Social Media Forensics for Hate Speech Opinion Mining

George Wafula Wanjala
School of Computing and Informatics,
University of Nairobi,
Kenya

Andrew M. Kahonge
School of Computing and Informatics,
University of Nairobi,
Kenya

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 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.

General Terms

Forensics, Opinion mining, LAMP.

Keywords

Web Forums, Social Media, Linux Apache MySQL PHP Python (LAMP), Hate Speech, Opinion Mining, Digital Forensics.

1. INTRODUCTION

As of May 2015, up to 62% of Kenya’s population had access to the Internet and a mobile tele density of 77% facilitated by three mobile service providers, that is, Safaricom, Airtel and Orange [1]. Social media forms a group of diverse and fast developing communication media which enables the production and sharing of information through collaborative means [2]. Hate speech is a verbal attack targeting a certain race, ethnicity, gender, religion or sexual orientation [3]. The lead up to the Kenya 2012 presidential elections saw the increased use of social media platforms by both citizens and politicians for campaign purposes, civic engagement, and communication, among others. Parallel to this was resurgence of hate speech online. This resulted in the Kenyan 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 [4]. Hate Speech on social media needs to be tracked, tackled and a 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. Social Media Forensics for Hate Speech Opinion Mining relates to the scientific application of cyber forensics tools to social media web forums to extract, identify and document hate speech.

With security agencies and communication authority personnel being overstretched, with less manpower, tools and cyber forensics expertise, the need for an automated and user-friendly system for social media hate speech forensics is crucial.

2. PREVIOUS EFFORTS IN OPINION MINING

Social Media Analytics for Radical Opinion formulated a procedure for radical opinion identification that was divided into three steps which included post collection, feature generation and model generation [5]. In the feature generation, they used F1, F2, F3, and F4 to denote syntactic, stylistic, content-specific, and lexicon features, respectively. Among them, F1, F2, and F3 are from machine learning approaches, and F4 is from a semantic-oriented approach. The third step was model generation, where they compared 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 that the classification model was developed [6].

Further previous work, Affect Analysis of Radical Contents on Web Forums Using SentiWordNet made an attempt towards sentiment analysis. The research was based on tagging and assigning positive and negative scores of each synset in each word [7]. The overall process consisted of data collection of two websites, Montada and Qawem, model building and result analysis. The data collection phase included ripping words from the forums. After that, 500 sentences of each forum were translated manually from Arabic to English for use in the experiment. Model building was written using Python programming language [8]. The model building phase was started by splitting sentences into words and reducing the high-frequency text (stop words) in the sentences. Words were stored in a bag of words (BOW) and part of speech (POS) was used, Lexicon, WordNet and SentiWordNet were used for assigning positive and negative scores of each synset in each word.

From the above previous work they have only shown the opinion mining analysis and generalization, tagging and categorization of social media comments. The research undertaken was to go further and extract the opinion, investigate, identify and flag the individual hate mongers and cyber criminals. It seeks to provide a platform of not only classifying the social media comments as positive or negative but to further go ahead and provide the digital cyber forensics evidence of such comments. The research seeks to provide the answer for social media hate speech opinion mining by providing the tools for digital data acquisition, forensics investigation and presentation of the digital artifacts. It provides a scientific approach for digital social media hate speech investigation and monitoring.

3. METHODOLOGY

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 [9]. It promotes adaptive planning, evolutionary development, early delivery, and continuous improvement, and it encourages rapid and flexible response to change [10]. Agile software development refers to a group of software development methodologies that promotes development iterations, open collaboration, and process adaptability throughout the life-cycle of the project. It chooses to do things in small increments, with minimal planning, rather than plan at length [11]. This methodology was employed for the development of the web interface, database, search query components and sentiment analysis ranking for hate speech. The methodology places emphasis on documentation (such as requirements documents and design documents) as well as source code [12]. The goal was to come up with a software application for social media forensics with respect to hate speech opinion mining. The block diagram below represents the conceptual model of the solution. The solution is an innovation using existing programming language and it is in four modules, namely the web crawler, compiler, database and search platform.

