saved in a database.

3.5 Implementation of the System

Bag of words technique was employed which uses wordlist (Sriram, 2010). Python is a general-purpose programming language that is simple, powerful high-level and very dynamic. To-date it has been widely used for functional and data processing in natural language using the Natural Language ToolKit (NLTK) library Manning, C., (2014). NLTK is a Python-based library which serves as a great companion when building programs and data classifiers. Sample data for tool training, testing of various classifiers and verification of the results was collected from Neik Sander sorted 5000 Tweets.

To address Kenya’s local problems associated with hate speech propagated on social media using ‘Sheng’, the researcher manually sorted a sample of training data sets with these local

dialects. Also, for provision of digital evidence in case need, a forensically sound copy of the collected Tweets is stored with a unique ID before and after analysis for the purpose of a successful prosecution (Taylor, Haggerty, Gresty, Almond & Berry 2014).

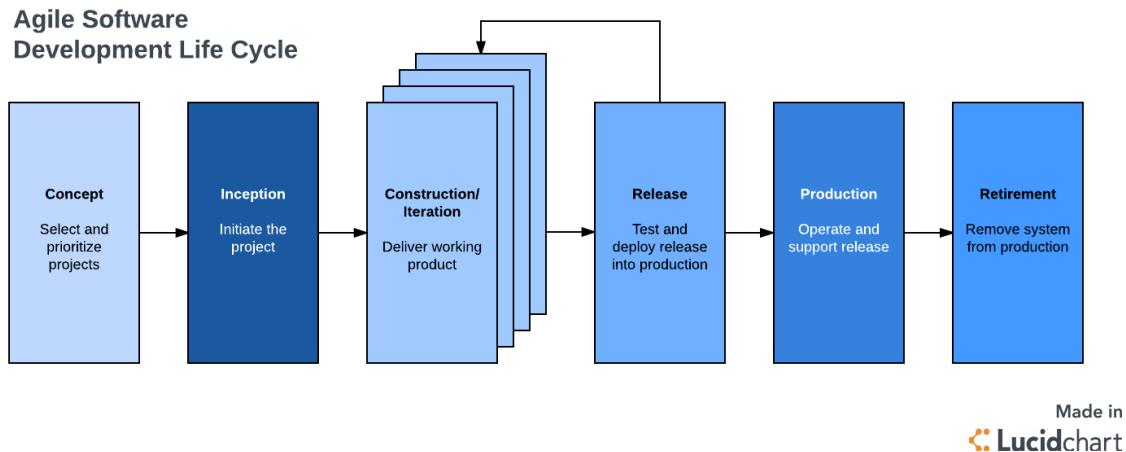
3.6 Overview of Agile Software Development Methodology

This research adopted Agile Software Development methodology. Studies show that Agile Software Development methodology is suited for the scenarios where the primary focus is developing an application without comprehensively knowing all specification and requirements (Sommerville, 2009). This fitted very well because NCIC did not have all the requirements at the time. Therefore, it was expected that adjustments would be made to specifications in response to knowledge gained as the development progressed and on the working prototype for hate speech analysis. Additionally, design and development of the tool focused on pegging the tool rather than comprehensive specifications and documentations due to time constraints. The Figure 3.1 below explains the steps taken in agile development of the application.

3.7 Agile Software Development Life Cycle (SDLC)

The Agile software development life cycle has six (6) stages, namely: Concept, inception, construction/iteration, release, production and retirement as shown in Figure 3.1 below:

Figure 3.1: The Stages of the Agile Software Development Life Cycle



Source: The Stages of the Agile Software Development Life Cycle.

Retrieved from <https://www.lucidchart.com/blog/agile-software-development-life-cycle>

3.7.1 Concept

It is the first stage of Agile SDLC. With close observation, it was identified that there were problems in the hate speech detection process, from detection to response of an incidence. The main problem identified was that manual detection was employed. With this in mind, consultations were made and it was agreed upon that there was a great need for a tool to automate the detection process. This research focused on the development of automated hate speech detection tool, which would facilitate hate speech detection process in Kenya in a more suitable manner guaranteeing timely response.

3.7.2 Inception

The researcher worked closely with NCIC hate speech monitoring analysts during the tool's development to determine the requirements that were needed during the development of this

tool. This included both functional and technical specifications. Using the functional specifications, technical specifications was developed and it detailed how the functional specifications would be implemented. The researcher carried out feasibility study to determine the possibility of developing the tool to completion in terms of technical, economic, legal and NCIC's scenarios at hand. Different tools were studied to understand the gap further and propose an adaptable solution for NCIC. With the plan sprint occurring in this stage, face to face interviews with the five NCIC social media monitoring analysts were conducted to establish if the tool in use by NCIC could automatically detect hate speech from social media. Initial meetings between the researcher and the NCIC social media monitoring analysts were held towards facilitating the tools development since they were the targeted audience.

3.7.3 Construction/ Iteration

At this stage, appropriate analyses were carried out to determine the relationship between the specifications identified by NCIC hate speech monitoring analysts and the actual tool to be developed. Planning standards and procedures were provided as per NCIC policy on software development tool management.

The development methodology focused on object-oriented development and was intended to cover the following, while adhering to the approved guidelines:

1. Web based platform; Django framework (Morales, D. R., 2018) was proposed but the researcher used existing tools and APIs.
2. Languages and framework – Python 3.0, JQuery languages and related plugins; system prototyping and module testing.

The immediate stakeholders were involved in this development stage in order to collaborate during the tool storming, Test Driven Design (TDD), confirming the first tests and documentation began. The first deployment was done in this stage.

3.7.4 Release/Transition

Also known as, the “End Game” the developed tool was released in this stage. There are several important aspects towards release, explained below.

- ❖ Testing and Quality Control

The objective of software testing was to see whether the tool worked well when integrated with external components like computer systems and other software within NCIC environment, as specified in the software requirements. The software would not be used on the developer’s computer system. With this in mind, testing was done on a computer system with specifications similar to those on which the software would be run. The immediate users and some representatives tested the final software to see if it was complete and it actually performed the functions it was supposed to perform.

The goal at this level was to evaluate whether the system had complied with all outlined requirements and see if it met quality standards. To ensure authentic system testing, independent testers were selected who had not played any role in development of the tool. Testing was performed in an environment that closely mirrored production. System testing verified that the application met the technical, functional, and business requirements that were set by the NCIC and the dissatisfactions presented by the previous tool as outlined in the annexed questionnaire.

- ❖ Prototyping and review

The developed and tested prototype was then reviewed by the five NCIC social media monitoring analysts in order to identify any gaps and areas of improvement. At this stage the specifications neither changed, nor new ones introduced. To make the tool very user friendly, a user interface was created.

❖ Rework

Failure to deal with defects and any known or suspected issues with the tool would lead to discrepancies, unauthentic results and possibly heightened tool failure rate. Therefore, any defects detected while carrying out the testing and quality control procedures were dealt with, to and in accordance with the intended design of the tool. No changes were identified during review.

❖ Finalise system and user documentation

The documentation written during the Construction/ Iteration stage was finalised in this stage to incorporate the system release. A documentation of the tool for both the general user and system administrator was generated in this phase to help them with firsthand information while using the tool.

❖ Training

The users were enrolled for training. Based on the purpose of this tool, only 5 users were trained in one sitting.

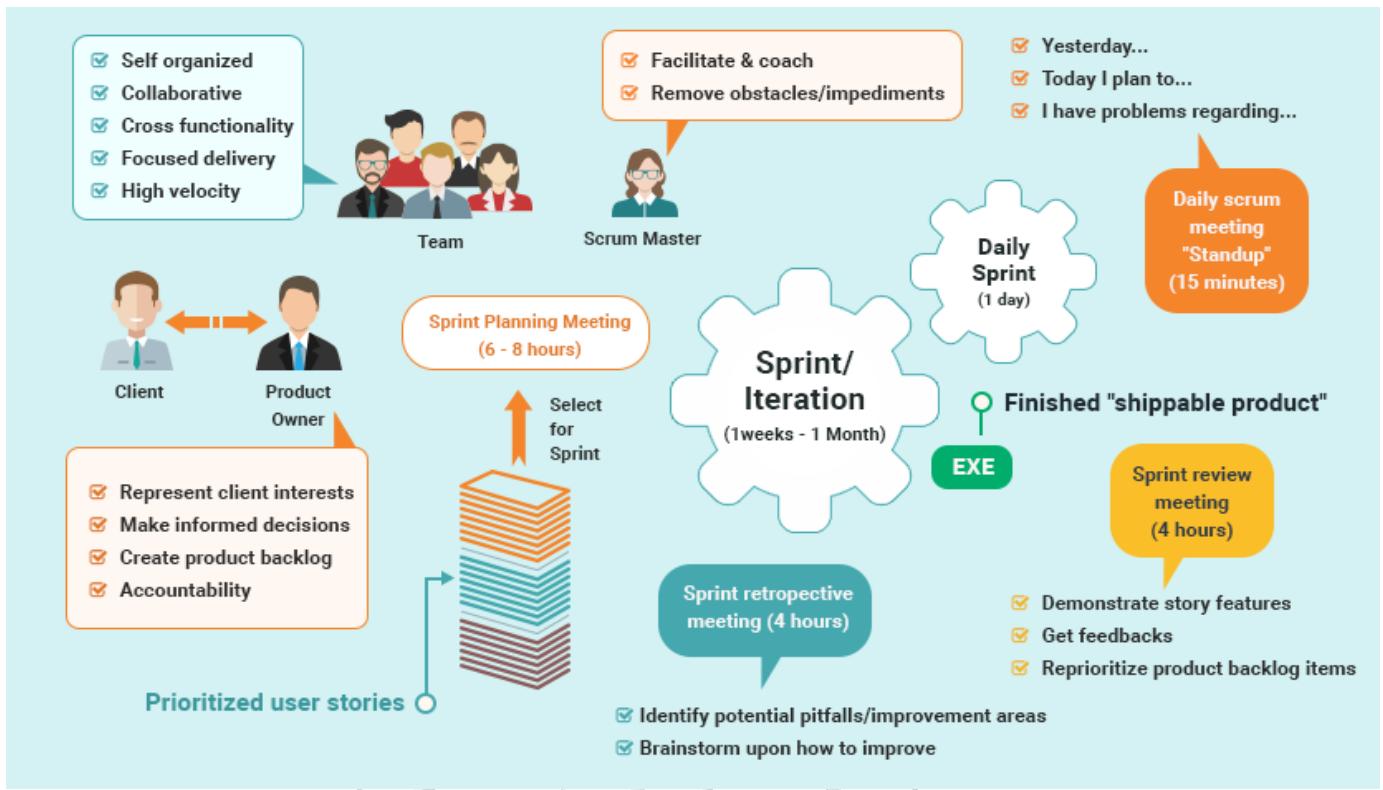
❖ Deploy the system - the iteration was then released into production.

3.7.5 Production

The developed and tested prototype was reviewed by potential users within NCIC who were randomly selected in order to identify gaps in functionality and areas of improvement. At

this stage the specifications were not changed, nor new ones introduced. No changes were identified during review.

Figure 3.2: Understanding the Agile Software Development Lifecycle and Process Workflow



Source: Understanding the Agile Software Development Lifecycle and Process Workflow.

Retrieved from <https://www.smartsheet.com/understanding-agile-software-development-lifecycle-and-process-workflow>

3.8 Validation

In order to ensure that the right feature types, feature weights, Sentimental detection, and analysis algorithm are chosen for training the hate speech detection tool, validation was required. Validation of the final application was accomplished by comparing the rate of false positive and false negative in the sample test data used.

CHAPTER FOUR: SYSTEM DESIGN AND ARCHITECTURE

4.1 Introduction

The discussed methods of natural language processing (NLP) using NTLK were applied in development of the tool to perform analysis of hate-speech on Twitter. This section provides a comprehensive logical and process architectural overview of the developed tool and the components that make up the core functions of the tool.

4.2 System Design

4.2.1 Functional Requirements

The functional requirements include:

1. Twitter REST API and Streaming API.
 - This is used in the initial stage of collecting training data for building a corpus
 - The Twitter API keys are used for authentication in the rest of the application
2. Python Programming Language
 - Provides regular expression library for cleaning the collected data before training and while analyzing.
 - This is used to implement the naïve bayes classifier using the Python NTLK library
 - Provides data analysis libraries for data visualization.
3. Debian Linux Distribution
 - Provides networking stack required for connection management
 - Ships with Whiptail library used for graphical user interface development
4. SQLite Database
 - Provide a storage media for the collected and analysed Tweets

5. Computing Hardware

- An Intel core-i5 with minimum speeds of 2.8 GHz and at least 4 GB of RAM
- Storage device of at least 500 GB and for better speeds use solid-state-drives (SSD).

4.2.2 Non-Functional Requirements

The properties and constraints of the system are specified by its non-functional requirements.

- i. The tool should provide real-time hate speech detection, analysis and monitoring on Kenyan social media space.
- ii. The tool should be time efficient, precise and realistic in its detection, monitoring and analysing of Tweets.
- iii. The tool should meet required standards of NCIC and as it is required by law of the same.
- iv. The processing and general analysis of the Tweets should follow some basic digital forensics principles for admissibility of evidence if prosecution is needed.

4.2.3 Software Requirements

i. Usability

The tool's intended and immediate users are NCIC social media monitoring analysts. This tool's ease of use should be guaranteed. The staff should undergo a training procedure scheduled for half a day within which they should be well conversant with the tool with a failure rate minimised to two errors per one hour of use. The users should be in a position to clearly understand and outline the analysed report issued by the tool.

ii. Scalability

The tool should be able to give the required throughput within the outlined time, and even well able to deal with data influx in case of emergencies without being overwhelmed. If there is an increase in the number of online Tweets to decipher, it should be able to deliver within the required timeline.

iii. Persistent Storage

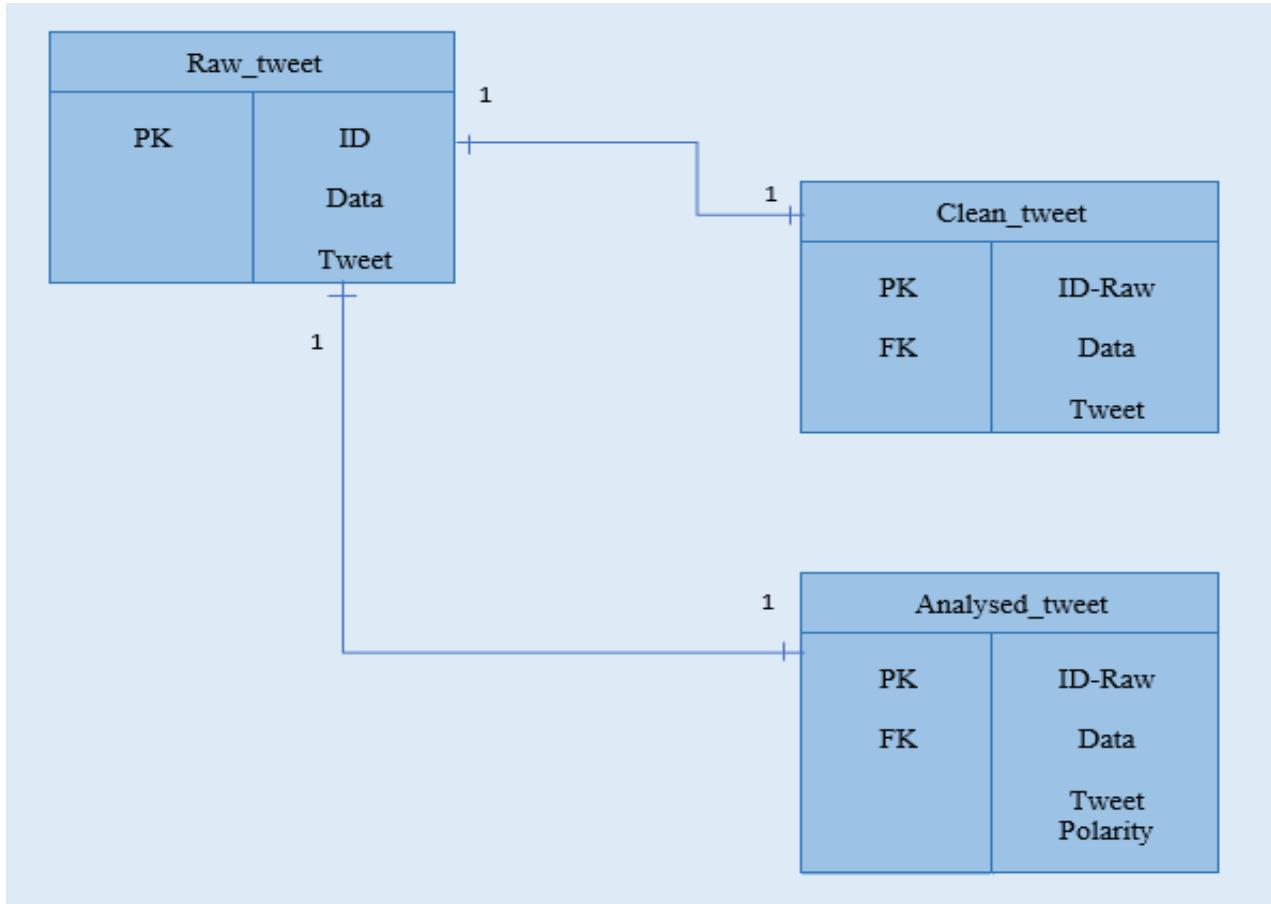
Permanent storage should be within the tools capacity, and within its ability to retrieve required information when required as evidence for legal purposes or as needed. This will give an opportunity for deep analysis and taking note of trends.

