I.J. Information Technology and Computer Science, **2021, 6, 29-47**
Published Online December 2021 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2021.06.03

Psychosocial Features for Hate Speech Detection in Code-switched Texts

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

Lawrence Muchemi[2] **and 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: 31 May 2021; Revised: 24 July 2021; Accepted: 09 August 2021; Published: 08 December 2021

Abstract: This study examines the problem of hate speech identification in codeswitched text from social media using a natural language processing approach. It explores different features in training nine models and empirically evaluates their predictiveness in identifying hate speech in a ~50k human-annotated dataset. The study espouses a novel approach to handle this challenge by introducing a hierarchical approach that employs Latent Dirichlet Analysis to generate topic models that help build a high-level Psychosocial feature set that we acronym PDC. PDC groups similar meaning words in word families, which is significant in capturing codeswitching during the preprocessing stage for supervised learning models. The high-level PDC features generated are based on a hate speech annotation framework [1] that is largely informed by the duplex theory of hate [2]. Results obtained from frequency-based models using the PDC feature on the dataset comprising of tweets generated during the 2012 and 2017 presidential elections in Kenya indicate an f-score of 83% (precision: 81%, recall: 85%) in identifying hate speech. The study is significant in that it publicly shares a unique codeswitched dataset for hate speech that is valuable for comparative studies. Secondly, it provides a methodology for building a novel PDC feature set to identify nuanced forms of hate speech, camouflaged in codeswitched data, which conventional methods could not adequately identify.

Index Terms: Hate Speech, Classification, Code-switching, Feature selection, Machine learning.

1. Introduction

Hate speech is a language that often expresses an attitude of prejudice or discrimination targeting an individual or group based on a protected characteristic like ethnicity, religion, or gender [1]. As a phenomenon, it deserves a lot more attention than it is getting today in society. There is increasingly more hate speech and subsequent hate crimes being witnessed around the world with rioting separatists in Hong Kong, gender and religious hate crimes in India, hateful attacks on people of African and Asian descent in the US, as well as ethnic hatred and genocides witnessed in some countries in Africa [3, 4, 5, 6]. During presidential elections, there are increasingly more campaign-related incidents that provoke online public reactions bordering hate speech. Notorious among these are negative ethnic sentiments invoked by politicians and often generating heated public reactions and counter-reactions by users on social media platforms [7]. In Kenya, the situation is exacerbated by the lack of specific policy frameworks to hold media companies, especially social media, responsible for the hate speech propagated on their platforms. Instead, the laws that were available during this study targeted specific users, with the bracket extended to include local administrators of network groups on social media like WhatsApp [8].

User-generated content on social media presents a significant challenge to conventional natural language processing, computational linguistics, and machine learning approaches and applications. It is noisy, irregular, full of duplicate and missing values, voluminous, variety in data types, real-time generated, codeswitched, and coupled with other challenges that come with big data. Codeswitching is a common social phenomenon that is highly indicative of group membership in social conversations [9]. Although it is regarded as informal communication, it is increasingly becoming the norm rather than the exception in everyday communication among bilingual and multilingual communities, more so on social media. Besides, codeswitching on social media regarding hate speech is seemingly the lingua franca for the in-group membership. It is perceived as strengthening cohesion in communicating to “our” people and distancing from the other people, especially those perceived to be the critical opponents. Codeswitching is also common as a way of emphasizing an idea or object in communication. Given that some of the social media networks can enable users to communicate anonymously, these platforms then become fertile grounds for hate speech proliferation.

Our study specifically addresses the problem of identifying hate speech in codeswitched text messages retrieved from social media. This is a challenging classification task that conventional methods have inadequately addressed, often by dropping the codeswitched text. An in-depth analysis of the data collected in this research revealed that some of the deep-seated hate on social media is often camouflaged in codeswitched text messages. Social media communication is often informal, and therefore not uncommon to find messages containing words alternating between multiple languages, especially among multilingual communities. This adds to the complexity involved in parsing sentences and performing a contextual analysis of words and phrases using traditional monolingual tools. The scarcity of native language resources, for example, corpora, parts-of-speech taggers, dictionaries, etc. [10], coupled with undocumented grammatical rules and uncoordinated research networks, seem to exacerbate the situation [11]. Therefore, deriving quality features from this kind of data for purposes of machine learning requires a new approach that mitigates the cracks in conventional data processing approaches. Consequently, the entire process involving the collection, annotation, and selection of quality features that best characterize hate speech in a codeswitching environment, for purposes of training a machine classifier, becomes complex and costly.

Previous research on hate speech identification has concentrated on monolingual datasets with English being the most frequent. However, communication on social medial platforms happens in many other regional and native under-resourced languages like Amharic, Bengali, Seneca, Swahili, and many more. Given that more than half of the world's population is multilingual [12], we postulate that this statistic is increasingly being mirrored on social media with the evidence of codeswitching in language communication. For example, in Kenya, nearly the entire native population is bilingual with the ability to speak in their mother tongue (L1), and Swahili or Kiswahili which is the national language(L2), and/or English which is the official language(L2) [13].

However, whenever there is codeswitching, whether, at word or sentence level, previous similar studies have considered this content as noisy data and opted to drop the entire sentence during the preprocessing stage. Instead of dropping, could there be another approach to better handle this increasingly popular phenomenon on social media? Our study seeks to answer this question by exploring various features to determine the distinguishing features for training a machine classifier to identify hate speech in codeswitched text messages from social media. Therefore, this study bridges the gap for codeswitched language datasets regarding automatic hate speech identification. To the best of our knowledge, this is the first study to collect and build a classifier for a codeswitched language dataset, specifically in English, Swahili, Sheng (slang), and some instances of words from native languages like Gikuyu and Luo. An example text message is,

"*We will swear in the rightful president (RAO) on 12/12.* ***Nyinyi Gikuyu mtabaki na uyo mwizi wenu****. Raila won votes from all 39 tribes*"[Translation of the Swahili codeswitched part: "You Kikuyus will be left with your thief"]

In this regard, the goal of our study was to explore a methodology that better captures key features inherent in subtle forms of hate speech, especially in codeswitched text messages, to enhance the performance of machine classification of big data. The primary objectives included the development of a hate speech conceptual framework, the building of a hate speech dataset from social media in Kenya, the training of a hate speech classification model, and the evaluation of the model. The contribution of this study is two-fold. First, it builds and publicly shares a codeswitched hate speech dataset that can be used by other researchers for comparative studies. Secondly, the study espouses a novel psychosocial feature subset that captures language-use based on the concept of psychosocial distancing, negative passion, commitment to hate, stereotyping, and hate as a story, to extract salient features that can be used to effectively train a machine classifier in identifying nuanced forms of hate speech in codeswitched text messages.

2. Literature Review

Previous studies have used various methods to understand the characteristics of hate speech in text messages. These include the use of hate theories and frameworks. The critical race theory has been used to build guidelines to annotate a corpus for racism [14]. Besides, one study developed a framework to analyze offensive messages in text documents[15]. However, the critical race theory in the previous studies has been limited to categorization based on race and the interplay of law and power. Therefore, the theory is inadequate in identifying other types of hate like gender, religion, disability, etc. Whereas some studies have developed frameworks, for example, to help identify offensive language[4, 15], they lack fundamental theoretical underpinning and are often limited to the use of word lists. Consequently, there is a need to fill this gap, which this study espouses by creating a wholistic hate speech framework comprising of psychosocial features that are informed by a solid theoretical foundation. This is intended to generalize well enough to identify other types of hate in social media messages.

There is a growing number of research activities going on in the hate speech domain including automated approaches to identify hate speech [14, 15, 16] and other related concepts like offensive language detection [17, 18], cyberbullying [19, 20], radicalization and Terrorism [21, 22]. The hate speech studies have approached the automatic classification problem as either a binary task or a multi-class classification task. The former approach is popular in many previous studies which also are particular on identifying the subtype of hate speech like racism [15, 23, 24] and anti-Semitism [16]. In a multi-class classification task, it is not just about identifying black and white but also recognizing the gray shade in the continuum of hate speech and not hate speech (Ok) messages. In this regard, the gray messages are captured by having an “offensive” class, which mirrors how a human annotator would ordinarily perceive and label messages. Previous studies involving multi-class classification include [26, 27, 28].

The review of these studies indicates the deployment of various features with different levels of success in improving the detection of hate speech in text messages. Primarily, these features can be categorized into two: high-level features, and low-level features. The high-level features are human-readable and often qualitative concepts in the text message, which a human annotator can identify and use to decide on the message’s class. These include syntactic, stylistic, semantic, and lexical features. Syntactic features include the length of the message, part-of-speech tags, and the use of imperatives. Stylistic features include the use of uppercase words, exclamation marks, emoticons, character and punctuation flooding as features [14]. Semantic features include associational terms, hate verbs, negative polarity, and the use of subjective nouns. Lexical features include word lists that comprise accusational and attributional terms [19, 29], abusive words [31], insults or flames [17, 31, 32], and offensive language [18, 25, 33] that include racist remarks [24, 34].