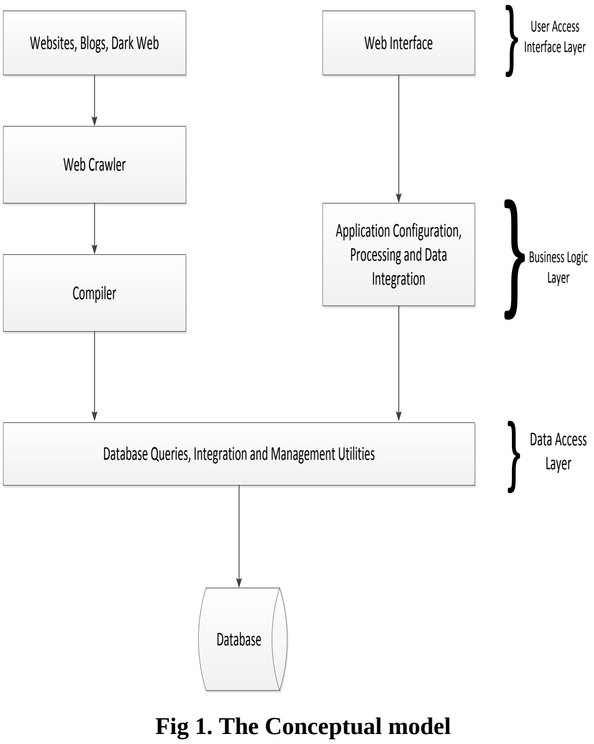

Websites, Blogs, Dark Web
Web Interface
User Access Interface Layer
Web Crawler
Application Configuration, Processing and Data Integration
Business Logic Layer
Compiler
Database Queries, Integration and Management Utilities
Data Access Layer
Database

Fig 1. The Conceptual model

3.1 Data Sources and Collection

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. With the data sources identified a means for data collection, harvesting, cleaning and verification was implemented. Harvesting data from social media is provided for by open source and several APIs which are specific to the social media site [13]. For this research, an open source web crawler was developed to be able to fetch and collect datum from all the data sources. Facebook and Twitter APIs were not used as they are specific, proprietary and limiting in terms of access to data. Python was the programming language that was used to perform web crawling and scraping [14]. 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.

3.2 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 [15]. The data preprocessing is broken down into 3 steps:

- **Tokenization.**

Tokenization involves splitting a string into its desired constituents seeking to isolate as much sentiment information as possible [16]. Tokenization helps in keeping the vocabulary as small as possible. Emoticons and abbreviations are identified as part of the tokenization process and treated as individual tokens.

- **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 [17].

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

A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural' [18].

3.3 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 [19].

This can later be used as a features for Text Classifiers. In this bag-of-words model you only take individual words into account and give each word a specific subjectivity score. This subjectivity score can be looked up in a sentiment lexicon. If the total score is negative the text will be classified as negative and if its positive the text will be classified as positive.

3.4 Algorithm

Natural Language Processing algorithms are many and widely used depending on the size of data and accuracy of performing sentiment analysis. Naïve Bayes usefulness in solving data mining classification problems and its application to various datasets and disparity as results were obtained [20].

Naive Bayes is super simple for 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.

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.

3.5 Evaluation

F1-score is a Machine Learning algorithm and statistical measure of test's accuracy combining both precision and recall [21]. 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:

1. Collect opinions from social media sites and place them in the MySQL Database
2. Identify the keyword set of potential hate speech and train the Natural language classifier.
3. Check the sentiment analysis in terms of Positive or Negative comment.

3.6 Prototype Validation

The below figure 2 illustrates the general design of the components of the system

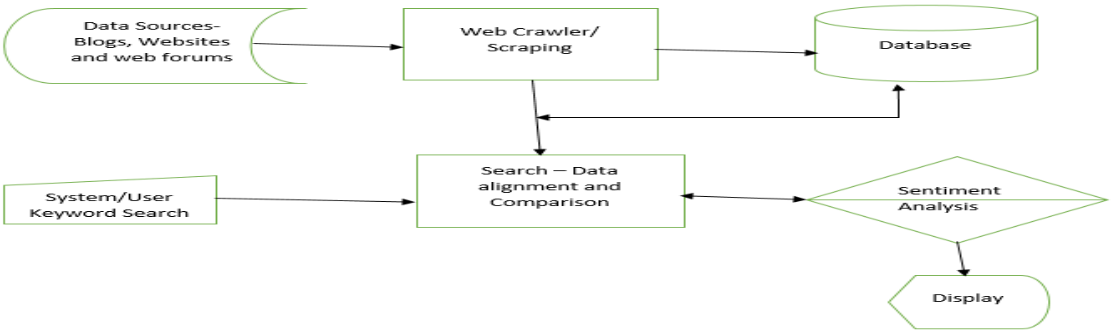

Data Sources- Blogs, Websites and web forums
Web Crawler/ Scraping
Database
Search – Data alignment and Comparison
System/User Keyword Search
Sentiment Analysis
Display