Not everything is stored but that which adds value. To begin with, a data model is defined and in this case an Entity Relationship Diagram (ERD) which basically consists of three entities. These three entities: raw_Tweet, clean_Tweet and analysis_Tweet are connected with a one to many relationship. A Tweet is stored as, raw Tweet, clean Tweet and analysed as either Positive, negative or a neutral Tweet. The Primary Key is the ID of the Tweet generated by Twitter. The ID identifies each Tweet thus ensuring that Tweets are no duplicated. To ensure that stored data is not cluttered, a Primary Key is identified which would uniquely identifies all table records.

A Primary Key is the same in all entities or can be the Foreign Key of child entities. Tweets either raw Tweets, cleaned Tweets and analysed Tweets are the ones stored. This are the entities of the database. Since entities in database become the tables, tables are created in which these Tweets can be stored. These tables include: raw_Tweets, cleaned_Tweets and Analysed Tweets. The common identifier through all these tables is ID, which is our Primary Key in table raw_Tweet and the Foreign Key of tables clean_Tweet and analysis_Tweet. These is as shown in Figure 4.1 where PK is Primary Key and FK is Foreign Key.

This is as shown in Figure 4.1. The schema is implemented in the developed tool and the relevant relationship enforced. The main table where data first got committed is on raw_Tweets. The integrity and relationship of table cleaned_Tweets and analysed_Tweets is enforced by the key index, ID.

Figure 4.1: The Database Schema



4.3 System Architecture

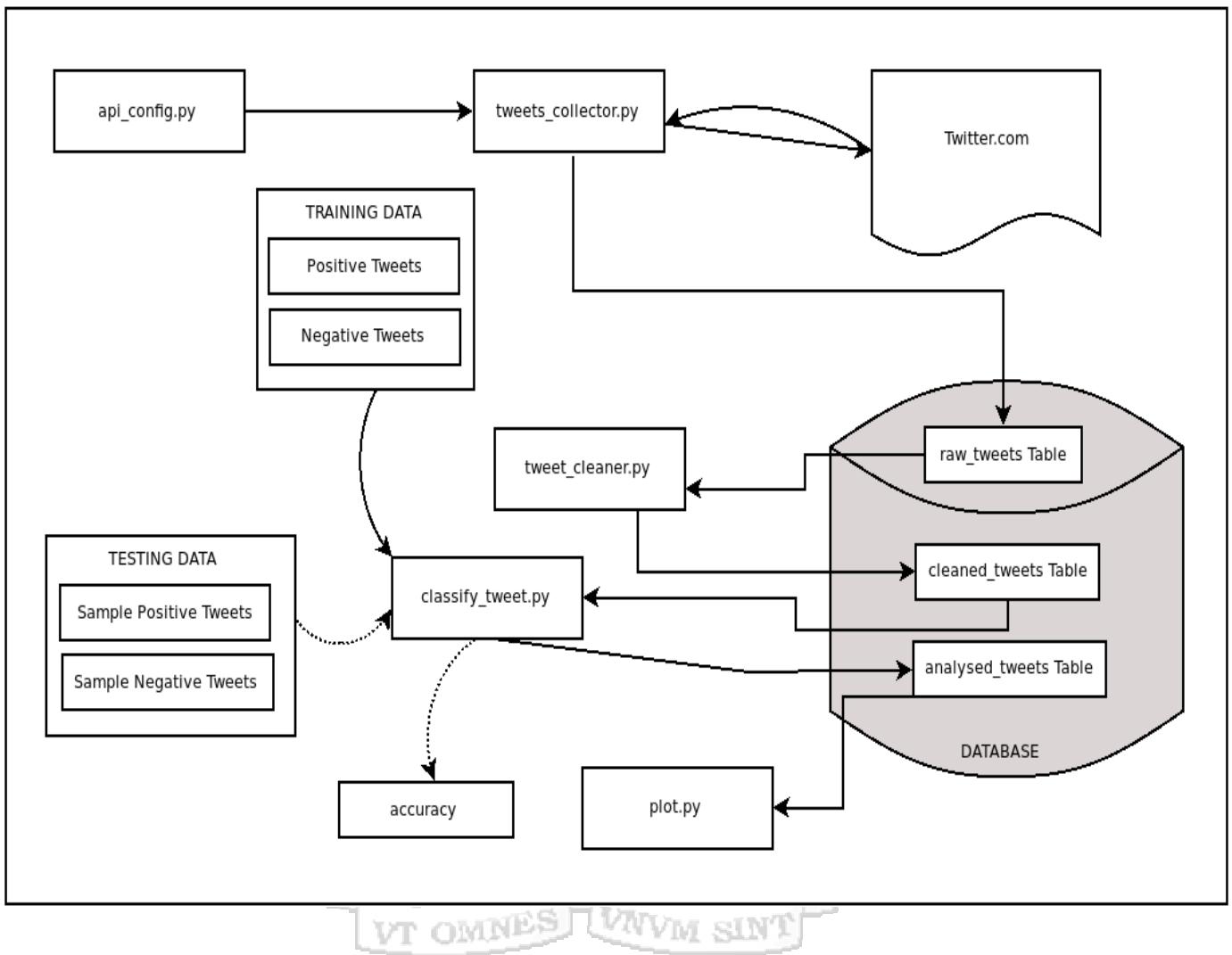
The system architecture depicts the hate speech detection, analysis and monitoring and the components of the tool. User supplied Twitter API keys were set on the Tweets collector module from API configuration file. The Tweets collector module connects to Twitter via a Streaming

API. It was also fitted with hate related key-words that are locally used; these key-words were used to filter through the stream of live Tweets from Twitter.

The collected Tweets were stored in a local database on a table named ‘raw_Tweets’. The raw Tweets were later parsed through a cleansing module (Tweet_cleaner) which striped off all unwanted text data and replaced user handles with a generic one. According to Twitter research data policy, no user information should be captured for research (“Restricted use of Twitter APIs”, 2018). These cleaned Tweets were stored back in the database, on a different table named ‘cleaned_Tweets’.

A custom Naïve Bayes sentimental classifier based on NTLK library was used to analyse and classify the cleaned Tweets from table ‘cleaned_Tweets’ (Vishal & Sonawane, 2016). The classifier was trained with a manually sorted set of positive and negative Tweets. Prior to deploying the classifier, a sample of test data was used to test for its accuracy score. A polarity tag was included for each analysed Tweet. If a Tweet had a higher frequency of negative words, it was tagged as a negative Tweet (neg) and if it had a higher frequency of positive words, it was tagged as a positive Tweet (pos). However, if both the negative and positive words were equal, the Tweet was treated as a neutral Tweet. These analysed Tweets were stored back on the database on a different table named ‘analysed_Tweets’. This is shown in Figure 4.2 below.

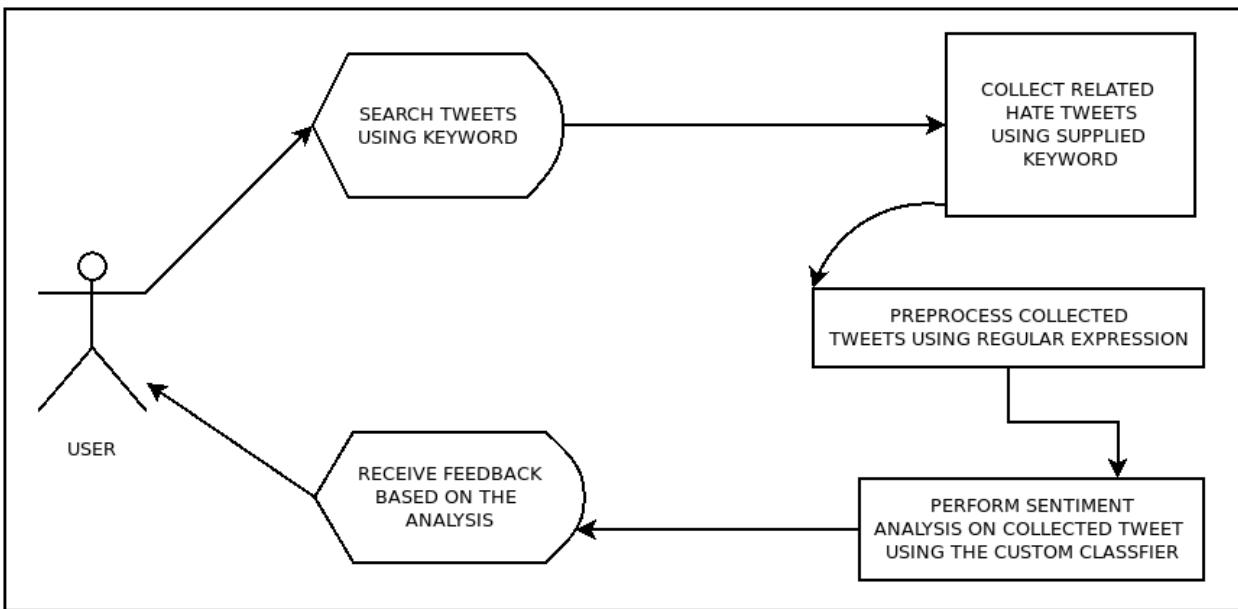
Figure 4.2: System Design Architecture



4.4 Use Case Diagram

The interaction between users and the tool is well depicted by using a Use Case Diagram as shown below in Figure 4.3.

Figure 4.3: Use Case Diagram



4.4.1 Detailed Use Case Descriptions

This section provides comprehensive descriptions of the use case in Figure 4.3 above.

Use case: Search hate speech Tweets, Retrieve hate speech Tweets.

a. Primary Actors

- User (Enter the Keyword to be used).
- Some of the keywords used are listed in Table 4.1 below.

Table 4.1: User Keywords

<ul style="list-style-type: none"> • Handcheque • Raila • uhuru • kikuyu thief • no raila • no peace • Wakikuyu ni wajinga • katakata • kill Raila 	<ul style="list-style-type: none"> • kill kikuyu • kunaswa • real raila • jaluo wajinga • muginki • chikoror • nasa tibim • no ruto no presi • uhurutothieves • wajaka ni wajinga 	<ul style="list-style-type: none"> • wakamba • maembe • no election no peace • wewe ndio kusema • handshake • uhuru-raila • luo lives matters • not my president
---	---	--

	<ul style="list-style-type: none"> • no Kamba no president 	<ul style="list-style-type: none"> • chubukati burns Kenya
--	---	---

- Twitter API (Fetch Tweets from Social media).
- Sentimental analysis (classification of hate-based Tweets).

b. Preconditions

- Internet access on platform being used by user.
- Search Hate Speech use case completed successfully.

c. Hate Speech Analysis platform

- System fetches Tweets from Twitter, in this case Twitter REST API and Streaming API matching the keywords provided by the user.
- Fetched Tweets are preprocessed.
- It analysis the Tweets and classifies as positive, negative or Neutral.
- Automatically save the Tweets after classification in a database.

d. User Feedback

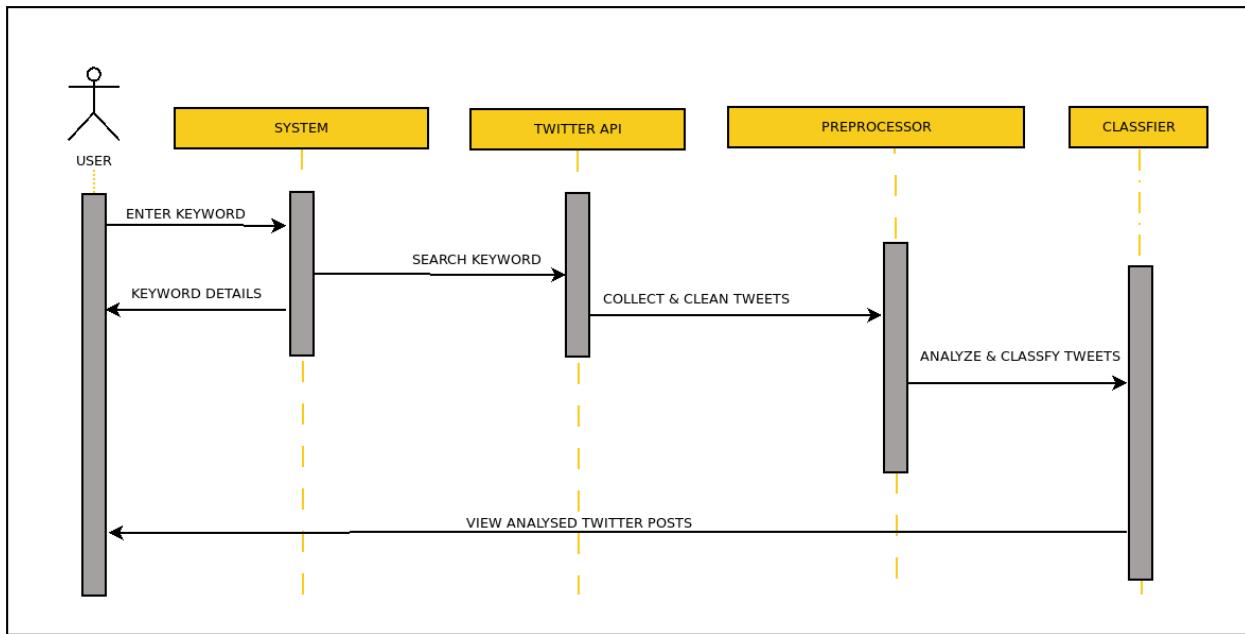
The User Views Tweets labelled as hate speech, analyses them then prepares the report and exit the tool.

4.5 Sequence Diagram

Sequence diagrams usually show the relation and interactions between users and the proposed tool and interactions between the internal components of the tool. A search parameter generated by the user is used on the social media to search for Tweets in relation to the keyword(s) used. Once the keywords are obtained, they are passed on from Twitter API to the Analysis platform for preprocessing. The Analysis platform is also known as the processor. When processed, they are passed on to the classifier for analysis and classification as either positive,

negative or neutral and results saved in the database. Finally, the generated result in the database is used to prepare a report depending on the person or topic of interest. This is as illustrated in Figure 4.4 below.

