



**UNIVERSITY OF NAIROBI**

SCHOOL OF COMPUTING AND INFORMATICS

**A Model for Classifying Hate Speech Text  
from Social Media Leveraging on Psycho-  
social Features and Machine Learning**

BY

**Ombui Edward Osoro**

**Reg. No P80/92844/2013**

**Supervisors:**

- 1. Dr. Lawrence Muchemi**
- 2. Prof. Peter Waiganjo Wagacha**

---

*Thesis Presented for the Award of Degree of Doctor of Philosophy in Computer Science*

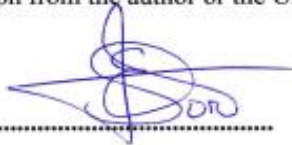
*School of Computing and Informatics*

*University of Nairobi, Kenya*

© 2020

**DECLARATION**

I hereby declare that this dissertation is my original work and where there are contributions or work from other individuals, they have been duly acknowledged. I confirm that this dissertation has not been submitted to any other institution of higher learning for the award of a degree. Therefore, the work presented in this dissertation, whether in part of full, may only be reproduced with permission from the author or the University of Nairobi.

**Signature:**  .....

**Date:** 20<sup>th</sup> Nov. 2020 .....

**Author:** Edward Oso Ombui

**Registration Number:** P80/92844/2013

School of Computing and Informatics

University of Nairobi, Kenya

This dissertation has been submitted for examination towards the fulfillment of requirements for the Doctor of Philosophy in Computer Science at the University of Nairobi with my approval as the University supervisor.

**Supervisor:**

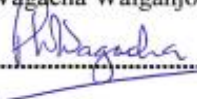
Dr. Lawrence Muchemi

**Signature:**  .....

**Date:** 20<sup>th</sup> Nov 2020 .....

**Supervisor:**

Prof. Peter Wagacha Waiganjo

**Signature:**  .....

**Date:** 20<sup>th</sup> Nov. 2020 .....

## ABSTRACT

Classifying brief text messages containing hate speech from the massive amount of content generated by social media users is a difficult undertaking. Social media data provides significant difficulties for conventional natural language processing approaches when it comes to obtaining high-quality features from noisy, highly dimensional, codeswitched, and large unstructured data. Additionally, a detailed assessment of past studies revealed a dearth of publicly available annotated datasets for comparative studies, a deficit of theoretical support for the annotation systems employed, and a scarcity of research on codeswitched data.

To overcome these shortcomings, this study takes a data-driven strategy to find qualitative and discriminatory characteristics in hate text messages from social media platforms. The objective is to use these attributes to construct a more effective machine classification model for detecting subtle hate speech text messages. Approximately 400k messages were crawled from social media during the 2017 Kenyan general election period, employing a combination of problematic hashtags, ethnic epithets, hate patterns, and messages from pro-hate user accounts. A random sample of 50k messages was manually classified by a team of 27 human annotators into three categories: Hate Speech, Offensive, or Neither. Subsequently, this dataset was condensed further by utilizing a hierarchical probability modeling technique to derive a psychosocial feature subset (PDC) informed by the conceptual framework. To analyze and select the best model, a grid search was conducted through all possible feature combinations using 5-fold cross-validation, with a tenth of the data set reserved for evaluation and to avoid over-fitting the model. According to the findings of the studies, the unique psychosocial feature set (PDC) was effective at identifying hate speech and outperformed traditional features when used to train the best classifier, namely the linear support vector machine algorithm, with an accuracy of 82.5 percent. The Passion (P) and Distance (D) factors were found to be most significant, with 74.3 percent and 74.2 percent accuracy, respectively. Further, the psychosocial feature framework generalized better than conventional features and classifiers in handling additional types of hate speech in codeswitched text messages.

This study makes three contributions. First, it provides a gold-standard annotated dataset that may be used for comparative studies by other researchers. Second, the study provided an empirical framework and methodology for identifying hate speech in short text messages that are anchored in theory. Thirdly, this approach was important in the development of a text classification model capable of effectively generalizing to various forms of hate speech on social media. Subsequently, the classifier's outputs could be utilized to influence evidence-based judgments by relevant security authorities and data-driven policy formation addressing the monitoring of hate speech on social media during future presidential elections in Kenya.

**Keywords:** *Hate Speech, Psychosocial features, Dimensionality reduction, Supervised learning, Codeswitching*

## ACKNOWLEDGMENT

My heartfelt thanks to my academic supervisors, Dr. Lawrence Muchemi and Prof. Peter Waiganjo Wagacha, who also served as academic advisors for my undergraduate and MSc studies, respectively. I consider myself fortunate to have worked with you over the years as you shaped my thinking on academic success, research, resilience, and life-long learning. I am grateful for your constant encouragement to complete this academic journey.

Additionally, I would like to express my gratitude to the Kenya Education Network for partially funding this project through the 2018 Computer Science/Information Systems Mini-Grants for Big Data. Likewise, I would like to express my heartfelt gratitude to my research team, Dr. Amos Gichamba, and Mr. Moses Karani, for their efforts that resulted in the successful completion of the study within the allotted time frame. I cannot fail to express my gratitude to my colleagues and students at Africa Nazarene University who assisted with the data annotation process.

Finally, I would like to express my gratitude to the leadership of the School of Computing and Informatics at the University of Nairobi for providing me with an excellent opportunity as a Ph.D. student researching big data to pursue a three-month exchange program at the University of Stavanger in 2019 through the Kenya-Norway Mobility program (KeNoMo). The research atmosphere at UiS was conducive and served as a significant motivation for completing the study's final mile, which included fine-tuning the classifiers, publishing the experiments, and preparing the thesis.

## DEDICATION

To God be the glory for the wonderful things He has accomplished. This endeavor has been completed via His grace and mercies. May this PhD be used in perpetuity to magnify and honor your holy name throughout your Kingdom.

Special thanks to my late father, who pushed me and longed to see me finish this thesis. May this day in paradise bring you a bigger smile.

To my darling wife, Julz, for partnering with me as my strongest supporter, ensuring that my thesis deadlines were met. To our gorgeous children Bree, Jojo, Kat, and Faus: may you follow in your father's footsteps and go further.

# Table of Contents

|  |           |
|--|-----------|
| ABSTRACT .....   | II        |
| ACKNOWLEDGMENT .....                                       | III       |
| DEDICATION .....   | IV        |
| LIST OF FIGURES .....                                      | VIII      |
| LIST OF TABLES .....                                       | X         |
| DEFINITION OF TERMINOLOGIES .....                          | XI        |
| LIST OF ABBREVIATIONS .....                                | XII       |
| <b>CHAPTER 1: INTRODUCTION .....</b>                       | <b>1</b>  |
| 1.1 BACKGROUND .....                                       | 1         |
| 1.2 PROBLEM STATEMENT .....                                | 4         |
| 1.3 OBJECTIVES OF THE STUDY .....                          | 5         |
| 1.3.1 General Objective .....                              | 6         |
| 1.3.2 Specific Objectives .....                            | 6         |
| 1.4 RESEARCH QUESTIONS .....                               | 6         |
| 1.5 RESEARCH SIGNIFICANCE .....                            | 7         |
| 1.6 RESEARCH JUSTIFICATION .....                           | 7         |
| 1.7 SCOPE OF THE STUDY .....                               | 8         |
| 1.8 STUDY ASSUMPTIONS .....                                | 8         |
| <b>CHAPTER 2: LITERATURE REVIEW .....</b>                  | <b>10</b> |
| 2.1 HATE SPEECH .....                                      | 10        |
| 2.1.1 Defining Hate Speech .....                           | 10        |
| 2.1.2 Operational Definition for Hate Speech .....         | 14        |
| 2.1.3 Dangers of Hate Speech .....                         | 15        |
| 2.1.4 Hate Speech in Kenya .....                           | 16        |
| 2.1.5 Hate Speech Laws in Kenya .....                      | 16        |
| 2.1.6 Hate Speech on Social Media Platforms in Kenya ..... | 17        |
| 2.2 THEORIES OF HATE IN SOCIAL PSYCHOLOGY .....            | 19        |
| 2.3 TEXT CLASSIFICATION .....                              | 23        |
| 2.3.1 Machine Learning .....                               | 23        |
| 2.3.1.1 Supervised Machine learning .....                  | 25        |
| 2.3.1.2 Unsupervised Machine learning .....                | 26        |
| 2.3.1.3 Semi-supervised Machine learning .....             | 26        |
| 2.3.1.4 Reinforcement learning .....                       | 27        |
| 2.3.2 Types of Machine Learning Algorithms .....           | 28        |
| 2.3.2.1 Linear machine learning .....                      | 28        |
| 2.3.2.2 Non-Linear Models .....                            | 30        |
| 2.3.2.2.1 Rule-based .....                                 | 30        |
| 2.3.2.2.2 Distance-based Algorithms .....                  | 31        |
| 2.3.2.2.3 Probabilistic Algorithms .....                   | 32        |
| 2.3.3 Deep Learning Algorithms .....                       | 35        |
| 2.4 FEATURES IN AUTOMATIC TEXT CLASSIFICATION .....        | 36        |
| 2.4.1 Dimensionality Reduction .....                       | 36        |

|  |           |
|--|-----------|
| 2.4.2 Feature Generation .....   | 37        |
| 2.5 METHODOLOGIES AND FEATURES USED IN PREVIOUS HATE SPEECH STUDIES .....      | 40        |
| 2.5.1 The Cross-Industry Standard Methodology for Data Mining (CRISP-DM) ..... | 46        |
| 2.6 HIGH-LEVEL FEATURES FOR TEXT CLASSIFICATION .....                          | 47        |
| 2.6.1 Psychosocial features .....  | 47        |
| 2.6.2 Linguistic features .....  | 50        |
| 2.6.3 Other Features .....   | 52        |
| 2.7 LOW-LEVEL FEATURES FOR TEXT CLASSIFICATION .....                           | 54        |
| 2.8 RESEARCH FRAMEWORK .....   | 56        |
| 2.8.1 Theoretical Framework .....  | 56        |
| 2.8.2 Conceptual Framework .....   | 58        |
| 2.8.3 Measurement using term frequency-inverse document frequency .....        | 61        |
| 2.9 SUMMARY OF FEATURES IN TEXT CLASSIFICATION .....                           | 62        |
| 2.10 SUMMARY .....   | 64        |
| <b>CHAPTER 3: RESEARCH METHODOLOGY .....</b>                                   | <b>65</b> |
| 3.1 RESEARCH METHODOLOGY .....   | 65        |
| 3.1.1 Mixed Methods Research Methodology .....                                 | 65        |
| 3.2 RESEARCH PHILOSOPHY .....  | 66        |
| 3.3 RESEARCH DESIGN .....  | 69        |
| 3.4 RESEARCH METHOD .....  | 71        |
| 3.4.1 Problem understanding .....  | 72        |
| 3.4.2 Data Understanding .....   | 74        |
| 3.4.3 Data Preparation .....   | 76        |
| 3.4.4 Feature selection and extraction .....                                   | 80        |
| 3.4.5 Modelling .....  | 83        |
| 3.4.6 Evaluation .....   | 85        |
| 3.4.7 Deployment .....   | 87        |
| 3.5 ENSURING VALIDITY AND RELIABILITY .....                                    | 88        |
| 3.6 ETHICAL CONSIDERATIONS .....   | 89        |
| 3.7 SUMMARY .....  | 90        |
| <b>CHAPTER 4: RESULTS AND FINDINGS .....</b>                                   | <b>91</b> |
| 4.1 PROBLEM UNDERSTANDING .....  | 91        |
| 4.1.1 Findings .....   | 93        |
| 4.2 DATA COLLECTION .....  | 94        |
| 4.3 DATA ANNOTATION .....  | 94        |
| 4.4 DATA UNDERSTANDING .....   | 97        |
| 4.4.1 Learning a class from examples .....                                     | 102       |
| 4.4.2 Probabilistic Hierarchical Modelling of Hate Speech .....                | 104       |
| 4.5 MODEL TRAINING AND EVALUATION .....  | 106       |
| 4.5.1 Experimental Results and Findings .....                                  | 107       |
| 4.5.2 Model Parameter Tuning .....   | 112       |
| 4.5.3 Evaluation of the Classification Models .....                            | 114       |

|   |            |
|---|------------|
| <b>CHAPTER 5: DISCUSSION AND CONCLUSION.....</b>                  | <b>90</b>  |
| 5.1 RESEARCH LIMITATIONS .....                                    | 90         |
| 5.2 DISCUSSION .....  | 91         |
| 5.2.1 <i>Developing a deep understanding of hate speech</i> ..... | 91         |
| 5.2.2 <i>Building a Hate Speech Conceptual Framework</i> .....    | 92         |
| 5.2.3 <i>Hate Speech Dataset from Social Media in Kenya</i> ..... | 93         |
| 5.2.4 <i>Training a Hate Speech Classification Model</i> .....    | 96         |
| 5.2.5 <i>Generalizability of the Classification Model</i> .....   | 103        |
| 5.3 PDC-BASED CLASSIFICATION MODEL .....                          | 107        |
| 5.4 CONCLUSION .....  | 111        |
| 5.5 FUTURE RESEARCH RECOMMENDATION .....                          | 112        |
| 5.6 THE RESEARCH CONTRIBUTIONS .....                              | 113        |
| <b>REFERENCES .....</b>   | <b>118</b> |
| <b>APPENDICES.....</b>  | <b>128</b> |
| APPENDIX A: LIST OF PUBLICATIONS .....                            | 128        |
| APPENDIX B: RESEARCH BUDGET .....                                 | 129        |
| APPENDIX C: ANNOTATION SCHEME.....                                | 130        |
| APPENDIX D: ANNOTATION PORTAL .....                               | 132        |
| APPENDIX E: HATE SPEECH CLASSIFIER PORTAL .....                   | 133        |
| APPENDIX F: TURNITIN REPORT.....                                  | 133        |



## LIST OF FIGURES

|   |    |
|---|----|
| Figure 2.1: Triangular Structure of Hate .....                        | 20 |
| Figure 2.2: Machine learning Function .....                           | 24 |
| Figure 2.3: Supervised learning.....                                  | 26 |
| Figure 2.4: Unsupervised Learning .....                               | 26 |
| Figure 2.5: Semi-supervised Learning.....                             | 27 |
| Figure 2.6: Reinforcement Learning .....                              | 27 |
| Figure 2.7: Machine learning Methods.....                             | 28 |
| Figure 2.8: Binary Classification adapter from [80].....              | 29 |
| Figure 2.9: Shallow learning.....                                     | 35 |
| Figure 2.10: Deep learning.....                                       | 35 |
| Figure 2.11: The curse of dimensionality (Adopted from [87]) .....    | 37 |
| Figure 2.12: The Supervised Machine learning framework.....           | 40 |
| Figure 2.13: Feature frequency from previous hate speech studies..... | 45 |
| Figure 2.14: Four-level breakdown of CRISP-DM methodology [119] ..... | 46 |
| Figure 2.15: High-level feature categories.....                       | 47 |
| Figure 2.16: Psychosocial features.....                               | 49 |
| Figure 2.17: Linguistic features.....                                 | 50 |
| Figure 2.18: Lexical Features.....                                    | 51 |
| Figure 2.19: Syntactic features .....                                 | 51 |
| Figure 2.20: Stylistic features.....                                  | 51 |
| Figure 2.21: Semantic Features .....                                  | 52 |
| Figure 2.22: Knowledge-based Features .....                           | 52 |
| Figure 2.23: Other features.....                                      | 53 |
| Figure 2.24: Combined High-level feature framework.....               | 53 |
| Figure 2.25: Low-level features .....                                 | 54 |
| Figure 2.26: Multidimensional Hate Speech Conceptual Framework.....   | 60 |
| Figure 2.27: Hierarchical Feature framework.....                      | 63 |
| Figure 3.1: Research workflow .....                                   | 72 |
| Figure 3.2: Data collection flowchart.....                            | 75 |
| Figure 3.3: The annotation portal.....                                | 88 |
| Figure 3.4: Evaluation Process Using Model Accuracy Estimation .....  | 86 |
| Figure 4.1: Frequency of verbs used in hate speech definitions .....  | 93 |
| Figure 4.2: Hate-specific content frequency .....                     | 93 |
| Figure 4.3: Percentage of annotated tweets .....                      | 95 |
| Figure 4.4: Rating on hate speech tweets .....                        | 95 |
| Figure 4.5: Types of Hate speech.....                                 | 95 |
| Figure 4.6: Hate speech features .....                                | 97 |
| Figure 4.7: Message length.....                                       | 98 |
| Figure 4.8: Class distribution.....                                   | 98 |
| Figure 4.9: Word frequency histogram.....                             | 99 |

|  |     |
|--|-----|
| Figure 4.10: Hate Speech histogram.....  | 99  |
| Figure 4.11: General Word frequency word cloud .....                           | 100 |
| Figure 4.12: Word frequency under the Hate Speech class .....                  | 100 |
| Figure 4.13: Correlation of terms to classes using chi-square.....             | 103 |
| Figure 4.14: End-to-End Pipeline for the Hate Speech Classification Model..... | 107 |
| Figure 4.15: PDC feature mapping to low-level features .....                   | 110 |
| Figure 4.16: Model Parameter Grid .....  | 112 |
| Figure 4.17: Confusion Matrix Based on 3 Classes .....                         | 115 |
| Figure 4.18: Confusion Matrix Based on 2 Classes .....                         | 115 |
| Figure 4.19: Confusion matrix for the balanced dataset .....                   | 89  |
| Figure 5.1: Conceptual framework of hate speech’s multidimensionality.....     | 101 |
| Figure 5.2: The covariance of n-grams to classes using chi-square.....         | 103 |
| Figure 5.3: The PDC-Based text classification framework .....                  | 108 |

## LIST OF TABLES

|  |     |
|--|-----|
| Table 2.1: Hate Speech Definitions.....  | 13  |
| Table 2.2: Theories concerning hate speech .....                                   | 21  |
| Table 2.3: Features versus training samples.....                                   | 37  |
| Table 2.4: A Summary of the reviewed hate speech studies and similar studies ..... | 44  |
| Table 2.5: A summary of high-level features used in previous studies .....         | 49  |
| Table 2.6: Other features used in previous studies.....                            | 56  |
| Table 2.7: Constructs from qualitative research on high-level features .....       | 57  |
| Table 2.8: Theoretical framework.....  | 57  |
| Table 2.9: The summary of the concepts.....  | 60  |
| Table 2.10: Multidimensionality of Hate Speech.....                                | 61  |
| Table 3.1: Example of Confusion Matrix.....  | 87  |
| Table 3.2: Summary of the research methodology .....                               | 90  |
| Table 4.1: Content analysis of hate speech definitions .....                       | 92  |
| Table 4.2: Raw Dataset Description.....  | 94  |
| Table 4.3: Preliminary annotations.....  | 95  |
| Table 4.4: Class distribution.....   | 98  |
| Table 4.5: Topic modeling for hate speech class .....                              | 105 |
| Table 4.6: Naïve Bayes classifier performance .....                                | 108 |
| Table 4.7: Linear Support Classifier performance .....                             | 108 |
| Table 4.8: Logistic Regression classifier performance .....                        | 109 |
| Table 4.9: Random Forest classifier performance .....                              | 109 |
| Table 4.10: Feature performance across nine classification models .....            | 111 |
| Table 4.11: Naïve Bayes with Grid Search .....                                     | 112 |
| Table 4.12:Support Vector Classifier with Grid Search.....                         | 113 |
| Table 4.13: Logistic Regression with Grid Search .....                             | 113 |
| Table 4.14:Random Forest with Grid Search .....                                    | 114 |
| Table 4.15: Performance on the balanced dataset .....                              | 88  |
| Table 5.1: Feature comparison across machine classifiers.....                      | 98  |

## DEFINITION OF TERMINOLOGIES

**Feature:** A unique, measurable property of text to experiment with.

**Label:** the target of a machine learning predictive model, aka the dependent variable.

**Binary Classification:** Classification that results in 2 predictable outcomes. E.g., Hate speech / not hate speech)

**Classifier:** An algorithm that maps the input data to a specific category or class

**Codeswitching:** alternation of words from two or more languages in the same message

**Corpus:** a collection of documents, e.g., a whole dataset of tweets

**Document:** a single entity or row in a dataset, e.g., a text message, a tweet.

**Ethnocentrism:** the attitude of prejudice or mistrust by an in-group member(s) towards out-group member(s) of a social group.

**Latent variable:** this is a variable that cannot be directly observed but gets measured by other variables that can be observed or measured.

**Learner:** The process that generates the classifier

**Multi-class classification:** Classification that results in more than two predictable outcomes. E.g., Hate speech, Offensive speech, Neither speech

**Noise:** any unwanted anomaly in the data that negatively impacts the learning of the model

**N-Gram:** a continuous sequence of n items from a given sample of text

**Observations:** these are the rows, records, samples, or instances in a dataset.

**Psychosocial:** the relationship between an individual's thoughts and expressions towards others in a social setting.

**Token:** a word, phrase, or symbol derived from a document through the process of tokenization.

**Tweets:** Short text messages posted on Twitter social media

**Tweet ID:** A unique identification number generated for each Tweet

**A Model:** A program that is automatically learned by the machine learning algorithm

## LIST OF ABBREVIATIONS

**BOW:** Bag of Words

**CGI:** Common Gateway Interface

**CNN:** Convolutional Neural Networks

**DT:** Decision Tree

**HAN:** Hierarchical Attention Network

**HS:** Hate Speech

**IDF:** Inverse Document Frequency

**KNN:** K-Nearest Neighbor

**LDA:** Latent Dirichlet allocation.

**LIWC:** Linguistic Inquiry and Word Count text analyzer.

**LLR:** Linear Logistic Regression,

**NB:** Naïve Bayes

**NCIC:** National Cohesion and Integration Commission

**NLTK:** Natural Language Tool Kit

**PDC:** Negative **P**assion, **D**istancing, and **C**ommitment to hate

**POS:** Part-of-Speech

**RF:** Random Forest

**SVM:** Support Vector Machine

**TF:** Term Frequency

**TF-IDF:** Term Frequency- Inverse Document Frequency

**WTF:** What the F\*#k

**XGB:** Extreme Gradient Boosting

## CHAPTER 1: INTRODUCTION

### 1.1 Background

This section discusses the background of the study, clarifies the problem statement, defines the main goal and the specific objectives of the study, describes the study's significance, and presents the research questions. The chapter concludes by outlining the significance, the justification, scope, and assumptions of the study, and provides an overview of the thesis organization.

The rise of hate speech on social media platforms has proven to be a difficult and intractable issue for some African governments in recent years, which have resolved to employ hard force to restrict it, particularly during electioneering periods. For example, the Democratic Republic of Congo shut down all social media immediately following the presidential election in 2018 [1]; Ethiopia shut down its Internet in 2016 in response to growing protests [2]; Uganda shut down social media during the 2016 presidential vote-counting [3]; and Kenya threatened to jail administrators of hate WhatsApp groups and hired human monitors to monitor social media ahead of the 2017 presidential elections [4]. Additionally, social media network businesses are under increasing pressure to improve their response to the spread of hate speech on their platforms, which have evolved into billboards for hate speech [5]. Thus far, social media companies have responded by constantly revising their hate-speech regulations to encompass areas of user material that previously provided loopholes for spewing hate speech. Twitter, for example, changed its hate speech policies to prohibit dehumanizing rhetoric and scaremongering stereotypes directed at certain groups [6].

While the advancement of the Internet and social media networks is lauded for providing a new avenue for people to publicly express their opinions, thoughts, and feelings [7],[8], these have, paradoxically, accelerated the production of online hate speech content, making manual monitoring impractical [10]. Today, the likelihood of an Internet user encountering instances of hate speech is increasing, particularly on blogs, newsgroup comments, online games, social media networks, and other interactive online public platforms [9], [10]. Additionally, written text has been demonstrated to be more persistent in its form than other media such as spoken speech [11]. Written content is easily spread to a larger audience and can swiftly escalate to offline social disorder, harm, and undiscovered difficulties outside of public online venues [12].

Governments, non-governmental organizations, and civil society organizations have increased their pressure on social media companies such as Facebook and Twitter to improve their moderation processes for policing hate speech content on their platforms. At the moment, the two social media platforms rely on users to report hostile content by clicking the report button located next to the objectionable content. Subsequently, content moderators review these instances and flag, hide, or remove them from the chat thread if they are found to violate the user agreement regulations and terms provided on their platforms [13], [14]. Given the rapidity with which users flag content on social media, in comparison to the over 2000 human languages utilized on these platforms, and the limited number of content moderators, it is almost difficult to review and flag every instance of reported hate speech.

Nonetheless, these platforms have evolved into a vehicle for the rapid and affordable dissemination of hate speech, including racism, derogatory ethnicity, religious attacks, insults, and sexist statements. According to a study conducted by a cloud-based web filtering and scanning service, 80 percent of blogs include inappropriate content[15]. This is a concerning report that should drive organizations and enterprises with an online presence to ensure that the material on their websites is monitored closely, lest they lose their online consumers' trust. Unfortunately, this is a highly difficult and often overwhelming task for a human being to do, all the more so considering the avalanche of hate speech that frequently ensues online following a trigger event [16]. In Kenya, hate speech on social media is particularly widespread during national election campaign seasons, particularly during presidential elections. Throughout this time period, there have been a growing number of campaign-related incidents across the country that have elicited online public reactions bordering on hate speech. Among the most egregious of these are politicians' invocations of negative ethnic feelings [17], which frequently elicit strong public reactions and counter-reactions from Kenyans on social media platforms. The situation is aggravated by Kenya's absence of explicit legislation holding media corporations, particularly social media platforms, accountable for hate speech spread on their platforms. Rather than that, the regulations available at the time of this investigation addressed individual users, with the bracket expanded to include local administrators of network groups on social media platforms such as WhatsApp [18].

In social discussions, codeswitching is a common social phenomena that is highly suggestive of group membership [19]. While it is considered informal communication, it is gradually becoming the rule rather than the exception in everyday contact amongst bilingual and multilingual

populations, particularly on social media. Additionally, codeswitching on social media in regards to hate speech appears to be the de facto lingua franca for in-group membership. It is interpreted as enhancing cohesion in communicating with "our" people and separating oneself from others, particularly those perceived to be critical adversaries. Additionally, code swapping is frequently used to emphasize an idea or item in communication. Given that some social media platforms allow users to interact anonymously, these platforms constitute fertile ground for the proliferation of hate speech.

User-generated content on social media poses a huge challenge to standard methodologies and applications in natural language processing, computational linguistics, and machine learning. It is noisy, irregular, replete with duplicate and missing values, massive, diverse in data kinds, generated in real time, and subject to all of the other issues associated with big data. When parsing sentences and performing contextual analysis on words and phrases using typical monolingual techniques, codeswitching creates a barrier. The scarcity of native language resources, such as corpora, parts-of-speech taggers, and dictionaries [20], in combination with unrecorded grammatical rules and disorganized research networks, appears to worsen the problem [21].

There is an increasing amount of research being conducted in the domain of hate speech, including automated approaches for detecting hate speech [22], [23], [24], as well as other related concepts such as offensive language detection [25], [26], cyberbullying [27], [28], radicalization and terrorism [29], [30].

