UNIVERSITY OF NAIROBI

COLLEGE OF BIOLOGICAL AND PHYSICAL SCIENCES

SCHOOL OF COMPUTING AND INFORMATICS

SOCIAL MEDIA FORENSICS FOR HATE SPEECH OPINION MINING

BY
WAFULA WANJALA GEORGE
P53/79539/2015

SUPERVISOR: DR ANDREW KAHONGE

October, 2016

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Distributed Computing

DECLARATION

The project, as presented in this document, is my original work and has not been presented for any other university award.

Signature: ____________________ Date: ______________________________

Wafula Wanjala George

P53/79538/2015

The Project has been submitted in partial fulfillment of the Requirements for the Degree of Master of Science in Computer Science at the University of Nairobi with my approval as the University Supervisor.

Signature: ________________________ Date: _____________________________

Dr. Andrew Kahonge

School of Computing and Informatics

ACKNOWLEDGEMENT

I want to thank God for granting me with the ability to complete this project. My lovely wife Florida and children Elianah, Eldad and Victoria, thank you for giving me time to be able to finish and work on this project. To my mum, your continual instruction and direction is unrelenting.

ABSTRACT

Social Media Hate Speech has continued to grow both locally and globally due to the increase of Online Social Media web forums like Facebook, Twitter and blogging. This has been propelled even further by smartphones and mobile data penetration locally. Global and Local terrorism has posed a vital question for technologists to investigate, prosecute, predict and prevent Social Media Hate Speech.

This study provides a social media digital forensics tool through the design, development and implementation of a software application. The objective of this study is to identify and analyze the different software techniques for hate speech monitoring and provide the best suited and customized application. The study will develop an application using Linux Apache MySQL PHP and Python. The application will use Scrapy Python page ranking algorithm to perform web crawling and the data will be placed in a MySQL database for data mining.

The application used Agile Software development methodology with twenty websites being the subject of interest. The websites will be the sample size to demonstrate how the application works together with the Python libraries as the framework for web crawling. MySQL data mining, database query application models will be used in performing the search of the lexicon of keywords for hate speech, Inferences from the data mined from crawled web pages will be drawn using Microsoft Excel 2016. Excel will be used for data analysis with the data being presented in tables and figures.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

1.1 Background

The Internet is continually providing the medium for communication, fundraising, training, media operations and recruitment. Web forums are online Internet websites where the owners of the webpage can post stories, ideologies, blogs; which in turn undergoes discussions from the community members and participants. Web forums such as Skype, Gtalk, WhatsApp, Twitter and Facebook have become valuable arenas for social interaction, discussion and communication online on the Internet.

Internet accessibility in Kenyan has continued to penetrate and grow in numbers, with 52% of the total population of 42 Million citizens being connected. 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 and mobile networks. Currently, Kenya's mobile tele density is placed at 77% facilitated by three mobile service providers, that is, Safaricom, Airtel and Orange (Nanfuka, 2014).

Obar (2015) defined social media as computer-mediated tools for creating, sharing and exchanging of information, photos, videos and career interests in the virtual communities and networks. Social Media services have been defined as Web 2.0 applications and web forums that emphasize on online generated content and interoperability (Darcy, 1999). This has led to web forums, blogs, wiki and media-sharing websites that grant online users the capability to communicate and respond to opinions on daily news, political debates, tribal and religion discussions.

Sutton (2009) formulates social media as diverse and developing online communication that allows the production and sharing of information (Sutton, 2009). Crisis communication, public relations, radicalization and hate speech has been acknowledged as the roles by which social media is aiding online communication. Social media forums such as Twitter have played a daily constant role in issuing daily political news, traffic updates, emergency warnings and alerts, supporting recovering efforts, situational awareness and real-time information.

Social Media, web forums and online conversations and debates have brought to the fore hate speech. The hate speech is due to the ethnic tensions which have polarized the country with each passing political season. Hate speech was in the limelight during the 2007 post-election results, as more citizens were using phone short messages to fuel, share destructive propaganda and support ethnic based violence of the different tribal communities depending on where you were aligned. ecame most apparent during the 2007 post-election period during which short message service (SMS) fueled political conflict and ethnic based violence. The number of reported causalities were not more than 1,200 and aleast 500,000 people being internally displaced. Current studies have shown the movement of hate speech from traditional person to person phone message in to social media platforms such as Facebook and Twitter. There was a swift change of political bickering and political tension from the physical to the digital cyber space. This was fueled much by the almost instant delivery of propaganda, agenda and tribal ethnic through social media forums such as Twitter, WhatspApp and Facebook.

The 2013 presidential debates and elections saw the maximization of social media for campaign purposes, voter communication and civic engagement. This in turn saw the resurgence of online hate speech. This resulted in the government enacting laws to control hate speech both online and offline through the Cybercrime and Computer Related Crimes Bill and the National Cohesion and Integration Act of 2008.

Operation Linda Nchi, a codename for the joint coordinated military operation of the Kenyan and Somali armies began on 16 October 2011 (Kron, 2013) in Somalia. This is after it had been believed that Al-Shaabab militants had kidnapped several foreign aid workers and tourists visiting the country in Lamu, a coastal town. Web forums have been used by radical groups for communication, sharing of information and their ideologies to the public. Such web forums form part of the Dark Web.

The Dark Web is made up of web forums, websites, chat groups and social media platforms that are used by terrorists, hate mongers, radicals, political, ethnic and extremist groups (Chen, 2006). The war on terrorism has been both a physical, psychological and virtual battle, with extremists using web forums to spread and influence their propaganda.

Hate speech is characterized by any form of verbal or non-verbal attack targeting a specific group of people. It can be motivated by racial, ethnic, gender, religion or sexual orientation. Hate speech is usually communicated through different media such as Internet, hand-held devices, newspapers, ,magazines, television, radio broadcasts, verbal person-to-person.

Hate Speech is defined by the National Cohesion and Integration 2008 Act as “speech that is threatening, abusive or insulting or involves the use of threatening, abusive or insulting words”

with the intention to stir up ethnic hatred or a likelihood that ethnic hatred will be stirred up. Hate Speech can be ethical or also radical when used by extremist on the basis of religion for recruiting people in terrorist acts. Hate speech is continuously evolving to the digital space and its being embodied on the social media through ethnical and radical groups and classification.
Social Media Forensics for Hate Speech Opinion Mining relates to the scientific application of cyber forensics tools to social media web forums in order to extract, identify and document hate speech.

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. The Law Enforcement and Communication authorities need to setup the war on the vice so as to avoid future problems. Social Media and Crime has shown a clear relationship that someone knew of the possible crime or planning of the crime. Wilson et. al (2012) argued that “Facebook and other social media platforms need to be positioned firmly within the criminologist’s gaze not only because of the wealth of data these platforms provide but also due to their significant popularity across all age groups and their influence on how people interact and communicate with one another. This is because social media is now intimately interacting with many of the issues that we as criminologists can and should be engaging with”.

The amount of big data that social media generates is immense, instant, dynamic and constantly evolving. The current law enforcement agencies and communication authorities are stretched and the need to provide an automated system is paramount.

The movement of extremist group towards the use of social media for recruiting Al-Shaabab militants is also of grave concern, as the recruiters are aware that there are no checks on the social media and cyber space. Currently, the security agencies depend on public to notify them of any hate speech comments on the Internet. The cyber hate mongers are quick enough to remove and de-register their social media accounts. Hence, the need of a solution which is more proactive and keeps a copy of the collected digital evidence.

1.2 Problem Statement

Hate Speech on social media needs to be tracked, tackled and means provided to apprehend the cyber criminals and hate speech mongers. The process of collecting and documenting online hate speech digital evidence should be optimized efficiently.

With security agencies and communication authority personnel being overstretched, with less manpower and tools, cyber expertise, the need for an automated and easier to use system for hate speech is crucial.

Using the Social Media platform, we will provide an application that will be able to mine social media opinions and easily present the results of possible hate speech crime together with the evidence and potential list of suspects. The system will invoke a web crawler that will collect all the web forum details and insert them in a database. Once the database is created, a script will be created to check the data against a list of potential keywords of hate speech. The script will search and provide the list of hate speech data together with the person of interest.

Communication authorities, security agencies and the country at large stands to benefit by placing checks on social media hate speech. The movement of hate speech mongers towards the digital cyber space needs to be addressed before it escalates further as was experienced regularly when political debates are held or government corruption cases are discussed.

1.3 Objectives of the Study

1.3.1 Overall Objective

The purpose of the study was to design, develop and implement a software application for hate speech monitoring and reporting.

1.3.2 Specific Objectives

i. To identify and analyze techniques used in hate speech monitoring and select the best suitable technique for creating a customized hate speech application.

ii. To develop an application that will combine hate speech keywords for data mining.
iii. To demonstrate and test the application while providing analysis on the hate speech websites being investigated.

1.4 Research Questions

The research answers the following:

i) How can Social Media web forums be monitored and forensic digital evidence collected?
ii) What are the digital investigative challenges that law enforcement are facing towards evidence carving?
iii) What are the methods used to collect the hate speech opinions and developing an algorithm?

1.5 Justification

Hate Speech on Social Media has been shown to directly influence and promote physical violent acts. Cyber criminals and hate mongers have been known to move from the cyber space to the actual physical world to promote, fund and finance violent crimes. Hence, the need for this monitoring tool to capture the digital evidence. The research proposal is a much needed approach to address the gap cyber criminals and hate mongers are taking advantage in the cyberspace arena. It's critical and a much needed approach to check the cyber space as such acts of hate speech are affecting and influencing different people, races, tribes and the entire country. The tool will assist law enforcement agencies to easily and readily make use of the features to capture data and digital evidence.

The algorithm developed will capture English, Swahili and Sheng hate speech keywords as means of increasing its relevance towards the local population and social media sites. With this tool in hand law enforcement and communication authorities will be on high alert and be able to bring down offensive social media web forums and thus reduce potential political, ethnical and tribal conflict. After all, prevention is better than cure when it comes to civil war and internal conflicts.

1.6 Assumptions and Limitations

The following assumptions and limitations are considered

(i) Its assumed that the cyber forensic investigator has access to the social media web forum or account or is a friend/follows the person of interest. The web forums are also thought to be open to the public and easily accessible on the Internet.

(ii) The potential risk of either a website or social media account being hacked is always present. With this in mind, it’s assumed that the social media web forums comments on a person of interest account or website are his/her publications and there was no hacking onto the user’s web forum or account.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

The continual pursuit of offenders in the cyber realm is an ever evolving practice and cyber forensics had been in the forefront. The need for newly released and better forensic tool is a constant desire any cyber forensics expert. Planning, Acquisition and reporting of cybercrime is more stretched when the digital evidence is not available locally on a hard disk drive or digital media, as the cyber forensics expert needs to rely on remote acquisition of hate speech digital evidence from social media web forums that are not hosted locally. Cyber hate mongers are always finding need ways to propagate their agenda with the use of more sophisticated IT tools and applications. Thus, cyber forensic experts have to come with more advanced tools, applications and systems that allow them to catch the hate speech mongers.

Rapport (2010) describes forensics for social media as any listening methods or solutions that allows technology and provides an algorithm for helping researchers and organizations to listen, collect data, document, analyse, interpret and respond to online conversations. This further collaborated by Branthwaite and Patterson (2011) who said “when compared to the traditional research approaches, the similarities that social media monitoring shares with quantitative research include large samples, numeric data and difficulty in assessing meanings, while among those it shares with qualitative approaches are the gathering of spontaneous views and opinions, and a need for rigorous semantic analyses”. Furthermore, social media monitoring can be thought to be a continuous daily process that offers insights, information, intelligence and communication for planners and research as found out by Hipperson (2010).

Traditionally, social media monitoring and quantitative approaches consist of sampling and standardization in big data research. Social Media data constitutes of what is known as big data. Branthwaite and Patterson (2011) argue that the research differences with qualitative approaches in big data samples is lack of direct contact, huge research targets due to huge data sets, differences in non-verbal cues, feedback, data collection and contextual information

Zailskaite-Jakste and Kuvykaite (2012) demonstrated social media monitoring as function of gauging, listening and analyzing the online environment. They went ahead to formulate that it may serve the purpose of evaluation, tracking the success of the hate speech message sent. Monitoring and evaluation of such web forums is paramount and critical for the general public, authorities and communication agencies. It through such initiative that hate speech can be mitigated if not eliminated.

Social Media monitoring can be considered as active methods of engaging citizens in online interaction, gathering online information and analyzing the data for useful purposes. A case in point is the example where health officials engaged online with a sample population to collect, gather and analyze online data through a survey on an e-health website (Laakso, Armstrong, & Usher, 2012). The research study undertaken for this project included social media monitoring of organic and interactive online conversations and communications on web forums.

2.1.1 Reasons for Monitoring

Monitoring is conducted to collect online data, analyze the data and make inferences on the analyzed data. Its from such research discussions that the general public, authorities and communication and regulatory authorities can be able to understand the online virtual environment that they are in (Zailskaite-Jakste & Kuvykaite, 2012). By understanding the environment it allows the general public to be educated and made aware of the dos and don’ts when it comes to hate speech communication. The authorities, communication and regulatory bodies will be able to come up with laws, rules and regulations for reporting and capturing the hate speech mongers. It allows a proper outline and structure for hate speech monitoring and reporting.

Kavanaugh et al. (2012) demonstrated the goal for social media monitoring is to allow a bird’s eye view and the phenomena of big picture evaluation and analysis. This is further collaborated by Bengston et. al (2009) when they conclude that monitoring allows a peep in the societal debate window and illuminates the stakeholder perceptions, wants and attitudes.

The driving force for monitoring may be related to public safety (Kavanaugh et al., 2012), product analysis (Deluca et al., 2012), political opinions, public reactions to policies (Sobkowicz et al., 2012), identification of radical opinions (Yang, Kiang, Ku, Chiu, &Li, 2011), hate speech, profiling (Keelan, Pavri, Balakrishnan, & Wilson, 2010) and falsified information (Campbell, Pitt, Parent, &Berthon, 2011).

2.1.2 Opinion mining Issues

Sentiment analysis research has continued to evolve as different approaches are taken towards to data accuracy and objectivity of the data collected. Social Media forms what is called Big Data and hence these huge data information and data sets offer a challenge when it comes to accuracy due to data outages and losses hence proving (Boyd & Crawford, 2012). Social media data is limited in archiving capability as such huge data sets require huge storage capability which is not possible for small time websites and web forums (Bruns and Liang, 2012). According to Boyd and Crawford (2012) researchers are more likely to study immediate past and the present data which will not provide the much needed data when it comes to drawing a pattern for hate speech and providing a cyber trail of the forensic artifact.

Sentiment Analysis and coming up with computer algorithms that automatically detect and classify hate speech has been flawed by the fact of contextual information. Contextual information has proven to be an issues mainly because when data is analysed and taken out of context it may lose or change their meaning (Boyd & Crawford, 2012). Furthermore, radicals and extremists are coming up with coded language that provides inaccurate information that promotes data inaccuracies (Lindsay, 2010).

Another concern when it comes to opinion mining is privacy of online social media accounts and data. In order to avoid social media accounts being hacked and tampered with, Social Media service providers are constantly checking their vulnerabilities and also notifying the account holders of any potential attempts of hacking. This has proved to be a bottleneck as social media accounts such as Facebook, Twitter and WhatsApp are coming with private ways of blocking access and making sure that the conversations are secure and less likely to be hacked (Boyd & Crawford, 2012).

2.2 Review of Existing Social Media Analytics

2.2.1 Social Media Analytics for Radical Opinion

Yang Ming et al 2011 formulated a procedure for radical opinion identification that can be divided into three steps, as shown in Fig.1. The first step is posting collection. A spidering program is developed to collect the online sources, that is, thread pages from the hate or extremist Web forums listed on watchdog organizations, such as the Anti-Defamation League (ADL) and Southern Poverty Law Center (SPLC). These watchdog organizations continuously update the lists of domestic and international extremist Web sites. Parsing programs are then developed to parse out the posting data from the raw webpages and store it in a relational database. Secondly, is the feature generation. Here, they used F1, F2, F3, and F4 to denote syntactic, stylistic, content-specific, and lexicon features, respectively. F1, F2, and F3 are natural machine learning approaches, and F4 is an approach that is semantic-oriented. The third step is model generation, where they compare the performances of three classifiers SVM, Naïve Bayes, and Adaboost. The best performing classifier was then used to classify new online messages. An iterative train-and-test process was needed to fine tune the model and, after the classification model was developed.