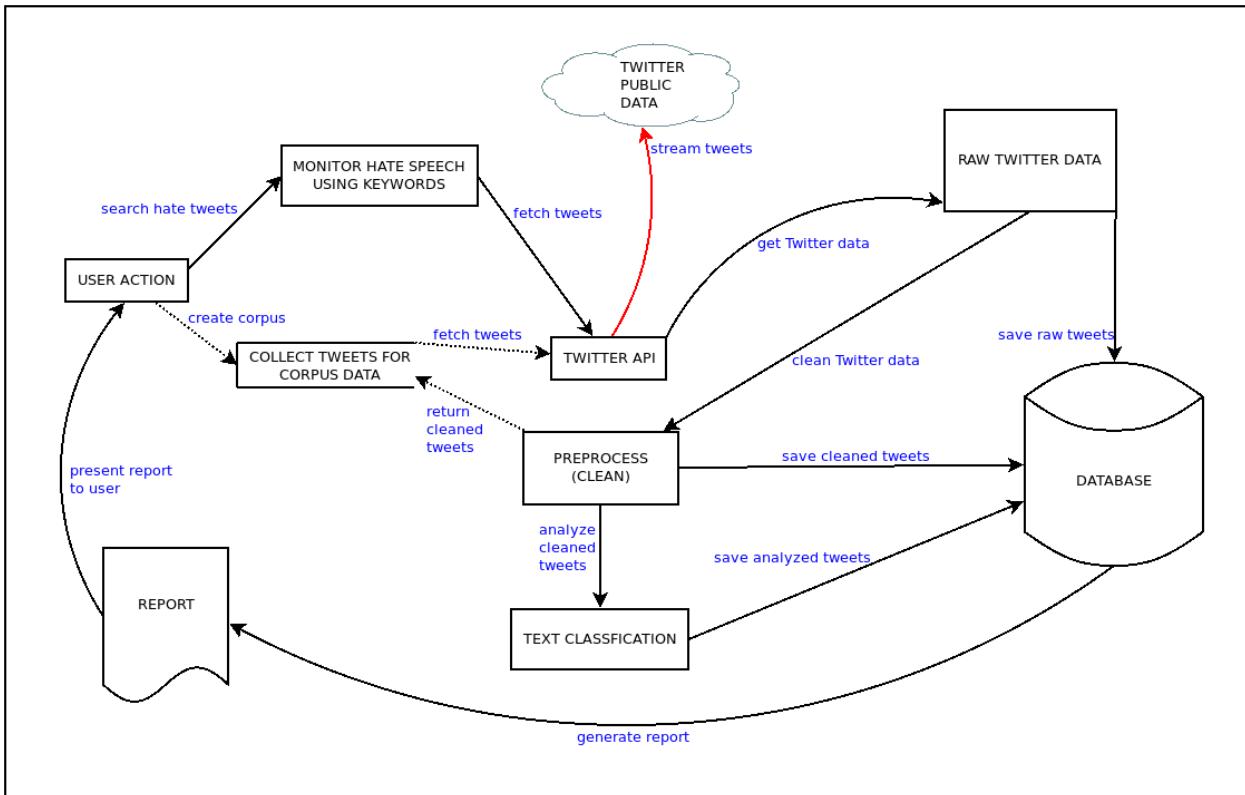
Figure 4.4: Sequence Diagram



4.6 System Analysis

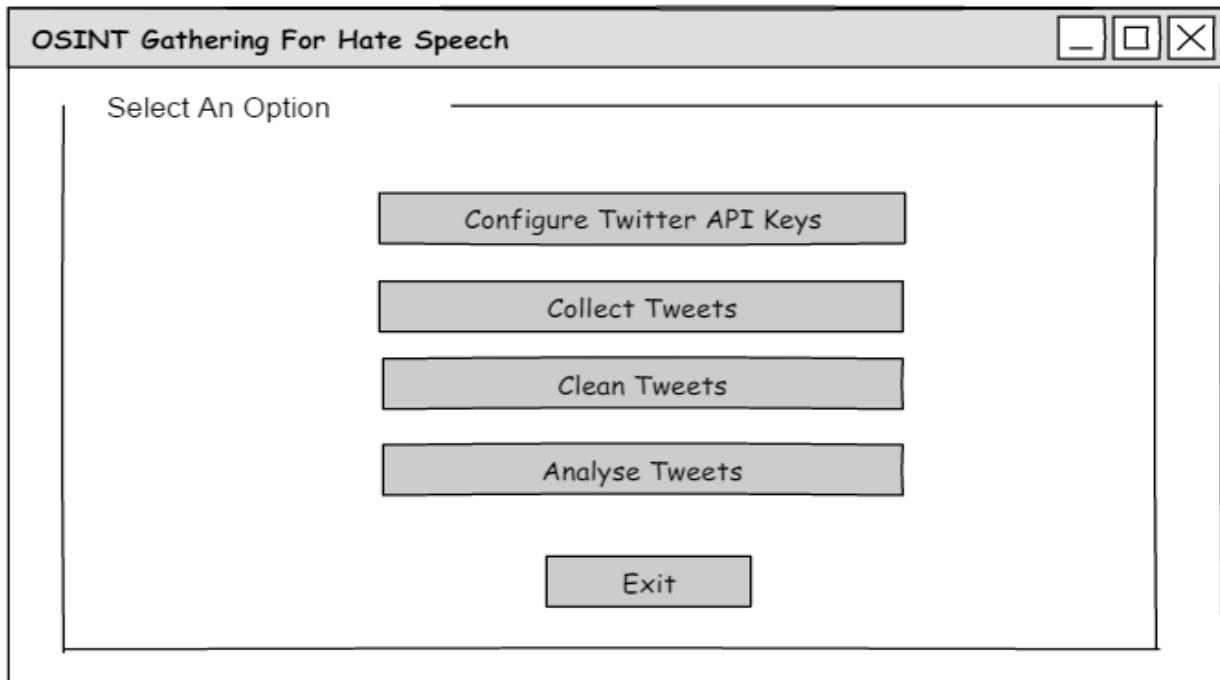
This describes the process of data collection, facts interpretation, problem identification and system decomposition into its interdepended components. With the research focused on the development of a hate speech detection, analysing and monitoring tool on social media in Kenya, this section outlines the user expectations i.e. services required and constraints under which the tool will operated and its development. This is as shown in the Data Flow Diagram Figure 4.5 below.

Figure 4.5: Data Flow Diagram



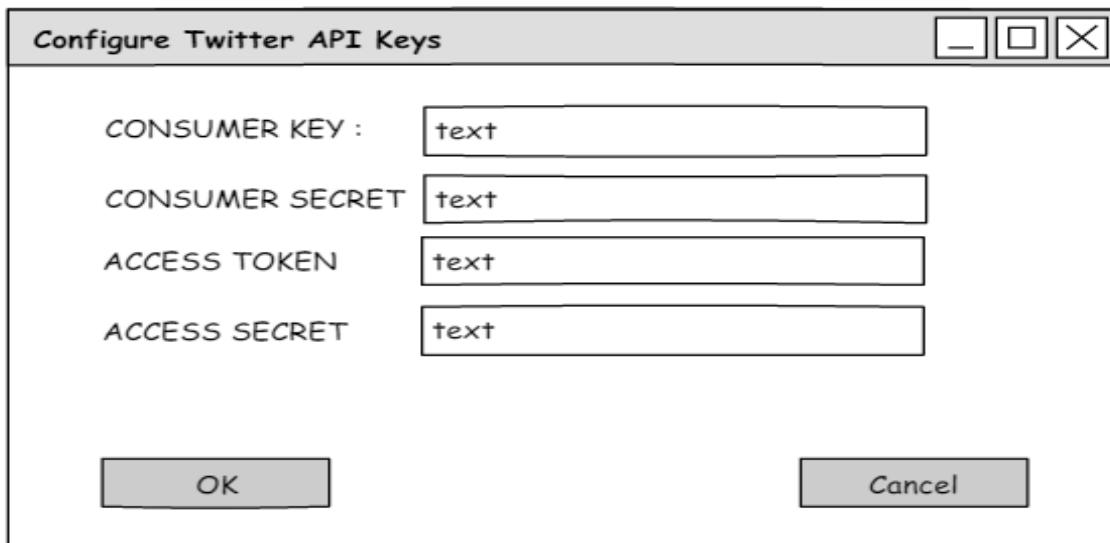
The Graphical User Interface (GUI) was designed using Wireframe. The first graphical user interface; OSINT gathering for hate speech as illustrated in Figure 4.6 below giving the user a choice of selecting the operation they intend to carry out.

Figure 4.6: OSINT Gathering for Hate Speech



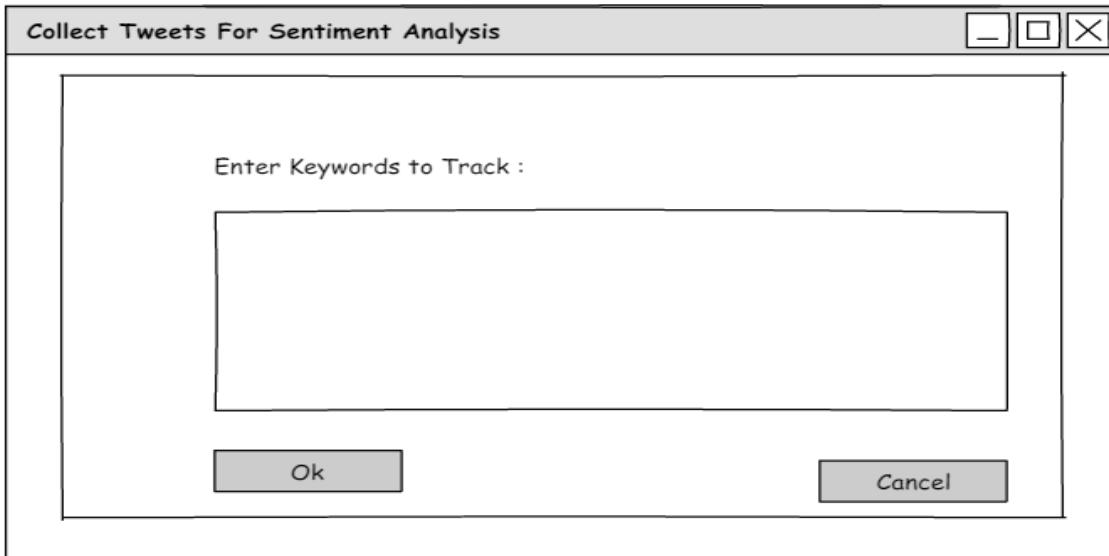
The second GUI as shown in Figure 4.7 is Configure Twitter API Keys. It is the interface bridging between the Twitter account and the application.

Figure 4.7: Configure Twitter API Keys



Data acquisition facilitated by Twitter Search APIs is illustrated by the third GUI, where the user is required to enter keywords to collect Tweets for sentimental analysis as shown in Figure 4.8 below.

Figure 4.8: Tweets Collection for Sentimental Analysis



The last GUI is cleaning Tweets designed to facilitate stripping off of unnecessary information as shown in Figure 4.9 below. This module stripes out the junk data that is not required during the analysis stage. Some of these junk data include URLs and long hashtags. User handles are replaced by a generic text. A custom preprocessor algorithm for cleaning the collected raw Tweets was developed. The Figure 4.10 below shows the summary of how the algorithm was implemented in Python. The grey boxes were accomplished using default regular expression library in Python while the ones in red boxes was accomplished by the use of NTLK Python Library.

Figure 4.9: Cleaning Tweets

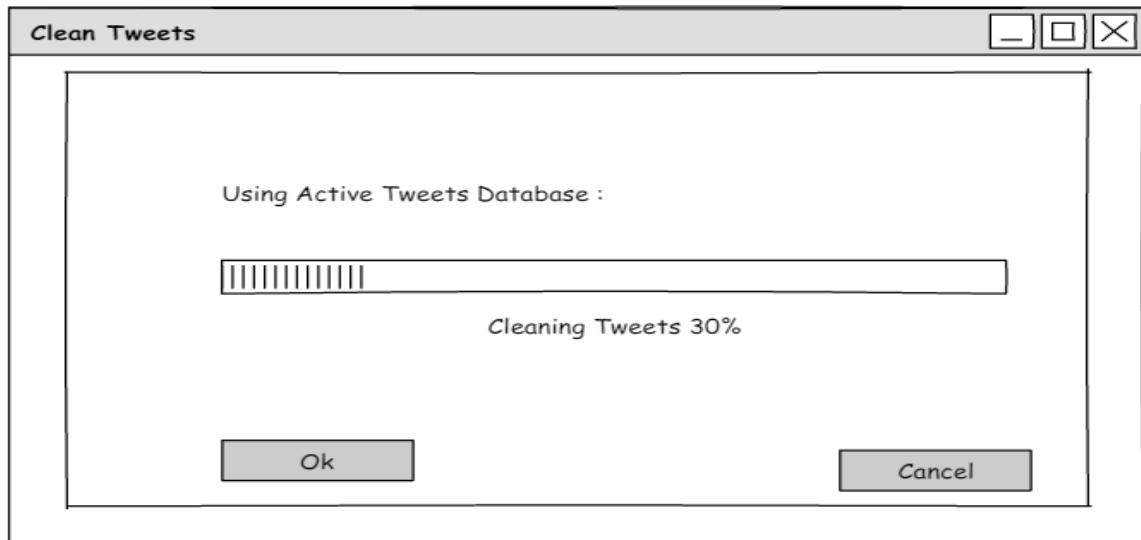
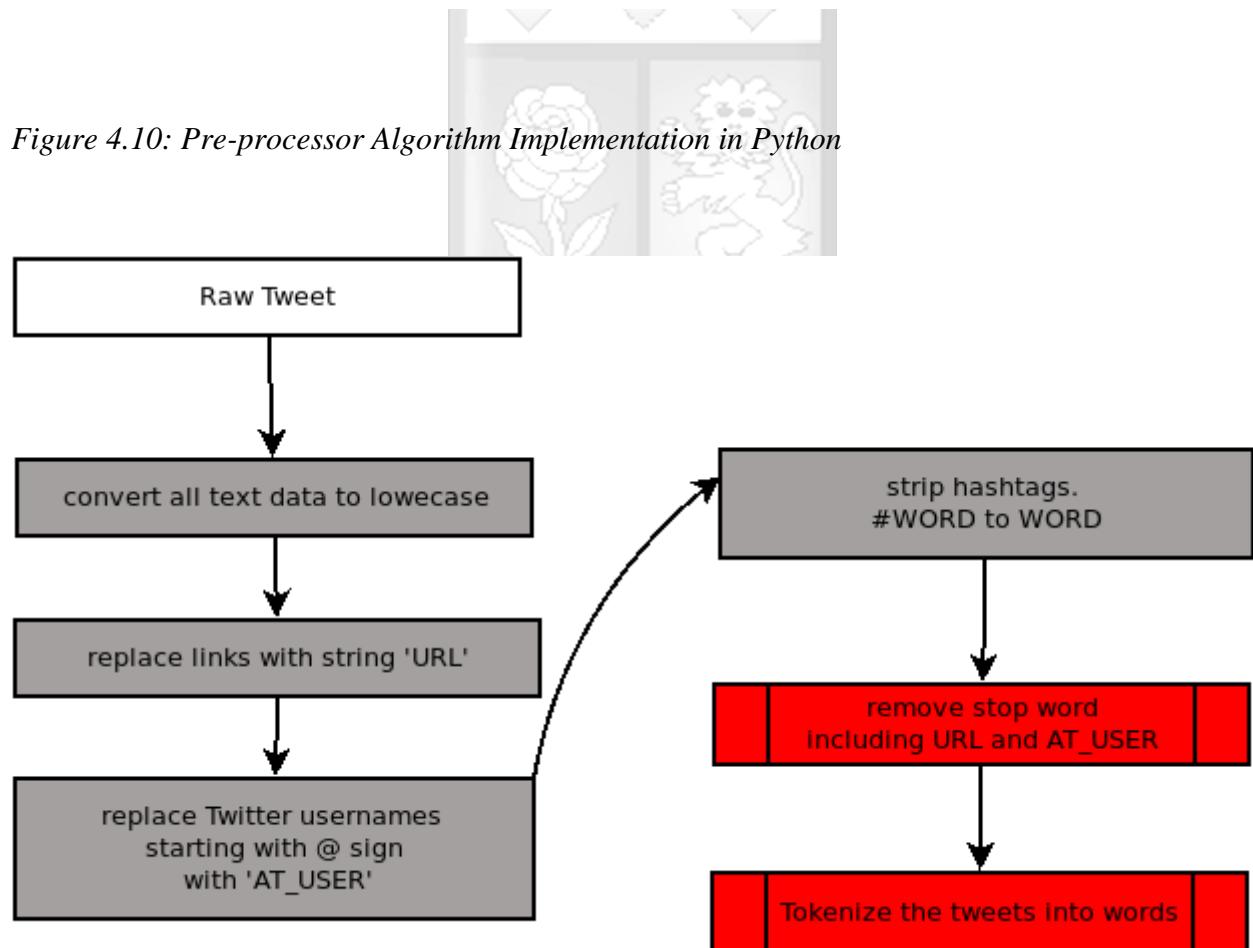


Figure 4.10: Pre-processor Algorithm Implementation in Python



CHAPTER FIVE: SYSTEM IMPLEMENTATION, TESTING AND RESULTS

5.1 Introduction

This chapter gives a clear description of the implementation, testing and validation of the prototype. It covers the steps an intended user would follow from the configuration to the generation of analysed Tweets. A simple menu-based dialog interface was created using Whiptail, a dialog box creator for Linux Shellscripts (Die, 2018).

