The current issue and full text archive of this journal is available on Emerald Insight at:
www.emeraldinsight.com/2514-9288.htm

Combating the challenges of social media hate speech in a polarized society

A Twitter ego lexalytics approach

Collins Udanor and Chinatu C. Anyanwu
Department of Computer Science, University of Nigeria Nsukka, Nsukka, Nigeria

Received 15 January 2019
Revised 1 May 2019
20 June 2019
Accepted 7 July 2019

Abstract

Purpose – Hate speech in recent times has become a troubling development. It has different meanings to different people in different cultures. The anonymity and ubiquity of the social media provides a breeding ground for hate speech and makes combating it seems like a lost battle. However, what may constitute a hate speech in a cultural or religious neutral society may not be perceived as such in a polarized multi-cultural and multi-religious society like Nigeria. Defining hate speech, therefore, may be contextual. Hate speech in Nigeria may be perceived along ethnic, religious and political boundaries. The purpose of this paper is to check for the presence of hate speech in social media platforms like Twitter, and to what degree is hate speech permissible, if available? It also intends to find out what monitoring mechanisms the social media platforms like Facebook and Twitter have put in place to combat hate speech. Lexalytics is a term coined by the authors from the words lexical analytics for the purpose of opinion mining unstructured texts like tweets.

Design/methodology/approach – This research developed a Python software called polarized opinions sentiment analyzer (POSA), adopting an ego social network analytics technique in which an individual's behavior is mined and described. POSA uses a customized *Python N-Gram* dictionary of local context-based terms that may be considered as hate terms. It then applied the Twitter API to stream tweets from popular and trending Nigerian Twitter handles in politics, ethnicity, religion, social activism, racism, etc., and filtered the tweets against the custom dictionary using unsupervised classification of the texts as either positive or negative sentiments. The outcome is visualized using tables, pie charts and word clouds. A similar implementation was also carried out using R-Studio codes and both results are compared and a t-test was applied to determine if there was a significant difference in the results. The research methodology can be classified as both qualitative and quantitative. Qualitative in terms of data classification, and quantitative in terms of being able to identify the results as either negative or positive from the computation of text to vector.

Findings – The findings from two sets of experiments on POSA and R are as follows: in the first experiment, the POSA software found that the Twitter handles analyzed contained between 33 and 55 percent hate contents, while the R results show hate contents ranging from 38 to 62 percent. Performing a t-test on both positive and negative scores for both POSA and R-studio, results reveal p-values of 0.389 and 0.289, respectively, on an α value of 0.05, implying that there is no significant difference in the results from POSA and R. During the second experiment performed on 11 local handles with 1,207 tweets, the authors deduce as follows: that the percentage of hate contents classified by POSA is 40 percent, while the percentage of hate contents classified by R is 51 percent. That the accuracy of hate speech classification predicted by POSA is 87 percent, while free speech is 86 percent. And the accuracy of hate speech classification predicted by R is 65 percent, while free speech is 74 percent. This study reveals that neither Twitter nor Facebook has an automated monitoring system for hate speech, and no benchmark is set to decide the level of hate contents allowed in a text. The monitoring is rather done by humans whose assessment is usually subjective and sometimes inconsistent.

Research limitations/implications – This study establishes the fact that hate speech is on the increase on social media. It also shows that hate mongers can actually be pinned down, with the contents of their messages. The POSA system can be used as a plug-in by Twitter to detect and stop hate speech on its platform. The study was limited to public Twitter handles only. N-grams are effective features for word-sense disambiguation, but when using N-grams, the feature vector could take on enormous proportions and in turn increasing sparsity of the feature vectors.

Practical implications – The findings of this study show that if urgent measures are not taken to combat hate speech there could be dare consequences, especially in highly polarized societies that are always heated up along religious and ethnic sentiments. On daily basis tempers are flaring in the social media over comments made by participants. This study has also demonstrated that it is possible to implement a technology that can track and terminate hate speech in a micro-blog like Twitter. This can also be extended to other social media platforms.

Social implications – This study will help to promote a more positive society, ensuring the social media is positively utilized to the benefit of mankind.

Data Technologies and Applications
Vol. 53 No. 4, 2019
pp. 501-527
© Emerald Publishing Limited
2514-9288
DOI 10.1108/DTA-01-2019-0007

Originality/value – The findings can be used by social media companies to monitor user behaviors, and pin hate crimes to specific persons. Governments and law enforcement bodies can also use the POSA application to track down hate peddlers.

Keywords Social media, Twitter, Sentiment analysis, Hate speech, Lexalytics, POSA

Paper type Research paper

1. Introduction

The result of the general elections held in Nigeria in 2015 clearly showed that the country was polarized along religious and ethnic boundaries (Udanor *et al.*, 2016), with the majority of the Northern Muslim populace voting for the winner, President Mohammadu Buhari, a Muslim, and the predominantly Christian southern part voting for the former president, Goodluck Jonathan, who lost as the incumbent. Since that time, the polity in the country has been heated up, with many taking to the social media to vent their misgivings, either bothering on political, ethnicity or religious issues.

The ubiquitousness of social media has brought about many challenges that have manifested themselves in a number of variations; hate speech is one example of such challenges (Leondro *et al.*, 2016). According to Jeremy (2012), hate speech may not have a specific definition, but in broad terms, can be seen as any form of activity that has no meaning other than communication that expresses anger for a person or group of individuals bothering on their gender, sex or race. The social media being an avenue where people easily pop-in to make new friends and express their diverse opinions on trending issues all over the world has recently become an avenue where people express their anger and hatred toward other people or the government in power. Sentiments in these media are being expressed in the form of name-calling, insinuations of race or tribal superiority, religious bigotry, abuses, or posting of inciting comments, images and videos, especially on WhatsApp. The one posting the sentiments may call it fighting a cause, while the one who receives it and is offended may call it hate speech. Some of the problems caused by hate speech may include the promotion of violence, discrimination, disintegration, communal wars and ultimately, loss of lives and properties.

Hate speech may be a spoken or written word that is offensive, threatening to an individual or a group based on a particular attribute of the persons being targeted. Hate speech is considered a crime in some countries (Anna and Michael, 2017), for example, in all western Europe where all states currently prohibit various forms of racist, sexist, anti-religious, homophobic or other intolerant speech (The Modern Law Review, 2006). The Penal Code in India also enforces prosecution against hate speech (Noorani, 1992), and in Nigeria (The Nation Nigeria, 2017), etc. Criminalizing hate speech is as a result of it consisting of incitement to an act of encouraging violence, discrimination and hostility. It is used to silence unfavorable opinions and suppress debate. Hate speech is also used as a popular tool to disseminate hatred especially speeches that are focused on religion. It is also used by immoral agents for propaganda based on hate speeches in a form that always becomes viral. Unfortunately, identifying the sources of the messages and holding the users responsible for the act might be difficult. This is because of the decentralized, anonymous and interactive structure of the new media which makes it an ideal platform to spread hate speeches (Aondover, 2018). Hate speech is characterized by the degree of attention given to the content and tone of the expression on social media, such as provocative words (racial or gender), through threats, defamation which includes libel and slander, feelings or attitudes of hatred, obscenity and so on (Jeremy, 2012).

Physical violence sometimes springs up as a result of assimilating hate contents, either spoken or written. One incident of hate speech violence that is still fresh is when a white supremacist Dylan Roof in 2015 was said to have killed nine African Americans at Charleston, just after he was received by the bible study group at a South Carolina Church.

His action was based on the white supremacist propaganda he read and downloaded directly into his brain from the internet. He is currently under a death sentence (Murphy, 2017).

Ethno-religious freedom of speech is being abused on social media platforms. This has prompted concerned authorities to now regulate this freedom of speech on social media sites. There is a need to track and tackle hate speeches on social media and provide a way of apprehending hate speech mongers. The collection and documentation process of hate speech using digital evidence online should be optimized efficiently. There is a need for an automated system to be able to implement such regulations (Levin and McDevitt, 1993; Agarwal and Sureka, 2014). An approach to implementing an automated system that monitors hate speech can be by using natural language processing (NLP) techniques.

Natural language is extremely rich in form and structure, and very ambiguous (Udanor and Bakpo, 2011); one input can mean many different things. Ambiguity can be at different levels. For example at lexical or word level, which implies different meanings for words. The purpose of lexical processing is to determine the meanings of individual words. The basic method is to look up in a database the meanings, that is, its lexicon. So, a lexicon is a language's inventory of lexemes. In sentiment analysis, a lexicon is the disambiguation processes after data extraction and cleaning. After which the processes of feature selection and sentiment analysis can commence. Sentiment analysis is an advanced form of data mining, which targets discovering the polarity of opinions in terms of positive, negative or neutral using NLP techniques. Lexalytics is a term coined by the authors from the words lexical analytics to imply using lexical analytic techniques for the mining of opinions on social media. This study follows a lexicon-based approach to sentiment analysis.

The contributions of this work are as follows.

Monitoring and detecting hate speeches in a social media platform like Twitter, as well as pinpointing who said what by the way of opinion mining. The paper is able to show both from literature survey and experiment that hate speech is on the increase in the social media; that hate speech is a new global challenge that threatens the peaceful coexistence of people in civilized societies, and that hate propagators do not show any form of remorse or repentance. The work can also be a veritable tool for relevant regulating agencies to monitor hate speech propagators, and sanction them accordingly, as the law may regulate. This study applied both interdisciplinary and computational social science approaches. The research methodology can be classified as both qualitative and quantitative. Qualitative in terms of data classification, and quantitative in terms of being able to identify the results as either negative or positive from the computation of text to vector. This research uses unsupervised methods for text analytics or text mining (TM) in unstructured texts like tweets. The rest of the paper is organized as follows: Section 2 takes an in-depth study of related works, in the process, pointing out the techniques to sentiment analysis from previous works, while Section 3 presents the research methods and the experiments. Section 4 discusses the results from the various experiments, while Section 5 lays the background for future works.

2. Review of concepts and literature

The menace of hate speech has prompted much attention in recent times. This is not surprising because hate speech has relatively been unchecked on social media sites for years, until recently when it has brought about many problems such as harassment, physical threats, embarrassments, terrorism, among others. This section begins with an overview of how social networks (SN) work, an understanding of sentiment analysis and common sentiment analysis techniques in use today, and concludes with related works on Twitter hate speech.

2.1 Understanding social network relationships

The SN connects people with friends and family in a social circle to freely express, make new friends and share thoughts or interests. The platform through which people share information online with other groups of people is referred to as social media. This platform has actually changed the way we communicate; get our daily news and much more. The advent of social media has amplified citizen journalism, making everyone a reporter (Cherian, 2018) of most often, and unverified information. The SN is also an articulation of social relationship/connections that is achieved among individuals called nodes (things, individual or people within the network), which are links (relationships or interactions), edges or ties that are connected by one or more types of specific inter-dependency, such as dislike, friendship, mention, retweet, etc.

Social network analytics (SNA), on the other hand, views social relationships as a network theory, which consists of ties and nodes (Udanor *et al.*, 2016). SNA is a method for measurement, representation and analysis of social structure, which seeks to predict the structure of relationships among social entities, as well as the impact of the said structure on other social phenomena (Butts, 2008; Collins). While sentiment analysis helps researchers to determine the opinion of the speakers or writers, SNA analyzes the relationships between people in many fields and activities (Udanor *et al.*, 2016). SNA allows the study of a network topology through the connections that are developed within it, thereby giving rise to a hierarchy of communities within the main topic. Also, SN like Facebook and Twitter allow tracking of the relationships in cases where knowledge is not mutual. That is, a node is a follower of another node and the number of followers defines the popularity of a node within the network in parts, but is not able to point out if this popularity is actually positive or negative.

