I.J. Information Technology and Computer Science, **2021, 3, 33-52**
Published Online June 2021 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2021.03.03

Building and Annotating a Codeswitched Hate Speech Corpora

Edward Ombui[1]
School of Science and Technology, Africa Nazarene University, Nairobi, Kenya
E-mail: eombui@anu.ac.ke

Lawrence Muchemi[2], **Peter Wagacha**[3]
School of Computing and Informatics, University of Nairobi, Nairobi, Kenya
E-mail: [2]lmuchemi@uonbi.ac.ke, [3]waiganjo@uonbi.ac.ke

Received: 23 February 2020; Revised: 26 August 2020; Accepted: 03 April 2021; Published: 08 June 2021

Abstract: Presidential campaign periods are a major trigger event for hate speech on social media in almost every country. A systematic review of previous studies indicates inadequate publicly available annotated datasets and hardly any evidence of theoretical underpinning for the annotation schemes used for hate speech identification. This situation stifles the development of empirically useful data for research, especially in supervised machine learning. This paper describes the methodology that was used to develop a multidimensional hate speech framework based on the duplex theory of hate [1] components that include distance, passion, commitment to hate, and hate as a story. Subsequently, an annotation scheme based on the framework was used to annotate a random sample of ~51k tweets from ~400k tweets that were collected during the August and October 2017 presidential campaign period in Kenya. This resulted in a gold-standard codeswitched dataset that could be used for comparative and empirical studies in supervised machine learning. The resulting classifiers trained on this dataset could be used to provide real-time monitoring of hate speech spikes on social media and inform data-driven decision-making by relevant security agencies in government.

Index Terms: Annotation scheme, Hate Speech, Dataset, distancing language, Code-switching.

1. Introduction

The negative ripples of hate speech on social media are daily amplified across geographical and legal jurisdictions over the Internet with appalling effects which often abrogate user experience and can easily escalate to off-line hate crimes[2]. Hate speech leads to social exclusion and negatively affects the mental and emotional well-being of the target groups as well as corrupting the thinking, attitude, and actions of the offenders. It instills fear and inhibits public participation on the part of the target group and an equally ballooning resentment that could escalate to physical violence at the opportunity of a trigger event [3]. Presidential elections as trigger events for hate speech are frequent not only in Kenya in regards to ethnic hate speech but also for racism in the USA during the 2020 presidential campaign periods and during president Obama's second term. Religious hate speech was evident too in India during prime minister Modi's second term campaigns. The magnitude of the effects of hate speech have been highlighted with an increased response by way of International, national, and corporate laws and policies to tackle hate speech in public environments, in schools, at the workplace, and in public online spaces like social media network platforms.

This phenomenon is closely associated with sentiment analysis, cyberbullying, body shaming, and other hateful attacks targeting individuals or groups based on belonging to a protected characteristic like race, ethnicity, religion, gender, and more.

Consequently, there is increasingly more interest, within the research community and intelligence agencies worldwide, in mining the new "oil", that is, social media data. This is evident by the growing number of academic conferences and workshops in big data analytics, computational linguistics, natural language processing, machine learning, deep learning, and the ballooning budget allocations by various governments towards monitoring social media activities, especially for purposes of national security. Social media companies are under pressure by various stakeholders to better respond to the online hate speech phenomenon. All social media networks have a user policy on hate speech content on their platforms. However, these companies mostly rely on users to flag such content which subsequently undergoes some element of manual review to establish whether it contravenes their hate speech policy. The respective hate speech policies are general statements that leave a lot of room for subjective interpretation by the reviewers.

The identification of hate speech in short text messages from the voluminous user-generated data on social media

is a nontrivial task. Classifying such unstructured data presents unprecedented challenges to conventional natural language processing techniques regarding extracting high-quality features from the noisy, highly dimensional, and often codeswitched data [4]. Consequently, this study embarked on achieving three objectives: To determine what constitutes hate speech; To build an annotated hate speech dataset from tweets posted during the 2017 elections in Kenya, and to develop an annotation framework for identifying salient features of hate speech in text messages.

Unlike previous studies that view hate speech from a single dimension, this research espouses a novel multidimensional hate speech framework that provides a deep understanding of the hate speech phenomenon. This is to identify salient features of hate in unstructured text data from social media. The study employs the Latent Dirichlet Allocation algorithm [5] during data preprocessing to automatically generate semantically meaningful topics that show the correlation between words and the various hate speech dimensions espoused in the framework. This is significant because it reduces the dimensionality of the input features which subsequently could be used to generate a dense vector representation of the resulting vocabulary to train a machine classification model.

This study builds upon and extends our previous study [6] by developing a multidimensional hate speech framework based on a solid theoretical underpinning. Besides, the framework is then used to inform the manual annotation of a corpus comprising unstructured text data from Twitter. Consequently, the framework proved effective in transforming the unstructured qualitative data into observable variables which could then be analyzed and interpreted quantitatively using computational techniques like machine learning. As a result, the hate speech framework was used to guide the development of a gold standard dataset comprising of ~50k tweets annotated into 3 classes i.e. hate speech, offensive, and neither. This dataset is publicly available (https://www.kaggle.com/edwardombui/hatespeechke) and could thereafter be useful for supervised machine learning experiments to train hate speech classifiers and for future comparative studies.

The rest of the paper is organized as follows. In the next section, we review previous similar research in building and manually annotating text corpora from social media for hate speech classification. In section 3, the methodology used to develop the hate speech annotation framework, data collection, annotation, and cleaning is presented. Section 4 presents the detailed results of the framework development, data collection, and annotation. Lastly, Section 5 discusses the implication of the results, the conclusion, and recommendations for future development.

2. Related Work

The review of the literature indicates a growing number of studies that build and manually annotate corpora for offensive language, sentiment analysis, and hate speech identification. The corpora in most of these studies are in English [6,7,8] and in European languages like Portuguese[9,10,11], German[12], Spanish [13], and Italian[14]. A few studies have built new datasets in Hindi [15], Arabic [16,17], and Amharic [18]. However, no study has built and annotated corpora for codeswitched data, a norm with multilingual communities on social media. Besides, several studies have annotated their corpora using binary categories, with a few studies [7,15] annotating using multiple categories. Our study employs a three-category manual annotation of the corpora using a similar methodology like [8]. However, unlike this study, the annotation framework developed in our study can be used to identify all types of hate speech and considers codeswitched messages in English and Swahili, the official and national languages in Kenya respectively. Although a previous study [19, 20] proposes a framework to help detect offensive tweets, it is however limited on Google's offensive word list and lacks a theoretical underpinning required to enhance research. The latter gap was foundational in our study resulting in the establishment of the multidimensional hate speech framework informed by various theories of hate as indicated in Table 2.

The two authors of the study [8] developed an annotation scheme that comprised of eleven guidelines based on the critical race theory and used these to generally identify offensive and racist tweets. A total of 16k tweets were annotated by the two researchers. They later employed an outside annotator to review the annotations, more so on their disagreement. There is no information given on whether the high inter-rater agreement of $\kappa = 0.84$ includes the outside reviewer's annotations and whether this score was before the involvement of the outside reviewer. Besides, a scrutiny of the eleven guidelines reveals some ambiguity. For example, guidelines number four, six, and seven indicate that a tweet is offensive if it criticizes a minority without a well-founded argument, by using a straw-man argument, or misrepresenting truth respectively. There is a very thin line, if any, in applying these guidelines separately to identify offensive tweets. Besides, the three guidelines could easily map into guidelines number two and three which indicate an attack on a minority or seeking to silence a minority, respectively. Unless specific examples are used under each of these annotation guidelines, it will be quite confusing for an independent team of human annotators to replicate the good inter-rater score. Alternatively, these guidelines could be reduced to the most salient that can easily be recalled by human annotators in identifying offensive language [9]. A subsequent paper by Waseem[21] underscores these concerns when a different team of annotators obtained an agreement score of $\kappa = 0.57$ and a much lower agreement with mean pairwise of $\kappa = 0.14$ considering both annotation groups based on a 42% sample that overlapped [8] previous annotations. Besides, non-English tweets were labeled as noise and consequently filtered out.

The subjective task of hate speech annotation of text messages is further evident by an even lower inter-rater agreement score, Fleiss's Kappa k= 0.17, by amateurs in a similar study by [10] annotating 5663 tweets in Portuguese.

Each message was annotated by three annotators as ground truth, just like in previous studies [10]. The study used the majority vote in their final annotations that resulted in 35% of messages labeled hate speech. Subsequently, these messages were further categorized into finer-classes in a hierarchical structure until no distinct groups could be established. Notably, the advantage of using the hierarchical classification over the flat classification include a deeper understanding of the hate speech sub-categories distinctively, and a better modeling of the relationships between the sub-categories[10]. Besides, the study also established the annotator agreement by these sub-categories which resulted in diverse scores indicating the difficulty in classifying specific types of hate speech as compared to others. Although the study employed the Rooted Directed Acyclic Graph (DAT) to represent the hierarchical structure of the hate speech subtypes and corresponding intersections, information regarding the exact annotation guidelines is inadequate. Further, the use of only one expert annotator to classify all the messages into the hierarchical class structure raises concerns of annotator biasness and ultimately the reliability of the hate speech dataset.

Other studies that gathered and annotated a non-English hate speech corpus include one for German[12], relating to the topic of the refugee crisis in Europe, and for Italian[14], targeting immigrants, Roma, and Muslim minorities. For the German corpus, the authors used ten offensive hashtags liked to the refugee crisis to collect a total of 13766 tweets which after sampling and preprocessing resulted in a corpus of ~500 tweets. The tweets were annotated internally by the six authors. They divided the dataset into six chunks, with each chunk getting annotated by a pair of the authors in rotation. A Krippendorff's inter-rater agreement score of $\alpha = .38$ was obtained. The low agreement score was attributed to the varying backgrounds and personal attitudes of the annotators[12]. A similar explanation was given for the low Inter annotator agreement achieved in the annotation of the Italian corpus comprising 1828 tweets [14].

3. Methodology

The study had three objectives, each of which determined its methodology. The first objective was to develop a deep understanding of hate speech which employed a qualitative content analysis on the various definitions of hate speech. The second objective was to develop a hate speech framework that also used content analysis to examine existing hate theories to identify salient themes of the hate speech phenomenon. The third objective was to build a dataset of hate speech by crawling Twitter social media to collect tweets during the 2017 presidential election period in Kenya. The hate speech framework was subsequently used to derive the annotation scheme to guide the team of human raters to classify these tweets into three predefined classes of hate speech, offensive, or neither.