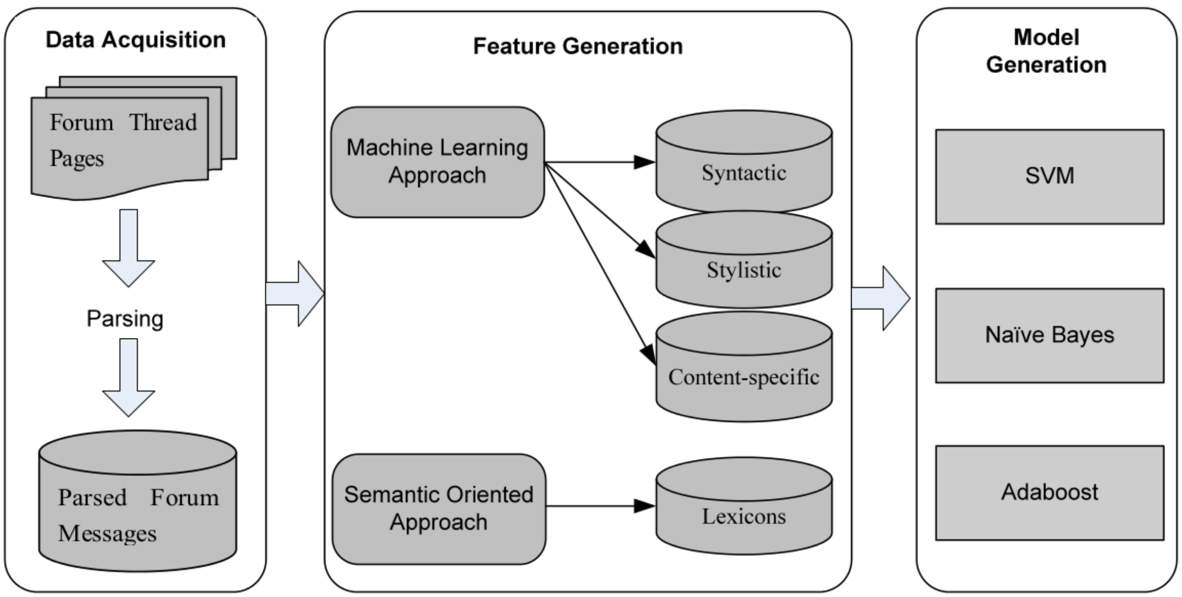

Figure 1: Radical Opinion Identification System

2.2.1.1 Datasets

In this research, online messages were collected from two hate group forums namely, www.nazi.org and www.resistance.org. The www.nazi.org is an English-language forum for an Aryan supremacist group that gained notoriety when a forum member was involved in a school shooting in 2004 (Abbasi et al. 2008a). The other website that propagates white, holistic and supremacist pro-Nazi Web forum. Using keywords such as kill, hit, stab, hang and lynch (Abbasi and Chen 2008b), they spidered 29,806 messages between September 2006 and April 2009.

Two domain experts were recruited to annotate the collected messages independently. The message annotation task performed by the two domain experts render Kappa (k) values of 0.90 for the www.nazi.org forum and 0.87 for the

www.resistance.org forum, suggesting sufficient inter-coder reliability. For each dataset, they manually selected 1,000 postings related to racial issues. In cross-validation the data sets are split into training, validation and testing data sets. Postings collected between September 2006 and August 2008 were divided and assigned a 70-30 split for cross-validation of the trained and tested data sets respectively. The training is normally used for building the classifiers, while the testing dataset was used to test the performance of the classifiers so as to fine tune their parameter settings.

2.2.1.2 Evaluation of Features Selected for Radical Opinion Identification

They conducted experiments to determine the appropriate text feature representation schema for radical opinion identification. The objective of the experiments was to determine the most effective feature sets to be implemented. They validate the effectiveness of the selected features via a two-prong approach: (1) gathering the frequency of each feature that appears and create a vector feature-set of the counts of all postings in the datasets; and (2) developing a discriminant function (using both the machine learning and semantic-oriented approaches) that uses the feature vectors to classify opinions (radical from non-radical).To extract F1, F2, and F3, a program was developed to search individual messages and compute the percentage of features that reflect the linguistic categories. The use of semantic-oriented approach (Dang 2009), specifically, the dictionary-based technique, to generated F4. They first performed Part-of-Speech (POS) tagging and then calculated the sentiment scores of the words by looking up sentiment-based lexicons, SentiWordNet (Denecke and Nejdl 2009; Esuli and Sebastiani 2006). The sentiment score is used to filter out less subjective words. Specifically, the use of midpoint of the 0-1 scale to differentiate and gauge the subjective words and objective words for a given POS. Words that have sentiment scores greater than 0.5 are kept as lexicon features. Table 1 lists some important lexicon features derived from the datasets. They started with a feature set (A) that contains only syntactic features (F1), the most general linguistic features, and gradually add features that are more specific. Stylistic features (F2) were added to form the second feature set (B), and content-specific features (F3) for the third feature set (C). Finally, lexicon features (F4) were added to form the fourth feature set (D). The feature sets evolve from generic linguistic features to domain-specific features. Three machine learning techniques, that is, SVM, Naïve Bayes (NB), and Adaboost, were used as classifiers for examining the performance of the different feature-sets. A 10-fold validation based on the training and testing dataset was utilized as means to compare the performances of the three classifiers.

Panel A: www.nazi.org		Panel B: www.resistance.org	
Word	Sentiment Score	Word	Sentiment Score
Hate	0.84	Kill	0.82
Beat	0.79	Lynch	0.77
Stab	0.71	Hang	0.69
Suck	0.68	Bomb	0.65

Table 1: Examples of important lexicon features in the selected feature set

Feature Set	Features	Panel A: www.nazi.org			Panel B: www.resistance.org		
		SVM	NB	Adaboost	SVM	NB	Adaboost
A	F1	0.711	0.493	0.521	0.694	0.457	0.507
B	F1+F2	0.804	0.574	0.599	0.781	0.546	0.579
C	F1+F2+F3	0.872	0.635	0.684	0.857	0.601	0.645
D	F1+F2+F3+F4	0.957	0.698	0.759	0.94	0.663	0.702

Table 2: F1 scores of the four different feature representation schemas

Feature Set	Features	Panel A: www.nazi.org			Panel B: www.resistance.org		
		SVM	NB	Adaboost	SVM	NB	Adaboost
A vs. B	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
B vs. C	< 0 .01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
C vs. D	< 0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01

Table 3:Paired t-tests of F1scores between the feature sets

Table 2 shows a comparison of the performance differences between the different feature sets for the same classifier. For both datasets, the F1 score improves with the addition of more feature types, and the SVM classifier consistently outperforms the other two classifiers regardless of the feature sets implemented. Paired t-tests are results were used to show statistical significance of the performance difference when additional features are added. As per the results, the added features significantly aid in the classifier's performance in all cases. Moreover, the lexicon features extracted by the semantic-oriented approach significantly improve the performance of radical opinion identification. The findings suggest that proper feature selection could significantly improve the effectiveness of the classifiers for radical opinion identification, and that both generic and domain-specific features of the data should be included. In addition, domain-specific features (i.e., content-specific features) could be utilized as a dictionary of key words to discriminate radical or violent contents.

2.2.2 Using SentiWordNet for Affect Analysis

Tawunrat Chalothorn and Jeremy Ellman in their study analyze an existing technique with a view to answer the effectiveness of SentiWordNet in detecting hate speech and emotions on the internet. Montada and Qawem web forums were chosen as both use the Arabic language and possess Islamic ideological content. The process involved collecting hate speech data, model building and data analysis of collected results. The data collection phase included takes the ripping words from the forums. 500 sentences of the ripped data are translated manually with Python programming language being used for model building. The phase for model building involves the splitting of sentences as words and reduction of the high-frequency text (stop words). The samples of high frequency texts can be found in Table 4. The bag of words (BOW) together with the part of speech (POS) tags were used, as shown in Table 3, for tagging and position alignment of each word in the sentence. Assigning of positive and negative scores of each synset in each word was done by SentiWordNet, WordNet and Lexicon.

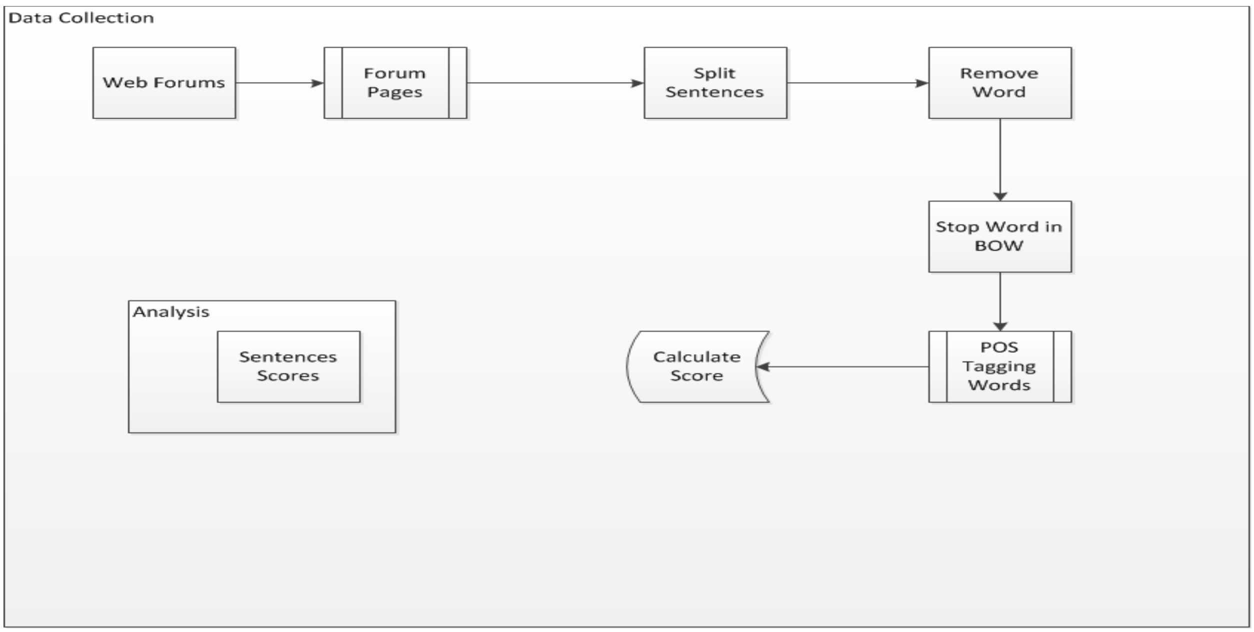

Figure 2: Framework of SentiWordNet

Stop Words
[‘I’,’me’,’my’,’myself’,’we’,’our’,’ourselves’,’you’,’your’,’yours’,’yourself’,’yourselves’,’he’,’him’,’his’,’himself’,’she’ ,’her’,’hers’,’herself’,’it’,’itself,’they’,’them’,’their’,’theirs’,’themselves’,’what’,’which’,’who’,’whom’,’this’,’that’, ‘these’,’those’,’am’,’is’,’are’,’was’,’were’,’be’,’been’,….]

Table 4: Sample of stop words

Sentiment analysis through model building was applied to the web forums Montada and Qawem for analysis of the results. The stop words were removed and the remainder sentences were analyzed. The search function was used for extraction of statistics and for fetching information on the number of words that were used in the forums, as shown in Fig. 2. The content was expected to be influenced by religion and ideology with results showing that the top 10 words user were on religion. "Allah" was found to be the most spoken content in both forums.

Qawem contained more frequency-related words on radical ideology than Montana, for example "curse" and "enemies". The figures below provide evidence of opinion mining and opinion analysis of webpost as percentages.

Negative opinions are shown to be less in the Montada web forum as compared to the Qawem forum. Of interest is the fact that the radical affect is quite stronger in the Qawem forum. Furthermore, 35% of the data analyzed in Qawem had a negative score in the range of 0.050 and 0.100, Montada had less than 15% of the data in the same score range. Lastly, the positive scores in the Montada web forum were much higher compared to those in the Qawem forum.

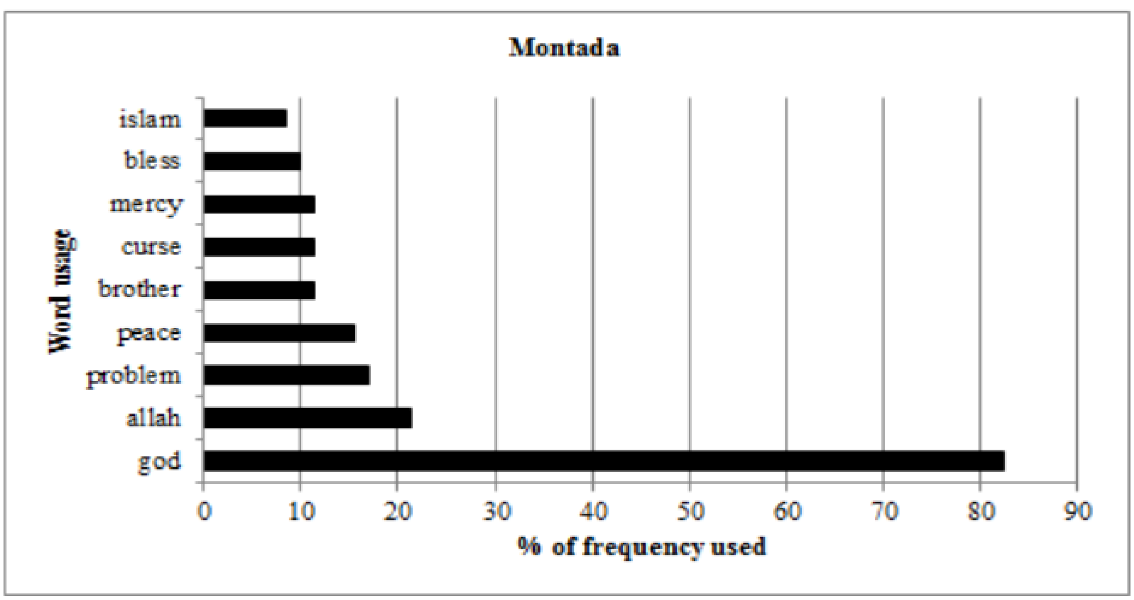

Figure 3: Graph depicting Montada word usage versus the frequency

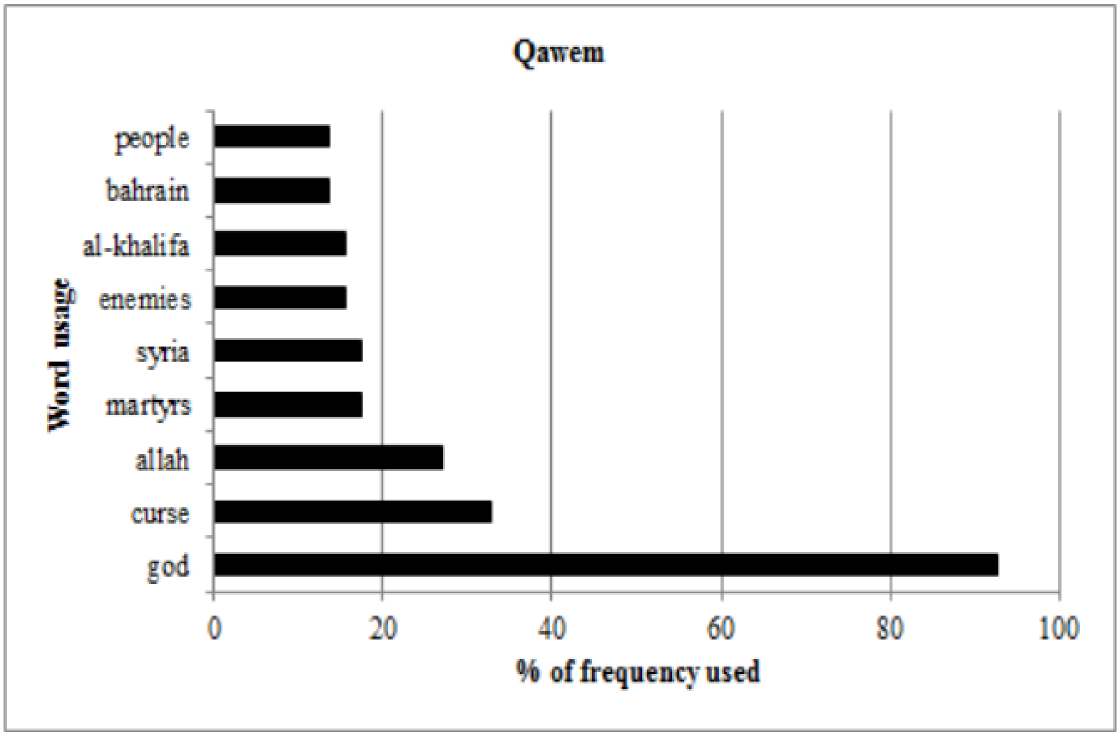

Figure 4: Graph depicting Qawem word usage versus the frequency

2.3 Literature Review Summary

According to Yang Ming article on Social Media Analytic they dwell mostly on static nature of data collection from the post of two websites. This serves as the basis of their empirical data analysis and conclusion. The research methodology on my end tends to encompass both web forums and blogs while targeting individuals with hate speech and radical comments.

The research seeks to address the gap on the two previous work implemented above. The previous work was disadvantaged as it mainly focused on getting data from already classified hate speech web forum. The research that I wish to undertake will comprise of both static already classified web forums and also dynamic collection of social media web forum. The concept undertaken will act as a sniffer on live traffic so as to be able to collect the digital evidence minutes after being posted. The approach is that the automated algorithm to behave as a watchdog and raise alarm on potential hate speech.