There are two basic kinds of social network analysis (SNA): ego network analysis of individual nodes, where the individual behavior is mined and described; complete network analysis which is an analysis of all the relationships among a set of nodes. The term “ego” denotes a person in a network who is connected to everyone else (Jeniffer, 2015). Facebook, for example, is built as an ego-centric network where an individual is allowed to discover his or her 1.5-degree network, but it does not allow people to discover their 2.0-degree network. SNA analyzes and measures the techniques of sub-groups like centrality, degree of closeness and betweenness for complete networks. Two nodes are said to be connected with each other if they interact regularly.

Ties are connections among individuals used for sharing information, experience and knowledge. These ties can be strong, weak or latent/absent based on the extent of interaction between the two nodes. Strong and weak ties are relevant to the interaction of social networking. For community formation, these ties (strong and weak) play a different role.

Strong tie (close friends) is referred to as someone within a circle of close friends and family. Weak ties (familiarity) are bridges that connect two different communities (SN). Absent/latent ties are expected existing connection but do not really exist. As SN grow, more attention is given to them. SNA is one of the disciplines that is applying scientific interests to identify patterns, locate, examine and analyze the structures of social entities and the variety of theories for explaining the pattern of the structure that arose from a long-standing practice (Stanley and Katherine, 1994).

In SN, we can encode the network data into low-dimensional representations. This is known as network embedding, whose purpose it is to ensure that the network topology structure and other attribute information are preserved. Network embedding can improve the performance of real-world applications such as in node classification and entity retrieval (Lizi *et al.*, 2017). In addition to applying network embedding, models for learning network representation can be classified (Tan *et al.*, 2019) into three sub-groups, such as representation look-up tables, auto-encoders and graph convolutional networks. Attributed network embedding is an attempt at learning representation for social actors, by preserving both the structural and attribute proximities of the nodes (Lizi *et al.*, 2017). Network embedding can also be in the form of RNN distributed representation learning embeddings for nodes in an SNA by applying skip-gram context for words in NLP (Vu and Parker, 2015). The purpose of this is based on the connectivity of nodes in the network and their attributes; they will be able to learn the embeddings for the nodes.

Graph embedding (Compagnon and Ollivier, 2017) is another form of network embedding method of SNA which uses the vector representation of nodes to build a bridge gap between SNA, data analytics and statistics. These vectors can be used as machine learning inputs when performing prediction, classification or clustering. Community member retrieval on social media using textual information (Aaron *et al.*, 2018) deals with users being able to find other users through social connections. A small set of account selected can form a query for finding other members of the community with similar interests.

This research work applies ego SNA techniques to detect social media abuses manifesting in the form of hate speeches, categorizes them, as well as identify the individuals responsible for spreading such sentiments. The Twitter API is flittered to extract tweets and using N-gram dictionary categorizes the tweets as positive or negative sentiment tweets.

2.2 Sentiment analysis techniques

Sentiment analysis is the analysis of opinion (such as like/dislike/mixed/do not know), feelings (such as happy/sad/angry), satisfaction rating using NLP and text analytics to determine, extract and analyze online social conversations and to determine the deep context as they apply to a topic.

Sentiment analysis has advanced the progress made in TM by bringing on board powerful algorithms and computational linguistics NLP techniques to extract information, meanings and opinions from texts such as unstructured social media data. Tweets are not considered “normal” pieces of text since the 140-character threshold imposes limitations in the length of words and phrases (Kontopoulos *et al.*, 2013). Sentiment refers to emotional effects and feelings a reader associates with the text or contents received through his sensory system. Sentiment analysis is being applied to customers product reviews on social media and the internet (Serrano-Guerrero *et al.*, 2015), politics (Oliveira *et al.*, 2016), education, events (Burnap *et al.*, 2014), shows, etc. Sentiment analysis also comes handy when trying to detect hate speech on social media. Sentiment can be perceived as positive, negative or neutral depending on the words expressed within a given post on a particular subject. Understanding the words and classifying sentiments into its categories is necessary to be able to perform sentiment analysis. Sentiment analysis lies between NLP and natural language understanding. There are challenges encountered in sentiment analysis: it requires a deep understanding of the explicit and implicit, regular and irregular, and syntactical and semantic language rules. And the following problems still remain unresolved: co-reference resolution, negation handling, anaphora resolution, named-entity recognition and word-sense disambiguation (Cambria *et al.*, 2013). Sentiment can be classified at various levels (The Nation Nigeria, 2017) such as text level, sentence level (involves two steps: subjectivity classification and sentiment classification and feature level classification). Sentiment analysis and opinion mining are techniques that can be used to extract information based on one’s own opinions from text documents. To find the overall contextual polarity in any provided text, sentiment analysis is used (Zia *et al.*, 2017).

Sentiment analysis can be broadly categorized into two methods: machine learning and linguistic methods (Yuan, 2017). Machine learning techniques, however, are more popular and some of the common supervised learning classification algorithms include support vector machine (SVM), Decision Trees, Neural Networks, etc. Recently, deep learning algorithm like recurrent neural network (RNN) has been applied to sentiment analysis.

Machine learning techniques require a huge amount of data for the training the model. This section examines a number of algorithms for hate speech sentiment analysis.

Supervised classification methods were applied on data sets (Malmasi and Zampieri, 2017; Liu and Zhang, 2012) using words N-grams, character N-grams and word skip grams as features in the system. In addition to using classification algorithms for hate speech detection, Williams *et al.* (2017) combined rule-based, probabilistic and spatial based classifier to find the best features for hate speech. Naïve Bayes classifier algorithm was applied to detect hateful speech on tweets (Okeyo *et al.*, 2018) using N-grams for documents structuring. In Peter and Leighton (2014/2017), supervised learning algorithms such as Naive Bayes (NB), SVM and k-Nearest Neighbors (kNN) were applied to hate speech detection. They first categorized opinions on tweets by applying SVM, NB and kNN and then their sentiment polarity was found. Same algorithms were applied in Okeyo *et al.* (2018) for N-gram text classification, who observed that Naïve Bayes model makes a strong assumption that the words in a text are independent, but these assumptions were violated clearly in natural language text. Saif *et al.* (2015, 2016) developed an approach to Twitter sentiment analysis called sentiCircle which takes repeated words and capture their meaning, updates their pre-assigned polarity and strength in sentiments lexicons on a different context in tweets accordingly. Burnap *et al.* (2014), in investigating the Woolwich attack in London 2013, used the zero-truncated negative binomial regression method to predict the survival of a tweet and identify emotive content specific to negative content within particular contexts, including racial and religious hate-related events. Featured reduction algorithms that use context-based knowledge and statistical knowledge (Tofighy and Fakhrahmad, 2018) as well as N-grams and statistical analysis to develop a Twitter-specific lexicon for sentiment analysis (Ghiassi *et al.*, 2013) are among the techniques in use.

2.3 Review of related literature

A report (Levin and McDevitt, 1993) has shown that 60 percent of hate crimes come from the largest groups of SN users, the youths. A study from the Pew Research Centre shows that 41 percent of American adults have experienced some form of harassment online. In the context of his study, Barthel *et al.* (2016) defined hate speech as online harassment that is strongly present in content areas such as race, ethnicity, religion and much more. Jeremy (2012) argued that the harm in hate speech is mainly a result of written speech rather than spoken speech, while Walker (1994) stated that hate speech lies within the freedom of expression, liberty, equality group right and dignity. They added that any objective definition of hate speech in a computer program that can be easily implemented can be contested. A lot of hate promoters' group use well known SN websites to spread extremist contents among their followers (Agarwal and Sureka, 2014). Hate message/information is on the increase and more people are joining, Angela *et al.* (2011) observed after examining issues on free speech that were posted by people on Facebook and users' attitudes toward delicate messages. A sample of such examination was carried out (Iftikhar *et al.*, 2016) on 200 Facebook users picked randomly from five Indian state terrorists. They collected data through questionnaires and analyzed them using the Kolmogorov–Smirnov *Z*-test. After examining issues on free speech on Facebook posted by people and their attitudes toward delicate/sensitive tweets, the authors found out that hate speech/information is on the increase and more people are still joining.

Most countries are suffering from the effects of hate speech, which Nigeria is not an exemption. Hate speech has led to violence and loss of lives in the country. Yemi Osinbajo, the Vice President of Nigeria, stated that hate speech is a terrorism spice. He proclaimed hate speech a crime at the National Security Conference in Aso Villa, Abuja. He further stated that in an attempt to control hate speech, the Federal Government has drawn a battle line on hate speech (The Nation Nigeria, 2017). Alakali *et al.* (2017) found out that social media users in Nigeria are aware that promoting hate speech and foul language has moral and legal implications in Nigeria, but they do not know what the punishments are. Joel (2012) is of the opinion that ethno-religious hate speech in Nigeria is a cumulative outcome of the country's history, it is over 250 vast cultural, and religious practices, fueled by the political landscape of the country which profits from these already existing diversities. The author lamented that, unfortunately, there are no legal frameworks in the country to enforce the ban on hate speech. Inflammatory hate speech catalyzes mass killings including genocide (Benesch, 2014). Benesch cited hate speech as being the major contributor to the genocide of Rwanda in 1994, and also predicted Nigeria as one of the countries at the risk of collective violence. Reddy (2002) showed how homophobia and hate speech threaten democracy and human rights in Africa through the context of language and gender. Maina (2010) traced the history of political violence in Kenya, noting that Kenyan politics is built along ethnic lines; therefore, hate speech against another tribe was what fueled the violence of the 1990s, and the 2008 mass killings, respectively. He further stated that technologies such as SMS, e-mail and blogging have clearly contributed to the "democratization" of hate speech and made it easier to spread without accountability.

Due to various political, cultural and historical reasons, hate speech has been a long-standing problem in the world. The less advanced form of hate speech as argued by Walker (1994) has led to the growth of more sophisticated reading of symbols and signs which are hard to examine. He demonstrated how these symbolic codes are historically and culturally imbued with meaning by using Nazi symbols appropriation and development as an example. He further argued that hate can be expressed and experienced symbolically.

Hate and racism in most SN are related. There have been several types of research/studies that have been investigating some suspected social networking sites (Zhou *et al.*, 2005). The networks were analyzed systematically by racism group and found out that these sites have a decentralized structure. A software was used by the authors to automate the analysis of the content of hate mongers' websites and their links on hate speech. They observed that the main objectives of these websites were to share ideas.

A review of online arguments and an overview of the USA currently for and against hate speech, internet regulations and important jurisdiction were presented in Ring (2013), where an attempt to find solutions to limit hateful speeches on social media was made. YouTube was used as an example to explore the ability and type of problem. They examined and evaluated the approaches that were proposed by other communication and legal scholars on whether or how to reduce online hate speech, and sought solutions to reduce online hate speech. Their results were reported to have reduced hate speech on SN site.

Ethical Journalism Network in several articles investigated the effect of hate speech on journalism and lamented that hate speech is a dilemma to journalists all over the world. According to one of the articles (Cherian, 2018), some journalists may ignore their operational procedures and amplify the voices of hate mongers, use their media outlets as ideological spokesmen and cheerleaders for forces of hate, from xenophobia to religious extremists. He further warns that the definition of hate speech may sometimes be fuzzy that it may lead to unwarranted censorship. The author further stated that there are vital distinctions to be made among the following examples in defining hate speech:

- incitement to cause harm such as negative discrimination and violence;
- expressions that hurt a community's feelings, including by insulting beliefs; and
- criticism of politicians and other powerful interests, exposing them to contempt.