3.1. Developing a deep understanding of hate speech

To develop a deep understanding of hate speech, several existing definitions of hate speech were methodically reviewed from the literature. The process started by looking at dictionary definitions and legal definitions of hate speech as found in national policy documents like constitutions and public acts. Thereafter, hate speech definitions from International agencies like the United Nations were reviewed. Finally, hate speech, as defined in user-content policy documents on websites of key social media networks were retrieved and compared. Generally, the study employed the content analysis methodology to establish emerging themes or commonalities in the various definitions. The key findings were that hate speech is an expression often comprising of a negative attitude, emotion, or sentiments. These could be emotions of anger, rage, revenge, fear, or hostility directed towards a person or group. Secondly, hate speech expression has a target that it seeks to distance from, whether a person or a group of people belonging to a protected characteristic like ethnicity, race, religion, etc. Thirdly, hate speech has an objective or purpose, which often is to threaten, offend, demean or devalue the target.

Further, a closer focus was drawn on the NCIC definition of hate speech to ensure that the phenomenon is contextualized to the Kenyan case. Besides, the researcher's participation in the hate speech training for human monitors organized by NCIC before the 2017 elections confirmed the seriousness to which the Kenyan government was treating the phenomenon, especially in anticipation of its propagation during the general election campaign period. This further justified the period as the most ideal trigger event for the collection of bigger volumes of hate speech data from social media in the country.

3.2. Building a Hate Speech Conceptual Framework

The objective of the research question here was to build a conceptual framework that captures salient features to identify subtle forms of hate speech in text messages, various theories relating to hate speech were examined. These included the social identity theory, self-categorization theory, speech act theory, the communication theory, critical race theory, Baumeister's theory, the integrated threat theory, the sociologist theory of homophile, and the triangular theory of hate. The concepts drawn from these theories of hate, as shown in Table 2, were used to build a strong theoretical underpinning, and subsequently a hate speech conceptual framework.

Generally, the process started by extracting key dimensions of hate from each theory. These dimensions were qualitatively analyzed and resulted in the identification of three primary concepts, i.e., psychosocial distancing, negative passion, and commitment to hate. These concepts eventually became foundational in building the hate speech conceptual framework. Subsequently, brainstorming was used to identify the seed features under each hate speech

concept in the framework. Besides, the concepts and specific features, which were methodically added, were evaluated by qualitatively analyzing their informativeness in capturing hate speech in sample text messages. The iterative process resulted in a psychosocial feature set, whose feature significance was empirically evaluated through topic modeling as described in section 4.4.

3.3. Data collection

The desirable data for acquisition consisted of text messages generated by Twitter users in Kenya during the country's August 2017 presidential election, plus the repeat election held in October 2017. Before this, Twitter API was used to build an application to collect tweets during the election days. A crawler built on python programming language was also used to complement Twitter API's limitations of two weeks' data collection window to acquire a formidable size of archival tweets comprising of tweets during the three months leading to the general elections and two weeks after the repeat election results were released. Historically, this period and the events surrounding it have been the most prominent trigger events leading to spikes in online hate speech.

The bootstrapping technique was used as a primary data acquisition strategy. This involved the use of seed words comprising of keywords (kw) associated with hate [21], phrase patterns (pp) with a connotation of hate[22], offensive hashtags (#) [20], and pro-hate user account names (un) to crawl social media networks. A summary of the process flow is shown in Fig.1.

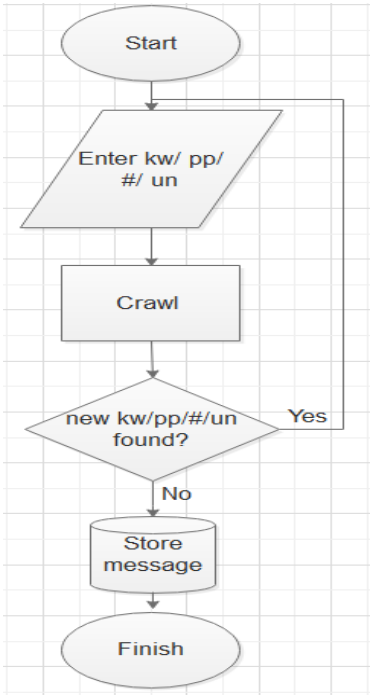

Fig.1. Data collection flowchart

Hateful keywords consisting of insults, profanities, discriminative, and offensive terms popularly used to culturally degrade or devalue a person or people based on their ethnic community in Kenya were used to search for messages on social media. These terms were used because they are likely to return messages containing hateful content and have equivalent terms listed in the hate speech lexicon found on Hatebase.org. Top among these included 'kihii,' which in Kikuyu, a Kenyan native language, is a term used to refer to an uncircumcised person in a devaluing manner. Most of these keywords were picked from tweets that online users had previously flagged as either offensive or hate speech. Therefore, using the snowballing technique, many other devaluing, offensive, and profane keywords were found and used to search for similar tweets. The same technique was employed to find other potentially hateful phrase patterns in the messages. For example, a phrase starting with "All <ethnic name> are. "

Hashtags are a unique feature on the Twitter application that is used to topically structure tweets. This feature was instrumental in collecting potentially hateful messages using 18 hateful hashtags. Some of these included #NoRailaNoPeace, #KenyattaFamilyOfLooters, #KillAllKikuyusToShunTribalism, #ArrestDuale, #itstimewepart, #kikuyuRepublic, #Maafakas, #CORDiots, #Ushenzi, #Kumabfu, #MaviYaKuku, #RailaTheWardogSince82.

Tweets from pro-hate user accounts[23], especially persons of influence in the society like politicians and famous local bloggers previously documented to have posted content bordering hate speech, were collected. First, a list of pro-hate speech politicians and bloggers was compiled as informed by the local dailies [19,20]. This included trending social media hashtags associated with the prolific user accounts. Subsequently, the names on the list were used to search for the verified account handles associated with the respective politicians and bloggers on Twitter. Using a tweet

crawler, these users' tweets were downloaded and saved into a database.

A. Sampling

Convenience sampling was used to collect data from Twitter's social media network. Unlike other social media networks, every post published on Twitter is publicly available and programmatically accessible, unless specified otherwise by the user in their settings. Besides, one does not need an account to access these tweets, and users can anonymously publish, like, dislike, and instantly forward the messages to a wide audience. These features and characteristics make Twitter susceptible and a favorable platform for hate speech propagation.

To create a study sample for annotation out of the big volume of the collected data, our study employed simple random sampling whereby ~50k tweets from ~400k raw tweets were selected for annotation. This sampling technique has been used to generate study samples from social media in previous studies [15,21].

B. Data Cleaning

To maximize the benefit of human annotation, the raw data set was subjected to cleaning to eliminate some noise. This involved the use of the natural language processing techniques like regular expressions (regex) in Python's natural language tool kit (NLTK) library to remove empty rows, duplicate messages, non-alphanumeric data, URLs, and replacing non-ASCII characters with space. The isalpha () function in Python was used to iterate over all tokens and filter out the standalone punctuation [24,25,26]. Spam messages consisting of advertisements riding on trending hashtags were also eliminated.

Besides, the length of the text message was considered in determining admissible tweets. For example, messages that generally had three or fewer characters were dropped. These mostly comprised of tweets that contained a single word or few characters which by themselves were contextually ambiguous. Moreover, these kinds of short messages contravened the guidelines enumerated in the annotation scheme for labeling a message into the predefined classes

Tweets in English, Swahili, and codeswitched text containing words from several Kenyan ethnic groups were included. There were a few other tweets in non-native languages that were removed as part of the noise signals that do not add value to the classification task.

Moreover, to protect the user identity of the message recipients and authors, all user mentions, for example, @martins, were replaced with a generic "USERNAME" tag, whereas the URL part that often contains the account names was filtered out. These were achieved by using the regular expressions library in Python.

3.4. Data annotation

A team of forty human annotators was recruited and trained on annotating the messages into three classes i.e. hate speech, offensive, and neither. The team comprised of undergraduate computer science students and members of staff in the ratio of 80:20, respectively. Convenience sampling was employed to get the final annotation team from the school of science and technology at Africa Nazarene University (ANU). The team's average age was twenty-three and consisted of a relatively balanced gender of 21 male and 19 female annotators. The nationality of the team members was skewed towards Kenya. The skewness was informed by the need to have annotators who could easily interpret the codeswitched nature of the corpus which comprised of messages in English, Swahili, and some other native languages in Kenya. The first training was based on the annotation scheme to establish a shared understanding of hate speech across the entire team. After that, the annotators were trained on the hate speech framework and how to annotate sample messages using a web-based annotation portal [27]. Moreover, the HS framework was displayed in the training room on an overhead projection. Besides, two subject matter experts were available in the training room during the preliminary annotation to help classify any ambiguous messages encountered by the amateur annotators.

The initial team of forty annotators was later trimmed to twenty-seven annotators. The selection was based on the individual performance and a signed commitment to annotate a target of at least three thousand messages for one week. The remuneration was pegged on meeting the target number of messages for the specified period, otherwise, no cash was awarded.

The annotation portal was designed to display one message per view, each annotated by a random team of three annotators. Fundamentally, once a third annotator picked the message, the system automatically locked it by making it inaccessible for annotation. The hate speech definition persisted above each new tweet to remind the human annotator of the shared definition. Four questions were asked under each tweet. First, to classify the tweet into the three predefined classes. The annotation options were presented as radio buttons that enable only one choice. By default, the choice was 'none'. The second question rated the choice of the first. The third question was intended to further code the type of hate speech identified. This choice allowed for multiple labels. The last question was used to identify the key feature or features of hate speech as defined in the conceptual framework. Similarly, the question allowed multiple labels. This is as shown in fig.2.

