Psychosocial Features for Identifying Hate Speech in Social Media Text

Edward Ombui[1*], Lawrence Muchemi[2], Peter Wagacha[3]

School of Science and Technology, Africa Nazarene University, Nairobi, Kenya;
School of Computing and Informatics, University of Nairobi, Nairobi, Kenya

ABSTRACT

This study uses natural language processing to identify hate speech in social media codeswitched text. It trains nine models and tests their predictiveness in recognizing hate speech in a 50k human-annotated dataset. The article proposes a novel hierarchical approach that leverages Latent Dirichlet Analysis to develop topic models that assist build a high-level Psychosocial feature set we call PDC. PDC organizes words into word families, which helps capture codeswitching during preprocessing for supervised learning models. Informed by the duplex theory of hate, the PDC features are based on a hate speech annotation framework. Frequency-based models employing the PDC feature on tweets from the 2012 and 2017 Kenyan presidential elections yielded an f-score of 83 percent (precision: 81 percent, recall: 85 percent) in recognizing hate speech. The study is notable because it publicly exposes a rich codeswitched dataset for comparative studies. Second, it describes how to create a novel PDC feature set to detect subtle types of hate speech hidden in codeswitched data that previous approaches could not detect.

Keywords: Hate Speech, Code-switching, Feature selection, Machine learning

1. INTRODUCTION

Hate speech is rhetoric that targets an individual or group based on protected characteristics like ethnicity, religion, or gender [1]. This needs a lot more attention than it is getting now. People of African and Asian heritage are being targeted with growing frequency in the US, as well as ethnic hatred and genocide in some African countries [3, 4, 5, 6]. Increasingly, campaign-related incidents trigger online public comments bordering on hate speech. Famously, politicians stir negative ethnic feelings, often provoking intense public reactions and counter-reactions on social media [5]. In Kenya, the lack of particular policy frameworks to hold media corporations, especially social media, accountable for hate speech promoted on their platforms exacerbates the situation. Instead, the legislation available during this study targeted individual users, including local administrators of social media networks like WhatsApp [6].

Social media user-generated content challenges traditional natural language processing, computational linguistics, and machine learning methodologies. In addition to being noisy and irregular, big data is also massive, diverse in data types, real-time created, and codeswitched. Codeswitching is a widespread social phenomenon that indicates group membership [7]. Codeswitching is becoming more common in everyday communication among multilingual societies, especially on social media. In addition, code-switching on social media appears to be the de facto in-group lingua franca. It is seen as boosting group cohesion and distancing oneself from others, especially perceived adversaries. Codeswitching is also commonly employed to accentuate a point in a message. Because some social media sites allow anonymous interaction, they are a breeding ground for hate speech.

Our work focuses on recognizing hate speech in codeswitched text messages received from social media. This is a difficult categorization issue that standard approaches have failed to address well, frequently by omitting the codeswitched text. An in-depth analysis of the data collected for this study indicated that some of the most pervasive hatred on social media is frequently disguised in coded text messages. Because social media communication is sometimes casual, it is not commonplace to encounter messages that alternate between various languages, particularly among multilingual communities. This complicates the task of sentence parsing and conducting contextual analysis on phrases using conventional monolingual methods. The scarcity of native language resources, such as parts-of-speech taggers, corpora, and dictionaries[8], along with undocumented grammar rules and disjointed research networks, appears to aggravate the situation [9]. As a result, extracting high-quality features from this type of data for machine learning applications requires a novel strategy that addresses the flaws in typical data processing techniques. As a result, the entire process of collecting, annotating, and selecting high-quality features that best describe hate speech in a codeswitching context for the goal of training a machine classifier becomes complex and expensive.

Previous research on hate speech identification has focused on monolingual datasets, with the most often used language being English. However, communication occurs on social media platforms in a variety of different regional and indigenous under-resourced languages, including Amharic, Bengali,

Seneca, and Swahili. Given that more than half of the world's population is multilingual [10], we hypothesize that this number is increasingly reflected on social media through evidence of language code-switching. For example, practically the whole indigenous population in Kenya is multilingual, able to communicate in their mother tongue (L1), Swahili or Kiswahili, the national language (L2), and/or English, the official language (L2) [13].

However, whenever codeswitching occurs, whether at the word or sentence level, prior similar research regarded this content as noisy data and chose to delete the entire phrase during the preprocessing stage. Rather than dropping, is there any way to properly manage this increasingly popular social media phenomenon? Our study tries to address this challenge by examining a variety of variables to uncover distinguishing characteristics for training a computer classifier to identify hate speech in codeswitched text messages extracted from social media. As a result, our study fills a void in the field of automatic hate speech recognition for codeswitched language datasets. To our knowledge, this is the first effort to gather and train a classifier for a dataset of codeswitched languages, specifically English, Swahili, Sheng (slang), and a few instances of vocabulary from indigenous languages such as Gikuyu and Luo. As an illustration of a text message, consider the following:

"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 context, the purpose of our work was to investigate a methodology that more accurately captures essential characteristics of nuanced types of hate speech, particularly in codeswitched text messages, to improve the effectiveness of machine categorization of large amounts of data. The primary objectives included the development of a conceptual framework for hate speech, the collection of a hate speech dataset from Kenyan social media, the training and evaluation of a hate speech classification model. This study makes two contributions. To begin, it creates and publicly releases a code-switched dataset of hate speech that may be used for comparative studies by other scholars. Second, the study presents a novel psychosocial feature subset that encompasses four characteristics of language use: psychosocial distance, stereotyping, commitment to hate, negative passion, and hatred as a narrative. These are then utilized to extract salient features that can be used to reliably train a machine classifier to recognize nuanced kinds of hate speech in codeswitched messages.

2. LITERATURE REVIEW

Earlier research has employed a variety of techniques to decipher hate speech characteristics in text messages. Among these are the application of hate theories and frameworks. Critical race theory was utilized to develop standards for annotating a corpus of racist texts. [11]. Additionally, one study established a framework for analyzing offensive messages contained in text documents[12]. However, past research on critical race theory has been limited to classification based on race and the interaction of law and power. As a result, the theory falls short of defining other forms of hatred, such as those motivated by gender, religion, or disability. While some researchers have devised frameworks to assist in identifying harmful language[4, 15], these frameworks lack essential theoretical underpinnings and frequently rely on word lists. As a result, there is a need to close this gap, which this study does by developing a wholistic hate speech framework comprised of psychosocial characteristics and grounded in sound theory. This is meant to be sufficiently comprehensive to allow for the identification of different forms of hate in social media messages.

There is a growing body of study being undertaken on hate speech, including automated methods for detecting hate speech[14, 15, 16] and other related topics such as offensive language identification[17, 18], cyberbullying [19, 20], radicalization and Terrorism [21, 22]. The studies on hate speech have handled the automatic classification problem in one of two ways: as a binary classification work or as a multi-class classification task. The former technique has been used in several earlier studies that focus on identifying subtypes of hate speech such as racism [15, 23, 24] and anti-Semitism [13]. It is not enough to recognize black and white in a multi-class classification test; it is also necessary to recognize the grey shade on the continuum of hate speech and non-hate speech (Ok) communications. Gray messages are recorded in this way by having an "offensive" class, which is analogous to how a human annotator would normally view and classify messages. Several previous studies have used multi-class categorization[26, 27, 28].