While the first can be seen as hate speech, the second raises ethical issues that need addressing, the third may be perceived as hate speech by the target elites like the police, military and politicians who would not want social criticism.

In some quarters, politicians were blamed for heightening hate speeches. For example, it was reported that (Cherian, 2018) race hate and religious abuse incidents rose to 41 percent the month after UK voted to quit the EU which led the Equality and Human Rights Commission to say that the politicians have polarized the country and legitimized hate.

To discover knowledge from social media data, especially Twitter, a lot of work is being done with the aid of sentiment analysis and opinion mining. Apart from tackling the hate speech menace, a good number of works have implemented Twitter data sentiment analysis, some for brand marketing, others for health monitoring, etc. Yoon *et al.* (2013) investigated the possibility of using Twitter TM and sentiment analysis methods to analyze physical activity, especially as it relates to health. They are of the opinion that related tweets enhanced the understanding of physical activity behaviors and their associated situational contexts, and this could be used as an alternative to traditional self-reports. This work is similar to the proposed work on the use of Twitter data but differs in purpose and in the application of the result. While they focused on sentiment analysis for physical health, the proposed work focuses on combating hate speech for psychological well-being. Brand-related tweets have also been retrieved and analyzed. For example, Ghiassi *et al.* (2013) introduced an approach to supervised feature reduction using N-grams and statistical analysis to develop a Twitter-specific lexicon for sentiment analysis of a brand to recognize emerging issues with their brand identity. The proposed work differs in terms of unsupervised approach but adopts the same N-grams text analytics for Twitter data. In addition to the two main approaches of lexicon-based and the machine learning-based (Yuan, 2017) employed in opinion mining, Kontopoulos *et al.* (2013) proposed an ontology-based technique to Twitter-based sentiment analysis that receives a sentiment grade for each distinct post, instead of a sentiment score as in the former approaches. They stressed the difficulty of opinion mining from tweets since tweets are filled with dynamic jargons. The proposed work varies in approach to the implementation of the former, even though both use unstructured twitter data, and instead of sentiment grades, our system uses vectorized sentiment scores. Besides Twitter opinion mining, the use of emotiblog to opinion mining is reported in Boldrini *et al.* (2012).

Gautam and Yadav (2014) used the Twitter data set, labeled the data sets using the uni-gram feature extraction technique and performed supervised learning sentiment analysis using classification algorithms, namely, NB, maximum entropy and SVM algorithms together with semantic-based WordNet. This differs from the proposed system in the approach of supervised learning as opposed to our unsupervised approach. Chen *et al.* (2015) were of the opinion that the more types and larger scale the emotional data are, the higher-accuracy the emotional analysis results will be. They, therefore, developed cloud-based wearable devices that detect emotional changes in the wearer. Once emotional changes have been detected, the most related wearable devices will be activated for data collection. The limitation of this approach is in not having enough devices to deploy, and in the energy consumption of the devices. Pang and Lee (2008) through a survey focused on methods that address the challenges of sentiment-aware applications, as opposed to the traditional systems that were fact based. They considered the questions of privacy, vulnerability to manipulation and whether or not reviews can have a measurable economic impact. Rigas *et al.* (2007) presented a fully automated bio signal based, user-independent emotion recognition method to detect fear, disgust and happiness, and used kNN and Random Forest to classify the emotions. The drawback to this approach is in the use of sophisticated equipment, which will require specialized expertise. It is not easily reproducible. The works of both Pang and Lee (2008) and Rigas *et al.* (2007) differ from the proposed work in the use of specialized hardware equipment. Similarly, Ekman (1999) against his earlier believe proposed automatic emotional processing systems which tend to use the minimal stimulus to activate an emotional response control system, and whose reaction must be with the highest speed. Dragoni *et al.* (2018) developed an application they called OntoSenticNet, which, according to them, uses a common sense ontology for sentiment analysis based on SenticNet, a semantic network which uses primitive concepts.

The challenges posed in performing sentiment analysis on short sentences like tweets stem from its insufficient context information. Dos Santos and Gatti (2014) proposed solving this problem using a character to sentence a convolutional neural network that extracts relevant features from words and sentences. Tang *et al.* (2014) pointed out that there is a problem with word embedding learning through a traditional neural network approach. To this, they proposed another approach to Twitter sentiment classification by using continuous word representation as features under supervised learning. The proposed system differs from the former in the implementation approach of using unsupervised learning from N-gram bag of words. Wiegand *et al.* (2010), after a survey on the role of negation in sentiment analysis, are of the opinion that negation modeling in sentiment analysis needs to be taken into consideration. Cambria *et al.* (2013) traced the evolution of opinion mining, pointing out that there is a progression from Coarse- to Fine-Grained Analysis, from Heuristics to Discourse Structure and from Keywords to Concepts.

Unlike the above works which focused on different applications of sentiment analysis using Twitter data, the following works are focused specifically on approaches to combating hate speech on the Twitter platform.

In determining the presence of hate speech on race, ethnicity or religious tweets, supervised machine learning and statistical model classifiers can aid policy and decision makers to monitor and forecast the likelihood of cyber hate spread (Burnap and Williams, 2015). The use of deep learning (Badjatiya *et al.*, 2017) to learn semantic word embeddings can be a way to handle the complexity in natural languages in order to classify a tweet as either racist or sexist. Using crowdsourcing, Waseem (2016) compared racism and sexism hate classification results of expert and amateur annotations, respectively, and observed that systems trained on expert annotator performed better than those of amateurs. An expert annotator is one who has both a theoretical and applied knowledge of hate speech. The retweet feature of Twitter can amplify hate speech. Using supervised learning, Kwok and Wang (2013) classified anti-black hate speech with big-rams and word-sense disambiguation. Four convolutional neural network models (Gambäck and Sikdar, 2017) employing word2vec, randomly generated word vectors and word vectors combined with character N-grams to determine the presence of racism or sexism or none in a tweet were implemented. Sometimes hate speech is expressed in the form of symbols, so it is necessary to analyze extra-linguistic features in conjunction with character N-grams for hate speech (Waseem and Hovy, 2016). There is the challenge of being able to distinguish between hate speech and other forms of offensive languages. Lexical detection has low precision (Davidson *et al.*, 2017) in that regard and supervised learning has failed to distinguish between the two. Lexical means are effective in identifying offensive terms but not so in identifying potential hate terms.

All the works on Twitter hate speech detection so far reviewed employed supervised learning methodology, apart from Gambäck and Sikdar (2017), who used unsupervised as one out of the other three supervised methods they used. Incidentally, they found that the unsupervised word2vec model without character N-grams approach outperformed the rest. The proposed work employs the unsupervised approach to sentiment analysis of terms extracted from tweets, using a customized dictionary of local context terms, in addition to the terms existing in the in-built English N-gram dictionary in Python.

3. Materials and methods

This research work developed a system called polarized opinions sentiment analyzer (POSA), which applies ego SNA techniques to detect social media abuses manifesting in the form of hate speeches, categorizing them, as well as identifying the individuals responsible for spreading such sentiments. The Twitter API is filtered to extract tweets and using N-gram dictionary classifies the tweets as positive or negative sentiment tweets. A similar implementation was also done using R-Studio and both results are compared. The following sub-sections describe the methods. The research employs unsupervised classification and adopts both qualitative and quantitative methodologies. Qualitative in terms of data classification, and quantitative in terms of being able to identify the results as either negative or positive from the computation of text to vector.

3.1 Social network analytics: a review of methodologies for analyzing hate speech

In order to examine the activities of hate speech, it is necessary to find an effective and efficient way of identifying hate speeches, extract text and learn about their relationships. The SN has provided a set of methods that will be used to analyze patterns of the structure of social entities. Due to the growing volume of social media contents, hate speech is also increasing alongside. Recently, sentiment analysis has attracted lots of attention on Twitter because of its wide application on public and commercial sectors (Saif *et al.*, 2015, 2016). The fundamental concept of network analysis and a range of methods that are currently used were reviewed, as well as issues associated with a collection of data, network analysis, and individual level comparison and analysis were also discussed.

Several types of SN were enumerated by Pokorny *et al.* (2012) as simple SNs, temporal SNs, large-scale SNs, virtual SNs, multi-layered SNs, multidimensional SNs, heterogeneous SNs (with two or many different types of nodes) and much more. They also stated that some SN have their own methods and measures developed. But the suitable analytical method for the above SNs is SNA. Further stated, SNA may be distinguished in several real application domains which include nodes classification for marketing purposes, organizational structural evaluation compared with communication structures in companies, hidden knowledge acquisition recommended systems for supporting users in Web 2.0, web forum analysis of social groups and, finally, their evolution prediction. Pokorny *et al.* (2012) extracted social communities and used appropriate algorithms and structural measures (SNA) for their statistics, calculation and data mining methods in their work.

SNA is the measuring and drawing of various characteristics of the distribution patterns of relational ties. It provides both mathematical and visual analyses of human relationships to represent the description of networks systematically. According to Mirigxin (2010), SNA is a method for examining social structures and that is why it is referred to as structural analysis. Also, SNA is one of the methods used to identify, locate entity, examine and analyze the network dynamics of these structures which can be presented either as a multigraph or graph.

To examine and analyze the usefulness of information from huge data on the SN, special graph-based mining tools will be required to easily model the structure of the SN. A number of such analysis features, tools and benefits were listed by Akhtar (2015). His work also presented comparative network analysis tools such as network Gephi (for interactive visualization and exploration of networks), Pajek (software for drawing networks, with analytic capabilities, etc.), IGraph (for creating and manipulating graph, which is based on the platform, algorithms complexity, execution time, Graph types, features of the graph and the input file format).

NLP is another approach that is readily used for SNA, especially in sentiment identification and classification. The increase in hate speech on SN motivated Anna and Michael (2017) to present a survey on key approaches that can recognize these speeches automatically using NLP and the limits to those approaches. They focused on outlining existing approaches in a systematic manner (character-level, the token-level approach) and extraction of features (such as simple surface features, word generation, sentiment analysis, lexical resource, knowledge-based features and much more). These tasks (sentiment analysis and lexical resource) were seen as supervised learning problems and they stated that a supervised learning approach should be applied in classification and utilization for hate speech detection. They also argued for a benchmark data set for the detection of hate speech for better comparability of different features and methods.

Fernando (2016) classified messages containing hate or violent speeches in a very precise way after an attack in 2015 against Charlie Hebdo in Paris. An important public reaction was generated after the disruptive events in SN, which led to the craving for the study of hate message and violent communication on Twitter. The author carried out a qualitative analysis using data mining method to classify the type of speech. Finally, the author observed that the disruptive event was as a result of communication that shows a textual pattern and spatial-temporal that was identified clearly. These motivations made the author to propose a methodology that classified messages containing hate or violent speech in a precise way.

With SNA tools, Angela *et al.* (2011) applied a content-based SNA approach in their work to find out people's interest in mailing list network. From the entire communication network, they extracted an overlapping topic that is related to subnetworks. Later, they combined it with TM to find out the extent to which sharing interest is connected with communication in two R-mailing lists (R-help and R-devel) where related questions of all kinds of application and development were discussed to describe people's interest. The authors found out that the communication efficiency is high in active mailing list than a mid-active list. In order to find the relationship between communication and interest sharing, they recommended that only the subject should be used as more noise is contained in its content.