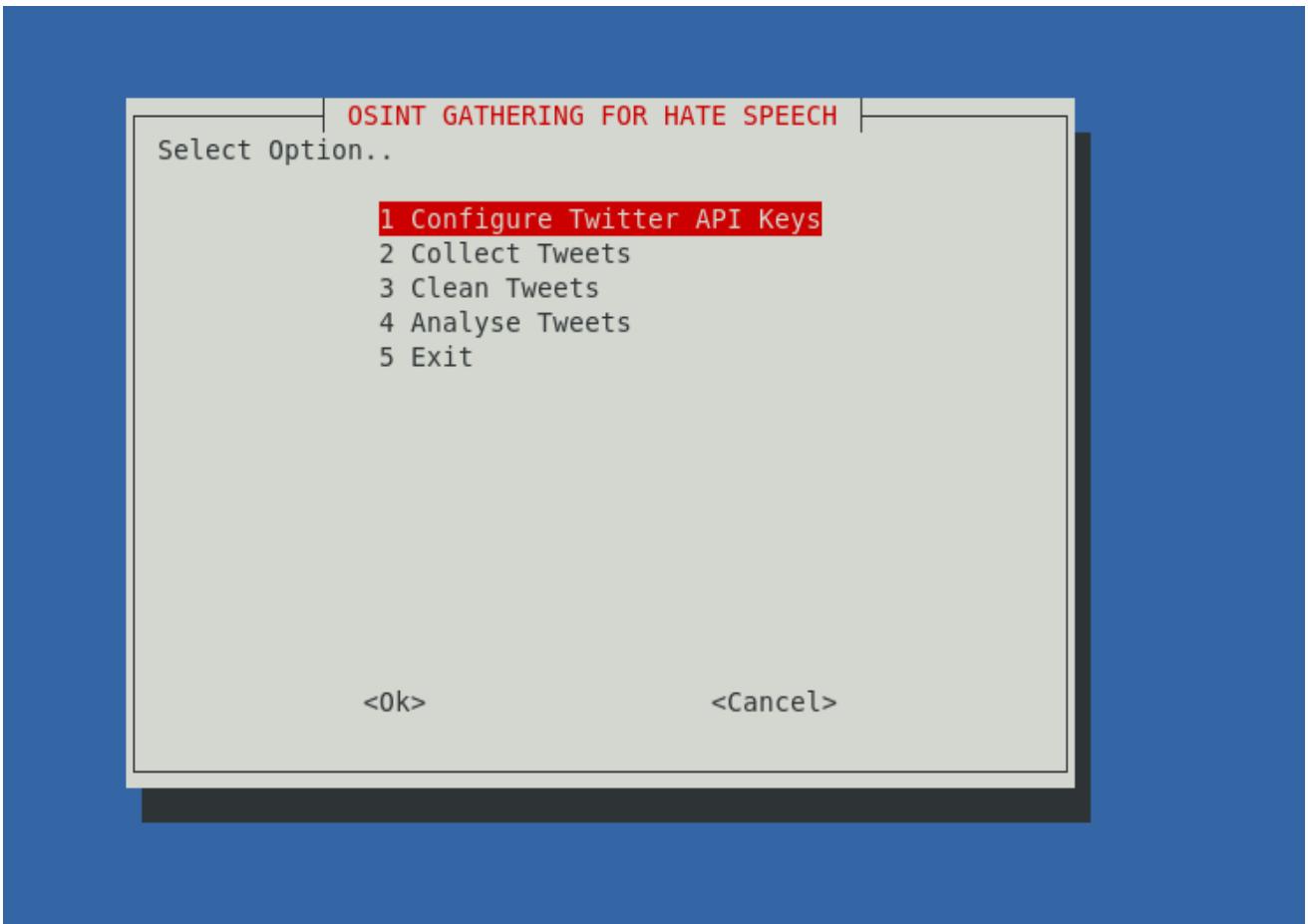
For validation purposes, several experiments were performed in order to identify the best feature types, feature weighting and machine learning algorithms to use for detecting, analysing and monitoring hate speech on social media.

5.2 System Implementation

5.2.1 Main Dialog Interface

The tool has a minimalist menu of options and can perform in a logical order of operations. These range from Twitter API setup to the final analysis of Tweets collected as shown in the Figure 5.1 below.

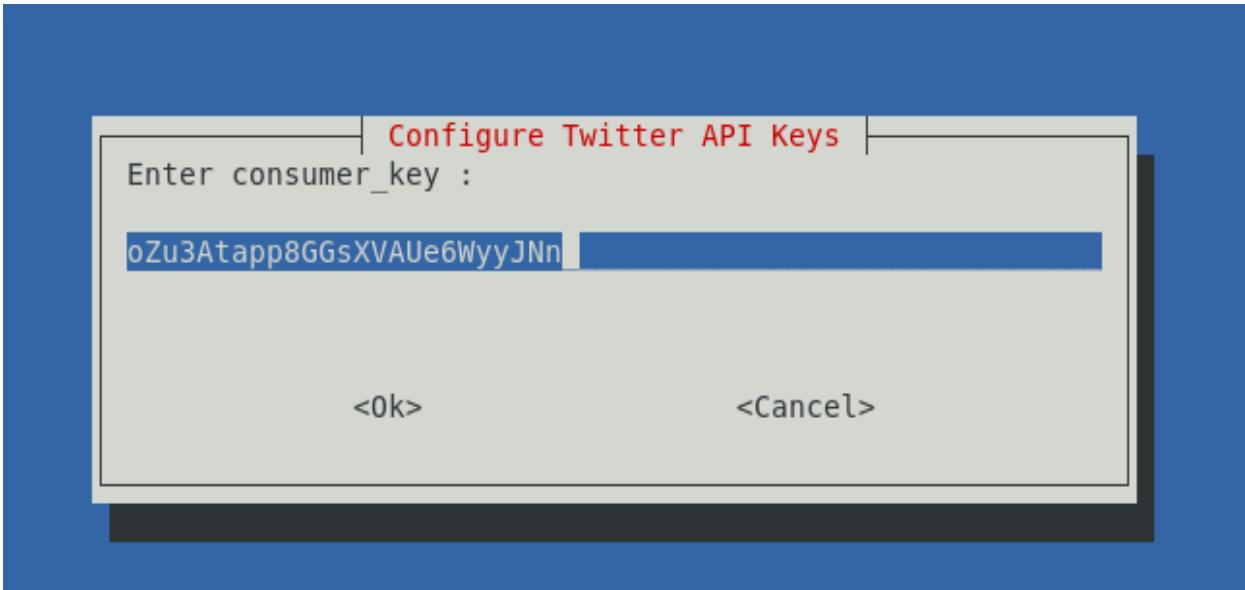
Figure 5.1: Main Dialog Interface



5.2.2 Twitter API Configuration

The configuration builder module of the application is created to act as an interface between the Twitter account and the application, the application was created successfully and it generated Consumer Key, Consumer Secret, Access Token and Access Token Secret to the developer to be used by the tool in collecting the data from Twitter as illustrated by Figure 5.2 below.

Figure 5.2: Twitter API Configuration



5.2.3 Collection of Tweets

Twitter data acquisition was implemented by the use of Twitter's public Streaming API. This API allows developers to have a long-lived connection that enables collection of almost real-time data as it is being posted. Part of the collected data was manually sorted and used as training dataset for the Naïve Bayes classifier. Twitter's Search REST API is different from the Streaming API, it supports short-lived connections and are rate-limited. REST API allow access to Twitter data such as status updates and user info regardless of time although Twitter limits access to a weekly archived data. All the test data used for this study was collected via the REST API.

Once the Twitter API Keys are configured, the tool is able to connect to Twitter and collect Tweets that are relevant to the study by filtering the stream of Tweets using the user supplied keywords as shown in the Figure 5.3 and 5.4 respectively.

Figure 5.3: Tweets Collection for Sentiment Analysis

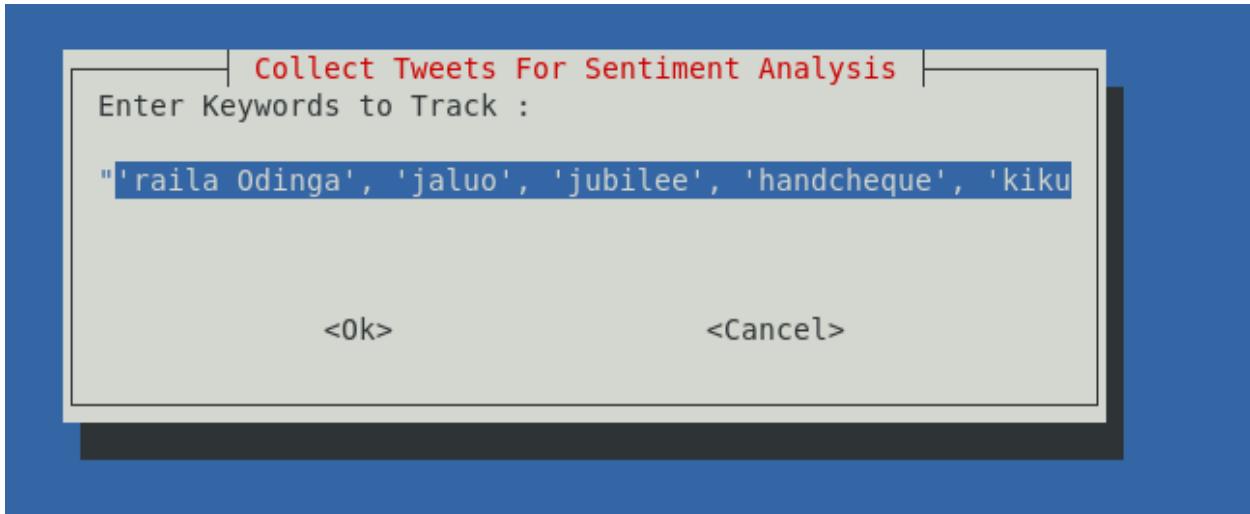


Figure 5.4: Collected Tweets the Output of Figure 5.3

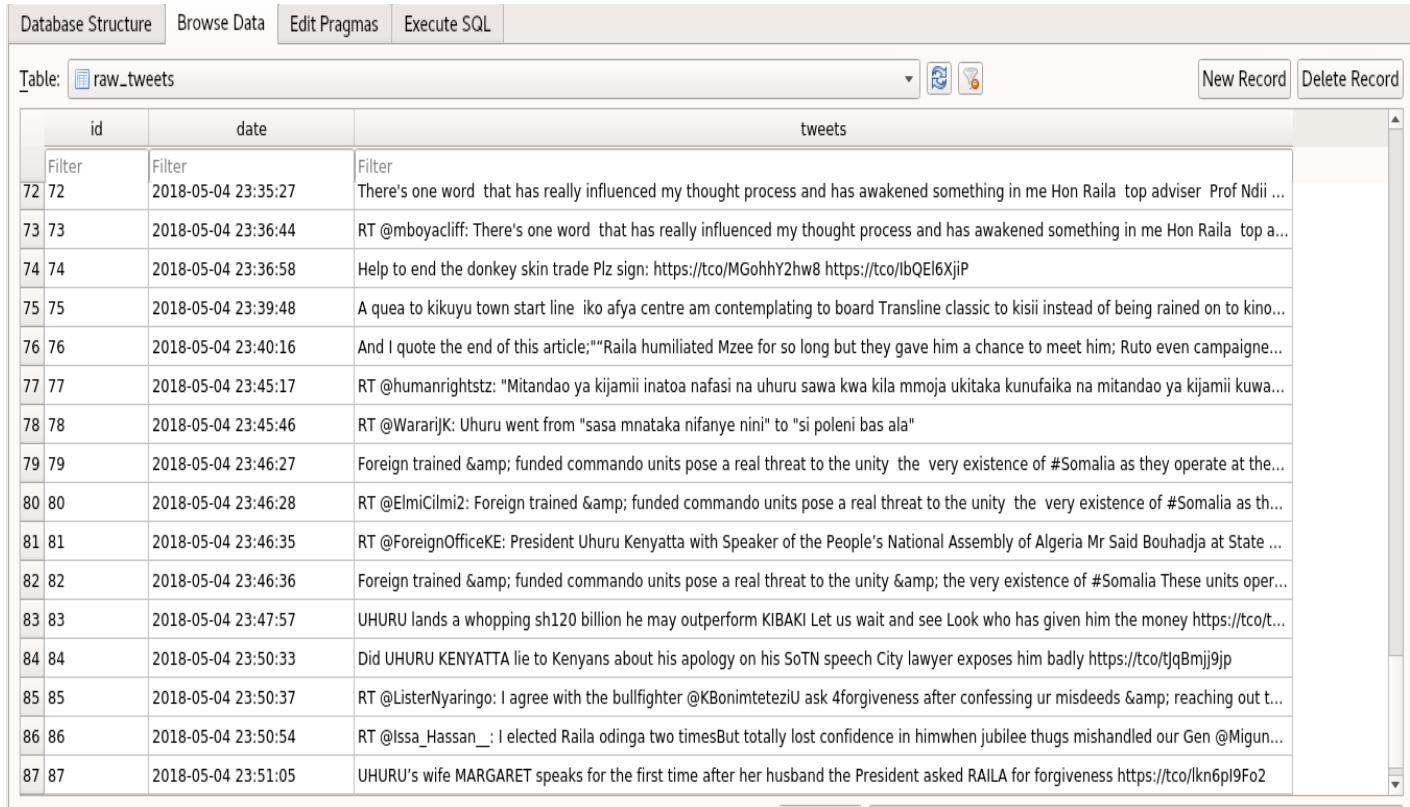
```

2018-05-04 23:30:48 RT @ListerNyaringo: I agree with the bullfighter @KBonimteteziU ask 4forgiveness after confessing ur misdeeds & reaching o
ut to those u of...
2018-05-04 23:31:16 President Uhuru meets East Africa Legislative Assembly Members
https://t.co/JcUVN5LszS
2018-05-04 23:32:24 Your are dreaming with the current set up don't be shocked https://t.co/dlpF2yQSiM
2018-05-04 23:32:49 RT @WarariJK: Uhuru went from "sasa mnataka nifanye nini" to "si poleni bas ala"
2018-05-04 23:35:27 There's one word that has really influenced my thought process and has awakened something in me
Hon Raila top adviser Prof Ndii said " To the youthwe are here solve past and present problems which does not https://t.co/LCDvITLXCx
2018-05-04 23:36:44 RT @mboyacliff: There's one word that has really influenced my thought process and has awakened something in me
Hon Raila top adviser...
2018-05-04 23:36:58 Help to end the donkey skin trade Plz sign: https://t.co/MGohhY2hw8 https://t.co/IbqEl6XjiP
2018-05-04 23:39:48 A quea to kikuyu town start line iko afya centre am contemplating to board Transline classic to Kisii instead of being rained
on to kinoisad life
2018-05-04 23:40:16 And I quote the end of this article;""Raila humiliated Mzee for so long but they gave him a chance to meet him; Ruto even camp
aigned with Mzee in the 2010 constitutional referendum but he was denied to meet Mzee" the analyst said" That is how human nature is
2018-05-04 23:45:17 RT @humanrightstz: "Mitandao ya kijamii inatoa nafasi na uhuru sawa kwa kila mmoja ukitaka kunufaika na mitandao ya kijamii ku
wa mkweli na...
2018-05-04 23:45:46 RT @WarariJK: Uhuru went from "sasa mnataka nifanye nini" to "si poleni bas ala"
2018-05-04 23:46:27 Foreign trained & funded commando units pose a real threat to the unity the very existence of #Somalia as they operate a
t the whims of their masters only Target is beyond Alshabab It is the vast untapped OIL deposits No one dares speak for this nation https://t.co/mQRuSEWKK
2018-05-04 23:46:28 RT @ElmiCilm2: Foreign trained & funded commando units pose a real threat to the unity the very existence of #Somalia a
s they operate at...
2018-05-04 23:46:35 RT @ForeignOfficeKE: President Uhuru Kenyatta with Speaker of the People's National Assembly of Algeria Mr Said Bouhadja at St
ate House N...
2018-05-04 23:46:36 Foreign trained & funded commando units pose a real threat to the unity & the very existence of #Somalia These units o
perate at the whims of their masters Remember the mutiny by UAE trained solders The target is beyond Alshabab It is simply the vast untapped OIL depo
sits
2018-05-04 23:47:57 UHURU lands a whopping sh120 billion he may outperform KIBAKI Let us wait and see Look who has given him the money https://t.co/thkk4IpSH9
```

After collecting enough Tweets, the tool then saves the raw Tweets in a local database, on

a table named raw_Tweets as show in the Figure 5.5 below.