Social Media forensics provides a framework for National Intelligence Services, Police and defence forces to draw patterns of individuals and organization that promote hate speech and radicalism. As per the previous review they worked on websites that promote hate speech and radicalism. The research that I tend to partake will seek to incorporate opinion mining on web forums and also target individuals and flag them as potential suspects. It will be a hybrid method of flagging individuals and also getting the possible locations of where the comments were posted. Furthermore, the algorithm will incorporate English, Swahili and the local dialect 'Sheng' languages. It will come up with a sample of keywords or stop words which will be used to query the data collected from social media hate speech. Thus increasing the sampling size of the variables used to verify hate speech.

This is critical for any cyber forensics expert as the location will also provide enough physical evidence to justify that a particular hate monger committed the crime; which is solely dependent on the possibility that the hate speech suspect has enabled location tagging on the comments posted.

On the Affect Analysis of Radical contents by using Sentient shows that there is a need of a lexical dictionary to be developed (Tawunrat, 2013). This approach showed that there is a need for a baseline to be used for assessing, classifying and interpreting hate speech. Hence, the research undertaken will be to define a set of words and comments that will be used as a search basis for the different radical web forums. The dictionary or set of words will be used by the algorithm to run queries on the database.

As earlier indicated on the different challenges faced towards sentiment analysis and opinion mining is that words can be taken out of context and understood in a different manner other than the intended purpose. This is a clear indication that the digital evidence collected cannot be absolute and thus the need of human intervention and interpretation. The research will use the automated tool to get the digital evidence, remove the clutter and reduce the digital evidence so as to enable the human resource to be able to decipher and understand on which context was the hate speech addressed.

The research will focus on solving the problem of automating cyber analysis and also providing a near real-time evidence of data posted in social media. This will aid the police to be able to tackle and address the situation faster. Currently, the police mostly rely on the public and informers to get information on potential target areas of either political violence or terrorist attack. Such a tool will aid in providing heat maps of potential locations and also will classify areas with respect to their political inclination and hate speech status.

Extremists and hate speech mongers are known to use social media and web forums to fuel their propaganda. The algorithm will capture the digital evidence and assist the cyber police and communication authorities to be able to lock down such web forums and websites. With terrorist groups recruiting more tech savvy individuals who use aliases and web browsers that hide their content, the need for the communication authority to draw plans, rules and regulations for ISPs which compel their consumers not to use their networks to propagate hate speech will be an added advantage.

The literature review used machine language techniques for classifying and providing a hate speech index of Dark Web forums. In as much as this provided the potential web forums of social media hate speech, the research undertaken will go a step further and collect all the data and classifying it. Thus, it will provide an easier format of representing the digital evidence and cyber trail.

The algorithm developed will provide a secure environment for all Kenyans so as to avoid potential political chaos or ethnic attacks. Since all these require communication, by placing such tools it will go a long way to capture, provide digital evidence and prosecute the cybercrime individuals.

2.4 Conceptual Design

The following is the conceptual design layout to be undertaken:

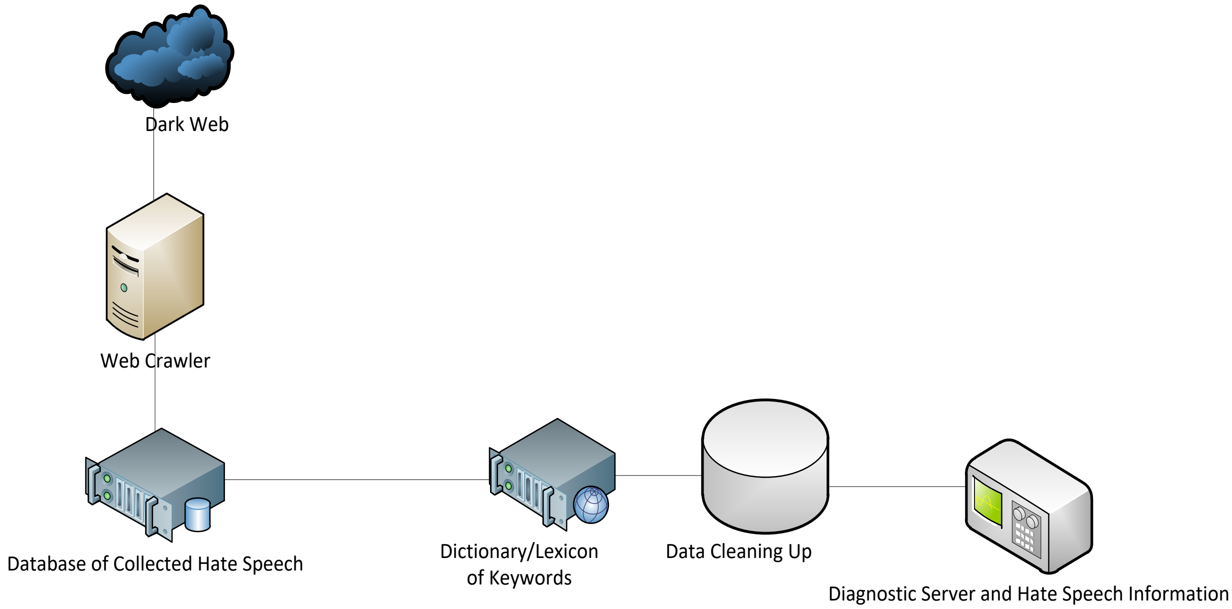

Figure 5: Conceptual Framework of Hate Speech Model

CHAPTER THREE: METHODOLOGY

3.1 Introduction

This chapter addresses the methods, procedures and instruments to be used by the researcher to gather data and analyze them. The chapter covers the development methodologies that were used in the study.

3.2 Software Development Life Cycle Methodology

3.2.1 Agile Software Development

"Agile software development describes a set of principles for software development under which requirements and solutions evolve through the collaborative effort of self-organizing cross-functional teams" (Collier, 2011). Larman (2004) demonstrated agile software development as the promotion of adaptive planning, development of evolutionary, early delivery, and continual improvement of software while encouraging rapid and flexible response to change.

Agile software development defines software development as a group of methodologies that promote iterations, open collaborating and adaptability of processes throughout the life-cycle of the development of the project. It allows the developer to perform small increments with minimal planning on a develop as you go basis, rather than planning the whole development at length. This minimizes the overall risk and encourages the project in adapting to changes more frequently. It places an emphasis on stakeholder and code developer involvement translating to the stakeholder being consulted about the product iterations and comments noted as software development proceeds.

This methodology was employed for the development of the web interface, database, search query components and sentiment analysis tools for hate speech ranking. Documentation, requirements documents and design documents, is a key part of the methodology as well as the source code. This encourages that the software development continues even if one or more members of the development team leaves the whole project does not collapse. The complete working design document is important as new team members or a completely different new team can be easily brought to speed on the progress of the design and easily familiarize themselves when undertaking the software development (Hughey, 2009).

3.3 Data Sources

All the data received on the Internet in Kenya would have been the appropriate sample specimen. However, because of the numerous daily Internet data, forums, websites and discussion chat. The researcher was to target the popular website and web forums.

A total of 20 public and popular websites, social media forums and blogs such as Twitter, Facebook and blog posts were to be considered to participate in the dark web collection of hate speech and radical data.

Some of the social media web forums researched on were on religion, tribe, community, education, universities, WhatsApp, Facebook and twitter. A set of keywords such as ‘Kill’, ‘Burn’, ‘Wezi’,’Fala’ will be used as search tools to query the database of the collected web forum data.

3.4 Data Collection

With the data sources identified a means for data collection, harvesting, cleaning and verification was implemented. The study mainly focuses on social media data from websites, Facebook, Twitter and blogs. Harvesting data from social media is provided for by open source and several APIs which are specific to the social media site. For this research, we developed an open source web crawler to be able to fetch and collect datum from all the data sources. We did not use APIs as they are specific and proprietary and limits in terms of access to data. Python was the progamming language that was used to perform web crawling and scraping. The use of Python was chosen due to its versatility, agility and previous studies have shown it to be a viable solution for web crawling, spidering, indexing and scraping.

The collected web crawled data are stored in the Mysql database before data preprocessing and transformation is applied. Labels of positive, negative and neutral are appended to the training set of social media data. The collected data includes text, emoticons and common acronyms.

3.5 Data Preparation and Transformation

The opinion mined data is usually unstructured and contains irrelevant and non-textual characters and thus requires to be prepared, processed and transformed for data evaluation and validation. Text preparation involves cleaning before analysis is performed (Rambocas et al, 2013). The data preprocessing is broken down into 3 steps:

1. **Tokenization**.

Tokenization involves splitting a string into its desired constituents seeking to isolate as much sentiment information as possible. Tokenization helps in keeping the vocabulary as small as possible. Tokenization is performed on emoticons and abbreviations and are identified and treated as individual tokens.

2. **Text Normalization.**

"Text normalization is the process of transforming text into a single canonical form that it might not have had before. Normalizing text before storing or processing it allows for separation of concerns, since input is guaranteed to be consistent before operations are performed on it. Text normalization requires being aware of what type of text is to be normalized and how it is to be processed afterwards; there is no all-purpose normalization procedure" (Sporat, 2011).
Abbreviations are noted and replaced by the meaning they represent, for example, LOL. Informal intensifiers are also determined such as character repetition and all capital words are made to small letters as they signify different ASCII values while character repetitions are reduced to three characters.

3. **Part-of-speech (POS) tagging**.

POS in corpus linguistics is also referred to as word-category disambiguation or tagging of words grammatical. It involves marking and classifying of words in a text (corpus) based on definition, context, its adjacent relationship with related words in a sentence, or paragraph (DeRose, 1988).

A Part-Of-Speech Tagger (POS Tagger) has been made and converted into a framework that reads text usually in English and assigns tokens to each word such as noun, verb, adjective

3.6 Bag of Words Model

“The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity” (Josef, 2009).

The bag-of-words model makes an unigram model of the text, this is done following the number of occurrences of each word and keeping a track of those occurrences. This forms part of the text classifiers by taking individual words into account and giving them a specific subjectivity score. The sentiment lexicon compares the subjectivity or objectivity score and grades the score as negative or positive.

3.7 Algorithm Considered and Justification

Natural Language Processing algorithms are many and widely used depending on the size of data and accuracy for performing sentiment analysis. Hence, Naïve Bayes usefulness in solving data mining classification problems as exposed from previous work as highlighted by the literature review was considered. Furthermore, its been applied to various datasets and disparity, results were obtained (Gopala and Bharath 2013), (Qasem et al. 2014).

Naive Bayes is super simple for performing counts. The Naïve Bayes classifier will converge quicker as long as the conditional independence is achieved hence you need less training data. A Naïve Bayes classifier does a great job even if the conditional assumption is not met. Its considered the best classifier when it comes to fast converge and ease of use.

The main disadvantage of a Naïve Bayes classifier is that it cannot learn interactions in different feature sets. Hence Naïve Bayes classifier was used as the Natural language processing algorithm for hate speech sentiment analysis.

3.8 Sentiment Analysis Prediction

“Sentiment analysis is the computational study of people's opinions, sentiments, emotions, and attitudes. This fascinating problem is increasingly important in business and society. It offers numerous research challenges but promises insight useful to anyone interested in opinion analysis and social media analysis” (Liu, 2016).

Once the data set is defined and evaluated, the bag of words forms the dictionary of words and the training set for sentiment analysis. With the classifier in place, the set of words are classified in terms of being either positive or negative together with range of subjectivity measured as 1.

3.9 Research Design

The research used the Dark Web, Blacklisted Social Media accounts and web forums together with Social Media users as inputs together with information Security Experts. These inputs were used to collect quantitative data and describe Social Media Hate Speech web forums characteristics.

The Dark Web was considered as a source of data collections when it came to randomly collect information that is likely to becoming form independent international web forums.

The blacklisted Social Media accounts and web forums which were potential mediums for hate speech were considered. This was so because a digital forensic expert sometimes information is relayed of potential suspects of hate speech. The researcher used this as a vital input to collect inculpatory and exculpatory digital evidence against the social media accounts and web forums.

Information Security Experts were also approached in order to validate and support the digital evidence, digital audit trail and quantitative data collected.

The collected output comprised of persons of interest and the digital evidence collected. The digital evidence comprised of the timestamps, digital signatures and digital cyber trail taken by the cyber-criminal and hate speech mongers.

The data collected will be through a web crawler which will take all the data, text and images from a potential web forum and place it in a database. The system comprises of a database, dictionary of words and a search engine. Once the data is collected and placed in the database. The system will have a dictionary of words and will highlight the hate speech words together with the full text of the information. It will also provide additional information of the timestamp and also location where the hate speech words were uttered. The dictionary of words is mainly a dataset of potential hate speech words such as kill, burn and others which will be used as the baseline for assessing the hate speech collected.

Hence with that the semi-automated monitoring tool providing the digital data and potential persons of interest as evidence of hate speech. The digital evidence collected will combine both the textual data and images.

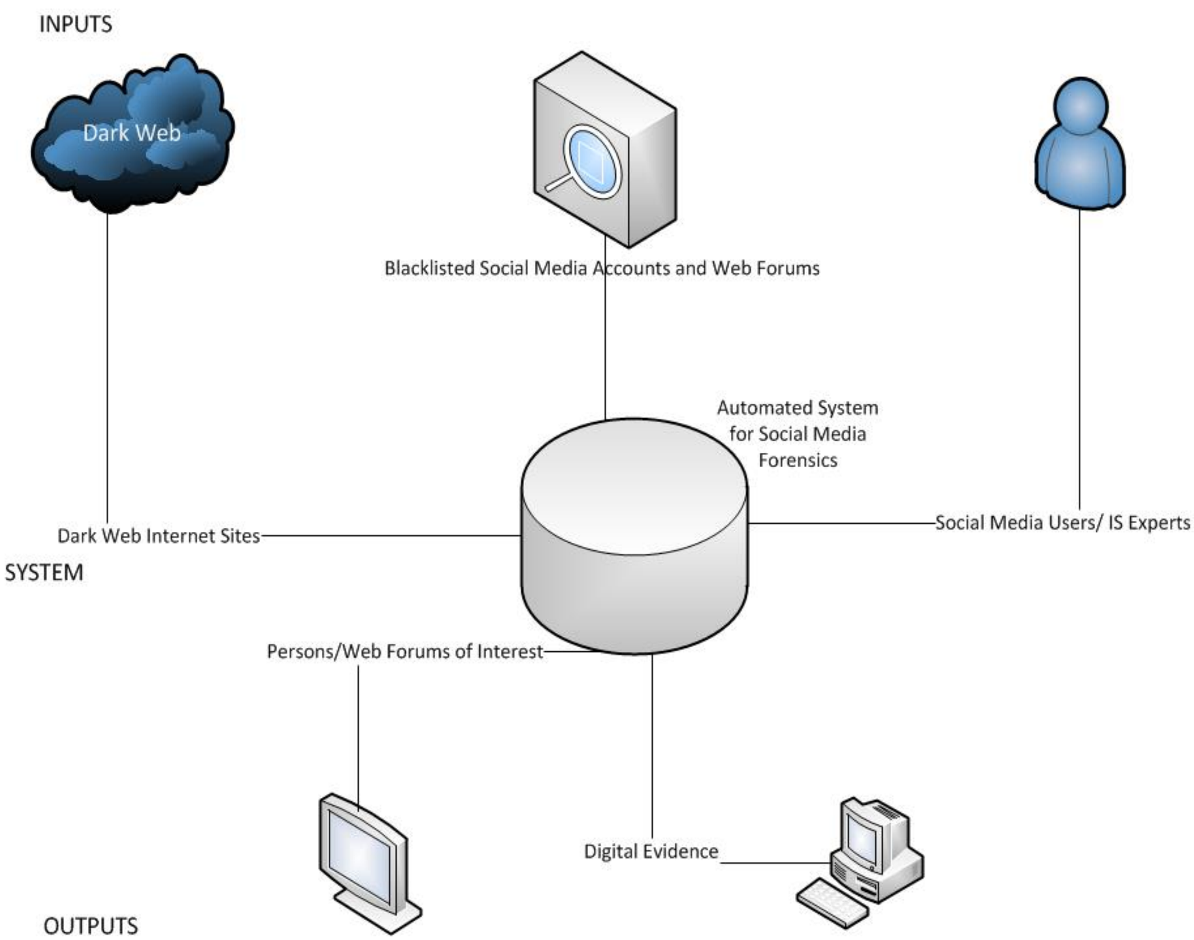

Figure 6: Inputs, System and Outputs for Data Collection

3.10 Prototype validation

The below figure illustrates the general architecture of the components of the system.