Fig 2. Prototype Elements of the Design

3.7 Web Crawler

The web crawler was designed in Python together with Scrapy, a Python framework for Web Crawling and Scraping [22]. The web crawler was designed to receive the URLs of the websites and blogs and retrieve data through web crawling and scraping. The application allowed indexing of the crawled and scraped websites or blogs or web forums. The web crawler is a web spidering program that collects the website URL, title and webpage body and sends it to the complier for presentation and processing.

3.8 Compiler

The compiler is the central processing unit of the platform and forms the core bond between the crawler and the database. It was developed using a combination of PHP and Python. It seeks to understand the data crawled from the website, performs data cleaning and integrates it to the database.

It should be able to:

- Understand the data scraped and crawled from the website.
- Cleans the data and removes non-ASCII characters together with the removal of html tags,
- stripping text of excess spaces and removal of stop words.
- Polls the data and aligns it for entry into the database.
- Perform sentiment analysis tags a URL as hate speech

3.9 Database

The MySQL relational database could store the indexed data from the scraped website, poll, relay and retrieve the search data with respective to the keyword used. The database was to act as a point of live acquisition of the digital data from the potential hate speech website. It's from the database where we perform the digital forensics of the the digital evidence collected from the scraped website.

The data on the database is presented on web interface with the title of the web forum, the full URL, ten kilobytes of the body of the website and the timestamp.

The search button was designed in PHP to perform keyword check on the MySQL database and provide the instances of the potential hate speech data. It would provide the web forum URL and a snippet of the content in the URL. The presentation of the data on a table would allow the cyber forensics investigator to dig deeper, to collect evidence in relations to the forensics investigation

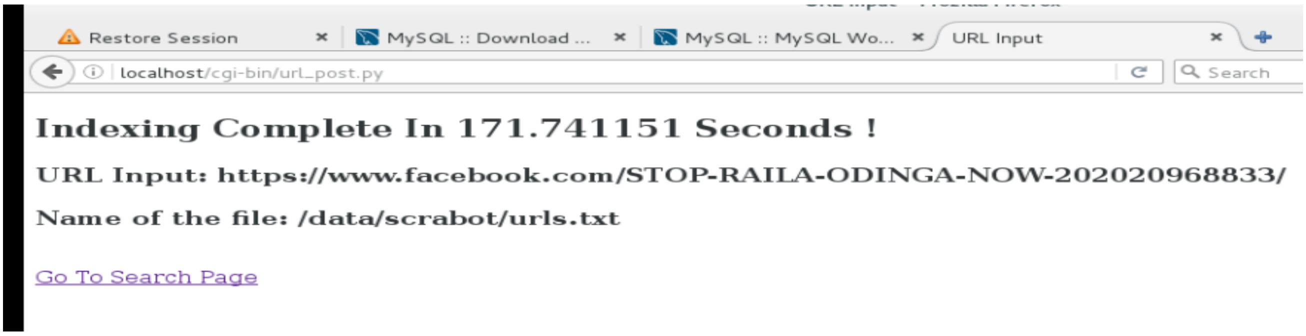
3.10 Non-Functional Requirements

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

- Administrator's graphical user interface should be easy to use and navigate.
- Performance requirements:

1) The system should have short response time during web crawling
2) Short data access time when keywords are being searched on the database.

- Operating system constraints: The application should not exhaust available computing resources or limit the behavior of the system hardware, memory and processor.
- The application needs at least the following 1GB ram, CPU 2,4GHz, Linux Centos 64bit system.



Indexing Complete In 171.741151 Seconds !