Other common feature representations include Bag of words (BoWs), N-grams, and word embeddings. BoWs often result in a high recall value, [36] but low precision due to false positives. This is because the mere presence of hate or offensive terms in the message skews the classification towards the hate speech class without considering the context usage of the term [15, 25, 29]. The N-gram features can exist in two levels: as character or word features. A key advantage of N-gram features is that they preserve context by keeping the word order in the original text. This feature has empirically shown better performance than BoWs in training machine classifiers [30, 36].

Low-level features are generally the extracted features amenable to machine processing, meaning they can be used directly by a machine-learning algorithm to train a model, unlike the original text format. These are often frequency counts based on BoWs, N-grams, and word embeddings representations that include count vectors, one-hot vector encodings, term frequency-inverse document frequency (TF-IDF), and dense vectors. Pre-trained word embeddings as dense vector representation are increasingly popular as the preferable features for training deep learning algorithms in hate speech detection studies [26, 37, 38]. The popularly used pre-trained embeddings include Global Vectors (GloVe) [40], FastText n-grams, and Word2Vec text representations, at both character, word, and sentence levels. A summary of the frequency of usage of both levels of features from the reviewed literature of previous hate speech studies is as shown in Fig. 1.

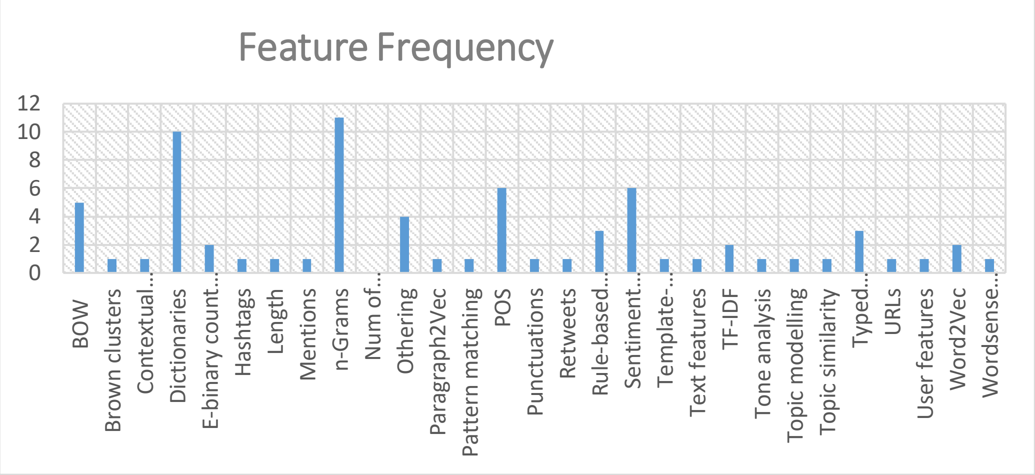

Fig.1. Feature usage frequency in the reviewed hate speech studies

Notably, from the reviewed literature, numerous studies were conducted using English datasets with very few similar studies in other languages like Dutch [26], Amharic [41], Arabic [41, 42], but none in Swahili.

Generally, the features used for text classification play a fundamental role in determining the effectiveness and accuracy of the trained model in discriminating between class instances. The features need to be identified, analyzed, and the most salient among them selected to inform the training of a machine classifier. From the review of literature in hate speech identification, it is apparent that there have been several features employed in previous studies regarding the classification of hate speech. However, these have often been convoluted therefore increasing the complexity of understanding them. This study theoretically and empirically breaks this complexity by dividing these features into two primary categories, i.e., high-level features and low-level features. The high-level features are easily comprehensible and directly identifiable by human annotators. These are further abstracted into psychosocial, linguistic, and App-specific features, as illustrated in Fig. 2. This abstraction introduces a new methodology that captures latent features, for example, the “othering” language, which has proved informative in capturing subtle forms of hate speech in a previous study [44]. Besides, our study espouses, through a holistic hate speech conceptual framework, that these latent features are easily identifiable through the psychosocial concepts and when combined, they become informative features for identification of subtler forms of hate speech, which conventional methods were inadequate in capturing, especially for supervised machine learning.

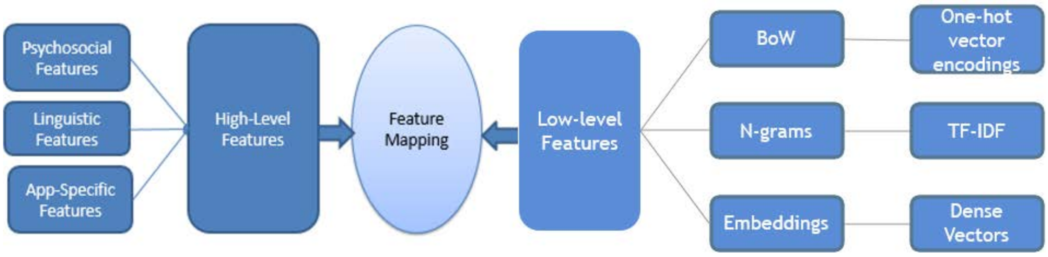

Fig.2. Feature Mapping from high-level to low-level features

3. Methodology

A mixed research method was used to address the four study objectives. First, through a qualitative approach, the content analysis method was used to establish the discriminant features of hate speech by exploring relevant themes emanating from various hate speech definitions and hate theories in the literature. These are concisely captured in the hate speech framework in fig.4. Subsequently, the framework was used to guide the collection and manual annotation of tweets into three predefined classes, i.e., Hate Speech, Offensive, or Neither. A quantitative approach was thereafter used to do text analysis to get word frequencies per class, and subsequently, other low-level features like TF-IDF and word frequency vectors were used to train the classifier model

A consolidated approach was taken to handle all processes from data preprocessing, data exploration and analysis, feature processing, model training, and actual classification using the Jupyter notebook integrated development environment. This was used to facilitate end-to-end model development and visualization of the data through the use of python programming (version 3.6.8) and machine learning libraries like the natural language tool kit (NLTK)- for data preprocessing, Pandas – to see and do various operations on data, Scikit-learn- for various kinds of machine learning models, Matplotlib – for data plotting, among other libraries.

The study used the ethnic group names of seven out of forty-two major tribes in Kenya that account for over 70% of the country’s population [45] as the study population parameter by crawling tweets containing Kikuyu, Luhya, Kalenjin, Luo, Kamba, Kisii, and Meru, including the Swahili versions of these as search key words. Besides, these ethnic names, in combination with other terms as guided by the multidimensional hate speech framework were used to collect and develop the raw dataset.

Unlike conventional research that uses traditional sampling approaches, the big-data projects use different sampling methods that computationally collect all available online content [46], for example by using a web crawler or Twitter API to collect a lot of messages from social media based on specific key words. Such methods are often devoid of various constraints that come with the traditional sampling approaches [47], which for example would have been inefficient and impractical to collect a sizeable volume of hate speech data from many social media users in Kenya for purposes of machine learning. To create a study sample for annotation out of the big volume of the collected data, our study employed simple random sampling. This sampling technique has been used to generate study samples from social media in previous studies [25, 47].

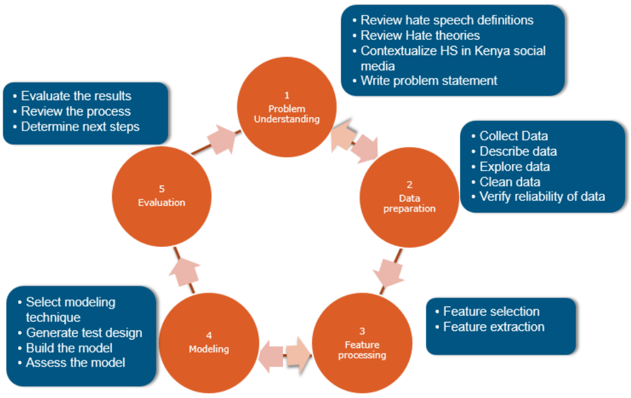

Fig.3. The five-step research workflow

The CRoss-Industry Standard Processes for Data Mining (CRISP-DM) [49] was used to inform the five workflow processes that were used to achieve the study's primary objective of establishing the salient features required to build a hate speech classifier for codeswitched messages. These included understanding the problem, data preparation, feature processing, modelling, and evaluation, as shown in Fig. 3.

The experimental process was guided by these five steps, which have previously proven to strengthen the exploratory undertakings in data analytics [50], an approach that mapped seamlessly to the process activities embedded in our study objectives. These are briefly discussed in the following subsections.

A. Problem understanding

The goal of this phase was to formulate a working definition of hate speech for the study by first seeking to understand the context under which the hate speech phenomenon exists to develop a solution that is practical within the natural setting of the problem. In this regard, relevant literature, both online and physical, was systematically reviewed to develop a deep understanding of the hate speech phenomenon, as it occurs on social media in Kenya. Using the snowballing technique, the cited papers that were referenced in the land-mark literature studies were perused to provide further insights. Besides, a qualitative analysis was done by analyzing the content of several hate theories, definitions of hate speech within user-content guidelines of various social media platforms, and legal definitions of hate speech in the Kenyan government policies, to derive relevant themes.

B. Data Preparation

This phase encompassed three key processes that included data collection, data annotation, and data cleaning.