Research that explored the innovative and cross-methodological computational approaches developed an investigative study of criminal behavior. Lettieri *et al.* (2017) combined agent-based modeling and SNA into an app they called CrimeMiner in order to support the study of criminal organizations as well as an experimental framework that seamlessly integrates visualization and document enhancement.

Network analysis and web mining techniques have been widely used in the analysis of website contents and structures of hate groups on social media, these techniques have not been applied in blogs in the study of hate groups. To address the problem, a framework was proposed that consists of four modules: network analysis, information extraction, blog spider and visualization. They applied this framework on Xanga which is a popular blog hosting sites to identify and analyze a selected set of 28 anti-Blacks hate groups. In these groups, their analysis result shows some interesting topological characteristics, demographical and identified at least two large communities on top of the smaller ones. They suggested that the framework proposed can be applied to blog analysis and can also be generally used in other domain (Chau and Jennifer, 2006).

In Batrinca and Treleaven (2015), a comprehensive review of software tools was provided for social networking media such as newsgroups, Wikis, blogs, chat and much more. The review was written for researchers that seek to analyze scrapping and analytics on social media in their research. They included an introduction to the scraping, data cleaning, storage and sentiment analysis of social media. As a result of the availability of SN APIs such as News Services APIs, Twitter and Facebook, they provided a methodology and critique of social media tools.

3.2 Methodology for analyzing hate speech in the proposed system

Sentiment analysis uses text analytics in understanding the polarity of sentences. The major task of the proposed system is to classify the polarity of a sentence in a retrieved tweet as negative or positive using the dictionary classifier, and then analyze the performance of the system using the word cloud, and pie charts visualizations. This work uses advanced NLP techniques by applying part of speech tags on collected and extracted tweets. As earlier stated, this research developed a Python Flask sentiment classifier application called POSA, using unsupervised learning or automatic classification. The processes involved in our methodology for the current research can be outlined as follows: data collection, data cleaning, data transformation, data visualization, selection of the algorithm, testing of the algorithm, using the system as shown in Figure 1. As a means of comparison, R-Studio codes are also written to implement text opinion mining of the same tweets retrieved by the Python program and visualized using the R-word cloud module and plots. Two sets of experiments were performed for both POSA and R. During the first POSA experiment, 30 Nigerian-based Twitter handles were mined (particularly for political, ethnics and racial tweets). For the second POSA experiment, 11 Nigerian top trending handles were examined with tweets covering the areas of politics, foreign affairs, religion, education, social activism, racism, etc. While for the first R experiment, the same data set used for the first POSA experiment was also used. The R TM package was used for this experiment. In the second R experiment, the same data set used for the second POSA experiment was also used, but this time the new sentimentr package was used. The results of both the Python and R text analytics are compared. The system architecture of POSA is shown in Figure 1.

Figure 2 shows a screenshot of the POSA dashboard.

Data collection from Twitter social network platform. Twitter is one of the most popular SN and micro-blogging platforms that allow posts of real-time messages from users known as tweets. It has over 300m active monthly users. This platform allows users to post contents such as facts, opinions, thoughts, references to images and other media. User posts in twitter are non-anonymous and in spite of its non-anonymity, users still find ways of using emotions, acronyms and other characters that express special meaning out of which hate speech is conveyed. Short messages are published on Twitter with a maximum of 140 characters (tweets). The platform also allows users to stream videos filmed live by other users and posted online. Twitter is not just a Social Network Corporation where users display and increase their social relationships, Twitter is now a real communication medium where users can choose a node and topic of reference based on their culture and interests (Fornacciari *et al.*, 2015).

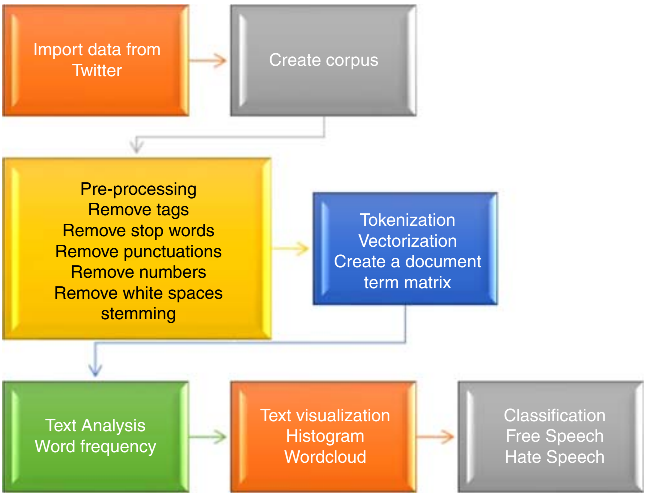
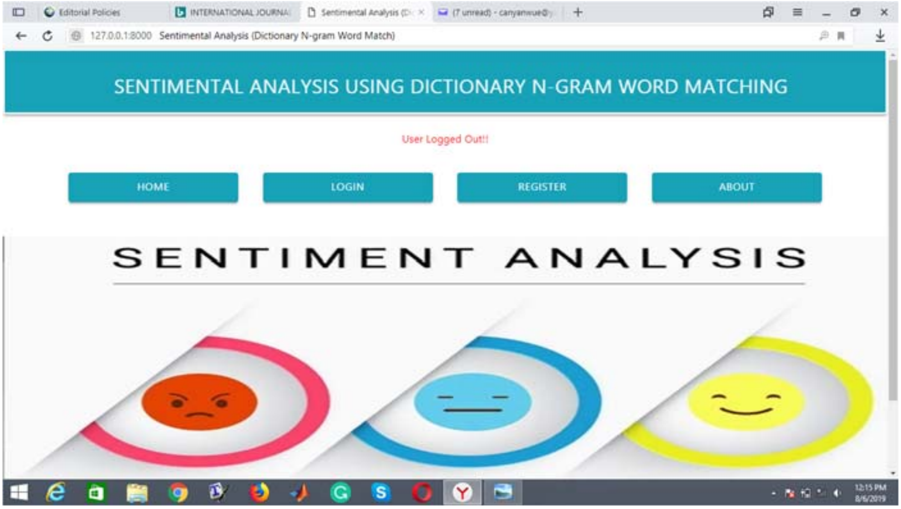


Figure 1.
System architecture



SENTIMENTAL ANALYSIS USING DICTIONARY N-GRAM WORD MATCHING

User Logged Out!!

SENTIMENT ANALYSIS

Figure 2.
Dashboard of POSA

Twitter's tweets were considered as a valuable resource for determination of hate speeches due to the following:

1. Twitter tweets are more descriptive, longer and publicly available, unlike Facebook.
2. Twitter is actually meant for people you want to communicate with not for family and friends.
3. A Twitter message contains a time zone that is often associated with the tweet.
4. Twitter is a tool for the promotion of brands and products as the producers cannot push their products to consumers without the consumers pushing back at them. For instance in customer service, Twitter is the easiest and quickest way to contact companies.
5. Twitter offers an up to the minute analysis of an outbreak due to its frequent high message posting.
6. Twitter has several sets of users and is also loved by celebrities.

Data collection with Twitter API. In Twitter, any user that creates an account can post and read tweets while anonymous users only read them. There are two application programming interfaces in Twitter micro-blogging service: Twitter Rest API and search API. An API is a software or software code intermediary that allows two applications to communicate with each other. For example, message APIs help our mobile phones to communicate with the internet to send or receive messages like instant messages, check the weather, etc. It is also the duty of an API to provide security to the user. APIs adhere to industry standards like the hypertext transfer protocol, HTTP or the more recent secured HTTPS, and REpresentational State Transfer, REST or RESTful web service protocol. REST uses HTTP requests to get, put, post and delete data. REST is generally preferred to the more robust simple open access protocol, SOAP because it uses less bandwidth, therefore making it more suitable for internet usage. RESTful APIs are used in most cloud websites such as Amazon, Google, Linked, Twitter, etc.:

- Twitter Rest API allows access to core Twitter data (Python Twitter API) to developers which include tweet texts, status data, update timelines and user information.

- The Search API allows the interaction of developers with Twitter Search and trends data. This is not our current interest but was used in Udanor *et al.* (2016).

Retrieving data with Twitter streaming API. API makes it easier for web services and computer programs to interact. APIs are provided by web services to be able to interact with their services. Four pieces of information are needed by a user application to establish a connection with the Twitter Streaming API: API key, API secret, access token and access token secret. These parameters were first obtained by the researcher, using his Twitter handle before the proposed application could gain access to Twitter public data.

3.3 Building the data collection module

Data collection is an essential requirement for classifiers. There was no readily available data set that was related to our work. This work focused mainly on SN data from top trending Twitter handles in Nigeria in the areas of news, politics, religion, social activism and ethnicity.

Python dictionary (Dictionary N-gram Word Matching Method) and python function that connect to open sourced Twitter API and SQLite 3 database were used to build our data collection server. Real-time tweets were streamed as unstructured texts and stored in the application's memory for immediate processing, ensuring there is no data redundancy. However, the user may want to store his result in the database. Tweets were extracted from mostly used hashtags and the messages collected were divided into positive and negative classes in the (POSA) Python program for matching with the word N-gram dictionary. The Python program was also used to develop a pie chart and word cloud modules for visualization and analysis. Python programming language was also used for indexing due to its versatility and agility. The following summarizes the steps involved in the system functionality:

(1) Data collection using Twitter API: large data sets of twitter which is publicly unavailable.

(2) Data pre-processing: this involves streaming of data, removing stop words, punctuations, etc.

(3) Tweets classification: the result from the above step classifies tweets to either positive or negative.

(4) Graphical representation of sentiments: the sentiment analysis result is presented using pie charts and word clouds.

The proposed system is developed using the Python tweet streaming API (Tweepy) tool, Python Flask and Word Cloud python modules. Python Flask and Word Cloud python modules are open-source tools and can be found on GitHub, or installed using the Python pip tool to an existing Python installation. The flask is a scripting language. It facilitates the development of a web-based program and creation of web pages using HTML5 for the frontend development. The local server is started and run from the command prompt. The backend code enables communication between the HTML5 and the Sqlite3 databases. The system is locally online, meaning an intranet service that allows multi-usage. The server enables authorized users to access data in the database. The bootstrapped interface consists of several functions of the cascading style sheet (CSS), JS and images. The CSS is used for the look and feel, while JavaScript is used for improving functionality and animations. Tweepy API allows a Python programmer to interact with the Twitter API platform. Tweepy is supported by OAuth which is a new basic authentication approach. The developer's web application that interfaces with the Python Twitter API requires the application key and application secret key as well.

In addition to the general N-Gram dictionary terms, the researchers built into the application, localized words that are commonly used to convey sentiments in the context of ethnicity, politics, religion and tribal sentiments in a polarized multi-cultural and multi-religious society like Nigeria. The total number of words in the negative term dictionary is 4,843, while the positive term dictionary contains 2,006 words. The application allows the user access to do:

1. Data access: it gives one the freedom to choose data on any topic.
2. Data query: it allows retrieval of data that is based on the query the user entered.
3. Graphical representation: users can visualize their results using a pie chart and word cloud.
4. Feature selection: the proposed system provides accurate feature selection on the components of the tweets to be analyzed and classified.