URL Input: https://www.facebook.com/STOP-RAILA-ODINGA-NOW-202020968833/

Name of the file: /data/scrabot/urls.txt

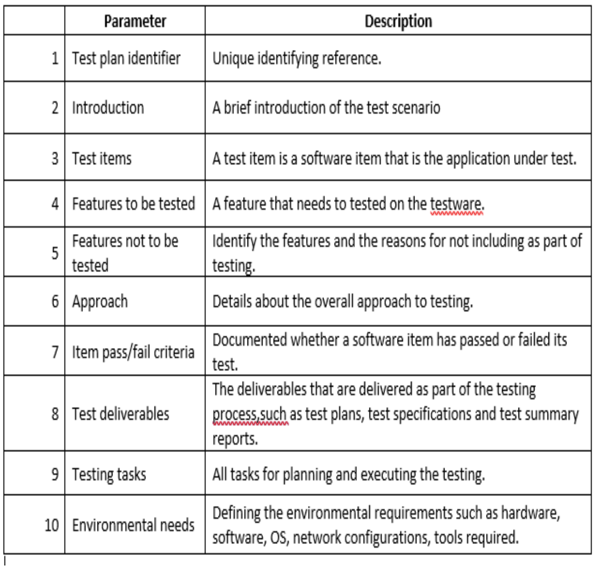
Go To Search Page

Fig 3. Completion of Index Process

4. TEST MANAGEMENT PLAN

Test management most commonly refers to the activity of managing the computer software testing process. A test management tool is software used to manage tests (automated or manual) that have been previously specified by a test procedure [23]. It is often associated with automation software. Test management tools often include requirement and/or specification management modules that allow automatic generation of the requirement test matrix (RTM), which is one of the main metrics to indicate functional coverage of a system under test (SUT) [24].

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



	Parameter	Description
1	Test plan identifier	Unique identifying reference.
2	Introduction	A brief introduction of the test scenario
3	Test items	A test item is a software item that is the application under test.
4	Features to be tested	A feature that needs to tested on the testware.
5	Features not to be tested	Identify the features and the reasons for not including as part of testing.
6	Approach	Details about the overall approach to testing.
7	Item pass/fail criteria	Documented whether a software item has passed or failed its test.
8	Test deliverables	The deliverables that are delivered as part of the testing process,such as test plans, test specifications and test summary reports.
9	Testing tasks	All tasks for planning and executing the testing.
10	Environmental needs	Defining the environmental requirements such as hardware, software, OS, network configurations, tools required.

Fig 4. Components for testing the Application

Test steps describe the execution steps and expected results that are documented against each one of those steps. Each step is marked pass or fail based on the comparison result between the expected and actual outcome [25]. While developing the test cases, we usually have the following fields:

- Test Scenario
- Test Steps
- Parameters
- Expected Result
- Actual Result

5. GENERAL PROCEDURE

The general procedure followed by the system is as depicted in figure 4. The website URL is entered on the home page of the web crawler. Once the URL is entered the application does an Internet search of the URL and performs web crawling and scraping. As the data is collected from the potential web forums its indexed and then placed in the MySQL database. The crawler sends the data to the compiler which contains algorithms for sentiment analysis and MySQL database scripts for storing the web crawled data. The completion of the indexing process is as shown on Fig 3 above.