Convenience sampling was used to collect tweets published during the August 2017 presidential campaign period in Kenya, which also included the repeat election in October 2017. The bootstrapping technique was used as a primary data acquisition strategy. This involved the use of seed words comprising of keywords associated with hate [51], phrase patterns with a connotation of hate[52], offensive hashtags, and pro-hate user accounts [17] to crawl Twitter social media network. Unlike other social media networks, messages published on Twitter are by default publicly available, topically structured, and programmatically accessible. Notably, several similar hate speech studies have used tweets[15, 28, 51]. Therefore, Twitter's API was used to build an application to collect tweets during the election week. A crawler built on python programming was also used to complement Twitter API's limitations of two weeks data collection window to acquire a formidable size of archived tweets dating as far back as the March 2013 Kenyan presidential elections. Besides, presidential campaign periods and events surrounding them are often prominent trigger events leading to spikes in online hate speech[54].

Data annotation, also known as data labeling, involved the use of human coders to manually assign a class to each message in the dataset. Through convenience sampling, an initial team comprising of forty undergraduate computer science students and members of staff in the ratio of 80:20 was recruited and trained on the annotation scheme. The gender of the team was relatively balanced with twenty-one male and nineteen female annotators with an average age of twenty-three. The nationality of the team members was skewed toward Kenyans because of the need to have annotators who could easily interpret the codeswitched nature of the corpus that comprised of messages in English, Swahili, and 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 given an orientation on how to annotate sample messages using a web-based annotation portal developed by the research team [11]. 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. Valuable feedback regarding the speed of the annotation portal was received from the first session. This was used to redesign the portal and expedite the annotation exercise by having each tweet annotated by a random team of three amateur annotators, unlike previously where each tweet had to be annotated by a specific team, with one being a subject matter expert (SME). The new design was informed by the slow annotation process in the first session and the need to expedite the annotation process to have a bigger labeled dataset for training the classifiers. Besides, this was hoped to better utilize and maximize the expertise of the team of human annotators within that short period.

The next process was data cleaning whose goal was to get a higher quality dataset by eliminating noise signals from the annotated dataset, which would otherwise negatively impact the training and overall performance of a machine learning model. The data cleaning process in this study involved the removal of Stopwords, duplicates, HTML characters, non-ASCII and corrupted characters, empty rows, emoticons, and punctuations. The data was also normalized by lowercasing all the words. This was made possible by the use of Python's natural language tool kit (NKTK), regular expressions, among other libraries.

C. Feature Engineering

The goal of this phase was to select a subset of informative and high-quality vocabulary from the highly dimensional output from the data preparation stage. Subsequently, this textual subset, comprising of high-level features, was to be transformed into some low-level numeric representation through feature extraction to be amenable for machine learning. This is because machine learning algorithms can only process numerical feature representations like vectors [55].

The high-level features comprised of two categories: the general lexicon from the data processing stage; and the PDC dictionary that contained five feature categories as informed by the multidimensional hate speech framework in Fig. 4. Both categories were mapped into BoW count frequencies, n-grams, and word embeddings. Subsequently, from these, three low-level features were extracted that included one-hot encodings, TF-IDFs, and dense vectors, respectively. The BoWs features were based on the frequency counts of term occurrences in each message. The n-grams were processed at both word and character-level with n= 2-to-5. These were derived using the count vectorizer in the Scikit-learn machine-learning library. TF-IDF features were used to comparatively find the significance of a specific term in a document and the whole dataset. The general idea here is to penalize words that appear too frequently across all documents. This is because they may not be informative enough to the model as compared to words that are distinct in specific documents but rare across all documents. Consequently, the TF-IDF vectors were generated for the respective levels. Concerning Word Embeddings, the GloVe pre-trained embeddings were used based on the 100d file of about 1 million word vectors, to transform each word to a similar high dimensional vector in these embeddings. The dataset containing the messages was first tokenized. Thereafter, using the transfer learning approach, each token was mapped to its respective embeddings.

The effectiveness of these features was evaluated by learning various classifiers and comparing their accuracy performance. Besides, additional features like the PoS and topic models were extracted from the general lexicon and tested for performance improvement of the classifiers.

Topic Models[56], as high-level features, were intentionally used for data exploration, linking the data to the conceptual framework, and more importantly as an automated process to inform the salient terms to include in the proceeding phase of generating the PDC word-family features. Therefore, the Latent Dirichlet Allocation (LDA) algorithm was used to generate twenty-three semantically meaningful topics or clusters from the large corpus of short-text messages from social media. PDC features are psycholinguistic features derived from the 3 dimensions of hate as explicated by the triangular theory of hate [2]. As high-level features, PDC espouses hate speech in 3 primary word families that are concept-based and language independents. Therefore, the language list can grow or shrink by adding or removing similar-meaning words in different languages in the respective word families. **P**assion word-family consists of words that express negative emotions of anger, fear, disgust, and contempt. These include threatening, abusive, derogatory, and other offensive words directed towards a target person or group that belong to protected characteristics like race, ethnicity, religion, etc. An example message is "to *hell with all <group>. They need to be swept from this country.*" Negative polarity and Sentiment analysis have been used in previous studies to detect passion instances [18, 33]. **D**istance word family consists of words that express psycho-social distance or proximity in inter-group or inter-person relationships, which is also referred to as "othering" language [53]. This is often indicated by a high-frequency usage of pronouns [55, 56, 57, 58]. For example, *"us," them," they," we," you,"* etc. An example of an actual tweet is "*Kambas also do not make good leaders...they are Cowards*". **C**ommitment word family consists of words or phrases that commit to blatantly hate on another person or group by devaluing. This can either be by referring to them using objects, insect or animal names, or generally seeing others as less superior, immature, or less human [61]. Moreover, this includes some of the code names only known and used by the in-group to refer to the members of the out-group. An example tweet from our dataset is "*Kikuyus Are Enemies of Luos Stop Making Music with This Cockroaches*".

The Scikit-learn library was used to encode all these high-level text features as input vector numbers for machine learning. Specifically, the *CountVectorizer* was used to convert the text messages to word count vectors, whereas the *Tfidf Vectorizer* was used to convert the text messages to word frequency vectors. In both cases, the messages in the dataset are first tokenized, and a vocabulary of known words is built. The output is an encoded vector with the length of the entire vocabulary. After that, each new text message is encoded as a fixed-length vector with the length of the vocabulary. For the CountVectorizer, the value in each position in the vector is filled with a frequency count of each word occurrence in the new text message. In case a word in the new text message is not included in the vocabulary, it gets ignored and therefore does not get a count in the resulting vector. The Tfidf Vectorizer calculates the word frequencies and gives a high score for frequent words within a document but downscales the most frequent words that cut across all the documents. The scores, often between 0 and 1, are used to assign frequency weightings in the vector while encoding new text messages.

D. Modeling

This process entailed model selection, training, and parameter tuning. Both conventional and deep learning algorithms were used to train the classification models. The specific choice of the machine learning algorithms was informed by a review of promising results from previous similar studies. The conventional machine learning algorithms included the Naïve Bayes, Support Vector Machine, Linear Logistic Regression, Decision Trees, and the K-Nearest. Besides, Bagging and Boosting models, that is, Forest (RF) and Extreme Gradient Boosting (XGB) were also used. The deep learning algorithms included the Convolutional Neural Networks and Hierarchical Attention networks. The numerous machine learning experiments conducted in this study used the equivalent models available in Python's Scikit-learn library of machine learning models.

A set of hyperparameters were identified and set up in a parameter grid when training each model. These were then automatically tweaked during the experiments using Grid search with 10-fold cross-validation [62] to score the combination of feature parameters and determine the best hyperparameters for the model. These included the value of the soft margin cost, C, the choice of kernel, and other estimator parameters. For example, for the nonlinear Support Vector Machine, the model's generalization in identifying various types of hate speech was tested by adjusting the soft margin cost, C, with lower penalty values between 0.001 to 1.0. To help the model find a nonlinear decision boundary, three popular kernels in literature were employed in the experiments, i.e., the linear, the Radial Basis Function (RBF), and the Polynomial. All these model parameters were based on the ones provided for in the SciKit- Learn libraries[62]. Besides, a pipeline was used to seamlessly combine these parameters with the vectorizer parameters each time the algorithm was run.

E. Evaluation

To simulate how our model was going to behave in the future, the input dataset was split into training and testing datasets. The confusion matrix was used to measure the accuracy performance of the trained models by comparing the predictions to the actual results based on the test dataset. The F-score, based on the weighted average of precision and recall values was used too. The highest prediction accuracy informed the choice of the best model in identifying the positive class, i.e., hate speech, holding ten percent as the validation data set, whereby the K-fold (10-fold) cross-validation was used for model testing.

An inter-rater reliability score was calculated based on the annotations done by the team of twenty-seven human annotators. Each tweet had to be annotated by at least three human annotators. Statistically, the 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 the case of a tie, a fourth annotator, who ideally was a subject matter expert, was introduced as a tie-breaker. Krippendorff's Alpha was chosen as an inter-rater reliability measure for the annotation exercise comprising of a team of twenty-seven novice annotators because it could deal with missing values and outliers [63].

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.