The majority of these research have focused on social media platforms, specifically Twitter and Facebook, because they make it easier and faster to access and collect user-generated data via their respective Application Programming Interfaces (APIs). Notably, these research generate their own datasets, and with the exception of Waseem and Hovy [10], few have specified the annotation criteria or theoretical underpinnings of the classification. Additionally, low-reliability ratings have been shown to have an effect on the quality of dataset annotation and, consequently, on the training and performance of classification models [31], [32]. In this regard, this research seeks to address this highlighted gap by proposing a theoretical framework for classifying tweets into three categories, namely those including hate speech, those containing merely offensive language, and those containing neither. As a result, training datasets will be of greater quality, and predicted accuracy in machine classification will improve. The triangle of hatred theory[33] denotes three



distinct components of hatred. These are the concepts of distance, passion, and choice/commitment. This theory is particularly useful in laying the groundwork for the three factors espoused in the annotation model, namely Distance, which is indicated by elements of exclusion or "otherness" [34], Passion, which is indicated by negative emotions and the use of derogatory language [35], and Decision/commitment, which is indicated by propaganda elements and target devaluation [36].

The qualitative research design is employed in this work to comprehend the phenomena of hate speech and its most descriptive aspects in order to successfully identify hate speech in brief text messages. This is based on a case study of hate speech on Kenyan social media during the campaign period leading up to the August general elections and the November 2017 repeat elections. Twitter was chosen as the social media network for this study due to its representativeness, scale, and accessibility of public access to tweets. A systematic evaluation of existing hate theories was done in order to define the proposed framework's thematic variables and to find the best discriminating indicators of hate speech. The reliability score assigned to human annotators was critical in evaluating the model's performance as a baseline for the same dataset.

This study makes two contributions: it creates a publicly available dataset that can be used for comparative studies by other researchers, and more importantly, it develops an empirical framework and methodology for developing a novel psychosocial feature set to aid in the machine classification of subtle forms of hate speech in short text messages. Additionally, unlike the majority of previous studies, in which the data language is monolingual and the context is not African, this study focuses on the case of Kenya, which has a growing population of multilingual social media users, as evidenced by codeswitched user content on these online public platforms.

## **1.2 Problem Statement**

Humans are capable of readily identifying related concepts and understanding meaning from text. However, automating these duties with computer software is rather tough, owing to the unstructured and ambiguous nature of human language communication, as well as codeswitching, which is a common occurrence in communication between multilingual communities on social media [37]. Additionally, manually sifting and analyzing user-generated data from social media is inefficient, time-consuming, expensive, and impractical given the volume, variety, veracity, and velocity of the data.

Recent years have seen an increase in the number of papers on hate speech, notably on automatic detection of hate speech utilizing text mining and natural language processing techniques [10], [31], [38], [39], [40], [41], and [53]. However, there are few publications on identifying hate speech in codeswitched communications. Additionally, present parsing approaches and other natural language processing tools are designed for monolingual datasets, making it difficult to handle codeswitched texts. Additionally, context-sensitive techniques perform poorly due to the syntactic and lexical difficulties introduced by the increased ambiguity introduced by the employment of words from many languages inside the same communication [37]. As a result, standard techniques are insufficient, and increased performance in detecting subtle types of hate speech on social media is required [41]. Automatic text classification is fundamentally about detecting and selecting features from natural languages that result in high-performance feature sets suitable for training simple, effective, and efficient classifier models.

Deriving quality features from social media data is a critical process that is fraught with difficulties due to the volume, pace, variety, truthfulness, and importance of big data. Additionally, it frequently demands a high level of domain experience, referred to as subject matter expertise (SME), which is expensive to employ. As a result, the entire process of collecting, annotating, and selecting high-quality features that best describe hate speech in a codeswitching environment for the goal of training a machine classifier is both difficult and costly.

In light of these difficulties, it is necessary to investigate a methodology that more accurately captures essential characteristics inherent in nuanced forms of hate speech to improve the effectiveness of the machine classification of huge data. This type of data, particularly that derived from user-generated content on social media, must be handled methodically to convert it from a highly dimensional and low-quality state to a low-dimensional and high-quality one suitable for training a machine classifier. As a result, it is hoped that the deployment of the machine classifier will improve data-driven decision-making by key stakeholders in national security by serving as early-warning systems that automatically monitor hate speech on social media, particularly during perennial trigger events such as presidential elections, referendums, and occurrences such as terrorist attacks and gender-based violence.

### **1.3 Objectives of the Study**

This section discusses the study's overall purpose and its specific objectives.

### 1.3.1 General Objective

The study's primary objective was to investigate the utility of a novel psychosocial feature set for detecting subtle forms of hate speech, particularly in codeswitched text messages, to train a machine classification model that can generalize to other types of hate speech shared on social media, using the 2017 Kenyan presidential elections as a trigger event.

### 1.3.2 Specific Objectives

Five objectives have been derived from the basic objective. These are intended to:

1. Gain a firm grasp of what constitutes hate speech.
2. Develop a conceptual framework for recognizing key characteristics of hate speech in text messages.
3. Create a dataset containing examples of hate speech on social media in Kenya.
4. Develop a machine classification model capable of detecting subtle kinds of hate speech embedded in code-switched text communications from social media.
5. Evaluate the classification model.

### 1.4 Research Questions

The major study question is: Does the novel psychosocial feature set improve machine classification of nuanced kinds of hate speech in code-switched text messages from social media? Among the specific inquiries are the following:

- i. What is considered to be hate speech?
- ii. How informative are psychosocial characteristics in text message discrimination?
- iii. How can we extract hate speech-containing text messages from social media to create a high-quality dataset representative of hate speech in Kenya?
- iv. How do these psychosocial characteristics stack up against the more traditional characteristics used to train machine learning models to accurately recognize hate speech?
- v. To what extent does our model generalize in terms of predicting different sorts of hate speech in social media text messages?

## 1.5 Research Significance

This research is timely in that it provides a mechanism for classifying hate speech that may be used to annotate text messages for supervised learning in machine classification. The classification framework postulates the existence of a composite of discriminant characteristics based on psychosocial ideas of separation, negative passion, and commitments to hatred. These provide useful features that enable the identification of more subtle types of hate speech that are concealed in codeswitched data and cannot be identified properly by traditional features.

Additionally, the study addresses a research gap in codeswitched hate speech datasets by creating an annotated dataset that will be made publically available for comparative research in hate speech categorization studies.

Unlike many previous efforts, the supervised model will be able to use a bootstrapping strategy to discover and categorize subsequent text documents, which will be especially beneficial in the absence of human annotators and the associated overhead associated with labeling new words or phrases.

These research questions assist in elucidating the larger significance and influence of the research findings on our understanding of hate speech. The findings will aid in developing strategies for preventing and intervening with hate speech in online public areas. Additionally, the classification approach, like others [42], will aid government intelligence agencies in automatically monitoring social media for surges in hate speech and other divisive communication.

Finally, the research findings should spark discussions that will inform policy development in the area of hate speech, particularly with regards to user-generated content in online public places.

## 1.6 Research Justification

While numerous studies have been conducted on the automatic identification of hate speech, few have highlighted a framework for the appropriate gathering and annotation of data for supervised machine learning.

There is a dearth of publicly accessible, credible datasets of annotated hate speech for comparative studies. This adds to the sluggish or haphazard establishment of hate speech policies based on empirical evidence. Numerous investigations have necessitated the creation of their own datasets, which is costly. Instead, the same work would have been directed toward inventing and refining existing machine learning techniques. While there are few publicly available datasets, those that

exist are domain-specific and may not generalize to other domains of hate speech. For instance, numerous research has been conducted on racism and anti-Semitism. Classifiers for racial, gender, or religious hate speech, on the other hand, will need to be retrained to achieve meaningful categorization performance.

Additionally, there is scant evidence of a framework with a valid theoretical foundation for guiding the acquisition and annotation of data for supervised machine learning. This project seeks to close this gap by developing a comprehensive annotation framework based on sound theory to guide the development of a classification model capable of efficiently discriminating hate speech in a codeswitching environment such as that prevalent on social media in Kenya.

### **1.7 Scope of the Study**

The content of hate speech that will be analyzed in this study is mostly text data. It excludes multimedia elements such as photos, graphics, movies, and audio.

Second, the model is guaranteed to perform very well in a specific context or domain, in this case, the entangled political and ethnic hate speech prevalent in Kenya and other African countries. Transfer-learning will be used to generalize the model to other dissimilar domains such as sports or medical, or settings such as irony, sarcasm, and idioms. However, given sufficient labeled data from those domains and situations, the model can be retrained to learn more salient characteristics and generate more accurate classification results.

### **1.8 Study Premises**

The study's central premise is that there exists an underlying function capable of exclusively mapping social media text messages to one of three specified types. This means that the training data has sufficient informative elements to appropriately categorize any communication as hate speech, offensive, or 'neither'. Second, the publicly accessible brief text messages on Twitter are reflective of similar messages on other social media networks. Thirdly, the words contained in a twitter message follow a consistent distribution that may be calculated probabilistically. Finally, because the project's objective is prediction rather than causal inference, the influence of confounding variables will remain constant from training to testing data throughout text classification[43].

### **1.9 Sinopsis.**

While advancements in Internet technology have resulted in numerous innovations and benefits in social networking and data communication, they have also exacerbated the prevalence of hate speech and other similar phenomena, such as cyberbullying, online harassment, and fake news in online public spaces[44][35]. Additionally, certain technological characteristics on these social media networks have enabled the bulk reposting, forwarding, and broadcasting of hate content in real-time to a global audience swiftly, anonymously, and economically.

Similar research in the past that employed computational tools to combat online hate speech has largely relied on monolingual datasets, with English being the most frequently used language. Additionally, as more multilingual groups spread throughout the world, codeswitching has become the standard rather than the exception while conversing on these social media sites. This, however, creates processing issues for monolingual dataset-aware language tools. This work makes two key contributions in this area. To begin, it constructs a gold-standard annotated dataset of codeswitched messages. Secondly, and perhaps most significantly, the study contributes to the machine classification literature by developing a novel hate speech framework grounded in psychosocial theories. This framework captures additional salient features that can be used to effectively train a machine classifier to identify nuanced forms of hate speech derived from the language of psychosocial distancing, negative passion, and commitment to hate.

The thesis is divided into five sections. The first chapter discusses the study's context, which includes the research problem, the purpose, objectives, and research questions, the study's significance, justification, scope, and assumptions. The second chapter discusses the state-of-the-art in identifying hate speech by analyzing prior similar research and compares them to the current study. The third chapter discusses the approach used to accomplish the study's objectives. The fourth chapter summarizes the findings of several research projects and machine learning experiments. The fifth chapter elaborates on the findings and provides an in-depth examination of the insights gained and how they might be applied to real-world problems.

## CHAPTER 2: LITERATURE REVIEW

### 2.1 Hate Speech

This chapter conducts a thorough overview of existing research on the identification of hate speech in text documents, more specifically in brief text messages common on social media sites. The chapter begins by defining hate speech and then discusses some of the more popular theoretical approaches or theories of hate speech. It then reviews some of the existing efforts to address automatic identification of hate speech and concludes by presenting the proposed conceptual framework that will guide the research.

The many definitions of hate speech are discussed in this paragraph, and the content analysis technique is used to identify some prevalent terminologies that could aid in establishing the study's core themes. Subsection 2.1.2 discusses some of the most useful theories of hate to have a thorough knowledge of the phenomena of hate speech. Subsection 2.1.3 details the framework upon which the empirical portion of this study is based.

#### 2.1.1 Defining Hate Speech

There are numerous definitions of hate speech in the literature, and there is no universal one. Nonetheless, other institutions, including international and domestic legislation, have attempted to define hate speech and even designated specific targets based on what is formally referred to as protected qualities, such as race, ethnic origin, religion, or gender. This section is not designed to redefine hate speech or to provide a universal definition, but rather to construct a workable definition to establish a shared understanding in this study. Existing definitions of hate speech can be found in dictionaries, existing legislation and policies of the government and non-government groups, as well as in published papers.

The following are some encyclopedia and dictionary definitions of hate speech:

Encyclopedia of the American Constitution [45].

*“Any prosecutorial, hateful, and degrading expression that conveys a message of group inferiority about a historically oppressed group.”*

The Oxford dictionary [46]

*“Abusive or threatening speech or writing that expresses prejudice against a particular group, especially on the basis of race, religion, or sexual orientation.”*

The Oxford English Dictionary [47]

*“A speech or address inciting hatred or intolerance, esp. towards a particular social group on the basis of ethnicity, religious beliefs, sexuality, etc.; (b) (as a mass noun) speech (or sometimes written material) inciting such hatred or intolerance.”*

The Merriam Webster dictionary [48]

*“Speech that is intended to insult, offend, or intimidate a person because of some trait (as race, religion, sexual orientation, national origin, or disability).”*

Some of the definitions of hate speech, as found in international agencies include:

UN’s International Committee on the Elimination of Racial Discrimination [49]

*“A form of other-directed speech which rejects the core human rights principles of human dignity and equality and seeks to degrade the standing of individuals and groups in the estimation of society.”*

The European Court of Human Rights [50]

*“All forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility towards minorities, migrants and people of immigrant origin.”*

Some of the definitions of hate speech from African governments and agencies include:

Kenya National Cohesion and Integration Commission (NCIC) Act, 2008 [51].

*“Words of incitement and hatred against individuals based on certain group characteristics they share. It includes speech that advocates or encourages violent acts against a specific group, and creates a climate of hate or prejudice, which may, in turn, foster the commission of hate crimes.”*

Some of the definitions of hate speech from social media companies include:

YouTube hate speech policy [14]

*“Content that promotes violence against or has the primary purpose of inciting hatred against individuals or groups based on certain attributes, such as race or ethnic origin, religion, disability, gender, age, veteran status, sexual orientation/gender identity.”*

Facebook hate speech policy [52]



*“Content that attacks people based on their actual or perceived race, ethnicity, national origin, religion, sex, gender or gender identity, sexual orientation, disability, or disease is not allowed. We do, however, allow clear attempts at humor or satire that might otherwise be considered a possible threat or attack. This includes content that many people may find to be in bad taste (example: jokes, stand-up comedy, popular song lyrics, etc.).”*

Twitter hateful conduct policy [13]

*“You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.”*

LinkedIn [53]

*“Do not use LinkedIn's services to promote or threaten violence or property damage, or for hate speech acts like attacking people because of their race, ethnicity, national origin, gender, sexual orientation, political or religious affiliations, or medical or physical condition. LinkedIn does not allow ads that include hate speech or show or promote violence or discrimination against others or damage to their property or are personal attacks on any individual, group, company, or organization or otherwise advocating against or targeting any individual, group, company, or organization.”*

This section also considers some of the definitions of hate speech from previous similar studies.

Gitari et al. [24] define hate speech as *“a kind of speech that demonstrates a clear intention to be hurtful, to incite harm, or to promote hatred.”*

According to Silva et al.,[41] hate speech is *“any offense motivated, in whole or in part, by the offender’s bias against an aspect of a group of people.”*

Davidson et al. [54] describe hate speech as *“ language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group.”*

The UMATI project [55] employs three of the five components of Benisch's [36] framework for defining dangerous speech as that which: first, targets a group of people based on their ethnicity, race, or other characteristics; second, devalues people by comparing or referring to them in terms

of insects, animals, or spots that contaminate the in-purity; group's and third serves as a call to action to chase, evict, riot, or kill.

Table 2.1 summarizes these definitions, their key concepts, and their objectives. This section summarizes definitions from a variety of stakeholders, each focusing on a different component of hate speech, as seen by the terminology used in the individual definitions.

Table 2.1: Hate Speech Definitions

| Entity                             | Key Concepts  | Target   |
|------------------------------------|---|--|
| Oxford dictionary                  | Abusive, threatening, prejudice   | Race, religion, or sexual orientation.   |
| Oxford English Dictionary          | Inciting hatred, intolerance  | Ethnicity, religious beliefs, sexuality,   |
| Merriam Webster dictionary         | Insult, offend, intimidate  | Race, religion, sexual orientation, national origin, or disability   |
| UN's ICED                          | Othering, degrade and rejects human dignity and equality  | Othering, individuals, and groups  |
| The European Court of Human Rights | Spread, incite, promote, justify hatred, Intolerance, aggressive nationalism, ethnocentrism, discrimination | Racial hatred, xenophobia, anti-Semitism, Intolerance  |
| NCIC Act, 2008.                    | Incitement, hatred, advocate for violence, a climate of hate, prejudice                                     | Ethnicity, color, race, nationality (including citizenship), or national origins   |
| YouTube                            | Promote violence, inciting hatred   | Race or ethnic origin, religion, disability, gender, age, veteran status, sexual orientation/gender identity                       |
| Facebook                           | Attacks people  | Race, ethnicity, national origin, religion, sex, gender or gender identity, sexual orientation, disability, or disease             |
| Twitter                            | Promote violence, directly attack, threaten, incite harm  | Race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease. |
| Susan Benesch,                     | Targets at a group, devaluing, call to action   | Race, ethnicity, gender, sexual orientation, religion  |
| Gitari et al., 2015                | Intention to be hurtful, to incite harm, or to promote hatred   | Nationality, race, ethnicity, gender   |
| Davidson et al., 2017              | Expresses hatred towards a targeted group, derogatory, to humiliate, or to insult                           | Race, gender, religion   |

According to the many definitions, hate speech has three distinct facets: To begin, it is an expression, whether nonverbal via body language or verbal via writing, visuals, or graphics, that threatens, incites, discriminates, degrades, attacks, intimidates, insults, offends, or stigmatizes. Second, these hate speech utterances can be classified into two categories: those directed directly at an individual and those directed at a group of individuals based on their membership in a protected social characteristic such as gender, ethnic origin, religion, or race [56]. Thirdly, the term is intended to arouse hostility, aggression, prejudice, intolerance, and unfavorable attitudes and feelings toward the subject.

### 2.1.2 Operational Definition for Hate Speech

The term hate speech is viewed linguistically as a compound noun and syntactically as an amalgamation of two distinct words that introduces a new but related semantic meaning.

The study develops a workable definition for hate speech based on an examination of the most frequently used words across the various definitions of hate speech in Table 2.1:

*Any text communication that expresses an attitude of [prejudice, discrimination, or hatred] targeting an individual or a group based on a protected social category [ including but not limited to race, religion, gender, ethnicity, nationality, sexual orientation, language, custom, class, nationality, political identification or disability].*

This definition is based on three essential observations gleaned from content analysis, namely that hate speech is:

A manifestation of one's attitude and emotions, i.e. Psychological dimensions

Has a target group or individual that it seeks to distance itself from, i.e. Aspects

social Has an aim, which is frequently to intimidate, offend, denigrate, or degrade

A deeper grasp of these three dimensions was sought and gained through the explanations in the following section on hate theories.

A working definition was required from the start to guide the annotation process for the social media text messages used in the study's experiment phase.

To contextualize the issue, the researchers focused on the most widespread kind of hate speech in Kenya, which is negative ethnicity, sometimes referred to as tribalism [57]. In Kenya, ethnic hate speech has been especially prominent on social media during presidential election seasons [4], [58].

Three basic classes were formed to guide the classification of social media communications based on the definitions of hate speech provided above. These are divided into three categories: hate speech, offensive speech, and neither. The following sections contain examples of each of these classes.

"Governor Matata is incredibly self-centered and naive." [Offensive]

"Kenyan governor Matata is incredibly self-centered and bumbling" [Hate speech]

"The governor's action was highly self-centered and rash." [Neither]

Generally, whenever a generalization is directed at a particular group, like in the second case above, it becomes undesirable and is thus classified as hate speech. However, if a reference is made to a single member of the social group, it is likely to be deemed "offensive." If the attack or insult is directed at neither a person nor a social group, it will be treated as "neither." Oftentimes, hate speech is more easily directed at a group than an individual, as a group is more abstract and hence more natural to be impersonal than an individual. For instance, during the Rwandan massacre, the Tutsi were dubbed "cockroaches." [59][60].

### **2.1.3 Dangers of Hate Speech**

The detrimental impact of hate speech cannot be overlooked any longer, considering that it is a forerunner to physical violence, fanaticism, crime, ethnic cleansing, and in certain cases, genocide [36]. Hate speech results in social marginalization have a detrimental effect on the mental and emotional well-being of target groups and corrupt the offenders' thinking, attitude, and behavior. It instills dread in the target group, limits public involvement, and fosters an equally burgeoning animosity that could erupt into physical violence in the event of a trigger event [16]. Presidential elections as trigger events for hate speech are common not only in Kenya, but also in the United States during President Obama's second term, and in India during Prime Minister Modi's second term campaigns. The magnitude of hate speech's effects or potential ripple effects has been underscored by increased response through international, national, and corporate laws and policies aimed at addressing hate speech in public spaces, schools, the workplace, and public online spaces such as social media networks platforms.

#### **2.1.4 Hate Speech in Kenya**

The history of electoral processes in Kenya, Africa, and generally around the world is exacerbating with the increased election and post-election violence, malpractices, and the proliferation of hate speech, especially during presidential elections, which are often regarded as a matter of life-and-death [61]. During this period, propaganda, negative ethnicity, stereotyping, and other aspects of hate speech are used by politicians to arouse the emotions of supporters, stifle the voices of the minority, and demonize opposition to gain political power and a following [62], [63]. The perpetual politicization of ethnicity has often stirred animosity among ethnic groups resulting in hate speech, violence, forceful eviction, destruction of property, and even death [64]. In Kenyan history, the climax of this narrative was during the 2007/2008 presidential elections whereby the country faced the worst ethnic profiling and post-election violence (PeV) that ultimately led to over 1200 deaths, about 300,000 internally displaced people, over 42,000 houses destroyed, destruction of crops, and looting of commercial outlets [65].

#### **2.1.5 Hate Speech Laws in Kenya**

Following the post-election violence in 2007/2008, increased focus on hate speech resulted in the examination of existing laws, the development of new regulations, and the establishment of government agencies to handle the matter. In this connection, the Kenya National Cohesion and Integration Commission (NCIC) was established by Parliament in 2008 to handle precisely issues of ethnic hate speech and community peacebuilding. Under Section 13 of the NCIC Act No. 12, 2008, hate speech is defined as speech that is threatening, abusive, or insulting toward a shared group feature such as ethnic origin [57].

Hate speech is protected under the country's constitution under chapter 4 of the Bill of Rights and article 33 on freedom of expression, but not incitement to violence, propaganda for war, hate speech, or support for ethnic provocation [66].

Other earlier laws that prohibit hate speech include the Kenya Information and Communications Amendment Act, 2013 under section 84D, which prohibits the electronic publication of obscene information, and section 27, which states that the freedom of the media does not extend to the dissemination of propaganda for war, incitement to violence, the dissemination of hate speech, or advocacy of hatred that constitutes ethnic incitement [67].

### 2.1.6 Hate Speech on Social Media Platforms in Kenya

The proliferation of online social media platforms has created new hurdles for monitoring and detecting offensive language, cyberbullying, online harassment, and hate speech generated in these online public spaces [56]. Additionally, unlike traditional print and electronic media channels, social media's user-generated content features enable anyone (not only politicians) to freely, publicly, and possibly anonymously create and post hate speech content to a bigger audience in a matter of seconds. Additionally, once a message is posted, it can quickly be shared, re-posted, and liked, exacerbating the effect and proliferation of hate speech and making it more difficult to contain or eliminate [11], [28].

Due to the numerous linguistic, theoretical, and technological challenges inherent in automatically processing the massive amounts of data generated by these online platforms, government agencies in some African countries have recently gone to extreme measures such as physically disconnecting Internet access or threatening to do so during electoral campaign periods in order to mitigate the impact of online hate speech. For instance, in 2017, the Democratic Republic of Congo's government ordered the shutdown of social media networks as President Kabila's mandate expired, while Ethiopia and Uganda cut down Internet access and the Kenyan government remained resolute in monitoring social media before elections [4], [68].

The NCIC primarily relies on user reports of hate speech on social media, supplemented by a small number of human monitors on social media sites. NCIC held a one-week introduction training for cohesion monitors in March 2017, just four months before Kenya's national elections. It taught 390 persons on spotting hate speech. Each county received 47 video recorders, and each cohesion monitor received a voice recorder to assist them in recording political rallies in their various constituencies around the country. These recordings were intended to be sent straight to a central server for the purpose of monitoring hate speech. This was accomplished through a program called UWIANO, a Swahili word for cohesiveness. Unlike previous general elections, which saw fragmented government efforts by its peace and conflict management agencies, namely the NCIC, the Criminal Investigation Department (CID), the National Police Service (NPS), and the Independent Electoral and Boundaries Commission (IEBC), UWIANO became the umbrella organization under which these various agencies coordinated their efforts to ensure peaceful, fair, and free general elections.

These kind of efforts taken by government entities prior to every other election serve as a reminder that hate speech frequently erupts during and after a single significant event [16]. Kenyans now have a platform for creating and consuming local content that was previously unavailable or limited but is now widely available and economically delivered in massive quantities. Additionally, access to high-quality Internet services and affordable smartphones, combined with social media features such as retweeting, reposting, likes, forward buttons, and emoticons, increases the speed with which ideas and sentiments can be shared publicly and even anonymously, albeit some bordering on hate speech. These and other characteristics make social media an ideal environment for the spread of hate speech.

### **2.1.7 Challenges of Monitoring Social Media**

The difficulties associated with monitoring social media range from a lack of enforcement of norms and regulations to insufficient monitoring tools, user behavior, and a general lack of understanding of what constitutes hate speech. Until the time of writing this thesis, social media firms relied heavily on users to report terms infractions, which are then manually reviewed. This has been linked to the complex balance between free expression and what constitutes hate speech, as well as to the always changing user behavior in evading hate speech-related measures. Facebook acknowledged the difficulty in catching content such as hate speech, owing to the dynamics of user behavior, such as when users repost previously marked information.

Due of the dynamic nature of internet users, they have developed techniques to circumvent hate "filters" in the majority of existing software. Several common techniques for evading lexical filters include adding or removing letters to obscure the offensive word, concatenating words without leaving any space between them to render them unintelligible to the filter, using abbreviations such as WTF, code-switching (using words from other languages) in light of the fact that the majority of filters are built on a single language, and the use of epithets and emoticons [22].

This motivates the hunt for a more robust alternative capable of managing these dynamics consistently across all time zones of the distributed user base.

The use of computer software as an alternative will be given precedence. However, this, like a human creature, has limitations. A human reviewer is capable of distinguishing offending language from useful text on the Internet, as well as contextualizing and analyzing the meaning of a word inside a sentence, which is not a straightforward operation for computer software.

Thus, this study contributes to the growing corpus of research addressing the difficulty of automatically monitoring and detecting hate speech on social media platforms, which, if not addressed swiftly, can easily devolve into violence and even genocide [69]. For instance, consider the 1994 Rwanda genocide, the bloodiest in African history, which was sparked by hate speech propagated on a national radio station [59].

## 2.2 Theories of Hate in Social Psychology

This section discusses some critical hate theories on hate speech intending to increase our understanding of the phenomena. The ultimate objective is to comprehend the essence of hatred and how it manifests in text messages. This is based on the assumption that there is a correlation between the way words are used in written text and the social psychology aspects of hatred [70]. The social identity theory, self-categorization theory, speech act theory, communication theory, critical race theory, Baumeister's theory, the integrated threat theory, the sociologist's homophile theory, and the triangular theory of hate are briefly examined in this context. The most informative theory is expanded upon.