According to the review of these studies, numerous features have been deployed with varying degrees of effectiveness in increasing the detection of hate speech in text messages. These features

can be classified into two types: high-level and low-level. The high-level properties of a text message are human-readable and frequently qualitative notions that a human annotator may detect and use to classify the message. This category encompasses syntactic, stylistic, semantic, and lexical characteristics. The length of the message, the usage of part-of-speech tags, and the use of imperatives are all syntactic aspects. Stylistic characteristics include the usage of capital letters, exclamation points, emoticons, and character and punctuation overloading [11]. Semantic features include associational terms, hate verbs, negative polarity, and the use of subjective nouns. Lexical features include word lists containing accusatory and attributional keywords[19, 29], abusive words [28], insults or flames [17, 31, 32], and offensive language [18, 25, 33] that include racial slurs [24, 34].

Other common representations of features include N-grams, Bag of words (BoWs), and word embeddings. BoWs frequently produce a high recall value but a low precision due to false positives [33]. This is because the mere appearance of hate or derogatory terms in a message automatically classifies it as hate speech, regardless of the context in which the term is used [15, 25, 29]. The N-gram features can exist in two levels: as a 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, in contrast to the original plaintext, are frequently retrieved features that are machine-processable, meaning they can be used instantly by a machine-learning algorithm to train models. These are frequency counts computed using BoWs, N-grams, and word embedding representations such as count vectors, term frequency-inverse document frequency (TF-IDF), one-hot vector encodings, and dense vectors. In studies on hate speech identification, pre-trained word embeddings as dense vector representations are becoming increasingly popular as the preferred feature set for training deep learning algorithms [26, 37, 38]. Word2Vec text representations, FastText n-grams, and Global Vectors (GloVe) [37] are all frequently used pre-trained embeddings at the sentence, word, and character levels. The frequency of use of both levels of features in prior hate speech research is summarised in Fig. 1.

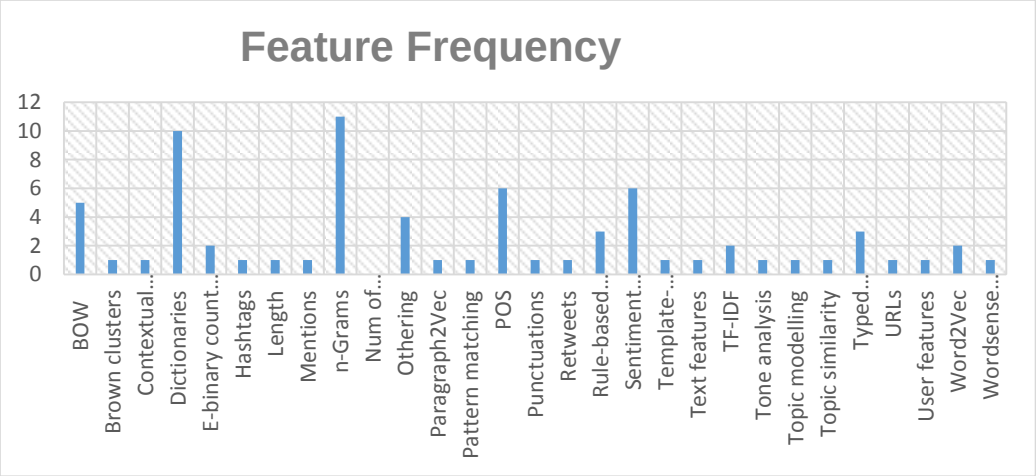

Fig. 1. Frequency of Features in hate speech studies

Notably, according to the reviewed literature, a large number of studies were conducted using English datasets, with just a small number of similar studies being conducted in other languages, such as Dutch [23], Amharic [38], Arabic [41, 42], but none in Swahili.

In general, the features used for text classification are crucial in determining the trained model's efficacy and accuracy at discriminating between occurrences of different classes. It is necessary to identify the features, examine them, and select the most significant among them to inform the training of a machine classifier. Various characteristics have been used in past research to categorize hate speech. These, on the other hand, have frequently been convoluted, thus increasing the difficulty of comprehending and applying them. This research both theoretically and practically breaks down the complexity of these qualities into two basic groups, namely, high-level features and low-level features, to better understand them. Individuals who annotate high-level features report that they are easily comprehendible and immediately identifiable. As demonstrated in Fig. 2, these are further subdivided into psychosocial, linguistic, and App-specific features. As a result of this abstraction, a new methodology for capturing latent traits, such as the "othering" language, is introduced, which has proven useful in catching subtle kinds of hate speech in a prior study[41]. Besides, using a

comprehensive hate speech conceptual framework, our study argues that these latent features are easily identifiable via psychosocial concepts and, when combined, become informative features for identifying subtler forms of hate speech that conventional methods, particularly supervised machine learning, were insufficient to capture in the first place.

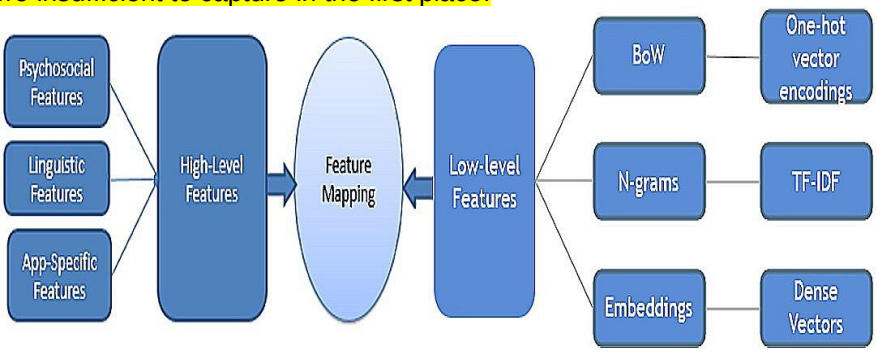

Fig.2 Mapping of Feature from high-level to low-level

3. METHODOLOGY

The four study objectives were addressed using a mixed-methods approach. To begin, a qualitative approach was utilized to define the discriminant characteristics of hate speech through an examination of important themes arising from diverse hate speech definitions and theories in the literature. These are succinctly expressed in fig.4's hate speech framework. Following that, the framework guided the collecting and manual annotation of tweets into three predetermined categories, namely Hate Speech, Offensive, or Neither. Following that, a quantitative approach was utilized to obtain word frequencies for each class, and other low-level features such as TF-IDF and word frequency vectors were employed to train the classifier model.

All processes, from data preprocessing to data analysis, feature processing, model training, and final classification, were consolidated using the Jupyter notebook integrated development environment. This was done by utilizing Python programming (version 3.6.8) and machine learning packages. These included the natural language toolkit (NLTK) for data processing, Pandas for data visualization, Scikit-learn for multiple kinds of machine learning models, and Matplotlib for data plotting, among others.

The study employed the ethnic group names of seven of Kenya's forty-two major tribes as the study population parameter [42], crawling tweets using the terms *Kikuyu, Luhya, Kalenjin, Luo, Kamba, Kisii, and Meru*, as well as their Swahili equivalents. Additionally, these ethnic names were employed in conjunction with other phrases to collect and create the raw dataset, as instructed by the multidimensional hate speech framework.

Unlike conventional research, which employs traditional sampling methods, big-data projects employ alternative sampling methods that computationally collect all available online content [43], such as by using a web crawler or Twitter API to collect a large number of messages from social media based on specific keywords. Such methods are frequently free of the limits associated with classic sampling methodologies [44], which would have made it inefficient and impractical to collect a significant volume of data on hate speech from a big number of social media users in Kenya for the intent of machine learning. Our study used simple random sampling to choose a study sample for annotation from the large volume of obtained data. In previous studies [25, 47], this sampling strategy was utilized to generate study samples from social media.