F. Ethical Consideration

The use of social media as the primary source of data for research often raises two primary concerns that include user consent and user identity protection [64]. 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. Therefore, our study focused on collecting only public tweets and retweets which do not need any formal consent or ethical approval. The user identity protection concern was addressed by replacing all user names and mentions with a generic USERNAME label to protect the identity of the online users.

4. Results

This section presents the results of the content analysis, data collection, and processing, modelling, evaluation, and generalizability of the models.

A. Hate Speech Conceptual Framework

The content analysis of several hate theories and hate speech definitions from various literature sources identified five primary dimensions of hate speech that include distancing language, negative passion, devaluation, subjectivity, and stereotyping.

Distancing, also known as othering language, was characterized by the high pronoun usage in the text, especially third-person plural nouns in English and Swahili. This concept has been used previously by several researchers to identify elements of hate speech [43, 51, 56, 63],. The concept of distancing is also evident in social media text whereby one social group displays an attitude of superiority over another or in the cases whereby they seclude themselves to maintain the "purity" of the group membership. For example, in Swahili, the term "madoadoa," which means "spots," was used in disseminating hate speech by some politicians about non-natives during the Kenyan post-election violence in 2007/2008.

The negative passion dimension was characterized by emotions of intense anger, rage, fear, and hostility towards the target individual or group. These were evident in the text that contained expletives such as curse words, obscenities, abusive, derogatory, and other offensive terms. Several previous studies have identified hate speech using this dimension [18, 33, 64]. Besides, the use of negative passion was also evident in text that incited violence towards an individual or a group because of belonging to a given protected social characteristic. Devaluation is a commitment to hate characterized by the use of demeaning words in text messages to refer to a target group using animal or insect

Terms [65,66]. For example, referring to the target group as maggots, cockroaches, rats, etc. This dimension has been used in other hate speech studies [4, 59].

Subjectivity was characterized by the use of faulty arguments that were biased, or propaganda through the use of quantifiers and certainty terms like “always”, “never”, “all”. Stereotyping was characterized by the reference to a person using their ethnic, racial, or religious group names. For example, Kikuyus, Luhyas, Kisiis, etc. These five dimensions and their relationships were built into a multidimensional hate speech framework [67] and as shown in Fig. 4.

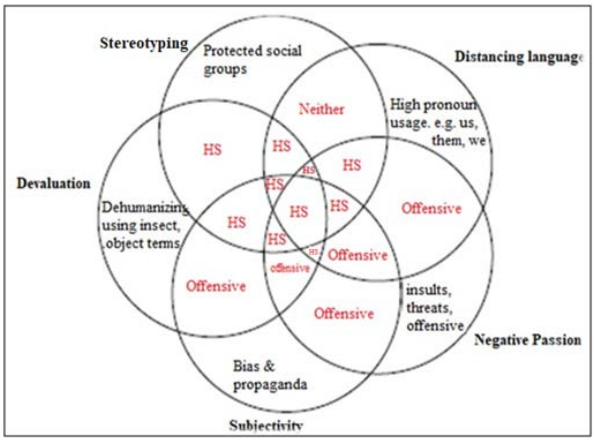

Fig.4. The multidimensional Hate Speech Conceptual Framework

B. *DATA*

Approximately 400k raw unprocessed messages were collected and stored in a comma-separated file (CSV) format. These mainly consisted of Twitter text messages, famously known as tweets, from the general elections in Kenya in August 2017, including a repeat election that was conducted 60 days later, in October 2017. Additional tweets were crawled from January to December 2017 plus the March 2013 general elections to build a sizeable raw corpus.

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

“yes I feel sorry for the dead people but ***bado lazima tu wakikuyu wakae*** *like the guilty ones even when we are doing nothing.”*

A summary description of the dataset is shown in Table 1.

Table 1. Raw Dataset Description

Description	Number of text Messages
Total number of raw text messages collected	401,211
Total number of text messages after preprocessing	398,000
Codeswitched: Swahili, English, and others.	29309

Out of the ~400k messages, 60k messages were randomly selected for annotation by a team of twenty-seven human annotators. Each tweet was annotated by a team of three annotators with the majority vote determining the class. Approximately 50k tweets were annotated, of which 6% comprised of hate speech, ~19% offensive, and 75% were labeled ‘neither’, as summarized in Table 2. The hate speech class was the minority. This is expected from such a big dataset from social media and was consistent with results from previous similar research [68]. One of the findings here was that ethnic hate speech is the predominant type of hate speech during election campaign periods in Kenya. Therefore, ethnic hate speech vocabulary could be profoundly and widely sought as a domain in building a classifier model for the Kenyan context. Secondly, unlike binary classification approaches, 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 [27].

Table 2. Class distribution of annotations

Class	Description	Count
0	Hate Speech	3094
1	Offensive	9401
2	Neither	37819
Total		**50314**

The inter-coder reliability score using Krippendorf's Alpha was 0.5207. This meant that half of the time the annotators were not exclusively in agreement. This is consistent with previous similar research with even a lower inter-rater score of 0.17 [69]. The low inter-rater agreement has previously been attributed to the diversity in personal sensitivities and social biases, coupled with the use of inexperienced but affordable annotators [70]. In our case, the score was exercabated by a few annotators who did not consistently attend the full annotation training and therefore introduced some teacher-noise in the annotations. Generally, the teacher noise coupled with the tacit knowledge and biases the team came with during annotation, despite the training, could form part of the latent attributes that got modeled as random components in the noise signal. Another reason could be the big percentage of missing annotations given that Krippendorff's Alpha assumes that each message is annotated by the entire team of annotators, in this case, twenty-seven, whereas the annotation portal was designed to have each message annotated by a random team of three annotators out of the twenty-seven. This design was motivated by the need to maximize the volume of annotations from the team of human coders using the minimal available resources. To better train a robust and unskewed classifier, random undersampling was done on the majority class, i.e. the 'neither' class, and also on the 'offensive' class. This resulted in a relatively balanced dataset of 9726k tweets with majority votes in the three classes. Further, another finer and balanced dataset comprising of 2537k tweets with only full agreement annotations was built. Both datasets were considered in the experiments for training the machine learning algorithms and are publicly available on Kaggle.

Topic modeling [56] based on the Latent Dirichlet Allocation (LDA) model was used to find deep underlying concepts of hate in the big corpus comprising of code switched text. LDA, a hierarchical probabilistic model, has successfully been used previously to identify topics related to cyberbullying [20]. 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 [56]. 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, the LDA algorithm 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 from the large corpus. These features were present in the twenty-three latent topics extracted as a "bag of words" closely associated with the hate speech class. 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.

C. Modeling

The study sought to answer the question of how informative the psychosocial (PDC) feature set was in comparison to the conventional high-level features in training a classifier for hate speech. The conventional features included the lexical features (LEX) that comprised the general lexicon from the input corpus. The specific features extracted from this included the BoWs and n-grams, Part of Speech features (POS), and Application-specific features (APP) like the frequency of retweets, likes, etc. Therefore, nine machine learning models were trained using these features and their performance compared to identify the best model for classifying subtle forms of hate speech in codeswitched messages. The features were tested independently and in combination using a feature combo to establish the best features and best performing model. The wrapper approach was used whereby the PDC feature set started with only a few features under the respective psychosocial categories as informed by the LIWC psychological word list [71]. The categorical features were added over time as more specific features were identified in messages previously reported as hate speech, coupled with the addition of translation equivalents to cater for codeswitched instances. Besides, the Lex features were quite sparse as compared to the dense and informative PDC features. This again can be explained by the random feature sampling approach used to extract the Lex features from the input dataset by the vectorizer, which contains a parameter for defining the number of features. With text, the higher the number of features the more complex the computation gets especially regarding the amount of memory and computation time required to process the highly sparse input vector. Generally, the PDC feature set, unlike the usually large and "diluted" Lex feature set, comprises of fewer but selectively high informative features, i.e., "concentrated features", to identify hate speech. As observed in Fig. 5, the addition of PDC to the conventional Lex feature set always led to better performance. Conversely, the addition of Lex features or the other features led to lower performance. This can be explained by the noise element introduced by these features and subsequently the sparseness of the new input vector in the classifiers.

Previous studies in hate speech identification have utilized lexical and other NLP-based features. However, these kinds of features by themselves will not be able to adequately capture hate speech in codeswitched messages. Therefore, classifier models that explicitly use these conventional features will underperform with a lot of false negatives, contrary to the actual representation of hate in social media messages.

The study's Psychosocial features (PDC) as well as other high-level features including linguistic features (PoS), general lexical features (n-grams), and App-specific features (App) like the length of a tweet, were used to train classifiers and their performance compared across the conventional text classification algorithm including the Naïve Bayes, Linear Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Decision Trees, Random Forest and extreme Gradient Boosting classification models. The models were learned using 10-fold cross-validation with a dataset ratio of 80% training features and 20% testing features. A grid search algorithm was used to compare and evaluate the models and feature categories yielding the highest accuracy performance across the seven machine learning algorithms. The Support Vector Machine model produced the highest accuracy of 76.2%, closely followed by the Linear Logistic Regression model with an accuracy of 75.8%. The accuracy scores for each model were based on 10-fold cross-validation and are well captured by the box and whisker plot in Fig. 6.