The social identity and self-categorization theories are similar in that they both describe how the in-group membership distances itself from the out-group membership, frequently by ridiculing or boasting about the in-group membership's superiority. This moral or cultural superiority toward foreigners or other ethnically diverse groups is commonly referred to as ethnocentrism [71]. The goal is to boost the in-pride, group's self-image, and esteem [72]. However, this frequently presents itself in hate speech as "we/us" versus "they/them" classification and membership generalization tendencies [34]. For instance, "Every Kikuyu is a thief." We will correct them this time".

Speech act theory is a type of meaning theory that describes how language statements influence the performance of various activities by commanding, warning, rebuking, and shouting. This theory helps to explain how hate speech messages contain call to action expressions, such as inciting members of the in-group to evict or chase members of the out-group [36]. Additionally, this explains the directions and commitment to preserving the in-'purity,' group's as seen by literature including hate terms.

The critical race theory (CRT) stresses or focuses on race-based categorisation and the relationship between law and power. The CRT is widely used in research on race and racism identification. CRT elucidates the connection between hate speech and elements of subjugation, humiliation,



humiliating, and injuring the target racial groups. In a prior work of a similar nature, CRT was used to establish 11 guidelines for annotating a corpus for racism [10].

Baumeister's thesis elucidates the source of hatred, as manifested in the desire for vengeance for historical wrongs. This idea assists in comprehending the presence of guilt or blame shift tendencies, greed calling, and ambition based on group membership in hate speech.

The integrated threat theory, also known as the intergroup threat theory, explains how risks are perceived or felt, which results in bias and withdrawal from out-group membership. In the case of hate speech, this enables an understanding of the danger markers in a message or the perception of outgroup membership as a threat to the in-group [73].

According to the sociologist's notion of homophily, humans will always follow their type. Lozano et al. [74] apply this theory to identify social group trends by examining the "mention or following" status of racist users, which is aided by Twitter's structural network qualities.

The labeling theory explains why the advantaged group has prejudices about the out-group. This is demonstrated in hate speech by belittling aspects, portraying the out-group as less mature, or utilizing object or animal names to code the out-group.

The triangle theory of hatred is the inverse of Sternberg's Triangular theory of love, which is frequently used to explain the fundamental components of love in interpersonal interactions [33]. According to this theory, love consists of three key components, which are depicted in a triangle: Intimacy, Passion, and Commitment.

Intimacy is defined under the idea as a sense of attachment, intimacy, and togetherness. Passion refers to the strong emotions felt when one is passionately or sexually attracted to another person. Commitment combines the elements of closeness and desire in order to forge a long-term commitment to one another.

The triangle of hate hypothesis elucidates three fundamental elements of hatred that span the seven distinct types of hatred. These are the elements of the Negation of Intimacy, which is ideally the concept of Distancing, Passion, and Commitment. Figure 2.1 illustrates them.

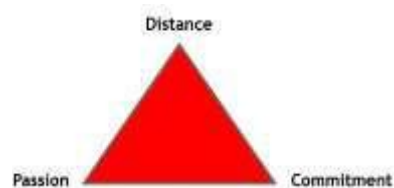


Figure 2.1: Triangular Structure of Hate

Additionally, the object of hatred can be an individual or a group. This theory is critical in this study because it serves to give the hate phenomena a structure or a form that is more easily comprehended by humans. More crucially, it provides a theoretical framework for the three components proclaimed in the study's annotation model, namely Distance, Passion, and Commitment, which provide the basis for the selection of features used in the experimental phase. Distance is represented by features of social isolation or "otherness"; passion is exhibited through unpleasant emotions expressed through the use of disparaging or insulting language; and commitment is expressed through feelings of contempt for the target, who is regarded as inhuman or subhuman. Table 2.2 summarizes these hypotheses and the key variables that underpin the hate speech investigation. In this sense, the social identity theory, as well as self-catastrophizing.

Table 2.2: Theories concerning hate speech

| <b>THEORY</b>  | <b>Description</b>   | <b>RELEVANT FEATURES</b>   |
|--|--|--|
| Social Identity theory                                   | Focused on issues of prejudice and discrimination. Promoting the in-group and one's self-identity while belittling the outgroup.   | Prejudice, Discrimination; in-out group, favoritism                                    |
| Self-categorization theory                               | The concept of In-group versus Out-group and the use of Us versus them.  | In-group vs. Out-group; Us vs. Them  |
| Speech act theory  | Elocutionary act, illocutionary act, perlocutionary act, call to action  | Directives, Declarative, Expressive  |
| Critical Race Theory                                     | The inter-centricity of race and racism; the challenge of dominant Ideology; the commitment to social justice; the centrality of experience knowledge; and the interdisciplinary perspective | Subordination, humiliation, degrading  |
| The duplex theory of hate                                | The negation of intimacy (distancing) in hate: Repulsion and disgust. Subhuman or inhuman, traitors, infidels. Passion; and Decision.  | The negation of intimacy; Passion; Decision; Story (culture belief: Stereotypes)       |
| Sociologist theory of homophile                          | Humans always follow their kind  | Stereotypes  |
| Integrated threat theory, (aka intergroup threat theory) | It attempts to describe the four components (realistic threats, symbolic threats, intergroup anxiety, and negative stereotypes.) that cause a perceived threat between social groups         | realistic threats, symbolic threats, intergroup anxiety, and harmful stereotypes       |
| Labeling Theory<br>By sociologist<br>Howard Becker       | Explains why certain people are viewed as different from or less worthy than others. Stereotypes, when used by people in power, can have very negative consequences.                         | Stereotypes or discrimination. Social prejudice, scapegoating, coding of the out-group |

The UMATI project [75] used Benesch's [36] paradigm for identifying harmful speech in establishing a system for monitoring and analyzing hate speech collected between October 2012 and March 2013, which coincided with an election campaign in Kenya. Five constructs are included in the framework that influence the dangerousness of communication. These factors include the speaker's power over the audience, the audience's receptiveness, the meaning of the speech act, such as a call to violence, the enabling social and historical backdrop, and the persuasive medium of dissemination. Additionally, the framework employs three indicators to identify dangerous speech: dehumanization of the target group through the use of an animal name or some code language, implying that the in-group faces a grave threat from the out-group, and implying that elements of the out-group are tainting the in-purity group's or integrity. All of them are based on the "distance" component in recognizing hate speech, which is characterized by the othering discourse of in-group against out-group, pure versus impure, or us versus them. Additionally, the concept of "othering" speech has been used to explain the exclusion caused by the usage of phrases such as "we" vs "them." Several scholars have already utilized this to identify aspects of hate speech [34], [76], [77], and [78].

Additionally, our research advocates for the term "distance," which broadens the definition of othering discourse to encompass perceived superiority, morality, maturity, and purity of the in-group relative to the out-group. Additionally, the distance factor effectively captures devaluation of the target group, when they are regarded to be less mature or inferior to infants, animals, or insects.

Is there a theory that encompasses all of the components of hate speech? There is currently no single theory that encompasses all of the variables. In the context of our investigation, Stenberg's [33] triangle theory of hatred appears to capture numerous critical characteristics that have been incorporated into other models. Distance, passion, and belief/culture are all facets of hatred. As a result, the dimensions of this theory were chosen as core variables in addition to what we may term secondary elements gathered from the other theories. Our study will employ these secondary components to construct a framework consisting of important factors for identifying hate speech in text data.

Although these are social science theories, they are critical in helping to crystallize our understanding of how hate is communicated through text texts. This becomes crucial when it

comes to coding the extracted characteristics in a way that classification algorithms in computer science and machine learning can employ.

## 2.3 Text Classification

The automatic detection of hate speech in text is a text classification challenge. As such, we begin by discussing the notion of text classification, the approaches and methods used, as well as some of the past research on automatic hate speech identification.

Text classification is often referred to as text categorization or text tagging. This is the process of organizing, structuring, and grouping text into preset categories. Classification of text can be performed manually or mechanically. The purpose of this study is to examine automatic text classification algorithms. This procedure necessitates the completion of essential activities such as data collection, data analysis, feature selection, feature building and weighting, model training, and model evaluation [79]. This work falls within the broad category of natural language processing, in which computer machines process human-generated speech in the form of text data. This is especially advantageous when dealing with large amounts of data, such as user-generated data from social media networks. Automatic categorization tasks can be accomplished using either hand-coded linguistic rules, as in a rule-based system, or labeled examples from historical data, as in machine learning, or a hybrid system integrating both approaches. While rule-based systems are intuitive to people, they are inefficient and time-consuming to maintain due to the large number of rules developed and the complexity introduced over time. According to the literature review, the most frequently used strategy is the application of machine learning.

### 2.3.1 Machine Learning

Fundamentally, machine learning comprises automatically and intelligently discovering patterns or regularities in features derived from training data with the goal of optimizing performance through forecasts or other useful approximations based on data or prior experience. Machine learning is a subfield of artificial intelligence that refers to systems that are capable of learning and adapting to changes in their environment without the intervention of a system creator [80]. Mathematically, machine learning is the act of analyzing data input ( $x$ ) in order to learn a function

(f) either by identifying relevant data patterns and aggregating relatively similar records, as in unsupervised learning, or by producing a specified output ( $y$ ), as in supervised learning. In many cases, input ( $x$ ) contains the independent variables, whereas output ( $y$ ) is the dependent variable

in this situation. As a result, machine learning methods are utilized to discover a function that maps input( $x$ ) to output ( $y$ ). In other words, machine learning involves two distinct and concurrent operations: learning performed by the algorithm and classification performed by the learnt function (model). The equation in Figure 2.2 summarizes this well mathematically. If the underlying function that transfers the input to the output is already identifiable, it is referred to as the target function. This is not always the case, as the output is frequently distorted by noise signals, impairing the learnt function's performance. As an optimization problem, machine learning's central objective is to select the best function from a group of candidate functions, also known as hypotheses, that produce the best estimates for mapping the input space to the output domain [81].

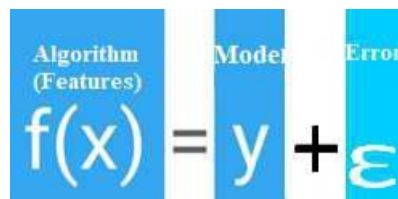

$$\begin{array}{|c|} \hline \text{Algorithm} \\ \text{(Features)} \\ \hline f(x) \\ \hline \end{array} = \begin{array}{|c|} \hline \text{Model} \\ \hline y \\ \hline \end{array} + \begin{array}{|c|} \hline \text{Error} \\ \hline \epsilon \\ \hline \end{array}$$

Figure 2.2: Function of Machine learning

There is frequently an irreducible error ( $\epsilon$ ) in machine learning that is independent of the input ( $x$ ), which explains why no machine learning function exists that learns a model from data with 100 percent prediction accuracy [82]. This error could be caused, for example, by insufficient features, which would have a detrimental effect on the learning process.

By default, machine learning produces two outputs, a process called binary classification. This is the most common type of learning assignment since it concentrates on a single desired output, such as whether a message is hate speech or not; otherwise, all other messages are considered not to be hate speech. Occasionally, machine learning algorithms specify more than two outputs. This is referred to as classification with many classes. For instance, determining whether an SMS message contains hate speech, is offensive, or contains neither.

Two difficulties frequently occur in machine learning. The first is concerned with learning and correctly identifying each case, which is particularly challenging when dealing with a limited dataset. This may appear to be a favorable option on the surface, but it introduces the issue of overfitting. This means that the classifier will perform poorly when confronted with fresh

occurrences that were not included in the training set. This is comparable to rote learning, in which students are instructed on a series of questions and their proper responses and then asked to complete the exact questions on the final test. Their outstanding performance will appear to be deceptive because they simply regurgitated the answers. Authentic learning will be demonstrated by their ability to respond appropriately to relevant but unseen questions. As a result, learning such a classifier is insufficient, as it will be unable to generalize successfully over previously unseen data. The second issue is learning a classifier over noisy or insufficient training data, which do not ensure that the underlying function accurately maps the inputs to the predicted output. This is referred to as underfitting. The classifier will perform badly as a result of its lack of training on relevant and informative features that aid in discriminating between instances belonging to various preset outputs.

Machine learning makes use of statistical inferences to process, analyze, and understand data patterns in small and large datasets using one of four major learning methodologies: supervised, unsupervised, semi-supervised, or reinforcement learning. These categories are frequently related to certain machine learning issues and techniques, as seen in Figure 2.7.

### **2.3.1.1 Supervised Machine learning**

Supervised machine learning is task-driven and requires not just providing input to the computer, but also training samples containing the intended result. The computer then automatically determines the relationship between the input and the predicted outcome by building a model from the input attributes mapped to the output label. The objective of supervised learning is to make some predictions about the outcomes of experiments given the input data. Figure 2.3 illustrates this. This type of learning is especially advantageous for tackling classification and regression problems. Typical classification issues include detecting identity fraud, medical and computer diagnostics, and identifying hate speech on social media. Several common regression problems include market forecasting, weather forecasting, population growth forecasting, and life expectancy estimation. The Nave Bayes, Support Vector Machines, and K-Nearest Neighbors are all examples of supervised learning algorithms that are frequently employed for classification tasks. Popular algorithms for regression issues include Decision trees, Random Forest, and Linear Regression [83]. Both challenges are amenable to neural network techniques.

A supervised machine learning method performs three main functions. The hypothesis function, the optimization function, and the cost function are all examples of these functions. Each of these functions is described using the Linear Regression technique as an example.

The hypothesis-generating function

A simple linear regression algorithm's hypothesis is typically expressed as



Figure 2.3: Supervised learning

### 2.3.1.2 Unsupervised Machine learning

Machine learning without supervision is data-driven. Unlike supervised learning, it relies entirely on the input data to determine the link between the various data instances and automatically builds clusters of data based on found patterns or structures.

Furthermore, the purpose of unsupervised learning is not necessarily prediction or the discovery of established patterns. Often, evaluation is conducted indirectly and qualitatively. Figure 2.4 illustrates this. This type of learning works extremely well with large amounts of rather unstructured data. Customer segmentation, targeted marketing, and recommender systems are all examples of common clustering difficulties. Latent Dirichlet Allocation (LDA), K-Means, K-Medoids, Fuzzy C-means, Hierarchical, and neural networks are all examples of algorithms. Additional strategies for unsupervised learning include the use of singular value decomposition (SVD) and Principal Component Analysis (PCA) (PCA).



Figure 2.4: Unsupervised Learning

### 2.3.1.3 Semi-supervised Machine learning

Semi-supervised machine learning lies on the continuum between supervised and unsupervised machine learning, in that the unsupervised approach's clusters can be concatenated or utilized

independently to label data for supervised classification. Figure 2.5 illustrates this. Transductive support vector machines, self-training, and generative models are all examples of algorithms [84].



Figure 2.5: Semi-supervised Learning

#### 2.3.1.4 Reinforcement learning

Reinforcement learning refers to the process by which an algorithm runs in a reward system, in which an agent is trained and motivated to follow suitable action steps that are cumulatively linked to some incentive. Positive reinforcement is used in the learning process to optimize rewards. Learning tasks, robot navigation, gaming artificial intelligence, and real-time judgments are all examples of common reinforcement learning difficulties. Figure 2.6 illustrates this.



Figure 2.6: Reinforcement Learning

Additionally, it is possible to merge two or more of the algorithm's trained classifier models. This is referred to as ensemble learning. The ensemble notion is based on the idea of training and testing several models learnt by the same machine algorithm and maximizing their cumulative strength, which frequently results in more robust models with lower variance than using a single learned model. Boosting techniques such as the Random Forest and bagging methods such as the Extreme Gradient Boosting are examples of ensembles.

Boosting methods enhance weak learners by utilizing weighted averages of data samples. This is accomplished through an aspect of "teamwork," in which the result of one model informs the input features for the next model. As a result, models that perform better receive higher weights. Additionally, this contributes to the reduction of model under-fitting bias.

Bagging algorithms generate numerous versions of the predictor model and combine them to create an aggregated predictor. This idea permits the reduction of variance from models that are



particularly prone to overfitting as a result of the increased accuracy obtained from the training data. Variation is introduced into the data by sampling and substituting the testing data for the various models.

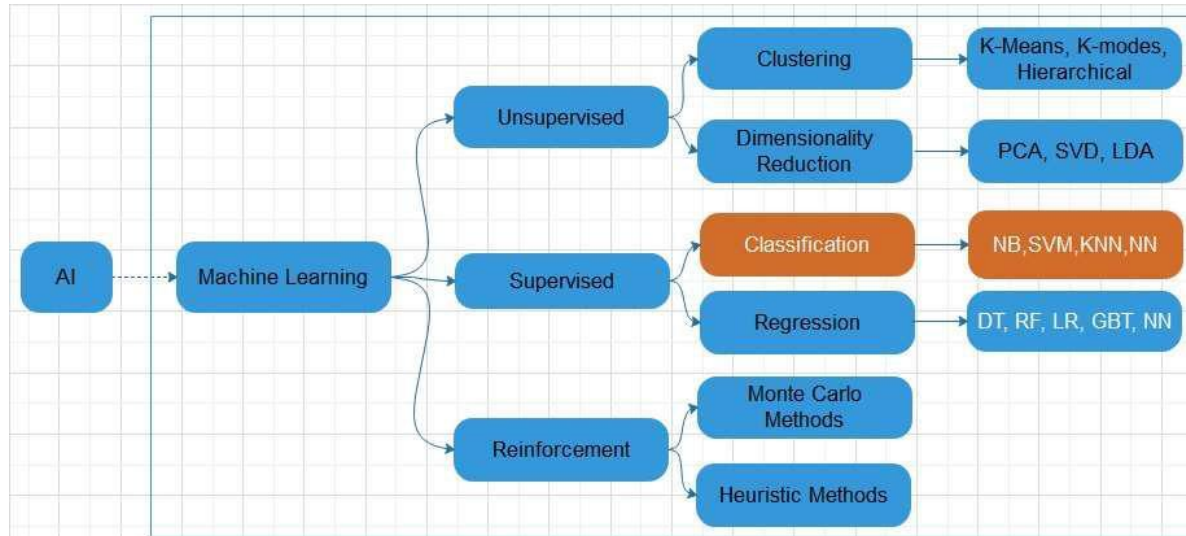


Figure 2.7: Methods of Machine learning

### 2.3.2 Algorithm Types for Machine Learning

Numerous machine learning techniques can be used to perform text categorization and regression tasks. They are all aimed at learning a model by inferring its shape from data. Often, the more training data there is, the higher the accuracy. However, the algorithms differ in their representations and coefficients, despite the fact that their shared objective is to optimize the set of coefficients that produces the best approximation of the target function [82]. The machine learning algorithm of choice is frequently determined by the issue area, the availability of computer resources, the empirical accuracy of categorization, and so forth [85]. Machine learning algorithms are frequently classified as linear, non-linear, or ensemble.

#### 2.3.2.1 Linear machine learning

Linear machine learning deals directly with numerical data in instance space's geometry, utilizing geometric concepts like as lines and planes to impart structure to space and so construct classification models. Non-numerical features must first be converted to a numerical representation that linear models may employ.

Linear models are desirable because they are straightforward, requiring only a fixed set of numeric parameters to be learned from data. Second, they are stable, which means that slight modifications in the training data have a limited impact on the learnt model. Thirdly, unlike other models, linear models rarely overfit the training data, owing to their ability to give decent results with minimal datasets. They are, nevertheless, prone to under-fitting.

Classification, regression, and probability estimation can all be accomplished using linear models. The least-squares classification, the perceptron, Logistic Regression, and the more famous Support Vector Machines are all examples of linear models.

### 2.3.2.1.1 Support Vector Machines

SVMs are used to learn a decent separating hyperplane for a high-dimensional instance space computationally. Separating data linearly in an instance space can imply a plethora of decision boundaries, with certain classifiers performing better than others. Support vectors are the training examples that are closest to the decision border. Thus, the SVM algorithm is used to determine the optimal decision boundary, or to 'draw the optimal line,' between positive and negative vectors, or between vectors belonging to distinct classes in the instance space. Figure 2.8 illustrates this. For instance, in a binary classification, the input  $x = (x_1, \dots, x_n)$  may represent the positive class, defined as  $f(x) \geq 0$ , or the negative class, defined as  $f(x) < 0$ .  $f(x)$  can be represented as a linear function of  $x$  as shown in equation 1.

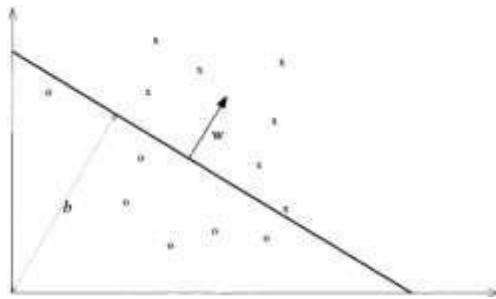


Figure 2.8: Binary Classification adapter from [81]

$$f(x) = (w \cdot x) + b \tag{eq. 1}$$

Generally, the decision boundary or line formed by the equation  $(w \cdot x) + b = 0$  divides the input space  $X$  into two half-spaces, as depicted in Figure 2.8. Positive examples occur above the line, whilst negative instances occur below the line.  $W$  is the vector whose direction perpendicular to the hyperplane is defined by  $t$ . Different values of  $b$  parallelize the movement of the line or hyperplane.

SVM performs well even with sparse instance vectors in high-dimensional feature spaces and eliminates the requirement for feature selection [86]. Additionally, it consumes less memory than other algorithms. This is especially true for text classification, where learning a classifier may require thousands of attributes comprised of words as features. Given the promise that SVM does not overfit and automatically selects the optimal parameters [86], the number of features is irrelevant for text classification in large datasets.

### **2.3.2.1.2 Linear Logistic Regression (LLR)**

The linear logistic regression algorithm is a classification procedure that is used to represent the connection between an independent variable and frequently categorized dependent variables. The advantage of logistic regression is that it displays the effect of multiple independent factors on a single outcome variable quickly. However, it is prone to underfit. Additionally, it performs well when no missing values exist and the predicted variable is binary..

### **2.3.2.2 Non-Linear Models**

Non-linear models are further classified into three subgroups based on their learning algorithms: rule-based, distance-based, and probabilistic.

#### **2.3.2.2.1 Rule-based**

This is a collection of rules and the conditions under which they can be used to get a given forecast. Combination can take one of two forms in supervised rule learning: A rule encapsulates a concept that applies to a collection of examples and produces a certain output, such as a class or label. This approach results in a model that is composed entirely of an ordered rule list. Alternatively, it can be done in reverse, with a class being selected first and then rule bodies covering a subset of the class's examples being identified. The collection of rules is referred to as an unordered rule set in

this technique. Both approaches frequently allow for the overlapping of rules to provide additional information to aid in ranking and probability estimate [87]. However, the list of regulations can become quite lengthy, and rules are notoriously difficult to keep.

#### **2.3.2.2.2 Distance-based Algorithms**

As with linear models, distance-based algorithms can be perceived geometrically. They classify objects using the concepts of exemplars, or prototypical instances, and neighbors, or the closest exemplars. The classifier first learns exemplars that minimize the squared Euclidean distance between each class using the training data and then applies the nearest exemplar decision rule to classify a new input [87]. In general, similarity between instances is quantified by their proximity along the instance space's coordinate axes. The Manhattan distance ( $p=1$ ), the Euclidean distance ( $p=2$ ), the Minkowski distance ( $p$ -norm), the Hamming distance (for binary strings), the Mahalanobis distance, and the edit or Levenshtein distance (for non-binary strings of unequal length) are all frequently used distance metrics [87]. A typical algorithm is the  $k$ -nearest neighbor algorithm.

The  $K$ -Nearest Neighbor algorithm employs lazy learning by classifying data points in relation to the most similar data instances. The term "lazy learning" refers to an algorithm that requires little training and makes no assumptions or generalizations. It is frequently used in simple recommendation systems, such as those found on e-commerce websites. However, its accuracy degrades as data points near the boundary line are added.

For unsupervised learning without a target variable, the distance metric, which ideally learns from exemplars and employs a distance-based decision rule, encodes the learning target in the form of compact clusters based on data point distances but with far-distant centroids, encodes the learning target in the form of compact clusters based on data point distances but with far-distant centroids. The advantage of clustering, particularly as a dendrogram, is that the number of clusters does not need to be known in advance and can be counted simply by inspecting the dendrogram. The disadvantage is that it cannot be used with large data sets due to its computational and memory requirements.

### 2.3.2.2.3 Algorithms Based on Probability

These employ probabilities to classify instances in accordance with their class probability distributions, as in probability estimation trees. The algorithm returns a probability distribution over the target output given an input (x) containing features (y). This is true when discriminative probabilistic models are used. The probability estimation tree is an example.

$P(Y|X)$ , in which Y denotes the dependent output variable and X denotes the independent input variables.

The second class of probabilistic algorithms is generative models, which are used to model the joint distribution  $P(Y, X)$ , where Y is the dependent target variable and X is the independent feature vector. They are referred to as 'generative' because a sample from the joint distribution can be used to generate new labeled data points. The posterior distribution is obtained using the following formalism [80]:

$$(1) \quad \frac{P(Y, X)}{\sum_{c \in C} P(Y, X)}$$

Learning is viewed as a process of reducing ambiguity in generative models by using the posterior distribution from a previous class as the prior for the next step. The Nave Bayes algorithm is an example of a generative probabilistic algorithm. In training the model, the parameters of the distributions used in the model are estimated.

Another popular probabilistic model is logistic regression, which combines a linear decision boundary with logistic steps by determining the optimal conditional probabilities.

### Algorithms for Non-linear Machine Learning

In general, there is no such thing as a one-size-fits-all machine learning algorithm. Different problems require distinct approaches that are optimized for specific machine learning algorithms. Classification, regression, anomaly detection, and clustering are all examples of common machine learning problems.

### Naïve Bayes (NB)

The Naïve Bayes algorithm is a widely used algorithm for processing text data in natural language. It is based on Bayes' conditional probability theorem. In general, Naive Bayes assumes that all features are unrelated to one another and thus contribute independently to the probability of an event occurring. There are three types of naive Bayes algorithms: Gaussian, Multinomial, and Bernoulli. The appropriate use of these types is determined by the nature of the attribute values, i.e., whether they are continuous, multinomial distributed, or distributed in terms of multiple features, each represented by a Boolean variable [88]. The advantage of the Naive Bayes algorithm is that it requires relatively little training data to estimate the required parameters. In comparison to more sophisticated methods, naive Bayes classifiers are simple, extremely fast, and highly scalable. However, the disadvantage of Naive Bayes is that it makes the assumption that all features are unrelated, making it an inefficient estimator and blind to feature relationships.

### **K-Nearest Neighbor (KNN)**

The guiding principle of KNN is that the closest data points in Euclidean space should share or belong to the same class. In general, if a new unlabeled instance exists, several nearby instances in the feature space, i.e., K, can be considered and used to infer its class. KNN is widely used in handwritten character classification, information retrieval, and pattern recognition in general.