The CRoss-Industry Standard Processes for Data Mining (CRISP-DM) [46] was utilized to inform the five workflow procedures needed to accomplish the study's primary objective of establishing the salient elements required to develop a hate speech classifier for codeswitched communications. These steps comprised defining the problem, preparing data, processing features, modeling, and evaluating, as seen in Fig. 3.

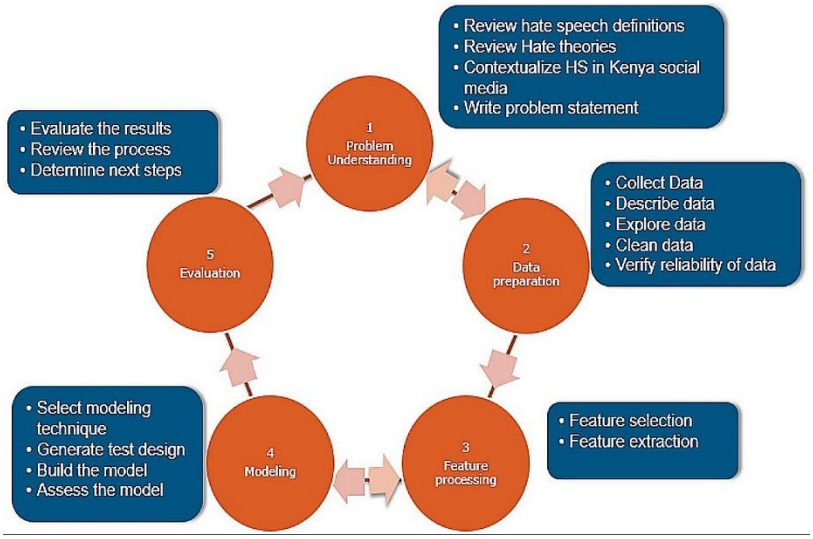

Fig. 3. The five-step research workflow

The experimental procedure was directed by these five phases, which have been shown to boost exploratory data analytics endeavors in the past[47], an approach that corresponded perfectly to the process activities incorporated in our study objectives. These are briefly covered in the subsections that follow.

3.1 Problem understanding

The purpose of this phase was to construct a working definition of hate speech for the study by first attempting to comprehend the environment in which the phenomena of hate speech exist to build a solution that is practicable within the problem's natural setting. In this regard, pertinent material, both online and in print, was rigorously researched to gain a thorough understanding of the problem of hate speech on social media in Kenya. The snowballing technique was employed to browse the cited articles that were referenced in the seminal literature research to gain further insights. Moreover, a qualitative analysis was conducted to identify relevant themes by examining the content of six hate theories, the definitions of hate speech contained in the user-content regulations of various social media platforms, and the legal standards of hate speech contained in Kenyan public policies.

3.2 Data Preparation

This phase involved data collecting, data annotation, and data cleansing. Convenience sampling was used to collect tweets during the Kenyan presidential campaign in August 2017 and the runoff election in October 2017. Bootstrapping was the major data collection strategy. The strategy used to crawl Twitter for messages included the use of hate-related keywords [48], phrase patterns having a negative connotation [49], pro-hate user accounts, and offensive hashtags [14]. In contrast to other social media platforms, Twitter messages are by default publicly visible, thematically organized, and programmatically accessible. Notably, several similar hate speech studies have used tweets[15, 28, 51]. As a result, the Twitter API was utilized to create an application that gathered tweets during election week. A crawler-based on Python programming was also utilized to supplement Twitter API's two-week data gathering window to obtain a massive amount of archived tweets dating back to the Kenyan presidential elections in March 2013. Furthermore, presidential campaign times and events are prominent trigger events that result in spikes in online hate speech [51].

Each message in the dataset was manually classified by human coders. An initial team of forty undergraduate computer science students and staff members (80:20) was recruited and trained on the annotation scheme by convenience sampling. The group comprised of twenty-one male and nineteen female annotators, with an average age of twenty-three. The team's nationality was weighted towards Kenyans since annotators needed to comprehend the corpus's code-switching nature, which includes texts in English, Swahili, and other indigenous languages. The initial training used the annotation technique to help the team understand hate speech. The annotators were then instructed on how to annotate example messages through the study team's web-based annotation portal [9].

The original team of forty annotators was reduced to twenty-seven having dropped outliers. One week of annotating at least three thousand messages was required for selection. The first session provided valuable feedback on the annotation portal's speed. Previously, every tweet had to be tagged by a

distinct team, one of which was a subject matter expert (SME). The revised design was inspired by the first session's sluggish annotation process and the necessity for a larger labeled dataset for machine classification training.

The second step was data cleaning, which aimed to improve the dataset's quality by removing noisy signals that could harm a machine learning model's training and overall performance. All words were lowercased to standardize the data, whereas empty rows, emojis, and punctuation were removed from the data using the regular expression library in Python. For example, the following python function was utilized to remove the emoticons, symbols, and pictographs in the tweets.

```
def remove_emoji(tweet):
emojis = re.compile("["\U0001F600-\U0001F64F"  # emoticons
                    u"\U0001F300-\U0001F5FF"]\n"  # symbols & pictographs)
return(emojis.sub(r'', tweet))
```

3.3 Feature Engineering

This phase's objective was to extract a subset of informative and high-quality vocabulary from the high-dimensional output in the data preparation stage. Following that, this textual subset comprised of high-level features needed to be translated into a low-level numeric representation amenable to machine learning via feature extraction. This is because machine learning algorithms are limited to processing numerical representations of features, such as vectors [52].

The high-level features were divided into two groups: the general vocabulary developed during the data processing step, and the PDC dictionary, which had five feature categories guided by the multidimensional hate speech framework seen in Fig. 4. Both categories were converted to BoW frequency counts, n-grams, and word embeddings. Following that, three low-level characteristics were retrieved from these: one-hot encodings, TF-IDFs, and dense vectors. The BoWs' features were determined based on the frequency of term occurrences in each communication. The n-grams were processed at the word and character levels, with n ranging from two to five. These were computed using the Scikit-learn machine-learning library's count vectorizer. The TF-IDF features were utilized to determine the relevance of a term both within a document and across the full dataset. The overarching objective is to penalize terms that appear inordinately frequent throughout all texts. This is because they may be less informative to the model than terms that are unique to particular documents but are found in all documents. As a result, the TF-IDF vectors for the various levels were created. Concerning Word Embeddings, the GloVe pre-trained embeddings were utilized to transform each word to a similar high-dimensional vector using the 100d file of around 1 million word vectors. To begin, the message dataset was tokenized. Following that, each token was mapped to its associated embeddings using the transfer learning approach.

The usefulness of these features was determined by training and comparing several classifiers. Besides, additional variables such as POS and topic models were extracted from the general lexicon and evaluated for their ability to improve the performance of classifiers.