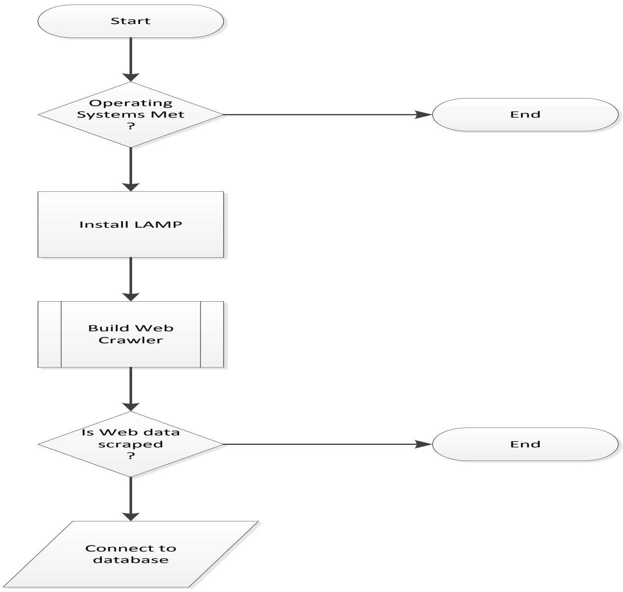
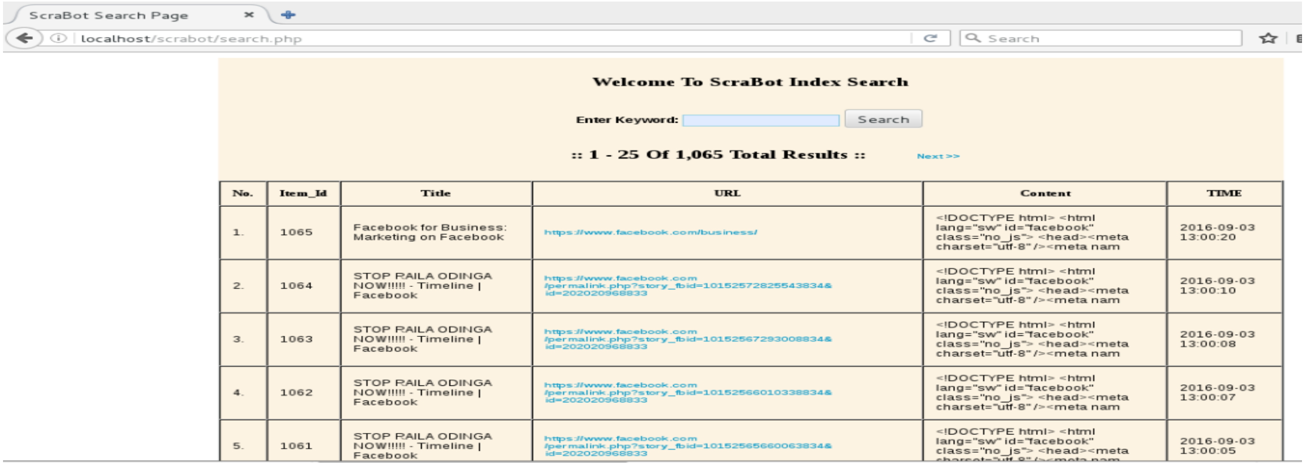


Fig 5. Flowchart System Design



Welcome To ScraBot Index Search

Enter Keyword: Search

:: 1 - 25 Of 1,065 Total Results :: Next >>

No.	Item_Id	Title	URL	Content	TIME
1.	1065	Facebook for Business: Marketing on Facebook	https://www.facebook.com/business/	<!DOCTYPE html> <html lang="sw" id="facebook" class="no_js"> <head><meta charset="utf-8" /><meta nam	2016-09-03 13:00:20
2.	1064	STOP RAILA ODINGA NOW!!!!! - Timeline | Facebook	https://www.facebook.com /permalink.php?story_fbid=10152572825543834& id=202020968833	<!DOCTYPE html> <html lang="sw" id="facebook" class="no_js"> <head><meta charset="utf-8" /><meta nam	2016-09-03 13:00:10
3.	1063	STOP RAILA ODINGA NOW!!!!! - Timeline | Facebook	https://www.facebook.com /permalink.php?story_fbid=10152567293008834& id=202020968833	<!DOCTYPE html> <html lang="sw" id="facebook" class="no_js"> <head><meta charset="utf-8" /><meta nam	2016-09-03 13:00:08
4.	1062	STOP RAILA ODINGA NOW!!!!! - Timeline | Facebook	https://www.facebook.com /permalink.php?story_fbid=10152566010338834& id=202020968833	<!DOCTYPE html> <html lang="sw" id="facebook" class="no_js"> <head><meta charset="utf-8" /><meta nam	2016-09-03 13:00:07
5.	1061	STOP RAILA ODINGA NOW!!!!! - Timeline | Facebook	https://www.facebook.com /permalink.php?story_fbid=10152565660063834& id=202020968833	<!DOCTYPE html> <html lang="sw" id="facebook" class="no_js"> <head><meta	2016-09-03 13:00:05

Fig 6. Search Process

With the data of the potential web forum scraped, crawled, cleaned and stored in the database, this allows the opportunity to perform digital forensic investigation. The search box provides and algorithm for querying the database with potential hate speech keywords as shown in Fig 6. It provides the total count of instances of person of interest and potential hate speech.

6. RESULTS

After full implementation and testing of the system, evaluation of the application was done with the aim to determine if the developed system is delivering the expected results. The following areas were evaluated to provide answers to the research questions set at the feasibility study of the project, which are in line with the project objectives and requirements.