Considering that the primary objective was to identify hate speech, the focus shifted to the performance of the models regarding the hate speech class. Therefore, only the accuracy performance for the hate speech class was extracted from the two promising models. The experimental results were based on the balanced dataset and are well summarized in Table 3.

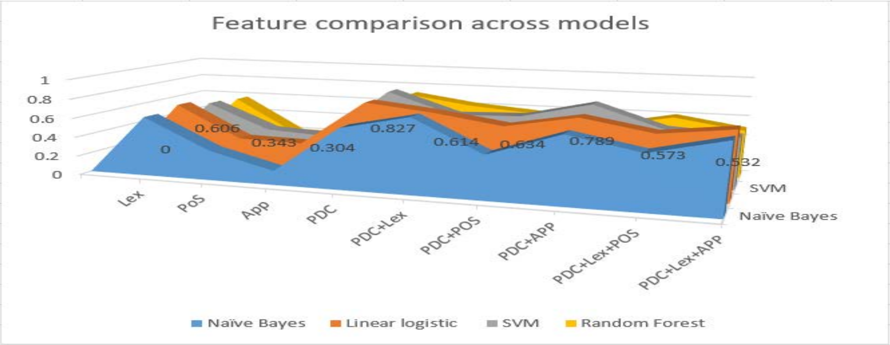

Fig.5. Feature Comparison across models

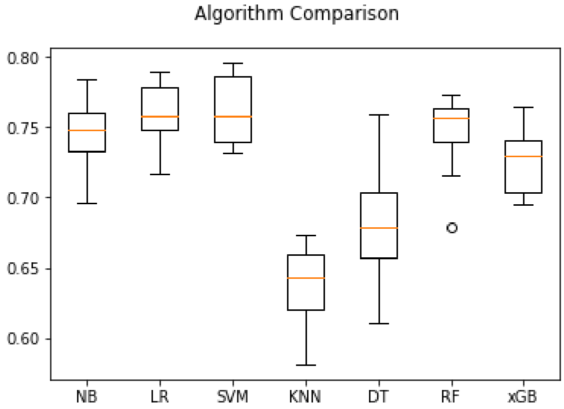

Fig.6. Accuracy performance comparison of the models

The co-occurrence of psychosocial features from the three domains, as defined by the multidimensional framework [67], provides a robust approach to identify hate speech in text messages. This approach overcomes the limitations of lexicon-based methods that mainly rely on the ability to identify hate by finding specific domain words in the message, often without taking cognizance of the syntactic patterns of hate, especially where codeswitching can be used to dodge the domain keywords.

Detection of hate speech is dependent on the existence of specific features, a common limitation with dictionary-based approaches because the model may not generalize in the absence of these features.

The key disadvantage with the general lexicon feature is its sparse vector representation which translates to the vectors having lots of zeros. This, therefore, requires more computational resources, especially memory, when modeling, which is a challenge, especially for the conventional machine learning algorithms.

The finding here is that the best features and classifiers in identifying hate speech based on accuracy performance are PDC features trained with linear SVC classifiers. Another finding is that PDC features had the most significant effect on the accuracy performance, specifically based on character-level n-grams, as compared to the word or phrase-level n-grams. This finding is in agreement with previous research in hate speech [14].

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 [30]. For example:

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

(1)

The element of social distancing was also prevalent in negative stereotypes where negative sentiments and generalizations were directed towards specific ethnic groups. For example:

*"We shall beat the **uncircumcised** hands down **Luos** will never rule Kenya. Be informed.*
*Raila CIC never ever **Luos** are south Sudanese"* (2)

Psychosocial features were also characterized by offensive and passionate words expressing emotions of anger, hate, fear, or hostility towards a target group. For example:

*"Arrest everyone mpaka their grand kids Kikuyus are **Mungikis** Luos are **Hooligans***
*Kambas are **witches** and Somalis are **Terrorists**.Twende kazi"* (3)

Some messages contained words bordering threats and incitement to violence towards a given social group. The use of uppercase letters, for example, message 4, was indicative of strong emotions and emphasis. Most of these messages contained codeswitching too which we have underlined in the message.

*" **Kisiis** are a DANGEROUS THREAT to our businesses **they** MUST be STOPPED"* (4)

*"@HonMoses_Kuria tel ur counter part **Kikuyus** are everywea na hawana mashamba. will chase **them** too"* (5)

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 retrieved from the dataset include messages 6 and 7:

*"We have never heard such from Central it means Luos are very thick and pathetic. Those are **bad tomatoes**"* (6)

*"Kikuyus Are Enemies Of Luos Stop Making Music With This **Cockroaches**"* (7)

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. This helps to better capture nuances of hate speech that would otherwise go through conventional filters, especially in previous frameworks[15] that only considered hate speech from one dimension. 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 for hate speech as shown in Fig. 4. 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 [71]. 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 of 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 in reference to 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[72]. 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 appeared 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 will new terms or campaign phrases be handled by the classifier in future elections? These are essential questions that should be answered, or at least trigger new discourse for future work.

D. Model Evaluation and Tuning

This phase was concerned with evaluating whether the classification model was working as expected and how to enhance the classifier’s performance through parameter tuning.

The best performing model out of the nine that were explored in the experiments was the SVM whose evaluation was done by generating its confusion matrix, and determining its F1, precision and recall score values. The SVM model had a uniform precision, recall, and F1 score of 0.77. Primarily, the study was keen on the model’s performance on the hypothesis class i.e. hate speech, which had a precision score of 0.81, a recall of 0.85, and an F1 score of 0.83. The full results are well summarized in Table 3.

Table 3. Evaluation of SVM model

Class	precision	recall	f1-score	support
0	0.81	0.85	0.83	203
1	0.70	0.71	0.71	226
2	0.79	0.75	0.77	206
accuracy			0.77	635
macro avg	0.77	0.77	0.77	635
Weighted avg	0.77	0.77	0.77	635

A general observation from the normalized confusion matrix in Fig. 7 is that there was more misclassification in the lower triangle of the matrix than the upper one. This indicates that the SVM model was more inclined towards classifying messages as hate speech or offensive more than how the human coders had originally annotated them.

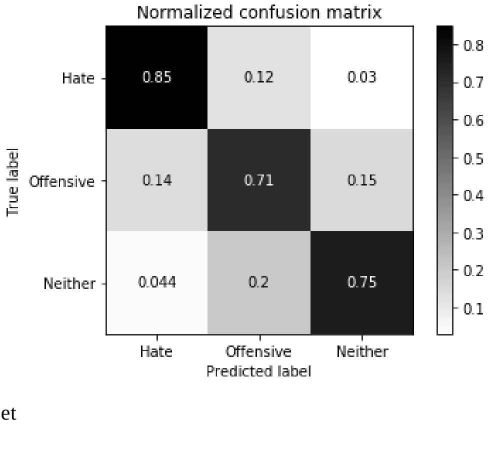

Fig.7. Confusion matrix for the balanced dataset

Looking at the first column of the matrix, the model correctly predicted 85% (recall) of the actual hate speech messages as true positives, whereas, 14% and 4% of offensive and “neither” messages respectively, were misclassified as hate messages. The misclassification, particularly of the offensive messages as false positives, can be explained by the shared characteristics of hate speech and offensive messages. From the multidimensional hate speech framework

[67], it is clear that hate speech contains offensive language, although not all offensive messages are necessarily hated speech unless they explicitly target a protected characteristic. Besides, the presence of offensive lexical terms would influence the classifier to label the message as hate speech, whereas the human annotator would consider the context and use their hindsight to label it otherwise. The misclassification could also be due to erroneous annotations influenced by the sensitivity of the human coder to day-to-day language use, long-held cultural, religious, and other social belief systems that are intrinsic [25, 69].

From the second column of the matrix, it is observed that the model correctly predicted 71% of the offensive messages, whereas 12% of true hate speech and 20% of "neither" messages were misclassified as offensive. This again could be explained by the inherent bias and subjective nature of the human annotator in this task.

The third column of the matrix shows that the model correctly predicted 75% of the messages as "neither", whereas 3% of true hate speech and 15% of offensive messages were misclassified as 'neither'.

Therefore, the most significant confusion for the model was in predicting messages belonging to the offensive class, whereby 12% and 20% of the true instances of hate speech and 'neither' were wrongly predicted as offensive. This too could be explained by the teacher noise introduced during annotation due to the varying sensitivity levels on the part of our annotators in what they individually considered to be offensive.

Concerning parameter tuning, the SVM classifier outperformed all the other classifiers with a soft margin, **C**= 0.1, probability=true, and a Gaussian Radial Basis Function (RBF) kernel, gamma (**γ**)=0.1, as the ideal hyper-parameter values. The justification for the C value was informed by the need to have a classification model that would generalize well over other types of hate speech. Therefore, that required the model to be trained to have more tolerance when establishing the decision boundary, which in machine learning is implemented by lowering the penalty incurred in the instances of model misclassification [73]. The choice of the kernel in SVM helped to best determine how the model generated a nonlinear decision boundary based on the features. The gamma (γ) hyper-parameter was important in determining the sensitivity of the decision boundary when presented with new features. A higher value of gamma means that new features will have a higher influence on the decision boundary, making it more twisted. Therefore, the lower values used for the soft margin and kernel hyperparameters were the most ideal for tuning the SVM classifier to handle the otherwise non-linearly separable case involving text data from social media. Besides, SVM classifiers are quite robust and record impressive predictions as models.