Topic Models[53], were used as high-level features for knowledge discovery, linking the data with the conceptual framework, and, more importantly, as an automated technique for identifying which significant terms to include in the later step of building PDC word-family features. As a result, twenty-three semantically relevant topics or clusters were generated using the Latent Dirichlet Allocation (LDA) method from a vast corpus of short-text communications from social media. PDC traits are psycholinguistic characteristics derived from the triangle theory of hate's three dimensions of hate [54]. PDC espouses hate speech in three basic word families, all of which are concept-based and language-independent. As a result, the language list can be expanded or contracted by adding or eliminating terms with comparable meanings in different languages from their respective word families. The passion word family includes expressions of negative emotions including fear, anger, disgust, and disdain. Threats and derogatory slurs directed to a specific individual or group based on protected characteristics like gender, ethnic origin, or religion are included in this category of offensive language. In previous studies, negative polarity and sentiment analysis were employed to discover instances of passionate emotions [18, 33]. The **D**istance word family comprises terms that represent psychological distancing or proximity in intergroup or interpersonal relationships, often known as "othering" language [50]. This is frequently suggested by a pronoun-heavy vocabulary like “we,” they”, and so on [55, 56, 57, 58]. An example message is “***We*** *make better leaders...****they*** *are Cowards*”. The **C**ommitment word family comprises terms that degrade others. This can be conveyed by portraying other people as objects, insects, animals, or by making broad generalizations about their depravity, immaturity, or callousness [59]. Additionally, this section contains several code names that are mostly understood by the in-group membership when referring to members of the out-group.

Using the Scikit-learn package, these high-level text properties were encoded as input vector values. The text messages were transformed to word count vectors using CountVectorizer and Tfidf Vectorizer. In every case, messages are tokenized first, then a lexicon of recognized terms is built. The result is a vector encoding the entire vocabulary. A fixed-length vector corresponding to the vocabulary's length is then used to encode new text messages. The CountVectorizer populates each vector position with a count of the occurrences of each word in the new text message. If a word in the resultant text message does not exist in the vocabulary, it is ignored and consequently excluded from the final vector. The Tfidf Vectorizer analyzes word frequencies and gives a high score to often occurring terms inside a paper, but degrades the most frequently occurring words across all publications. When encoding new text messages, the scores, which are often between 0 and 1, are utilized to weight the vector's frequencies.

3.4 Modeling

This procedure involved the selection of a model, training, and parameter tweaking. The classification models were trained using both conventional and deep learning strategies. A review of promising findings from prior similar investigations influenced the precise choice of machine learning techniques. The Support Vector Machine, Naive Bayes, Decision Trees, Linear Logistic Regression, and K-Nearest Neighbor were all instances of well-known machine learning techniques. Additionally, Bagging and Boosting models were utilized, namely Extreme Gradient Boosting, and Random Forest. Convolutional Neural Networks and Hierarchical Attention Networks were used as deep learning techniques. Numerous machine learning experiments were conducted in this work using models equivalent to those found in Python's Scikit-learn machine learning model library.

Each model was trained using a collection of hyperparameters that were discovered and placed in a parameter grid. These were then updated automatically during the testing using Grid search and tenfold cross-validation[60] to score the combination of feature parameters and determine the model's ideal hyperparameters. These included the soft margin cost, C, the kernel, and additional estimator parameters. For the nonlinear Support Vector Machine, for example, generalization was assessed by adjusting the soft margin cost, C, to lower penalty values ranging from 0.001 to 1.0. Three widely utilized kernels from the literature were used in the trials to assist the model in establishing a nonlinear decision boundary: Radial Basis Function, linear, and Polynomial. The SciKit-Learn packages were used to choose all of these model parameters. Additionally, each time the algorithm was run, a pipeline was employed to smoothly merge these parameters with the vectorizer settings.

3.5 Evaluation

To mimic how our model might behave in the future, we partitioned the input dataset into training and testing sets. The confusion matrix was used to assess the trained models' accuracy performance by comparing predicted values to actual values from the test dataset. Additionally, the F-score was employed, which is based on the weighted average of precision and recall values. The highest prediction accuracy guided the selection of the optimal model for identifying the positive class, i.e., hate speech, with 10% serving as the validation data set, and model testing performed using K-fold (10-fold) cross-validation.

In this study, an inter-rater reliability score was calculated premised on the annotations provided by a team of twenty-seven human annotators. At least three human annotators were required to annotate each tweet. Statistically, the mode decided the tweet's class, implying that the tweet's class was selected by two or more votes. An additional annotator was introduced as a tie-breaker in the case of a tie. This fourth annotator ought preferably to be a subject matter expert. Krippendorff's Alpha was adopted as an inter-rater measure of reliability for the annotation exercise since it could deal with incomplete data and misfits. [61].

The triangulation approach was used to establish the construct and prediction validity of the data and framework concepts. This involved analyzing the performance of various traditional and deep learning machine learning techniques to select the optimal feature set for training our classifier.

3.6 Ethical Consideration

It is common for researchers to highlight concerns about user consent and user identity protection when using social media as the major source of data[62]. But unlike the privacy settings on other social networking sites, all messages sent on Twitter are made publicly viewable by default. As a result, the study focused on public tweets and retweets which do not require formal consent. Online users' identities were protected by substituting user names with a generic label.

4. Results and Discussion

The findings from this original research article are presented in this section. These are based on the content analysis, data collection, data pre-processing, model training, model evaluation and generalizability.

4.1 Theoretical Framework for Hate Speech

Distancing language, negative passion, devaluation, subjectivity, and stereotyping were identified as five key features of hate speech based on content analysis of many hate theories and hate speech definitions from various literature sources.

The use of pronouns in the text, particularly third-person plural nouns in English and Swahili, was indicative of distancing, also known as othering language. Several scholars have already utilized this idea to identify aspects of hate speech [43, 51, 56, 63]. Distancing is also obvious in social media messaging where one social group asserts superiority over another or isolates itself to safeguard the "purity" of the group membership. For example, during Kenya's post-election violence in 2007/2008, the Swahili phrase "madoadoa," which means "spots," was used to disseminate hate speech about non-natives by some politicians.

The negative passion dimension was marked by powerful feelings of hatred, fear, and antagonism toward the target individual or group. The material includes expletives such as swear words, obscenities, abusive, disparaging, and other offensive languages. This dimension has been used in several earlier research to identify hate speech [18, 33, 64]. Furthermore, negative passion was obvious in writing that incited violence against a person or a group because they shared a protected social feature. Devaluation is a hate-filled commitment characterized by the use of insulting language in text messages to refer to a target group as animals or insects. Using terms like maggots, cockroaches, rats, and so on to describe the target population. Other hate speech studies[4, 59] have employed this dimension.

Subjectivity was defined by the use of biassed or defective arguments, as well as the use of quantifiers and certainty phrases such as "always," "never," and "all." Stereotyping was defined as the use of a person's ethnic, racial, or religious group names to refer to them. Kikuyus, Luhyas, Kisiis, and other ethnic groups are examples. These five aspects, as well as their interactions, were included in the multidimensional hate speech framework, as depicted in Fig. 4.

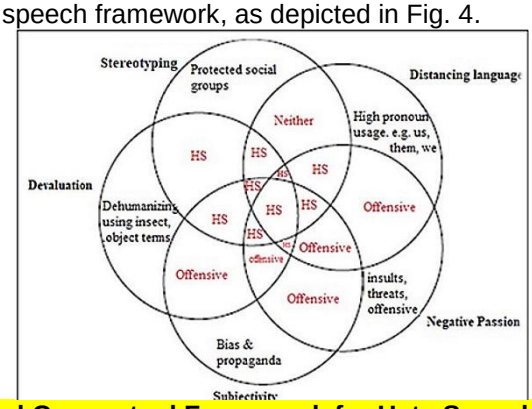

Fig. 4. The multidimensional Conceptual Framework for Hate Speech

4.2 DATA

Around 400k tweets, mostly comprising tweets posted in the 2017 Kenyan presidential elections, were collected and stored in a comma-separated file format.

Languages represented in the sample included English, Swahili, and code-switched communications, with the majority of code-switched messages being English-Swahili in nature. As an illustration,

"*We will not relent,* ***hata kama watatumi****a police to rig the elections.*" [Eng-Swa Codeswitched "....., ***even though they will use*** the police...."]