6.1 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.

Table 1. Showing the Sentiment Analysis based on being Positive or Negative

Title	Link	Positive Sentiment	Negative Sentiment
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

6.2 Potential Hate Speech Facebook Group

To ascertain this, sample data of the web forums and the amount of traffic in regards to Hate speech and subject matter on potential political topics was collected.

The statistics for the crawled web page and the count of the mentioned against a particular topic has been shown below. The frequency of potential keywords were recorded in a table in MS Excel. Using MS SQL server data query and analysis tools together with the search function of the application, the data was analyzed and the following table presents a summary of the statistics taken in this experiment.

Table 2. Summary of the Facebook Group Hate Speech

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

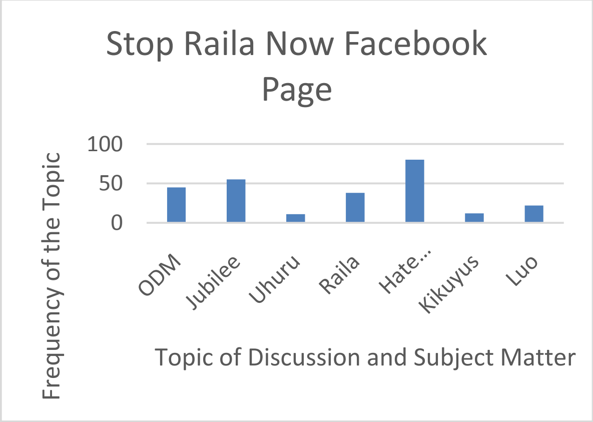


Fig 9. Graph showing the frequency versus the topic being discussed

During the digital forensics investigation, the collected hate speech data on the "Stop Raila Odinga Facebook" page offered evidence of the potential hate mongers. The above graph showed that hate speech and Jubilee, a political party, had a high frequency. This clearly shows that supporters of Jubilee were discussing inflammatory topics against the opposing political party.

6.3 Measuring web forum statistics with respect to the hate speech discussions

In carrying out the digital forensics, the twenty websites which were used as the sample size for hate speech monitoring were plotted against the topic of discussion as a ranking technique. The below graph shows the digital forensics evidence on hate speech collected by the application.

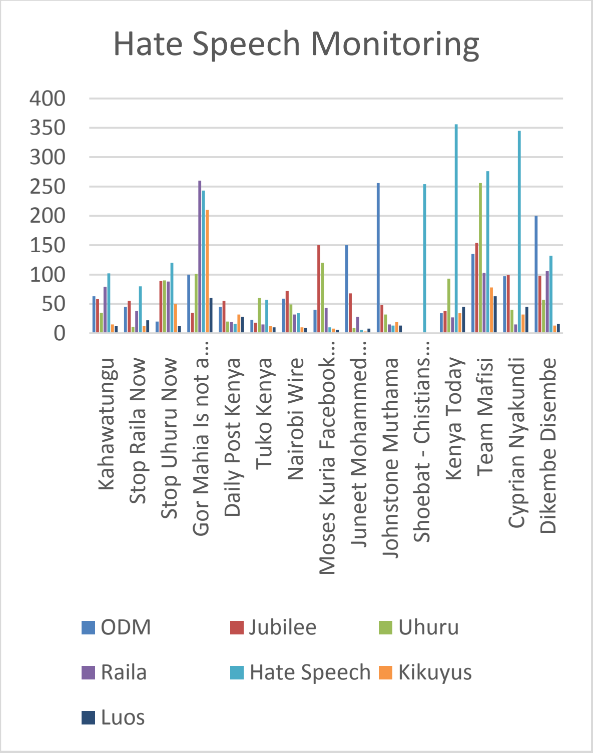


Fig 10. Graph showing the different social media web forums and the topics being discussed

6.4 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.