3.4 Pre-processing of data

In the proposed (POSA) system, Twitter API is used to retrieve tweets from Twitter based on the user query. The tweets collected are subjected to pre-processing. The raw text is a bit messy so before we can do any analytics we need to clean things up. Cleaning the data requires removing some unwanted characters such as hashtags, https, special characters and replace with space or no space, as well as converting all texts to lowercase. This is done with Python codes.

3.5 Data transformation by tokenization and vectorization

An N-gram is a sequence of tokens derived by breaking sentences into uni-gram, bi-grams or tri-grams (words or phrases) that may appear in a text corpus. In machine learning tasks like NLP, we usually need to generate N-grams from input sentences. It parses each document in the corpus or bag of words into tokens, and then into all possible N-grams (combinations of sequential tokens). In python, an input sentence is just a string of characters. Python has built-in functions (Albert, 2018) that can be used to quickly generate N-grams, like the NLTK Tokenizer. Dictionary classifier is a unique pattern which is the main algorithm. This is due to the fact that the keys must be ordered and the values are made unique, while dictionary keys are inherently unordered and the values are not checked between keys.

Dictionary N-gram Search Word Matching Method is applied to the stored custom data dictionary to match the words retrieved from the tweets by the POSA system. The results of the Dictionary N-gram Search Word Matching is presented as numeric data (i.e. vectorization) and the sentiment represented graphically in pie charts or as text data are shown graphically as a word cloud. The purpose of vectorization is for the data to make sense to the machine learning algorithm, which can be achieved by converting the entire data into a one column matrix.

The data dictionary consists of a list of words (positive and negative):

- Number of positive words: 2,006.
- Number of negative words: 4,783.

Sample dictionary words which we have used to classify tweets into positive and negative sentiments are listed in Table I.

	Proposed word dictionary
1	Positive: abound, abundance, admire, honored, love, hopeful, rich and cool
	Negative: abnormal, abuse, addict, ail, suck, bad, terrible and hate
2	Positive: cute, crisp, deft, delight, delicacy, dote, divine, good and enjoy
	Negative: buggy, bull, cancer, careless, cold, choke and comical
Statistics based	Positive: pretty, love, best, wonderful and good
	Negative: hate, stupid, bad and waste

Table I. Proposed word dictionary for the classification of tweets (positive and negative)

3.6 R supervised sentiment classification

As a means of comparing the results obtained from the POSA system, the authors also used the same tweets as text files in a corpus made up of three files ("political tweets.txt," "ethnic tweets.txt" and "racism tweets.txt") in conducting the first experiment using the TM package, while the second experiment read in a .csv file containing a corpus of tweets retrieved from 11 trending Nigeria-based Twitter handles containing 1,207 observations and applied the (sentimentr) library.

R is an open-source application with many contributors developing many powerful tools like packages, installed and imported as libraries by users for analysis. The sentimentr package version 2.7.1 is designed and maintained by Tyler (Rinker 2019; Renard, 2018) to quickly calculate text polarity sentiment at the sentence level and optionally aggregate by rows or grouping variable(s). The package uses a lexicon approach to sentiment analysis. The sentiment package is a response to the needs with sentiment detection that were not addressed by the current R tools.

According to Rinker, sentimentr attempts to take into account valence shifters (i.e. negators, amplifiers (intensifiers), de-amplifiers (downtoners) and adversative conjunctions) while maintaining speed. Simply put, sentimentr is an augmented dictionary lookup. The next questions address why it matters.

So what are these valence shifters?. Rinker gave the following examples in explaining what valence shifters are: a negator flips the sign of a polarized word (e.g. "I do not like it."). See lexicon::hash_valence_shifters[y = = 1] for examples. An amplifier (intensifier) increases the impact of a polarized word (e.g. "I really like it."). See lexicon::hash_valence_shifters[y = = 2] for examples. A de-amplifier (downtoner) reduces the impact of a polarized word (e.g. "I hardly like it."). See lexicon::hash_valence_shifters[y = = 3] for examples. Adversative conjunction overrules the previous clause containing a polarized word (e.g. "I like it but it's not worth it."). See lexicon::hash_valence_shifters[y = = 4] for examples.

The sentiment works this way, each paragraph ($p_i = \{s_1, s_2, \ldots, s_n\}$) composed of sentences, is broken into element sentences ($s_i, j = \{w_1, w_2, \ldots, w_n\}$) where w are the words within sentences. Each sentence (s_j) is broken into an ordered bag of words. The text is then cleaned by removing punctuation, with the exception of pause punctuations (commas, colons and semicolons) which are considered a word within the sentence. The words in each sentence ($w_{i,\ j,\ k}$) are searched and compared to a dictionary of polarized words (e.g. a combined and augmented version of the syuzhet package and Rinker's augmented dictionaries in the lexicon package). Positive ($w^{+}_{i,\ j,\ k}$) and negative ($w^{-}_{i,\ j,\ k}$) words are tagged with a +1 and −1, respectively (or other positive/negative weighting if the user provides the sentiment dictionary).

Sentimentr uses two main commands to quickly analyze the sentiment of a sentence, the "sentiment(sometext)" and "sentiment_by(sometext)" commands. The former returns the positive, negative and neutral sentiments as numeric values sentence by sentence. For example, values of zero means neutral, negative values, e.g. −1, indicate negative sentiment, while values like 0.1, 0.5, etc. indicate positive sentiment. The sentiment_by command is used to get an aggregate sentiment measure for the entire text under review. It returns the sentiment object in the form of a data.table, which includes the following columns:

- element_id – the id number of the review;
- word_count – the word count of the review;
- sd – the standard deviation of the sentiment score of the sentences in the review; and
- ave_sentiment – the average sentiment score of the sentences in the review.

The most interesting variable is the ave_sentiment, which is the sentiment of the review in one number. The number can take positive or negative values and expresses the valence and the polarity of the sentiment.

The sentiment package can also enable one to visualize the sentiment of a text using the ggplot2 package, as well as display the positive and negative words found in the text. The output of the analysis of our tweets is shown in the Results section. Word clouds were also generated for both R-Studio experiments also for visualization of the results.

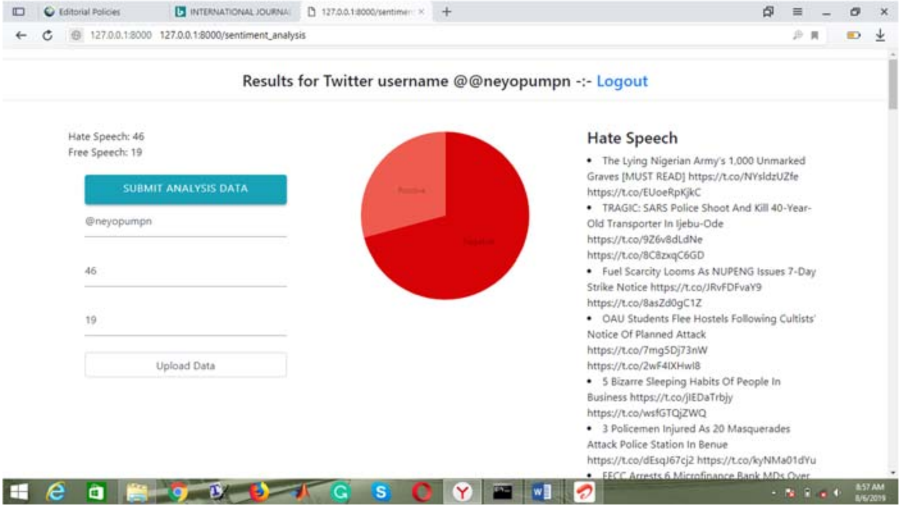
4. Results and discussion

In this section, the results of both the POSA Python software and the R-Studio lexical analytics are presented. The result of the POSA lexical analytics is shown first, then that of the R software. Finally, comparison tables are provided for the sentiment analysis results of both software, and the discussion follows thereafter.

4.1 The POSA result

The POSA software, in addition to being able to stream live tweets from any public Twitter handle, is also able to display the results in a form containing a pie chart which converts the sentiments scores into numerical values and in percentages of positive to negative, a sample screenshot is shown in Figure 3. It is also able to analyze and display the texts in a word cloud as shown in Figure 4.

Table II shows the summary of the results generated from different Twitter handles covering the three areas under investigation (politics, ethnicity and racism) during the first POSA experiment. Table III shows the summation of the three categories of tweets, their percentage scores and their sentiment scores, as well as the sentiment classification, while it shows the corresponding result of the first R experiment.



Results for Twitter username @@neyopumpn -:- Logout

Hate Speech: 46
Free Speech: 19

@neyopumpn

46

19

Hate Speech

- The Lying Nigerian Army's 1,000 Unmarked Graves [MUST READ] https://t.co/NYsldzUZfe https://t.co/EUoeRpKjkC
- TRAGIC: SARS Police Shoot And Kill 40-Year-Old Transporter In Ijebu-Ode https://t.co/9Z6v8dLdNe https://t.co/8C8zxqC6GD
- Fuel Scarcity Looms As NUPENG Issues 7-Day Strike Notice https://t.co/JRvFDFvaY9 https://t.co/8asZd0gC1Z
- OAU Students Flee Hostels Following Cultists' Notice Of Planned Attack https://t.co/7mg5Dj73nW https://t.co/2wF4IXHwl8
- 5 Bizarre Sleeping Habits Of People In Business https://t.co/jIEDaTrbjy https://t.co/wsfGTQjZWQ
- 3 Policemen Injured As 20 Masquerades Attack Police Station In Benue https://t.co/dEsqJ67cj2 https://t.co/kyNMa01dYu
- EFCC Arrests 6 Microfinance Bank MDs Over

Figure 3.
Pie chart analysis of a POSA result

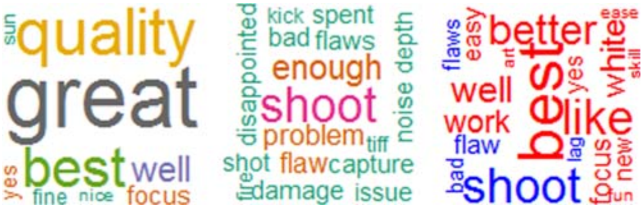


Figure 4.
Word cloud representation of the POSA (positive and negative) terms

In Tables II and III, the sentiment score is obtained by subtracting the negative sentiment score from the positive score in order to classify the sentiment. From Table II, both ethnics and political tweets retrieved and analyzed are classified as positive sentiments, while the racism tweets are classified as a negative sentiment. But in Table III, political tweets returned the negative sentiment. A comparison of the results of both software (from Tables II and III) shows that in both, only the ethnics result has the same sentiment classification of positive, while the political and racism tweets have reverse results.

The result of the second POSA experiment conducted on 11 top trending Nigeria Twitter handles with tweets covering the areas of politics, foreign affairs, religion, education, social activism, racism, etc., is shown in Table IV. The numbers reflect the amount of hate/free speech found in each tweet, as classified by POSA. The data on this table are plotted on the histogram in Figure 5.

Table V shows the summary of the scores in Table IV, such as percentage of hate to free speech, as well as the prediction accuracy. These data were also used to conduct a similar experiment in R-Studio. The results are shown in Table VI.