Figure 5.5: Raw Tweets Saved in Table Raw_Tweets



Database Structure			Browse Data	Edit Pragmas	Execute SQL
Table: raw_tweets					
id	date	tweets			
72	Filter	Filter	Filter		
72	2018-05-04 23:35:27	There's one word that has really influenced my thought process and has awakened something in me Hon Raila top adviser Prof Ndii ...			
73	73	2018-05-04 23:36:44	RT @mboyacliff: There's one word that has really influenced my thought process and has awakened something in me Hon Raila top a...		
74	74	2018-05-04 23:36:58	Help to end the donkey skin trade Plz sign: https://t.co/MGohhY2hw8 https://t.co/lbQE16XjiP		
75	75	2018-05-04 23:39:48	A quea to kikuyu town start line iko afya centre am contemplating to board Transline classic to kisii instead of being rained on to kino...		
76	76	2018-05-04 23:40:16	And I quote the end of this article; "Raila humiliated Mzee for so long but they gave him a chance to meet him; Ruto even campaigne...		
77	77	2018-05-04 23:45:17	RT @humanrightsztz: "Mitandao ya kijamii inatoa nafasi na uhuru sawa kwa kila mmoja ukitaka kunufaika na mitandao ya kijamii kuwa...		
78	78	2018-05-04 23:45:46	RT @Wararijk: Uhuru went from "sasa mnataka nifanye nini" to "si poleni bas ala"		
79	79	2018-05-04 23:46:27	Foreign trained & funded commando units pose a real threat to the unity the very existence of #Somalia as they operate at the...		
80	80	2018-05-04 23:46:28	RT @ElmiCilmi2: Foreign trained & funded commando units pose a real threat to the unity the very existence of #Somalia as th...		
81	81	2018-05-04 23:46:35	RT @ForeignOfficeKE: President Uhuru Kenyatta with Speaker of the People's National Assembly of Algeria Mr Said Bouhadja at State ...		
82	82	2018-05-04 23:46:36	Foreign trained & funded commando units pose a real threat to the unity & the very existence of #Somalia These units oper...		
83	83	2018-05-04 23:47:57	UHURU lands a whopping sh120 billion he may outperform KIBAKI Let us wait and see Look who has given him the money https://t.co/t...		
84	84	2018-05-04 23:50:33	Did UHURU KENYATTA lie to Kenyans about his apology on his SoTN speech City lawyer exposes him badly https://t.co/tjqBmj9jp		
85	85	2018-05-04 23:50:37	RT @ListerNyaringo: I agree with the bullfighter @KBonimteteziU ask 4forgiveness after confessing ur misdeeds & reaching out t...		
86	86	2018-05-04 23:50:54	RT @Issa_Hassan__: I elected Raila odinga two timesBut totally lost confidence in himwhen jubilee thugs mishandled our Gen @Migun...		
87	87	2018-05-04 23:51:05	UHURU's wife MARGARET speaks for the first time after her husband the President asked RAILA for forgiveness https://t.co/lkn6pl9Fo2		

5.2.4 Cleaning of Tweets

The raw Tweets collected in the previous module as indicated in 5.2.3, usually contain other unrelated junk data as well as personal information regarding the accounts that created the collected Tweets. Twitter has regulations on how to use its data for research, and one of its requirement is to conceal any Twitter account information (“Restricted use of Twitter APIs”, 2018).

This module stripes out the junk data that is not required during the analysis stage as shown in the Figure 5.6 below. Some of these junk data include URLs and long hashtags. User handles are replaced by a generic text. When this module is finished cleaning the Tweets from

raw_Tweets table, it saves the cleaned Tweets to a different table named cleaned_Tweets, in the same database as shown in the Figure 5.6 and 5.7 below.

Figure 5.6: Cleaning Tweets

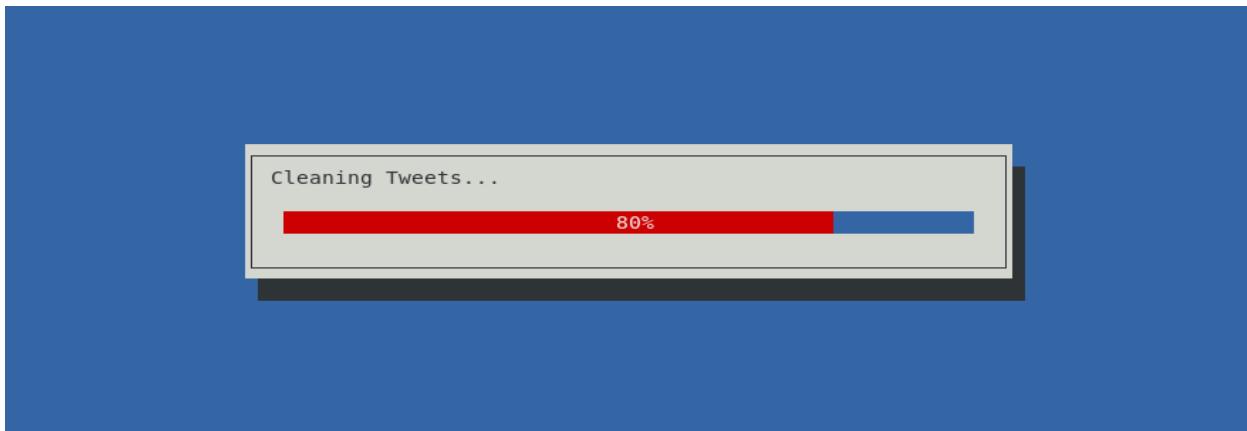


Figure 5.7: Cleaned Tweets Saved in Table Cleaned_Tweets

Database Structure			Browse Data	Edit Pragmas	Execute SQL
			Table: cleaned_tweets		New Record Delete Record
	id	date	clean_tweet		
	id	date	clean_tweet		
75	75	2018-05-04 23:3...	a quea to kikuyu town start line iko afya centre am contemplating to board transline classic to kisii instead of being rained on to kinoosad life		
76	76	2018-05-04 23:4...	and i quote the end of this article; "raila humiliated mzee for so long but they gave him a chance to meet him; ruto even campaigned with mzee in the 201...		
77	77	2018-05-04 23:4...	rt at_user "mitandao ya kijamii inatoa nafasi na uhuru sawa kwa kila mmoja ukitaka kunufaika na mitandao ya kijamii kuwa mkweli na...		
78	78	2018-05-04 23:4...	rt at_user uhuru went from "sasa mnataka nifanye nini" to "si poleni bas ala		
79	79	2018-05-04 23:4...	foreign trained & funded commando units pose a real threat to the unity the very existence of somalia as they operate at the whims of their masters o...		
80	80	2018-05-04 23:4...	rt at_user foreign trained & funded commando units pose a real threat to the unity the very existence of somalia as they operate at...		
81	81	2018-05-04 23:4...	rt at_user president uhuru kenyatta with speaker of the people's national assembly of algeria mr said bouhadja at state house n...		
82	82	2018-05-04 23:4...	foreign trained & funded commando units pose a real threat to the unity & the very existence of somalia these units operate at the whims of thei...		
83	83	2018-05-04 23:4...	uhuru lands a whopping sh120 billion he may outperform kibaki let us wait and see look who has given him the money url		
84	84	2018-05-04 23:5...	did uhuru kenyatta lie to kenyans about his apology on his sotn speech city lawyer exposes him badly url		
85	85	2018-05-04 23:5...	rt at_user i agree with the bullfighter at_user ask 4forgiveness after confessing ur misdeeds & reaching out to those u of...		
86	86	2018-05-04 23:5...	rt at_user i elected raila odinga two timesbut totally lost confidence in himwhen jubilee thugs mishandled our gen at_user		
87	87	2018-05-04 23:5...	uhuru's wife margaret speaks for the first time after her husband the president asked raila for forgiveness url		
88	88	2018-05-04 23:5...	uhuru doubled expenditure on hospitality report shows :: kenya - url url		
89	89	2018-05-04 23:5...	ali skalled & uhuru - nge hamba (original mix)		
90	90	2018-05-04 23:5...	rt at_user *looks at avi* *looks at bio* this tweet checks out url		

5.2.5 Preprocessing of Tweets

The cleaned Tweets are ready to be parsed on the classifier. There are three possible results from the classifier, negative, positive or neutral. Prior to this analysis, the classifier was trained by a set of positive and negative that are manually sorted and later tested with a sample of Tweets to calculate its accuracy level. The trained Naïve Bayes based-classifier was used to classify the Tweets. If the frequency of negative sentiments in a Tweet is high, it is tagged as negative (neg), if the frequency of the positive words is high it is tagged as a positive Tweet (pos). However, if the frequency of positive and negative words is the almost the same, it is treated as a neutral Tweet as captured in the Figure 5.8 below.

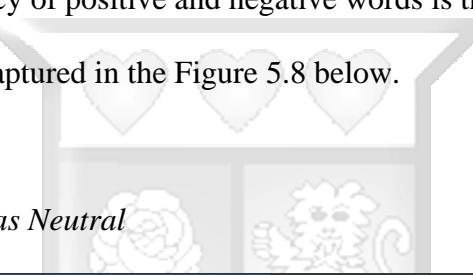


Figure 5.8: Tweets Classified as Neutral

```
(71, u'2018-05-04 23:32:49', u'rt at_user uhuru went from "sasa mnataka nifanye nini" to "si poleni bas ala', 'neg')
(72, u'2018-05-04 23:35:27', u'there's one word that has really influenced my thought process and has awakened something in me hon raila top adviser prof ndii said " t o the youthwe are here solve past and present problems which does not url', 'neg')
(73, u'2018-05-04 23:36:44', u'rt at_user there's one word that has really influenced my thought process and has awakened something in me hon raila top adviser\u2026', 'neg')
(74, u'2018-05-04 23:36:58', u'help to end the donkey skin trade plz sign: url url', 'neg')
(75, u'2018-05-04 23:39:48', u'a quea to kikuyu town start line iko afya centre am contemplating to board transline classic to kisii instead of being rained on to kinoo sad life', 'neg')
(76, u'2018-05-04 23:40:16', u'and i quote the end of this article;"\u201craila humiliated mzee for so long but they gave him a chance to meet him; ruto even campaigned with mzee in the 2010 constitutional referendum but he was denied to meet mzee\u201d the analyst said" that is how human nature is', 'neg')
(77, u'2018-05-04 23:45:17', u'rt at_user "mitandao ya kijamii inatoa nafasi na uhuru sawa kwa kila mmoja ukitaka kunufaika na mitandao ya kijamii kuwa mkweli na\u2026', 'neg')
(78, u'2018-05-04 23:45:46', u'rt at_user uhuru went from "sasa mnataka nifanye nini" to "si poleni bas ala', 'neg')
(79, u'2018-05-04 23:46:27', u'foreign trained & funded commando units pose a real threat to the unity the very existence of somalia as they operate at the whims of their masters only target is beyond alshabab it is the vast untapped oil deposits no one dares speak for this nation url', 'neg')
(80, u'2018-05-04 23:46:28', u'rt at_user foreign trained & funded commando units pose a real threat to the unity the very existence of somalia as they operate at\u2026', 'neg')
(81, u'2018-05-04 23:46:35', u'rt at_user president uhuru kenyatta with speaker of the people\u2019s national assembly of algeria mr said bouhadja at state house n\u2026', 'neg')
(82, u'2018-05-04 23:46:36', u'foreign trained & funded commando units pose a real threat to the unity & the very existence of somalia these units operate at th e whims of their masters remember the mutiny by uae trained solders the target is beyond alshabab it is simply the vast untapped oil deposits', 'neg')
(83, u'2018-05-04 23:47:57', u'uhuru lands a whopping sh120 billion he may outperform kibaki let us wait and see look who has given him the money url', 'neg')
(84, u'2018-05-04 23:50:33', u'did uhuru kenyatta lie to kenyans about his apology on his sotn speech city lawyer exposes him badly url', 'neg')
(85, u'2018-05-04 23:50:37', u'rt at_user i agree with the bullfighter at_user ask 4forgiveness after confessing ur misdeeds & reaching out to those u of\u2026', 'neg')
(86, u'2018-05-04 23:50:54', u'rt at_user i elected raila odinga two timesbut totally lost confidence in himwhen jubilee thugs mishandled our gen at_user', 'neg')
(87, u'2018-05-04 23:51:05', u'uhuru\u2019s wife margaret speaks for the first time after her husband the president asked raila for forgiveness url', 'neg')
(88, u'2018-05-04 23:54:09', u'uhuru doubled expenditure on hospitality report shows :: kenya - url url', 'neg')
(89, u'2018-05-04 23:54:42', u'ali skalled & uhuru - nge hamba (original mix)', 'neg')
(90, u'2018-05-04 23:56:22', u'rt at_user *looks at avi* *looks at bio* this tweet checks out url', 'neg')
root@nghula:/root/Desktop/DataSpeech #
```

Once classified as Positive, Negative or neutral, each Tweet is tagged with its polarity. If it was classified as positive it is tagged positive. Negative and neutral classified Tweets are tagged negative and neutral respectively. All Tweets are then saved in the database as analysed Tweets of the entered keyword. These new tagged Tweets are stored in a separate table named analysed_Tweets along with their polarity as shown below in Figure 5.9.

Figure 5.9: New Tagged Tweets Stored in Analysed_Tweets Table

Database Structure				Browse Data	Edit Pragmas	Execute SQL
Table: analysed_tweets						
	id	date	analysed_tweet	polarity		
	Filter	Filter	Filter	Filter	neg	
78	69	2018-05-04 ...	president uhuru meets east africa legislative assembly members url		neg	
79	70	2018-05-04 ...	your are dreaming with the current set up don't be shocked url		neutral	
80	71	2018-05-04 ...	rt at_user uhuru went from "sasa mnataka nifanye nini" to "si poleni bas ala		neg	
81	72	2018-05-04 ...	there's one word that has really influenced my thought process and has awakened something ...	neg		
82	73	2018-05-04 ...	rt at_user there's one word that has really influenced my thought process and has awakened ...	neg		
83	74	2018-05-04 ...	help to end the donkey skin trade plz sign: url url	neg		
84	75	2018-05-04 ...	a quea to kikuyu town start line iko afya centre am contemplating to board transline classic to...	pos		
85	76	2018-05-04 ...	and i quote the end of this article;""raila humiliated mzee for so long but they gave him a cha...	neg		
86	77	2018-05-04 ...	rt at_user "mitandao ya kijamii inatoa safasi na uhuru sawa kwa kila mmoja ukitaka kunufaika...	pos		
87	78	2018-05-04 ...	rt at_user uhuru went from "sasa mnataka nifanye nini" to "si poleni bas ala	neg		
88	79	2018-05-04 ...	foreign trained & funded commando units pose a real threat to the unity the very existen...	neg		
89	80	2018-05-04 ...	rt at_user foreign trained & funded commando units pose a real threat to the unity the v...	pos		
90	81	2018-05-04 ...	rt at_user president uhuru kenyatta with speaker of the people's national assembly of algeria ...	neg		
91	82	2018-05-04 ...	foreign trained & funded commando units pose a real threat to the unity & the very ...	neutral		
92	83	2018-05-04 ...	uhuru lands a whopping sh120 billion he may outperform kibaki let us wait and see look who ...	pos		
93	84	2018-05-04 ...	did uhuru kenyatta lie to kenyans about his apology on his sotn speech city lawyer exposes hi...	neg		

5.2.6 Training Data

People post messages on Twitter in vernacular language, English, Kiswahili or a mixture of both. The training process is supposed to reveal hidden dependencies and patterns in the Twitter data to be analysed. Training data was collected for use in training the Naïve Bayes classifier.