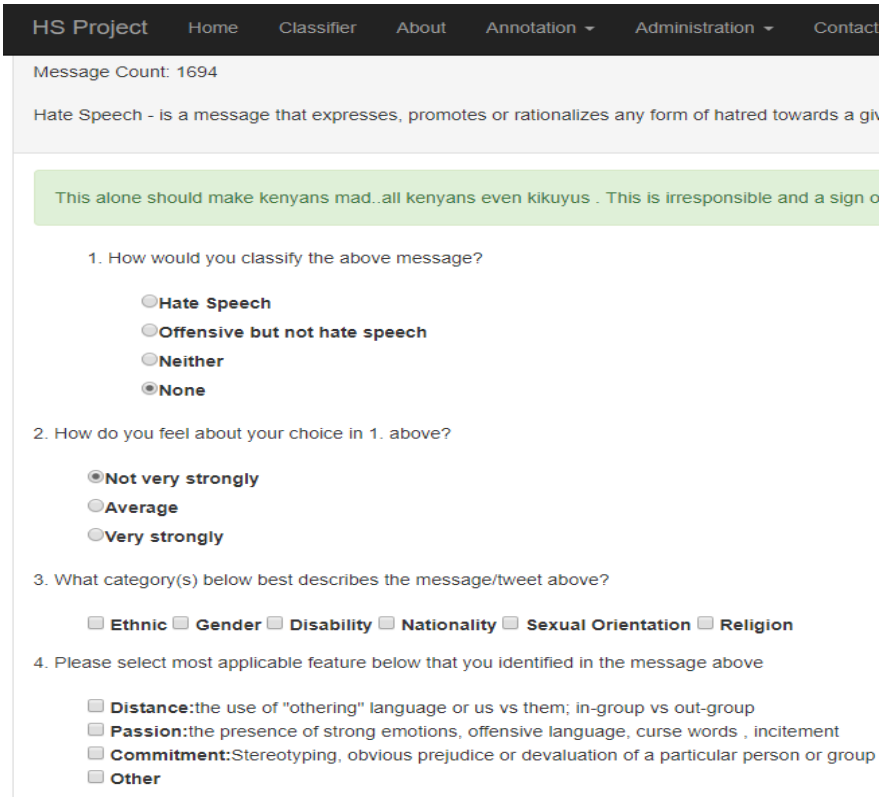

Fig.2. The annotation portal used to label each tweet

Valuable feedback regarding the speed of the annotation portal was received from the preliminary annotation of ~1k messages. Therefore, questions 2,3, and 4 in Fig. 2 were dropped. The new design was informed by the slow annotation process in the preliminary session and the need to expedite the annotation process to have a bigger labeled dataset that could subsequently be adequate to train a machine classifier. Besides, this was hoped to better utilize and maximize the expertise of the team of human annotators within the short period of one week. The reliability of the annotations, regarding establishing the extent to which the team agreed on the class for each message was initially set to be measured using the Krippendorff's alpha [28], because it could accommodate any number of human raters and handle incomplete data, even with relatively small data samples.

A message's class was generally determined by a majority vote. If there was no consensus, a fourth annotator comprising of a subject matter expert would act as the tie-breaker to determine the class.

3.5. Evaluation

The objective of the evaluation was twofold: To link the collected data to the hate speech concepts defined in the multidimensional framework, and to assess the reliability of data annotations. Topic modeling[5] based on the Latent Dirichlet Allocation (LDA) model was used to establish whether our code switched corpus contained the deep underlying concepts of hate espoused in our hate speech framework. LDA, a hierarchical probabilistic model, has successfully been used previously to identify topics related to cyberbullying [29]. LDA was used to model each word in the corpus as a finite mixture over a set of underlying topics e.g. Passion, Distancing, etc. which, in turn, could be modeled over an infinite possibility of topics representative of the hate speech corpus [5].

The reliability of the data annotations was initially determined by Krippendorff's alpha, an inter-rater reliability score [28].

A. Ensuring Validity and Reliability

Content validity was ensured by using three human-raters to label each message based on the annotation scheme implemented on the annotation portal. The scheme was informed by the study's hate speech framework that was grounded on the duplex theory of hate [1]. Besides, the definition of hate speech persisted above the frame that displayed each new message for annotation on the annotation portal as evident in Fig. 2.

An inter-rater reliability score was calculated based on the annotations done by a team of 27 human annotators. Each tweet had to be annotated by at least 3 human annotators. The statistical mode was the determining factor for the class of the tweet, meaning that the class of the tweet was determined by two or more votes. In case of a tie, implying that there was no agreement among the team of three human annotators, a fourth annotator would be introduced as a tie-breaker, who ideally was a subject matter expert. The Krippendorff's Alpha was chosen as an inter-rater reliability

measure for the annotation exercise comprising of the team of 27 novice annotators because it could deal with missing values and robust to deal with outliers [28]. To further validate the reliability of the novice annotators, a second annotation comprising of one subject matter expert was carried out on 9k sampled tweets. The Cohen Kappa was used to score the reliability of the annotations.

The construct and predictive validity of the research data and framework features were established through the triangulation approach. This involved comparing performance results from various conventional and deep learning machine learning algorithms to determine the best feature set to train our classifier.

3.6. Ethical Considerations

There are some ethical practicalities of using social media as the primary source of data for research. These online platforms, like Facebook, Instagram, and Twitter, are increasingly being used by people of different demographics to share opinions, feelings, and intimate sentiments. This, therefore, raises two primary concerns on user consent and user identity protection when collecting such kind of data. In the first case, the issue of user consent for messages posted on social media, specifically Twitter, has been debated previously[30]. However, unlike the other social media platforms that are private by default, messages posted on Twitter are publicly accessible by default unless the user turns on the privacy settings, which only allows users who follow them to access their tweets. This is the reason why a lot of academic research has been conducted using public tweets [31]. Needless to say, it will be practically impossible to get user consent from user accounts that generate thousands and possibly millions of tweets that could be collected using either the Twitter streaming API or even archival tweets [31]. Besides, tweets could be posted anonymously, or users would have left or deleted their accounts, but the retweets could still be available. This, too, makes it unfeasible to reach out to get consent, if at all that would have been necessary. In this regard, the study focused on collecting only public tweets and retweets which do not need any formal consent or ethical approval.

Regarding the issue of user identity protection, all user names and mentions were replaced with a generic USERNAME label to protect the identity of online users. Only tweet IDs will be used to publicly share the dataset following the Twitter privacy and data sharing policy [32].

4. Results and Discussion

The results were based on the three research objectives which are presented in sections 4.1, 4.2, and 4.3 respectively.

4.1. Developing a deep understanding of hate speech

To gain a deeper understanding of what constitutes hate speech, several definitions of hate speech were reviewed, including dictionary definitions, legal definitions, and hate speech definitions on user policy documents on social media networks. Content analysis was conducted by highlighting similarities and differences in these hate speech definitions, as shown in Table 1.

Table 1. Content analysis of hate speech definitions

Source	Reference to											Target Attributes									
	Violence	Attack	threatening	prejudice	Insult/Abusive	Offend	Intimidate	Incite	Discriminate	Intolerance	Degrade	Race	Ethnic	Religious	Sexual orientation	Disability	Nationality	Gender	Disease	Political	Age
Oxford dictionary		x	x	x	x							x		x	x						
Oxford English dictionary								x		x			x	x	x						
Merriam-Webster					x	x	x					x		x	x	x	x				
UN's International com											x										
European Court of Human Rights								x	x	x		x									
Kenya NCIC Act of 2008	x		x	x	x			x				x	x				x				
BCC South Africa	x							x				x	x	x	x	x	x	x			x
YouTube –	x							x				x	x	x	x	x	x	x			x
Facebook	x	x	x						x			x	x	x	x	x	x	x	x		
Twitter	x	x	x					x				x	x	x	x	x	x	x	x		x
LinkedIn	x	x	x						x			x	x	x	x	x	x	x	x	x	

Besides, the verb frequencies and targets of hate derived from the content analysis exercise were aggregated, as shown in Fig. 3. A higher number of definitions viewed hate speech as inciting or threatening speech.

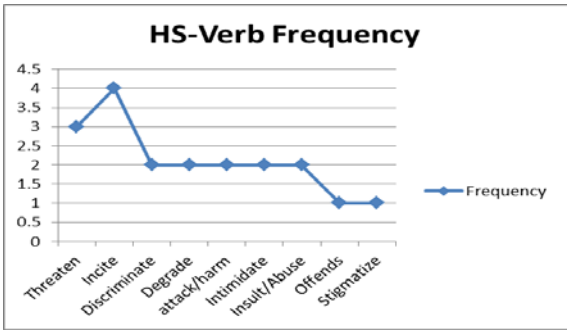

Fig.3. Frequency of verbs used in hate speech definitions

Concerning the specific content and salient characteristics, hate speech is intended to provoke hatred, violence, and prejudice. This is shown in Fig. 4.

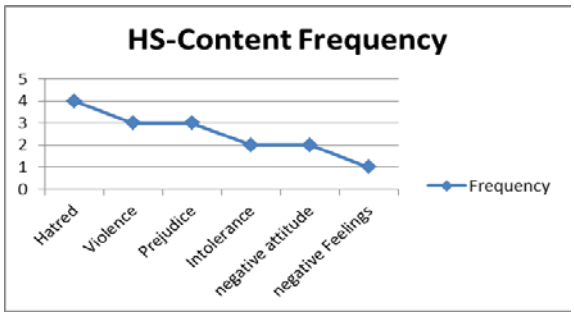

Fig.4. Hate-specific content frequency

A. Findings

From the analysis of the various definitions, the finding was that hate speech has three key facets: First, it is an expression, whether nonverbal through body signs or verbal through oral or written format including text, images, or graphics that threatens, incites, discriminates, degrades, attacks, intimidates, insults, offends or stigmatizes. Secondly, hate speech expression has a target that it seeks to distance from, whether a person or a group of people belonging to a protected characteristic like ethnicity, race, religion, etc. Thirdly, hate speech has an objective or purpose, which often is to threaten, offend, demean or devalue the target.

There was no universal definition of hate speech [11,19]. Besides, it was observed that most of the hate speech definitions were derived from the legal perspective as enshrined in respective country policy documents. Therefore, the study's working definition of hate speech encapsulated the NCIC definition, i.e., Hate speech is any communication that expresses distancing language (prejudice, discrimination, or hatred) targeting an individual or a group based on their membership to a protected social category (including race, religion, gender, ethnicity, nationality, sexual orientation, or disability).

Fundamentally, it is only after a proper understanding of the hate speech phenomenon and its characteristics that it can then be easily defined and provide valuable insight on how to identify it automatically.