KNN is a robust algorithm that works well with noisy training data and is simple to implement. It performs admirably in practice, even with large training datasets, and is easily extensible with additional training examples. However, the cost of computing each Euclidean distance between data points in the entire training dataset is frequently prohibitively high. As a result, the demand for additional storage space increases. Additionally, KNN is prone to succumb to the curse of dimension.

### **Decision Tree (DT)**

A Decision Tree is a hierarchical data structure that consists of a root node representing the entire document, internal nodes representing subsets of the document divided by an individual attribute or term, branches from these representing the weights assigned to each term in the document, and leaves representing the categories [89]. The primary advantage of decision tree algorithms is their visual nature, which makes them easier to comprehend than neural networks. Additionally,

decision tree algorithms can work with unbalanced numerical and categorical data. However, the trees can become complex and are difficult to generalize. Additionally, a small data variation can result in significant tree transformations.

### **Random Forest (RF)**

The random forest algorithm consists of multiple decision trees that work in concert to create an ensemble of supervised learning classifiers for classification or regression tasks. The best prediction is determined through a voting process in which the mode is chosen. The advantage of this approach is that it minimizes overfitting and frequently produces higher accuracy than decision trees. This approach, however, is complicated, difficult to implement, and slow in terms of real-time prediction.

### **Stochastic Gradient Descent (SGD)**

SGD is an optimization algorithm that employs multiple iterations to determine the best approximation of the gradient slope's lowest point. It is an excellent technique for solving large-scale machine learning problems and a necessary precondition for deep learning algorithms [90]. SGD is both effective and simple to implement. It does, however, make extensive use of hyperparameters and is extremely sensitive to feature scaling [91].

### **Latent Dirichlet Allocation**

This algorithm is a member of the unsupervised algorithm class and is frequently used to organize a large corpus of diverse text documents into topics. The topics are then refined into a collection of related words. This is especially useful for imparting structure to a previously unstructured complex corpus, such as text from social media [92]. LDA does not perform precise text classification but rather identifies word relationships. It can be used as a preprocessor to generate dense vector representations for use as input to supervised learning [93].

### 2.3.3 Deep Learning Algorithms

Deep learning is a technique that utilizes multilayered artificial neural networks that mimic the neurons in the human brain to solve extremely complex problems that require the analysis of large amounts of data using artificial intelligence, specifically machine learning. Deep learning algorithms fall into two broad categories: shallow algorithms and deep learning algorithms. Shallow learning algorithms, in their simplest form, consist of three layers: an input layer, a single intermediary layer, and an output layer. Figure 2.9 illustrates a case of shallow learning. Many conventional machine learning algorithms are shallow learning algorithms that use this simple architecture but are sufficiently powerful to solve the majority of traditionally structured and semi-structured problems.

The architecture of deep learning algorithms comprises an input layer, multiple intermediary layers, and an output layer. Figure 2.10 illustrates this. These algorithms are effective at processing semi-structured and unstructured signals, including real-time voice and image signals from CCTV cameras, as well as large amounts of unstructured data from social media. Convolutional Neural Networks (CNN), Recursive Neural Networks (RNN), and Hierarchical Attention Networks are all examples of deep learning algorithms (HAN).

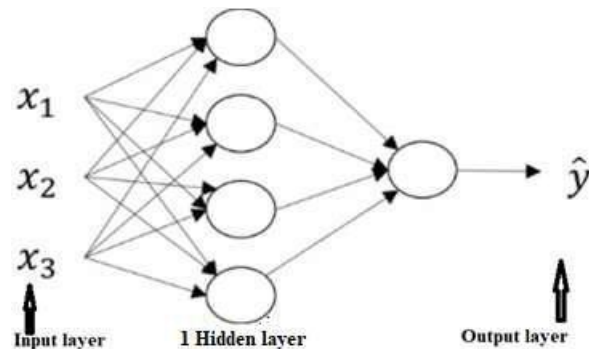


Figure 2.9: Shallow learning

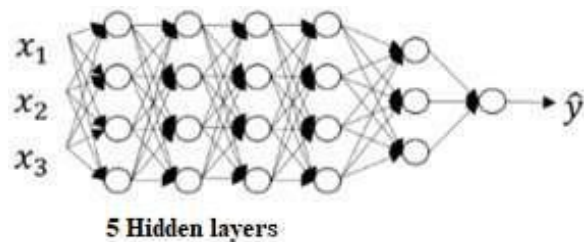


Figure 2.10: Deep learning



## 2.4 Features in Automatic Text Classification

When it comes to text classification, words serve as the initial features, which are then associated with measurements. The term "features" refers to the mappings between the input attribute space and a particular feature domain. Fundamentally, features are the workhorses of machine learning [87]. The classification model's performance is optimally determined by the appropriateness of the features used to train it.

In general, features can be classified into three types: quantitative, ordinal, and categorical. Quantitative characteristics are also referred to as 'continuous' characteristics. These correspond to real numbers. For instance, an individual's weight, height, and age can be converted to a subset of real numbers such as kilograms, feet, and years. Ordinal characteristics are composed of an ordered set but lack a scale. For instance, characteristics that are frequently used interchangeably with adjectives such as high, low, or medium. The mode, median, and quantiles are frequently used statistical measures of these characteristics. Categorical features, alternatively referred to as nominal features, lack a statistical scale or order. However, they are frequently represented using the mode, for example, through the use of Boolean values: true and false. As a result, certain machine learning algorithms perform better with certain feature types.

In unstructured text processing, the predominant feature type is nominal, with words serving as the data's most significant and unique feature. Other features are frequently generated from these initial vocabulary words, which may or may not preserve the initial word order or adjacency. Certain characteristics result in excellent classifier performance, while others have a detrimental effect or have no effect. The former is commonly referred to as pertinent features, whereas the latter are referred to as irrelevant or redundant features. As a result, the goal is frequently to seek out and select only those features that are sufficiently discriminative and informative to enable rapid and accurate classification. This is referred to as dimension reduction.

### 2.4.1 Dimensionality Reduction

The number of input dimensions and the size of the data sample both contribute significantly to the complexity of the classification task for the machine learning algorithm [80]. At its most extreme, this complexity is referred to as the curse of dimensionality. Generally, as more features are added, the performance of the classifier improves up to a point; thereafter, adding features without increasing the number of training instances or records results in performance degradation [94]. As illustrated in Figure 2.11.

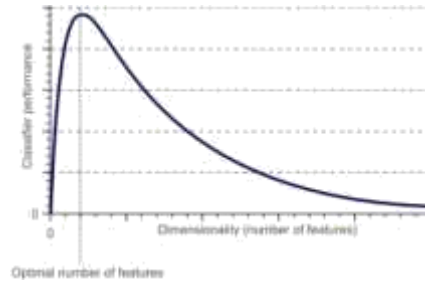


Figure 2.11: The curse of dimensionality (Adopted from [87])

For instance, a task with three features and approximately 100 instances or records is simpler to learn and classify than one with 3000 features and over one million records. In the case of a spreadsheet, features refer to column data that shares similar attributes and has a relatively similar meaning. One of the columns will be the target attribute for supervised machine learning. The rows represent data samples or instances, whereas the cells represent a single data value. This is illustrated in Table 2.3. Complexity develops as a result of the increased demand for resources such as processing time and memory space.

Feature selection and feature extraction are two critical techniques for reducing the dimensionality of the inputs. These are necessary preprocessing steps for removing irrelevant, noisy, and low-value features, resulting in smaller, low-dimensional, high-quality features for learning a classifier. Additionally, the subset of features frequently results in lower requirements for computation time and memory space. Second, simpler models can be used on smaller datasets, resulting in superior performance due to lower variance. Thirdly, using the subset makes it easier to comprehend and extract deeper insights from the dataset. This can be accomplished through the use of visual aids such as charts.

Table 2.3: Features versus training samples

|            | Feature 1 | Feature 2 | Feature 3 | Target-feature |
|------------|-----------|-----------|-----------|----------------|
| Record 1   | Hate      | Kill      | them      | hate speech    |
| Record 2   | Fool      | Fuck      | Stupid    | Offensive      |
| Record ... | ---       | ---       | ---       |                |
| Record 100 | Love      | Happy     | Kind      | Neither        |

### 2.4.2 Generating Features

Machine learning algorithms, on the other hand, do not process raw text data. To begin, the data must be transformed into a numeric representation, such as a vector or matrix, as illustrated in

Figure 2.12. This is accomplished using a technique known as feature engineering. This entails both feature selection and feature extraction processes in machine learning.

#### **2.4.2.1 Selection of Features**

Feature selection entails determining the most informative subset,  $d$ , of the original set of  $D$  dimensions. Fundamentally, it is a process of searching for and identifying the best features from a set of input features that will improve a predictive model's performance [82]. This is advantageous because extracting fewer features from a large feature space results in lower modeling costs in terms of system memory and computation time. This improves the model's performance as a result. Certain input features are omitted during feature selection because their addition introduces noise and reduces the model's accuracy. Otherwise, without feature selection, the maximum number of possible subsets of  $D$  features is  $2^D$ , which grows exponentially with the number of  $D$  features, making it prohibitively expensive and impractical, as previously stated. The selection of features is primarily accomplished through the use of algorithms that employ either the wrapper or the filter method.

##### **2.4.2.1.1 Wrapper Method**

This is the preferred method for determining which features add the most value when combined with other features to form feature subsets. The wrapper approach is used to classify the algorithms used in feature subset selection. The approach entails two local search procedures that utilize either a forward- or backward-selection method. The forward selection method begins with an empty feature set and gradually adds features one at a time, stopping only when no further improvements in model performance are observed. Whereas the backward selection approach begins with the entire set of features and gradually eliminates them one at a time until a significant drop in model performance occurs. As a result, this method entails developing multiple models with various combinations of input features, with the optimal features being those that produce the highest accuracy or some other specified performance metric. The disadvantage of this method is that it can become quite costly in terms of the computational resources required to search for an exponentially large number of feature subsets.

##### **2.4.2.1.2 Filter Method**

This method makes use of statistical procedures to determine which input features best correlate with the target output based on the statistical score assigned to each feature. This is also referred to as the univariate statistical measure due to the fact that the features are evaluated independently.

However, with this approach, it is possible to have redundant features, which are valuable on their own but add little value when combined with other features. As a result, the primary disadvantage of the filter method is that, while numerous features are selected, they introduce the problem of collinearity [95].

Numerous univariate statistical measures are used in the method of filter feature selection. Frequently, the choice is dictated by the data types of the input and output features. Pearson's correlation coefficient and Spearman's rank coefficient, for example, are ideal for numerical input and output characteristics. Kendall's rank coefficient and ANOVA correlation coefficient are optimal for numerical input and categorical output. The Chi-Square and Mutual Information tests are both frequently used correlation measures for categorical input and output features. Additionally, the Mutual Information correlation measure applies to both categorical and numerical data. Finally, the Relief feature selection method iteratively samples a random input  $x$  and determines its nearest hit  $h$  and nearest miss  $m$  to narrow the gap between  $h$  and  $m$ . The Euclidean distance is frequently used to describe quantitative features, whereas the Hamming distance is frequently used to describe categorical features [87].

Fundamentally, filtering algorithms use heuristics to determine the relevance of features outside of the predictive model during the preprocessing stage [95]. Within the supervised method, a well-known algorithm is linear discriminant analysis [96], which linearly combines features that correlate within a given class (intra-class) or separates classes (inter-class) to reduce the dimensionality space and achieve maximum discriminability in a classification task [97]. Generally, the validation or testing data set is distinct from the training data set. This is used to ensure that the classifier has a high degree of generalization. The decision trees are another approachable and simple-to-understand algorithm for feature selection (DT). While generating the decision tree, DTs can select features [80].

#### **2.4.2.2 Extraction of Features**

Feature extraction is the process of generating a new set of dimensions,  $d$ , which is frequently a combination of features from the  $D$  dimensions. Depending on the nature of the output data, feature extraction can be performed using supervised or unsupervised methods. Principal component analysis [98] is a popular technique for unsupervised feature extraction because it maximizes the variance between sample points in order to form clusters of data instances.

Additional useful linear projection methods under unsupervised techniques include factor analysis [99], which is useful for identifying latent variables in data, and multidimensional scaling[100], which provides a visual representation of the degree of similarity between data instances in Cartesian space.

As mentioned previously, the features generated as a result of these processes must be sufficiently discriminative and informative to train a good classifier. The performance of the classifier is proportional to the quality of the features used[87]. Despite this, there is no universal feature set for everything. Thus, it is critical to investigate the salient features in each classification task to assist the classifier in categorizing the documents meaningfully, especially when using supervised machine learning.

Several popular features in the literature have been demonstrated to be significant with varying degrees of success in a variety of classification tasks. These include the following:

- Psychosocial characteristics include emotions, passion, prejudice, and othering.
- Lexical, syntactical, and stylistic characteristics
- Application-specific (software) features Geographic location; preferences; and mentions, These features can generally be classified into two broad categories: surface-level features, which we refer to as high-level features in our study, and deep-level features, which we refer to as low-level features in our study.

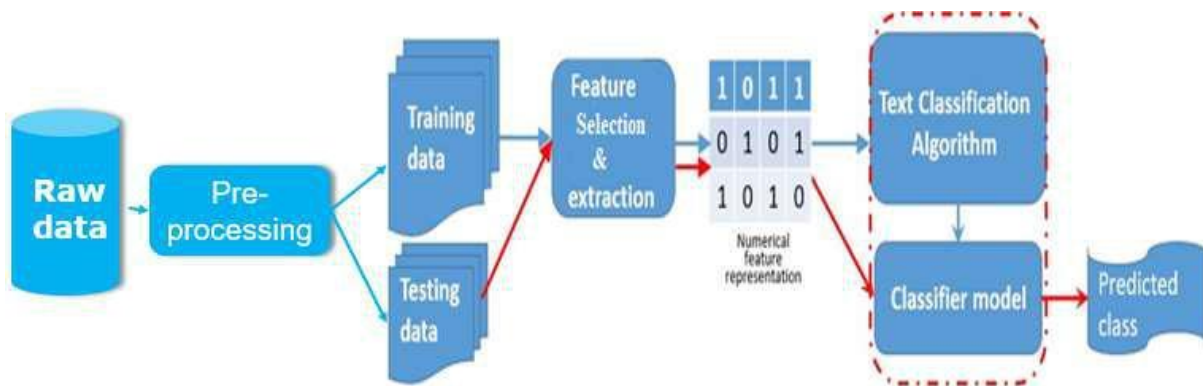


Figure 2.12: The Supervised Machine learning model

## 2.5 Previous Hate Speech Studies’ Methodologies and Features

Computational methods for detecting hate speech have been studied in the pre-social media era. However, this study focuses on some of the most recent studies published within the last seven

years before the publication of this report. This includes the UMATI project[75], which used a framework for identifying dangerous speech in social media to identify hate speech.

Saleem et al. [101] propose a method based on the self-identification of hate communities. Their approach presupposes that all content produced by these types of organizations is hate speech. They assume that the content is not annotated and can be used to train the classifier in this manner. Ross et al. [32] place a premium on the annotation process and its critical role in generating reliable data for training a hate speech detection system. They collected approximately 13k tweets over two months using obvious hate speech tags. Around 500 hate speech tweets were reliably obtained from these.

Waseem et al. [10] develop rules for annotating a corpus of 16k tweets using critical race theory. Numerous features, such as character n-grams, gender, geographic location, and word-length distribution, are evaluated for their efficacy in improving hate speech detection. They report that character-level n-gram features had a significant effect, whereas demographic features, except for gender, did not affect performance.

Paula [31] uses a set of rules to manually annotate a Portuguese dataset culled from Twitter and trains a classifier to detect hate speech using n-gram features. The paper documents the majority of existing research in hate speech detection that utilizes text mining features and hate speech-specific features, such as "othering" discourse, through a systematic review of the literature.

Davidson et al. [54] develop a multi-class classifier for discriminating between hate speech and other forms of offensive language. They were able to collect and annotate tweets as hate speech, offensive language, or neither using a crowd-sourcing approach. n-grams plus TF-IDF, POS tags, binary and count indicators for hashtags, mentions, retweets, URLs, and tweet-length were all used in this study.

Mondal et al. [41] develop a methodology for identifying and quantifying hate speech in online social media using pattern recognition. For instance, as demonstrated by their word tree, the high frequency of "I" and "They" collocations was indicative of hate speech. Whisper and Twitter were used to collect data for their experiment.

Burnap and Williams [34] create a data-driven blended model capable of automatically detecting cyberhate in Twitter data by identifying co-occurring tokens and the classification categories to which they are likely to belong. They use the Crowd Flower service to have their training data annotated by humans. Their blended model is said to be sufficiently robust to classify a variety of

hate targets, such as race, sexual orientation, and disability. This blended model is composed of three major components: a Bag of Words, a lexicon of hateful terms, and typed dependencies that detect "othering discourse" by capturing the grammatical and syntactic relationships between words. Burnap and William [12] also used trigrams to identify features such as "othering" and incitement. They discovered that the most efficient feature set for classifying cyberhate was a combination of n-gram typed dependencies and n-gram hateful terms in their experiment. Gitari et al. [24] employ subjectivity detection, a technique frequently used in sentiment analysis, to detect and rate hate speech sentiments in web forums and blogs on three levels, namely strongly dislike, weakly dislike, and no dislike. The two primary features used to train their classifier for hate speech detection were dictionary and corpus-based features pertaining to the semantic classes of race, nationality, and religion. Their method begins with a rule-based approach for extracting subjective sentences from a corpus. Second, subjective and semantic features of words are extracted from step one to construct a hate lexicon. Thirdly, using bootstrapping techniques, hate-related verbs and noun patterns pertaining to race, nationality, and religion are extracted and added to the lexicon created in step 2. Finally, a classifier was trained and evaluated for detecting hate speech in documents using these features.

Warner and Julia [22] use a supervised approach to train an SVM classifier to detect anti-Semitic speech in online text using three main features: word n-grams, brown clusters, and word frequency occurrence in a 10-word window. They used Yarowsky's [102] template-based strategy to extract the features that improved classification accuracy by 94%, precision by 68%, recall by 60%, and an F1 measure of 0.6375. Their experiment with binary features, in which they assigned a value of -1 to negative features and +1 to positive features, revealed no effect. Additionally, adjusting the SVM soft margin parameter had little effect.

Lozano et al. [74] were able to identify patterns and detect racist users and content using the Twitter following/mention networks. Their experiment on approximately 84k unique Twitter users identified racist users using two primary metrics: a sentiment word count and a racist score. In essence, a user was labeled racist if their tweet received a high negativity score and contained racially charged language.

Badjatiya et al. [103] use deep learning to address the problem of detecting hate speech in tweets. This was a problem of multiple classes, with three categories, namely racist, sexist, or neither. The researchers conducted experiments using 16k annotated tweets to train deep neural networks,

including Convolutional Neural Networks (CNN), Long Short-Term Memory Networks (LSTMs), and FastText. Additionally, experiments were conducted using conventional classifiers such as logistic regression, random forest, Support Vector Machine (SVM), and Gradient Boosted Decision Trees (GBDTs), all of which included embeddings. Additionally, character n-grams, TF-IDF vectors, and BoW vectors were used in the baseline methods. They report that semantic word embeddings learned with GBDTs outperform char and word n-grams in terms of accuracy by 18 F1 points.

Their research made two significant contributions: it demonstrated the successful application of semantic embeddings such as char n-grams, word TFIDF values, BoW vectors, and task-specific embeddings learned using FastText, CNN, and LSTMs.

Waseem and Hovy [10] discuss the difficulty of identifying hate speech in tweets, particularly racist and sexist remarks. Additionally, this was a multi-class problem. The study addressed a critical gap by utilizing demographic data as a predictive feature to aid in the improvement of the hate speech classifier's accuracy. Generally, the study examined the effect of various extra-linguistic features such as gender, length, and location on a classifier trained using character n-grams. They describe the process of developing an annotation scheme comprised of eleven steps based on critical race theory that was used to annotate 16k tweets. As a result, they were able to identify a few prolific users and frequently occurring terms in tweets containing hate speech, as well as references to specific entities such as hashtags. The study made three significant contributions: it developed a hate speech annotation scheme with a detailed data preprocessing methodology, it annotated a dataset of 16k tweets, it investigated the most predictive hate speech features, and it discovered that geographic and word-length distribution had no significant effect on performance and rarely improved performance over character-level features. They discovered that only the gender feature had a marginally positive effect on performance out of the several user description features they tested. Additionally, n-grams were used. To determine the degree to which the features were informative, the model's coefficients for each feature were added together over the tenfold cross-validation period. In general, demographic information had little effect on the classifier's performance.

Table 2.4 summarizes these studies and others on related topics such as cyberbullying and offensive language.



Table 2.4: A Summary of the reviewed hate speech studies and similar studies

| Research Area                  | Authors  |
|--------------------------------|--|
| General Hate Speech            | Fortuna, 2017; Badjatiya & Varma,2017; Davidson et al., 2017; Mondal et.al,2017 ; Schmidt A,Wiegand M,2017; Burnap & Williams, 2016; Silva et al.,2016; Gagliardone et.al,2015 ; Gitari et al., 2015; Djuric, et al.,2015; Kwok & Wang, 2013; Blarcum et.al,2005 |
| Racism                         | Lozano et.al, 2017; Tulkens, et al 2016; Cisneros & Nakayama, 2015; Chaudhry, 2015; Djuric et.al, 2015; Kwok, Wang, 2013; Greevy, Smeaton, 2004  |
| Anti-Semitism                  | Warner w,Hirschberg,2012   |
| Religion                       | Burnap & Williams, 2015  |
| Cyberbullying                  | Van Hee et al.,2015; Dadvar 2014 Dinakar et al.,2012; Yin et al., 2009   |
| Abusive & Hostile Messages     | Nobata et.al,2016 Chen et al.,2012; Sood et al., 2012; Mahmud et.al,2008   |
| Offensive language             | Davidson et.al,2017, Chen et.al., 2012, Razavi et.al,2010, Xu & Zhu ,2010; Sood et.al, 2012 Mahmud et.al., 2008.   |
| Radicalization/Terrorism       | Correa d, Sureka A,2013, Smith et al,2008, Last et.al, 2006  |
| othering language              | Burnap & Williams, 2015; Alorainy et al.,2019  |
| Hate Group detection           | Ting et.al,2013  |
| Anti-social behavior detection | Douce ,2017; Cheng et.al,2015  |
| Fake News detection            | Popat et. al., 2018, Potthast et.al., 2018   |
| Authorship Detection           | Li et.al,2016 ; Zheng, et.al,2006  |

According to this study's review of previous work in automatic hate speech detection, ranging from Spertus's work in the late 1990s to 2019 publications, eight features consisting of both high- and low-level features have been widely employed. The n-gram feature is the most popular of these. This means that these characteristics have been examined in more than two empirical studies of hate speech in the past. This is illustrated in Figure 2.13.

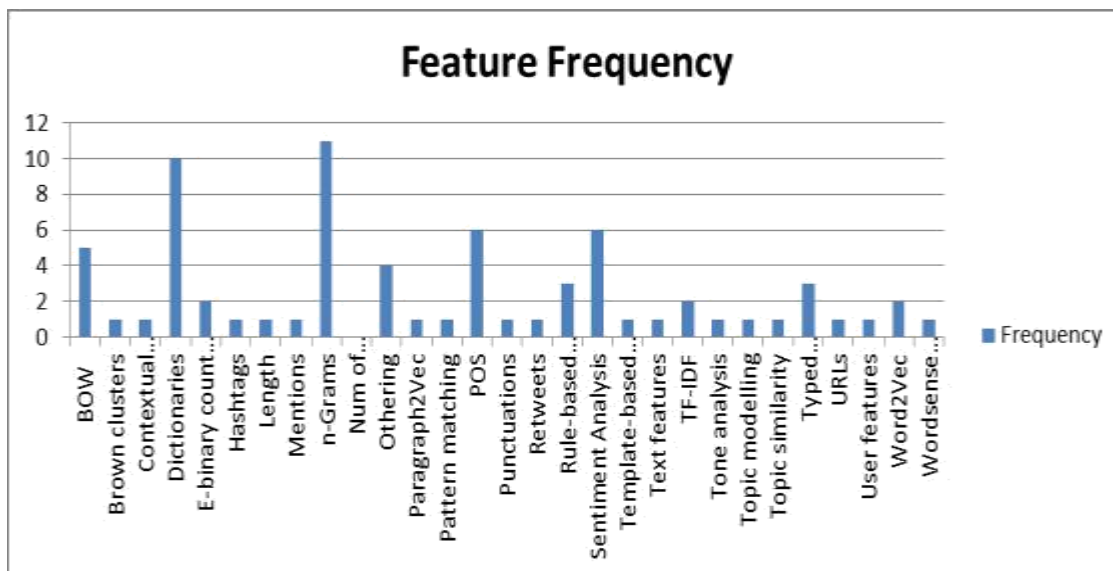


Figure 2.13: Feature frequency from previous hate speech studies

Individual studies have found varying levels of success based on the characteristics discussed above. BoW is not recommended as a primary feature if the context is critical because the word order is not preserved and each word is given equal weight. When some of the words appear frequently across classes, the classifier will not be able to discriminate properly between them, resulting in misclassification due to a significant number of false positives. N-grams, particularly bigrams, have been shown to have considerable performance effects when employed as features because they tend to keep context. Additionally, they work effectively when paired with other features. Typed dependencies as features have also been observed to capture syntactic grammatical ties between words in a message well, making them good classification features. Word embeddings, in which each word in the vocabulary is turned into a real-value vector in a preset vector space, with comparable meaning words having similar representations, have become a popular feature in recent studies. Word embeddings are widely utilized in deep learning and have already been used to detect hate speech in several research. Djuric et al., for example, used comment embeddings with paragraph2vec to improve their classifier's performance in terms of training time and memory requirements. Word embeddings provide a number of advantages over one-hot encoding techniques, including the use of a densely distributed representation for each word, as opposed to the sparse word representation used by the latter [104]. Furthermore, the number of traits is frequently less than the size of the lexicon.

### 2.5.1 Cross-Industry Standard Methodology for Data Mining (CRISP-DM)

CRISP-DM is a non-proprietary, open-source, and widely used standard process approach for data mining and other initiatives. The next competing methodology, Sample, Explore, Modify, Model, and Assess (SEMMA)[105], is also more extensive in terms of its steps. Unlike SEMMA, which has five phases, CRISP-DM includes six iterative and interrelated phases that begin with business understanding, which is renamed to problem understanding in this study for clarity in the context of the hate speech domain. Each phase is made up of tasks that are further subdivided into specific subtasks, resulting in a hierarchical structure with four levels of abstraction. The first level is the most abstract, with six general phases: problem identification, data identification, data preparation, modeling, evaluation, and deployment. The second level comprises of generic tasks that are accommodating or general enough to cover any machine learning issue domain under each phase. For example, regardless of the domain, data exploration is a general duty during the data understanding phase. For each general task, the third level comprises of specific tasks. Data visualization, for example, will be a distinct task under data exploration. The process instances, which are records or results of certain tasks, make up the last level. A pie chart or histogram depicting numerous data frequency units, for example, can be the result of data visualization. Figure 2.14 depicts a summary of the hierarchical design in detail.

The methodology was deemed to be proportional to the depth required to complete the objectives of this study based on these distinct, iterative, and precise phases in CRISP-DM, and so became the methodology on which this study built.