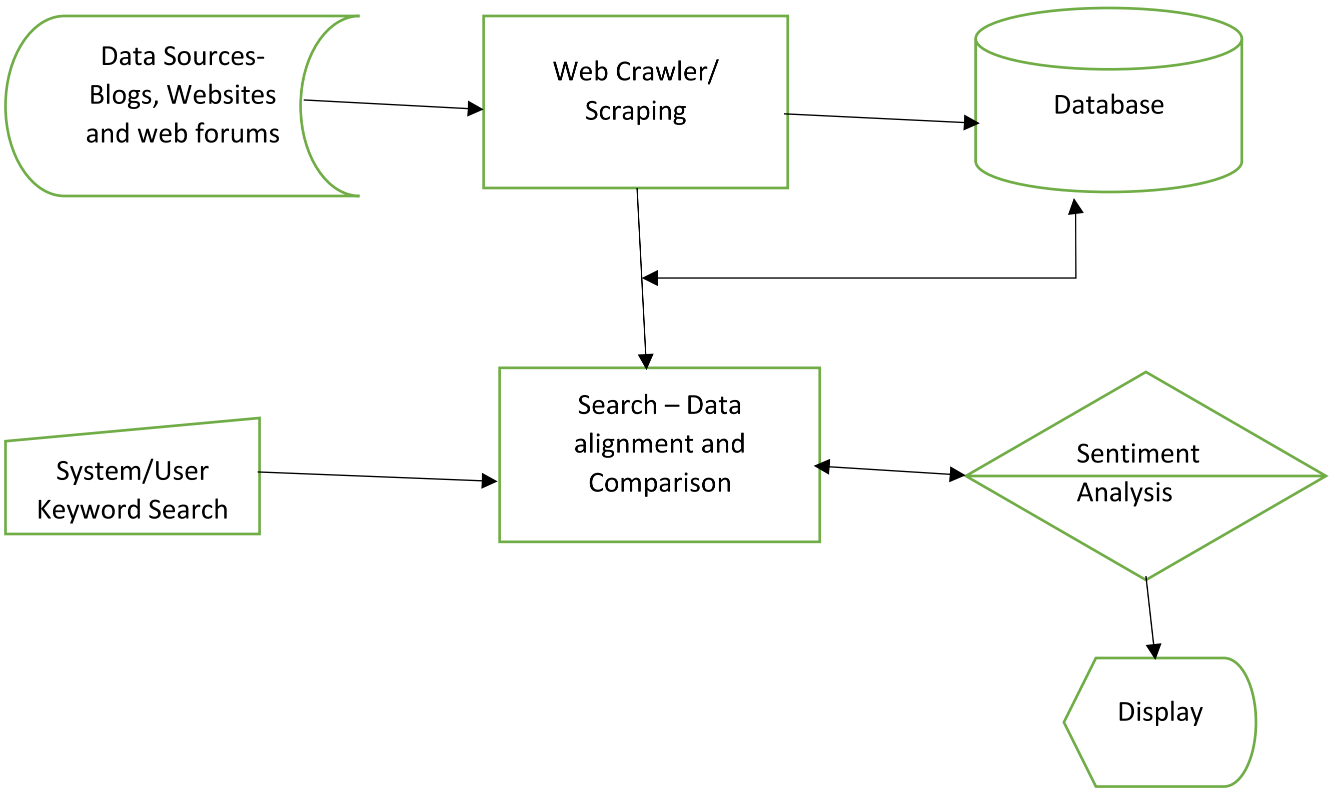

Figure 7: Prototype Elements for the Application

3.11 Evaluation

F1-score is a Machine Learning algorithm and statistical measure of test's accuracy combining both precision and recall. It has an inbuilt NLTK with metrics that are used to measure accuracy, precision and recall.

For determining the accuracy of a single Classifier, or comparing the results of different Classifier, the F-score is usually used. This F-score is given by

$$F = \frac{2pr}{p + r}$$

where p is the precision and r is the recall. The precision is the number of correctly classified examples divided by the total number of classified examples. The recall is the number of correctly classified examples divided by the actual number of examples in the training set.

The overall solution should have the capacity to:

I. Collect opinions from social media sites and place them in the Mysql Database

II. Identify the keyword set of potential hate speech and train the Natural language classifier.

III. Check the sentiment analysis in terms of Positive or Negative comment.

3.12 Test Management Plan

Graham (2012) defines test management as all those activities for testing, analyzing and managing the software testing process. He goes further and demonstrates that the test management tool is the procedure and methodology for automated or manual testing of software. These tools include requirement, specification and description of management modules that allow automatic generation of the requirement test matrix (RTM). RTM is one of the main metrics for functional coverage of the system under test (SUT).

The test management plan was used as the methodology for testing the various components of application as shown below.

	Parameter	Description
1	Identifier	The unique reference.
2	Introduction	A brief explanation of the test scenario
3	Items	Software element of the application under test.
4	Tested Feature	The software feature uniquely being tested.
5	Non-features	Features that will not be tested and reason behind
6	Methodology	Overall detailed method of testing.
7	Pass/Fail	Has the feature passed or failed the test
8	Test deliverables	Acceptance Test Procedures, Test plans, Specification and Summary
9	Tasks	Procedure for executing the tasks
10	Environment	The environmental requirements both hardware and software

Table 5: Showing the components of Test Management Plan

The Test Steps describe the procedures and methods for executing, testing and the possible expected results. The results are marked as pass or fail based on the comparison result between the expected and actual outcome. The test cases usually comprise of the following fields:

- Scenario
- Procedure
- Variables
- Expected Result
- Actual Result

3.13 Data Analysis

In the research great lengths have been made to assess the data to the best it can be done. Automatic filtering and analysis has been implemented so as to reduce the data set and making continual assessment and monitoring of Hate Speech. The manual data filtering and analysis will seek to seek to provide a contextual lookup of the data that has been flagged as Hate Speech.

As we are aware that with words and information posted in social media, there is a need to understand the context in which the word was used. For example, bomb is a potential word for flagging down as Hate Speech but if used in a different context it can be an affirmation. Such as in the case a friend tells a friend you are the bomb, meaning that the other friend is a cool and nice friend.

Hence, the need for a semi-automated tool is critical as it provides the constant automatic monitoring and ease of work. But also the human knowledge element is important to remove the clutter and jargon flagged as hate speech by the system.

The data collected from the 20 potential hate speech social media web forums, Facebook and twitter accounts through the spidering program will form the database. The raw data will undergo filtering by using the keywords as search tools which will be run as queries on the database. Once the filtered data is collected it will be mapped against hate speech monger together with the total counts of hate speech and exact phrase of hate speech.

The qualitative analysis will be undertaken by the Cyber Forensic expert to check if the conversation was taken out of context. Once the quantitative and qualitative analysis is complete the data is confirmed as the true reflection of the data.

3.14 Summary

This research thesis target was to investigate the social media web forums for hate speech and report and blacklist them. Additionally, it was to provide the forensic digital evidence. The research was intended to collect dark web information and from it mine the opinions of hate speech and classify them.

The results from the inputs would pave the way for the development of the system and framework for collecting hate speech, mining hate speech opinions form the data collected and to provide a systematic approach of the digital evidence collected.

CHAPTER FOUR: ANALYSIS, DESIGN AND IMPLEMENTATION

4.1 Introduction

This chapter covers system's requirements, design and finally the implementation of the system.

4.2 Systems Analysis

4.2.1 Functional Requirements

The following are the functional requirements for the prototype components:

Web Crawler

1. Should be able to receive the URLs of the website and blogs and retrieve data through web crawling and scraping
2. Should allow indexing of the crawled and scraped websites or blogs or web forums.

Compiler

The Interpreter is the central processing unit of the platform and form the core bond between the crawler and the database. It seeks to understand the data crawled from the website, cleaned and integrate it to the database.

It should be able to:

1. Understand the data scraped and crawled from the website.
2. Cleans the data and removes non ASCII characters together with the removal of html tags,
3. Stripping text of excess spaces and removal of stop words.
4. Polls the data and aligns it for entry into the database.
5. Perform text classification, preprocessing, tokenization and use of stop words of the web crawled data.
6. Perform sentiment analysis on the collected web crawled data.

Database

The relational database should be able to do the following:

1. Allows and store the indexed data scraped from the website.
2. Poll, relay and retrieve the search data with respective to the keyword used.

Search

It will should be able to do the following based on the keyword text being queried:

1. Search the database for the title of the URL and the content with respect to the keyword
2. Present the searched data in form of the URL and content body being requested.

4.4.2 Non-Functional Requirements

The non-functional requirements specify systems' properties and constraints. The following are the non-functional requirements for the application:

1. Administrator's graphical user interface should be easy to use and navigate.
2. Performance requirements:
   a. The system should have short response time during web crawling
   b. Short data access time when keywords are being searched on the database.
3. Operating system constraints: The application should not exhaust available computing resources or limit the behavior of the system hardware, memory and processor.
4. The application needs at least the following 1GB ram, CPU 2,4GHz, Linux Centos 64bit system.

4.3 System Design

4.3.1 Architectural System Design

The Figure 9 illustrates the detailed architectural design of the system.

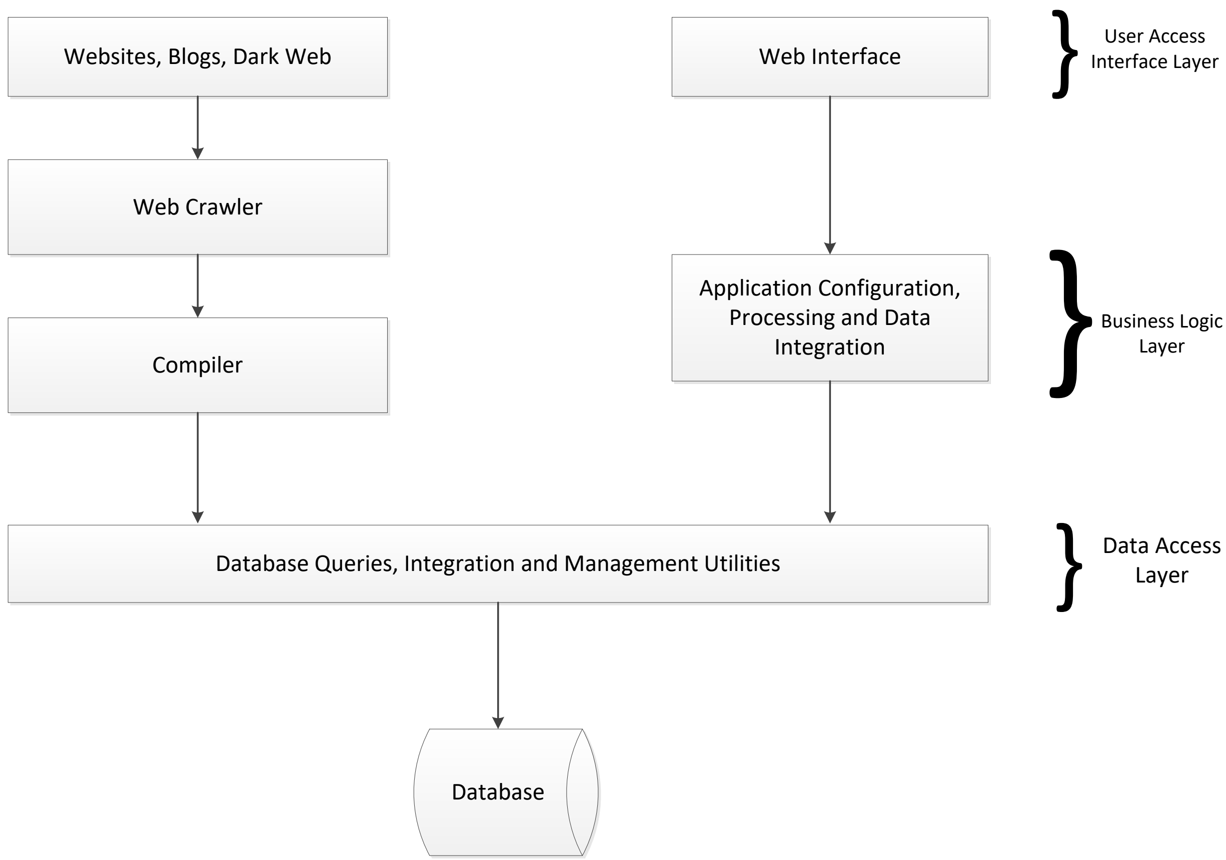

Figure 8: Architectural System Design

1. The **User Access Interface Layer** is the first impression of a software system from the user's point of view. The User Access Interface was mainly to perform two functions:

- Accepting the user's input
- Displaying the output

2. The **Business Logic Layer** or domain logic is the part of the application that encodes the real-world business rules that determine how data can be created, displayed, stored, and changed (Minksy, 2005). This the core part of the system and has the workflow and

business rules for the system. It includes the application and configuration of the web crawler and working logic of the compiler or interpreter.

3. The **Data Access Layer** serves retrieval requests from the upper layers these include; data storage, data retrieval and data crosschecking.
4. The **Data Layer** this is the data store for all data pertaining to the application, includes support related data and applicant data.

4.4 Use Cases

This phase captures use cases from the initial system requirements and restructures them as a sequence diagram (Figure 10).

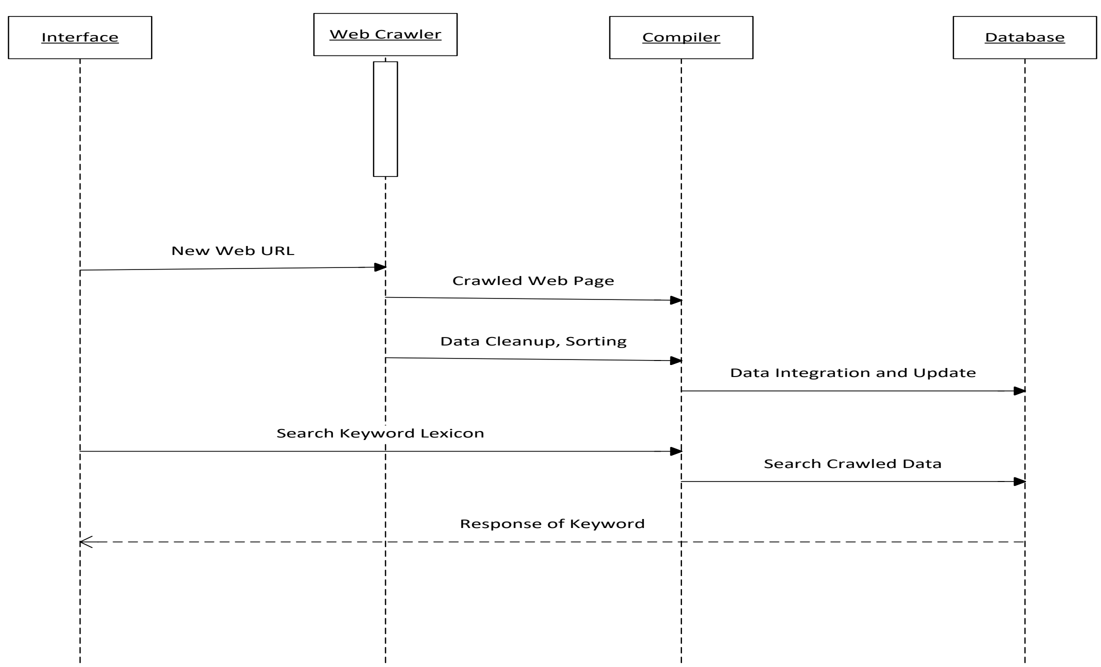

Figure 9: Sequence Use Cases

4.3.4 Logic/Data flow of the Application

Application Initialization

Figure 11 illustrates the logic that pertains to initialization of the application.

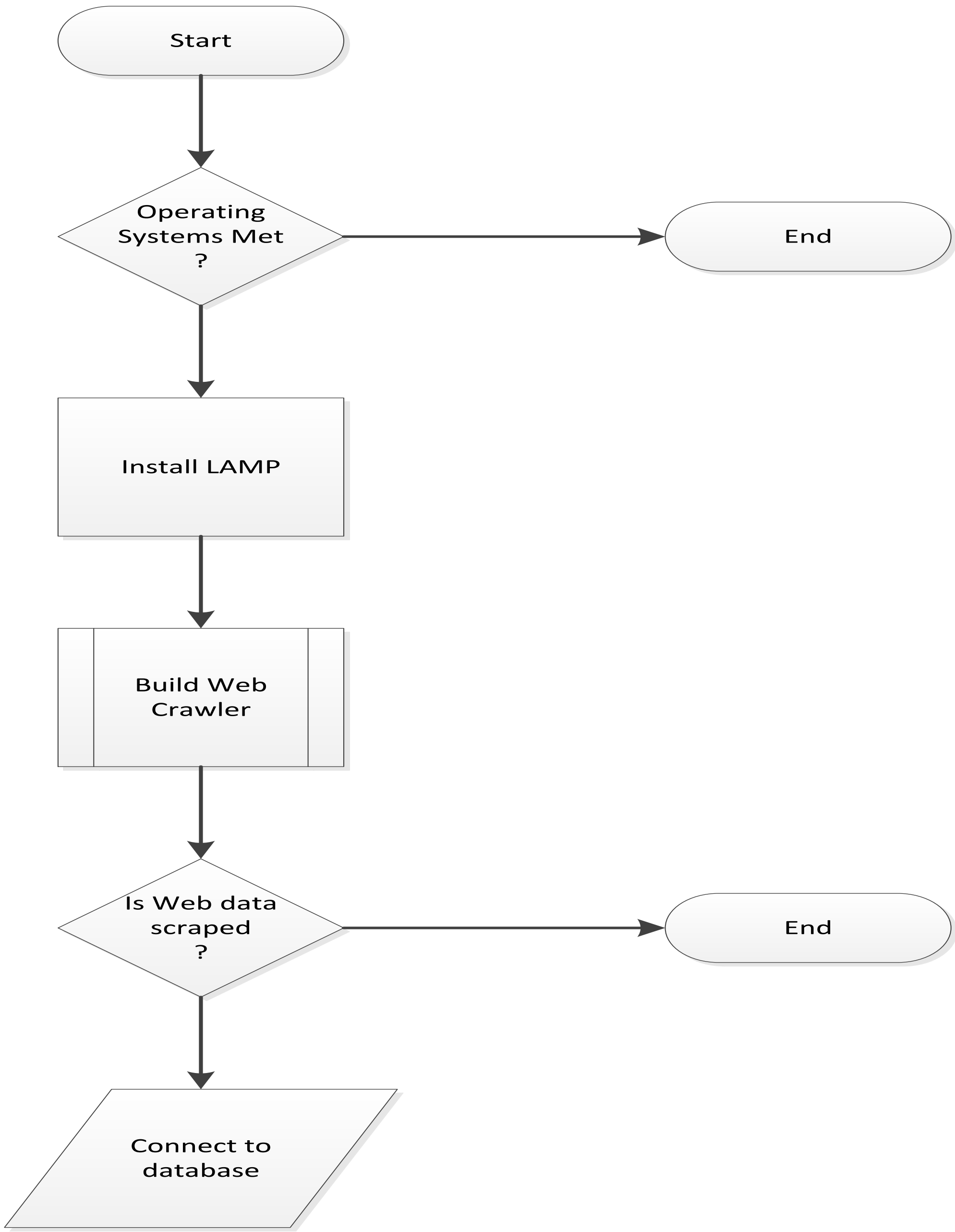

Figure 10: Application Initialization Diagram

The system requirements that are required to be met for application to run include installation and initialization of apache, MySQL and python.