4.2. Building a Hate Speech Conceptual Framework

Besides, existing hate speech studies and theories of hate from the field of sociology and psychology were analyzed and some constructs were identified. The goal here was to understand the nature of hate and how it manifests itself in text messages. This was on the premise that there is a relationship between word usage in written text and social psychology characteristics of hate [33]. In this regard, the social identity theory, self-categorization theory, speech act theory, the communication theory, critical race theory, Baumeister's theory, the integrated threat theory, the sociologist theory of homophile, and the triangular theory of hate were analyzed, the general and specific constructs identified, and applicable hate speech studies mapped. This is as summarized in Table 2.

Table 2. Constructs from qualitative research on high-level features

Category	General Construct	Specific Construct	Description	Theory	Studies
Psychosocial (High-level)Features	Distance	Othering	Us vs Them; In-group vs Out-group	Self-categorization theory	[BW16]; [DAKAA15]; [BW14]; (Van Dijk 2002, p.150) Coupland,2010; UMATI project
		Prejudice – Stereotypes	Implicit biases, -ve stereotypes (an over-generalized belief about a particular category of people) Anti-"group"	Social Identity theory	[Warner & Hirschberg, 2012] Waseem&Hovey,2016;
	Passion	Emotions Negative Polarity/ sentiments	Anger, Fear, Disgust, Contempt	Integrated threat theory	Dinakar et al. 2012;Chen et al,2012; [BW16];Nobata et.al 2016;Spertus,97; Stephens,2013;Gitari et.al,2015 Swati & sureka,2015;Ting et.al,2013; Warner w,Hirschberg,2012
		Derogatory language	Insults, Abuses, Offensive		Nobata et.al,2016; Spertus,1997; Mahmud et.al,2008; Chen, Ying, Yilu Zhou, Sencun Zhu, and Heng Xu (2012) Sood et.al,2012; Razavi et.al,2010; Xu z, Zhu s,2010
		Incitement	Call-to-Action to harm target	Speech act theory	UMATI Project
	Commitment	Devaluation	Comparison to animals, insects, things	Susan Benisch Framework for	UMATI project
		Stereotyping	Negative attitude towards target	Baumeister's theory (revenge)	s [WH12a] a [WH12b] [SMC+16]
	Hate as a story	Prejudice – Stereotypes	Implicit biases, -ve stereotypes (an over-generalized belief about a particular category of people) Anti-"group"	Social Identity theory	[Warner & Hirschberg, 2012] Waseem&Hovey,2016;

The triangular theory of hate [1], unlike the other hate theories, was found to be most comprehensive and general enough to accommodate the multiple dimensions of hate which were independently expressed in the other theories as summarized in Table 3. Therefore, the theory was considered most appropriate because it had the highest explanatory power regarding the hate phenomenon. Besides, from the various definitions, it is apparent that hate speech has a clear target; otherwise, the message will be considered to belong to the offensive class. Therefore, the three factors of words that express distance, negative passion, and commitment to hate seamlessly translate to the salient concepts or variables that would inform whether a message could be humanly identified and classified as hate speech, offensive, or neither.

Based on empirical results obtained from the qualitative analysis on sample hate speech messages labeled by human annotators, it was further noted that the mere presence of one concept could not adequately discriminate a message into the positive class, i.e., hate speech. For example, the concept of distancing language could be indicated by the presence of pronoun dichotomies in a message, whereas negative passion could be indicated by the presence of words alluding to negative sentiments or offensive language. However, independently, these concepts could not qualify a message as hate speech. For example, "*We will not accept hawa wasee to treat us like shit kwa nchi yetu*". The codeswitched message contains pronoun dichotomies, i.e., *"We," "hawa," and "us"*, and an offensive term, i.e., *"shit"*. For a human annotator, it is apparent that the author of the message is aggravated and the message is emitting negative passion. However, the target of the hate, i.e., *"wasee"*, is not clear and cannot be established as belonging to a protected characteristic like ethnicity, nationality, religion, etc., to be positively identified as hate speech. Therefore, the concept of distancing language which is ideally indicated by pronoun dichotomies had to be further qualified by clearly establishing the target subject as belonging to a protected social group. This was best captured by the concept of stereotyping as indicated by the team of human annotators and elaborated in Baumeister's theory of revenge. Besides, the team of human annotators also pointed out the deficiency of the initial three concepts in exhaustively capturing the concept of bias or propaganda targeted at individuals or groups sharing a common social characteristic. Therefore, these two additional concepts, having gone through a systematic qualitative analysis involving human annotators, coupled with the idea of concept intersection, helped to generate the multidimensional framework summarized in Table 3 and illustrated by the Venn diagram in Fig.5.

In regards to operationalization, the five variables were measured by their respective term frequencies-inverse document frequencies (TF-IDF). The specific variables under each concept were primarily drawn from the set of emotional, cognitive, and psychological word lists available in the LIWC2015 dictionary [33]. Examples of these are shown in Table 3.

Table 3. The summary of the concepts

Concept Name	Description	Indicators	Examples of variable
Distancing	Negation of Intimacy by the use of othering language	High pronoun usage in the text, especially third person plural nouns	They, them, their, she, he, us, we
Negative passion	The use of negative sentiments and offensive language	Emotions of anger, use of offensive, insulting, threatening, sexual, and swear words	Damn, fuck, piss, kill, stop, hate, annoying, ugly, nasty, horny, uncircumcised
Devaluation	Commitment to hate the target by use of demeaning language	Use of subhuman, object, animal, or insect names to degrade a person(s)	Cockroach, maggots, Rats, dog, bitch, fish, madoadoa, bitch, pussy, foreskin
Subjectivity	Use of faulty arguments	Bias & propaganda using quantifiers and certainty	Always, never, all, many, much
Stereotyping	Hate directed on the target based on a protected social group	Presence of ethnic, racial, religious names	Kikuyus, Luos, Merus, Kalenjins, Luhyas, Kambas, Kisiis, Maasai, Muslims, Hindus,

Consequently, the task of a human rater was to annotate a text message into the three classes, i.e., hate speech, offensive, neither, based on the scheme in Table 4 as derived from the hate speech conceptual framework in fig. 5. The concept of distancing could manifest as discrimination and othering language. Discrimination is primarily based on reference to protected social groups, including ethnic group names like Kikuyus, Luos, Kalenjin, etc. Othering language category includes common noun pairs like “us - them; we – they.” Negative passion consists of insults, threats, pejoratives, and other offensive terms like “Fuck, stupid, kill, chase, etc.”. Subjectivity consists of biased arguments that are one-sided and cannot be substantiated and therefore become propaganda. Commitment to hate is often evident by expressions of certainty, generalization, and devaluation. The presence of words like ‘never,’ ‘always,’ in a message are clear indicators of certainty. Generalization is evident in phrase patterns that start with ‘all <tribe> …’. Devaluation consists of the use of dehumanizing terms to refer to the target. For example, the use of insect, object, or animal terms like “maggots, cockroaches, foreskin, etc.”. These are well summarized in the multidimensional hate speech framework in Fig. 5.

The multidimensional hate speech conceptual framework indicates ten instances in which a message could be considered to be hate speech. The first nine concept combinations directly reference a protected social group, whereas the tenth one is an indirect reference, often obscured in a devaluing word known to the in-group membership. Out of these, one involves a five-concepts overlap, five involve a four-concepts overlap, three involve a three-concepts overlap, and another one involves a two-concepts overlap.

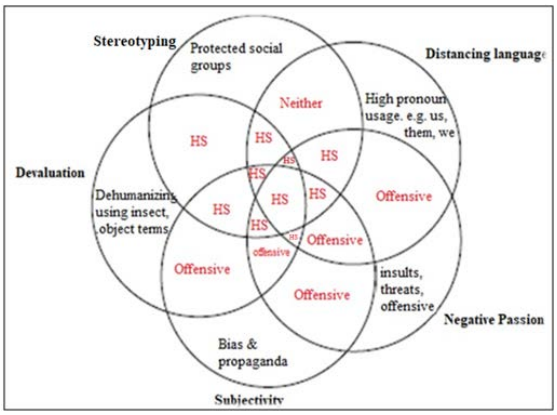

Fig.5. Multidimensional Hate Speech Conceptual Framework

Therefore, given a tweet, the human annotator looks for indicators of distance (D) and passion (P) or commitment (C). Hate speech is based on D+P or D+C or D+P+C, whereby distance is targeting a person or group based on them belonging to a protected characteristic like ethnicity. For example, “Uhuru Kenyatta is a hopeless drunkard. We are tired of this Kikuyu president “. The reference to the president being “Kikuyu,” an ethnic group, makes this outrightly hate speech. The complete list of feature combinations that result in hate speech is summarized in Table 4.

Table 4. Multidimensionality of Hate Speech

No	Class	Concept Combination
1	Hate Speech	Distancing + Stereotype + Devaluation
2	Hate Speech	Distancing + Stereotype + Negative passion
3	Hate Speech	Distancing + Stereotype + Negative passion + Devaluation
4	Hate Speech	Distancing + Stereotype + Negative passion + Subjectivity
5	Hate Speech	Distancing + Stereotype + Negative passion + Subjectivity + Devaluation
6	Hate Speech	Stereotype + Negative passion + Subjectivity + Devaluation
7	Hate Speech	Stereotype + Subjectivity + Devaluation
8	Hate Speech	Distancing + Stereotype + Subjectivity + Devaluation
9	Hate Speech	Devaluation + Stereotype
10	Hate Speech	Distancing + Negative passion + Subjectivity + Devaluation
11	Offensive	Devaluation + Subjectivity + Negative passion
12	Offensive	Distancing Language + Negative Passion + Subjectivity
13	Offensive	Distancing Language + Negative Passion
14	Offensive	Negative Passion + Subjectivity
15	Offensive	Negative Passion + Subjectivity +devaluation

Offensive, just like hate speech, could be based on the three different combinations but not referencing a protected social characteristic, whether directly or indirectly. For example, “Uhuru Kenyatta is a hopeless drunkard. We are tired of this guy “The premise will be treated as offensive but not hate speech.