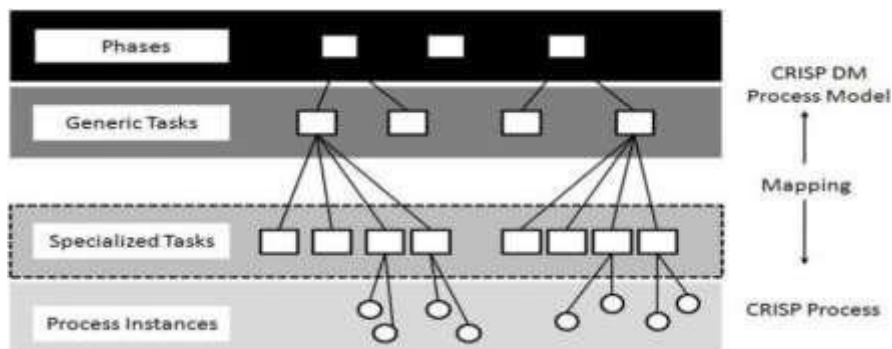


Figure 2.14: Four-level breakdown of CRISP-DM methodology [119]

## 2.6 Text classification High-Level Features

These are mostly qualitative concepts in the text message that is legible by humans. These characteristics can be extracted or detected directly from the text document and utilized to help a human annotator decide which class or group the documents belong to. A machine classifier, on the other hand, cannot use these features directly. They must first be translated into a machine-learning-friendly format. The high-level features can be classified into psychosocial features, general linguistic features, and application-specific features based on a review of past work on computerized hate speech detection. Figure 2.15 shows how this works.

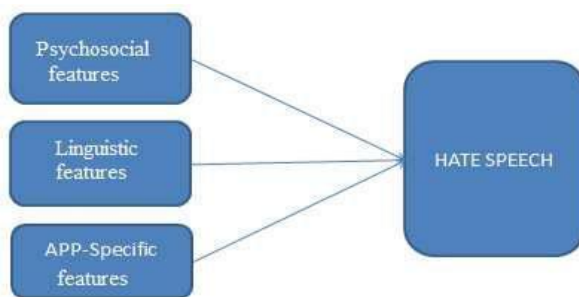


Figure 2.15: High-level feature categories

### 2.6.1 Psychosocial features

The psychosocial elements are made up of three major concepts that served as a conceptual framework or lens through which any communication could be assessed for hate speech on a qualitative level. The three dimensions of hate explained in the triangular theory of hate influenced these ideas the most. They include the denial of closeness (distance), desire, and hatred commitment. Figure 2.16 illustrates them.

#### 2.6.1.1 Distance

This is the psychological distance produced when a person is viewed as a member of the out-group rather than an individual, whereas members of the in-group are seen as individuals. The concept of distance is the polar opposite of closeness or intimacy. It is characterized by stratification and exclusivity, in which an individual or target group is viewed as distinct, eliciting feelings of hatred, disgust, or aversion. Other categories used to describe distance in prior studies include "othering," in-group against out-group, us versus them, insider versus outsider, pure versus impure, and pure versus impure[12], [77],[76], [75].

Othering speech is defined by concepts such as "us" versus "them," superiority and inferiority, in-group and out-group within social groups, and superiority and inferiority.

The use of "othering discourse" in text messages as a factor in recognizing racism [40] and antisocial conduct has been used in several earlier research in the domain of automatic hate speech detection.

Distancing is also obvious in situations where one social group assumes 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.

#### **2.6.1.2 Passion**

The passion component is defined by powerful feelings of hatred, fear, and animosity directed at the target individual or group. Expletives such as curse words, obscenities, abusive, disparaging, and other objectionable expressions can also be used to communicate these feelings [26, 106, 107, 26]. If not controlled, these feelings can easily devolve into instigation and violence against an individual or a group as a result of belonging to a protected social feature.

#### **2.6.1.3 Commitment**

The triangle of hate theory [33] explains this attribute of hatred. Long-held views or culture about a certain social group have a tremendous influence on the commitment or decision to hate. Cultural indoctrination or prejudiced beliefs that become the dominant ideology often inform an individual's or group's biased attitude toward another. This is characterized by devaluation and degradation of the target group, such as seeing the target group as subhuman or comparing them to animals, insects, or objects [36], [108].

The use of disparaging words that refer to a group of people using animal or insect language, such as maggots, cockroaches, rats, and so on, is a prevalent feature of devaluation.



Figure 2.16: Psychosocial features

Table 2.5 provides a summary of the psychosocial features employed in prior studies, including the psychosocial feature, its description, and particular examples of the feature.

Table 2.5: A summary of high-level features used in previous studies

| Feature                                      | Description  | Example   | Source  |
|--|--|---|---|
| Othering/<br>Social categorization           | Us versus Them language<br>In-group vs Out-group<br>Strategic Pronoun use, Predicate use (signaling proximity and distance )                                 | Immigrant, Alien, foreigners. "send them home"  | Burnap & William, 2016; Coupland,2010;Semin,2009          |
| Dehumanization (degradation and humiliation) | the process of depriving a person or group of positive human qualities. Use of metaphors, insect or animal names. Used to vilify the target.                 | Animalistic: Childlike, immature, backwards, maggots, cockroaches<br>Mechanistic (labeled as inert, cold, rigid, passive, fungible, or superficial) | Haslam (2006)   |
| Stereotyping                                 | Societal beliefs, cultural meanings or prior assumption or preconception about a certain social group or individual preconceptions about the targeted groups | Superior, Pure, Immoral, Evil, lazy, white, black.  | W. Warner and H. Julia,2012<br>Kwok & Wang, 2013          |
| Faulty Arguments                             | A person or group that is made to bear blame for others.   | Shift of responsibility Intense blaming and scapegoating  | Benesch, 2012   |
| Inciting violence                            | Call to action   | Accusation and Threats  | Benesch, 2012   |
| Offensive language:                          | Offensive, insulting, abusive language<br>Pejoratives,Profanities and obscenities,   | Bitch, fuck, dick, stupid, foolish  | Kwok & Wang, 2013<br>Xiang et al,2012<br>Chen et al, 2012 |
| Accusation and Threats                       | Reference to painful social or historical context  | Land allocation, water resources, job opportunities   | Kwok & Wang, 2013<br>Benesch                              |
| Powerful speaker                             | Powerful speaker with a high degree of influence over the audience   | Politician, cultural and religious figures, celebrities/entertainers  | Benesch,2012  |

## 2.6.2 Linguistic features

Previous hate speech research has found five key language characteristics. Lexical, syntactic, semantic, stylistic, and knowledge-based aspects are among them. They are the most obvious characteristics that can be extracted from the raw text. Figure 2.17 depicts these.

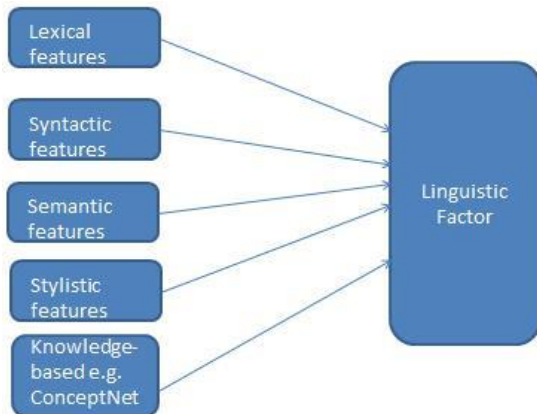


Figure 2.17: Linguistic features

### 2.6.2.1 Lexical features

In prior investigations, the most common features in the automatic identification of hate speech were lexical terms, word lists, and dictionaries. These include the use of accusatory and attributional terminology [12], [27], abusive language [35], insults or fires [109],[25],[110], and unpleasant language [54], [26], [107], including racial statements [40],[111].

Unit features such as n-grams are then broken down into lexical features (both character and word n-grams). Unigrams, bigrams, trigrams, and collocations are all included in the term n-gram. Bags of words (BoWs) and dictionaries are further lexical features that typically result in a high recall value, as indicated in earlier machine learning experiments including hate speech classification [112]. Figure 2.18 is a good example of this.

One drawback of dictionaries and the BoWs characteristics is that they frequently have low precision due to a high number of false positives. The inclusion of a hate speech or offensive term in a text message will automatically classify it as hate speech, regardless of how the term is used in context [12], [23], or [54]. As a result, N-grams have been employed to preserve context and have proven to be superior to BoWs [35], [113].

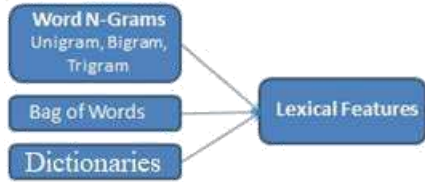


Figure 2.18: Lexical Features

### 2.6.2.2 Syntactic features

Patterns in sentence structure are examples of syntactic characteristics. Parts of speech (PoS) notations, which show nouns, verbs, and adjectives, as well as parse structures, punctuation frequencies, and unique syntactic structures, are just a few examples. PoS has already been utilized as a feature, but not as a strong independent one, to detect hate speech-related word categories such as pronouns [54], [114]. Figure 2.19 depicts this.

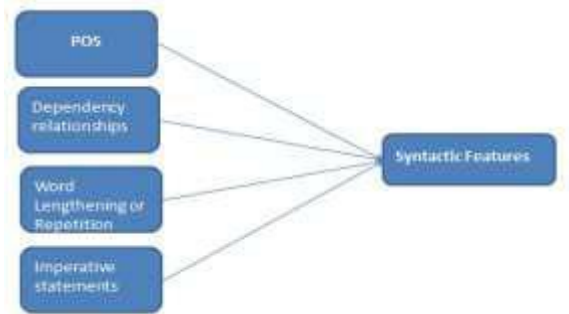


Figure 2.19: Syntactic features

### 2.6.2.3 Stylistic features

Capitalization, exclamation marks, emoticons, certain word categories or lexicon, average word and sentence length, including total amount of characters, and punctuation mark frequencies are all stylistic elements. PoS tag frequencies and punctuation mark frequencies have also been employed as stylistic features in prior investigations [40]. Figure 2.20 depicts these.

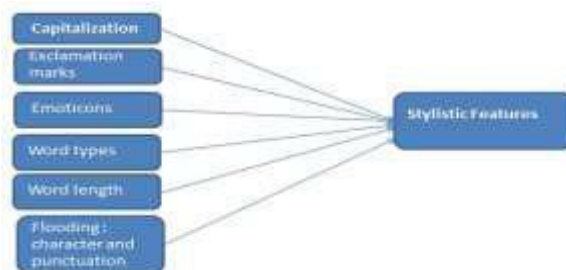


Figure 2.20: Stylistic features



### 2.6.2.4 Semantic Features

The semantic meaning of concepts or words is indicated by semantic characteristics. Associational terms, subject nouns, hate verbs, and negative polarity are examples of semantic traits that have been used to classify hate speech. Figure 2.21 depicts these.

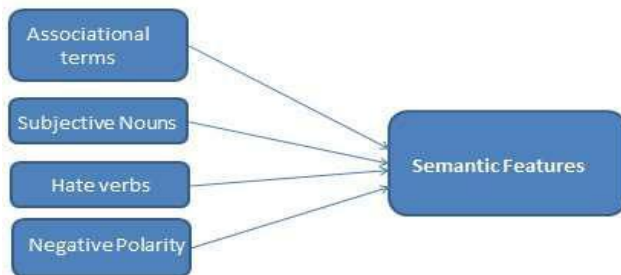


Figure 2.21: Semantic Features

### 2.6.2.5 Knowledge-based Features

There are distinct types of words that are important in each knowledge domain. This is critical when creating a corpus to train and evaluate a classification system because it has a significant impact on its performance. For example, a corpus compiled from social media data during an election campaign will almost certainly contain a higher frequency of phrases or attitudes related to current events than one compiled during the Olympics, which will almost certainly contain more keywords related to sports and athletics. Figure 2.22 illustrates this.

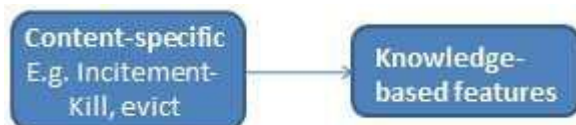


Figure 2.22: Knowledge-based Features

### 2.6.3 Other Features

Gender, locality, popularity level, and ethnicity of the message's author are examples of factors that aren't related to specific linguistic traits. Other characteristics are more dependent on the capabilities of the social media site in question. App-specific features are elements that are unique to each software application and have been demonstrated to contribute to the spread of hate speech. One can, for example, retweet or use hashtags connected to hate speech on Twitter. Furthermore, some studies have found that the type of URL in a tweet, the number of mentions or followers, the length of a tweet, geolocation, and user demographics can all be indicators of hate. Figure 2.23

depicts these elements. All of the features discussed above have been used in past studies, with varying degrees of success, in other sectors.

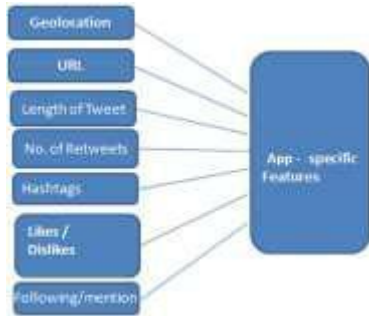


Figure 2.23: Other features

The psychosocial, linguistic, and App-specific levels can be abstracted from a list of the numerous high-level variables linked with hate speech identification documented in the literature. Figure 2.24 depicts this.

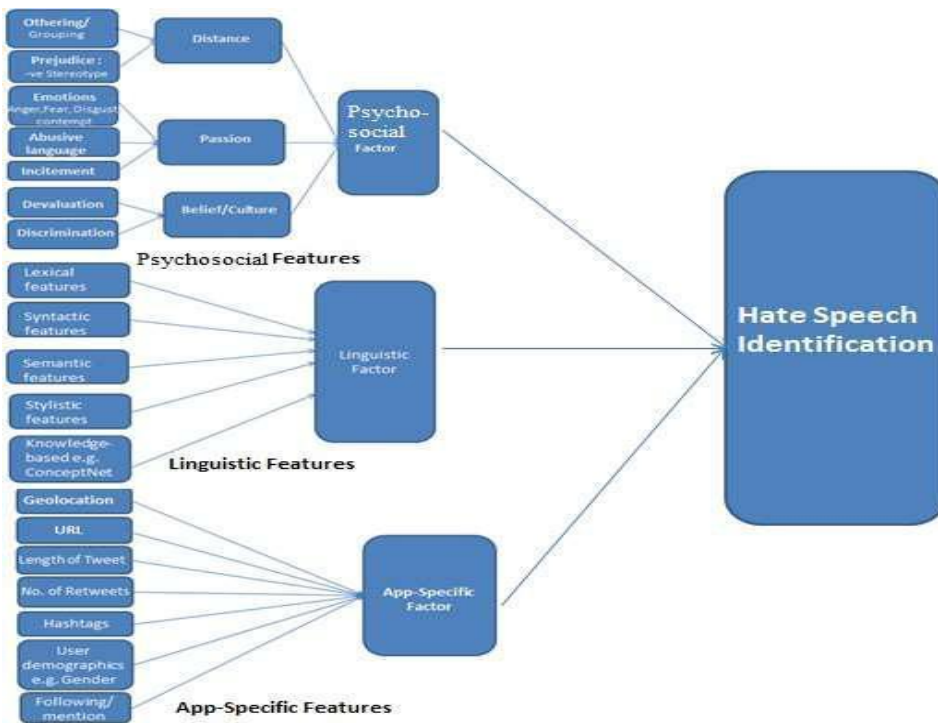


Figure 2.24: High-level feature abstraction

As a result, the same feature might be detected differently depending on the abstraction level, increasing the possibility of feature overlap, which is useful for collecting latent hate speech traits that frequently escape detection using conventional feature selection approaches. For instance, the

presence of pronoun dichotomies might convey a sense of separation. These pronouns are linguistic traits that can be recorded automatically by a text tagger that recognizes parts of speech. However, based on the annotation exercise conducted in this study, the most intuitive feature for training human annotators to identify hate speech was psychosocial concepts, such as identifying negative passion, stereotyping, devaluation, and distancing language in a message as potential hate speech markers or features.

## 2.7 Low-level Features for Text Classification

To categorize some text input into predetermined categories, any classification system can use these features directly. The feature extraction representation and the numeric feature representation are used to further categorize these features. Text characteristics are extracted using a specific approach, which is represented in the feature extraction representation. BoW, n-grams, and word embedding are some of the most prevalent approaches for extracting features. Numeric representation is the lowest degree of feature abstraction. A machine-learning algorithm can use these features directly to train a model, because they can be processed by machines. One-hot vector encodings, term frequency-inverse document frequency, and dense vectors are all examples of this type of data. Word embeddings are represented as dense vectors and the BoW is represented as an array of one-hot encoders. Figure 2.25 shows a summary of the mappings.

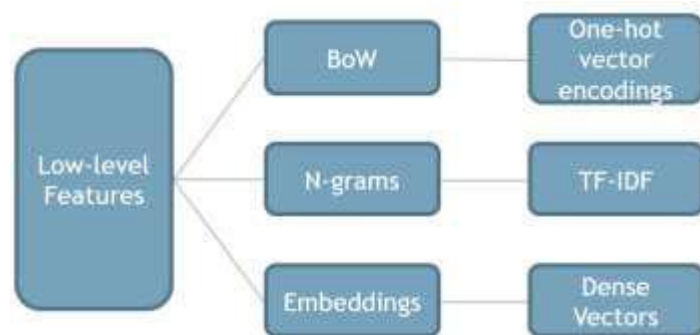


Figure 2.25: Low-level features

## Term Frequency-Inverse Document Frequency

The TF-IDF is a score that indicates how important a phrase in a document is over the entire corpus. It's a compound term that combines two concepts: the term frequency and the Inverse Document Frequency. The term frequency refers to the number of times a word appears in a document relative to the overall number of words.

$$TF-IDF = \frac{TF}{IDF}$$

The log of the number of documents in the corpus divided by the number of documents containing the word yields the Inverse Document Frequency..

Input documents can be handled at three levels: phrase level (n-grams), word level (words), and character level (characters). As a result, each of these input tokens' levels can be translated into their TF-IDF feature vectors. A matrix will be produced at the n-gram level to represent the n-gram scores. A matrix will be created at the word level to represent the scores of each phrase in the various publications. A matrix will be created at the character level to reflect the TF-IDF scores of n-gram characters in the dataset.

## Word Embeddings

The technique of encoding text documents or words as dense vectors is known as word embedding. Word embeddings, unlike Bow, may learn the word order in the vector space from nearby words in the original text message. Transfer learning with pre-trained word embeddings or training word embeddings from scratch using the input dataset are the two main methods for generating word embeddings. Word2Vec, GloVe, and FastText are some of the most popular pre-trained word embeddings.

Table 2.6 summarizes other features utilized in prior studies that occasionally overlap with the ones mentioned above.

Table 2.6: Other features used in previous studies

| Feature description           | Feature Example   | Author  |
|-------------------------------|---|---|
| Hate speech specific features | Othering language, Race, Behavior, Physical, Sexual orientation, Class, Gender, Ethnicity, Disability, Religion, and “other.” | Burnap and William, 2016; Burnap and William, 2015; Van Dijk, 2002; Coupland, 2010; UMATI project |
|                               | Speaker Perpetrator Characteristics;  | Burnap & William, 2016,   |
|                               | Objectivity Subjectivity of the language,   | Gitari et.al,2015; Warner & Hirschberg,2012   |
|                               | Declarations of the superiority of the in-group   | Warner & Hirschberg,2012  |
|                               | Focus on particular stereotypes e.g. Race, Class, Gender, Ethnicity   | Warner & Hirschberg,2012; Silva et.al,2016  |
|                               | Intersectionism of oppression   | Burnap & William, 2016  |
| Trigger Events                | major incidents: ‘trigger’ events, e.g., presidential elections   | King & Sutton, 2013<br>Thelwall, Buckley, & Paltogou, 2011<br>Burnap & William, 2015              |

## 2.8 Research Framework

This section explains the study's theoretical framework and the resulting conceptual framework.

### 2.8.1 Theoretical Framework

To identify hate speech in short text messages, this framework uses psychological observations, linguistic, and non-linguistic elements. The framework looks for hate speech phrases and the context in which they are employed, as well as condescending syntactic structures or patterns, such as imperative assertions. Table 2.7 summarizes four general ideas derived from psychological factors used in prior investigations, as well as the theories that underpin them.

Table 2.7: Constructs from qualitative research on high-level features

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

Table 2.8 summarizes the theoretical framework, with columns indicating the overall concept, specific construct, brief description, supporting theory, and past investigations that have documented the application of the specific features.

Table 2.8: Theoretical framework

| Category                         | General Concept    | Specific construct              | Description   | Theory                                       |
|----------------------------------|--------------------|---------------------------------|---|--|
| Psychosocial High-level Features | Distancing         | Othering                        | Dichotomies of In-group Vs Out-group; Us vs Them.                                       | Self-categorization theory                   |
|                                  |                    | Prejudice                       | Implicit biases, discriminatory attitude, Anti-group                                    | Social Identity theory                       |
|                                  | Passion            | Negative Polarity or sentiments | Emotions of anger, fear, disgust, and contempt towards the target group. Negation words | Integrated threat theory                     |
|                                  |                    | Offensive language              | Insults, curses, Abuses, derogatory   |  |
|                                  |                    | Incitement                      | Call-to-action to harm the target   | Speech act theory                            |
|                                  | Commitment to hate | Devaluation                     | Comparison of humans to animals, insects, or objects                                    | Susan Benesch Framework for dangerous speech |
|                                  | Hate as a story    | Stereotyping                    | Negative attitude towards the target  | Baumeister's theory (revenge)                |
|                                  |                    | Prejudice Fault Arguments       | Implicit biases Referencing painful historical/social issues                            | Social Identity theory                       |

### 2.8.2 Conceptual Framework

In research, a conceptual framework is a visual representation of the links between variables or concepts in a network structure with a clear indication of the dependent, independent, and intermediary variables, based on existing theories. It is a crucial tool for determining precisely what will be researched in the study. The conceptual underpinning for this thesis will be a schematic explanation of hate speech with a clear representation of the numerous components or characteristics of hate speech that are frequently visible in brief text messages.

The hate triangle in Figure 2.1 represents the triangular theory of hate [33], which has three basic dimensions: distance, passion, and commitment. These aspects were chosen as the core variables for empirically developing the comprehensive hate speech framework in Figure 2.26 throughout the earlier stages of the investigation. The wrapper method in feature selection for machine learning, as well as the necessity to keep feature dimensionality low, an ideal in machine learning practice, drove this strategy to systematically add features one by one. Furthermore, text feature sets have been demonstrated to increase exponentially, resulting in the dimensionality curse in machine learning [104]. In addition, unlike the other theories analyzed, the triangle theory of hate was the most complete and expounded hate from numerous aspects, the majority of which encompassed the concepts in the other theories reviewed, as stated in table 2.8. As a result, the hypothesis was deemed the most appropriate because it provided the greatest explanatory power for the occurrence of hatred. Furthermore, it is clear from the numerous definitions that hate speech has a definite target; otherwise, the message will be classified as offensive. As a result, the three characteristics of distance, negative passion, and dedication to hatred effortlessly transition to the original qualities or variables that will determine whether a communication can be categorized as hate speech, offensive, or neither.

It was also found that the mere existence of one idea could not sufficiently distinguish a message into the positive class, i.e., hate speech, based on empirical results from the early trials and qualitative analysis on sample hate speech messages classified by human annotators. The presence of pronoun dichotomies in a message, for example, could suggest the concept of distancing language, whereas the presence of terms alluding to negative sentiments or unpleasant language could imply negative passion. These notions, on their own, could not classify a message as hate speech. "*We will not accept hawa wasee treating us like crap kwa nchi yetu,*" for example. The message contains pronoun dichotomies, such as "*We,*" "*hawa,*" and "*us,*" as well as an

inflammatory phrase, "crap." For a human annotator, it is clear that the message's author is irritated, and the message exudes negative passion. However, the target of the hatred, i.e., "wasee," is unclear and cannot be identified as belonging to a protected characteristic such as ethnicity, nationality, religion, or other factors that would allow hate speech to be positively identified. As a result, the concept of distancing language, which is best represented by pronoun dichotomies, has to be qualified further by clearly establishing the target subject as a member of a protected social group. The concept of stereotyping, as suggested by the team of human annotators (see appendix D), and extended in Baumeister's retribution theory, best captured this. Furthermore, the team of human annotators pointed out that the first three ideas fall short of fully conveying the concept of prejudice or propaganda directed at individuals or groups who have a common social trait. As a result, these two additional concepts, combined with the idea of concept intersection, helped to develop the multidimensional framework reported in Table 2.9 and illustrated by the Venn diagram in Figure 2.26, which had undergone a rigorous qualitative study including human annotators.

The five variables will be measured by their respective term frequencies-inverse document frequencies in terms of operationalization (TF-IDF). The specific variables under each concept were mostly selected from the LIWC2015 dictionary's set of emotional, cognitive, and psychological word lists [70]. Table 2.9 shows the details.



Table 2.9: The summary of the concepts

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

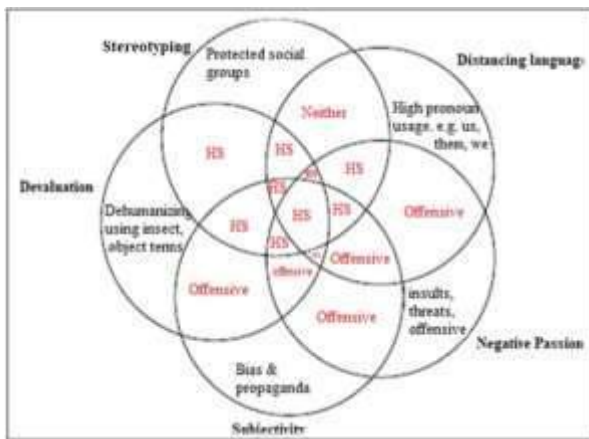


Figure 2.26: Multidimensional Hate Speech Conceptual Framework

There are 10 instances in which a text message will be labeled hate speech according to the multidimensional hate speech conceptual framework. The first nine idea combinations link to a protected social group explicitly, however, the tenth is a reference that is often veiled by a devaluation term that is only known by in-group members. Five have a four-concept overlap, three have a three-concept overlap, one has a five-concept overlap, and one has a two-concept overlap. Hate speech is defined as the use of insults or harsh language in combination with a protected trait. Table 2.10 summarizes this as well as additional idea combinations.

Table 2.10: Multidimensionality of Hate Speech

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

### 2.8.3 Measurement using term frequency-inverse document frequency

Based on the five primary variables in the conceptual framework, TF-IDF was the primary indicator for the independent variables that are indicative of hate speech. The words required to be vectorized, or changed from high-level features to low-level numerical features like TF-IDF, in order to accomplish this. Because machine learning algorithms can only understand and analyze numerical features, this is the case. But, exactly, what is TF-IDF? In machine learning, the TF-IDF is a feature weighting factor. The term frequency (TF), which specifies how frequently a word appears in a message or document, is multiplied by the number of times the word appears over the entire corpus in the TF-IDF (IDF). Equation 1 illustrates this. In this regard, very common words such as determiners such as 'the' or 'is,' as well as any other stop words, are punished and ranked low because they do not provide useful information to aid the classifier in distinguishing between classes. Because they occur often across all categories, they contribute to the cacophony. However, if a word is only found in a single category or cluster, it is given a greater weight factor and is ranked highly.