NTLK library (Manning, C., 2014). was used to collect training data since it has a very large corpus with structured text files useful in training models.

About 100,000 Tweets were collected using Twitter Search API for dates between October 2016 and 2018 May. The searched Tweets were sorted in their respective key words and later cleaned into either positive or negative training dataset. The training data is special and crucial to this study. Previous studies that employed use of sentiment analysis show no record of collecting or using localized training datasets that accurately identify the lingo Kenyans use while on social media platform like Twitter. Figure 5.10 shows the training datasets saved as text files and Figure 5.11 shows content of Kikuyu-Mugiki dataset.

Training data is used in a supervised environment to train tools to map training examples to their corresponding targets. Once training is well implemented, the algorithm can then generalise the training data to new data correctly. The training data used in the proposed system is a representative sample of the hate related data shared on Twitter. The data used to train the Naïve Bayes classifier is a collection of both negative and positive Tweets.

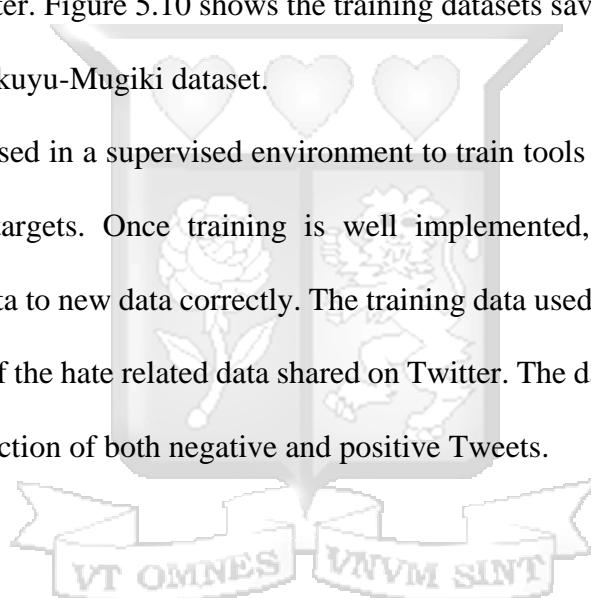


Figure 5.10: Training Dataset

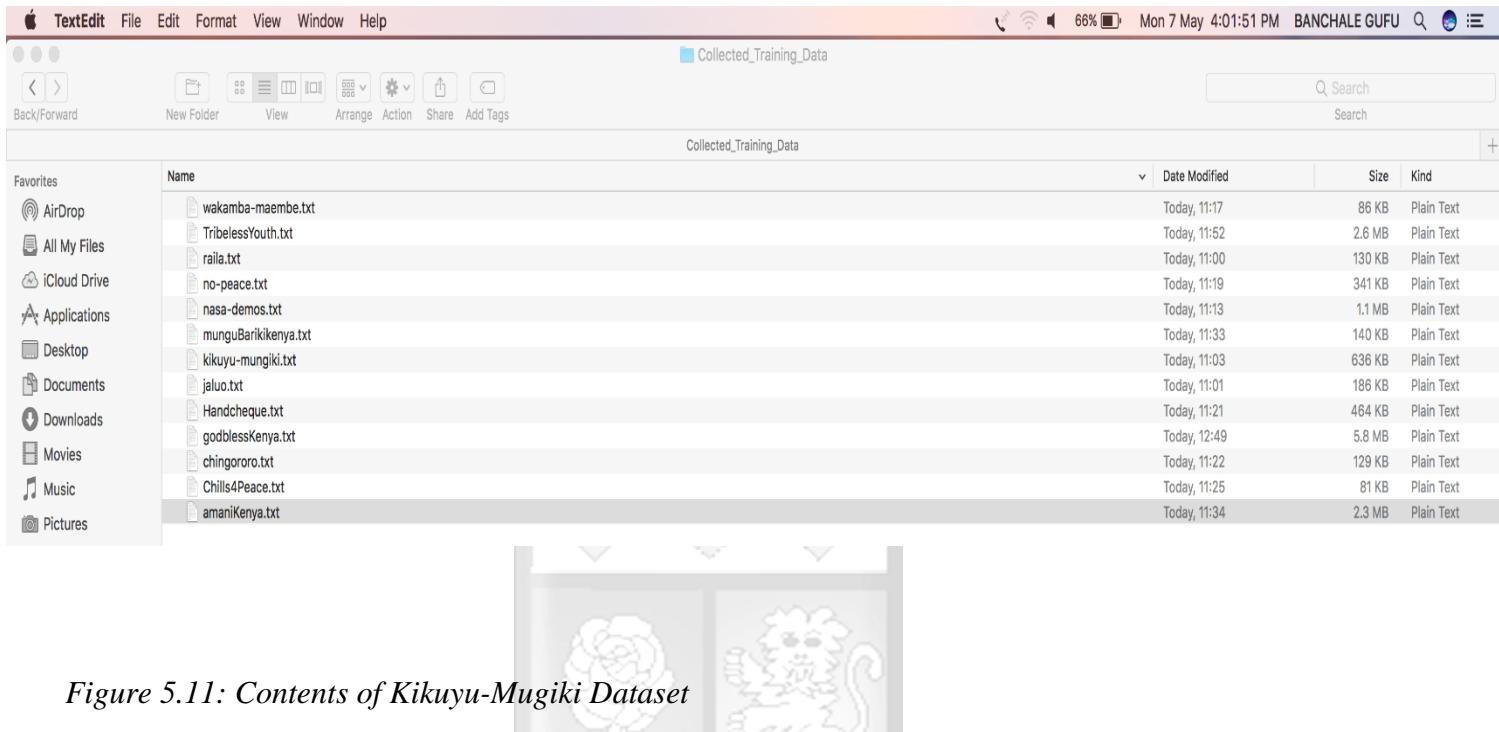


Figure 5.11: Contents of Kikuyu-Mugiki Dataset

kikuyu-mungiki - Notepad

File Edit Format View Help

```
[INFO:requests.packages.urllib3.connectionpool:Starting new HTTPS connection (1): twitter.com]INFO:root:1 [2018-04-24 11:20:14]
ya mafia and some blogger for hire.INFO:root:6 [2018-03-27 11:17:41] - This country of looters like @AnneWaiguru and killers
ot:10 [2018-03-05 22:44:35] - Someone told me that "Mungiki" means m'gikuyu? (M'kikuyu) So Kikuyu's is this facts or nah?INFO
T FOR ITOnce they are done misusing the Mungiki outfit they will kill the Kikuyu teenagers in their hundreds to cheat the pu
against Kitui Governor Charity Ngilu.~Siginon Freight Truck Driver.INFO:root:22 [2018-02-26 22:52:50] - After investigations
I did not know that the sand cartels are this big. How can they sponsor such a widespread demo? Trust me, no Kikuyu, Mungiki
alled mungiki- u need to grow balls @C_NyaKundiH - that lorry burnt is someones hard work and pain NKT!!INFO:root:31 [2018-0
iang!]INFO:root:34 [2018-02-17 10:55:27] - But if not for Raila si Kikuyu wangeisha in as much as Uhuru paid his gang mungik
erious forum? Those jeans wouldn't even fit @RickRoss. Ongezeni nyama, decks na viti mfanye io gathering iwe bash malaya nyin
killed. Ask Kimathi or King'ara whose dream was to make Mungiki 21st century Mau Mau with a cause for land, freedom and just
ungiki) na kalenjinINFO:root:46 [2018-02-07 05:13:37] - Mungiki is recognized by kikuyu'sINFO:root:47 [2018-02-06 19:35:36]
a kikuyu word which means multitude. A group of people assembled at a place is called Mungiki.I thought u should know.INFO:r
t he does as long as they continue enjoying Kikuyu privilege which only amounts to not getting shot and tear gasssed by the
police station in frustration after NASA brigade stayed put #MigunaMigunaArrested #UhuruMustGo #MigunaMigunaFreedomhttps://twi
s nonsense. #MigunaMigunaArrestedINFO:root:64 [2018-02-02 15:31:58] - Illegitimate Jubilee government having panicked jas ar
21:03:37] - Infact kikuyu is mungiki and mungiki is kikuyu a group that should be banned in kenyaINFO:root:70 [2018-01-30 17
ng themselves and those celebrating this will end up cryingINFO:root:74 [2018-01-29 19:07:37] - Kikuyu na uchawi..... Kitael
's talk about it.When a Maasai/Moran sheaths his knife/panga it's traditional and allowed.But if you as a Kikuyu,tries that
...INFO:root:84 [2018-01-08 13:02:23] - If tribalism is what ills Kenya why didn't the Kikuyu tycoons that raised billions f
ungiki. https://twitter.com/adwarizobill/status/949739443929296896 ...INFO:root:89 [2018-01-06 21:46:03] - There is a dictiona
other tribes? Mungiki mentalityINFO:root:93 [2018-01-03 12:34:53] - Stupid kikuyu mungikiINFO:root:94 [2018-01-03 11:11:13]

```

5.3 Testing

The custom `get_word_features` module uses classification to do part-of-speech tagging.

Features are extracted from words, and then passed to an internal classifier. The classifier classifies the features and returns a label, in this case, a polarity tag using a feature detector.

The feature detector finds multiple length suffixes, does some regular expression matching, and looks at the unigram, bigram, and trigram history to produce a fairly complete set of features for each word. The feature sets it produces are used to train the internal classifier, and for classifying words into part-of-speech tags.

The Naïve Bayes algorithm is inherited from the NTLK library and only implements a `feature_detector()` method. All the training and tagging is done in `ClassifierBasedTagger`. Training of the `NaiveBayesClassifier` class is done with the training datasets in the folder named `training_set`. Once this classifier is trained, it is used to classify word features produced by the `feature_detector()` method. Figure 5.12 below illustrates the testing process.

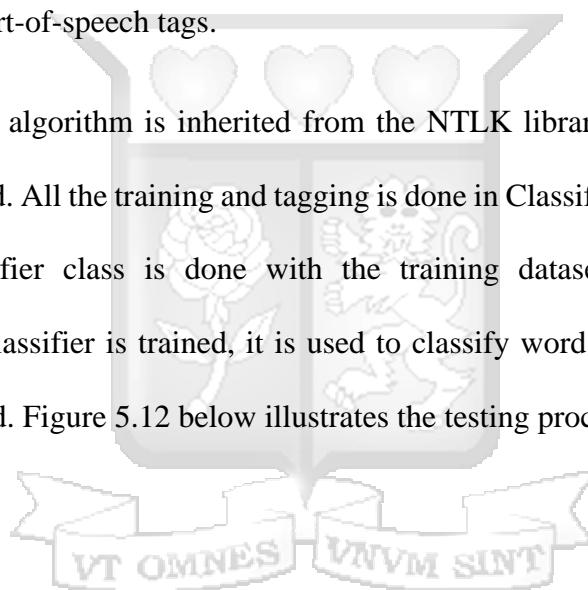
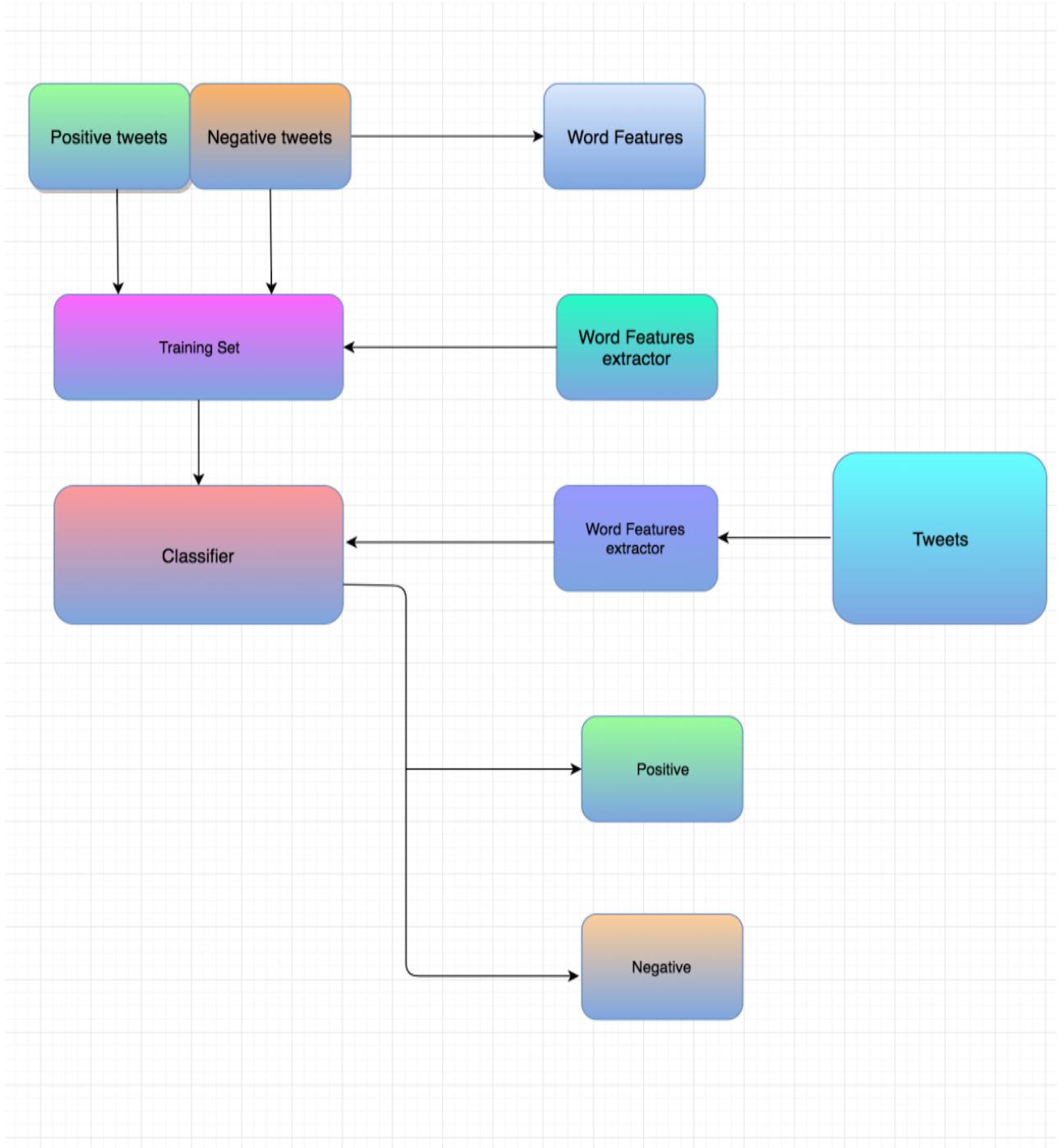


Figure 5.12: Testing Process



The quality and accuracy of the classifier depends on the amount of training sets it is fed.

If adequate training datasets is not supplied, the classifier will not be able to tag Tweets correctly

during the analysis. It is therefore advised to provide large datasets as possible to the trainer module.

5.4 Validation

5.4.1 Manual Validation

A quick manual verification process was selected to check how the tool's classifier will perform once subjected to random Tweets. A Python script named `Manual_validation.py` in Figure 5.13 was used for this specific purpose. It imports the classifier and two Tweets (one positive and another negative) which are parsed by the classifier for analysis and the predictions returned and printed on the screen as shown in Figure 5.14 to prove the accuracy and function validation for the tool created.