Application Logic

Figure 12 shows the data/logic flow during social media forensics; the steps followed during a typical

run of the application.

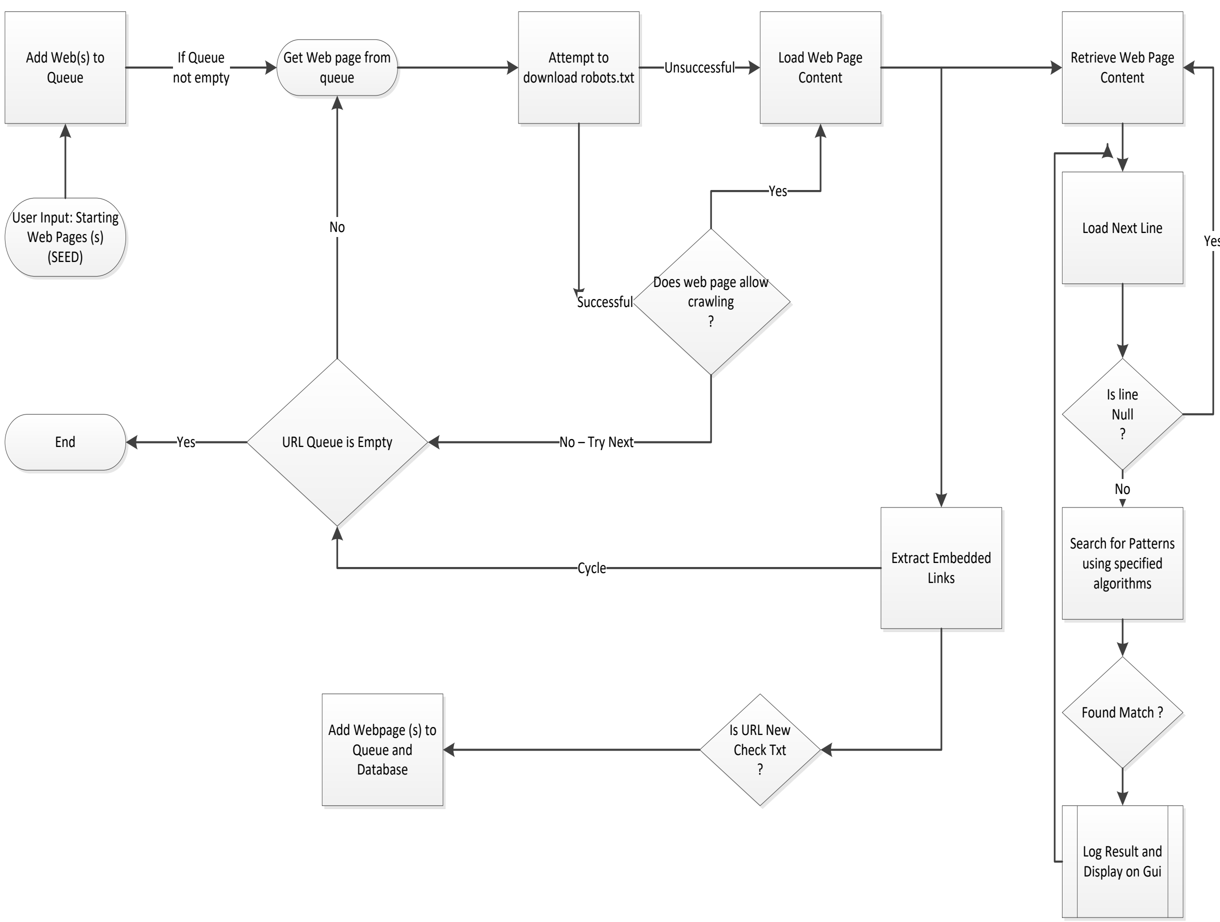

Figure 11: Architectural System Design

Figure 13 shows the logic flow during the retrieval of an applicant's details based on the index number numerical string provided in the email text.

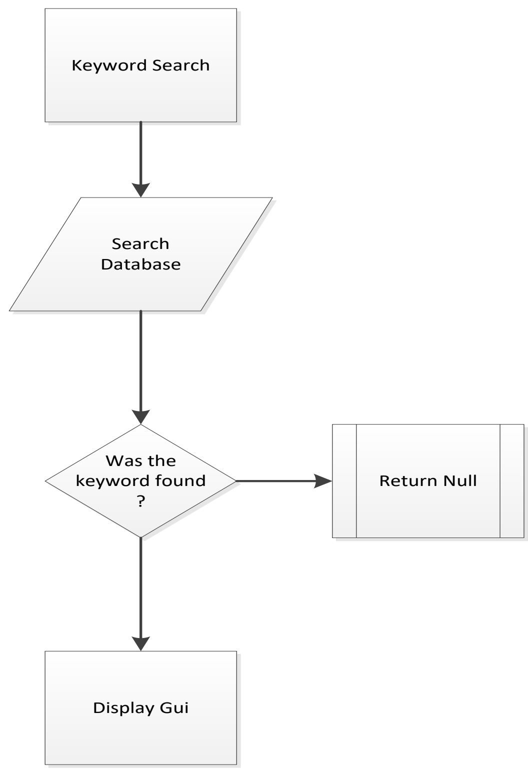

Figure 12: Flowchart Depicting the retrieval process

4.3.5 The Database Model

4.3.5.1 Support Database

The below is the database model of the application.

The MySQL database had one table called scrabots and the columns on the table include.

Title was a variable character and includes the title of the web page and the heading info of the URL

URL was also a variable character and was the column in which the URL for each corresponding scraped website was placed.

Body was the data on the website and which comprised of the content.

Date stamp was the date and time when the scraping had been done.

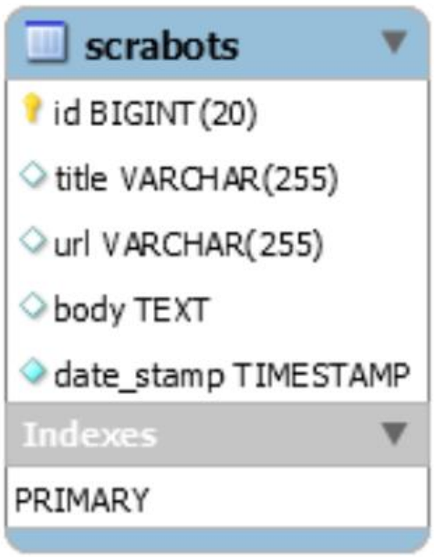

Figure 13: Schema for the Scrabot Database

4.4 System Implementation

4.4.1 The User Front End

The user front end was implemented using LAMP (Linux Apache MySQL PHP). The PHP/HTML was used to design the web interface. Linux and PHP scripts supported the backend application running on Python and Scrapy spiders.

4.4.2 The Application Logic

The application was implemented with Python programming language and MySQL as the database on the CentOS 7 Linux platform.

4.4.3 Web Crawler

The web crawler was designed in Python together with Scrapy, a Python library for Web Crawling and Scraping.

4.4.4 Compiler

It was designed using Linux, PHP and python scripts. MySQL queries and commands were also used to be able to input the crawled data into the MySQL database.

4.4.5 Search

Mainly used MySQL queries to search the MySQL database, HTML and PHP scripts to present the data on the web page.

CHAPTER FIVE: RESULTS AND DISCUSSION

This chapter presents and discusses the results obtained from study.

5.1 Presentation of Results

The Social Media Forensics was designed and developed to accept the URL and perform the digital forensics of copying the data and placing it in a database.

Sentiment Analysis Automation

The application was designed to perform automatic sentiment analysis on the basis of begin positive or negative and the value subjectivity being calculated as shown below.

The data was represented as the title of the website, the corresponding URL, the positive and negative ranking together with the measure of sentiment.

From the below sample of the Kahawatungu Website (www.kahawatungu.com) we check a Title named "Senior UNEP employees celebrate Tom Cholmondeley for killing Kenyans" which has a Positive of 0.3 and a negative 0.4.

('"gor mahia\'s favourite couple breakup citing differences"	"http://www.kahawatungu.com/2015/01/14/gor-mahias-favourite-couple-breakup-citing-irreconcilable-differences/"\n'	'Positive: 0.571428571429'	'Negative: 0.285714285714')
('"sofapaka players stranded in nairobi after failing to raise airfare to zimbabwe -"	"http://www.kahawatungu.com/2015/02/25/sofapaka-players-stranded-in-nairobi-after-failing-to-raise-airfare-to-zimbabwe/"\n'	'Positive: 0.230769230769'	'Negative: 0.461538461538')
('""dr james mwangi sabotaged my business when i refused his sexual advances" - esther passaris -"	"http://www.kahawatungu.com/2016/08/01/dr-james-mwangi-sabotaged-business-refused-sexual-advances-esther-passaris/"\n'	'Positive: 0.25'	'Negative: 0.25')
('"cj willy mutunga fades into the night like a witch - donald kipkorir -"	"http://www.kahawatungu.com/2016/06/13/cj-willy-mutunga-fades-night-like-witch-donald-kipkorir/"\n'	'Positive: 0.285714285714'	'Negative: 0.357142857143')
('"kahawa tungu (@kahawatungu) | twitter"	"https://twitter.com/kahawatungu"\n'	'Positive: 0.4'	'Negative: 0.2')
('"kahawa tungu - bitter!! sweet!!"	"http://www.kahawatungu.com"\n'	'Positive: 0.4'	'Negative: 0.2')
('"senior unep employees celebrate tom cholmondeley for killing kenyans -"	"http://www.kahawatungu.com/2016/09/02/senior-unep-employees-celebrate-tom-cholmondeley-killing-kenyans/"\n'	'Positive: 0.3'	'Negative: 0.4')
('"facebook developers workshop in nairobi scheduled on 26th august -"	"http://www.kahawatungu.com/2016/08/02/facebook-developers-workshop-nairobi-scheduled-26th-august/"\n'	'Positive: 0.2'	'Negative: 0.4')
('"family of a four year old gilr defiled in bomet by a 50 year old is looking for help -"	"http://www.kahawatungu.com/2016/08/22/family-four-year-old-gilr-defiled-bomet-50-year-old-looking-help/"\n'	'Positive: 0.2'	'Negative: 0.45')

Table 6: Showing the Sentiment Analysis based on being Positive or Negative

Example 1

The application is given a Facebook URL of a potential dark web forum for scraping.

The forensic investigator clicks on the submit button so as to perform the web scrawling.

Figure 14: Welcome Page for Scrabot Web Crawler

Once the web crawling and indexing of the web forum is complete the below response is shown.

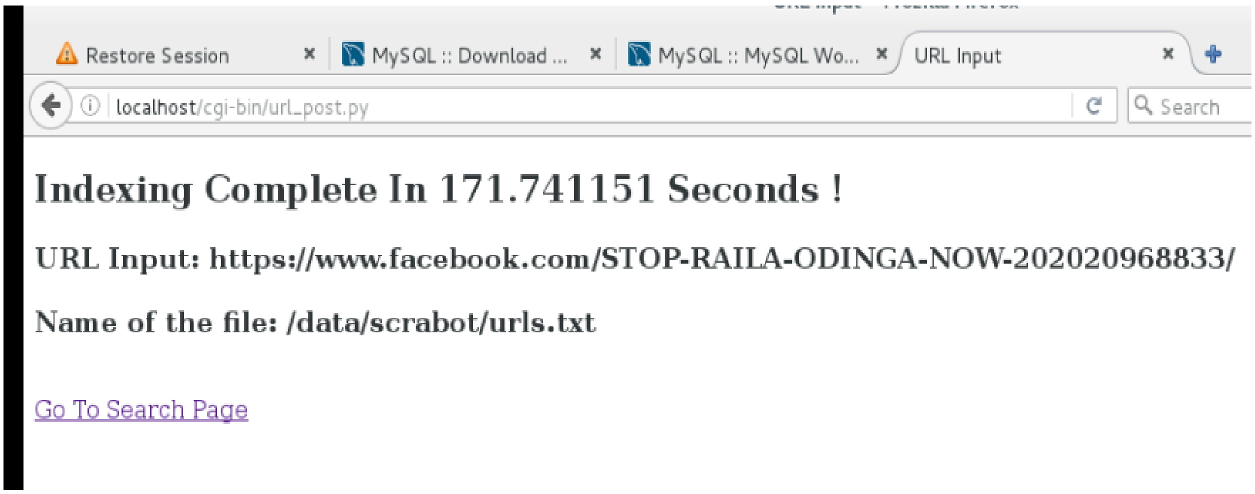

Figure 15: Indexing Process Completion

The forensic investigator is given the below information:

URL Input: The full URL that was used to web crawling and scraping.

Name of file: The server location of the URL and data crawled.

Go to Search Page: The link for performing the keyword search of the potential hate speech.

Search Page

The forensic investigator is presented with a keyword search page where they can enter the text to be searched.

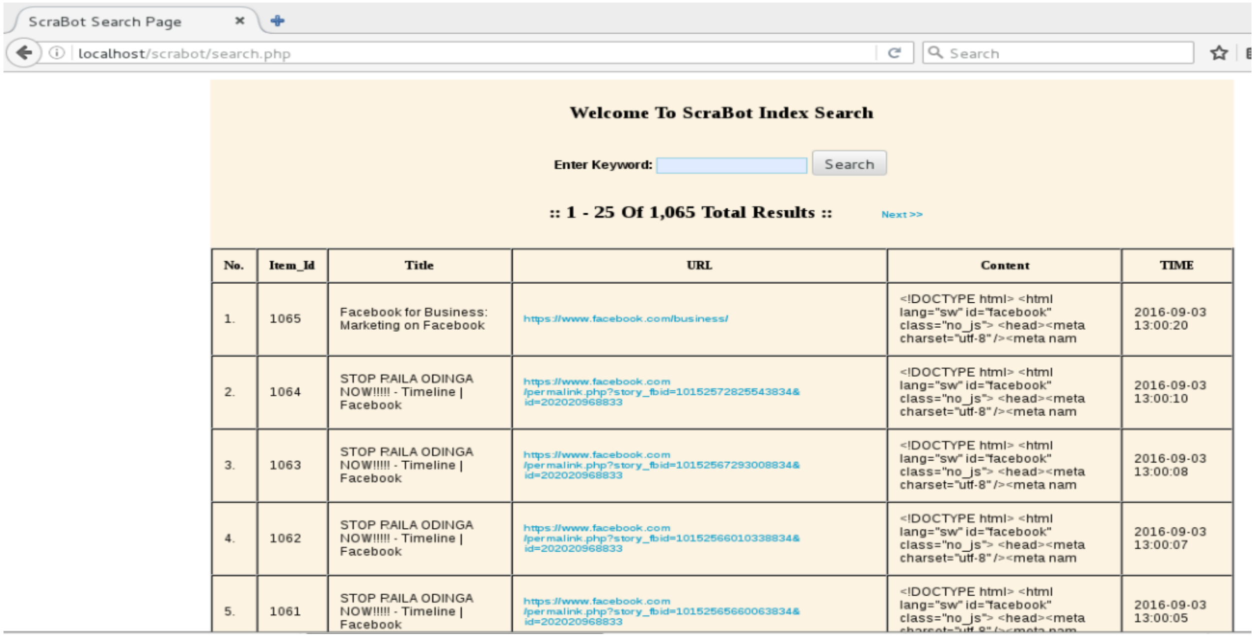

Figure 16: Index Search Process

Keyword Entry

The digital forensic investigator enters the keyword to be searched.

From the keyword shown there are 69 total results for the keyword “raila” searched.