As a result, given an annotated dataset in which each message, consisting of many words, is allocated to a class, the statistical model may transform high-level PDC characteristics to low-

level PDC features. Text vectorization and statistical inference are used by TF-IDF to learn and properly categorize fresh unseen messages into predefined categories or classes.

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D)$$

where

$$tf(t, d) = \log(1 + freq(t, d))$$

$$idf(t, D) = \log\left(\frac{N}{count(d \in D: t \in d)}\right)$$

Equation 1: *TF-IDF formula (Adopted from [115])*

## 2.9 Summary of Features in Text Classification

In general, the features employed for text classification play a critical role in defining the trained model's efficacy and accuracy in distinguishing across class instances. To inform the training of a machine classifier, the features must be found, examined, and the most important among them chosen. Following a study of the literature on hate speech identification, it is clear that past studies on hate speech classification used a variety of criteria. However, they are frequently muddled, which adds to the difficulty of comprehending them. This study divides these features into two basic groups, high-level features and low-level features, to theoretically and empirically break down their complexity. Human annotators can quickly understand and directly identify the high-level features. As shown in Figure 2.24's high-level feature abstraction architecture, these are further abstracted into psychological, linguistic, and App-specific aspects. This abstraction presents a novel methodology that catches latent traits, such as the "othering" language, which has previously been found to be useful in collecting subtle kinds of hate speech [78]. Furthermore, our research asserts, via the holistic hate speech conceptual framework in Figure 2.26, that these latent features are easily identifiable through psychosocial concepts, and that when combined according to the scheme in Table 2.10, they become informative features for positive identification of subtler forms of hate speech, which conventional methods, particularly supervised machine learning, were inadequate in capturing.

Furthermore, these high-level text elements must be recorded, organized, and converted into a numeric representation suitable for machine learning and classification. As seen in Figure 2.25, these features are referred to as low-level features and are generally informed by text mining techniques. Previous research has found that these two major tiers of features are sometimes counted and treated as independent features, which adds to the difficulty of comprehension. The larger picture presented by this study is two-fold. To begin, the study splits text classification

features into high-level and low-level features based on the features' human vs machine understandability and interpretability. Second, the research demonstrates the relationship between important traits that make up the two tiers. High-level lexical characteristics, for example, are translated into n-grams and then into TF-IDF low-level features.

In conclusion, this research proposes a holistic framework that links high-level properties to low-level features that may be used to develop a machine learning model for social media data classification. Human annotators can label the data for supervised learning using the high-level characteristics, which are human comprehensible. Low-level characteristics, on the other hand, which are represented numerically, can be employed directly by the machine learning algorithm to train a model. Figure 2.27 depicts this.

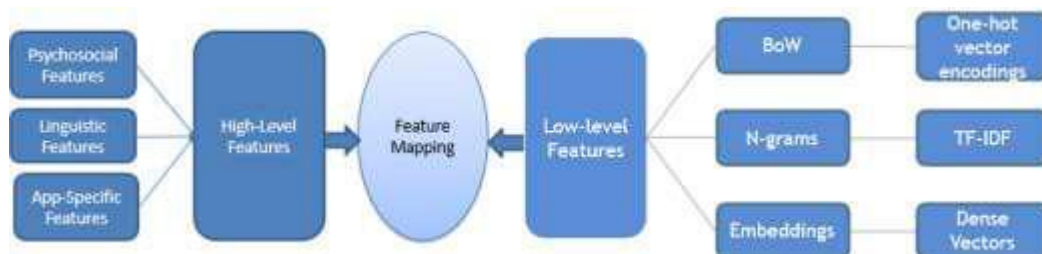


Figure 2.27: Hierarchical Feature framework

As a result, the primary goal is to identify techniques and approaches (low-level features) that best capture the high-level features of hate speech, or what the research considers to be the most important hate speech variables, such as othering, group solidarity, incitement, derogatory language use, and so on.

## 2.10 Summary

To identify nuanced kinds of hate speech, this study proposes a novel psychosocial feature set based on language use around the concepts of psychosocial distancing, negative passion, dedication to hatred, stereotyping, and hate as a tale. These ideas are well-founded in theory and detailed in section 2.8's conceptual framework. Furthermore, the study presents a simple and effective method for qualitatively identifying and analyzing hate speech in short text documents using human-readable high-level psychosocial features, namely PDC-based features, which can then be mapped to machine-readable lower-level features such as Term Frequency-Inverse Document Frequency (TF-IDF) and one-hot encoding vectors for training a machine classifier. Previous hate speech detection research has relied on lexical and other NLP-based features. These types of capabilities, on their own, will not be able to capture hate speech in codeswitched messages effectively. As a result, classifier models that explicitly use these traditional traits would underperform, resulting in a high number of false negatives, contrary to how hate is expressed in social media postings.

The psychosocial (PDC) aspects are intended to be beneficial in two ways. To begin, the feature set must be sufficiently informative to improve classification performance. Second, the PDC feature set is substantially smaller than traditional approaches that use the TF-IDF to represent the whole input lexicon. Unlike the sparse input vector of the general lexicon, this substantially lowers the sparseness and dimensionality of the original features, making PDC a great feature selection strategy with a dense input vector length. Furthermore, the efficiency of the PDC design as a qualitative feature selection method for codeswitched text categorization of nuanced kinds of hate speech will contribute to overall machine classification efforts.

## CHAPTER 3: RESEARCH METHODOLOGY

### 3.1 Research Methodology

The first section of this chapter defines the terms "research" and "research methodology." The research philosophy is discussed in the next section, section 3.1. Section 3.2 describes the research design in the following sections. The study process, including the methods used to construct the text categorization model, is covered in Section 3.3. The research validity and reliability are discussed in Section 3.4. The ethical research considerations are highlighted in section 3.5, and the chapter is summarized in section 3.6.

The term "research methodology" is made up of two words: research and methodology. The term 'research' is derived from the old French noun 'recherche' [116], which refers to a rigorous study conducted to discover and document new information about a subject. The term 'methodology' is derived from the Latin term 'methodologiae' [48], which means "way to progress" [117] or "clear path to achieve identical outcomes." As a result, the term "research methodology" will be used in this study to refer to a thorough investigation that specifies the sequence of repeatable activities that were used to investigate and establish facts and new knowledge about hate speech to improve machine learning performance in automatic text classification.

#### 3.1.1 Mixed Methods Research Methodology

The use of two research methodologies, either concurrently or in sequence, is part of a mixed methods research methodology. This might be done in parallel using triangulation, in which different methodologies are utilized at the same time to seek convergence in the results and therefore gain credibility. The output of one technique is used to enrich the understanding of the phenomena and as an input of the second method, which enriches the entire study. In addition, mixed methods can be utilized to uncover inconsistencies or new viewpoints in the results of another method. Credibility, context, example, utility, diversity of viewpoints, confirmation, and discovery are among the five characteristics of these arguments [118]. The use of qualitative and quantitative approaches to data gathering, analysis, and inference is a popular mixed method [119]. Furthermore, mixed methods are especially beneficial in the conclusion section, when the researcher legitimizes the study by addressing its research questions with validity and then demonstrating a contribution to the existing knowledge domain [120].

In this study, a mixed research method was used. First, using a qualitative approach, content analysis was utilized to determine the discriminant characteristics of hate speech phenomena by looking at pertinent hate theories in the literature, as summarized in Table 2.8. The framework was subsequently turned into a web tool, which nine human annotators used to categorize 48k short text messages (tweets) into one of three predetermined categories: Hate Speech, Offensive, or Neither.

Following that, a quantitative approach was employed to conduct text analysis in order to determine word frequencies by class.

The classifier model was then trained using other low-level variables such as TF-IDF and word frequency vectors.

Using the Jupyter notebook integrated development environment, all activities from data pretreatment to data exploration and analysis, feature processing, model training, and actual classification were handled in a consolidated manner. This was used to facilitate end-to-end model development and data visualization through the use of python programming (version 3.6.8) and machine learning libraries such as the natural language tool kit (NLTK) for data preprocessing, Pandas for seeing and doing various operations on data, Scikit-learn for various kinds of machine learning models, and Matplotlib for data plotting, among other libraries.

### **3.2 Research Philosophy**

A research philosophy, in general, is a lens through which a phenomenon is examined. The assumptions and procedures that will be employed in the gathering, processing, and analysis of data, as well as how knowledge will be created from the data, are all influenced by research philosophy [121].

Every science, whether natural sciences such as physical, biological, psychological, and geological sciences or artificial sciences such as architectural design, mathematics, engineering, and computer science, is based on a specific phenomenon, or domain of empirical reality. In this sense, computer science as an example of artificial science is based on autonomous computing as a class of reality[122]. The human mind has been the fundamental source of inspiration for computation: how the mind works, how it assesses, how it makes judgments, how it calculates, and how it stores and recalls events, activities, tasks, and things. As a result, the concept of

automatic computing and its diverse applications in the virtual world has continuously expanded from the time of Charles Babbage, an English and Mathematics professor credited with creating the foundational framework of a computer in the 19th century [123], to the present day. More and more better, faster, and complete techniques/methods for autonomously gathering, manipulating, transforming, storing, and retrieving data are being discovered. Unlike natural disciplines such as biology, chemistry, physics, and geology, which study the world as it is, computer science investigates the artificial world, or virtual world. Virtual items, events, tasks, and activities exist in the virtual world that are almost identical to, if not identical to, natural world objects. This is what we generally refer to as software on a computer.

So, in terms of computer science, what is reality? Is natural-world reality comparable to computer-science reality? Natural science, as previously said, is a science of "as is," whereas computer science is a science of both "as is" and "as ought to be." Natural science, for example, will look into the human mind and its diverse functions as they are, which is essentially an end in itself. Computer science, on the other hand, will not only analyze but also imitate human mental operations, notably in terms of signal input, processing, and storage. This is frequently done in order to optimize or reduce some objective function in relation to mental activities. This could include increasing compute capacity, memory, and other resources while lowering costs, biases, and other errors. This will be done in the context of computer science's software.

Furthermore, computer science contains a physical reality component in the form of computer hardware. This, like any other physical object, can be touched, felt, and smelt, and is subject to the physical laws of nature. We will concentrate on the computer science component of software that deals with completely abstract things in this research. As a result, since this thesis is about computer science and its importance, it will focus on the theoretical aspects of computer science relating to the abstract psychosocial phenomenon of hate speech as it occurs in the virtual world of social media networks. To summarize, the ultimate motivation is to mimic the human system's "hardware" and "software." Hardware will be every part of the human body that can be touched, such as hands and internal organs, whereas software will be the human mind's thought patterns, sentiments, attitude, and signal computational capabilities. Just as a human being's body is required for all of these 'invisible' aspects to function, the software aspect of computer science is also dependent on computer hardware to function.



So, what qualifies as computer science expertise? We believe that this is not an attempt to redefine computer science knowledge, nor is it an attempt to construct a new definition. However, a brief response is provided here for the goals of guiding this study and generating coherence in this thesis regarding the broader body of knowledge in computer science. Scientific knowledge, on the other hand, is the knowledge that is acquired systematically and can be shown with the same results when the process of investigation is repeated independently. Observation, reasoning, and experimenting are frequently used as methods of investigation. These strategies allow real-world events or things to be rationally explained in terms of existing ideas, conceptions, hypotheses, and other evidence [122].

The virtual objects and events under investigation, namely hate speech in text messages from social media during elections, will be rationally explained through the construction of hypotheses and the testing of various features and classifier models against the ground truth (validation set) through experimentation, observation, and reasoning. The study's main goal is to figure out how people express hate in text messages on social media, as well as which traits best reflect hate messages, to train a classifier that can automatically distinguish hate speech text from other texts on social media.

The epistemological pluralism approach was used to guide the study through various experiments in which different feature combinations and machine learning algorithms were investigated to determine the best features and ideal classification algorithm for detecting hate in a codeswitched text dataset.

Given the empirical character of this work, a positivist approach was used, which is a simpler approach to computing. Because it encourages the use of scientific procedures for knowledge generation that are replicable and generalizable, the positivist approach is a good fit for our research. Second, it encourages the search for causal linkages, which is critical in our research when it comes to classifying noisy text messages from social media. Furthermore, the classification job in this work uses statistical inferencing based on logic and mathematics to process and analyze the text data statistically, which is consistent with the positivist approach. As a result, the replicability may be measured and the validity can be tested.

### 3.3 Research Design

The road plan that places and guides the study in answering the research questions and achieving the research objectives is known as a research design. It aids in the proactive protection of research integrity by limiting or avoiding any potential sources of threats or biases that could endanger the credibility of the study's results and conclusions [124].

The methodologies utilized to address the research questions about problem description, data collecting, feature identification, model development, and model evaluation largely determined the research strategy used in this work. The underlying research problem was text classification, which produces mostly qualitative results. The research looked at the topic through the lens of computer science, which is generally slanted toward using mathematical approaches to automatically detect the underlying function in text corpora to produce a model. As a result, a mixed-method approach was utilized, which incorporated both qualitative and quantitative epistemologies and was principally guided by the study objectives. Content analysis was used to discover essential terms from multiple definitions of hate speech and existing hate theories, which then guided the development of the study's conceptual framework. The annotation technique for guiding the team of human raters to accurately categorize the messages with either of the three preset classes of hate speech, offensive, or neither was developed using the same architecture. Furthermore, the study's hypotheses were derived from qualitative observations made throughout the data exploration phase. To derive quantitative inferences from the annotated data and develop a classifier that intelligibly predicts new messages as belonging to the three established classes, statistical machine learning models were used. Both techniques were required to provide a comprehensive picture of the hate speech phenomena and to influence the study's experimental design in terms of identifying prominent elements and training a computer system to automatically detect hate speech in text messages.

The document, phrase, word, and character levels were the units of analysis and assessment. Each brief communication was considered as a document at the document level, especially since the length of a short text message on social media is limited to less than 150 characters. The n-gram word lengths 2, 3, and 4 were tested at the phrase level. Each term was regarded as a feature at the word level. N-gram character lengths of 2 to 7 were examined at the character level. For each level of analysis, the word and character frequencies were calculated.

Aside from that, there was an experimental study in which data was collected for a year, encompassing Kenya's 2017 national elections. Hate speech has been known to rise on social media during trigger events such as presidential campaign periods, which can extend several months before and after the official election results are announced. In Kenya, brief text messages potentially containing hate speech from social media were gathered throughout the presidential campaign season in August 2017, which also included a rerun election later that year.

Given the empirical nature of the study, which involves recognizing hate speech from the social media content, it was crucial to first have a strong knowledge of the hate speech phenomena, as informed by several hate theories from sociology and psychology detailed in section 2.2. The theoretical framework that influenced the annotation scheme and standards was built on these foundations.

Human annotators used a deductive technique to establish the classification category of each communication and label it, as instructed by the annotation framework [125] based on the three aspects of the triangular theory of hate [33]. Despite this, the study's assumptions were developed using an inductive technique based on the findings of a preliminary investigation. These were subsequently used as the foundation for the study's future experiments. Furthermore, the machine classifier used in this study is designed to learn from examples of classified text messages before using inductive inference to categorize fresh, unseen text messages. This inductive learning principle is essential in the development of any automated machine that uses prior knowledge from specific examples to progress to greater generalization while keeping high performance [126].

In addition, comprehensive literature analysis and Internet search for hate speech-related laws, rules, and user policies from major social media networks and periodicals were undertaken. This was done descriptively, focusing on the instance of Kenya concerning user-generated content bordering on hate speech on social media, which was sparked by the 2017 presidential elections. Keywords were extracted from these hate speech definitions using content analysis. Table 4.1 summarizes this nicely.

To collect messages, a variety of search tactics including key hate terms, pro-hate user accounts, and problematic hashtags were employed to create a hate speech dataset from the ground up.

Following that, experiments for the text classification task were conducted, in which statistical models in machine learning were used to build classifiers that make inferences from sample data in classifying codeswitched messages from social media into three predefined classes: hate speech, offensive, or neither.

### 3.4 Research Method

In data mining initiatives, two methodological frameworks are commonly utilized. These have been created as the industry standards for data mining initiatives, encompassing the guiding phases [127]. The Sample, Explore, Modify, Model, and Assess (SEMMA)[105] framework and the Cross-Industry Standard Processes for Data Mining (CRISP-DM) [128] are two examples. Although traditional Knowledge Discovery in Databases (KDD) precedes these two, data mining practitioners dislike it because of constraints such as the inability to learn automatically over time[129]. KDD and SEMMA have five steps in their processes, but CRISP-DM has six. Although the process stages in KDD and SEMMA have distinct names, they are comparable. The first process stage of KDD, for example, is the Selection stage, which entails creating a target dataset. This is similar to SEMMA's first stage, Sample, which comprises collecting a data sample from a larger dataset that contains informative properties to aid data mining. Preprocessing in KDD and Explore in SEMMA, for example, are comparable, as are the Transformation stage with the Modify stage, data mining stage with the model stage, and the Evaluation stage with the Assess stage, respectively.

The Cross-Industry Standard Processes for Data Mining [128] methodology best aligned itself to these to build and explore various features and models for hate speech identification in short text messages, based on the exploratory nature of this study's objectives, as well as the practicality, affordability, and accessibility to resources. Furthermore, unlike the traditional product-oriented software development cycle, the CRISP-DM is designed especially for exploratory data analytics research [129], which is the focus of this research. In summary, the approach for the study is divided into five parts and is based on industry-standard data mining procedures. Figure 3.1 summarizes and illustrates these points.



Figure 3.1: Research workflow

### 3.4.1 Problem understanding

This phase's purpose is to assist in the formulation of a problem statement by first attempting to comprehend the domain or environment in which the hate speech issue occurs. This allows you to stay focused on providing real solutions within the context of the problem or opportunity.

In this context, relevant literature was thoroughly researched in order to gain a thorough grasp of the hate speech phenomena as it manifests itself on social media in Kenya, as well as to investigate similar past studies. There was a review of both online and physical books and journals. Keywords were used in search engines like Google Scholar, university publication repositories, and other online publication databases as part of the online literature search approach. The cited publications that were referenced in the landmark research were looked up online using the snowballing technique.

Several hate theories were also investigated, as well as the influence of the problem on social media. Furthermore, a content analysis was performed on the hate theories descriptions, as well as the various definitions of hate speech provided by social media network companies on their respective user-content guidelines webpages, and legal definitions derived from constitutions or related hate speech regulations in some countries, with Kenya serving as a case study. Furthermore, the researcher's interactions with officials from NCIC, the government agency in charge of hate speech issues [130], and KENET, the principal Internet Service Provider for all higher learning institutions in Kenya [131], shed more light on hate speech as a phenomena in

Kenya. As a result, the rest of the research was guided by an operational definition of hate speech. In addition, essential factors that characterize hate speech were derived from hate theories to inform the conceptual framework's development.

By crawling tweets containing these ethnic names as key terms, the study employed the ethnic names of seven out of forty-two major tribes in Kenya that account for over 70% of the country's population [132] as the study population parameter. The Kikuyu, Luhya, Kalenjin, Luo, Kamba, Kisii, and Meru, as well as Kiswahili and "nick-named" variations of these ethnic groupings, were among them. Furthermore, the raw dataset was collected and developed using these ethnic names in combination with other terms as indicated by the study's multidimensional architecture in figure 2.26.

#### **3.4.1.2 Population**

By crawling tweets containing these ethnic names as key terms, the study employed the ethnic names of seven out of forty-two major tribes in Kenya that account for over 70% of the country's population [132] as the study population parameter. The Kikuyu, Luhya, Kalenjin, Luo, Kamba, Kisii, and Meru, as well as Kiswahili and "nick-named" variations of these ethnic groupings, were among them. Furthermore, the raw dataset was collected and developed using these ethnic names in combination with other terms as indicated by the study's multidimensional architecture in figure 2.26.

#### **3.4.1.3 Sampling**

Unlike traditional research, big-data initiatives employ various sampling strategies to computationally capture all available online content [133], such as employing a web crawler or Twitter API to collect a large number of messages from social media based on specified key terms. Such methods are frequently free of the constraints that come with standard sampling methodologies [134], such as the inefficiency and impracticality of collecting a large volume of hate speech data from many Kenyan social media users for machine learning reasons. Our work used simple random sampling to establish a study sample for annotation from the large volume of data collected. Previous research [54] [135] employed this sampling strategy to obtain study samples from social media.

Data from the Twitter social media network was collected via convenience sampling. Unlike other social media platforms, Twitter makes every post public and programmatically accessible unless the user specifies differently in their settings. Furthermore, accessing these tweets does not require an account, and anyone can anonymously publish, like, dislike, and rapidly transmit the messages to a large audience. Because of these traits and characteristics, the platform is vulnerable to the spread of hate speech.

### 3.4.2 Data Understanding

This is the second step of CRISP-DM, which checks the data quality using the result from the problem understanding phase as input. The conceptual framework was one of the most important inputs, and its variables were crucial in shaping the data gathering process. The dataset properties, such as the dataset size, data columns, data kinds, class distribution, data frequencies, the mean, and other statistical information about the data, are often studied during this phase. During these data exploration tasks, the Pandas library was utilized to compute and present the data in data frames. The data was also plotted in charts using the Matplotlib software. The most often occurring terms in the dataset, for example, were plotted on a histogram, whereas the most frequently occurring words per class were plotted on a word cloud.

#### 3.4.2.1 Data collection

This is a critical step of the research, the heart of the study, which verifies the findings [136]. The performance of the trained model is directly proportional to the quantity and quality of data collected. The desired data for acquisition included tweets from Kenya's presidential campaign in August 2017, which includes a second election in October 2017. Previously, the Twitter API was used to create an app that gathered tweets during election days. A crawler based on Python programming was also used to supplement Twitter API's two-week data collection window in order to obtain a massive amount of archival tweets, which included tweets from the March 2013 general elections and the four months leading up to March, as well as two months after the results were announced. This time period and the events surrounding it have been the most prominent trigger events in the past, resulting in large spikes in online hate speech.

As a key data collecting strategy, the bootstrapping technique was adopted. To explore social media networks, seed words consisting of hate-related keywords (kw), phrase patterns (pp) with a connotation of hatred (138), offensive hashtags (#), and pro-hate user account names (un) were used. Figure 3.2 depicts a summary of the process flow.

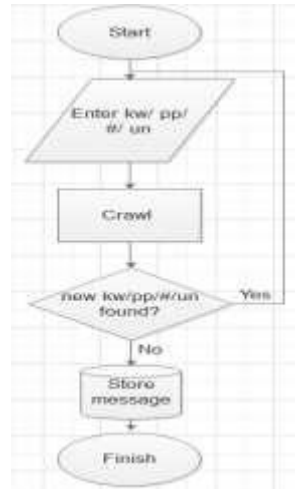


Figure 3.2: Data collection flowchart

#### 3.4.2.1.1 Use of Keywords and Phrase patterns

Hateful keywords were used to search for messages on social media, including insults, profanities, discriminatory, and offensive expressions commonly used to culturally denigrate or devalue a person or persons based on their ethnic community in Kenya. These terms were chosen because they are more likely to return messages with hostile content and have similar terms in the Hatebase.org hate speech lexicon. The term 'kihii,' which is a Kikuyu term that devalues an uncircumcised individual, was at the top of the list. The majority of these terms were taken from earlier tweets that had been labeled as offensive or hate speech by online users. As a result, many more degrading, abusive, and profane terms were discovered and utilized to search for similar tweets utilizing the snowballing technique. Other possibly abusive term patterns in tweets were discovered using the same method. For instance, a phrase beginning with “All <ethnic name>.”

#### 3.4.2.1.2 Use of problematic Hashtags

Hashtags are a unique feature of the Twitter app that are used to organize messages by topic. Using 18 offensive hashtags, such as #KillAllKikuyus, this strategy was crucial in collecting potentially hateful messages.



### **3.4.2.1.3 Use of Pro-hate speech user accounts**

Additionally, messages made by pro-hate user accounts[23], particularly by people of importance in society such as legislators and well-known local bloggers who had previously been documented for posting content verging on hate speech, were gathered. First, according to local newspapers [139], [140], a list of pro-hate speech politicians and bloggers was compiled. This includes hashtags that were trending on social media and were linked to individual user accounts. Following that, the names on the list were used to look up verified Twitter account handles associated with the politicians and bloggers in question. The tweets of these users were captured and saved into a database using Twitter's API.

### **3.4.3 Data Preparation**

This was the most time-consuming step of the project, but it was also the most important in properly preparing the data for the machine learning phases that followed. Data preparation was followed by data labeling, which required annotating a sample of the messages gathered. Following that, the annotated dataset was cleaned. Following that, the data was converted to a numeric representation using feature extraction and vectorization to meet the numerical input requirements for machine learning and deep learning models.

#### **3.4.3.1 Data annotation**

Human labor was used to give a class to each message in the dataset during data annotation, also known as data labeling. A group of forty human annotators was enlisted and given training on how to annotate the messages. The team was made up of 80 percent undergraduate computer science students and 20 percent staff employees. The final annotation team from Africa Nazarene University's school of science and technology was selected through convenience sampling (ANU). The average age of the crew was twenty-three, with a gender balance of 21 male and 19 female annotators. The team's nationality was heavily biased toward Kenya. The skewness stemmed from the necessity for annotators who could quickly decipher the codeswitched nature of the corpus, which included texts in English, Swahili, and a few other Kenyan native languages. The first training was based on the annotation method to ensure that everyone on the team had a common understanding of hate speech. The annotators were then instructed on how to use a web-based annotation portal built by the research team to annotate sample texts [21]. The original crew of forty annotators was eventually reduced to twenty-seven. Individual performance was considered,

as well as a signed pledge to annotate at least three thousand mails in one week. The first session yielded useful feedback on the annotation portal's speed. This was utilized to restructure the portal and speed up the annotation process by having a random team of three amateur annotators annotate each tweet, rather than having each tweet annotated by a specific team, one of whom was a subject matter expert (SME). The new design was inspired by the previous session's slow annotation process and the necessity to speed up the annotation process in order to have a larger labeled dataset to train the classifiers. Furthermore, it was thought that by doing so, the team of human annotators would be able to better utilize and optimize their experience in that short amount of time. Krippendorff's alpha [141] was used to assess the annotations' dependability in terms of determining the extent to which the team agreed on the class for each message. Because it can accommodate any number of human raters and is strong enough to accept missing data, even with very small data samples, this inter-rater reliability score is widely employed in content analysis.

The class of a tweet was finally determined when two or three raters agreed on it independently. If no consensus could be reached, a fourth annotator, a subject matter expert, would function as a tie-breaker to establish the tweet's class.

### **3.4.3.2 Data Cleaning**

Data cleaning is an essential step in the machine learning process because it removes noisy signals that would otherwise degrade the training and, as a result, the overall performance of a classifier model. Natural language processing techniques such as tokenization, stemming, and lemmatization was used to clean the data in this study. Regular expressions (regex) were used to eliminate HTML characters, non-ASCII and corrupted characters, empty rows, duplication, emoticons, stop words, and punctuations, among other things. Quote marks, commas, apostrophes, exclamation marks, and other punctuation marks were commonly used. All of the terms were also lowercased to normalize the data.