Table 3. Showing the Web Forums and the Hate Speech Count

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

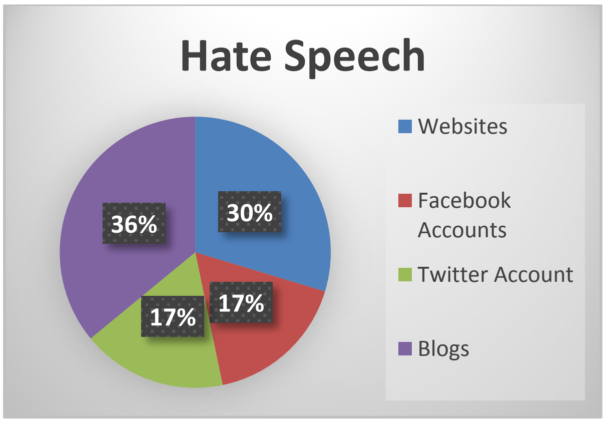


Fig11. Graph showing the different social media web forums

6.5 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.

Table 4. Access Efficiency of the Web Application

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

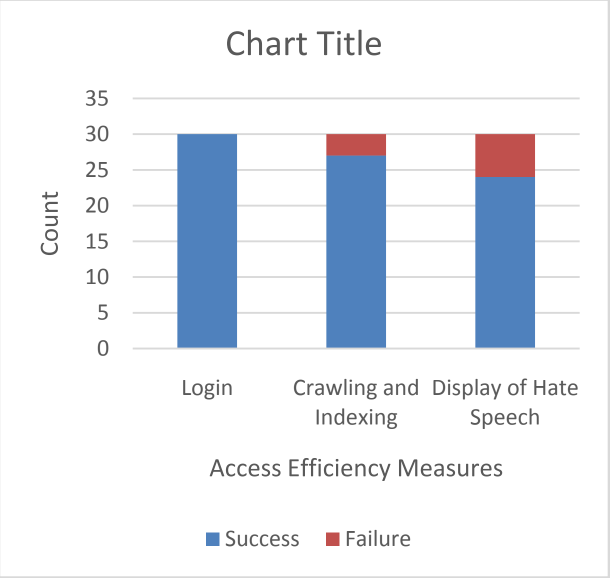


Fig 12. Measure of success for the web database

6.6 Test Management Results

The test management plan was used as a guide in terms of testing the application which included unit testing, usability, use case testing, volume testing, vulnerability and performance setting it. The application was tested with regards to the parameters under each test scenario with a scale of 1 to 10.

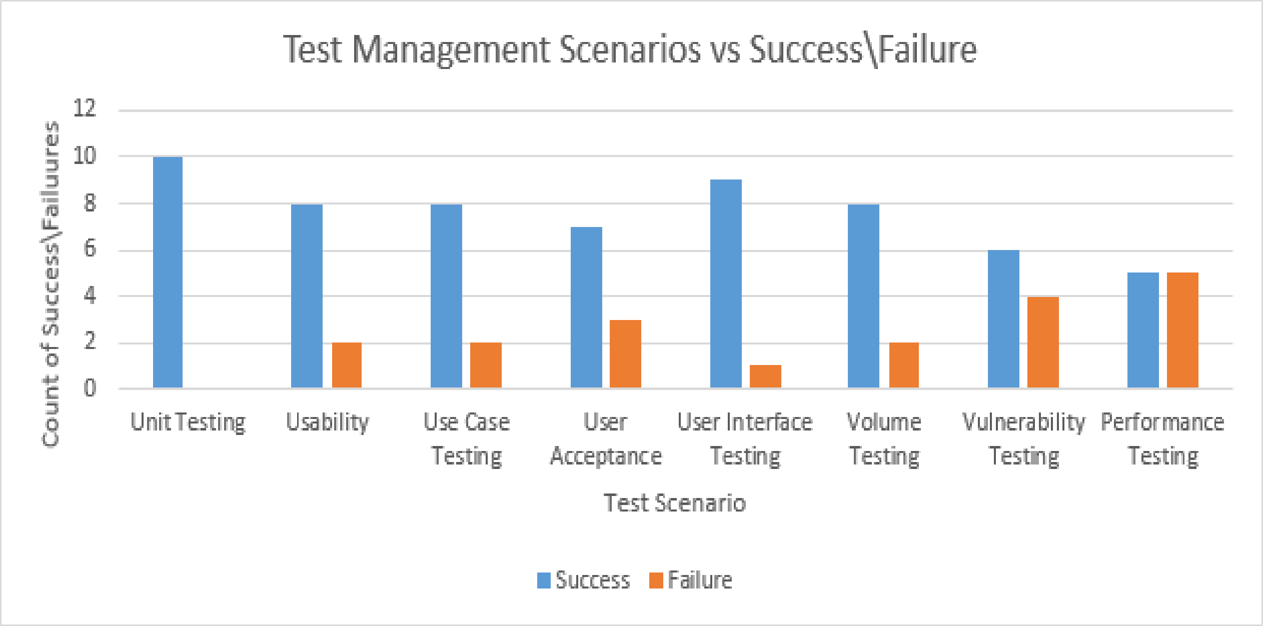


Fig 13: Scenarios of the Test Cases.