Table 1 provides a high-level summary of the dataset's characteristics.

Table 1. Raw Dataset Description

Description	Number of	Examples

	Messages	
Messages collected in raw form	401,211	;2017-12-08 18:09;6;13;"Kikuyus are ILLEGAL immigrants all over. Canâ€™t ‰ã□®é€ ";;;;"939149879514943488"; https:// twitter.com /xxxx3210/ status â€¦
Number of preprocessed text messages	398,000	kikuyus are illegal immigrants all over. Can t
Number of Codeswitched messages	29309	***Wacha Ujinga*** that is not what the president said! Luos are the worst tribalists. [Swa-Eng translation: Don't be foolish]

From the 400k messages, a sample of 60k was randomly selected for annotation. A team of three annotators annotated each tweet, with the class selected by a majority vote. As seen in Table 2, around 50k tweets were annotated, with 6% containing hate speech, 19% including offensive language, and 75% classed as 'neither.' The class of hate speech was in the minority. This is to be expected given the size of the social media dataset and is consistent with a prior study [68]. Analysis of the data indicated that ethnic hatred is the most prominent form of hate speech in Kenya during electioneering periods. Thus, ethnic hate speech vocabulary might be substantially and broadly employed to create a classifier model for the Kenyan context. Second, in contrast to binary classifier frameworks, the addition of the "offensive" class made it easier to explicitly distinguish hate and offensive messages, thus reducing the probability of falsely labeling tweets as hate speech, which is a frequent problem during annotation exercises and allowing for more accurate classification [24].

Table 2. Annotation Class distribution

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

Using Krippendorf's Alpha, the inter-coder reliability score was 0.5207. This meant that the annotators were not in complete agreement half of the time. This is in line with prior studies[65] that had a lower inter-rater score of 0.17. The low inter-rater agreement has been attributed to a variety of personal sensitivities and societal prejudices, as well as the use of inexperienced but inexpensive annotators [66]. In our example, a few annotators who did not consistently attend the complete annotation course injected some teacher-noise into the annotations, lowering the score. Despite the training, the teacher noise, together with the team's tacit knowledge and biases during annotation, might form part of the latent qualities that were modeled as random components in the noise signal. Another reason for a large number of missing annotations could be 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 from the twenty-seven. The necessity to maximize the volume of comments from a team of human coders while using the fewest resources possible inspired this approach. Random undersampling was used on the majority class, i.e. the 'neither' class, as well as the 'offensive' class, to better train a robust and unskewed classifier. This yielded a dataset of 9726k tweets that was fairly balanced, with majority votes in each of the three classes. In addition, a finer and more balanced dataset of 2537k tweets with just full agreement annotations was created. Both datasets were used to train machine learning algorithms in the studies and are freely available on Kaggle.

Using topic modeling, the Latent Dirichlet Allocation (LDA) model helped to uncover deep underlying concepts of hate in a huge corpus of code-switched text. [53], just as previously used to identify cyberbullying subjects [17]. Each word in the corpus is represented by LDA as a finite mixture of underlying hate concepts, modeled over an unlimited number of topics characteristic of a text document[53]. This contributes to the establishment of a probabilistic model for the codeswitched corpus, which assigns a high probability to messages that are highly related to the corpus' membership and other messages that are comparable to them. As a result, the LDA technique is effective for preprocessing data and as a first-level statistical tool for automatically detecting and extracting passion, distancing, and discriminative (PDC) features from the huge corpus in this work. These characteristics were found in the twenty-three latent subjects recovered as a "bag of words" that were closely related to the hate speech category. The use of LDA, on the other hand, revealed

the limits of the bag-of-words approach, where the order of words is not maintained and hence the context or word meaning is not preserved.

4.3 Modeling

The purpose of this work was to determine the predictive power of the psychosocial (PDC) feature set as compared to the traditional high-level features when training a classifier for hate speech. The conventional features included the lexical features (LEX) that defined the input corpus's generic lexicon. The BoWs and n-grams were retrieved, as well as Part of Speech (POS) and Application-specific (APP) variables such as the frequency of retweets and likes. As a result, nine machine learning models were trained and their performance compared to establish the optimal model for detecting subtle forms of hate speech in codeswitched messages. To determine the optimal features and model performance, the features were evaluated singly and in combination using a feature combo. The wrapper technique was adopted, in which the PDC feature set began with only a few elements classified as psychosocial, as determined by the LIWC psychological word list[67]. The categories elements were expanded over time when additional features were uncovered in previously reported hate speech texts, as well as translation equivalents to account for codeswitched occurrences. Additionally, the Lex features were lacking in comparison to the comprehensive and instructive PDC elements. This can be explained again by the vectorizer's random feature sampling technique for extracting the Lex features from the input dataset, which includes a parameter for specifying the number of features. With text, as the number of features increases, the computation becomes more complicated, particularly in terms of memory and compute time necessary to handle the highly sparse input vector. In general, unlike the often wide and "diluted" Lex feature set, the PDC feature set consists of fewer but highly informative characteristics for hate speech identification. As illustrated in Fig. 5, supplementing the usual Lex feature set with PDC always resulted in improved performance. On the other hand, the inclusion of Lex or other features resulted in a decrease in performance. This is because these features generate noise, which results in the sparsity of the incoming data vector in classifiers.

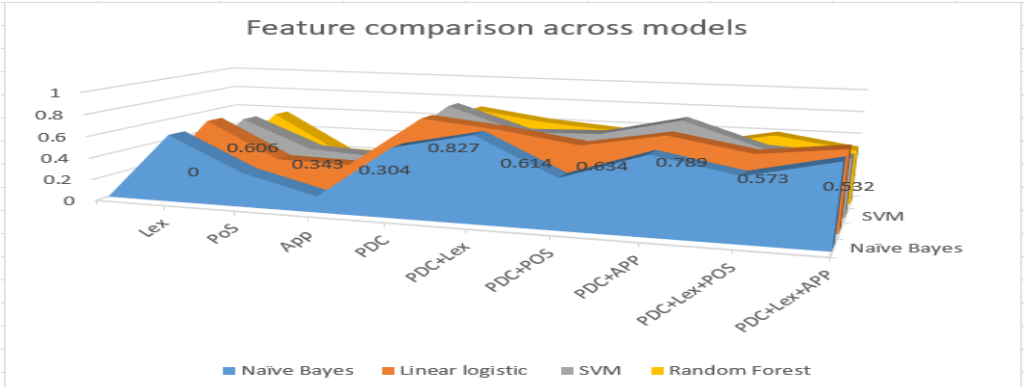

Fig. 5. Feature Comparison across models

Previous research on hate speech identification has relied on lexical and other natural language processing (NLP) elements. However, these kinds of capabilities will fall short of successfully capturing hate speech in codeswitched messages on their own. As a result, classifier models that make explicit use of these traditional features will underperform, producing a high number of false negatives, in contrast to how hate is expressed in social media postings.

Psychosocial features (PDC) from the study, plus other high-level features such as general lexical features, linguistic features, and application-specific features (App), such as tweet length, were utilized in training the classifiers. The performance of these features was compared to that of other conventional text classification algorithms such as Linear Logistic Regression, Naive Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest, and Decision Trees. The models were trained using tenfold cross-validation on a dataset with an 80/20 ratio of training to testing characteristics. The models and feature categories with the highest accuracy performance across the seven machine learning techniques were compared and evaluated using a grid search algorithm. The Support Vector Machine model achieved the best accuracy of 76.2 percent, followed closely by the Linear Logistic Regression model, which achieved 75.8 percent accuracy. The accuracy scores for each model were determined using a tenfold cross-validation procedure and are well represented by the box and whisker plot in Fig. 6.