Any other message falling outside of these boundaries was considered “neither.” In principle, class learning is optimum when features are specific to a class and not universal. The feature description is shared by all instances of a class and none with other competing classes [34]. However, analysis of the covariance structure of the unigrams, bigrams, and trigrams concerning the three classes using the Chi-square significance test, revealed a different pattern than earlier thought. Ethnic names frequently appeared across the three classes with Kikuyu and Luo (including their equivalent Swahili language translations, i.e., Wakikuyu, Wajaluo) being the most frequent respectively. This is as shown in Fig. 6. Therefore, this meant that ethnic names, by themselves, were not a powerful feature to use to train the classifier to discriminate between the classes. This was contrary to our initial thought; however, the ethnic names were valuable when used in combination with the other concepts, especially in identifying the target of hate; therefore, they could not be entirely discarded.

Fig.6. Correlation of terms to classes using chi-square

4.2. Building the Hate Speech Dataset

A total of 401211k raw messages were collected and stored in a comma-separated file (CSV) format. These consisted of text messages, i.e. tweets, from the general elections in Kenya in August 2017, including a repeat election that was conducted 60 days later, in October 2017. Out of these, ~60k tweets were randomly sampled for annotation.

Online hate speech has been known to spike immediately after a major event affecting a largely targeted populace, for example, a terrorist attack, or the periodic presidential campaigns [35]. Based on this, the choice of the 2017 presidential election campaign period in Kenya became an ideal data collection period for hate speech. First, the country has a history of the perpetuation of negative ethnicity during past presidential elections. Besides, the August 2017

elections presented a unique and prolonged data collection period occasioned by the repeat elections in October 2017. Secondly, existing research indicates that there is often a higher volume of hate speech generated during trigger events; in this case, presidential campaigns, and not much after that [3]. A sizeable dataset comprising approximately 400k messages was collected during the 2017 general election period.

The research hypothesized that hate speech could be effectively crawled online by using the combination of problematic hashtags, pro-hate user accounts, offensive words, and phrase patterns.

The analysis of the hate content revealed the effectiveness of the use of problematic hashtags, offensive terms, pro-hate user accounts, and phrase patterns in scrapping the ~400k tweets from the January to December 2017 election period. Unlike text messages from other social media networks, tweets were purposively chosen because they are often topically structured, publicly available, and programmatically accessible via Twitter APIs [36], python tweet collection libraries, and even using custom-built crawlers. First, it was possible to collect and build a big dataset of text from the publicly available tweets, unlike most of the other social media. By this, we mean that we did not have to have a Twitter account to access public tweets, whereas we had to create an account in the other social media networks to access data, which was constraining. Secondly, Twitter data was programmatically accessible using a tweet crawler and an application built using Twitter's API. Thirdly, the use of hashtags enabled the collection of all related tweets to a given topic. For example, the hashtag #killallkikuyus generated a lot of hateful responses. Also, the nature of the platform allowed wide participation that covered all demographics, of which the minority would otherwise not have had a voice in the conventional platforms. Besides, several previous comparable studies in hate speech have used Twitter data [8,15,18,37,38].

Data preprocessing involved the filtering of tweets in Kenyan non-native languages apart from English. For each tweet, only the ID and the text message sections were retained, whereas all other parts like the dates, URL, and user account name were dropped because they often do not make a significant contribution to the information required to classify a tweet [39]. Although the tweet ID does not add valuable information, it was, however, retained so that the dataset could be shared publicly as tweet IDs, that is, in conformity to Twitter data sharing policy[32].

A. Data Annotation

A total of 152403 annotations were done by the team of 27 amateur annotators on 50994k tweets out of the sample of ~60k tweets that were availed for annotation. Out of these, 50656 tweets, that is 99.7%, were each annotated by a random team of 3 annotators. Besides, 97 tweets were each annotated by a team of two annotators, and 241 tweets were labeled by only one annotator. The discrepancy in the team annotations was due to the initial annotations that were captured during the training of annotators. Only the 2 and 3 rater-team annotations were considered while the 1 rater annotations were dropped. This was to avoid introducing further human bias and keep the data as reliable as possible by using only annotations done by a team. A majority vote was used to decide on the class of the message. However, 3156 messages did not have a majority vote, representing 6% of the annotated messages. These were dropped resulting in a new total of 47838 labeled messages. Out of these, 3125 messages were labeled hate speech, 8379 messages were labeled offensive, while 36334 messages were labeled neither. The class distribution for the annotated dataset was unbalanced, with the majority class being the "neither" class and the minority class being hate speech. A summary of the class distribution is presented in Table 5.

Table 5. Class distribution

Class	Description	Count	Percentage
0	Hate Speech	3125	6.5%
1	Offensive	8379	17.5%
2	Neither	36334	76%
Total		47838	100%

This distribution was unsurprising because it is an actual representation of the population of messages posted on social media and is consistent with results from previous similar research [40]. One of the findings here was that ethnic hate speech is the predominant type of hate speech during election campaign periods in Kenya. Secondly, unlike binary classification, the introduction of the "offensive" class helped to clearly distinguish between hate and offensive messages, thus reducing the chances of mislabeling tweets as hate speech, a common flaw during annotation exercises [19].

There were a total of 23554 messages that had a full agreement among the team of annotators, which translates to 46% of the total annotated messages. Out of these, 3.5% was labeled hate speech, 6.5% offensive, and 90% neither. A further 24284 messages had a majority vote of 2 out of 3 raters, which is approximately 48% of all the annotated messages. Out of these, 10% was hate speech, 28% offensive, and 62% neither. Besides, 3156 messages, representing 6% of the annotated messages, did not have a majority vote. This is well summarized in Table 6.

Table 6. Summary of the Annotations

Annotation Agreement	Hate	Offensive	Neither	Total
Full Agreement by all raters	830	1524	21200	23554
Majority 2 out of 3 raters	2295	6855	15134	24284
No agreement	3156			
Total Messages	3125	8379	36334	50994

The inter-annotator agreement was initially intended to be calculated using Krippendorff's Alpha because it measures the extent of agreement of any number of raters and allows for missing data[28]. Krippendorff's Alpha assumed that a set of messages are rated by the entire team of annotators, in our case, the group of 27 annotators. However, this would result in substantial effort on the part of the annotators and consequently diminish the annotation output without making significant improvements in the annotation reliability. Therefore, the inter-rater agreement could not be performed using Krippendorff's alpha. This challenge was also noted in a previous similar study [37]. With this hindsight, the annotation portal was designed to assign a message to a random team of 3 annotators, as compared to involving all the annotators in rating each message in the dataset. Moreover, the need to have a significant amount of annotated messages was more desirable for purposes of supervised machine learning.

Low inter-rater reliability scores have, in previous studies, been attributed to several reasons including the use of affordable but inexperienced annotators, personal sensitivities and biases, plus the lack of a clear annotation scheme [12,21]. When annotators erroneously label a message as hate speech when it should not be, or vice versa, this introduces noise signals, specifically teacher noise[34]. This was evident in our study, where some of the annotators who were colleagues, did not attend the full annotation training due to work constraints. We found that the involvement of colleagues in research sometimes becomes a challenge, mainly if they are primarily motivated by monetary incentives attached to the research activities. Generally, the teacher noise coupled with the tacit knowledge and biases they come with during annotation, despite the training, could form part of the latent attributes modeled as random components in the noise signal.

Generally, the size of the annotated dataset in this study, comprising of approximately 48k useable tweets, was not only adequate but surpassed the size of datasets used in previous similar studies in hate speech that had 13k [12], 16k [21], and 21k [19] respectively. Besides, future work would consider including a stricter annotator recruitment criterion and an extended training session to exclude outliers. Moreover, the annotation activity could use an iterative approach, whereby the messages that are selected could be repeated randomly in different cycles to evaluate whether the human annotators are consistent with the specified annotation scheme. This will be vital in identifying and eliminating outliers for purposes of improving inter-coder reliability and, subsequently, the performance of machine classifiers trained on that labeled dataset. That notwithstanding, it might be worth establishing, considering, and accommodating the beliefs, values, and theories already held by the human annotators concerning the phenomenon under study at the onset, rather than imposing an annotation scheme based solely on existing literature or methods, and the researcher's assumptions. This could generate a more realistic inter-coder reliability performance. Besides, the annotation tool [27] was designed to overcome the limitations of Krippendorff's[28] inter-coder reliability methodology that results in higher costs and slower annotations when subjected to the annotation of big datasets. The annotation task demonstrates how challenging the classification task is even for human annotators.

4.4. Data Exploration

Generally, the annotated dataset consisted of English, Swahili, and codeswitched messages, with English-Swahili forming the bulk of the code-switched messages. For example,

"*Kisiis are the weirdest people well apart from their constant noise making their men are very stingy while their women are spendthrifts*" (Example 1)

"*Some kyuks think Luos are chickens inafaa tu wachinjwe some people should have never been born shule ni muhimu pia*!" (Example 2)

"*Hio ya kirumi kiria giatigirwo bururu uyu gitingireka tuatho ni ihii means we cannot be ruled by luos who are uncircumcised*" (Example 3)

Example 1 demonstrates a hate speech message that is purely in English. The generalization and use of adjectives associated with a negative connotation to describe the membership of the Kisii ethnic group makes it categorically hate speech.

Example 2 demonstrates English-Swahili codeswitching while example 3 is in Kikuyu and English. The bold text in example 1 in Swahili translates to "they should be slaughtered" and "school is important", respectively. In the same example, the coined term "Kyuks" is used to refer to the Kikuyus, the largest ethnic group in Kenya.

This dataset was later cleaned by removing stop words apart from the English and Swahili pronouns, which according to our hate speech framework could be indicative of "othering" language. The resulting histogram, as shown in Fig. 7, indicated ethnic group names like Kikuyus, Luos, and Kalenjins as most frequent, whereas the presidential

contender's names of Uhuru and Raila, plus popular bloggers names occurring frequently too. Besides, other frequent words included passion words e.g. hate, kill, and like, and hashtag names like #noreformsnoelections, #luolivesmatter, and #electionboycottke.