E. Generalizability of the Classification Model

The question of model generalizability was pivotal in this research and was used to overarch all the other objectives and the experiments in the study. Therefore, from the onset, the study sought a deep understanding of the hate speech phenomenon and its salient characteristics as informed by relevant theories in the field of psychology and sociology. This resulted in a multidimensional hate speech framework that was used to guide the data collection and data annotation activities. Although the data that was collected during the 2017 presidential elections in Kenya, mainly containing the ethnic type of hate speech, was used to train the machine classifier in this study, it, however, does not limit it to classifying ethnic hate. First, from the various experiments conducted, the best performing classifier regarding generalizability, was trained on a balanced dataset. Generally, a classifier trained on a balanced dataset comprising of an equal or almost equal number of class instances will not be skewed towards any given class as compared to the classifier that is trained on a dataset that is biased towards the majority class [74]. Secondly, our model was built on the multidimensional hate speech framework that is conceptually universal regarding hate. Therefore, it should generalize to other types of hate speech and in any language, as long as it is retrained with positive instances of that type of hate speech. Consequently, our model was able to positively identify other types of hate speech from new unseen messages, for example: " ***Kill*** *all* ***those Muslims*** *to eradicate terrorism*" (Religious hate); "***Wtf****! Eastleigh explosion.* ***Wasomali*** *warudi kwao"* (Nationality hate); *"Thot the 'summerbreak' is over?* ***hawa wazungu*** *waende zao bana! kazi kutuchafulia ma lightskins wetu nkt eyesore galore"* (Racial hate); *"****Women*** *are some of the most* ***corrupt*** *individuals when placed in positions of power. "* (Gender hate).

These messages contained three key features of hate speech defined in the hate speech conceptual framework. These include negative passion e.g., *Kill, Wtf*; distancing language by the use of plural pronouns, e.g., *those,' hawa' (*these*),* and stereotyping by mentioning a protected characteristic e.g. *Muslims, Wasomali (*Somalis*), Wazungu (*Whites*), Women.* The combination of these features in one message ultimately triggers the hate speech flag.

Fundamentally, the hate speech conceptual framework helps to demarcate the hypothesis class $\mathcal{H}$,i.e., Hate Speech, from which the hate speech instances could be mapped to. Therefore, the work of the machine learning algorithm is to establish the specific hypothesis, h ϵ $\mathcal{H}$, that closely approximates hate speech.

The question of model generalizability also addresses the dynamic nature of language and how to handle future terms that are not part of the training set. Here, the concern is whether the hypothesis will hold for future unseen examples that were not part of the training set, for example, instances in the test or validation data presented using cross-validation. This can be handled by inducing a class S such that h = S. This means that S must only contain all positive examples of hate speech. Alternatively, a general hypothesis, G that contains all the positive examples of hate speech without any false examples, can be used. The algorithm can be retrained using the G-set that accommodates instances of the new terms, as well as increase the margin, which will result in increased distance between the boundary and the closest instances, as shown in a previous study [72].

5. PDC-Based Classification Model

The study espouses a new text classification framework that uses a combination of psycho-social features (PDC) based on the language connotating negative passion, social distancing, and commitment to hate, as the primary informative concepts for identifying subtle forms of hate speech. These qualitative concepts, well established in hate theories like the duplex theory of hate, offer a rich mechanism of capturing these elusive hateful expressions, especially when camouflaged in codeswitching, which conventional methods were inadequate in capturing.

The PDC-based classification model is based on supervised machine learning. It comprises three main components that include data pre-processing, feature engineering, and model building and evaluation. These are illustrated in Fig. 8. The data preprocessing component contains two subcomponents, i.e., the data annotation and data preprocessing. Typical of supervised machine learning, the PDC-based model's input comprises labeled data that could be designed for either binary classification or multi-class classification problems. By this, we mean that the data could be annotated using only two labels, for example, positive or negative in the case of binary classification, or more than two, for example, high, medium, and low, in the case of the multi-class classification. The raw data input is often labeled by human annotators based on some annotation scheme. In this study, the annotation scheme was based on a strong theoretical underpinning that is well elaborate in our previous study [1] and as illustrated in the first zoomed-out component in Fig. 8. For example, the PDC-based multi-dimensional framework can capture the use of devaluation in a codeswitched message such as, "*Do not make music with **those cockroaches**, hiyo ndiyo dawa ['that's the medicine needed'] to silence **them***". Several ethnic devaluation names in Kenya are well understood and used by the in-group membership when referring to out-groups. For example, the use of "foreskins", or "fish", often used to imply and belittle the Luo ethnic group that does not traditionally practice circumcision. The use of stereotypes translates into the use of language in a subtle hateful manner without necessarily using obvious hateful lexicons. For example, the use of compound terms like "*money lovers*", "*tire thieves*", or "*night runners*" to refer to the Kikuyu, Kamba, and Kisii ethnic groups, respectively. These subtle forms of hate speech, especially when codeswitching is applied, often go undetected through the conventional filters.

The data pre-processing sub-unit involves tokenization and data cleaning of the annotated text which is often noisy. The standard data cleaning steps are carried out including dropping of punctuations, duplicates, empty strings, non-alphanumeric characters, lowercasing, stemming, and removal of Stopwords. However, unlike conventional models that indiscriminately drop all pronouns in the preprocessing step, the PDC-based model retains the pronouns when removing the Stopwords. This is because the occurrence of pronoun dichotomies in a message has previously been proved to be informative features in indicating the "othering" language [44], which is a hate speech concept under psychosocial distancing. For example, " ***We** shall not allow **them** to cross river Tana. Punda hao!"*

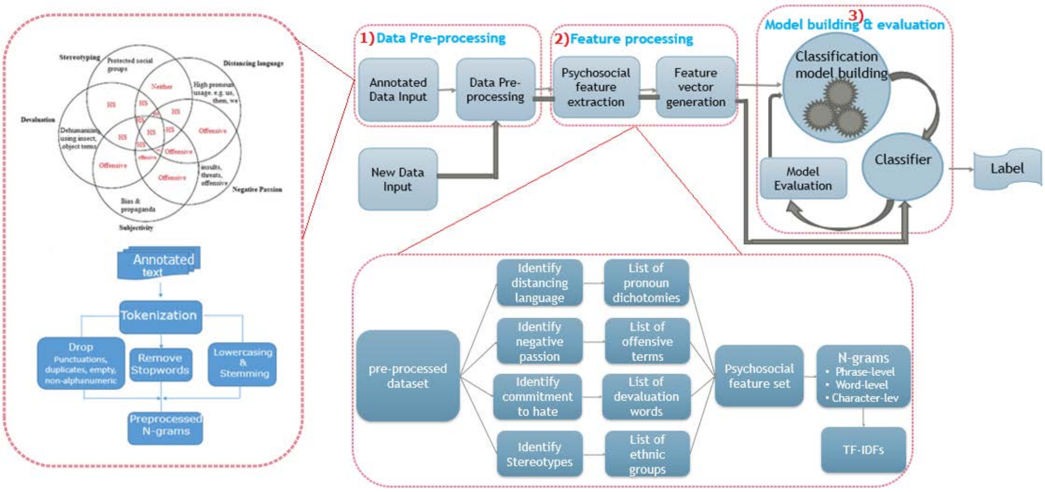

Fig.8. The PDC-Based text classification Model

The output from the first component is the pre-processed dataset which has been stripped off of the regular noise signals and normalized with lowercasing and stemming. This consequently achieves a significant reduction in dimensionality in comparison to the initial raw annotated input. However, textual data presents a challenge because the words or tokens are often the primary features. Therefore, this feature set transforms into a high dimensional feature space whose input vector to the machine learning algorithm, in component 3, will be very sparse with many zeros, subsequently requiring more computational time and memory. The PDC-based model resolves this challenge, in component 2 which comprises the PDC vocabulary learning subcomponent and feature vector generation subcomponent. The first subcomponent filters the pre-processed dataset by extracting vocabulary indicative of the language of psychosocial distancing, negative passion, commitment to hate, and stereotyping, to form their respective lists. The seed features under each list were primarily informed by features that proved useful in similar problems in literature [4, 15, 51]and the ones drawn from the psychological word categories in the Linguistic Inquiry Word Count analyzer[71]. Besides, the respective feature categories were populated with terms that correlated to the classes from the topic models that were automatically generated using the Latent Dirichlet Allocation technique. Structurally, the five feature categories were organized into a table based on word families whereby the first column indicated the word-family, with subsequent columns containing the word forms or features, whereas the rows stored the meanings. Through bootstrapping, new words or similar-meaning words in other languages i.e. codeswitched words could easily be added under the respective feature columns. The structural form is as summarized in Table 4.