7. CONCLUSION

The research project provided an opportunity for performing digital forensics for hate speech on social media web forums. Previous work has shown that machine learning algorithms have been used for filtering, tokenization and categorization of sentiment analysis. This has shown that there exists a knowledge gap as machine learning delves more in grouping rather than individual comments. The software application comes in as a potential solution for hate speech monitoring. It provides an avenue to allow security agencies and communication authority personnel to check possible hate speech web forums.

Results from the forensics data collected verify that social media can be monitored and hate speech collected as forensics digital data. The application has shown how the data is stored and presented as hate speech for use by the forensics investigator. The tool has shown the capability to draw big data associated by Social Media web forums.

From the analysis of the results there was high content of hate speech in blogs as compared to websites and Facebook accounts on social media. This is so because blogs are usually updated on a weekly basis as compared to Facebook and Twitter posts which are generally emotive and hate speech traffic is usually during specific political event.

The ultimate objective is to enable authorized users of the system to be able to access digital evidence. Evaluations carried out to measure the success level of users accessing the hate speech database returned positive results. Legitimate users could login the web page, perform enquiries and display detailed hate speech information instantly be reported to authorized personnel and agencies as and when they occur with all the necessary details to guide the authorities in rescue operations.

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.

Overall, 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.

The application has demonstrated the acquisition of the digital data from the web forums, identification of the potential hate speech using keywords and extraction and documentation of the digital evidence.

7.1 Further 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.

8. REFERENCES

[1] 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.

[2] 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.

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

[4] Kenya Parliament (2014). Article 19. Kenya Cyber Crime and Computer Related Crimes bill.

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

[6] 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.

[7] 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.

[8] Van Rossum, G., 2007, June. Python Programming Language. In USENIX Annual Technical Conference (Vol. 41)

[9] 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?"

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

[11] Sultan, M. A., Salazar, C., & Sumner, T. (2016). Fast and Easy Short Answer Grading with High Accuracy. In Proceedings of NAACL-HLT (pp. 1070-1075).

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

[13] 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.

[14] Shkapenyuk, V. and Suel, T., 2002. Design and implementation of a high-performance distributed web crawler. In Data Engineering, 2002. Proceedings. 18th International Conference on (pp. 357-368). IEEE.

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

[16] Pantone, P., Adding Sentiment Analysis support to the NLTK Python Platform.

[17] Bollen, J., Mao, H. and Pepe, A., 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. ICWSM, 11, pp.450-453.

[18] Pak, A. and Paroubek, P., 2010, May. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREc (Vol. 10, pp. 1320-1326).

[19] Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y. and Potts, C., 2011, June. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 142-150). Association for Computational Linguistics.

[20] Tan, S., Cheng, X., Wang, Y. and Xu, H., 2009, April. Adapting naive bayes to domain adaptation for sentiment

analysis. In European Conference on Information Retrieval (pp. 337-349). Springer Berlin Heidelberg.

[21] Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S. and Androutsopoulos, I., 2015, June. Semeval-2015 task 12: Aspect based sentiment analysis. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado (pp. 486-495).

[22] Smedt, T.D. and Daelemans, W., 2012. Pattern for python. Journal of Machine Learning Research, 13(Jun), pp.2063-2067.

[23] Aurum, A., Jeffery, R., Wohlin, C. and Handzic, M. eds., 2013. Managing software engineering knowledge. Springer Science & Business Media.

[24] Gintis, N., Ixia, 2014. Methods, systems, and computer readable media for providing user interfaces for specification of system under test (sut) and network tap topology and for presenting topology specific test results. U.S. Patent Application 14/452,205.

[25] Shree Divya, S. and Chitra, P., 2014. common automation testing framework using selenium. Journal on Software Engineering, 9(1).