Given that the major purpose was to identify hate speech, the emphasis moved to the models' performance in identifying hate speech. As a result, just the accuracy performance of the two

promising models was recovered. The experimental outcomes are described in Table 3 based on the balanced dataset.

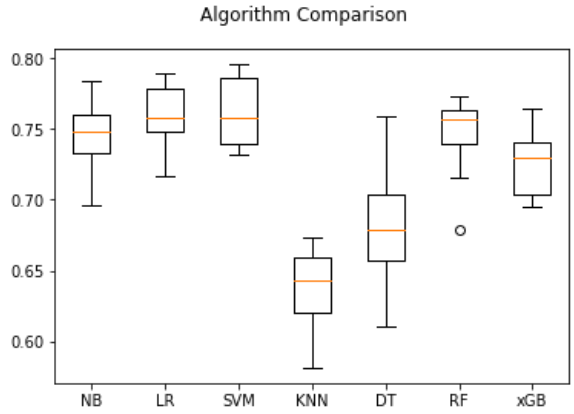

Fig. 6. Accuracy performance comparison of the models

The co-occurrence of psychosocial characteristics, as outlined by the multidimensional framework [68], enables the identification of hate speech in text messages to be robust. This technique overcomes the constraints of lexicon-based solutions, which rely primarily on the capacity to detect hate by identifying domain-specific words within messages, frequently with little regard for hate's syntactic patterns, particularly if codeswitching is used to circumvent the domains keywords.
The discovery of hate speech is dependent on the existence of certain features, which is a typical weakness of dictionary-based algorithms, as the model would not generalize if these features are excluded.
The generic lexicon feature's biggest disadvantage would be its sparse vector representations, which results in large numbers of zeros in the resulting vectors. Therefore, modeling demands more computer resources, particularly memory, which is a difficulty for typical machine learning techniques.
According to the findings, the best accurate features and classification systems to identify hate speech are PDC-based features trained using the linear SVC classifier. Additionally, PDC characteristics had a greater effect on accuracy performance when character-level n-grams were used rather than the word- or phrase-level n-grams. This outcome corroborates another study on hate speech [11].

4.4 Model Evaluation and Tuning

This phase focused on determining whether the classification model performed as expected and on determining how to improve the classifier's performance via parameter tuning.
The SVM was found to be the highest performing model out of the nine studied in the studies, as determined by constructing its confusion matrix and determining its F1, precision, and recall score values. The SVM model performed consistently well in terms of precision, recall, and F1 score, with an average accuracy of 0.77, and an error rate of 0.11. The study was particularly impressed with the model's performance on the hypotheses class, namely hate speech, which had a precision score of 0.81, a recall score of 0.85, and an F1 score of 0.83. The complete findings are given 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 the lower triangle of the matrix had more misclassification than the higher triangle. This suggests that the SVM algorithm was more predisposed than the human coders to label texts as hate speech or offensive.

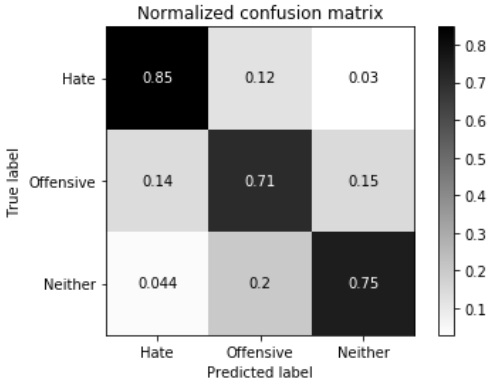

Fig. 7. Confusion matrix for the balanced dataset

According to the first column of the matrix, the model accurately predicted 85 percent (recall) of actual hate speech messages as true positives but misclassified 14% and 4% of offensive and "neither" messages as hate messages, respectively. The misclassification, particularly of hostile communications as false positives, can be explained by the fact that hate speech and offensive messages have some traits. According to the multidimensional framework for hate speech[68], hate speech includes offensive language, although not all offensive statements are inherently hate speech unless they expressly target a protected feature. Additionally, the presence of objectionable lexical phrases would lead the classifier to classify the communication as hate speech, whereas the human annotator would evaluate the context and use hindsight to classify it differently. The misclassification could also be the result of incorrect annotations impacted by the human coder's sensitivity to everyday language use, long-held cultural, religious, and other social belief systems [25, 69].

According to the second column of the matrix, the model correctly predicted 71% of offensive messages, while misclassifying 12% of actual hate speech and 20% of "neither" messages as offensive. This could be explained, once again, by the inherent bias and subjectivity of the human annotator in this work.

The third column of the matrix indicates that the model accurately forecasted 75% of messages as "neither," but misclassified 3% of actual hate speech and 15% of offensive messages as 'neither'.

As a result, the model had the most difficulty in identifying messages classified as offensive, with 12 percent and 20% of real cases of hate speech, respectively, and 'neither' being incorrectly projected as offensive. This, too, might be explained by instructor noise added during annotation as a result of our annotators' differing degrees of sensitivity to what they regarded to be objectionable.

The optimum hyper-parameter settings for the SVM classifier were C=0.1, gamma ()=0.1, and probability=true for the Gaussian Radial Basis Function (RBF) kernel. To create a model that could be applied to a wide range of hate speech, the C value was selected. Since the model required to be trained for more tolerance when setting the decision boundary, machine learning was used to reduce the penalty for model misclassification[69]. The kernel is chosen for SVM assisted in determining the optimum way for the model to build a nonlinear decision boundary based on the features. Adjusting the decision boundary's responsiveness to new features relied on the hyper-parameter gamma (). A larger gamma value indicates that additional features will exert a greater influence on the decision border, flexing it. Lower values for the soft margin and kernel hyperparameters were found to be the most effective for designing the SVM classifier to manage the rather non-linearly separable problem of text data from social media. Additionally, SVM classifiers are quite resilient and produce good predictions when used as models.

4.5 Generalizability of the Classification Model

The subject of model generalizability was central to this research and served as a lens through which all other aims and trials were viewed. Thus, from the start, the study sought a thorough knowledge of the phenomena of hate speech and its conspicuous traits as informed by relevant psychological and sociological theories. This resulted in the development of a multidimensional framework for hate speech, which was utilized to guide the data gathering and annotation operations. Although the data from the 2017 Kenyan presidential elections, which primarily contained ethnic hate speech, was utilized in this study to train the machine classification model, it is not limited to ethnic hate. To begin, the top-performing classifier in terms of generalizability was trained using a balanced dataset using data from the various studies. By and large, a classifier trained using a balanced dataset, with an equal or nearly equal number of instances of each class, will not be skewed toward any particular class, in contrast to a classifier trained using a dataset biased toward the majority class[70]. Second,