Table 4. PDC Conceptual lookup table

Word-Family	**Word Form (Features)**					
	Feature1			**Feature …**	**…**	**Feature n**
Negative **P**assion	F_{L1}	F_{L2}	F_{Ln}			
Distancing						
Commitment (Devaluation)						
(Stereotyping)						
(Subjectivity)						

This psychosocial feature set, PDC, could then be processed at various levels, i.e., phrase, word, or character level, and transformed into numerical feature vectors, in this case, TF-IDFs. As a feature selection and representation method, TF-IDF is designed to rank tokens based on their importance regarding the entire corpus so that tokens at both extremes, i.e., words that are very frequent or rare across documents, are penalized because they are regarded as irrelevant or outliers. Therefore, the resulting TF-IDF feature vector based on the high-level PDC feature set is dense and a more amenable input for classification model building, in component 3. Here, a suite of machine learning algorithms, informed by their performance in identifying hate speech in previous similar studies, are trained on the TF-IDF input vector to build their respective classifiers. Often, it is hard to tell beforehand what machine learning algorithm will be ideal for the classification problem. Therefore, it is a common practice in machine learning, to try several algorithms, starting with simpler ones, to establish the most ideal for the specific machine learning task [74]. The best classifier model is evaluated and tested based on its results and performance. There are two ways of doing the evaluation. First, the Chi-Square feature scoring method is used to compute the correlation between the features and the class, i.e., the text vector and the label column value. Secondly, the confusion matrix is used to compute the precision, recall, and subsequently the accuracy of the trained model using the testing dataset. Finally, a new text message is given as input to the pre-processing sub-component. It does not have to be annotated. However, it also has to go through the feature processing component, and equally transformed it into its TF-IDF vector representation. Subsequently, the vector is directly presented to the classifier and the predicted class label is given as the output. This is as shown in Fig. 8.

In summary, experiments were conducted to validate our approach of using psycho-social concepts drawn from existing hate theories in psychology and sociology to develop a novel psycho-social feature set, which we refer to as PDC. The PDC feature set was subsequently transformed to TF-IDF vectors to learn a classification model for identifying subtle forms of hate speech, especially in codeswitched data. The results from our classifier were compared to the baseline, which was the human inter-rater reliability score for the same annotated dataset, and did much better by over ~27% in classification accuracy. The classifier was further tested on its generalization on an unseen dataset for racist, religious, and nationality hateful comments. The results were comparatively at par with the state-of-the-art baseline classifiers for similar classification of hate speech. However, due to the use of different datasets, especially where the emphasis in this study was in codeswitched data, it would be unrealistic to directly compare with the publicly available monolingual datasets. Besides, this further demonstrated the ability of the psycho-social features in being robust to generalize to other types of hate speech, like the racist comments.

6. Conclusion

The contribution of this study is three-fold in that it provides a gold-standard annotated dataset that can be used by other researchers for comparative studies. Secondly, the study developed an empirical hate speech framework and methodology explicitly grounded in theory for building a novel psychosocial feature set for identifying nuanced forms of hate speech in short text messages. Thirdly, this framework proved useful in developing a text classification model that can effectively generalize to identify other types of hate speech on social media. Subsequently, accrued results from the deployed hate speech classifier could be used to inform evidence-based decisions by relevant security agencies, and data-driven policy formulation regarding monitoring of hate speech on social media during future presidential elections in Kenya.

The psychosocial feature set utilizes language-use around the concept of psychosocial distancing, negative passion, commitment to hate, stereotyping, and hate as a story to identify nuanced forms of hate speech, which conventional methods could not identify, especially in codeswitched data. These concepts are well-grounded in the duplex theory of hate[2] and summarized in the conceptual framework in fig.4. Besides, the study presents a simple and effective method for qualitatively identifying and analyzing hate speech in short text documents by the use of human-readable high-level psychosocial features, that is PDC-based features, which can subsequently be mapped to machine-readable lower-level features like Term Frequency-Inverse Document Frequency (TF-IDF) and one-hot encoding vectors for training a machine classifier. Previous studies in hate speech identification have utilized lexical and other NLP-based features. However, these kinds of features, in isolation, are not adequate in capturing the full range of hate speech in codeswitched messages. Therefore, classifier models that explicitly use these conventional features will underperform with a lot of false negatives, contrary to the actual representation of hate in social media texts.

The use of the psychosocial (PDC) features is designed to be effective in two key ways. First, the feature set ought to be informative enough to enhance classification performance. Secondly, the size of the PDC feature set is much lower than the conventional methods employing the TF-IDF comprising of the entire input lexicon. This fundamentally reduces the sparseness and dimensionality of the original features, making PDC an excellent feature selection technique with a dense input vector length, unlike the sparse input vector of the general lexicon. Besides, the PDC design's effectiveness as a qualitative feature selection method for codeswitched text classification of nuanced forms of hate speech will contribute to the general machine classification efforts.

The comparison of the various features in training the nine machine learning models in this study indicates that the PDC features, using character-level n-grams, are the most discriminative in the classification of hate speech in codeswitched text messages. The best performance was achieved with the length of n=3 to 5 characters respectively, using the SVM classifier. This could be explained by the high level of language independence of character n-gram features, which also makes the portability of feature extractors effortless between languages [75]. Besides, this feature has proved to be most salient in the authorship categorization task [76]. The downside of character n-grams is that they increase the dimensionality of the feature space, especially with very large datasets. Nonetheless, their performance is superior to conventional n-grams in relatively smaller datasets with conventional machine learning algorithms running on moderate computer hardware specifications.

Future work will consider going beyond the current discrete representation of the PDC features where the words exist as atomic symbols to distributed representation where dense vectors could be used to represent the word families to accommodate synonyms, hypernyms, and codeswitching appearing in their context words.

Acknowledgments

We would like to thank the Kenya Education Network (KeNet) for funding the study and the wonderful team from Africa Nazarene University for annotating the dataset.

References

[1] E. Ombui, L. Muchemi, and M. Karani, "Annotation Framework for Hate Speech Identification in Tweets: Case Study of Tweets during Kenyan Elections."

[2] R. Sternberg and K. Sternberg, "The Duplex Theory of Hate I: The Triangular Theory of the Structure of Hate. In The Nature of Hate," *Cambridge Univ. Press*, pp. 51–77, 2008.

[3] A. Des Forges, "Leave None To Tell The Story: Genocide in Rwanda," *New York Hum. Rights Watch*, 1999.

[4] S. Benesch, "Dangerous Speech: A Proposal to Prevent Group Violence," 2012.

[5] L. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber, "Analyzing the Targets of Hate in Online Social Media," in *Tenth International AAAI Conference on Web and Social Media,* 2016, pp. 687–690.

[6] R. Hatzipanagos, "How online hate turns into real-life violence," *The Washington Post*, Washington, 30-Nov-2018.

[7] R. Ajulu, "Politicised Ethnicity, Competitive Politics and Conflict in Kenya: A Historical Perspective," *Afr. Stud.*, vol. 61, no. 2, pp. 251–268, 2002.

[8] P. Makori, "Whatsapp admins face jail in crackdown to curb hate-speech," *Business Today*, 17-Jul-2017.

[9] S. Madonsela, "A critical analysis of the use of code-switching in Nhlapho's novel Imbali YemaNgcamane," *South African J. African Lang.*, vol. 34, no. 2, pp. 167–174, 2014.

[10] E. Ombui and L. Muchemi, "Wiring Kenyan Languages for the Global Virtual Age: An audit of the Human Language Technology Resources," *Int. J. Sci. Res. Innov. Technol.*, vol. 2, no. 2, pp. 35–42, 2015.

[11] M. Karani, E. Ombui, and A. Gichamba, "The Design and Development of a Custom Text Annotator," in *IEEE Africon*, 2019.

[12] A. I. Ansaldo, K. Marcotte, L. Scherer, and G. Raboyeau, "Language therapy and bilingual aphasia: Clinical Implications of psycholinguistic and neuroimaging research," *J. Neurolinguistics*, vol. 21, 539–55, 2018.

[13] L. Muaka, "Language Perceptions and Identity among Kenyan Speakers," in *Proceedings of the 40th Annual Conference on African Linguistics*, 2011.

[14] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter.," in *In*

Proceedings of NAACL-HLT, 2016, pp. 88–93.

[15] 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.

[16] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *Language in Social Media (LSM 2012)*, 2012.

[17] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," *AAAI*, 2013.

[18] D. N. Gitari, Z. Zuping, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection.," *J. Multimed. Ubiquitous Eng.*, vol. 4, no. 10, pp. 215–230, 2015.

[19] E. Spertus, "Smokey: Automatic recognition of hostile Messages," in *IAAI*, 1997.

[20] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *The fourth ASE/IEEE international conference on social computing (SocialCom 2012)*, 2012.

[21] D. K, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying.," *ACM Trans Interact Intell Syst*, vol. 3, no. 2, 2012.

[22] C. Van Hee and G. De Pauw, "Automatic Detection and Prevention of Cyberbullying," in *The First International Conference on Human and Social Analytics*, 2015.

[23] S. Agarwal and A. Sureka, "Using KNN and SVM Based One-Class Classifier for Detecting Online Radicalization on Twitter," in *The 11th International Conference on Distributed Computing and Internet Technology*, 2015, pp. 431–442.