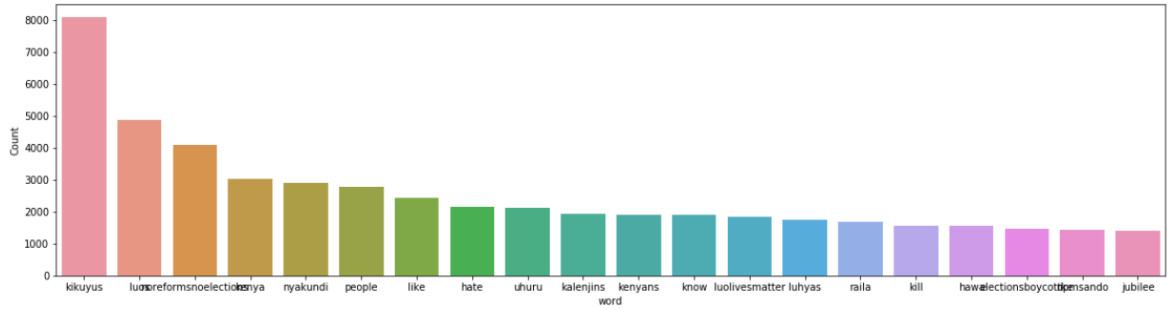

Fig.7. Hate Speech histogram

Besides, based on word frequency per the class distribution, it was apparent that the ethnic group names were most frequent across the three classes with Kikuyu, Luo, Kalenjin, Luhya, Kisii, and Kamba being most frequent respectively. This is as shown in the word cloud in Fig. 8.

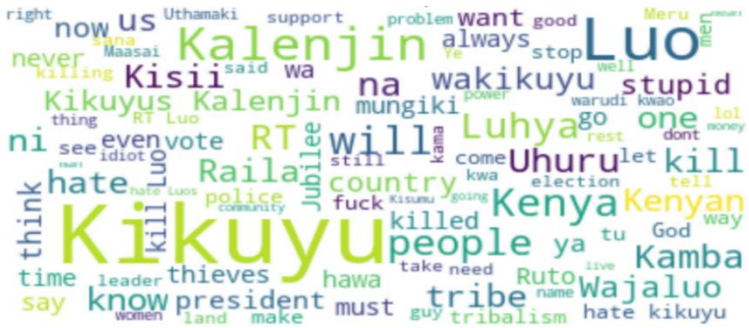

Fig.8. General Word frequency word cloud

After excluding ethnic group names, the new word frequency for the hate speech labeled messages was now dominated by the PDC words like "kill," "thieves," "stupid," " hate," etc. This is as shown by the word cloud in Fig.9.

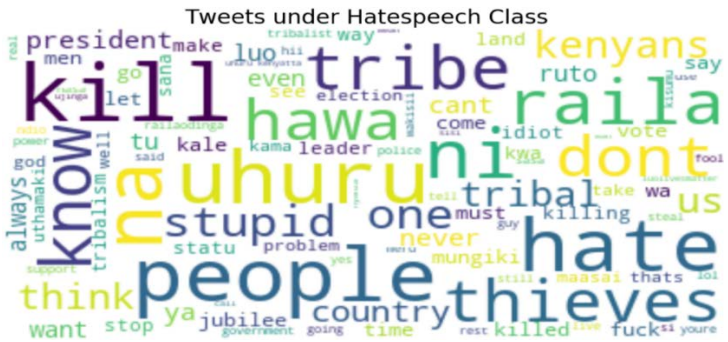

Fig.9. Word frequency under the Hate Speech class

The systematic review of government documents consisted of the constitution and other legal regulations that encompass hate speech. Hate speech in Kenyan law is under articles 33(2) subsection (c) and (d). It is also referenced with the freedom of expression [41]. Besides, hate speech is also reiterated in the penal code, section 77(3)(e), and the NCIC act section 13 that outlaws discrimination based on ethnicity [42].

Negative ethnicity is the most rampant type of hate speech in Kenya. The peak of ethnic hatred recorded in the country was the post-election violence witnessed in the country in 2007, immediately after the announcement of the presidential results. The media's contribution to the incitement that heightened the tension proceeding the violence was evident [43].

Further, using the participant observation approach, the researcher was able to acquire a more in-depth understanding of hate speech markers and the methodologies that the concerned authorities had planned to use to capture the evidence of hate speech. This was done by taking part in the March 2017 hate speech human monitors training workshop that was conducted by the National Cohesion and Integration Commission (NCIC)[44] in

conjunction with the communication authority of Kenya (CA), and the Kenya Police in preparation for monitoring and collecting evidence for hate speech during the campaign periods that preceded the 2017 general elections [45]. The monitoring was mostly manual based on recordings of political rallies using voice recorders, video cameras, and manually perusing through popular social media platforms.

Besides, the researcher's interactions with government agencies like the National Cohesion and Integration Commission, which is in charge of matters related to hate speech [44], and the Kenya Education Network that is the primary Internet Service Provider of all tertiary learning institutions in the Country [46] were able to offer more insight into the challenges of monitoring the phenomenon. From this phase, a working definition of hate speech was derived, that is, any message that discriminates devalues, or uses offensive language targeting a person or a group of people based on belonging to a protected characteristic like race, ethnicity, religion, gender, etc.

Exploratory data analysis further helped reveal the unigrams, bigrams, and trigrams that were most correlated to the hate speech class (0), offensive class (1), and Neither class (2). According to the study's conceptual framework, the presence of terms that express distancing or othering language, negative passion, commitment to devaluation, and propaganda were most characteristic of hate speech. For example, negative passion is evident in the terms *"stupid," "fuck," 'kihii,'* etc. Distancing words are evident by the frequency of pronoun terms *"Hawa," "wewe,"* and the tribe names. Commitment to propaganda hate is evident by trigram like *"Uhuru steals everything," "Yule jamaa wa vitendawili,"* etc. Besides, the codeswitching phenomenon is evident by the presence of several Swahili, and native terms like *"hawa', ni wajinga sana,' 'kihii,'* etc. is apparent in fig. 6 (correlation without ethnic groups). *'ni,' 'na,''ya,' 'wa,'* were frequent but these consist of the English equivalent Stopwords "is,'' and,'' of,' respectively. This is an excellent example of how existing standard libraries like Stopwords, will not be able to capture such noise in codeswitched text data. These Swahili Stopwords were added in the inclusion set of Stopwords during data cleaning and therefore dropped because they were not adding any valuable information to the classification process.

4.4. Evaluation of the HS Framework

Topic modeling[5] based on the Latent Dirichlet Allocation (LDA) model was used to find deep underlying concepts of hate in our code-switched dataset. LDA, a hierarchical probabilistic model, has successfully been used previously to identify topics related to cyberbullying [29]. LDA models each word in the corpus as a finite mixture over a set of underlying Passion, Distancing, and Commitment (PDC) topics, which in turn are modeled over an infinite possibility of topics representative of a text document [5]. This helps to establish a probabilistic model over the codeswitched corpus that will assign high probabilities to messages closely linked to the membership of the corpus and other messages that are similar to these. Therefore, LDA was explicitly used to extract a "bag of words" into twenty-three latent topics closely associated with the hate speech class and bearing the feature characteristics of the study's conceptual framework. These are as shown in the twenty-three rows in Table 7. The green cells indicate a legally protected characteristic; in this case, the ethnic group names in Kenya and the nationality. The purple cells are individual names, mainly the presidential contenders/politicians and one popular blogger. The blue cells are also groupings but not falling under the protected characteristic category. These include police, government, country, and nation. The yellow cells indicate the "distancing" or "othering" features often characterized by frequent pronoun usage. The red cells indicate the "passion" features characterized by harmful and offensive words. Each topic in Table 7 shows a combination of the passion, distancing, and commitment features, which are reflective of the salient feature in the hate speech conceptual framework developed in this study. Therefore, the use of the LDA topic models proved helpful in quickly exploring and revealing the embedded PDC thematic structure, just as previously used to identify topics in tweets related to bullying [47].

However, the use of LDA presented the limitations of the bag-of-words technique, which does not maintain word order; therefore, word-meaning or context is not preserved. Pragmatically, the reliance on LDA as the primary approach proved inadequate regarding text classification. That notwithstanding, its usage in this study was very useful in data preprocessing and proved helpful as a first-level statistical approach in automatically identifying and extracting passion, distancing, and discriminative (PDC) features as topics from the large corpus. These topics were learned by the model based on the deep underlying concepts in the big data from social media, evidently mirroring the PDC features explicated in espoused hate speech framework.

Table 7. Topic modeling for hate speech class

Topic	1	Kikuyus	thieves	Kenyans	Why	tribal	tribes	Uhuru	country	All	What
Topic	2	Luos	kill	tribal	They	Uhuru	sue	Why	killed	Luo	police
Topic	3	hate	speech	passion	love	Raila	reason	Kamba	dont	way	Luhya
Topic	4	luos	kill	Why	luhyas	killing	police	Kikuyu	dont	mungiki	luo
Topic	5	like	just	Nyakundi	This	Raila	shit	said	Well	time	did
Topic	6	ni	kihii	ya	tu	sana	wa	kama	hawa	wewe	ama
Topic	7	You	think	kill	stupid	Nyakundi	guys	right	sick	know	Kikuyu
Topic	8	people	country	Kisiis	Kambas	The	think	violent	tribal	stupid	nation
Topic	9	Wajaluo	mawe	na	si	wajinga	ujinga	sana	tu	hawana	ndio
Topic	10	don	know	need	want	They	care	Kenyans	chase	women	dont
Topic	11	Luhyas	These	Jubilee	food	cowards	Luhya	stupid	They	poor	supporting
Topic	12	Kenyans	tribes	Kikuyus	kikuyus	IEBC	https	heard	talking	Kuria	ujinga
Topic	13	just	said	election	support	Ruto	hate	Your	chase	world	does
Topic	14	like	kwa	governmer	truth	nyakundi	shit	feel	coming	http	10
Topic	15	people	kill	luhyas	Kikuyu	Luo	Kikuyus	kambas	killing	power	police
Topic	16	We	Maasai	country	going	Mara	hear	free	community	ur	fools
Topic	17	hate	They	All	When	We	luo	won	fuck	better	nonsense
Topic	18	Luos	Kikuyus	Kenya	think	thieves	say	stupid	Well	country	bad
Topic	19	kikuyus	tribal	tribes	think	thieves	kenyans	kikuyu	country	vote	said
Topic	20	luos	luo	raila	tribal	killing	stupid	kisumu	nyanza	poor	killed
Topic	21	ni	kihii	ya	tu	wewe	kama	ule	wa	wembe	hao
Topic	22	wajaluo	wa	nini	ndio	wote	ya	sana	ujinga	tu	sio
Topic	23	people	country	kisiis	think	want	shall	good	don	kambas	killed