Figure 5.13: Python Script Named `Manual_Validation.py`

```
1 from classifier import *
2
3 def classify_tweet(tweet):
4     return classifier.classify(extract_features(tweet.split()))
5
6
7 manual_test_positive_tweet = "I am so happy about the sacrifice that was made between President Kenyatta and Raila Odinga"
8 manual_test_negative_tweet = "Kikuyus are helped steal votes from us. We haven't pushed this away."
9
10 print "Sentiment analysis of the manual tested positive tweet is : %s" % classify_tweet(manual_test_positive_tweet)
11 print "Sentiment analysis of the manual tested negative tweet is : %s" % classify_tweet(manual_test_negative_tweet)
12 |
```

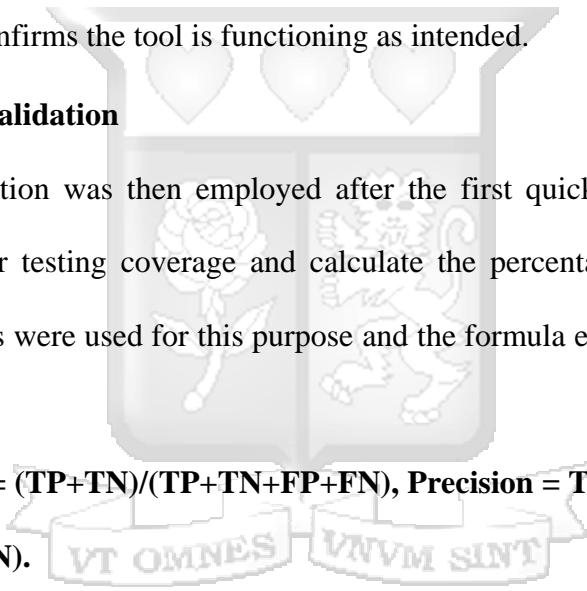
Figure 5.14: Predictions of Python Script in Figure 5.13

```
1 # coding=utf-8
2 # Author - B. Gufu
3 # @ilabafrika, Strathmore University
4
5 from classifier import *
6
7 def classify_tweet(tweet):
8     return classifier.classify(extract_features(tweet.split()))
9
10
11 manual_test_positive_tweet = "I am so happy about the sacrifice that was made between President Kenyatta and Raila Odinga"
12 manual_test_negative_tweet = "Kikuyus helped the crooks steal votes from us. We haven't pushed this away."
13
14 print "Sentiment analysis of the manual tested positive tweet is : %s" % classify_tweet(manual_test_positive_tweet)
15 print "Sentiment analysis of the manual tested negative tweet is : %s" % classify_tweet(manual_test_negative_tweet)
```

The tool can correctly classify the two subjected Tweets in their respective expected outcome. This quickly confirms the tool is functioning as intended.

5.4.2 Automated Validation

Automated validation was then employed after the first quick manual testing. This is important to add a wider testing coverage and calculate the percentage level of accuracy. A collection of 1000 Tweets were used for this purpose and the formula employed for the testing is as follows:


$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \text{ and Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

TP – True Positive, TN – True Negative, FP – False Positive, FN – False Negative.

TP refers to instances of hate speech text that were correctly identified as hate speech whereas TN are instances of non-hate speech text correctly predicted as non-hate speech. FP are the instances of non-hate speech text incorrectly classified as hate speech whereas FN are instances of hate speech text incorrectly determined to be non-hate speech (Feldman & Sanger, 2007).

A Python script that implements this formula was created and named as Automatic_validation.py then supplied with two files with test data (located in the folder test_data of the project) a collection of negative Tweets as negative.txt and a collection of positive Tweets

as positive.txt. Each of these test files had 500 Tweets resulting to a total of 1000 Tweets. The Table 5.1 below shows a summary of the automatic testing results.

Table 5.1: Summary of the Automatic Testing Results

Instance	Total
Total Number of Instances	1000
Total Instances of Positive Tweets	500
Total Instances of negative Tweets	500
Total Instance of Neutral Tweets	0
Percentage of instances of classified positive	34%
Percentage of instance classified negative	52%
Percentage of instance classified neutral	14%



CHAPTER SIX: DISCUSSION OF RESULTS

6.1 Introduction

This chapter discusses results achieved in the previous chapters. It clearly highlights the results obtained in this study and how the objectives were met.

6.2 Discussion

Obtained results greatly met the research objectives: to identify the challenges of National Cohesion and Integration Commission (NCIC) in the detection, monitoring and analysis of hate speech, review existing tools that are used in hate speech detection and to identify gaps that exists in these tools then design, develop, test and implement an automated tool for detecting analysing and monitoring hate speech in Kenya. Several challenges faced by NCIC and gaps in the present hate speech detection and reporting tools in use were identified during this study. Some of the identified challenges were: overwhelming of the tool with junk data, poor storage design, lack of real time hate speech detection and very limited number of Tweets classified in one attempt, high rate of false positives and the use of classifiers only trained for English contexts.

It was also discovered that hate speech detection and monitoring was mostly done manually with most developed tools like Umati were semi-automated. With this in mind the end result was that the detection and monitoring tools were prone to human errors since analysis was done separately by a different standalone tool or manually. Tools like Perspective API and Speech Blocker have a greater limitation since their designated users are the social media users who would be warned of the impact of their words to others. In addition to these, NCIC has previously failed to prosecute suspects of hate speech due to lack of proper digital evidence. The proposed and designed tool mitigated these identified challenges and gaps.

Several keywords were used which generated very accurate and precise results. Also, from the literature reviewed, it is worth to highlight that no publicly available training data is available for such a study presented herein. A collection of more than 100,000 Tweets were cleaned to create a training dataset that accurately captures out local problems.

The tools storage capacity proved to be great as Tweets storage occurred in several stages. Raw Tweets once collected were saved separately in raw_Tweets table as shown in Figure 5.5. After the cleaning process, cleaned Tweets were saved in cleaned-Tweets table and tagged Tweets after classification saved in analysed_Tweets table as shown in Figure 5.7 and 5.9 respectively. This data archiving is important for a couple of reasons, creating a historical collection of raw Tweets which can be shared by other researchers in the academia and for the purpose of proof in case another researcher would like to improve the tool created or recreate the logical steps implemented in the tool.

The tool developed collects publicly available live Tweets via the Twitter Streaming API. Collected Tweets are cleaned using regular expression by stripping out unnecessary information and replacing user personal identifiable information with generic text strings. Cleaned Tweets makes the classification process more efficient. Note that all the processes after the user entered the keywords were automated thus meeting the tools main objective. The tool also facilitated further monitoring of given keywords till the user choose to terminate the process.

The developed tool portrayed ease of use, timely analysis and real-time hate speech detection and analysis. The tool's post collection, preprocessing, analysis and classification is done using a Naïve Bayes algorithm adaptable from the NLTK library which performs the sentimental analysis. All these processes where done with minimal time, to run and generate a report for keywords entered by the user. The result was generated and stored in the database.

CHAPTER SEVEN: CONCLUSIONS AND RECOMMENDATIONS

7.1 Conclusions

The purpose of this research was to develop an automated tool to detect, analyse and monitor hate speech on Kenyan social media space using a machine learning technique. Relevant literature review was carried out in order to elaborate on foundational principles of opinion mining. In addition, interviews and questionnaires were carried out facilitating the same. Five NCIC staff were interviewed. The scope of this research is Kenya and Twitter as the social media platform for the case study. Tools relevant to hate speech detection, analysis and monitoring like Umati and Uchaguzi Online Monitoring tools were also reviewed.

The final developed tool was able to cover all the identified gaps discussed in this research. Python programming language was used to conceptualize the algorithms and system design models developed in this research. The text classification algorithm developed used was Naïve Bayes approach with the assistance of NTLK Python library. Data cleaning algorithm developed used text processing techniques such as regular expression to eliminate and strip off unwanted data.

Forensically sound processes were adopted in designing the tool, the collection and preservation of the Twitter data was done in consideration of admissibility of digital evidence in a court of law. Presenting only the processed Tweets in a court of law as evidence, is not admissible. The original raw Tweets should be available if further analysis is requested and even if the Tweet in question is deleted.

7.2 Recommendations

The accuracy of the tool is impressive, however, for a wider coverage of language variation used by Kenyans on Twitter, more custom training data sets need to be developed. A lot of data in corpus means a larger set features to be extracted and be used for feature vectors while processing live Tweets. More computing power will be needed when analysing larger volumes of Tweets. An operational decision should be made on how long to archive the collected Tweets. Due to the redundancy required for forensic design, too much data will accumulate in a short time.

For the benefit of the academia, this tool also provides data sets that can be shared with other researchers working on sentiment analysis and opinion mining for hate speech detection in Kenya. The developed classifier can be reused for similar routines in other social media platforms as long as the connection API along with the data formats are identified, data is cleaned and ready for analysis.

7.3 Future Work

Hate speech in Kenya can be expressed on Social media in more than one language since not all Tweets are in English. Social media users have taken to extensive use of their own coded language which posed a challenge during testing. I would recommend a self-learning tool that would discover new word use, its mode of use and probable meaning. Once this is obtained, the tool could present this to the user for review before use in Tweets classification. As the trend is, most people nowadays do not always use the actual words but shortened version of the word or user generated acronyms.

For a more comprehensive detection tool media analysis accompany the text data should be subjected to analysis. Often, most Tweets also contain captioned images and GIF videos which may also contain expressions of hate speech.



REFERENCES

- Adedoyin-Olowe, M., Gaber, M. M., & Stahl, F. (2013). A survey of data mining techniques for social media analysis. *arXiv preprint arXiv:1312.4617*.
- Akl, A. (2016). Hate speech plugin gives internet trolls a chance to pause. Retrieved from <https://blogs.voanews.com/techtonics/2016/10/21/hate-speech-plugin-gives-internet-trolls-a-chance-to-pause/>
- Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *International Language Resources and Evaluation*, (pp. 2200-2204).
- Bai, J., Nie, J.-Y., & Paradis, F. (2004). Using Language Tools for Text Classification. *Asia Information Retrieval Symposium. Montreal*.
- Bifet, A., & Frank, E. (2010, October). Sentiment knowledge discovery in Twitter streaming data. In *International conference on discovery science* (pp. 1-15). Springer, Berlin, Heidelberg.
- Borzi, V., Faro, S., Pavone, A., & Sansone, S. (2015). Prior Polarity Lexical Resources for the Italian Language. *arXiv preprint arXiv:1507.00133*.
- Chan, J. (2012). Uchaguzi: A Case Study. Harvard Humanitarian Initiative and Knight Foundation. Retrieved from http://www.knightfoundation.org/media/uploads/media_pdfs/uchaguzi-121024131001-phpapp02.pdf
- Chauhan, S., & Panda, N. K. (2015). *Hacking Web Intelligence: Open Source Intelligence and Web Reconnaissance Concepts and Techniques*. Syngress.
- Communications Authority of Kenya. (2013). *Kenya Information and Communication Amendment Act 2013* [Ebook]. Nairobi. Retrieved from <http://www.ca.go.ke/index.php/sector-legislation>
- Communications Authority of Kenya. (2017). *Sector Statistics Report Q1 2017/2018*[Ebook]. Nairobi. Retrieved from <http://www.ca.go.ke/images/downloads/STATISTICS/Sector%20Statistics%20Report%20Q1%202017-18.pdf>
- Die. (2018). Whiptail(1) – Linux man page. Retrieved from <https://linux.die.net/man/1/whiptail>
- Deng, L., & Wiebe, J. (2015). Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic tools. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 179-189).
- Deshpande, P. M., Atreyee, D. E. Y., Joshi, S., & Xing, S. (2018). U.S. Patent No. 9,886,711. Washington, DC: U.S. Patent and Trademark Office.
- Dezyre. (2016). Top-10-machine-learning-algorithms, 202. Retrieved from www.dezyre.com/article/
- Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.
- Greer, D., Hamon, Y. (2011). Agile software development. Software: Practice and Experience, 41(9), 943-944. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/spe.1100/full>
- Gross, J., & Suttor, N. (2013). Getting Found-Using social media to build your research profile.
- Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine Learning-Based Sentiment Analysis for Twitter Accounts. *Mathematical and Computational Applications*, 23(1), 11.

- Hatzivassiloglou, V., Wiebe, J. (2000). Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In: *Proceedings of the 18th International Conference on Computational Linguistics*, New Brunswick, NJ
- Hitz, L., Blackburn. B. (2017). *The State of Social Marketing 2017 Annual Report*. Retrieved from https://get.simplymeasured.com/rs/135-YGJ-288/images/SM_StateOfSocial-2017.pdf
- Joachims, T. (1998). Text Categorization with Support Vector Machine: Learning with Many Relevant Features. *Proceedings of the 10th European Conference on Machine Learning* (pp. 137-142). London: Springer-Verlag.
- Kaplan, A. M., & Haenlein, M. (2011). The early bird catches the news: Nine things you should know about micro-blogging. *Business horizons*, 54(2).
- Kennedy, A., Inkpen, D. (2016). Sentiment Classification of Movie and Product Reviews Using Contextual Valence Shifters, Computational Intelligence.
- Kinnunen, T. (2017). Hate speech detection. Retrieved from <https://futurice.com/blog/hate-speech-detection>
- Lichterman, J. (2017). This tool from Google parent Alphabet tries to tackle “toxic” comments through machine learning. Retrieved from <http://www.niemanlab.org/2017/02/this-tool-from-google-parent-alphabet-tries-to-tackle-toxic-comments-through-machine-learning/>
- Liombart, R. Ò., & Duran, C. J. (2017). Using machine learning techniques for sentiment analysis.
- Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1), 1-167.
- Maloba, W. (2013). *Use of regular expressions for multilingual detection of hate speech in Kenya*. (Published MSc. thesis) MMTI Theses and Dissertations (2013). (2198)
- Martini, B., Do, Q., & Raymond Choo, K. K. (2016). Digital forensics in the cloud era: The decline of passwords and the need for legal reform. *Trends & Issues in Crime & Criminal Justice*, (512).
- Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55-60).
- Morales D. R. (2018). Django-comments-xtd Documentation. Release 2.0.1. Retrieved from <https://media.readthedocs.org/pdf/django-comments-xtd/latest/django-comments-xtd.pdf>
- Mutahi, P., & Kimari, B. (2017). *The Impact of Social Media and Digital Technology on Electoral Violence in Kenya*. IDS.
- National Council for Law Reporting. (2008). National Cohesion and Integration Act. Nairobi.
- Ogada, K., Mwangi, W., & Cheruiyot, W. (2015). N-gram Based Text Categorization Method for Improved Data Mining. *Journal of Information Engineering and Applications*, 5(8), 35-43
- Omenya, R. (2013) *Uchaguzi Kenya 2013: Monitoring & Evaluation*. iHub Research and HIVOS. Retrieved from http://www.ihub.co.ke/ihubresearch/jb_UchaguziMEFinalReportpdf2013-7-5-14-24-09.pdf
- Peng, F. (2003). Augmenting Naïve Bayes Classifiers with Statistical. *University of Massachusetts, Computer Science Department Faculty Publication Series*.
- Published: Ondingi O Nyambane ‘prosecuting hate speech in Kenya’ published dissertation, University of Nairobi, 2012 Ondingi O Nyambane.
- Restricted use of Twitter APIs. (2018). Retrieved from <https://developer.Twitter.com/en/developer-terms/more-on-restricted-use-cases>

- Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures*, 12(2-3), 140-157.
- Sambuli, N., Morara, F., & Mahihu, C. (2013). *Monitoring Online Dangerous Speech in Kenya*. Nairobi: Umati.
- Satapathy, S. C., Govardhan, A., Raju, K. S., & Mandal, J. K. (Eds.). (2014). Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) (Vol. 1). Springer.
- Silva, L. A., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016, March). Analysing the Targets of Hate in Online Social Media. In ICWSM (pp. 687-690).
- Skyrme, D. (2007). Knowledge networking: Creating the collaborative enterprise. Routledge.
- Social@Ogilvy. (2015). *Social Media in Africa*. Retrieved from https://social.ogilvy.com/wp-content/uploads/Social-Media-in-Africa_Infographic.pdf
- Sriram, B., Fuhr, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010, July). Short text classification in Twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 841-842). ACM.
- Standard. (2016, December 05). *How app assists parents manage child's phone*. Retrieved from <https://www.standardmedia.co.ke/business/article/2000225838/how-app-assists-parents-manage-child-s-phone>
- State cracks down on 176 social media accounts over hate speech. (2017, July 28). Retrieved from https://www.the-star.co.ke/news/2017/07/28/state-cracks-down-on-176-social-media-accounts-over-hate-speech_c1606015
- Strachan, A. L. (2014). Interventions to counter hate speech. GSDRC Applied Research Services, 23.
- Taylor, M., Haggerty, J., Gresty, D., Almond, P., & Berry, T. (2014). Forensic investigation of social networking applications. *Network Security*, 2014(11), 9-16.
- Thakkar, H., and Patel D. (2015) Approaches for Sentiment Analysis on Social media: A State-of-Art study.
- The Bloggers Association of Kenya (BAKE). (2015). *The State of Blogging & Social Media in Kenya 2015 Report* [Ebook]. Nairobi. Retrieved from <http://www.monitor.co.ke/wp-content/uploads/2015/06/The-State-of-Blogging-and-Social-Media-in-Kenya-2015-report.pdf>
- The Stages of the Agile Software Development Life Cycle. (2017, December 01). Retrieved March 14, 2018, Retrieved from <https://www.lucidchart.com/blog/agile-software-development-life-cycle>
- Turney, P. D., (2002) "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews classification of reviews", *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 417–424.
- Understanding the Agile Software Development Lifecycle and Process Workflow. (2017, October 19). Retrieved March 14, 2018, Retrieved from <https://www.smartsheet.com/understanding-agile-software-development-lifecycle-and-process-workflow>
- Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012, July). A system for real-time Twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 115-120). Association for Computational Linguistics.

- Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language resources and evaluation, 39(2-3), 165-210.
- Yahyaoui, M. (2001). Toward an Arabic web page classifier, Master project. AUI.
- Vaishnavi, V., & Kuechler, W. (2004). Design research in information systems.



APPENDIX A: Interview Guide

STRATHMORE UNIVERSITY FACULTY OF INFORMATION TECHNOLOGY MASTER OF SCIENCE IN INFORMATION SYSTEMS SECURITY

Research Questionnaire

I am a graduate student at the Strathmore University, Faculty of Information Technology.

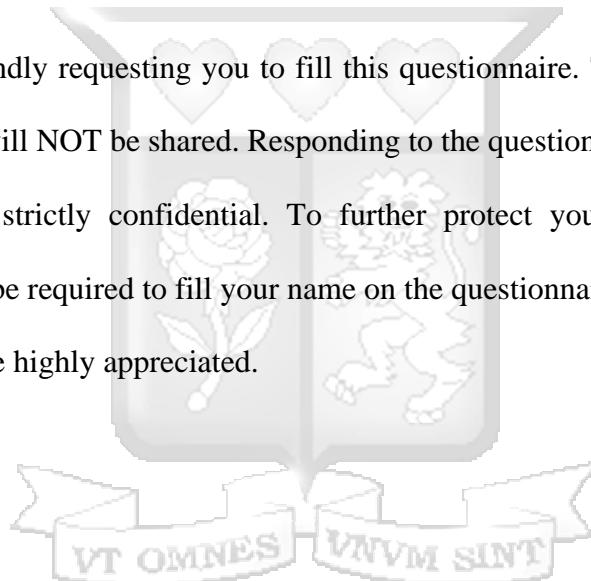
I am conducting a research in partial fulfilment of a Masters in Information System Security (MISS). My research aims at developing an Open Source Intelligence Gathering tool for Hate Speech in Kenya.

I am therefore kindly requesting you to fill this questionnaire. This survey is strictly for academic purposes and will NOT be shared. Responding to the questionnaire is voluntary and the responses will be kept strictly confidential. To further protect your opinions and enhance anonymity, you will not be required to fill your name on the questionnaire.

Your co-operation will be highly appreciated.

Yours Faithfully:

.....
Banchale A. Gufu



Date.....

Questionnaire NO.....

The following interview guide was used to in a personal interview with staff members of the NCIC to find out the challenges faced in monitoring hate speech on social media.

Questions

1. Do you have any systems in place to help monitor hate speech on social media?

.....
.....
.....
.....

2. Do you currently monitor hate speech on social media?

- a. If yes, how do you monitor hate speech on social media?

.....
.....
.....
.....

- b. If not, why?

.....
.....
.....

3. Which social media sites do you monitor?

1.
2.

3.

4.

others.....

3. How often do you monitor the social media sites for hate speech?

.....

4. What challenges do you face while monitoring hate speech on social media?



5. Do you use automated tools to detect, analyse and monitor hate speech on social media?



6. Have you identified any gaps in the tools you use for hate speech detection, analysis and monitoring?

a. If yes, which are they?

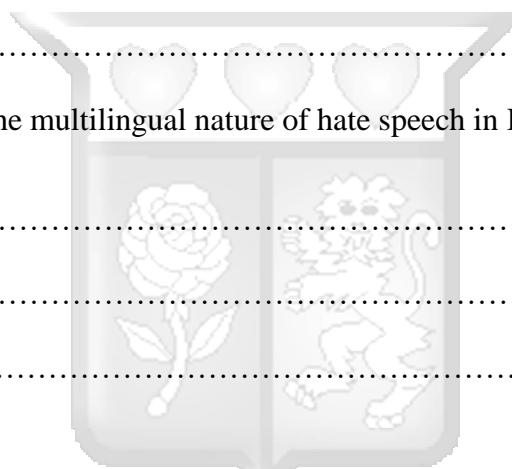
.....
.....
.....
.....

7. How do you analyse collected data from social media?

.....
.....
.....

8. What challenges do you face while analysing collected data from social media?

.....
.....
.....



9. How do you deal with the multilingual nature of hate speech in Kenya on social media?

.....
.....
.....

10. Which are the most frequent terms found in hate speech text?

.....
.....
.....

11. Which organisation(s) do you collaborate with in the detection, analysis and monitoring of hate speech on social media?

.....
.....

APPENDIX B: Python Program

a) General Python Source Code followed by the four processes as shown in the GUI

```
99 function main(){
100
101 OPTIONS=$(whiptail --title "OSINT GATHERING FOR HATE SPEECH" --menu "Select Option.. " 20 60 8 \
102 "1" "Configure Twitter API Keys" \
103 "2" "Collect Tweets" \
104 "3" "Clean Tweets" \
105 "4" "Analyse Tweets" \
106 "5" "Exit" 3>&1 1>&2 2>&3)
107
108 exitstatus=$?
109 exception_handler $exitstatus "[!] Failed To Start Menu"
110
111 if [ $OPTIONS = 1 ]; then
112     API_CONFIG
113
114     exitstatus=$?
115     exception_handler $exitstatus "[!] Failed To Configure KEYS"
116
117 elif [ $OPTIONS = 4 ]; then
118     ANALYZE_TWEETS
119
120     exitstatus=$?
121     exception_handler $exitstatus "[!] Failed To Analyze Tweets"
122
123 elif [ $OPTIONS = 2 ]; then
124     COLLECT_TWEETS
125
126     exitstatus=$?
127     exception_handler $exitstatus "[!] Failed To Collect Tweets"
128
129
130 }
```



b) Source Code of Twitter Credentials

```
1 ## Twitter credentials
2
3 consumer_key = 'PsnaQDBsgjBGF9eLR3LsCxhm2'
4 consumer_secret = 'uJwtkFHH0kyybRJWi0r0Z1tYRxH0QIBfbnbZ3RS0FSKdeREiT'
5 access_token = '263544909-twl0h6AE9yGDTxb18JHmgumRKU2xJUsEIaReiWHn'
6 access_token_secret = 'T52ZojWaTlYr4w9azHRJVA89Bf7dRumTd47dj8WGvNfnd'
7
8
```

c) Twitter Authentication and Twitter Streaming API Connection Source Code

```
#This handles Twitter authentication and the connection to Twitter Streaming API
auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
```

d) Tweets Collection Source Code

```
4
5 from tweepy.streaming import StreamListener
6 import tweepy.streaming
7 from tweepy import OAuthHandler
8 from tweepy import Stream
9
10 from api_config import *
11 import sqlite3
12
13 class CustomStreamListener(StreamListener):
14
15     def on_status(self, status):
16         db_file = 'Database/tweets.db'
17         con = sqlite3.connect(db_file)
18         #con.text_factory = str
19         con.text_factory = lambda x: unicode(x, 'utf-8', 'ignore')
20         cur = con.cursor()
21
22         tweet_date = status.created_at
23         try:
24             text = status.extended_tweet["full_text"]
25         except AttributeError:
26             text = status.text
27         tweet_text = text.encode('utf-8').translate(None, '!.,?')
28
29         print "%s\t%s" % (tweet_date, tweet_text)
30         cur.execute("INSERT INTO raw_tweets(date, tweets) VALUES (?, ?)", (tweet_date, tweet_text))
31         con.commit()
32         con.close()
33
34 #This handles Twitter authentication and the connection to Twitter Streaming API
35 auth = OAuthHandler(consumer_key, consumer_secret)
36 auth.set_access_token(access_token, access_token_secret)
37
38 #This line filter Twitter Streams to capture data by the keywords
39 streaming_api = tweepy.streaming.Stream(auth, CustomStreamListener(),
40                                         timeout=60, tweet_mode='extended')
41 streaming_api.filter(track=['alshabab', 'mjinga', 'jeuri', 'kikuyu', 'jaluo',
42                           'handcheque', 'raila', 'thererealaila', 'uhuru'].asvnc=True)
```

e) Tweets Cleaning Code

```
1 import re
2 import sqlite3 as lite
3
4 db_file = 'Database/tweets.db'
5 con = lite.connect(db_file)
6 con.text_factory = lambda x: unicode(x, 'utf-8', 'ignore')
7 cur = con.cursor()
8
9 def export(filename, data, p):
10    with open(filename, p) as output:
11        for line in data:
12            output.write(line)
13
14 def cleanTweets(tweet):
15    cleantweet = re.sub('((www\.[^\s]+)|(https?://[^\s]+))','URL',cleantweet)
16    cleantweet = re.sub('@[^\s]+','AT_USER',cleantweet)
17    cleantweet = re.sub('[\s]+', ' ', cleantweet)
18    cleantweet = re.sub(r'#[^\s]+', r'\1', cleantweet)
19    cleantweet = cleantweet.strip('\"')
20    cleantweet = cleantweet.lower()
21    return cleantweet
22
23 all_cleaned_tweets = []
24
25 for row in cur.execute('SELECT * FROM raw_tweets;'):
26    twit = row[2]
27    clean_tweet = cleanTweets(twit)
28
29    print(row[0], row[1], clean_tweet)
30
31    tweet = (row[0], row[1], clean_tweet)
32    all_cleaned_tweets.append(tweet)
33
34 for item in all_cleaned_tweets:
35    cur.execute("INSERT INTO cleaned_tweets(id, date, clean_tweet) VALUES (?,?,?)", (item[0], item[1], item[2]))
36    con.commit()
37
38 con.close()
```

f) Tweets Preprocessing Source Code

```
1 from classifier import *
2 import sqlite3 as lite
3
4 posDB = 'test_data/positive_test.txt'
5 negDB = 'test_data/negative_test.txt'
6 posScore = []
7 negScore = []
8 neuScore = []
9
10 def plolarityCount(ourscore):
11     if ourscore > 0:
12         print "positive tweet : " + str(ourscore)
13         posScore.append(ourscore)
14     elif ourscore < 0:
15         print "Negative tweet : " + str(score)
16         negScore.append(ourscore)
17     else:
18         print "Neautral : " + str(score)
19         neuScore.append(ourscore)
20
21 def processTweet(tweet):
22     tweet = re.sub('((www\.[^\s]+)|(https?://[^\s]+))','URL',tweet)
23     tweet = re.sub('@[^\s]+','AT_USER',tweet)
24     tweet = re.sub('[\s]+', ' ', tweet)
25     tweet = re.sub(r'#[^\s]+', r'\1', tweet)
26     tweet = tweet.strip('\'')
27     tweet = tweet.lower()
28     return tweet
29
```

VT OMNEPSU UNIVM SINI

g) Tweets Classification Code

```
1 from classifier import *
2 import sqlite3 as lite
3
4 db_file = 'Database/tweets.db'
5 con = lite.connect(db_file)
6 con.text_factory = lambda x: unicode(x, 'utf-8', 'ignore')
7 cur = con.cursor()
8
9
10 def classify_tweet(tweet):
11     return classifier.classify(extract_features(tweet.split()))
12
13 all_analysed_tweets = []
14
15 for row in cur.execute('SELECT * FROM cleaned_tweets;'):
16     twit = row[2]
17     sentiment = classify_tweet(twit)
18
19     if (sentiment != "neg") and (sentiment != "pos"):
20         sentiment = "neutral"
21
22     polarised_tweet = (row[0], row[1], row[2], sentiment)
23     all_analysed_tweets.append(polarised_tweet)
24
25
26 for item in all_analysed_tweets:
27     print item
28     cur.execute("INSERT INTO analysed_tweets(id, date, analysed_tweet, polarity) VALUES (?,?,?,?,?)",
29                 (item[0], item[1], item[2], item[3]))
30     con.commit()
31
32
33 con.close()
```

h) Training Data

```
1 # coding=utf-8
2 # Author - B. Gufu
3 # @ilabafrica, Strathmore University
4
5 import os
6
7 def getTrainData():
8     positives, negatives, traindata = [], [], []
9     for filename in os.listdir("training_set"):
10         if filename == "positive_tweets.txt":
11             with open('training_set/'+filename) as f:
12                 positives = [(tweet, 'pos') for tweet in f.readlines()]
13         if filename == "negative_tweets.txt":
14             with open('training_set/'+filename) as f:
15                 negatives = [(tweet, 'neg') for tweet in f.readlines()]
16
17     for (words, sentiment) in negatives + positives:
18         words_filtered = [e for e in words.split() if len(e) > 2]
19         traindata.append((words_filtered, sentiment))
20
21
22 return traindata
```

i) Analyses Source Code

i. Part 1

```
1 from classifier import *
2 import sqlite3 as lite
3
4 posDB = 'test_data/positive_test.txt'
5 negDB = 'test_data/negative_test.txt'
6 posScore = []
7 negScore = []
8 neuScore = []
9
10 def polarityCount(ourscore):
11     if ourscore > 0:
12         print "positive tweet : " + str(ourscore)
13         posScore.append(ourscore)
14     elif ourscore < 0:
15         print "Negative tweet : " + str(score)
16         negScore.append(ourscore)
17     else:
18         print "Neautral : " + str(score)
19         neuScore.append(ourscore)
20
21 def processTweet(tweet):
22     tweet = re.sub('((www\.[^\s]+)|(https?://[^\s]+))', 'URL', tweet)
23     tweet = re.sub('@[^\s]+', 'AT_USER', tweet)
24     tweet = re.sub('[\s]+', ' ', tweet)
25     tweet = re.sub(r'#[^\s]+', r'\1', tweet)
26     tweet = tweet.strip('\'')
27     tweet = tweet.lower()
28     return tweet
29
```



ii. Part 2

```
21 def processTweet(tweet):
22     tweet = re.sub('((www\.[^\s]+)|(https?://[^\s]+))', 'URL', tweet)
23     tweet = re.sub('@[^\s]+', 'AT_USER', tweet)
24     tweet = re.sub('[\s]+', ' ', tweet)
25     tweet = re.sub(r'#[^\s]+', r'\1', tweet)
26     tweet = tweet.strip('\'')
27     tweet = tweet.lower()
28     return tweet
29
30 def classify_tweet(tweet):
31     return classifier.classify(extract_features(tweet.split()))
32
33 pos = open(posDB, 'r')
34 pos = pos.read()
35 negs = open(negDB, 'r')
36 negs = negs.read()
37
38 for line in pos:
39     ans = classify_tweet(line)
40     polarityCount(ans)
41
42 for line in negs:
43     ans = classify_tweet(line)
44     polarityCount(ans)
```