our model is based on a multidimensional framework for hate speech that is conceptually universal. As a result, it should generalize to different forms of hate speech and in any language, provided that it is retrained with positive examples of that form of hate speech. As a result, our approach was able to positively identify other forms of hate speech in previously unknown messages, such as, " ***Kill*** *all* ***those Muslims*** *to eradicate terrorism*" (Religious hatred);
*"**Wtf**! Eastleigh explosion. **Wasomali** warudi kwao [Somalis should go back home]"* (Nationality hatred); *"Thot the 'summer break is over? **hawa wazungu** waende zao bana! kazi kutuchafulia ma lightskins wetu nkt eyesore galore" [....these white tourists should leave! They are spoiling our girls and are an eyesore]* (Racial hatred);
*"**Women** are some of the most **corrupt** individuals when placed in positions of power. "* (Gender hatred).
These messages exhibited three major characteristics of hate speech as specified in the conceptual framework for hate speech. These include negative passions such as Kill, Wtf; distancing language through the use of plural pronouns such as that,' *hawa'* (these); and stereotyping by the use of protected characteristics such as Muslims, *Wasomali* (Somalis), *Wazungu* (Whites), and Women. When these characteristics are combined in a single message, the hate speech flag is raised.
Fundamentally, the conceptual framework for hate speech aids in delineating the hypothesis class, $\mathcal{H}$, to which hate speech occurrences can be mapped. As such, the algorithm's task is to identify the precise hypothesis, h ϵ $\mathcal{H}$, that closely reflects hate speech.
The generalizability of the model also addresses the fluid nature of language and how to manage future phrases absent in the training set. The question here is whether the hypothesis holds for previously unknown cases that were not included in the training set, such as instances in the test data provided via cross-validation. This could be fixed by introducing a class *S* that contains the feature *h* equal to *S*. This implies that S must consist entirely of positive examples of hate speech. Conversely, a broad hypothesis, *G*, might be utilized that encompasses all acceptable instances of hate speech while excluding the incorrect ones. As demonstrated in a recent study, the algorithm may be retrained using a *G*-set that includes examples of the new phrases, plus increasing the margin, consequently increasing the distance between the boundary and the nearest occurrences. [71].

5. A CLASSIFICATION MODEL BASED ON PDC FEATURES

This study advocates for a novel text categorization framework that employs a mix of psychosocial features (PDCs) premised on language meant to convey negative passion, psychological distancing, and commitment to hate as the principal characteristics for distinguishing subtle forms of hateful speech. These qualitative characteristics are well documented in theories of hate such as the duplex theory of hate. They provide a rich framework for detecting elusive hateful utterances, particularly when concealed by codeswitching, a phenomenon that was undetectable by previous techniques.
The classification model based on PDCs features employs a supervised learning approach. It is divided into three major components, namely data pre-processing, feature engineering, model construction, and evaluation. These are depicted in Figure 8. The component of data pre-preprocessing is further divided into two subcomponents, namely data annotation, and data preparation. As is customary in supervised machine learning, the input to the PDC-based model is labeled data that can be used to solve binary or multi-class classifications. This means that the data can be annotated using two labels, such as positive or negative, or with multiple labels, such as low, medium, or high respectively. Human annotators frequently label the raw data input according to some annotation strategy. The annotation scheme used in this investigation was based on a strong theoretical foundation, as detailed in our prior study [1], as well as shown in Fig. 8, the leftmost zoomed-out component.
For instance, the multi-dimensional framework based on the PDC can detect the use of devaluation in a codeswitched message such as, "*Take not lightly* ***these cockroaches****, wanaweza kutuibia kura ['they might steal our votes] and oppress* ***us***". In Kenya, certain ethnic devaluation terms are well recognized and frequently used among in-group membership to refer to out-groups. For instance, the term, "foreskins", is frequently used to infer and denigrate the Luo ethnic group, which does not practice circumcision traditionally. The usage of stereotypes is observable in the subtle use of harsh words without the use of overtly nasty lexicons. For instance, the Kikuyu, Kamba, and Kisii ethnic groups are referred to by compound phrases such as, "*money lovers*", "*tire thieves*", or "*night runners*". These nuanced kinds of hate speech, particularly when codeswitching is used, frequently pass unnoticed by conventional filters.
The data preprocessing sub-unit is responsible for tokenizing and cleaning the nosy text. Standard data cleaning procedures are followed, including the elimination of punctuations, stopwords, non-

alphanumeric characters, duplication, lowercasing, empty records, and stemming. Unlike traditional models, which remove pronouns uniformly throughout the preprocessing phase, the PDC-based model maintains the pronouns while deleting other stopwords. This is because the presence of pronoun dichotomies in messages is indicative of the "othering" language[41][58], which in our frameworks maps to the concept of psychological distancing.

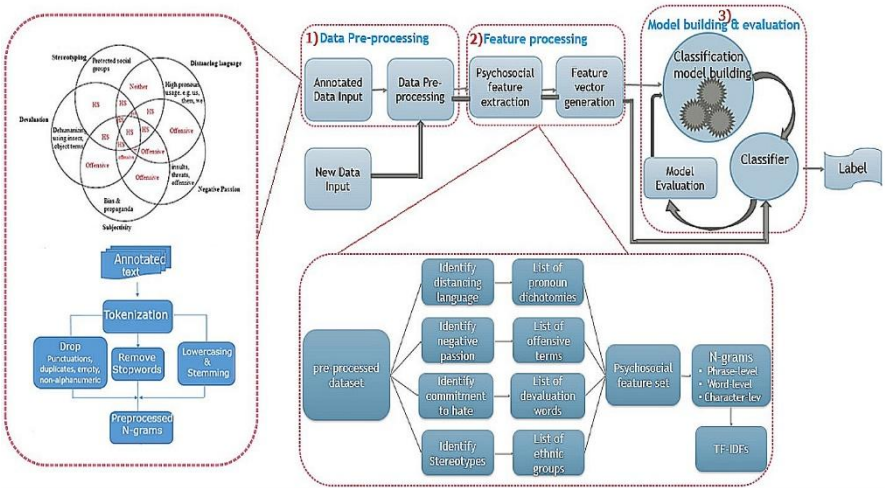

Fig. 8: PDC-Based Text Classifier

The first component produces a clean dataset that has been de-noised and normalized using lowercasing and stemming. As a result, the dimensionality of the input is significantly reduced in contrast to the initial raw annotated input. However, textual data offers a barrier because the major attributes are frequently the words or tokens. As a result, this feature set converts into a high dimensional feature space, with a sparse input vector to the machine learning method in component 3. This results in additional processing effort and memory. This difficulty is addressed by the PDC-based model in component 2, which consists of learning the PDC vocabulary and the feature vector generation subcomponents. To produce their respective lists, subcomponent one filters the pre-processed dataset by extracting words indicative of psychological distancing, hatred commitment, negative passion, and stereotyping. The seed attributes for each list were primarily inspired by qualities that are effective in similar challenges in the literature [4, 15, 51] and by psychological word categories in the Linguistic Inquiry Word Count analyzer [67]. Besides, the respective feature categories were populated with terms that were connected to the classes in the Latent Dirichlet Allocation topic models that were constructed automatically. The five feature categories were structured structurally in a table based on word families, with the first column indicating the word family, successive columns holding the word forms or features, and rows containing the meanings. Bootstrapping enables the addition of new words or terms with similar meanings in other languages, i.e. codeswitched words, to the corresponding feature columns. As described in Table 4, the structural shape is as follows.