Computing classifier accuracy. The following formula is used to compute the classification accuracy:

$$\text{Percentage accuracy} = \left(\text{No. of correctly predicted polarity}/\text{Total no. of polarity in the corpus}\right) \times 100.$$

Using the count( ) function, we were able to obtain the total number of both negative and positive polarities from the tweets. We also conducted a physical examination to ensure that

Table II.
Result of first POSA (Python SW) sentiment analysis

S/N	Type of Tweet	Positive sentiment score	Negative sentiment score	Sentiment score	Classification of sentiment
1	Ethics	142 (66.7%)	71 (33.4%)	71	Positive
2	Politics	94 (62.7%)	56 (37.3%)	38	Positive
3	Racism	67 (44.7%)	83 (55.3%)	−16	Negative

Table III.
Result of first R sentiment analysis

S/N	Type of Tweet	Positive sentiment score	Negative sentiment score	Sentiment score	Classification of sentiment
1	Ethics	179 (59%)	120 (40.1%)	59	Positive
2	Politics	75 (37.1%)	127 (62.9%)	−52	Negative
3	Racism	123 (62%)	75 (38%)	48	Positive

Table IV.
The result of the second POSA experiment

Handles	Hate speech	Free speech
@Mssxxxxx	14	43
@Omxxxxxx	17	30
@gixxxxxxx	17	28
@gbxxxxxxx	14	24
@Abxxxxxx	15	31
@blxxxxxxx	9	43
@Aixxxxxx	21	32
@ayxxxxxx	25	25
@Mrxxxxxx	38	10
@Ayxxxxxx	17	24
@Chxxxxxx	20	19

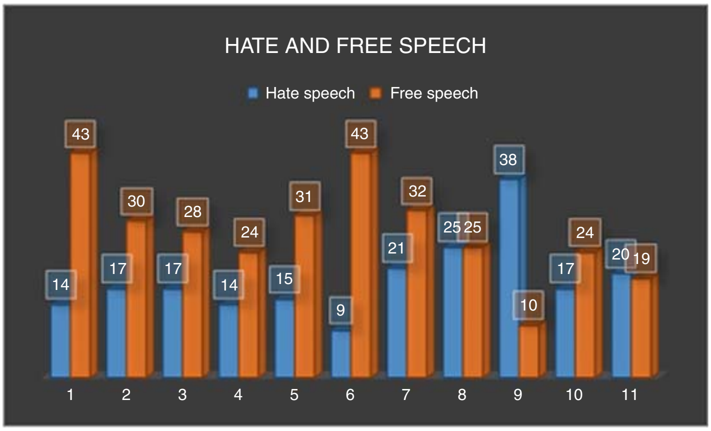


Figure 5.
Hate vs free speech in second POSA experiment

Contents	Negative terms	Positive terms
Summed sentiments	207	309
%	40.1	59.9
Accuracy (%)	207/237×100 = 87.3	309/360×100 = 85.8

Table V.
Second POSA sentiment analysis result

Contents	Negative terms	Positive terms
Summed sentiment	155	148
%	51.1	48.9
Accuracy (%)	155/237×100 = 65.4	266/360×100 = 74

Table VI.
Second R-Studio sentiment analysis result

there was no repetition of terms. The results of the computations for both POSA and R experiments are shown in the last rows of Tables V and VI, respectively.

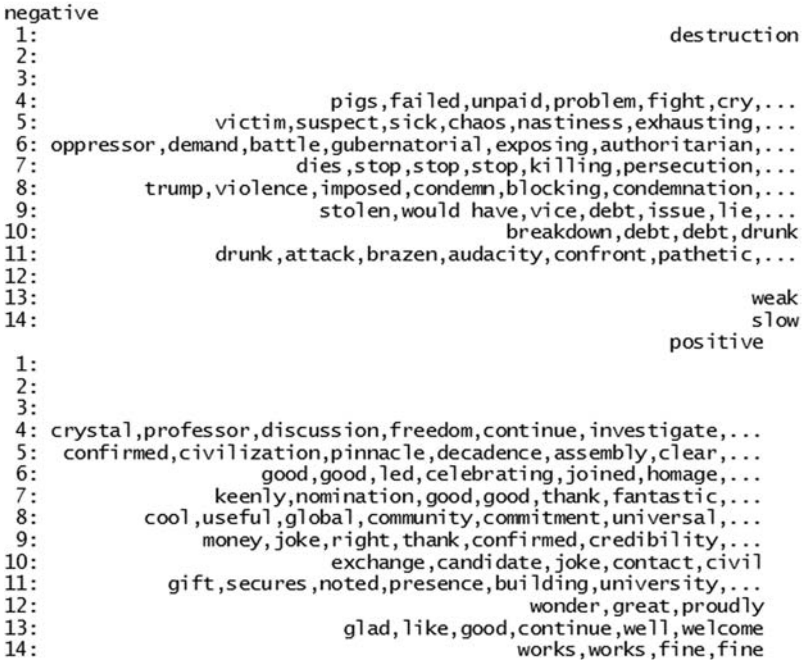
Figure 6 shows a sample print out of the sentiment polarities as predicted by the sentimentr library. Notice that some terms were repeated. This was sorted out during the physical examination.

Figures 7 and 8 show the negative and positive sentiment polarities plots, respectively, from the R experiment.

4.2 Discussion

In the first experiments, a comparison of Tables II and III shows that, in Table II, POSA returned positive for ethics (71) and politics (38), respectively, and negative (−16) for racism, R returned positive for ethics (59), negative (−52) for politics and positive for racism (48), respectively. Comparing both results, we find out that the POSA software found that the Twitter handles analyzed contained between 33 and 55 percent hate contents, while the R-Studio shows hate content ranging from 38 to 62 percent.

Performing a *t*-test on both positive and negative scores for both POSA and R-studio, results reveal *p*-values of 0.389 and 0.289 for the positive and negative scores, respectively, with an α value of 0.05. This implies that there is no significant difference in the results obtained from our POSA software when compared with the one obtained from the standard



```
negative
 1:                                                       destruction
 2:
 3:
 4:                          pigs,failed,unpaid,problem,fight,cry,...
 5:               victim,suspect,sick,chaos,nastiness,exhausting,...
 6: oppressor,demand,battle,gubernatorial,exposing,authoritarian,...
 7:                     dies,stop,stop,stop,killing,persecution,...
 8:          trump,violence,imposed,condemn,blocking,condemnation,...
 9:                       stolen,would have,vice,debt,issue,lie,...
10:                                     breakdown,debt,debt,drunk
11:             drunk,attack,brazen,audacity,confront,pathetic,...
12:
13:                                                              weak
14:                                                              slow
                                                          positive
 1:
 2:
 3:
 4: crystal,professor,discussion,freedom,continue,investigate,...
 5:  confirmed,civilization,pinnacle,decadence,assembly,clear,...
 6:                   good,good,led,celebrating,joined,homage,...
 7:               keenly,nomination,good,good,thank,fantastic,...
 8:         cool,useful,global,community,commitment,universal,...
 9:             money,joke,right,thank,confirmed,credibility,...
10:                       exchange,candidate,joke,contact,civil
11:           gift,secures,noted,presence,building,university,...
12:                                        wonder,great,proudly
13:                          glad,like,good,continue,well,welcome
14:                                        works,works,fine,fine
```

Figure 6.
Sample negative and positive sentiments retrieved from tweets

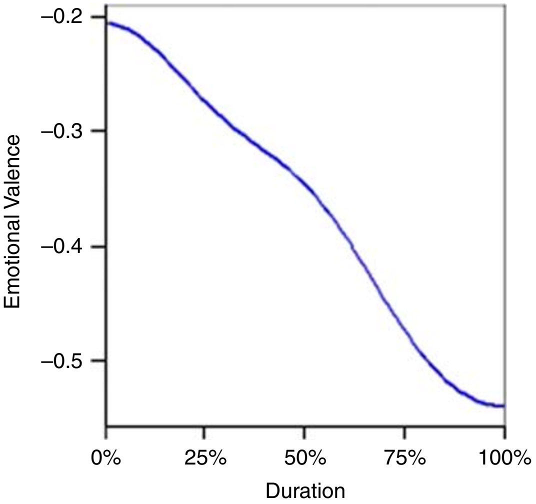


Figure 7.
Negative sentiment plot

R-Studio software. This is a measure of the goodness of the POSA app in being able to classify as good as R.

During the second experiment, we deduce from Tables V and VI, respectively, as follows:

(1) that the percentage of hate contents classified by POSA is 40 percent;

(2) that the percentage of hate contents classified by R is 51 percent;

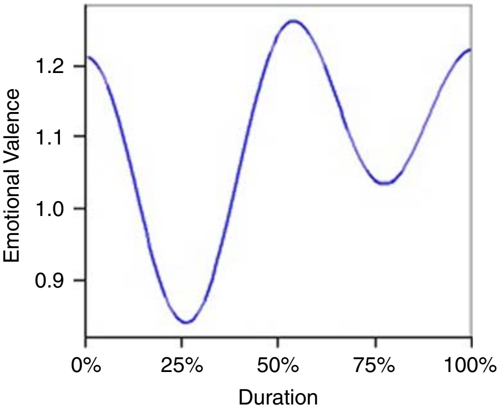


Figure 8.
Positive sentiment plot

(3) that the accuracy of hate speech classification predicted by POSA is 87 percent, while free speech is 86 percent; and

(4) that the accuracy of hate speech classification predicted by R is 65 percent, while free speech is 74 percent.

It should be noted that during the second experiments, different dictionaries were used for the POSA and the R sentiment analysis, respectively, unlike the first experiments where the same dictionary was used. R uses a combined and augmented version of the syuzhet package and Rinker's augmented dictionaries in the lexicon package. The POSA system showed a better classification accuracy because of the localized and customized dictionary that was used.

With the percentage of hate contents in all the experiments ranging between 30 and 60, we set out to know what the acceptable hate contents in the social media should be. And what monitoring mechanisms the social media platforms like Facebook and Twitter have put in place to combat hate speech.

The US first amendment, on hate speech, states that determining when conduct crosses the line of targeted harassment, or threat, or creating a hostile environment, is a legal question that requires an examination on a case-by-case basis (American Library Association, 2017). In other words, there is no strict boundary or parameter to measure the level of hate content. That gives room to subjectivity. The same first amendment says that "Hate speech" does not have a legal definition under US law, just as there is no legal definition for rudeness, evil ideas, unpatriotic speech or any other kind of speech that people might condemn. With a setting like this, it may be difficult for social media companies to draw a hard line on what may be called "extreme" hate speech. Facebook has hired over 7,500 people whose job it is to watch out for users' (Barbara, 2018; Waseem and Hovy, 2016) complaints on what they term hate speech. Then Facebook will have to take down such offensive posts. Many a time, after some users protests for taking down their posts, Facebook restored them, admitting that it was done in error. This gives room for inconsistency.

Twitter agrees that they do not have a means of automatically monitoring and combating hate speech.

This led to the company's hiring of academics from institutions like Oxford (Archer, 2018) to help combat the "intolerant discourse" and monitor the health of the SN. It is part of Twitter's move to create algorithms that will better distinguish between hate speech and conversations that break the "norms of politeness." According to Sophia (2018), Jack Dorsey, Twitter's CEO in 2017, said, "We see voices being silenced on Twitter every day. We've been working to counteract this for the past 2 years […] We prioritized this in 2016. We updated our policies and increased the size of our teams. It wasn't enough," while Mark Zuckerberg, Facebook Founder in 2018, says, "It's much easier to build an AI system that can detect a nipple that it is to determine what is linguistically hate speech."

Our results also show that there is a consistent growth in hate contents on social media. They also indicate that the growing negative contents on the new media could be a reflection of societal ills making a transition from physical to electronic platforms, just as bank armed robbery metamorphosed into online identity theft of credit card information, and other financial transactions because the world has gone cashless. Regarding our finding on the rise in the number of hate contents being released on social media, we compare this with the findings of other studies in that regard.