#### **3.4.3.2.1 Tokenization**

The tokenization method includes using whitespaces, newlines, tabs, and other delimiters to separate raw text messages into phrases and then into a list of individual words, also known as tokens. This was significant because, in terms of machine learning, computers can digest token units far more readily and quickly than the original corpus documents. Standard abbreviations and hyphenated terms, the majority of which were in English, were likewise handled via tokenization.

The hyphenated words were broken up into two tokens. Apart from the default, which is in the standard English list, the NLTK word tokenizer function was customized by adding to the list several typical codeswitched abbreviations.

#### **3.4.3.2.2 Removing Stop words**

With the exception of third-person pronouns, which, as indicated in the study's conceptual model, would be highly symptomatic of hate speech, the NLTK corpus stop words library was used to eliminate all English stop words. Swahili equivalent stop words like "ni," which is the Swahili equivalent of the English stop word "is," were added to the English stop word list. Stop words, in general, have been found not to contribute meaning to a phrase's deeper meaning [104], which explains why they are frequently filtered out.

#### **3.4.3.2.3 Filtering out punctuations**

Duplicate messages, single letters, non-alphanumeric data, HTML elements, dates, emoji, and URLs were removed using regular expression methods in Python's natural language tool kit (NLTK) module, and punctuations and other non-ASCII characters were replaced with a space. To loop over all tokens and filter out the solitary punctuation, the Python `isalpha()` function was utilized. Spam messages containing advertising based on popular hashtags were also removed. The following is an example of a message with corrupted characters: “Ã¶ ë¼¼¼¼¼¼¼¼¼¼.” Advertisements based on a popular hashtag include the following: “*#NoReformsNoElections Apple launch iPhoneX.*”

#### **3.4.3.2.4 Stemming**

Stemming is the process of reducing a word's inflectional forms to its base form, which usually results in a smaller vocabulary. Words like hate, hated, hating, and hates, for example, are reduced to the stem word hate. The Porter Stemmer method in the NLTK was used to stem the data for this investigation.

#### **3.4.3.2.5 Case Normalization**

Case normalization was another preprocessing activity, in which all of the text messages were changed to a single case, which was lowercase. This was accomplished by using Python's `lower()` function on each word.

#### 3.4.3.2.6 Additional Filtering

In addition, the length of the text message was taken into account when selecting whether tweets were acceptable. Messages with three or less characters, for example, were eliminated. These were generally tweets with a single word or a few characters that were contextually confusing on their own. Furthermore, the requirements specified in the annotation system for categorizing a message into the preset classes were broken by these types of brief communications. "c," "ok," "DAAMN!" and "I'll" are some examples of messages. There were also instances where messages were confusing due to the usage of a single number, symbol, or one-word acronym. "2546," "WTF," "Smh.nkt!" and "#" were examples of messages.

The message section of a tweet can typically be no more than 140 characters long. In contrast to the findings of this investigation, the longest tweet recorded had 991 characters and was made up of concatenated URLs. Although this was an exception, it could be explained by the latest Twitter design's enlarged capacity of 280 characters in the message area. The 280 character limit applies solely to the message part; anything after that, such as an attached URL address, can cause the tweet to become too long. Given that the length of a message has previously been shown to be less helpful for categorization as a machine learning feature [54], this study chose to just evaluate the message component of the tweet and ignore the URL section.

There were tweets in English, Swahili, and codeswitched text including words from numerous Kenyan ethnic groups. A few more tweets in Asian languages were eliminated as part of the noise signals because they didn't contribute any value to the classification process.

Furthermore, all user mentions, such as @martins, were replaced with a generic "USERNAME" tag, whereas the URL section, which often contains account names, was filtered out to safeguard the user identity of message recipients and authors. Regular expressions were used to achieve these results.

#### Split into Training and Testing datasets

Splitting the dataset into training and testing data samples is an important step in data preparation. For the classifier to understand the underlying data distribution, the training dataset is frequently given a larger fraction than the testing dataset. The proportions are frequently determined by the magnitude of the data available. If the original data set is enormous, for example, just a tiny percentage of it will be needed for evaluation. The 80:20 and 70:30 ratios are two popular training

and testing proportions. In this study, 80 percent of the dataset was used for training, and the remaining 20% was used to test and evaluate the classifier model. The train test split library in Scikit learn was used to split the data.

Separating the two data sets is a machine learning technique for avoiding overfitting, which occurs when a model becomes deceptively good by "regurgitating answers" from memory. This is because, during testing, the data samples that were viewed during training are submitted to the model once more. As a result, the model performs admirably during training but horribly when exposed to new or unknown data samples.

### **3.4.3.3 Exploratory Data Analysis**

The data exploration step entailed using quantitative and visual approaches to examine and comprehend the dataset in general by looking for patterns in data types, class distribution, word frequencies, missing values, and other aspects. After an in-depth analysis that succinctly determines the interpretation and correctness of the conclusions regarding the machine classification of hate speech text, these are frequently visually plotted using word clouds for text data, pie charts or bar graphs, or any other statistical chart to provide some high-level insight into data patterns and other characteristics that will give confidence to the kind of expected results. The chart graphics were created using the Scikit-Matplot package. The quantitative approach also aids in displaying the class distribution, describing the counts, mean standard, max, and min of the data, and describing the counts, mean standard, max, and min of the data. Furthermore, the researcher will be able to ask the proper questions without biasing the studies with faulty data assumptions. The class and tweet message columns in the dataset used in this study were the two key columns of interest.

### **3.4.4 The Selection and Extraction of Features**

Following the meticulous cleaning of the raw text messages as described above, the following phase entailed feature selection and extraction. Feature selection aids in the extraction of a set of informative and high-quality words for machine learning from a larger raw corpus. The data cleaning phase, which comprised punctuation filtering, case normalization, stemming, and stop

word removal, was the initial step in reducing noise signals and improving the input vocabulary's quality. This language must then be converted to a numeric representation before being used as input for machine learning algorithms. This is due to the fact that machine learning algorithms can only interpret numerical data, such as vectors of numbers [104]. As a result, the text messages in this study required to be transformed into machine learning feature representations.

During the trials, four low characteristics were largely used, and their performance was compared by learning several classifiers. TF-IDF, BoW word count frequencies, word embeddings, and PDC-TF-IDF characteristics were among them. PoS as features and topic models as high-level features were also used in the studies as extra features. The BoWs features are essentially frequency counts of phrase occurrences in a tweet. The count vectorizer in the Scikit-learn machine learning library was used to create these. The relevance of a specific term in a document and the entire dataset was compared using TF-IDF characteristics. The basic concept is to punish terms that appear too frequently across all documents because they may not be as relevant to the model as words that are unique in individual texts but uncommon across all documents. The TF-IDF of a term  $t$  in document  $d$  is calculated mathematically as follows:

$$TF-IDF(d, t) = tf(tf) * idf(d, t)$$

The input tokens were processed on three levels: phrase level (represented by n-grams), word level (represented by words), and character level (represented by characters). As a result, TF-IDF vectors for the various levels were created. For each level, a feature matrix was created in general. The GloVe pre-trained embeddings were employed as the key features for Word Embeddings, based on the 100d file containing about 1 million word vectors. The messages were initially tokenized in the dataset. Following that, each token was mapped to its appropriate embeddings using the transfer learning method. Topic Models[92] were employed as high-level features for data exploration, data connecting to the conceptual framework, and, more crucially, as an automated procedure to inform the salient words to include in the subsequent phase to build the PDC word-family features. From a vast dataset of short-text messages from social media, the Latent Dirichlet Allocation (LDA) method was utilized to identify 23 semantically relevant topics or clusters. Table 4.5 shows the results. PDC characteristics are psycholinguistic qualities derived from the triangle theory of hatred's three dimensions of hate [33]. PDC promotes hate speech through three core word families that are both concept-based and language agnostic. As a result, by adding or eliminating similar-meaning terms in other languages in the relevant word families,

the language list might increase or shrink. Words of the passion word family reflect negative emotions such as anger, fear, disgust, and contempt. Threatening, abusive, insulting, and other offensive words directed at a target person or group based on protected traits such as race, ethnicity, religion, and so on are examples. "To heck with all group>," for example, is an example message. They must be expelled from the country." Previous research have employed negative polarity and sentiment analysis to detect passion episodes [26], [107]. The distance word family, often known as "othering" language, consists of terms that communicate psycho-social distance or proximity in inter-group or inter-person connections [34]. The use of pronouns [56], [76], [142], [143] is frequently indicative of this. For instance, "us, them, they, us, you," and so on. "Kambas likewise do not make good leaders...they are Cowards," as an example of a real tweet. The commitment word family is made of words or phrases that pledge to openly depreciate another person or group. This can take the form of utilizing objects, bugs, or animal names to refer to them, or just considering others as less superior, immature, or human [108]. Furthermore, this contains some of the code names that are only known and used by the in-group to refer to out-group members. Here's an example of a tweet from our database: "*Luos, your Enemies Are Kikuyus, please stop making music with these Cockroaches*".

All of these high-level text attributes were encoded as input vector values for machine learning using the Scikit-learn toolkit. The text messages were converted to word count vectors using the CountVectorizer, and the text messages were converted to word frequency vectors using the Tfidf Vectorizer. In both situations, the dataset's messages are tokenized first, and a vocabulary of known words is created. The result is an encoded vector containing the whole vocabulary's length. Following that, each new text message is encoded as a fixed-length vector with the vocabulary's length. The value at each place in the vector is filled with a frequency count of each word occurrence in the new text message for the CountVectorizer. If a word in the new text message is not in the vocabulary, it is ignored and hence does not receive a count in the final vector. The Tfidf Vectorizer calculates word frequencies and assigns a high score to often occurring terms inside a text, but downscales the most frequently occurring words across all papers. When encoding new text messages, the scores, which are usually between 0 and 1, are utilized to provide frequency weightings to the vector.

### 3.4.5 Modelling

Modeling comprises three main steps: the selection of the model, the training of the model, and the tuning of the model's parameters.

#### 3.4.5.1 Model selection

There are various models to choose from, however they are essentially divided into two types: unsupervised and supervised. When a model is given unlabeled data as input, it finds some pattern or structure in the data on its own, determining which data points are more linked and building clusters. In supervised learning, the machine has access to both the input data and the expected outputs, allowing it to do classification or regression. There is a prediction in this case, whereas there is none in unsupervised learning. The dataset in this work is made up of annotated messages, which informs our decision to use supervised learning. These models can be further classified as regression or classification models within supervised models. Regression models are used to analyze non-discrete data, such as a range of real numbers from -1 to 1. Classification is the process of mapping certain input data to a discrete set of values or classes, such as hate speech or neither.

Furthermore, classification models can be classified into two types: traditional and deep learning. The classifier models were learned using both types. The encouraging findings from past similar work guided the selection of specific models for each kind. The Naive Bayes, Support Vector Machine, Linear Logistic Regression, Decision Trees, and K-Nearest were among the traditional machine learning methods. Furthermore, the Random Forest (RF) and Extreme Gradient Boosting (XGB) Bagging and Boosting models were applied. Convolutional Neural Networks and Hierarchical Attention Networks were investigated in terms of deep learning. For the deep learning models, the default settings were used with only minor fine-tuning. All of the machine learning tests were carried out with similar models from the Python Scikit-learn library of machine learning models.



### 3.4.5.2 Model Training

During this phase, the data was used to help the model improve its ability to detect hate speech in text messages.

$$Y = m * x + b \quad Y = m * x + b \quad Y = m *$$

Where  $y$  is the output,  $m$  is the slope,  $x$  is the input, and  $b$  is the y-intercept.

The slope ( $m$ ) and the intercept ( $b$ ), are the variables to alter during training. Where  $x$  is the input and  $y$  denotes the projected output. Because there might be multiple features included in machine learning, there could be a variety of slopes( $m$ ). A weight matrix is frequently used to collect the slope values ( $w$ ). In the same way, the biases are organized into a biases matrix ( $b$ ).

To forecast the output, the training procedure begins by initializing  $w$  and  $b$  with some random values. Initially, the anticipated value may reflect poor model performance. However, by modifying the parameter values,  $w$  and  $b$ , in succeeding cycles and comparing them to the expected values, this performance can be improved. This is frequently shown on a confusion matrix, a table that is commonly used to visualize classification model performance.

### 3.4.5.3 Tuning the Model's Parameters

Parameter tuning is performed to see if the trained model may be improved further. During training, the default parameter values are frequently accepted. These settings can now be modified to determine if they make a substantial difference in the model's performance. The number of iterations during training, in which the model is exposed to the data several times, is an example of a parameter that can be tweaked. The margin cost, learning rate, and kernel choice are some of the other characteristics that can be tweaked. These hyperparameters influence not just the model's performance but also the length of time it takes to train it, with longer training times implying higher costs.

A collection of hyperparameters was discovered and set up in a parameter grid for each model for this study. During the tests, these were automatically changed using Grid search with cross-validation [60] to score parameters and find the model's ideal hyperparameters. The value of the soft margin cost,  $C$ , the kernel choice, and other estimator parameters were among them. The

generalization of the nonlinear Support Vector Machine in identifying various sorts of hate speech, for example, was investigated by modifying the soft margin cost,  $C$ , with lower penalty values ranging from 0.001 to 1.0. Three common kernels from the literature, the linear, the Radial Basis Function (RBF), and the Polynomial, were used in the tests to help the model establish a nonlinear decision limit. All of these model parameters were derived from those in the SciKit-Learn libraries. Furthermore, each time the algorithm was run, a pipeline was employed to smoothly merge these parameters with the vectorizer settings.

### 3.4.6 Evaluation

This step comes before the training phase and is used to evaluate the model's performance by determining whether it is accurate to reality. To assess how well the trained model can make right predictions, it is exposed to new unlabeled examples that it has never seen before. Given the supervised approach to machine learning, the goal of the study was to develop a model that could accurately categorize hate speech from unseen text messages while also estimating its generalizability. This is a representation of how the model was supposed to perform in the real world, or the "ground truth," as it is known in science. To objectively and quantitatively evaluate the performance of each model, classification accuracy and F-score based on the weighted average of precision and recall values were utilized. Furthermore, heat maps and other visualization techniques were used to create the confusion matrix reports.

The accuracy evaluation was done to determine the likelihood of the classifier being right. The accuracy value was calculated using the formula below

$$\text{Accuracy (1 - Error)} = \frac{\text{Tp} + \text{Tn}}{\text{Cp} + \text{Cn}}$$

$T_p$  and  $T_n$  were the projected true positive and true negative occurrences, respectively, whereas  $C_p$  and  $C_n$  were the total counts of real positive ( $T_p + T_n$ ) and actual negative ( $F_p + F_n$ ) instances. The percentage of correct positive predictions was used to determine how often the classifier would be correct when predicting a message as hate speech. Precision was calculated using the following formula:

$$\text{Precision} = \frac{\text{Tp}}{\text{Tp} + \text{Fp}}$$

The percentage of positive instances anticipated as positive is known as recall, also known as model sensitivity. It is calculated like this:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

In essence, three evaluations were required: an assessment of the data annotations in terms of inter-rater reliability, an assessment of features in terms of how discriminative each feature would be in a text classification task, and an assessment of the trained model. The annotations' inter-rater reliability was assessed using Krippendorff's alpha score [141]. The labeled dataset was divided into training and testing datasets in an 80:20 ratio, with 80 percent used to train the machine learning model and 20 percent used to evaluate the learnt model's performance. In addition, a cross-validation process based on random samples from the whole labeled data set was used to produce 5-folds to test and evaluate the models' performance. When working with a single model and relatively smaller datasets, 10-fold cross-validation is typically used. However, the large number of models and instances used in this study would otherwise increase computational time and memory. Hate speech classification models were trained using seven traditional machine learning algorithms and two neural network algorithms. Following that, these models were compared, and the model with the best prediction accuracy in identifying the positive class, i.e., hate speech, was chosen based on the validation data set, i.e., K-fold cross-validation. Grid search was also utilized to learn the models by determining the optimal model hyperparameters and feature combinations. Figure 3.3 depicts the model evaluation approach.

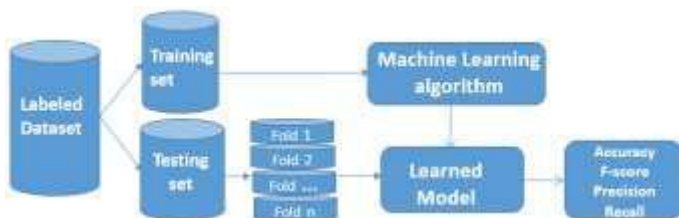


Figure 3.3: Model Accuracy Estimation in the Evaluation Process

By comparing the predictions to the actual results based on the annotated test dataset, the confusion matrix was utilized to assess the classifier's accuracy. Table 3.1, for example, is a confusion matrix for determining if a message contains hate speech or not. The true class of the text message is in the first column, while the predicted class by the classifier is in the second. True positive (Tp) indicates that the classifier correctly predicted that the message would contain hate speech, and it

did. True negative (Tn) indicates that the classifier predicted that the message would not contain hate speech, and it did not. The term false-positive (Fp) denotes that the classifier anticipated that the message would contain hate speech, but it did not. This is also known as a type I error, and it is the term used in the study to indicate when the null hypothesis is rejected when it is true. False-negative (Fn) means that the classifier predicted that the message would not be hate speech, but it was. This is also known as a type II error, and it refers to situations in which the null hypothesis is not rejected even when it is false.

Table 3.1: Example of Confusion Matrix

| True Class | Predicted Class |          |   |
|------------|-----------------|----------|---|
|            | Positive        | Negative |   |
| Positive   | Tp              | Fn       | P |
| Negative   | Fp              | Tn       | N |
|            |                 | RR       |   |

Separate classifiers for hate speech, offensive speech, and neither were trained using the One-Versus-All (OVA) architecture. The highest predicted probability from all of the classifiers was used to establish the class label for each message.

In terms of model accuracy, the ideal performance is 1.0, the average performance is 0.5, and the worst performance is 0.0, which indicates the model is always wrong.

Benchmarking with the inter-rater reliability score of human annotators who labeled the same dataset was the major approach for determining the success of the final choice of the classifier model. As a result, based on the initial K-Alpha produced by the human annotators, a threshold value of more than 0.5 was applied. Any result with a likelihood greater than this was positively classified as hate speech; otherwise, it was not.

### 3.4.7 Deployment

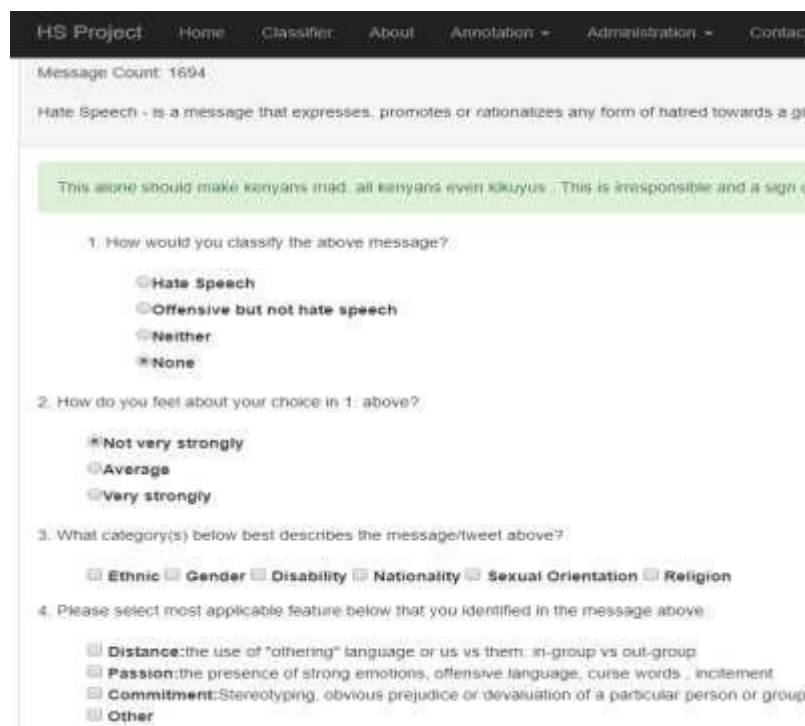
This is the stage at which the machine learning model is really put to work in order to generate a return on investment [129]. The predictive model is realized in this study by categorizing some genuine unseen text messages into hate speech, offensive, or neither categories. The general public can use a web site with a public Internet Protocol address to view and test the classifier by entering

in fresh messages or pasting copied messages from social media to see the projected class. Appendix E illustrates this.

### 3.5 Validity and Reliability Assurance

Construct validity is a criterion for determining if a measuring method is appropriate for the specific construct being tested rather than another. This necessitates the use of indicators and measurements that are based on theory or existing knowledge. The content validity of a measuring method is used to validate that it captures all of the construct's attributes. Validity will be harmed if crucial features are missing. Criterion validity is a method of comparing or calculating the results of one study to the outcomes of similar studies [144].

Three human raters labeled each message based on the annotation scheme provided on the annotation site to assure content validity (appendix D). The conceptual foundation in section 2.8.2, which is based on the duplex theory of hate [33], guided the scheme. Furthermore, the definition of hate speech remained visible above the frame that presented each new message available for annotation on the annotation portal. Figure 3.4 depicts this.



The screenshot shows the 'HS Project' annotation portal. At the top, there is a navigation bar with links for Home, Classifier, About, Annotation, Administration, and Contact. Below the navigation bar, the page displays 'Message Count: 1694'. A definition of Hate Speech is provided: 'Hate Speech - is a message that expresses, promotes or rationalizes any form of hatred towards a group'. The message to be annotated is: 'This alone should make kenyans mad. all kenyans even sikuyus . This is irresponsible and a sign of...'. The form below the message asks for classification and features. Question 1: 'How would you classify the above message?' with radio button options: Hate Speech, Offensive but not hate speech, Neither, and None. Question 2: 'How do you feel about your choice in 1: above?' with radio button options: Not very strongly, Average, and Very strongly. Question 3: 'What category(s) below best describes the message/tweet above?' with checkboxes for Ethnic, Gender, Disability, Nationality, Sexual Orientation, and Religion. Question 4: 'Please select most applicable feature below that you identified in the message above:' with checkboxes for Distance (the use of "othering" language or us vs them: in-group vs out-group), Passion (the presence of strong emotions, offensive language, curse words, incitement), Commitment (Stereotyping, obvious prejudice or devaluation of a particular person or group), and Other.

Figure 3.4: The annotation portal

Based on the annotations made by the team of 27 human annotators, an inter-rater reliability score was generated. At least three human annotators were required to annotate each tweet. The mode was the determining factor for the tweet's class statistically, meaning that the tweet's class was determined by two or more votes. In the event of a tie, meaning that the team of three human annotators could not agree, a fourth annotator, ideally a subject matter expert, would be introduced as a tie-breaker. Because it could deal with missing values and was robust enough to deal with outliers, Krippendorff's Alpha was chosen as an inter-rater reliability measure for the annotation exercise with a team of 27 rookie annotators [141]. A second annotation, consisting of one subject matter expert, was performed on 9k sampled tweets to further validate the novice annotators' reliability. The Cohen Kappa was used to assess the annotations' reliability.

The triangulation approach was used to determine the construct and prediction validity of the research data and framework elements. To find the appropriate feature set to train our classifier, we compared performance results from various conventional and deep learning machine learning techniques.

### 3.6 Ethical Considerations

Using social media as a primary source of data for research has certain ethical implications. People of all demographics are increasingly using internet platforms like Facebook, Instagram, and Twitter to share their thoughts, feelings, and private thoughts. As a result, while collecting such data, there are two key concerns: user consent and user identity protection. In the first case, the topic of user consent for messages posted on social media, particularly Twitter, has already been debated [145]. Unlike other social networking sites, however, messages made on Twitter are public by default unless the user puts on the privacy settings, which only allow individuals who follow them to see their tweets. This is one of the reasons why public tweets have been used in a lot of academic studies [146]. Needless to say, obtaining user consent from accounts that create hundreds, if not millions, of tweets that could be gathered using either the Twitter streaming API or archival tweets [146] will be nearly impossible. Furthermore, tweets may be made anonymously, or people may have left or canceled their accounts, but retweets may still be accessible. This, too, makes it impossible to reach out for consent, if consent was ever required. The study concentrated on gathering solely public tweets and retweets, which do not require formal consent or ethical approval.

To safeguard the identities of internet users, all user names and mentions were replaced with a generic USERNAME label. The tweets were accessed using Twitter APIs for developers. Following Twitter's privacy and data sharing policy [147], only tweet IDs will be used to publicly publish the information.

### 3.7 Summary

The steps of problem understanding, data understanding, data preprocessing, modeling, assessment, and deployment were reviewed in this chapter, as well as research methodology, research philosophy, research design, and research methodologies using CRISP-DM. Concerns about validity and dependability were addressed, and ethical considerations were thoroughly presented in each segment.

In conclusion, the research technique was systematically constructed, with each study objective linked to a specific research question, the appropriate research design, and the anticipated outcomes. Table 3.2 summarizes these findings.

Table 3.2: Summary of the research methodology

| Research Objective  | Research Question   | Research methodology                                     | Expected results  |
|---|---|--|---|
| Develop a deep understanding of what constitutes hate speech  | What constitutes hate speech?   | Literature review<br>Content analysis                    | Working definition of hate speech. Hate speech themes     |
| Establish a gold-standard annotated codeswitched dataset of hate speech from social media in Kenya  | How can we extract relevant text messages containing hate speech from social media during past general elections to build a high-quality hate speech dataset? | Literature review<br>Experimental design<br>Survey       | Complete human-annotated dataset                          |
| Explore the salient features that discriminate hate speech messages from other messages   | What are the salient characteristics of text messages containing hate speech?   | Literature review<br>Descriptive and experimental design | Feature subset<br>Feature vectors                         |
| Investigate the feasibility of a framework for building classification models to automatically analyze and identify hate speech in a codeswitched text environment. | To what extent does our framework effectively learn a classification model to accurately predict hate speech in the codeswitched text?                        | Descriptive and Experimental design                      | Conceptual framework<br>Classifier model                  |
| Evaluate the performance of the model   | How does the classification model's performance enrich the understanding of hate speech in Kenya during general elections?                                    | Experimental and descriptive design                      | Accuracy, precision, recall results, and Confusion matrix |

## CHAPTER 4: RESULTS AND FINDINGS

The outcomes and findings from the numerous activities and experiments outlined in chapter three are presented in this chapter. The outcomes of the problem understanding phase are presented in the first part. The second element is a descriptive examination of data comprehension. The outcomes of data preparation are presented in the third section. The model-building outcomes are described in the fourth section. The model evaluation results and findings are presented in the final section.

### 4.1 Problem Understanding

This phase was primarily focused on addressing the first objective, which was to determine what constitutes hate speech. Several definitions of hate speech, including dictionary definitions, legal definitions, and hate speech definitions on user policy documents on social media networks, were examined in this regard. In addition, a review of current theories of hate was conducted to have a better understanding of the phenomena as it was expressed on social media.