Green: Protected Characteristics e.g. Ethnic names and nationalities

Yellow: Distancing features: othering e.g. You, they, we, hawa

Red: Passion features: offensive terms e.g. thieves, kill, fools, stupid, chase, kihii

Purple: Individuals Blue: Other characteristics e.g. Police; Jubilee, IEBC

The psychosocial features were primarily informed by the presence of words or concepts in the message that sought to distance from the target or object of hate. The presence of "othering" discourse in the text message was evident in the usage of pronoun terms such as 'us', 'them,' and other pronoun dichotomies such as 'we,' 'they' which became particularly helpful in identifying hate, just like in a previous study [37]. Example messages included 1 and 2:

"#RailaInMeru Merus are betraying us. Let's defrock them from GEMA." (1)

"*Jubilee is another nusu mkate govt. It's between Kikuyus & R. Valley. We will punish them. We are not happy #TheBigQuestion*" (2)

The element of social distancing was also prevalent in negative stereotypes where negative sentiments and generalizations were directed towards specific ethnic groups. Examples of actual messages include 3 and 4:

"*Kambas also do not make good leaders...they are Cowards*" (3)

"*We shall beat the uncircumcised hands down Luos will never rule Kenya. Be informed. Raila CIC never ever Luos are south Sudanese*" (4)

Psychosocial features were also characterized by offensive and passionate words expressing emotions of anger, hate, fear, or hostility towards a target group. Examples of actual messages include 5 and 6:

"*Arrest everyone mpaka their grand kids Kikuyus are Mungikis Luos are Hooligans Kambas are witches and Somalis are Terrorists.Twende kazi*" (5)

"*Luos and their culture are generally STUPID...People could not pay for your XRAYS will automatically offer RAMS and BULLS in your funeral*" (6)

Some messages contained words bordering threats and incitement to violence towards a given social group. The use of uppercase letters, for example, message 7, was indicative of strong emotions and emphasis. This, too, was the case with codeswitching in message 8. Other examples of messages include 9, and 10.

" *Kisiis are a DANGEROUS THREAT to our businesses they MUST be STOPPED*" (7)

"@USER_NAME tel ur counter part kikuyus are everywea na hawana mashamba.will chase them too" (8)

"*And tell Kambas we are waiting for you come general elections you will not cross River Tana bridge*." (9)

"*Luos are not the whole nation. Only your tribe want war we gonna give it to you man.we will make you extinct if you start it*" (10)

Psychosocial features indicative of the commitment to hate were characterized by words that devalued or demeaned the target. Common among these were words that referred to the target as being immature or equated them to insects, animals, or objects.

Examples of messages from the dataset include 11 and 12:

"*We have never heard such from Central it means Luos are very thick and pathetic. Those are bad tomatoes*" (11)

"*Kikuyus Are Enemies Of Luos Stop Making Music With This Cockroaches*" (12)

Moreover, some of these doubled up as coded language meant to hate on the target using terms or phrases whose meaning was well understood by the in-group, but not obvious with the out-group membership.

These high-level psychosocial features were foundational in developing the initial conceptual framework of the study. The framework was continually revised throughout the study to reflect empirical findings that emerged from the various experiments that were conducted. Some of the significant findings in this regard included the realization that hate speech is multidimensional. From the multiple examples of annotated and automatically identified messages containing hate speech, it was apparent that there was an underlying pattern consisting of messages that discriminated, distanced, used negative passion, were subjective or devalued a person or group of people based on their intrinsic characteristics like ethnicity, gender, etc. Any message lacking these dimensions, particularly the identification of the target based on their ethnicity, was considered to be either offensive or neither. This is well summarized in the multidimensional framework of hate speech as shown by the Venn diagram in Fig. 5. It exhaustively captured the five salient concepts that portray the multidimensionality of hate speech.

The presence and frequency of pronouns in messages have, in the past, been shown to identify the quality of relationships [33]. For example, the use of first-person pronouns like ‘we,’ ‘us,’ ‘our,’ is indicative of closeness and a high-quality relationship among the in-group membership and the general group identity. Whereas, the use of second-person ‘you,’ and especially the third-person pronoun ‘them,’ is indicative of social distancing and lower-quality relationships. A significant finding was that when these pronouns were used, in reference to a protected characteristic, coupled with the other concepts of devaluation, negative passion, or subjectivity, hate speech was extant.

The primary objective of the study was to learn the class, “hate speech” to identify positive instances in a codeswitched text dataset. There were ~50k examples of tweets already labeled into three categories, i.e., hate speech, offensive, neither. As discussed in the conceptual framework section, the annotations were based on the three psychosocial features comprising negative Passion, Distance, and Commitment (PDC). Given a tweet, the human annotator looked for indicators of distance (D) and passion (P) or commitment (C). Hate speech was based on D+P or D+C or D+P+C combinations, whereby psychosocial distancing was targeting a person or a group based on them belonging to a protected characteristic like ethnicity. For example, “Kenyarra is a foolish Kikuyu president. “The reference to the president’s ethnicity, i.e., from the Kikuyu ethnicity, would classify the message as a true positive.

Offensive, just like hate speech, could be based on the three different combinations but not about a protected social characteristic, whether directly or indirectly. For example, “Kenyarra is a foolish drunk. “The premise will be treated as offensive but not as hate speech.

Any other message falling outside of these boundaries will be considered “neither.” In principle, class learning is optimum when features are unique to a class. Fundamentally, the feature description is shared by all instances of a class and none with other competing classes[34]. However, an investigation into the differences in the distributions of class features within the same class and the dependence between class features using the Chi-square revealed a different pattern than earlier thought. Ethnic names frequently appear across the three classes, with Kikuyu, Luo, and Kalenjin (including their respective Swahili language versions) being the most frequent, respectively. Therefore, this means that ethnic names are not a strong feature to use to train a classifier to discriminate between the three classes. This is contrary to our initial thought; however, if this is to be ground-truthed, the presence of ethnic names and negative passion often borders hate speech.

After qualitatively analyzing sample hate messages from the dataset, it is apparent that to classify a message as hate speech, it must contain indicators of negative passion (P) or commitment (C), not just the mere presence of ethnic names or pronouns. The question remains, is there an exhaustive list of the indicators belonging to the set P, D, and C? Do the elements in these sets change over time? For example, given the ambiguous nature of language use, especially in codeswitched texts, does a popular term in a given election campaign persist to the next? If not, how are new terms unique to another election campaign handled by the classifier? These are essential questions that should be resolved, if not at least commence a new discussion for future work.

5. Conclusion

The study sought a deep understanding of the hate speech phenomenon and its salient characteristics as informed by relevant hate theories in the field of psychology and sociology. This resulted in a multidimensional hate speech conceptual framework that is universal and can therefore generalize to any type of hate speech. Therefore, this novel framework will be helpful to other researchers interested in doing a similar study to guide the collection and annotation of hate speech data in any domain or language.

Identifying hate speech in short text messages generated on social media platforms is a challenging classification problem [12,14]. This problem is further compounded by the lack of a universal definition of hate speech, making it an

ill-defined phenomenon[12]. Besides, the process of annotating messages by human annotators is not devoid of annotator bias and subjectivity, therefore making it difficult to formalize [21].

Developing a gold-standard code-switched dataset generated by a multilingual social media community is a first of its kind by our study. Previous studies have often neglected this crucial aspect of a natural and increasingly evident codeswitching phenomenon among multilinguals, by preprocessing monolingual datasets. Secondly, the low inter-rater agreement score reported in previous annotation studies shows how much bias and subjectivity are introduced into the annotation process despite having some form of annotation scheme. This further indicates how emotive hate speech is and the challenge it presents to human annotators who already have some intrinsic knowledge informed by their ethnic and political biases. Although non-Kenyans without any ethnic and political preferences would seemingly appear a better annotation team, however, they will be constrained by a lack of the same intrinsic knowledge to decipher the semantic meaning of the code-switched text messages. So, is this an out-of-reach problem to solve? It may seem so, but the study has already proved critical methodological approaches that definitely will be useful in augmenting the human judgment regarding improved computational time and memory of the machine classifier in classifying codeswitched hate speech related messages from the big data generated from social media; a challenge otherwise unfeasible with human annotators.

Topic modeling was a useful method to identify the latent semantic representations underlying the data from social media. Besides, it enabled the automatic exploration and identification of the hate concepts specified in the study's conceptual framework. Further, the topic modeling technique helped to generate a deeper understanding of the underlying latent factors to the various topics or clusters of hate words, which otherwise would have been unidentified by the conventional methods. The qualitative text analysis using topic modeling enabled the researcher to identify additional salient features to the hate speech framework.

One limitation in this study was that data collection was primarily based on one social media, i.e., Twitter. This was primarily due to the constraint regarding access to data and particularly, obtaining user consent to the messages posted, often privately, to the in-group membership on WhatsApp, Facebook, and other social media platforms. Essentially, all messages posted on Twitter social media are public by default, unless specified otherwise by the user in their Twitter settings. Taking this into account, the study was able to confidently and programmatically scrape relevant public tweets without worrying about the bridge of copyright legislation. However, the question is: could the results obtained in this study be generalized as representative of text data from the other social media networks? This, therefore, would best be answered by future research with the availability of sufficient data from the other social media networks. Besides, the next immediate step for this work is to use the code switched dataset and the psychosocial features espoused in the multidimensional framework to train various machine learning algorithms to establish the best classifier for hate speech. This will also help determine the performance of the new psychosocial features in comparison to conventional features used in previous studies.

Acknowledgments

We would like to thank the Kenya Education Network (KeNet) that funded the data collection and annotation exercise in this study under the Big data cohort.