Iftikhar *et al.* (2016) in a survey of hate vs free speech on social media in India agree that the posting of hate speech is on the increase and many more are joining. Brett (2018), while investigating social media users' willingness to help combat extreme speech on social media, discovered that the participants in general did not show any willingness to censor extreme speech on the social media, rather they mocked the attempts being made by social media companies to combat hate languages. Oksanen *et al.* (2014) in an attempt to investigate the effect of exposure to online hate among young social media users between ages 15 and 18 years observed that the negative impact is multi-faceted. They observed that 67 percent of young users have been exposed to hate materials primarily focused on sexual orientation, physical appearance and ethnicity mostly on Facebook and YouTube. This has led to offline victimization, poor family attachments and general unhappiness. In a similar work to investigate the effect of exposure to online hate material on social trust among Finnish youths (Matti *et al.*, 2015), it was revealed that witnessing negative images and writings reduces both generalized and particularized trust. The authors opined that exposure to hate materials has a more negative effect on relationships with acquaintances than in a general context. Diane (2008) feared the negative impact of race hate campaigns on UK media and observed that government policies and approach to the anti-terrorism "war" have not helped matters either. The author observed that this campaign has led to Islamophobia.

Gambäck and Sikdar (2017) using unsupervised learning on word2vec convolutional neural network model found the result outperforming their counterpart supervised models built on randomly generated word vectors, and word vectors combined with character N-grams, respectively, with an average *F*-score of 78.69.

5. Conclusion

This study sets out to investigate the reality of hate speech in social media, Twitter especially using unsupervised text classification. The results of the study establish the fact that hate speech is on the increase in social media platforms, and hate peddlers do not show any form of repentance. It also shows that hate mongers can actually be pinned down, with the contents of their messages. The findings can be used by social media companies to monitor user behaviors, and pin hate crimes to specific persons. Governments and law enforcement agencies can also use the POSA application to track down hate peddlers.

5.1 Future work

The study was limited to public Twitter handles only. The study only focused on the polarity of tweets, it did not include subjectivity analysis. N-grams are effective features for word-sense disambiguation, but when using N-grams, the feature vector could take on enormous proportions and in turn increasing sparsity of the feature vectors. Future work will focus on supervised machine and deep learning approaches, using word2vec RNN sentiment analysis on Tensor Flow framework. The result will be compared with other supervised classification algorithms like SVM, decision trees, Naïve Bayes, etc.

References

Aaron, J., Shobhit, H. and Mari, O. (2018), "Community member retrieval on social media using textual information", *Proceedings of NAACL-HLT 2018, Association for Computational Linguistics, New Orleans, LA, June 1–6*, pp. 595-601.

Agarwal, S. and Sureka, A. (2014), "A focused crawler for mining hate and extremism promoting videos on YouTube", *Proceedings of the 25th ACM Conference on Hypertext and Social Media, Santiago*, pp. 294-296.

Akhtar, N. (2015), "Network analysis tools", *4th International Conference on Communication Systems and Network Technologies, Bhopal, April 7–9*, doi: 10.1109/CSNT.2014.83.

Alakali, T.T., Faga, H. and Mbursa, J. (2017), "Audience perception of hate speech and foul language in the social media in Nigeria: implications for morality and law", *Academicus International Scientific Journal*, Vol. VIII No. 15, pp. 166-183.

Albert, A.Y. (2018), "Notes on machine learning and A.I. generating N-grams from sentences python", available at: www.albertauyeung.com/post/generating-ngrams-python/ (accessed June 3, 2018).

American Library Association (2017), "Hate speech and hate crime", available at: www.ala.org/advocacy/intfreedom/hate (accessed April 2, 2019).

Angela, B., Ingo, F., Kint, H. and Patrick, M. (2011), "Content-based social network analysis of mailing lists", *The R Journal*, Vol. 3 No. 1, pp. 11-18.

Anna, S. and Michael, W. (2017), "A survey on hate speech detection using natural language processing", *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media, Valencia*, pp. 1-10.

Aondover, E.M. (2018), "Curbing hate speeches on social media: in letters", available at: http://thenationonlineng.net/curbing-hate-speeches-social-media/ (accessed March 31, 2018).

Archer, J. (2018), "The telegraph technology intelligence", Twitter hires academics to monitor its "health" and combat hate speech, available at: www.telegraph.co.uk/technology/2018/07/30/twitter-hires-academics-monitor-healthand-combat-hate-speech/ (accessed April 28, 2019).

Badjatiya, P., Gupta, S., Gupta, M. and Varma, V. (2017), "Deep learning for hate speech detection in tweets", *Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, April*, pp. 759-760.

Barbara, O. (2018), "Facebook says it's getting better at removing hate speech", available at: https://phys.org/news/2018-11-facebook-speech.html (accessed April 19, 2019).

Barthel, M., Shearer, E., Gottfried, J. and Mitchell, A. (2016), "The evolving role of news on Twitter and Facebook", Pew Research Center's Journalism Project, available at: www.pewresearch.org/wp-content/uploads/sites/8/2015/07/Twitter-and-News-Survey-Report-FINAL2.pdf (accessed February 4, 2016).

Batrinca, B. and Treleaven, P.C. (2015), "Social media analytics: a survey of techniques, tools, and platforms", *AI & Society*, Vol. 30 No. 1, pp. 89-116, doi: 10.1007/s00146-014-0549-4.

Benesch, S. (2014), "Countering dangerous speech: new ideas for genocide prevention", working paper, Dangerous Speech Project, United States Holocaust Memorial Museum, Washington, DC, available at: https://dangerousspeech.org/ (accessed May 15, 2018).

Boldrini, E., Balahur, A., Martínez-Barco, P. and Montoyo, A. (2012), "Using EmotiBlog to annotate and analyze subjectivity in the new textual genres", *Data Mining Knowledge Discovery*, Vol. 25 No. 3, pp. 603-634, available at: https://doi.org/10.1007/s10618-012-0259-9

Brett, G.J. (2018), "Tolerating and managing extreme speech on social media", *Internet Research*, Vol. 28 No. 5, pp. 1275-1291, available at: https://doi.org/10.1108/IntR-03-2017-0100

Burnap, P. and Williams, M.L. (2015), "Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making", *Policy & Internet*, Vol. 7 No. 2, pp. 223-242.

Burnap, P., Williams, M.L., Sloan, L., Rana, O.F., Housley, W., Edwards, A., Knight, V.A., Procter, R. and Voss, A. (2014), "Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack", *Social Network Analysis and Mining*, Vol. 4, pp. 1-14.

Butts, C.T. (2008), "Social network analysis: a methodological introduction", *Asian Journal of Psychology*, Vol. 11 No. 1, pp. 13-41.

Cambria, E., Schuller, B., Xia, Y. and Havasi, C. (2013), "New avenues in opinion mining and sentiment analysis", *IEEE Intelligent Systems*, Vol. 28 No. 2, pp. 15-21.

Chau, M. and Jennifer, X. (2006), "A framework for locating and analyzing hate groups in blogs", *PACIS 2006 Proceedings*, *Kuala Lumpur, MA*, *July 6–9*, available at: http://aisel.aisnet.org/pacis2006/60 or https://pdfs.semanticscholar.org/d30e/becfabb830368fa9b3623e7cc20f19aff433.pdf

Chen, M., Zhang, Y., Li, Y., Hassan, M.M. and Alamri, A. (2015), "AIWAC: affective interaction through wearable computing and cloud technology", *IEEE Wireless Communications*, Vol. 22 No. 1, pp. 20-27.

Cherian, G. (2018), "HATE SPEECH: a dilemma for journalists the world over", available at: https://ethicaljournalismnetwork.org/resources/publications/ethics-in-the-news/hate-speech (accessed August 23, 2018).

Compagnon, P. and Ollivier, K. (2017), "Graph embeddings for social network analysis: state of the art", available at: www.researchgate.net/publication/331714802_Graph_Embeddings_for_Social_Network_Analysis_State_of_the_Art (accessed April 19, 2019).

Davidson, T., Warmsley, D., Macy, M. and Weber, I. (2017), "Automated hate speech detection and the problem of offensive language", *11th International AAAI Conference on Web and Social Media Montréal, Québec, Canada, The AAAI Press*, *Palo Alto, CA*, *May 15–18*.

Diane, F. (2008), "Islamophobia: examining causal links between the media and 'race hate' from 'below' ", *International Journal of Sociology and Social Policy*, Vol. 28 Nos 11/12, pp. 564-578, available at: https://doi.org/10.1108/01443330810915251

Dos Santos, C. and Gatti, M. (2014), "Deep convolutional neural networks for sentiment analysis of short texts", *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 69-78.

Dragoni, M., Poria, S. and Cambria, E. (2018), "OntoSenticNet: a commonsense ontology for sentiment analysis", *IEEE Intelligent Systems*, Vol. 33 No. 3, pp. 77-85.

Ekman, P. (1999), "Basic emotions", in Dalgleish, T. and Power, M. (Eds), *Handbook of Cognition and Emotion*, Vol. 98 Nos 45-60, John Wiley & Sons, p. 16.

Fernando, M.-L. (2016), "Cyber hate speech on twitter: analyzing disruptive events from social media to build a violent communication and hate speech taxonomy", *International Journal of Design & Nature and Ecodynamics*, Vol. 11 No. 3, pp. 406-415, doi: 10.2495/DNE-V11-N3-406-415.

Fornacciari, P., Monica, M. and Michele, T. (2015), *Social Network and Sentiment Analysis on Twitter: Towards a Combined Approach*, KDWeb.

Gambäck, B. and Sikdar, U.K. (2017), "Using convolutional neural networks to classify hate-speech", *Proceedings of the First Workshop on Abusive Language*, *August*, pp. 85-90.

Gautam, G. and Yadav, D. (2014), "Sentiment analysis of twitter data using machine learning approaches and semantic analysis", *2014 Seventh International Conference on Contemporary Computing (IC3), IEEE, August*, pp. 437-442.

Ghiassi, M., Skinner, J. and Zimbra, D. (2013), "Twitter brand sentiment analysis: a hybrid system using n-gram analysis and dynamic artificial neural network", *Expert Systems with Applications*, Vol. 40 No. 16, pp. 6266-6282.

Iftikhar, A., Roshan Raini, L. and Faizia, S. (2016), "Free vs hate speech on social media: the Indian perspective", *Journal of Information, Communication & Ethics in Society*, Vol. 14 No. 4, pp. 350-363, available at: https://doi.org/10.110/jices-06-2015-0016

Jennifer, G. (2015), "Introduction to social media investigation", available at: www.sciencedirect.com/topics/computer-science/egocentric-network (accessed December 22, 2018).

Jeremy, W. (2012), *The Harm in Hate Speech*, Harvard University Press, Cambridge, MA, p. 304.

Joel, J. (2012), "Ethnopaulism and ethno-religious hate speech in Nigeria enabling policies for responding to 'hate speech' in Practice, 2012", available at: e-learning.ceu.hu/user/view.php?id=4190&course=1181 (accessed November 18, 2018).

Kontopoulos, E., Berberidis, C., Dergiades, T. and Bassiliades, N. (2013), "Ontology-based sentiment analysis of Twitter posts, expert systems with applications (2013)", available at: http://dx.doi.org/10.1016/j.eswa.2013.01.001 (accessed March 18, 2019).