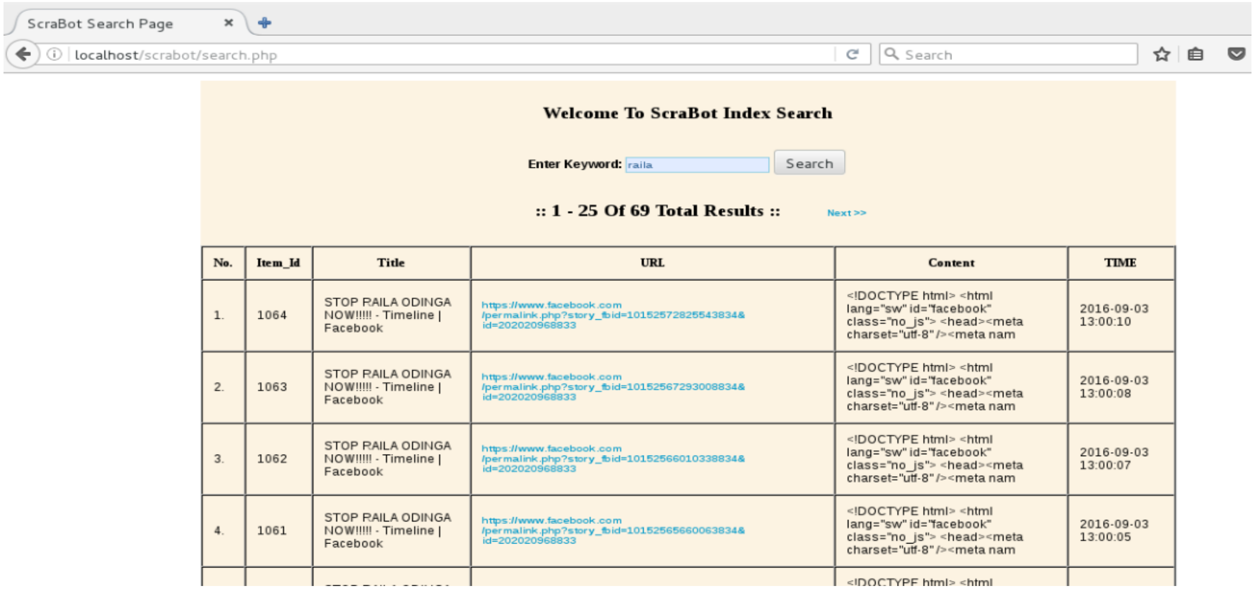

Figure 17: Search Index Response

From the 69 total results a person called Gerald is flagged as a potential hate speech monger and the database is searched further.

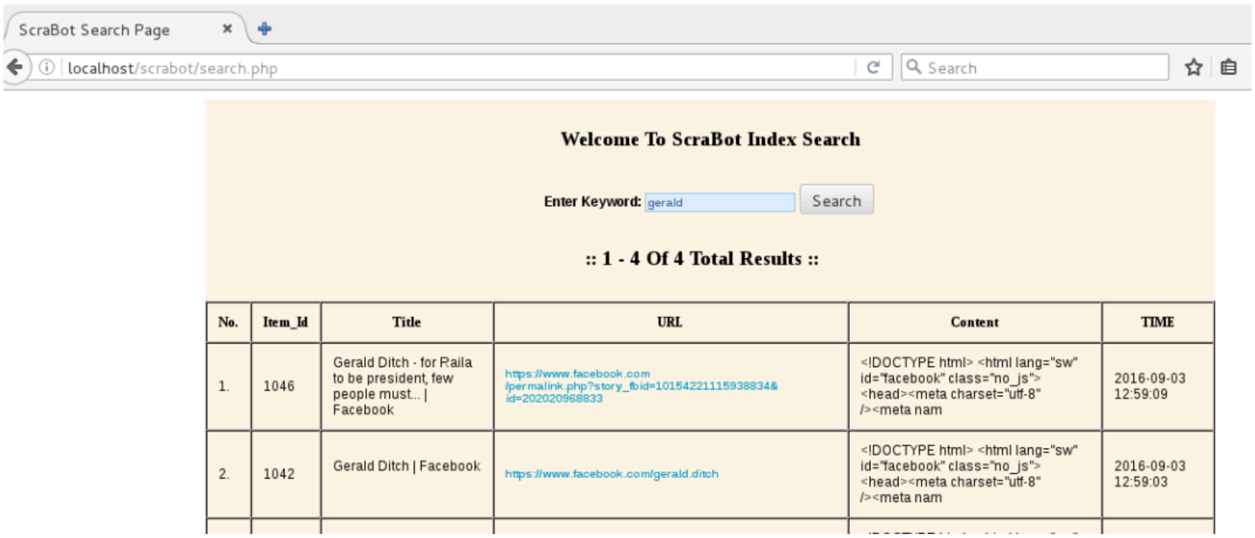

Figure 18: Search Index Process Results

From the searched results for Gerald Ditch, we get the title of the Comment and the full URL of the website where the comment in regards to the "STOP RAILA ODINGA NOW" Facebook page. The full content which has been minimized to 100 letters for ease of display and the time when the data was crawled and stored on the database.

URL

The forensic investigator goes further and clicks on the full URL and is directed to the Facebook comment of the potential hate speech monger.

Figure 19: Website showing the hate monger actual text

Blog

The Social Media Forensics Tool was performed also on a popular blog to check the possibility of hate speech and the content being said on the mentioned web forum.

The URL Input was a web forum called KahawaTungu owned by Robert Alai, a prominent supporter of Orange Democratic Movement, a local Kenya political party.

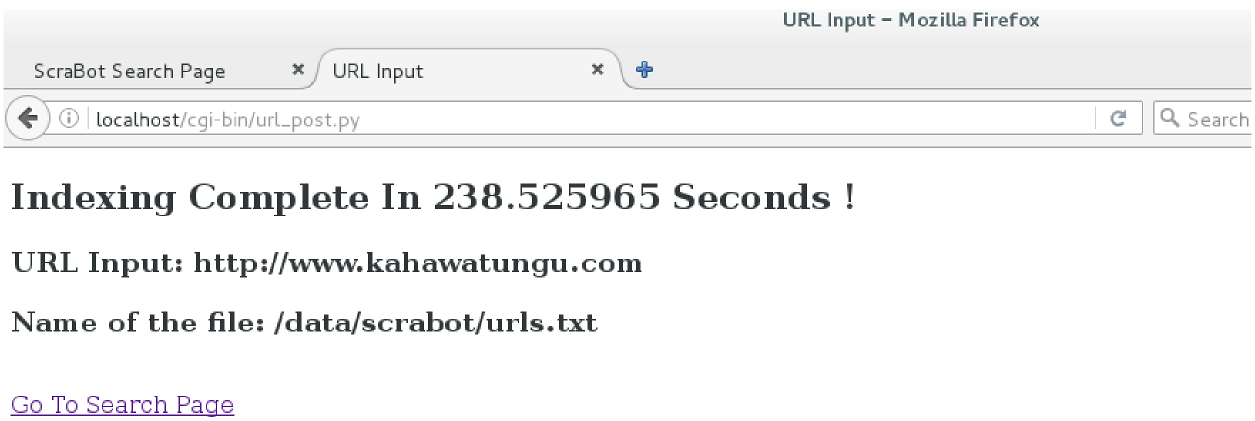

Figure 20: Kahawatungu Indexing Completion Process

Search Keyword

The keyword was searched and the below were the potential hate speech comments.

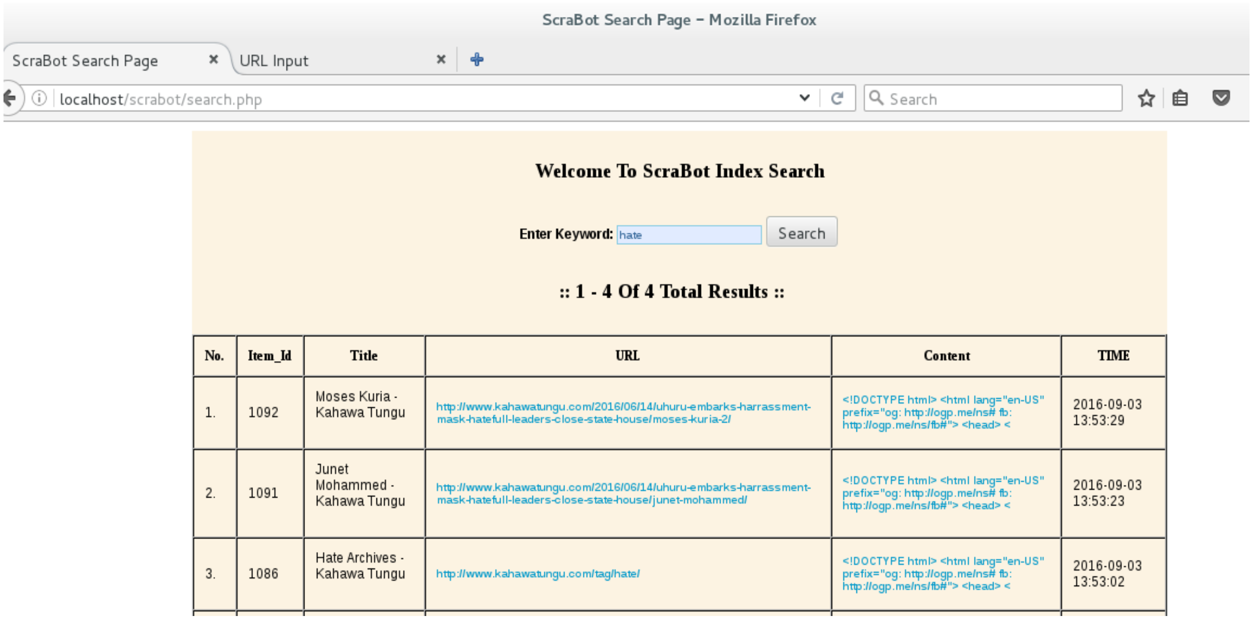

Figure 21: Kahawatungu Index Search Response

The digital forensics investigator clicks on the URL to be directed to the actual page.

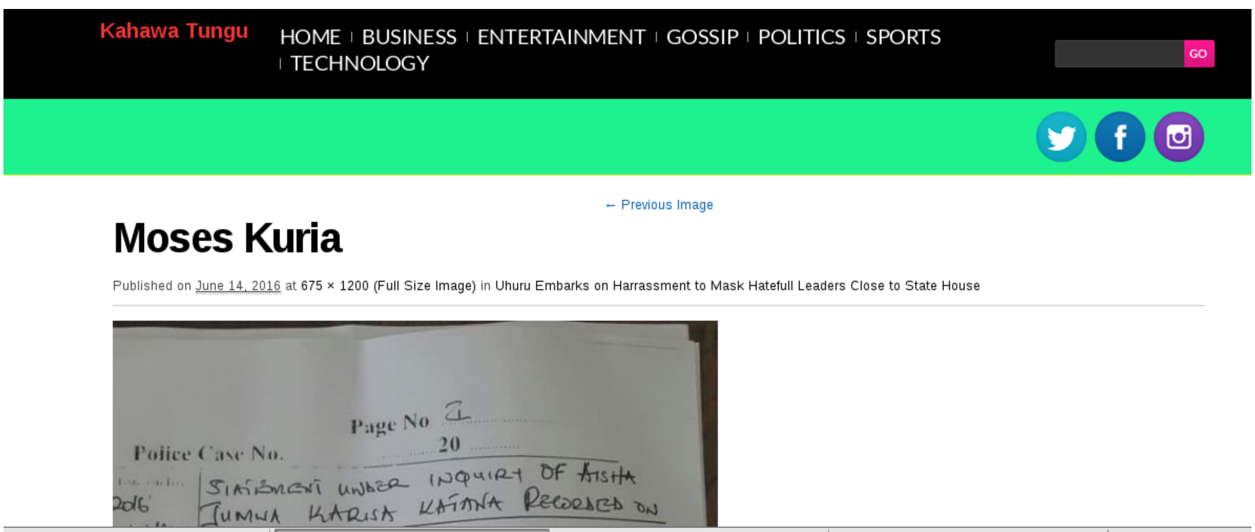

Figure 22: Website for Popular politician and the hate speech offence

Social Media Forensics Statistics

Here are the sample data of the web forums and the amount of traffic in regards to Hate speech and subject matter on potential political topics.

The statistics for the crawled web page and the count of the mentioned times a particular topic has been highlighted.

Daily Post Kenya		
URL: www.kenyan-post.com		
	TOPIC	COUNT
	ODM	45
	Jubilee	55
	Uhuru	20
	Raila	19
	Hate Speech	16
	Kikuyus	32
	Luo	28

Table 7: Daily Post Opinion and Sentiment Analysis

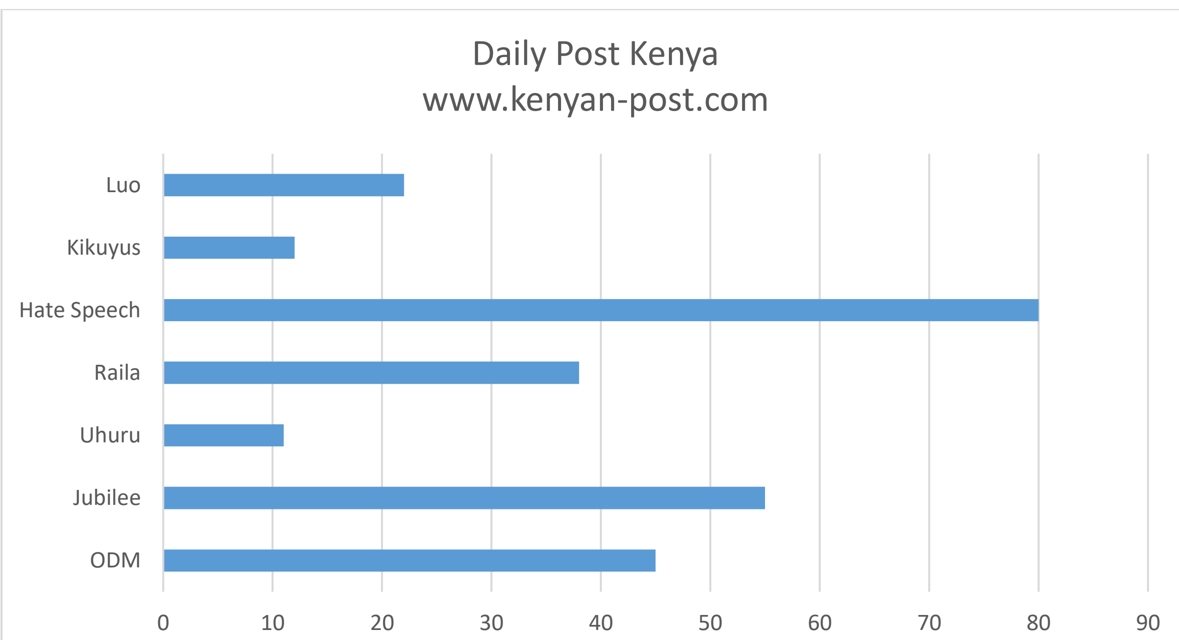

Figure 23: Graph showing Sentiment versus Count for Daily Post Website

From the above data there is a clear indication that the owner of the website is propagating hate speech directed to a particular group in this case the Luo and Jubilee being the political party being supported in this scenario.

Facebook Page

The "Stop Raila Now" Facebook page was full of inflammatory comments against the ODM, flag bearer. The Social Media Forensics tools captured the data as shown below.

Stop Raila Odinga Now Facebook Group		
URL: https://www.facebook.com/STOP-RAILA-ODINGA-NOW-202020968833		
	TOPIC	COUNT
	ODM	45
	Jubilee	55
	Uhuru	11
	Raila	38
	Hate Speech	80
	Kikuyus	12
	Luo	22

Table 8: Table showing Facebook Group Sentiment Analysis

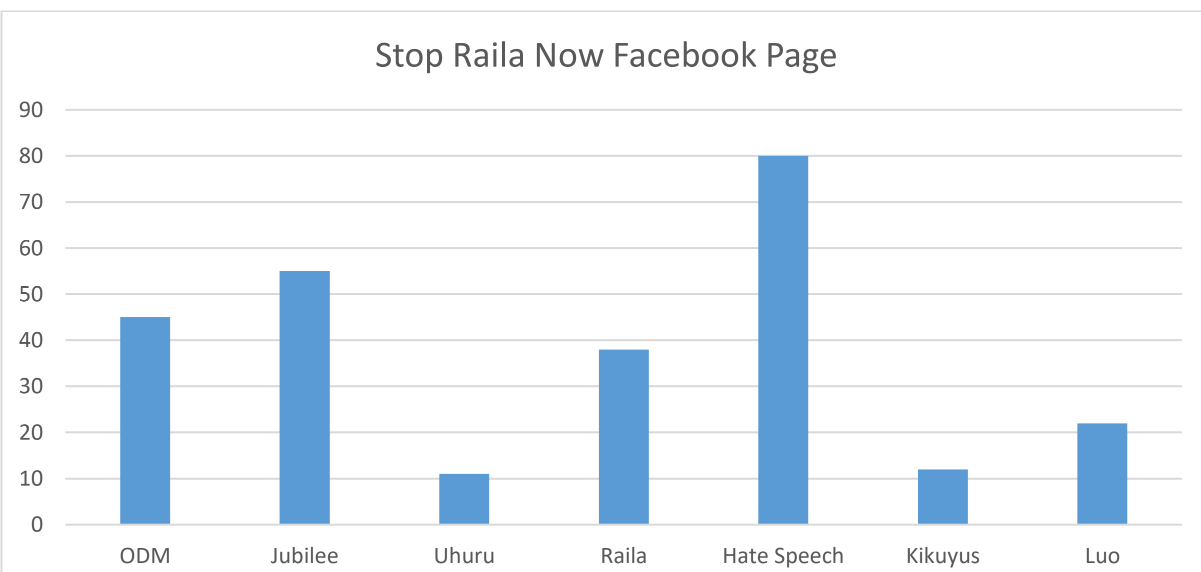

Figure 24: Graph showing Sentiment Count versus Topic

Web Forums Statistics

The Social Media Forensics Tool took the data for the all the web forums and the keywords used for searching the hate speech data.