References

[1] R. Sternberg, K. Sternberg, The Duplex Theory of Hate I: The Triangular Theory of the Structure of Hate. In The Nature of Hate, Cambridge Univ. Press. (2008) 51–77. https://doi.org/0.1017/CBO9780511818707.004.
[2] A. Des Forges, Leave None To Tell The Story: Genocide in Rwanda, New York Hum. Rights Watch. (1999).
[3] R.. King, G.M. Sutton, High Times for Hate Crime: Explaining the Temporal Clustering of Hate Motivated Offending, Criminology. 51 (2013) 71–94.
[4] E. Ombui, L. Muchemi, P. Wagacha, Hate Speech Detection in Code-switched Text Messages, in: 3rd Int. Symp. Multidiscip. Stud. Innov. Technol., IEEE, Ankara, 2019. https://ieeexplore.ieee.org/document/8932845/.
[5] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[6] E. Ombui, M. Karani, L. Muchemi, Annotation Framework for Hate Speech Identification in Tweets: Case Study of Tweets during Kenyan Elections, in: IST-2019, 2019.
[7] P. Burnap, M.L. Williams, Us and them: identifying cyber hate on twitter across multiple protected characteristics., EPJ Data Sci. (2016).
[8] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter., in: Proc. NAACL-HLT, 2016: pp. 88–93.
[9] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive Language Detection in Online User Content, in: 25th Int. Conf. World Wide Web, 2016: pp. 145–153.
[10] P. Fortuna, L. da Silva, Jo˜ao Rocha Soler-Company, Juan Wanner, S. Nunes, A Hierarchically-Labeled Portuguese HateSpeech Dataset, in: Proc. Third Work. Abus. Lang. Online, ACL, 2019: pp. 94–104. https://www.aclweb.org/anthology/W19-3510.pdf.
[11] V.P. de Pelle, Rogers Prates Moreira, Offensive Comments in the Brazilian Web: a dataset and baseline results, in: 6th Brazilian Work. Soc. Netw. Anal. Min., 2017. http://www.each.usp.br/digiampietri/BraSNAM/2017/p04.pdf.
[12] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, M. Wojatzki, Measuring the Reliability of Hate Speech Annotations:

The Case of the European Refugee Crisis, Arxiv:1701.08118. 1 (2017). https://arxiv.org/pdf/1701.08118.pdf.
[13] E. Fersini, P. Rosso, M. Anzovino, Misogyny, Overview of the task on automatic Identification, in: E Third Work. Eval. Hum. Lang. Technol. Iber. Lang., 2018. http://ceur-ws.org/Vol-2150/overview-AMI.pdf.
[14] F. Poletto, M. Stranisci, M. Sanguinetti, V. Patti, C. Bosco, Hate speech annotation: Analysis of an italian Twitter corpus., in: CEUR WS, 2018: pp. 1–6. http://ceur-ws.org/Vol-2006/paper024.pdf.
[15] R. Kumar, A.K. Ojha, S. Malmasi, M. Zampieri, Benchmarking Aggression Identification in Social Media, in: Proc. First Work. Trolling, Aggress. Cyberbullying, ACL, 2018: pp. 1–11. https://www.aclweb.org/anthology/W18-4401/.
[16] Donia Gamal, Marco Alfonse, El-Sayed M. El-Horbaty, Abdel-Badeeh M.Salem, "Twitter Benchmark Dataset for Arabic Sentiment Analysis", International Journal of Modern Education and Computer Science, Vol.11, No.1, pp. 33-38, 2019.
[17] Afnan Atiah Alsolamy, Muazzam Ahmed Siddiqui, Imtiaz Hussain Khan, " A Corpus Based Approach to Build Arabic Sentiment Lexicon", International Journal of Information Engineering and Electronic Business, Vol.11, No.6, pp. 16-23, 2019.
[18] Alemu Kumilachew Tegegnie, Adane Nega Tarekegn, Tamir Anteneh Alemu,"A Comparative Study of Flat and Hierarchical Classification for Amharic News Text Using SVM", International Journal of Information Engineering and Electronic Business, Vol.9, No.3, pp.36-42, 2017.
[19] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated Hate Speech Detection and the Problem of Offensive Language, in: ICWSM, 2017.
[20] Priya Gupta, Aditi Kamra, Richa Thakral, Mayank Aggarwal, Sohail Bhatti, Vishal Jain, "A Proposed Framework to Analyze Abusive Tweets on the Social Networks", International Journal of Modern Education and Computer Science, Vol.10, No.1, pp. 46-56, 2018.
[21] Z. Waseem, Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter, in: EMNLP Work. NLP CSS, 2016: pp. 138–142.
[22] W. Warner, J. Hirschberg, Detecting Hate Speech on the World Wide Web, in: Lang. Soc. Media (LSM 2012), 2012.
[23] I. Kwok, Y. Wang, Locate the hate: Detecting tweets against blacks, AAAI. (2013).
[24] Three Kenyan politicians arrested over “hate speech,” Telegr. (2010). https://www.telegraph.co.uk/news/worldnews/africaandindianocean/kenya/7831369/Three-Kenyan-politicians-arrested-over-hate-speech.html.
[25] Kenyan authorities arrest blogger after posts on alleged official corruption, CPJ. (2018). https://cpj.org/x/72de.
[26] P.. Cavazos-Rehg, M.J. Krauss, S. Sowles, S. Connolly, C. Rosas, M. Bharadwaj, L. Bierut, A content analysis of depression-related tweets, Comput. Hum. Behav. 54 (2016) 351–357. https://doi.org/10.1016/j.chb.2015.08.023.
[27] M. Karani, E. Ombui, A. Gichamba, The Design and Development of a Custom Text Annotator, in: IEEE Africon, 2019.
[28] K. Krippendorff, Computing Krippendorff’s Alpha-Reliability, Univ. Pennsylvania Sch. (2011). mhttp://repository.upenn.edu/asc_papers/43.
[29] Y. Chen, Y. Zhou, S. Zhu, H. Xu, Detecting offensive language in social media to protect adolescent online safety, in: Fourth ASE/IEEE Int. Conf. Soc. Comput. (SocialCom 2012), Amstadam, 2012.
[30] W. Clyne, S. Pezaro, K. Deeny, R. Kneasfsey, Using Social Media to Generate and Collect Primary Data: The #ShowsWorkplaceCompassion Twitter Research Campaign, JMIR Public Heal. Surveill. 4 (2018) e41. https://doi.org/10.2196/publichealth.7686.
[31] W. Ahmed, P. Bath, G. Demartini, Using Twitter as a data source: An overview of ethical, legal and methodological challenges, in: Second (Ed.), Ethics Online Res. Adv. Res. Ethics Integr., Emerald, 2017: pp. 79–107.
[32] Twitter Privacy Policy, Twitter, Inc. (2018). https://twitter.com/en/privacy (accessed October 26, 2019).
[33] Y.R. Tausczik, J.W. Pennebaker, The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods, J. Lang. Soc. Psychol. 1 (2010). https://doi.org/10.1177/0261927X09351676.
[34] E. Alpaydin, Introduction to Machine Learning, 2nd Editio, The MIT Press, London, 2010.
[35] M.L. Williams, P. Burnap, Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data, Br. J. Criminol. 56 (2016) 211–238.
[36] Twitter, About Twitter’s APIs, (n.d.). https://help.twitter.com/en/rules-and-policies/twitter-api (accessed November 25, 2020).
[37] P. Burnap, M.L. Williams, Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making, Policy & Internet. 2 (2015) 223–242.
[38] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep Learning for Hate Speech Detection in Tweets, in: 2017 Int. World Wide Web Conf. Comm., 2017.
[39] S. Joshi, D. Deshpande, Twitter Sentiment Analysis System, Int. J. Comput. Appl. 180 (2018). https://arxiv.org/ftp/arxiv/papers/1807/1807.07752.pdf.
[40] A. Schmidt, M. Wiegand, A Survey on Hate Speech Detection using Natural Language Processing, SocialNLP@EACL. (2017). https://doi.org/10.18653/v1/w17-1101.
[41] K. Constitution, THE CONSTITUTION OF KENYA, 2010, LAWS OF KENYA, Kenya, 2010. http://extwprlegs1.fao.org/docs/pdf/ken127322.pdf.
[42] KLR, NATIONAL COHESION AND INTEGRATION ACT NO.12 of 2008, National Council for Law, Kenya, 2012. http://kenyalaw.org/kl/fileadmin/pdfdownloads/Acts/NationalCohesionandIntegrationAct_No12of2008.pdf.
[43] M. Makinen, M.W. Kuira, Social Media and Post-Election Crisis in Kenya, Inf. Commun. Technol. - Africa. 13 (2008). https://repository.upenn.edu/cgi/viewcontent.cgi?article=1012&context=ictafrica.
[44] NCIC, Functions of the Commission, (2019). https://cohesion.or.ke/index.php/about-us/functions-of-the-commission (accessed September 16, 2019).
[45] R. Damary, NCIC deploys peace monitors to arrest triggers of election chaos, Star. (2017). https://www.the-star.co.ke/news/2017-04-13-ncic-deploys-peace-monitors-to-arrest-triggers-of-election-chaos/.
[46] Kenet, Kenya Education Network, (2018). https://cert.kenet.or.ke/node/4 (accessed September 16, 2016).

[47] J.-M. Xu, K.-S. Jun, X. Zhu, A. Bellmor, Learning from bullying traces in social media, in: Proc. 2012 Conf. North Am. Chapter Assoc. Comput. Linguist., Association for Computational Linguistics, 2012: pp. 656–666.

Authors' Profiles

Edward Ombui is a lecturer at the School of Science and Technology, Africa Nazarene University, Kenya. He is a PhD candidate at the University of Nairobi. His education includes an MSc –Applied Computer Science, University of Nairobi, and BSc Computer Science, Africa Nazarene University. His research interests are in Artificial Intelligence, Natural language processing, Machine learning, and Machine Translation. He has published extensively on IEEE, the African Academy of Languages, among other journals. His professional membership includes the Computer Society of Kenya, the Association for computational Linguistics, IEEE, and the African Language Technology group.

Lawrence Muchemi holds a PhD in Computer Science and is a senior lecturer at the School of Computing and Informatics, the University of Nairobi, Kenya. His current research interests include Data Mining, Natural Language Processing, Artificial Intelligence, and Machine learning. He is an experienced and licensed Engineer since 1995. He has taught at various universities in Kenya which include Jomo Kenyatta University of Agriculture and Technology, Africa Nazarene University where he was the head of the department, and currently at the University of Nairobi.

Peter Wagacha is a Professor of Computer Science at the School of Computing and Informatics, the University of Nairobi, Kenya. His research interests and work includes human language technology, health informatics, mobility, and intelligent systems. He has published in refereed journals and conferences.

How to cite this paper: Edward Ombui, Lawrence Muchemi, Peter Wagacha, "Building and Annotating a Codeswitched Hate Speech Corpora", International Journal of Information Technology and Computer Science(IJITCS), Vol.13, No.3, pp.33-52, 2021. DOI: 10.5815/ijitcs.2021.03.03