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, can then be analyzed at multiple levels, such as word, character, or phrase, and translated into numerical feature vectors, namely, TF-IDFs. As both a feature selection and representation technique, TF-IDF is intended to prioritize tokens based on their relevance across the corpus, penalizing tokens at either extreme, that is, words that are extremely frequent or extremely infrequent across documents, for being deemed unimportant or outliers. Consequently, the TF-IDF feature vector created from the high-level PDC set is dense, rendering it a better input for the classification model building of component 3. A suite of machine learning algorithms is trained on the TF-IDF input vector to construct their respective classifiers, based on their performance in recognizing

hate speech in prior similar studies. Often, it is difficult to predict which machine learning algorithm will be optimal for a classification task in advance. As a result, it is usual practice in machine learning to experiment with multiple algorithms, beginning with the simplest, to choose the optimal one for the particular machine learning problem[70]. The optimal classifier model is reviewed and tested based on its performance and results. There are two methods for evaluating. To begin, the Chi-Square feature scoring method is used to calculate the correlation of the features and the classes, Second, the confusion matrix is utilized to calculate the precision, recall, and, ultimately, the accuracy of the trained model when applied to the testing dataset. Eventually, the pre-processing sub-component is given a fresh text message as input. It is not required to include annotations. It must, however, pass through the feature processing component and be converted to its TF-IDF vector representation. Following that, the vector is passed straight to the classifier, which outputs the projected class label. This is illustrated in Figure 8.

In summary, tests were done to confirm our method of utilizing psychosocial concepts borrowed from established theories of hate in psychology and sociology to construct a novel psychosocial feature set, dubbed PDC. Following that, the PDC feature set was converted to TF-IDF vectors to train a classifier for detecting nuanced kinds of hate speech, particularly in codeswitched text documents. The classifier performed significantly better than the baseline, in our case, the human inter-rater reliability score, by over 27% in classification accuracy. The classifier's generalization was subsequently evaluated using unseen data comprising of religious, racist, and nationality-based abusive comments. The results were comparable to those obtained using state-of-the-art baseline classifiers for the classification of similar types of hate speech. Conversely, given the variety of datasets used in this study, particularly the emphasis on codeswitched data, it would be unrealistic to directly compare to publically available monolingual datasets. Additionally, the study indicated the psychosocial characteristics' capacity to generalize to other forms of hate speech, such as racist insults.

6. CONCLUSION

This study provided a gold-standard annotated dataset that can be utilized by other studies for comparative experiments. It established a multidimensional analytical framework and methodology for identifying subtle forms of hate speech. Consequently, the classifier's outputs could be utilized to influence evidence-based judgments by national security agencies.

Psychosocial (PDC) concepts are intended to be useful in improving classifier performance. The PDC feature set is significantly less than that of standard approaches that use the TF-IDF to represent the complete input lexicon. This drastically lowers the sparseness and dimensionality of the original features.

The PDC features are the most discriminative in classifying hate speech in codeswitched text messages. The best performance was obtained with n=3 to 5 characters, respectively. This could be explained by the great degree of language independence of character n-gram features.

Future work will consider moving away from the current discrete representation of PDC features in which words exist as atomic symbols and toward a distributed representation in which dense vectors could be used to represent word families to accommodate synonyms, hypernyms, and codeswitching in their context words.

ACKNOWLEDGMENTS

This research would not have been possible without financing from the Kenyan Education Network (KeNet) and the team of annotators from Africa Nazarene University.

COMPETING INTERESTS

The authors declare that there are no existing competing interests.

AUTHORS' CONTRIBUTIONS

Author 1 designed the study, performed the machine learning experiments, and wrote the first draft of the manuscript. . Author 2 and Author 3 analyzed the experimental data and proofread the draft. All authors read and approved the final manuscript.

REFERENCES

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

[2] S. Benesch, “Dangerous Speech: A Proposal to Prevent Group Violence,” 2012.

[3] 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.

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

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

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

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] G. Priya, K. Aditi, T. Richa, A. Mayank, B. Sohail, and J. Vishal, “A Proposed Framework to Analyze Abusive Tweets on the Social Networks,” *Int. J. Mod. Educ. Comput. Sci.*, vol. 10, no. 1, pp. 46–56, 2018.

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

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

[15] 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.

[16] E. Spertus, “Smokey: Automatic recognition of hostile Messages,” in *IAAI*, 1997.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “AutomatedHateSpeechDetectionandtheProblemofOffensiveLanguage,” in *ICWSM*, 2017.

[25] 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.

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

[27] 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.
[28] C. Nobata, J. Tetreault, A. Thomas, Y. Mehrdad, and Y. Chang, “Abusive Language Detection in Online User Content,” in *25th International Conference on World Wide Web*, 2016, pp. 145–153.
[29] P. . O’Sullivan and A. . Flanagin, “Reconceptualizing ‘flaming’ and other problematic messages,” *New Media Soc.*, vol. 5, pp. 69–94, 2003.
[30] 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.
[31] A. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, “Offensive Language Detection Using Multi-level Classification,” *Springer*, p. 1627, 2010.
[32] I. Chaudhry, “Hashtagging hate: Using Twitter to track racism online,” *First Monday 20(2)*, 2015. .
[33] 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.
[34] 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.
[35] M. Hasanuzzaman, G. Dias, and A. Way, “DemographicWordEmbeddingsforRacismDetectiononTwitter,” in *Proceedings of The 8th International Joint Conference on Natural Language Processing,* 2017, pp. 926–936.
[36] 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.
[37] 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].
[38] Z. Mossie and J.-H. Wang, “SOCIAL NETWORK HATE SPEECH DETECTION FOR AMHARIC LANGUAGE,” in *COMIT*, 2018, pp. 41–55.
[39] 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.
[40] 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.
[41] 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.
[42] K. N. B. of Statistics, “2019 Kenya Population and Housing Census Volume I: Population by County and Sub-County,” 2019.
[43] H. Kim, S. Jang, Mo, S.-H. Kim, and A. Wan, “Evaluating Sampling Methods for Content Analysis of Twitter Data,” *Sage*, 2018.
[44] 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.
[45] P. . Cavazos-Rehg *et al.*, “A content analysis of depression-related tweets,” *Comput. Hum. Behav.*, vol. 54, pp. 351–357, 2016.
[46] C. Shearer, *The CRISP-DM model: the new blueprint for data mining*. 2000.
[47] 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.
[48] 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.
[49] W. Warner and J. Hirschberg, “Detecting Hate Speech on the World Wide Web,” in *Language in Social Media (LSM 2012)*, 2012.
[50] P. Burnap and M. L. Williams, “Us and them: identifying cyber hate on Twitter across multiple protected characteristics.,” *EPJ Data Sci.*, 2016.
[51] 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.
[52] J. Brownlee, *Deep Learning for Natural Language Processing*, V1.2. 2018.
[53] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
[54] 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.
[55] 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.
[56] N. Coupland, "'Other' representation, Society and Language." John Benjamins Publishing, 2010.
[57] G. R. Semin, "Linguistic Markers of Social Distance and Proximity." 2009.
[58] 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.
[59] N. Haslam, "Dehumanization: An integrative review," *Personal. Soc. Psychol. Rev.*, vol. 10, pp. 252–64, 2006.
[60] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[61] K. Krippendorff, "Computing Krippendorff's Alpha-Reliability," *University of Pennsylvania ScholarlyCommons*, 2011. [Online]. Available: mhttp://repository.upenn.edu/asc_papers/43.
[62] 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.
[63] V. Dijk and A. Teun, "Discourse and racism, The Blackwell companion to racial and ethnic studies," pp. 145–159, 2002.
[64] 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.
[65] 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.
[66] 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.
[67] 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.
[68] E. Ombui, L. Muchemi, and P. Wagacha, "Building and Annotating a Codeswitched Hate Speech Corpora," *Int. J. Inf. Technol. Comput. Sci.*, no. 3, pp. 33–52, 2021.
[69] 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].
[70] J. Brownlee, *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*. 2016.
[71] E. Alpaydin, *Introduction to Machine Learning*, 2nd Editio. London: The MIT Press, 2010.
[72] 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.
[73] 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.