Website	ODM	Jubilee	Uhuru	Raila	Hate Speech	Kikuyus	Luos
Kahawatungu	63	58	35	79	102	15	12
Stop Raila Now	45	55	11	38	80	12	22
Stop Uhuru Now	20	89	90	88	120	50	12
Gor Mahia Is not a club it’s a lifestyle	100	35	101	260	243	210	60
Daily Post Kenya	45	55	20	19	16	32	28
Tuko Kenya	23	18	60	15	57	12	10
Nairobi Wire	59	72	49	32	34	10	9
Moses Kuria Facebook Page	40	150	120	43	10	8	6
Juneet Mohammed Facebook Page	150	68	9	28	6	3	8
Johnstone Muthama	256	48	32	15	13	19	13
Shoebat - Chistians Army					254		
Kenya Today	34	38	93	27	356	34	45
Team Mafisi	135	154	256	103	276	78	63
Cyprian Nyakundi	97	99	40	15	345	32	45
Dikembe Disembe	200	98	57	106	132	13	16

Table 9: Web Forum Statistics of Sentiment and Subject of discussion

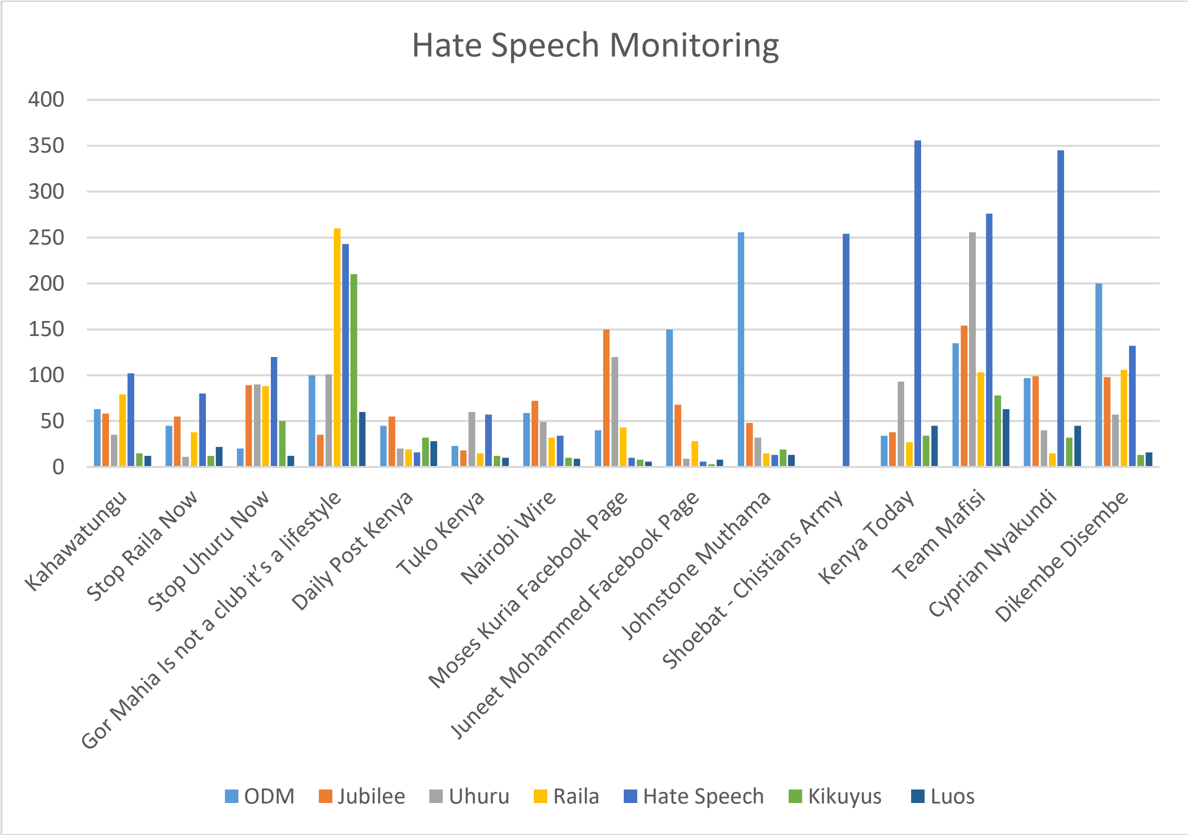

Figure 25: Graph Hate Speech Monitoring Statistics

Hate Speech Medium Classification

The below data shows the relationship of the medium used to propagate hate speech. From the below data is evident that Blogs are the main mediums of hate speech mongering followed by Websites and Facebook Accounts. The data is a clear indication the hate speech mongers are known as the digital evidence clearly shows below

Web Forums	Hate Speech
Websites	819
Facebook Accounts	472
Twitter Account	477
Blogs	993

Table 10: Showing the Web Forums and the Hate Speech Count

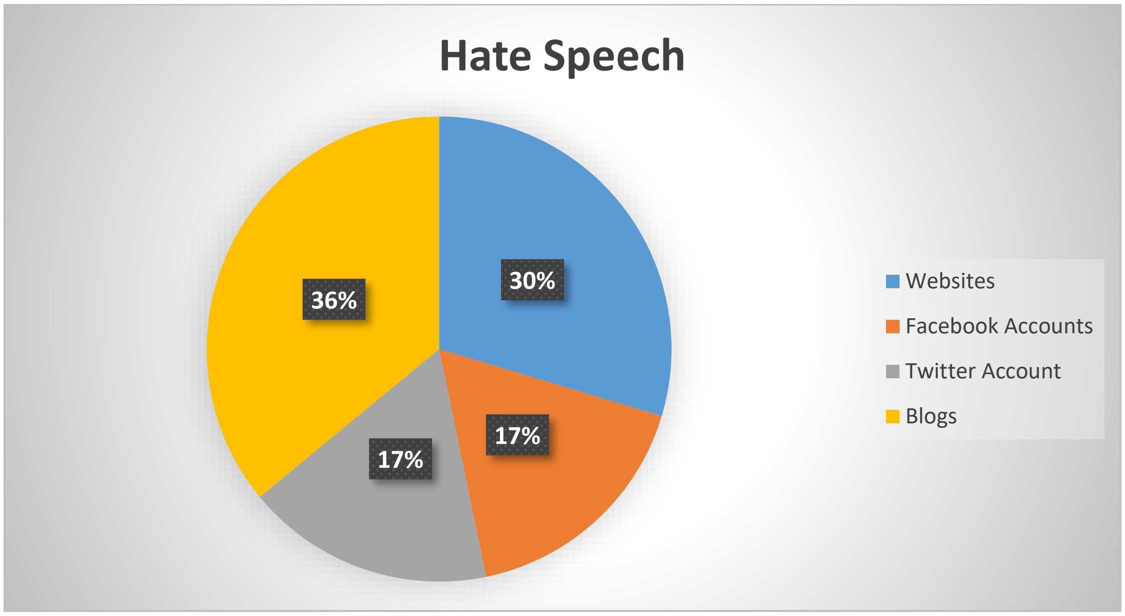

Figure 26: Pie Chart Showing the Hate Speech Sentiment versus the Web Forum

Test Management Plan

As per the defined test plan management plan, the following.

Test Scenario	Test Steps	Parameters	Expected Results	Actual Results	Success	Failure
Unit Testing	-Database -Web Crawler -Web Interface	-Login -Passwords -Uptime	-Successful Logins -Wrong Password Notification	-Worked fine	10	0
Usability	-Level of Skill Required -Time required -User Attitude	User Friendliness Ease of directions and buttons	-Ease of use -Tech Savvy	-No Skill required -Less instructions on the manual	8	2
Use Case Testing	-Actors and if transactions	-Need for supervision	The ease of how	Achieved	8	2

	are carried out	-Is the process automatic	transactions are handled			
User Acceptance	-Application works	User acceptance for hate speech	User accepts it as a viable solution	Achieved	7	3
User Interface Testing	-User friendliness -GUI Testing	-Buttons execution -Display - Understandable	Achieved	Achieved	9	1
Volume Testing	-Check data loss -System Response Time -Data stored correctly -Check if data is overwritten	-DB access -DB Backup scripts -Time it takes for DB queries	-No dataloss -DB backups done everyday -Select and Insertion queries to run faster	Achieved expect database backup	8	2
Vulnerability Testing	-Strength of Passwords -Logins	- Cyprotographic Tools	Check the passwords and logins	No system to check how strong the password is	6	4
Performance Testing	-Load Test -Stress Test -Soak Test -Spike Test	-Database load -Multiple processes running -Users increased	-Database total simultaneous logins	Partly	5	5

Table 11: Test Management Results

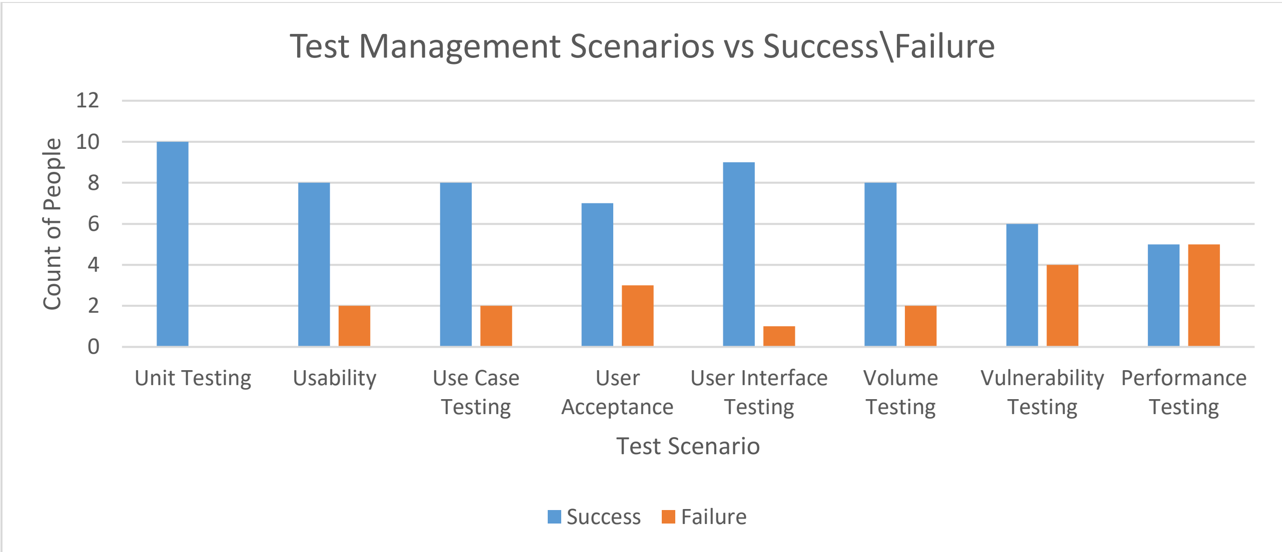

Figure 27: Test Management Scenarios

Discussion of the results

From the above results and analysis using the set of keywords the frequency of the potential hate speech data was used as a means for sentiment and opinion analysis. The frequency of the keywords and personalities mentioned was used as a gauge for flagging and blacklist some of the websites. This was important as it allowed the digital forensics investigator to be able to flag on the website in relations to the level of how offensive the website is.

CHAPTER SIX: CONCLUSIONS AND FUTURE WORK

6.1 Achievements

The objectives of the study were discussed and the below were the procedures used to achieve them:

i. To identify and analyze techniques used in hate speech monitoring and select the best suitable technique for creating a customized hate speech application. To achieve this objective, different techniques for hate speech monitoring were analyzed which included hate speech using machine language, natural processing algorithms and classifiers

Algorithm	Problem Type	Results Interpretable	Easy to explain	Predictive accuracy	Training speed	Prediction speed	Tuning Needed	Small number of	Noise Handling	Feature interaction	Calibrated probabiliti	Parametric ?	Scaling	Algorithm
KNN	Either	Yes	Yes	Lower	Fast	Depends on n	Minimal	No	No	No	Yes	No	Yes	KNN
Linear regression	Regression	Yes	Yes	Lower	Fast	Fast	None (excluding regularizati on)	Yes	No	No	N/A	Yes	No (unless regularized)	Linear regression
Logistic regression	Classificati on	Somewhat	Somewhat	Lower	Fast	Fast	None (excluding regularizati on)	Yes	No	No	Yes	Yes	No (unless regularized)	Logistic regression
Naive Bayes	Classificati on	Somewhat	Somewhat	Lower	Fast (excluding feature extraction)	Fast	Some for feature extraction	Yes	Yes	No	No	Yes	No	Naive Bayes
Decision trees	Either	Somewhat	Somewhat	Lower	Fast	Fast	Some	No	No	Yes	Possibly	No	No	Decision trees
Random Forests	Either	A little	No	Higher	Slow	Moderate	Some	No	Yes (unless noise ratio is very high)	Yes	Possibly	No	No	Random Forests
AdaBoost	Either	A little	No	Higher	Slow	Fast	Some	No	Yes	Yes	Possibly	No	No	AdaBoost
Neural networks	Either	No	No	Higher	Slow	Fast	Lots	No	Yes	Yes	Possibly	No	Yes	Neural networks

Table 12:Differences between the Natural Processing Algorithms

From the above list of classifiers used for natural language processing algorithms we can see that Naive Bayes is super simple and performing a bunch of counts. If the Naïve Bayes conditional independence assumption holds, a Naive Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data. And even if the Naïve Bayes assumption doesn't hold, a Naïve Bayes classifier still often does a great job in practice. Its a good bet if someone wants something fast and easy that performs pretty well. Its main disadvantage is that it can't learn interactions between features. Hence Naïve Bayes classifier was used as the best Natural language processing and Support Vector machines algorithms for hate speech sentiment analysis

ii. To develop an application that will combine English, Swahili and "Sheng" hate speech keywords for data mining. An application was designed, implemented and deployed using LAMP (Linux Apache MySQL PHP and Python). The algorithm was designed to perform web crawling with the web scrapped data being placed in a database in preparation for searching. The application went further to provide how the crawled database was queries and the indexed search results provided.

    The application combined the Natural Language Tool Kit (NLTK) such as the Naïve Bayes classifier together with support vector machine learning tools to automate sentiment analysis based on the title of the potential website. Furthermore, Matplotlib, a python graphing library tool, was used to show the scatter plot of the sentiment data.

iii. To demonstrate and test the application while providing analysis on the hate speech websites being investigated. A set of keywords that were used to search the stored crawled websites formed the basis for analysis. The frequency of the potential hate speech keywords was used to plot and quantify the sentiment of the web forum.

    This was represented statistically as shown on the table below where the application sentiment analysis was evaluated and cross-validation performed on the premise of testing the precision, recall, F1 score and support with the average accuracy score of the sentiment analysis tool as 75.89%.

It was known and shown that the tool depended heavily on the dictionary of words and keywords for opinion mining. Furthermore, the training of the dataset was important for the accuracy in terms of positive or negative comments to be reached.

	Precision	Recall	f1-score	support
Negative	0.71	0.91	0.8	409
Positive	0.86	0.59	0.7	375
Avg/Total	0.78	0.76	0.75	784
The Accuracy Score is 75.89%				

Table 13: Cross Validation of the Sentiment Analysis

Measuring access of web database

The measures of efficiency considered were: -

- Successful login
- Successful web crawling, indexing and data storage.
- Successful display of potential hate speech on the page

Thirty logins were done and data on the above measures was taken and recorded on an Excel table. The following is a summary of the software application access efficiency data as analyzed using Excel.