[24] M. Last, A. Markov, and A. Kandel, "Multi-lingual Detection of Terrorist Content on the Web," in *International Workshop on Intelligence and Security Informatics*, 2006.

[25] E. Lozano, J. Cedeno, G. Castillo, F. Layedra, H. Lasso, and C. Vaca, "Requiem for online harassers: Identifying racism from political tweets," in *Fourth International Conference on eDemocracy & eGovernment (ICEDEG)*, 2017.

[26] S. Tulkens, L. Hilte, E. Lodewyckx, B. Verhoeven, and W. Daelemans, "The Automated Detection of Racist Discourse in Dutch Social Media," *CoRR, abs/1608.08738*, 2016.

[27] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "AutomatedHateSpeechDetectionandtheProblemofOffensiveLanguage," in *ICWSM*, 2017.

[28] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep Learning for Hate Speech Detection in Tweets," in *2017 International World Wide Web Conference Committee*, 2017.

[29] P. Fortuna, "Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes," University of Porto, 2017.

[30] P. Burnap and M. L. Williams, "Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making," *Policy & Internet*, vol. 2, no. 7, pp. 223–242, 2015.

[31] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive Language Detection in Online User Content," in *25th International Conference on World Wide Web*, 2016, pp. 145–153.

[32] P. . O'Sullivan and A. . Flanagin, "Reconceptualizing 'flaming' and other problematic messages," *New Media Soc.*, vol. 5, pp. 69–94, 2003.

[33] A. Mahmud, K. . Ahmed, and M. Khan, "Detecting Flames and Insults in Text," in *In Proceedings of the 6th International Conference on Natural Language Processing*, 2008.

[34] A. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, "Offensive Language Detection Using Multi-level Classification," *Springer*, p. 1627, 2010.

[35] I. Chaudhry, "Hashtagging hate: Using twitter to track racism online," *First Monday 20(2)*, 2015. .

[36] S. Liu and T. Forss, "New classification models for detecting Hate and Violence web content," in *7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, 2015, pp. 487–495.

[37] A. Gaydhani, V. Doma, S. Kendre, and L. Bhagwat, "Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach," 2018.

[38] M. Hasanuzzaman, G. Dias, and A. Way, "DemographicWordEmbeddingsforRacismDetectiononTwitter," in *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, 2017, pp. 926–936.

[39] N. Djuric, J. Zhou, M. Morris, Robin Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in *In Proceedings of the 24th InternationalConferenceonWorldWideWeb(WWW2015)*, 2015, pp. 29–30.

[40] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," 2014. [Online]. Available: https://nlp.stanford.edu/pubs/glove.pdf. [Accessed: 19-Sep-2019].

[41] Z. Mossie and J.-H. Wang, "SOCIAL NETWORK HATE SPEECH DETECTION FOR AMHARIC LANGUAGE," in *COMIT*, 2018, pp. 41–55.

[42] A. Al-Hassan and H. Al-Dosari, "Detection of Hate Speech in Social Networks: A Survey on Multilingual Corpus," in *6th International Conference on Computer Science and Information Technology*, 2019.

[43] D. Gamal, M. Alfonse, M. E.-H. El-Sayed, and A.-B. M.Salem, "Twitter Benchmark Dataset for Arabic Sentiment Analysis," *Int. J. Mod. Educ. Comput. Sci.*, vol. 11, no. 1, pp. 33–38, 2019.

[44] W. Alorainy, P. Burnap, H. Liu, and M. L. Williams, "'The Enemy Among Us': Detecting Cyber HateSpeech with Threats-based Othering Language Embeddings," *ACM*, 2019.

[45] K. N. B. of Statistics, "2019 Kenya Population and Housing Census Volume I: Population by County and Sub-County," 2019.

[46] H. Kim, S. Jang, Mo, S.-H. Kim, and A. Wan, "Evaluating Sampling Methods for Content Analysis of Twitter Data," *Sage*, 2018.

[47] A. E. Kim, H. M. Hansen, J. Murphy, A. K. Richards, J. Duke, and J. A. Allen, "Methodological Considerations in analyzing Twitter data," *J. Natl. Cancer Inst.*, vol. 47, pp. 140–146, 2013.

[48] P. . Cavazos-Rehg *et al.*, "A content analysis of depression-related tweets," *Comput. Hum. Behav.*, vol. 54, pp. 351–357, 2016.

[49] C. Shearer, *The CRISP-DM model: the new blueprint for data mining*. 2000.

[50] F. Provost and T. Fawcett, *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*, First Edit. O'Reilly Media, Inc., 2013.

[51] Z. Waseem, “Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter,” in *EMNLP Workshop on NLP and CSS*, 2016, pp. 138–142.

[52] W. Warner and J. Hirschberg, “Detecting Hate Speech on the World Wide Web,” in *Language in Social Media (LSM 2012)*, 2012.

[53] P. Burnap and M. L. Williams, “Us and them: identifying cyber hate on twitter across multiple protected characteristics.,” *EPJ Data Sci.*, 2016.

[54] R. . King and G. M. Sutton, “High Times for Hate Crime: Explaining the Temporal Clustering of Hate Motivated Offending,” *Criminology*, vol. 51, no. 4, pp. 71–94, 2013.

[55] J. Brownlee, *Deep Learning for Natural Language Processing*, V1.2. 2018.

[56] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[57] M. Elsherief, V. Kulkarni, D. Nguyen, W. Wang, and E. Belding, “Hate lingo: A target-based linguistic analysis of hate speech in social media,” in *12th International AAAI Conference on Web and Social Media*, 2018, pp. 42–51.

[58] N. Coupland, “‘Other’ representation, Society and Language.” John Benjamins Publishing, 2010.

[59] G. R. Semin, “Linguistic Markers of Social Distance and Proximity.” 2009.

[60] M. Cikara, M. M. Botvinick, and S. T. Fiske, “Us versus them: Social identity shapes neural responses to intergroup competition and harm,” *Psychol. Sci.*, vol. 22, no. 3, pp. 306–313, 2011.

[61] N. Haslam, “Dehumanization: An integrative review,” *Personal. Soc. Psychol. Rev.*, vol. 10, pp. 252–64, 2006.

[62] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[63] K. Krippendorff, “Computing Krippendorff’s Alpha-Reliability,” *University of Pennsylvania ScholarlyCommons*, 2011. [Online]. Available: mhttp://repository.upenn.edu/asc_papers/43.

[64] W. Clyne, S. Pezaro, K. Deeny, and R. Kneasfsey, “Using Social Media to Generate and Collect Primary Data: The #ShowsWorkplaceCompassion Twitter Research Campaign,” *JMIR Public Heal. Surveill*, vol. 4, no. 2, p. e41, 2018.

[65] V. Dijk and A. Teun, “Discourse and racism, The Blackwell companion to racial and ethnic studies,” pp. 145–159, 2002.

[66] S. Sood, J. Antin, and E. Churchill, “Profanity use in online communities,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1481–1490.

[67] Edward Ombui, Lawrence Muchemi, Peter Wagacha, "Building and Annotating a Codeswitched Hate Speech Corpora", International Journal of Information Technology and Computer Science, Vol.13, No.3, pp.33-52, 2021.

[68] A. Schmidt and M. Wiegand, “A Survey on Hate Speech Detection using Natural Language Processing,” *SocialNLP@EACL*, 2017.

[69] P. Fortuna, L. da Silva, Jo˜ao Rocha Soler-Company, Juan Wanner, and S. Nunes, “A Hierarchically-Labeled Portuguese HateSpeech Dataset,” in *Proceedings of the Third Workshop on Abusive Language Online*, 2019, pp. 94–104.

[70] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, “Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis,” *arxiv:1701.08118*, vol. 1, 2017.

[71] Y. R. Tausczik and J. W. Pennebaker, “The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods,” *J. Lang. Soc. Psychol.*, vol. 1, no. 29, 2010.

[72] E. Alpaydin, *Introduction to Machine Learning*, 2nd Editio. London: The MIT Press, 2010.

[73] L. Chen, “Support Vector Machine — Simply Explained,” *Towards Data Science*, 2019. [Online]. Available: https://towardsdatascience.com/support-vector-machine-simply-explained-fee28eba5496. [Accessed: 02-Apr-2020].

[74] J. Brownlee, *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*. 2016.

[75] F. Peng, D. Schuurmans, V. Keselj, and S. Wang, “Language independent authorship attribution with character level n-grams,” in *10th Conference of the European Chapter of the Association for Computational Linguistics*, 2003, pp. 267–274.

[76] J. Kruczek, P. Kruczek, and M. Kuta, “Are n-gram Categories Helpful in Text Classification?,” in *International Conference on Computational Science*, 2020, pp. 524–537.

Authors’ Profiles

Edward Ombui is a lecturer at the School of Science and Technology, Africa Nazarene University, Kenya. He is a Ph.D. candidate at the University of Nairobi. His education includes an MSc in 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 Ph.D. 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 include 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 Wagach, "Psychosocial Features for Hate Speech Detection in Code-switched Texts", International Journal of Information Technology and Computer Science(IJITCS), Vol.13, No.6, pp.29-47, 2021. DOI: 10.5815/ijitcs.2021.06.03