Kwok, I. and Wang, Y. (2013), "Locate the hate: detecting tweets against blacks", *Twenty-Seventh AAAI Conference on Artificial Intelligence*, *Bellevue, Washington*, *July 14-18*.

Leondro, S., Maniack, M., Denzil, C., Fabricro, B. and Lngmar, W. (2016), "Analyzing the targets of hate in online social media", available at: https://arxiv.org/pdf/1603.07709.pdf (accessed November 6, 2018).

Lettieri, N., Altamura, A., Malandrino, D. and Punzo, V. (2017), "Agents shaping networks shaping agents: integrating social network analysis and agent-based modeling in computational crime research", in Oliveira, E., Gama, J., Vale, Z. and Lopes Cardoso, H. (Eds), *Progress in Artificial Intelligence, EPIA 2017*, Vol. 10423, Lecture Notes in Computer Science, Springer, Cham.

Levin, J. and McDevitt, J. (1993), *Hate Crimes: The Rising Tide of Bigotry and Bloodshed*, Plenum Press, New York, NY.

Liu, B. and Zhang, L. (2012), "A survey of opinions mining and sentiment analysis in mining text data", in Aggarwal, C. and Zhai, C. (Eds), *Mining Text Data*, Springer, Boston, MA, pp. 415-463.

Lizi, L., Xiangnan, H., Hanwang, Z. and Tat-Seng, C. (2017), "Attributed social network embedding", *Journal of Latex Class Files*, Vol. 14 No. 8, pp. 2257-2270.

Maina, K. (2010), "Speech, power and violence: hate speech and the political crisis in Kenya", available at: www.ushmm.org/m/pdfs/20100423-speech-power-violence-kiai.pdf (accessed December 22, 2018).

Malmasi, S. and Zampieri, M. (2017), "Detecting hate speech in social media", *Proceedings of Recent Advances in Natural Language Processing*, *Varna*, pp. 467-472, available at: https://doi.org/10.26615/97-954-452-049-6_062 (accessed June 4, 2018).

Matti, N., Pekka, R., James, H., Emma, H. and Atte, O. (2015), "Exposure to online hate material and social trust among Finnish youth", *Information Technology & People*, Vol. 28 No. 3, pp. 607-622, available at: https://doi.org/10.1108/ITP-09-2014-0198

Mirigxin, Z. (2010), "Social network analysis: history, concepts, and research", in Furht, B. (Ed.), *Handbook of Social Network Technologies & Applications*, Springer, Boston, MA, pp. 3-21.

Murphy, J. (2017), "A brief analysis of the free speech vs. Hate Speech Debate_Stand", available at: www.standleague.org/blog/a-brief-analysis-of-the-free-speech-vs-hatespeech-debate.html (accessed December 11, 2017).

Noorani, A.G. (1992), "Hate speech and free speech", *Economic and Political Weekly*, Vol. 27 No. 46, p. 2456.

Okeyo, G., Ogada, K. and Rimiru, R. (2018), "Using Naïve Bayes algorithm in detection of hate tweets", *International Journal of Scientific and Research Publications*, Vol. 8 No. 3, pp. 99-107.

Oksanen, A., Hawdon, J., Holkeri, E., Näsi, M. and Räsänen, P. (2014), "Exposure to online hate among young social media users", *Soul of Society: A Focus on the Lives of Children & Youth*, Sociological Studies of Children and Youth, Vol. 18, Emerald Group Publishing Limited, pp. 253-273, https://doi.org/10.1108/S1537-466120140000018021

Oliveira, D.J.S., Bermejo, P.H.D.S. and Santos, P.A.D. (2016), "Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls", *Journal of Information Technology & Politics*, doi: 10.1080/19331681.2016.1214094.

Pang, B. and Lee, L. (2008), "Opinion mining and sentiment analysis", *Foundations and Trends® in Information Retrieval*, Vol. 2 Nos 1-2, pp. 1-135.

Peter, B. and Leighton, W.M. (2014/2017), "Hate speech, machine classification and statistical modelling of information flows on twitter: interpretation and communication for policy decision making", Internet, Policy & politics, Oxford, September 26.

Pokorny, J., Snasel, V. and Richta, K. (2012), "Social network analysis: selected methods and applications", *Proceedings of the Dateso 2012 Workshop*, p. 151.

Reddy, V. (2002), "Perverts and sodomites: homophobia as hate speech in Africa", *Southern African Linguistics and Applied Language Studies*, Vol. 20 No. 3, pp. 163-175, doi: 10.2989/16073610209486308.

Renard, M. (2018), "Doing your first sentiment analysis in R with sentimentr Oct 2, 2018", available at: https://medium.com/@mattifuchs/doing-your-first-sentiment-analysis-in-r-with-sentimentr-167855445132 (accessed January 16, 2019).

Rigas, G., Katsis, C.D., Ganiatsas, G. and Fotiadis, D.I. (2007), "A user independent, biosignal based, emotion recognition method", *International Conference on User Modeling, Springer*, *Berlin and Heidelberg*, *July*, pp. 314-318.

Ring, C.E. (2013), "Hate speech in social media: an exploration of the problem and its proposed solutions", Journalism & Mass Communication Graduate Theses & Dissertations No. 15, available at: https://scholar.colorado.edu/jour_gradetds/15 (accessed December 2018).

Rinker, T. (2019), "Dictionary based sentiment analysis that considers valence shifters", available at: https://github.com/trinker/sentimentr (accessed April 22, 2019).

Saif, H., He, Y., Fernandez, M. and Alain, H. (2015), "Contextual semantics for sentiment analysis of Twitter", *The 13th International Semantic Web Conference*, *Riva del Garda*, *October 19-23*.

Saif, H., He, Y., Fernandez, M. and Alani, H. (2016), "Contextual semantics for sentiment analysis of Twitter", *Information Processing & Management*, Vol. 52 No. 1, pp. 5-19.

Serrano-Guerrero, J., Olivas, J.A., Romero, F.P. and Herrera-Viedma, E. (2015), "Sentiment analysis: a review and comparative analysis of web services", *Information Sciences*, Vol. 311, pp. 18-38.

Sophia, K. (2018), "Monitoring and tagging hate speech in social media", *IFRRO Athens World Congress Connecting the Dots: The Future of Collective Management*, Macedonia, October 23, available at: http://ifrro.org/sites/default/files/datascouting_sophia_karakeva_oct2018.pdf

Stanley, W. and Katherine, F. (1994), "Social network analysis in the social & behavioral science", in Wasserman, S. and Galaskiewicz, J. (Eds), *Social Network Analysis: Methods & Applications*, ISBN 9780521387071, Cambridge University Press and Sage Publications, London, pp. 1-27.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T. and Qin, B. (2014), "Learning sentiment-specific word embedding for twitter sentiment classification", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 1555-1565.

The Modern Law Review (2006), *The Modern Law Review*, Vol. 69 No. 4, pp. 543-582.

The Nation Nigeria (2017), "Hate speech", Editorial, August 30, available at: https://thenationonlineng.net/hate-speech/ (accessed April 2018).

Tofighy, S. and Fakhrahmad, S.M. (2018), "A proposed scheme for sentiment analysis: effective feature reduction based on statistical information of SentiWordNet", *Kybernetes*, Vol. 47 No. 5, pp. 957-984, available at: https://doi.org/10.1108/K-06-2017-0229

Udanor, C.N. and Bakpo, F.S. (2011), *Artificial Intelligence with Prolog Programming*, JTC Publishers, Enugu, pp. 124-137.

Udanor, C.N., Aneke, S.O. and Ogbuokiri, B.O. (2016), "Determining social media influences of the politics of developing countries using social network analytics", *Emerald Insight Program: Electronic Library and Information Systems*, Vol. 50 No. 4, pp. 481-507.

Vu, T. and Parker, D.S. (2015), "Node embeddings in social network analysis", *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, *Paris*, *August 25–28*.

Walker, S. (1994), *Hate Speech: The History of an American Controversy*, University of Nebraska, Lincoln.

Waseem, Z. (2016), "Are you a racist or am I seeing things? Annotator influence on hate speech detection on twitter", *Proceedings of the 1st Workshop on NLP and Computational Social Science*, pp. 138-142.

Waseem, Z. and Hovy, D. (2016), “Hateful symbols or hateful people? Predictive features for hate speech detection on twitter”, *Proceedings of the NAACL Student Research Workshop*, *June*, pp. 88-93.

Wiegand, M., Balahur, A., Roth, B., Klakow, D. and Montoyo, A. (2010), “A survey on the role of negation in sentiment analysis”, *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pp. 60-68.

Williams, M.L., Burnap, P. and Sloan, L. (2017), “Towards an ethical framework for publishing Twitter data in social research: taking into account users’ views, online context and algorithmic estimation”, *Sociology*.

Yoon, S., Elhadad, N. and Bakken, S. (2013), “A practical approach for content mining of Tweets”, *American Journal of Preventive Medicine*, Vol. 45 No. 1, pp. 122-129, doi: 10.1016/j.amepre.2013.02.025.

Yuan, B. (2017), “Sentiment analytics: lexicons construction and analysis”, Masters Theses No. 7668, available at: https://scholarsmine.mst.edu/masters_theses/7668 (accessed April 18, 2019).

Zhou, Y., Reid, E., Qin, J., Chen, H. and Lai, G. (2005), “US domestic extremist groups on the web: link and content analysis”, *IEEE Intelligent Systems*, Vol. 20 No. 5, pp. 44-51.

Zia, T., Shehbaz Akram, M., Saqib Nawaz, M., Shahzad, B., Abdullatif, A.M., Mustapha, R.U. and Lali, M.I. (2017), “Identification of hatred speeches on Twitter”, *International Journal of Advances in Electronics and Computer Science*, Vol. 4 No. 1, pp. 46-51.

Further reading

Burnap, P. and Williams, M.L. (2014), “Hate speech, machine classification and statistical modeling of information flows on twitter: interpretation and communication for policy decision making”, *Proceedings of the Internet, Policy & Politics Conferences*, *Oxford*.

Carter Butts, T. (2000), *“Social Network Analysis: A Methodological Introduction”. Department of Sociology and Institute for Mathematical Behavioral Sciences*, University of California, Irvine, CA, pp. 92697-95100.

Chen, H. and Chau, M. (2004), “Web mining: machine learning for web applications”, *Annual Review of Information Science and Technology*, Vol. 38 No. 1, pp. 289-329.

Package “sentimentr” (2019), “Calculate text polarity sentiment”, version 2.7.1, available at: https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf (accessed March 22, 2019).

Qiaoyu, T., Ninghao, L. and Xia, H. (2019), “Deep representation learning for social network analysis”, *Frontiers in Big Data*, Vol. 2 No. 2, doi: 10.3389/fdata.2019.00002, ISSN = 2624-909X, available at: www.frontiersin.org/article/10.3389/fdata.2019.00002

Rinker, T.W. (2018), “Sentimentr: calculate text polarity sentiment”, version 2.6.1, available at: http://github.com/trinker/sentimentr (accessed April 22, 2019).

Scott Richard, W. and Davis Gerald, F. (2003), *Networks In and Around Organization*, ISBN 0-13-195893-3, Organizing Pearson Prentice Hall, Upper Saddle River, NJ.

Whillock, R.K. and Slayden, D. (1995), *Hate Speech*, Sage Publications, Thousand Oaks, CA, pp. ix-xi.

Corresponding author

Collins Udanor can be contacted at: collins.udanor@unn.edu.ng