Access Efficiency Measures	Success	Failure
Login	30	0
Crawling, Indexing	27	3
Display of hate speech	24	6

Table 14: Efficiency Access Tests on the Application

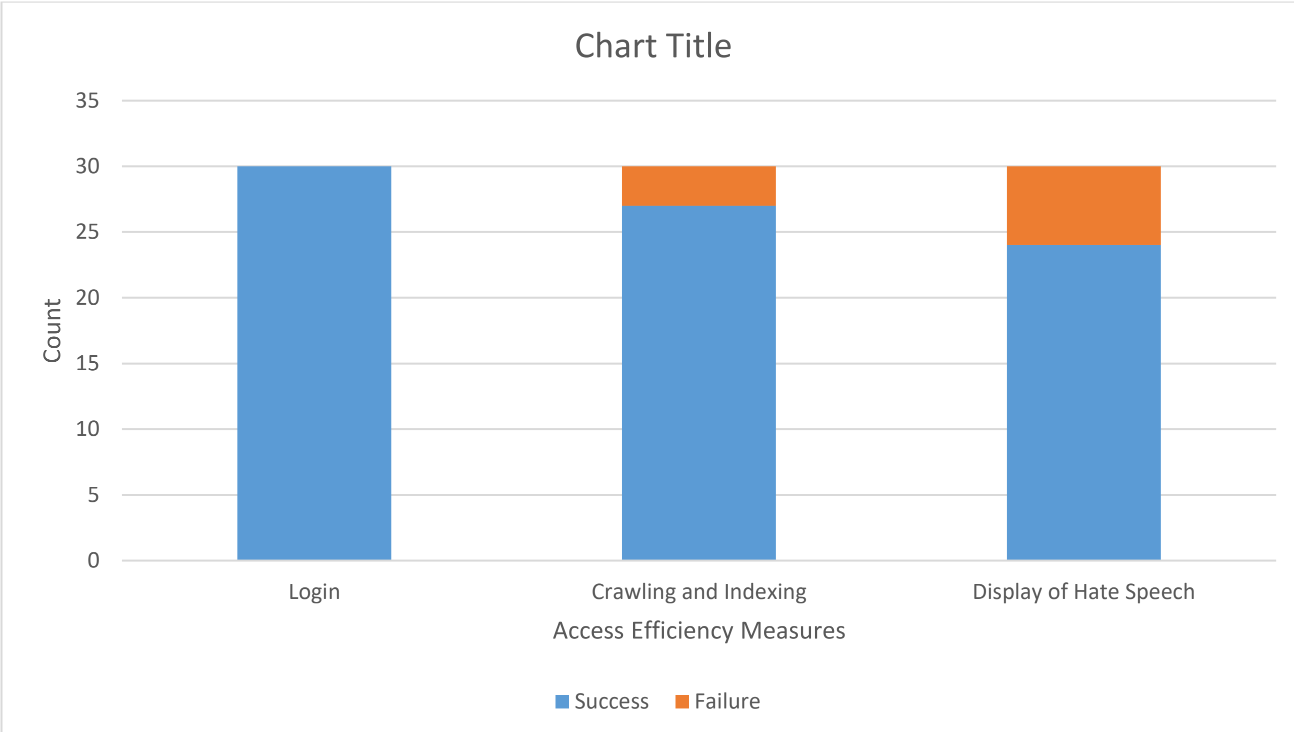

Figure 28: Measure of Access of the Application's General Outlook

In the end an application together with an online occurrence form was developed and easily customized for the purposed of providing a Social Media Forensics Tool for hate speech monitoring.

6.2 Limitations

The study encountered the various limitations:

Operating System: Due to the huge volumes of data that web forums such as websites and blogs possess there was need of a much faster operating system so as to be able to perform web crawling faster and to go deeper in the web scraping.

High Data Volumes: Social Media possess a lot of data and there was need to deploy faster database schemas and mechanism for data storage. This will allow for faster indexing and processing. Social Media is referred to a Big Data and thus there was need to employ different data storage mechanisms.

Programming Language: Since different programming languages such as PHP, Python and Shell Scripting were used to develop the application, a lot of time was consumed for testing and debugging application related issues.

Time: Developing the application required a lot of time in analyzing the web crawled data and finding the best way to represent the scraped data. The searching of the database and indexed data together with the presentation of the data also consumed a lot of time.

6.3 Conclusions

It was discovered during the web crawling process of the application, the body of the web forum was being tagged together with other html tags and this made it even harder when it comes to presenting the website content. The presentation of the website content inside a table required the need to limit the body of the website so that it can be viewed on the search page.

All in all, the objectives of the study were achieved but there was possibility of presenting the data even much better and faster. The research has provided the platform to show hate speech can be monitored and shown how data can be scrapped from web forums and stored as evidence.

Forensics as defined is the application of scientific methods and techniques in order to recover data from electronic/digital media. It involves the preservation, identification, extraction, documentation and interpretation of computer or digital media for evidence and/or root cause analysis in a legally acceptable manner. The application has demonstrated the acquisition of the digital data from the web forums, identification of the potential hate speech by the use of keywords and extraction and documentation of the digital evidence.

6.4 Future Work

Future studies should focus on improving on the duration taken for web crawling and data presentation. Social Media is part of big data and hence the need for huge databases with faster processing speeds can be applied to improve on the monitoring.

The need for faster operating systems which will perform even more deeper web crawling and indexing can studied in the future.

The presentation of the searched data was manual and future studies can further automate the process so that the system can present the data as business intelligence models and present graphs showing the real-time data. The system can go further by employing machine intelligence algorithms for predictive analysis

REFERENCES

A. Khan and B. Baharudin, "Sentiment classification using sentence-level semantic orientation of opinion terms from blogs," presented at the National Postgraduate Conference (NPC), 2011.

Abbasi, A. and CHEN, H. 2007b. Analysis of affect intensities in extremist group forums. In Intelligence and Security Informatics. E. Reid and H. Chen, Eds. Springer (forthcoming).

Abbasi, A. and Chen, H. 2005. Identification and comparison of extremist-group Web forum messages using authorship analysis. IEEE Intell. Syst. 20, 5, 67–75.

Abbasi, A. and Chen, H. 2006.Visualizing authorship for identification. In Proceedings of the 4th IEEE International Conference on Intelligence and Security Informatics, San Diego, CA, 60–71.

Abbasi, A. and Chen, H. 2007a. Affect intensity analysis of Dark Web forums. In Proceedings of the 5th IEEE International Conference on Intelligence and Security Informatics, New Brunswick, NJ, 282–288.

Baumrin, Julian (2011). Internet Hate Speech and the First Amendment, Revisited

Chalothorn, T. and Ellman, J. (2013) 'Affect Analysis of Radical Contents on Web Forums Using SentiWordNet'. International Journal of Innovation Management and Technology, 4 (1). pp. 122-124.

Collier, Ken W. (2011). Agile Analytics: A Value-Driven Approach to Business Intelligence and Data Warehousing. Pearson Education. pp. 121 ff. ISBN 9780321669544. "What is a self-organizing team?"

Crilley, K. 2001. Information warfare: New battle fields, terrorists, propaganda, and the Internet. Aslib Proc. 53, 7, 250–264.

DeRose, Steven J. 1988. "Grammatical category disambiguation by statistical optimization." Computational Linguistics 14(1): 31–39.

Glaser, J., Dixit, J., and Green, D. P. 2002.Studying hate crime with the Internet: What makes racists advocate racial violence? J. Social Issues 58, 1, 177–193

H. Chen, “Intelligence and Security Informatics for International Security,” Information Sharing and Data Mining, Springer, 2006.

Hughey, Douglas (2009). Comparing Traditional Systems Analysis and Design with Agile Methodologies. University of Missouri – St. Louis. Retrieved 11 August 2014.

Juliet, N.N. (2014). Internet Freedom in Kenya: Balancing Hate Speech and Free: Launch of the Internet Freedoms in East Africa 2014, 4(1), 21-25.

Kenya Parliament (2014). Article 19. Kenya CyberCrime and Computer Related Crimes bill.

Larman, Craig (2004). Agile and Iterative Development: A Manager's Guide. Addison-Wesley. p. 27. ISBN 978-0-13-111155-4.

Richard Sproat and Steven Bedrick (September 2011). "CS506/606: Txt Nrmlztn". Retrieved October 2, 2012.

Sivic, Josef (April 2009). "Efficient visual search of videos cast as text retrieval" (PDF). IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 31, NO. 4. IEEE. pp. 591–605.

Sutton, J. N. (2009). Social media monitoring and the democratic national convention: New tasks and emergent processes. Journal of Homeland Security and Emergency Management, 6(1), 1-20.

Valkanas, G., Katakis, I., Gunopulos, D. and Stefanidis, A., 2014, August. Mining twitter data with resource constraints. In Proceedings of the 2014 IEEE/WIC/ACM International Joint

Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01 (pp. 157-164). IEEE Computer Society.

Yang, M., Kiang, M., Ku, Y. (2011). Journal of Homeland Security and Emergency Management. Volume 8, Issue 1

Zhao, J., Liu, K. and Xu, L., 2016. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions Bing Liu (University of Illinois at Chicago) Cambridge University Press, 2015, 381 pp.; hardcover, ISBN 9781107017894

Zhou, Y., Reid, E., Qin, J., Chen, H., and Lai, G. 2005. U.S. extremist groups on the Web: Link and content analysis. IEEE Intell. Syst. 20, 5, 44–51

APPENDIX A: USER GUIDE

In order to run the application a user would require to follow the following steps to install the application on a machine

1. System Requirements

i. Windows or Linux Operating system

ii. 64-bit system

iii. RAM 3GB

iv. CPU 2.4 GHz

2. Installation Requirements

i. Apache version >= 2.4

ii. MySQL version >= 5.6

iii. PHP version >= 5.4.16

iv. PHP extensions;

v. Python version >= 2.7.5

3. Installation

i. Install Apache, MySQL and PHP, then install and load the indicated Python extensions on the machine.

ii. Install Python on the machine then install all the indicated Python libraries, on Linux configure the Apache server.

iii. Install the Scrapy Python web crawler and configure the Spiders and the Robots for Web crawling.

iv. Extract and copy the application files to the web root of the Apache server.

v. Create a database for the web crawled data and import the scrabot schema

a. spider: scrabot.sql

vi. Customize the application settings i.e. database and Scrapy web crawler.

vii. To run the application enter the URL http://localhost/scrabot on your browser the following should be displayed, click the Submit button.

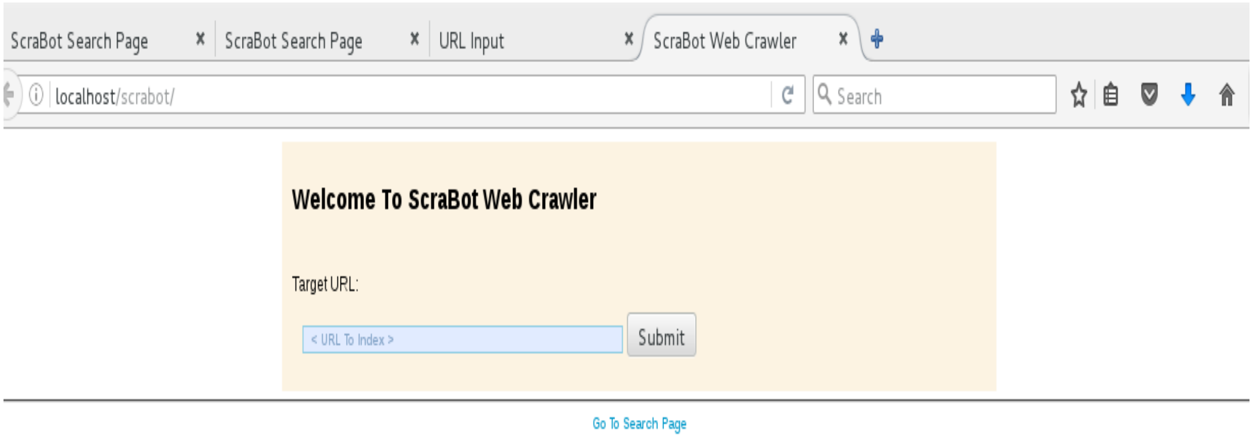

Figure 29: Scrabot Welcome Page for Target URL

To run the web crawling the below mentioned Linux script was run.

```
[root@localhost scrabot]# cat run_scrabot.sh
#!/bin/bash

scrapy runspider scrabot.py -o scrabot.json
[root@localhost scrabot]#
```

Figure 30: Scrabot run startup script

The below script was run which initiates the web crawler.

```
[root@localhost scrabot]# ./run_scrabot.sh
2016-09-04 02:23:58 [scrapy] INFO: Scrapy 1.1.1 started (bot: scrapybot)
2016-09-04 02:23:58 [scrapy] INFO: Overridden settings: {'FEED_FORMAT': 'json', 'FEED_URI': 'scrabot.json'}
2016-09-04 02:23:58 [scrapy] INFO: Enabled extensions:
['scrapy.extensions.feedexport.FeedExporter',
 'scrapy.extensions.logstats.LogStats',
 'scrapy.extensions.corestats.CoreStats']
2016-09-04 02:23:58 [scrapy] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
 'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
 'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.chunked.ChunkedTransferMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
2016-09-04 02:23:58 [scrapy] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2016-09-04 02:23:58 [scrapy] INFO: Enabled item pipelines:
[]
```

Figure 31: Scrabot Startup Script Output

```
2016-09-04 02:23:58 [scrapy] INFO: Spider opened
2016-09-04 02:23:58 [scrapy] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)
2016-09-04 02:24:00 [scrapy] DEBUG: Crawled (200) <GET https://blog.scrapinghub.com> (referer: None)
2016-09-04 02:24:00 [scrapy] DEBUG: Crawled (200) <GET https://blog.scrapinghub.com/category/tools/> (referer: https://blog.scrapinghub.com)
2016-09-04 02:24:01 [scrapy] DEBUG: Scraped from <200 https://blog.scrapinghub.com/category/tools/>
{'title': u'Scrapy + MonkeyLearn: Textual Analysis of Web\xa0Data on'}
2016-09-04 02:24:01 [scrapy] DEBUG: Scraped from <200 https://blog.scrapinghub.com/category/tools/>
{'title': u'Introducing Scrapy Cloud\xa02.0 on'}
2016-09-04 02:24:01 [scrapy] DEBUG: Scraped from <200 https://blog.scrapinghub.com/category/tools/>
{'title': u'Scrapy Tips from the Pros: Part\xa01 on'}
2016-09-04 02:24:01 [scrapy] DEBUG: Scraped from <200 https://blog.scrapinghub.com/category/tools/>
{'title': u'Why we moved to\xa0Slack on'}
2016-09-04 02:24:01 [scrapy] DEBUG: Crawled (404) <GET http://www.example.com/category/keyboards?id=123> (referer: https://blog.scrapinghub.
```

```
{'title': u'Announcing Portia, the Open Source Visual Web\xa0Scraper! on'}
2016-09-04 02:24:02 [scrapy] DEBUG: Scraped from <200 https://blog.scrapinghub.com/category/autoscraping/>
{'title': u'Introducing Dash on'}
2016-09-04 02:24:02 [scrapy] DEBUG: Scraped from <200 https://blog.scrapinghub.com/category/autoscraping/>
{'title': u'Spiders activity graphs on'}
2016-09-04 02:24:02 [scrapy] DEBUG: Scraped from <200 https://blog.scrapinghub.com/category/autoscraping/>
{'title': u'Autoscraping casts a wider\xa0net on'}
2016-09-04 02:24:02 [scrapy] INFO: Closing spider (finished)
2016-09-04 02:24:02 [scrapy] INFO: Stored json feed (80 items) in: scrabot.json
2016-09-04 02:24:02 [scrapy] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 5502,
 'downloader/request_count': 20,
 'downloader/request_method_count/GET': 20,
 'downloader/response_bytes': 224020,
 'downloader/response_count': 20,
 'downloader/response_status_count/200': 18,
 'downloader/response_status_count/404': 2,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2016, 9, 3, 23, 24, 2, 516009),
 'item_scraped_count': 80,
 'log_count/DEBUG': 102,
 'log_count/INFO': 8,
 'request_depth_max': 1,
 'response_received_count': 20,
 'scheduler/dequeued': 20,
 'scheduler/dequeued/memory': 20,
 'scheduler/enqueued': 20,
```

Figure 32: Continued Scrabot Startup Script Output

APPENDIX B: SENTIMENT ANALYSIS CODE

Below is the sample of the Python code used for sentiment analysis

```
[root@localhost html]# cat sentiment.py
import nltk.classify.util
import fileinput
import sys
from nltk.classify import NaiveBayesClassifier
from nltk.corpus import names

def word_feats(words):
   return dict([(word, True) for word in words])

positive_vocab = [ 'fantastic', 'good', 'nice', 'great', ':)' ]
negative_vocab = [ 'bad', 'sexual', 'terrible','useless', 'hate', 'kill' ,'crafty', 'hateful', 'killing', ':(' ]
neutral_vocab = [ 'movie','the','sound','was','is','actors','did','know','words','not' ]

positive_features = [(word_feats(pos), 'pos') for pos in positive_vocab]
negative_features = [(word_feats(neg), 'neg') for neg in negative_vocab]
neutral_features = [(word_feats(neu), 'neu') for neu in neutral_vocab]

train_set = negative_features + positive_features + neutral_features

classifier = NaiveBayesClassifier.train(train_set)

#for line in file:
#     sentence = file.readline()
# Predict
```

```
neg = 0
pos = 0

## Open the file with read only permit
f = open('/var/www/html/result_line.txt')
## Read the first line
#line = f.readline()

## If the file is not empty keep reading line one at a time
## till the file is empty
#for line in f.readlines(): with open("title.txt","r")  as infile:
#    for line in iter(f):
#for line in iter(f):
#        sentence=f.readline(
#        line=f.readline()
sentence=f.readline()

sentence = sentence.lower()
words = sentence.split()
for word in words:
    classResult = classifier.classify( word_feats(word))
    if classResult == 'neg':
        neg = neg + 1
    if classResult == 'pos':
        pos = pos + 1
#fh = open('out.txt', 'w')
print (sentence, 'Positive: ' + str(float(pos)/len(words)) ,'Negative: ' + str(float(neg)/len(words)))
#log = open("out.txt", "w")
#print >> log.close()
```

```
[root@localhost html]#
```