Furthermore, the NCIC commission had underlined the necessity for automated monitoring of social media for hate speech before the 2017 Kenyan presidential elections in the chairman's report [148]. The increased number of online hate speech incidents, as well as the commission's previous concerns with successfully prosecuting hate speech, owing to a lack of evidence to support convictions in hate speech instances, prompted this decision.

As stated in Table 4.1, content analysis was undertaken by identifying similarities and variations in these hate speech criteria. The core reference or dimension of hate, as well as the object of hate, were among the subjects explored. An important finding was that all of the definitions appeared to be affected by and generated from a legal perspective, which may be insufficiently detailed or limited in scope [32]. These were also influenced by the opposing perspectives of the hate speech originator and the target or victims of hate speech [149].



Table 4.1: Content analysis of hate speech definitions

| Source                         | Reference to |         |             |           |                 |         |           |                |             |         | Target Attributes |        |           |                       |            |             |        |               |           |     |
|--------------------------------|--------------|---------|-------------|-----------|-----------------|---------|-----------|----------------|-------------|---------|-------------------|--------|-----------|-----------------------|------------|-------------|--------|---------------|-----------|-----|
|                                | Violence     | Attacks | threatening | prejudice | Insubordination | Offence | Inimidate | Discrimination | Intolerance | Degrade | Race              | Ethnic | Religious | Sexual or orientation | Disability | Nationality | Gender | Dissemination | Political | Age |
| Oxford dictionary              |              | X       | X           | X         | X               |         |           |                |             |         | X                 |        | X         | X                     |            |             |        |               |           |     |
| Oxford English dictionary      |              |         |             |           |                 |         |           | X              | X           |         |                   | X      | X         | X                     |            |             |        |               |           |     |
| Merriam-Webster                |              |         |             |           | X               | X       | X         |                |             |         | X                 |        | X         | X                     | X          | X           |        |               |           |     |
| UN's International com         |              |         |             |           |                 |         |           |                |             | X       |                   |        |           |                       |            |             |        |               |           |     |
| European Court of Human Rights |              |         |             |           |                 |         |           | X              | X           | X       | X                 |        |           |                       |            |             |        |               |           |     |
| NCIC Act 2008                  | X            |         | X           | X         | X               |         |           | X              |             |         | X                 | X      |           |                       |            | X           |        |               |           |     |
| BCC South Africa               | X            |         |             |           |                 |         |           | X              |             |         | X                 | X      | X         | X                     | X          | X           | X      |               |           | X   |
| YouTube –                      | X            |         |             |           |                 |         |           | X              |             |         | X                 | X      | X         | X                     | X          | X           | X      |               |           | X   |
| Facebook                       | X            | X       | X           |           |                 |         |           |                | X           |         | X                 | X      | X         | X                     | X          | X           | X      | X             |           |     |
| Twitter                        | X            | X       | X           |           |                 |         |           | X              |             |         | X                 | X      | X         | X                     | X          | X           | X      | X             |           | X   |
| LinkedIn                       | X            | X       | X           |           |                 |         |           |                | X           |         | X                 | X      | X         | X                     | X          | X           | X      | X             | X         |     |

In addition, as illustrated in Figures 4.1 and 4.2, the verb frequencies and hate targets acquired from the content analysis exercise were aggregated. Hate speech was defined in the majority of definitions as inciting or threatening statements. Figure 4.1 summarizes this information.

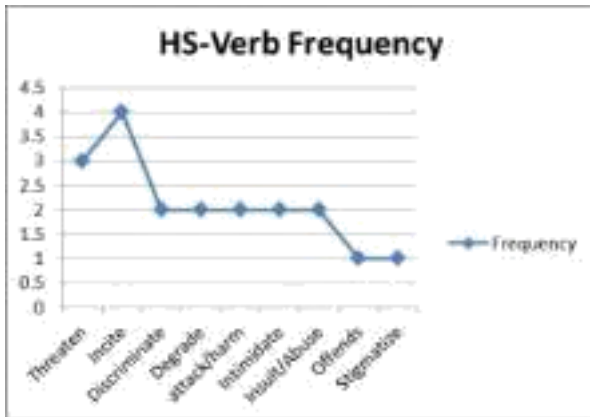


Figure 4.1: Verb Frequency in hate speech definitions

Hate speech is meant to incite hatred, violence, and prejudice through its content and prominent qualities. Figure 4.2 depicts this.

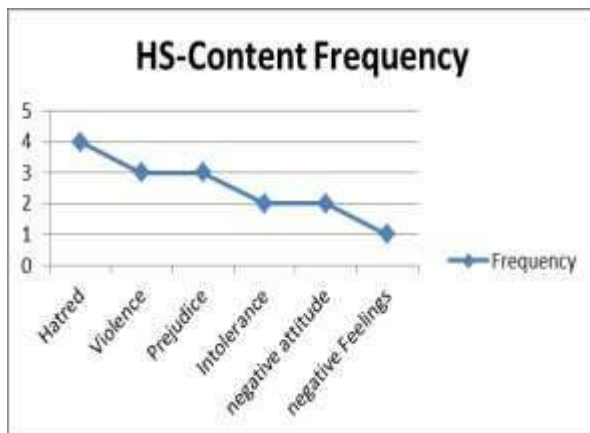


Figure 4.2: Hate-specific content frequency

### 4.1.1 Findings

The examination of the various definitions revealed that hate speech has three distinct aspects: First, it is a nonverbal expression, such as body language, or a verbal expression, such as text, images, or graphics, that incite, attacks, threatens, intimidates, discriminates, degrades, offends, insults, or stigmatizes. Second, hate speech expressions have targets, which might be an individual or a group of individuals who share a protected social feature like ethnicity, race, gender, religion, and others [56]. Third, the expression is intended to incite hostility, violence, prejudice, intolerance, a negative attitude toward the target, and unfavorable feelings toward the target.

Fundamentally, only with a thorough grasp of the hate speech phenomena and its characteristics can it be simply recognized and significant insight into how to recognize it automatically be gained.

#### 4.2 Data Collection

Approximately 400k unprocessed messages were gathered and saved in a comma-delimited file (CSV) format. These primarily comprised of Twitter text messages, often known as tweets, from Kenya's general elections in August 2017, as well as a follow-up election held 60 days later in October 2017. To build a large raw corpus, additional tweets were crawled from January to December 2017 as well as the March 2013 general elections.

The dataset included English, Swahili, and code-switched messages, with the majority of the code-switched messages being English-Swahili. "Yes, I feel terrible for the deceased, but bado lazima tu wakikuyu wakae like the guilty ones, even while we are doing nothing," for example.

Table 4.2 provides a summary description of the dataset.

*Table 4.2: Raw Dataset Description*

| Description                           | Number of text Messages |
|---------------------------------------|-------------------------|
| Total collected text messages         | 401,211                 |
| Total preprocessed text messages      | 398,000                 |
| Codeswitched: English, Swahili, other | 29309                   |

#### 4.3 Data Annotation

Two data annotation tasks were done. First, 9 human annotators annotated preliminary data to validate the conceptual framework and the annotation tool. The initial annotation sampled 20k messages from the raw dataset. The team annotated 4,931 messages in one day, with three annotators each message. As a consequence, 903 communications were identified as hate speech, 520 as offensive, and 3140 as neither. The annotators could not agree on a fourth category of 368 messages, meaning each message was annotated differently by each annotator. Table 4.3 summarizes the annotated texts, with the first row representing hate speech, the second obnoxious,

and the third representing “neither”. The fourth row of “Draw” was added to cover the case where the annotators couldn't agree on the message's class. In terms of percentage, 18% of texts were identified as hate speech, 11% as offensive, 64% as neither, and 7% as “draw.” Figure 4.3 depicts this.

Table 4.3: Preliminary annotations

| Class Label  | Count of Tweets |
|--------------|-----------------|
| Hate Speech  | 903             |
| Offensive    | 520             |
| Neither      | 3140            |
| Draw         | 368             |
| <b>Total</b> | <b>4931</b>     |

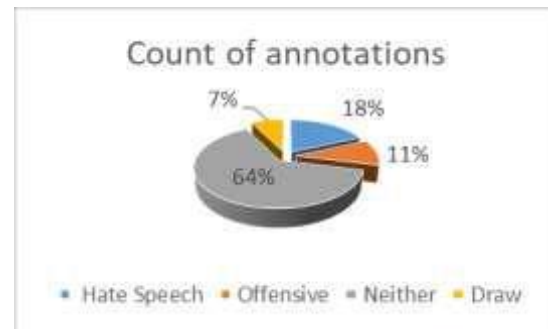


Figure 4.3: Annotated tweets

The hatefulness of the 903 messages marked as hate speech ranged from 76 percent weakly vicious to 4 percent averagely hateful to 20 percent severely hateful. Figure 4.4 shows this. In Kenya, 95% of hate speech messages were based on ethnicity, 4% on nationality, and 1% on sexual orientation, religion, gender, and disability. Figure 4.5 summarizes this.

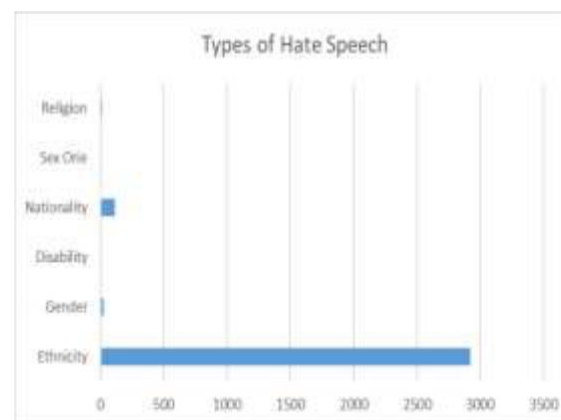
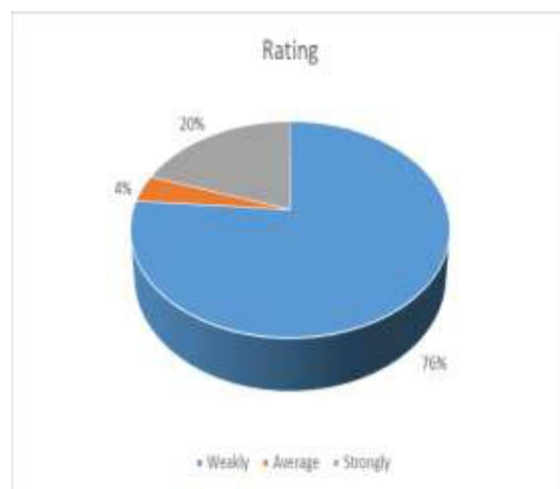


Figure 4.4: Hate speech tweets' Rating

Figure 4.5: Types of Hate speech

A fourth question looked into the three aspects of hate and how they could be utilized to identify hate speech in texts. Distancing was recognized in 41% of hate speech texts, Passion

in 30%, and dedication in 23%. Beyond the three predefined aspects, annotators recognized 'Other' features, which comprised the remaining 6%, with propaganda being the most commonly identified, followed by degrading phrases.

Figure 4.6 shows this.

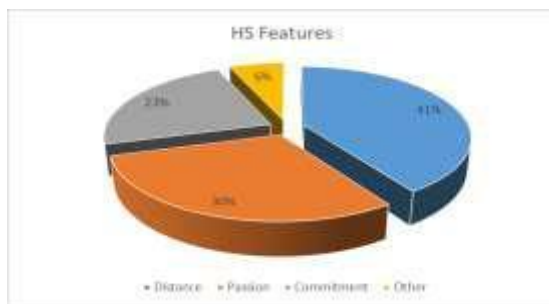


Figure 4.6 Features of Hate speech

The extent to which the 4931 messages' annotations were agreed upon by the team of annotators was measured using Krippendorff's alpha[141], which had a score of 0.5027.

#### 4.4 Data Understanding

Every tweet followed a particular pattern, which was discovered using a Panda data frame. The tweet message was encased in double quotes, the tweet ID, the tweet URL with subsections containing the username, and the tweet ID again at the end. For example,

*;2017-12-16 14:51;1;2; " Many non-Kikuyus are unfair to the Kikuyus, presuming they are all die-hard Uhuru supporters because they are arch tribalist!! "; "941999135271587840"; <https://twitter.com/hassanY78584268/status/941999135271587840>*

The distribution of message word counts is visualized in Figure 4.7, which shows the skewness of the average due to outliers.

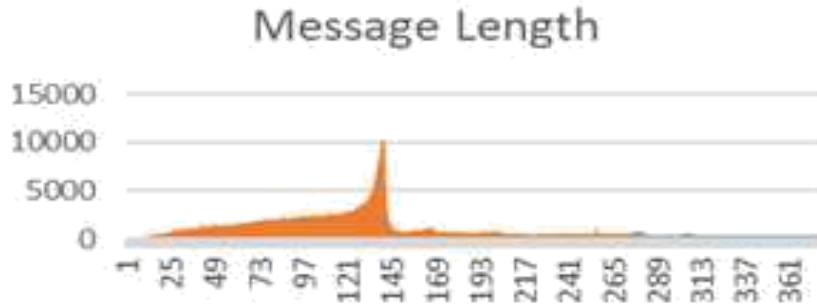


Figure 4.7: Message length

### CLASS DISTRIBUTION

The class distribution for the annotated dataset was uneven, with 75% of the tweets skewed towards the “neither” class. A summary is presented in Table 4.4 and illustrated in Figure 4.8.

Table 4.4: Class distribution

| Class        | Description | Count        |
|--------------|-------------|--------------|
| 0            | Hate Speech | 3094         |
| 1            | Offensive   | 9401         |
| 2            | Neither     | 37819        |
| <b>Total</b> |             | <b>50314</b> |

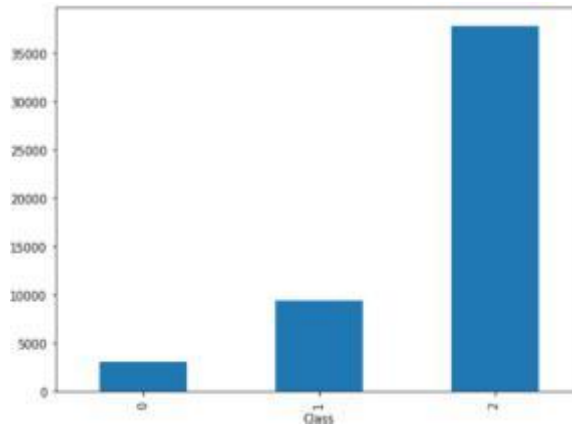


Figure 4.8: Text Class distribution

The annotated dataset's most common words were English stop words like "the," "are," "to," "and," "is," "we," "of," "a," and so on. This is depicted in Figure 4.9's histogram. Stop words frequently add to the noise signals and do not provide relevant material to aid in classifier training. As a result, they were filtered out.

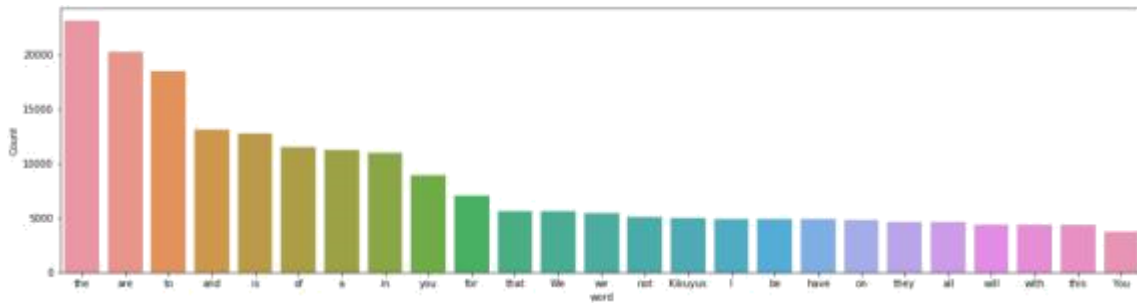


Figure 4.9: Histogram of Word frequency

After that, the dataset was cleaned by deleting stop words, punctuations, non-alphanumeric characters, and single characters, except for plural English pronouns. Ethnic group names like Kikuyus, Luos, and Kalenjins were the most common, as shown on the histogram in Figure 4.10. The names of well-known politicians like Uhuru and Raila, as well as well-known bloggers, appeared regularly. Hate, and kill, were among the most often used passion terms. #electionboycottke, #noreformsnoelections, #luolivesmatter, and were among the most popular hashtags.

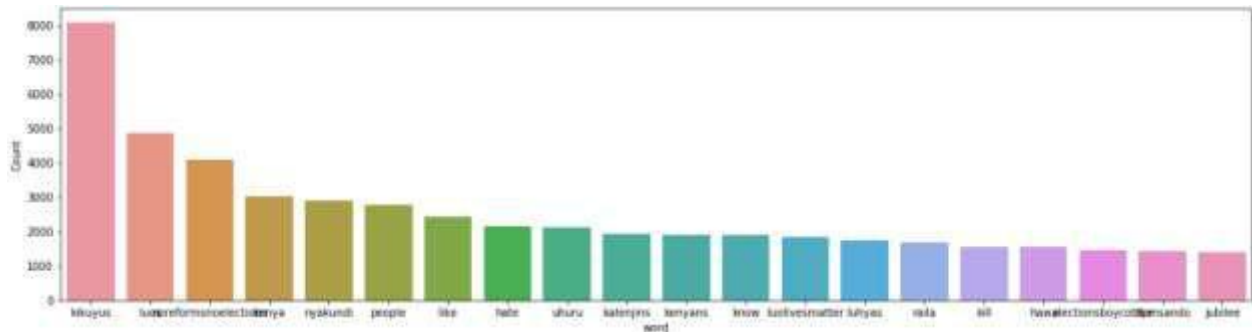


Figure 4.10: Hate Speech histogram

Furthermore, word frequency by class distribution revealed that ethnic group names were the most common throughout the three classes, with Luo, Kikuyu, Kalenjin, Kisii, Luhya, and Kamba being the most common. This is depicted in Figure 4.11 by the word cloud.

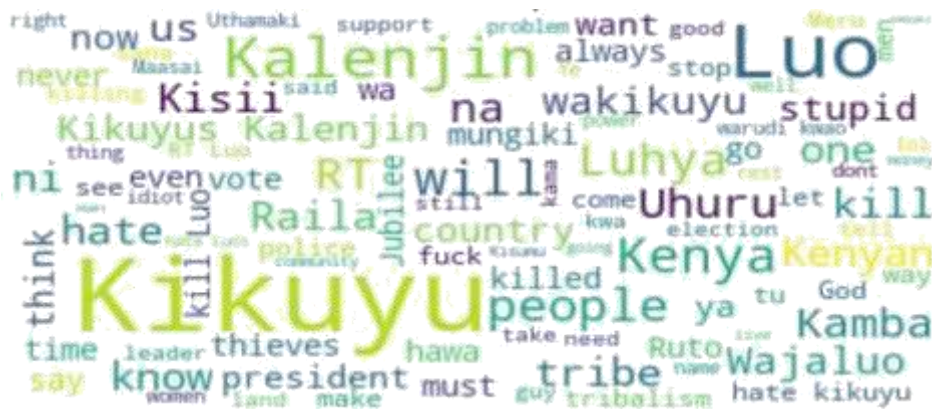


Figure 4.11: General Word frequency word cloud

After eliminating ethnic group names, PDC terms such as "kill," "thieves," "dumb," "hate," and others dominated the revised word frequency for hate speech classified messages. The word cloud in Figure 4.12 demonstrates this.

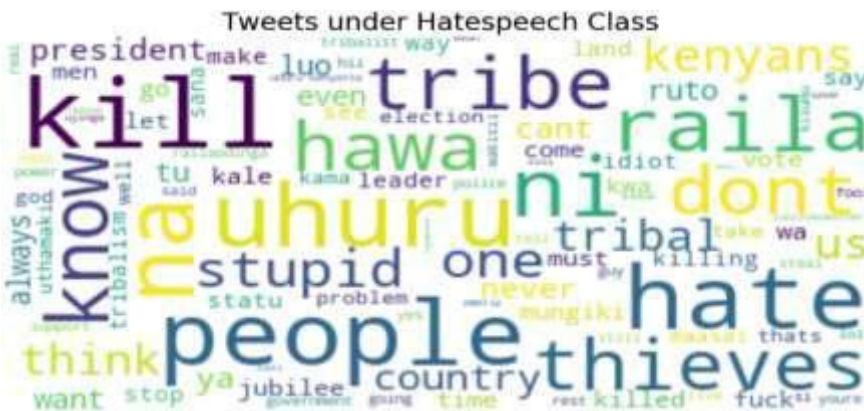


Figure 4.12: Word frequency under the Hate Speech class

The constitution and other legal provisions that cover hate speech were included in the systematic assessment of official documents. Article 33(2) subsection (c) of Kenyan law prohibits hate speech (d). It's also mentioned in relation to freedom of expression [66]. Hate speech is likewise prohibited under section 77(3)(e) of the penal code and section 13 of the NCIC Act, which prohibits discrimination based on ethnicity [57].

Hate speech based on negative ethnicity is the most common in Kenya. The post-election violence in 2007, which occurred soon after the presidential results were announced, was the pinnacle of ethnic hate in the country. The media's role in inciting the heightened tensions that preceded the violence was clear [150].



Content analysis was conducted on multiple definitions of hate speech published in top social media networks' user policy guidelines, research articles, and dictionary-identified standard phrases, as shown in Table 2.1.

Furthermore, the researcher was able to gain a more in-depth understanding of hate speech markers and the procedures that the relevant authorities planned to utilize to record evidence of hate speech by using the participant observation strategy. This was accomplished by attending the National Cohesion and Integration Commission's (NCIC)[130] hate speech human monitors training workshop in March 2017, which was held in collaboration with Kenya's Communication Authority (CA) and the Kenya Police in preparation for monitoring and collecting evidence for hate speech during the campaign periods leading up to the 2017 general elections [148]. The majority of the surveillance was done manually, employing voice recorders, video cameras, and manually scouring popular social media sites for records of political gatherings.

Furthermore, the researcher's interactions with government agencies such as the National Cohesion and Integration Commission, which is in charge of hate speech issues [130], and the Kenya Education Network, which is the primary Internet Service Provider for all tertiary learning institutions in the country [131], provided additional insight into the difficulties of monitoring the phenomenon. A working definition of hate speech was established from this phase, which is any statement that discriminates, devalues, or employs offensive language against a person or a group of persons based on protected characteristics such as race, ethnicity, religion, gender, and so on.

The unigrams, bigrams, and trigrams that were most connected to the hate speech class (0), offensive class (1), and neither class were discovered by exploratory data analysis (2). Hate speech was defined by the existence of expressions that reflect distancing or othering language, negative passion, dedication to devaluation, and propaganda, according to the study's conceptual framework. Negative passion can be seen in expressions like "*stupid*," "*fuck*," "*kihii*," and so on. The pronoun phrases "*Hawa*," "*wewe*," and the tribe names are also used frequently as distancing words. Trigrams like "*Uhuru snatches everything*," "*Yule jamaa wa vitendawili*," and others show a commitment to promoting hatred. In addition, the presence of various Swahili and native phrases like as "*hawa*," "*ni wajinga sana*," "*kihii*," and others demonstrates the codeswitching phenomena (correlation without ethnic). '*ni*,' '*na*,' '*ya*,' '*wa*,' were common, but these are English equivalent

Stopwords "is" and "of," respectively. This is a great example of how existing standard libraries, such as Stopwords, would fail to capture noise in codeswitched text data. These Swahili Stopwords were included in the inclusive set of Stopwords during data cleaning but were later removed because they didn't provide any useful information to the classification process.

#### 4.4.1 Learning a class from examples

The purpose in this scenario was to learn the class "hate speech" after receiving a text message (a tweet). There were other examples of tweets that had already been classified as hate speech, inflammatory, or neither by the team of human annotators. The study's annotation framework, which is based on three key characteristics of negative passion, distancing language, and dedication to hate, influenced their decision (PDC). Distancing is further subdivided into two main categories: discrimination and the use of othering language. Protected social groups, such as the Kikuyus, Luos, and Kalenjin, are included in the discrimination category. The category of othering language contains common noun pairings such as "us - them; we – they." Insults, threats, derogatory phrases, and other offensive terms such as "fuck, stupid, murder, chase, etc." make up negative passion. Subjectivity consists of one-sided, one-sided arguments that cannot be supported and hence become propaganda. Expressions of certainty, generalization, and devaluation are frequently used to demonstrate a commitment to hate. Words like 'never,' 'always,' and 'always,' when used in a message, are unambiguous markers of certainty. Phrase patterns that begin with 'all tribe>...' show generalization. The use of dehumanizing language to refer to the target is known as devaluation. For example, "cockroaches, foreskin, maggots, etc." are insect, item, or animal terminology. Figure 5.21 shows a Venn diagram that nicely summarizes these points. As a result, the human annotator examines a tweet for indicators of distance (D) and passion (P) or commitment (C). Hate speech is based on D+P, D+C, or D+P+C, in which a person or group is targeted based on their belonging to a protected trait such as ethnicity. "Uhuru Kenyatta is a hopeless alcoholic," for example. "We've had enough of this Kikuyu president." The fact that the president is a "Kikuyu," a Kenyan ethnic minority, qualifies this as hate speech. Table 2.10 summarizes the whole list of characteristic combinations that result in hate speech.

Offensive speech, like hate speech, can be based on any of the three combinations without explicitly or indirectly addressing a protected social trait. "Uhuru Kenyatta is a hopeless alcoholic,"

for example. We've had enough of him. “ Although the notion will be considered insulting, it will not be considered hate speech.

Any message that could not fit within these parameters was labeled "neither." In theory, when traits are distinctive to a class rather than universal, class learning is at its best. All instances of a class share the feature description, but none of the competing classes do [80]. However, using the Chi-square significance test, the study of the covariance structure of the unigrams, bigrams, and trigrams for the three classes revealed a different pattern than previously anticipated. Ethnic names were common in all three groups, with Kikuyu and Luo (together with their Swahili language equivalents, i.e. Wakikuyu, Wajaluo) being the most common in that order. Figure 4.13 illustrates this. As a result, ethnic names were not a powerful feature to employ to train the classifier to discriminate between the classes on their own. This was counter to our earlier assumptions; yet, ethnic names were useful when combined with the other notions, particularly in identifying the subject of hatred, so they couldn't be completely dismissed.

| Hate speech  | Offensive  | Neither  |
|--|--|--|
| # '0':   | # '1':   | # '2':   |
| <ul style="list-style-type: none"> <li>. Most correlated unigrams:</li> <li>. fuck</li> <li>. hulligans</li> <li>. thieves</li> <li>. noreformsnoelections</li> <li>. stupid</li> <li>. kill</li> <li>. kikuyus</li> <li>. Most correlated bigrams:</li> <li>. ni kihii</li> <li>. kikuyus think</li> <li>. kill kikuyus</li> <li>. small dicks</li> <li>. luo kill</li> <li>. kill luo</li> <li>. kikuyus kalenjin</li> <li>. Most correlated Trigrams:</li> <li>. ruto join raila</li> <li>. kenya kikuyus kalenjin</li> <li>. presid kikuyus kalenjin</li> <li>. tribal kikuyus sue</li> <li>. luo tribal kikuyus</li> <li>. stones uon students</li> <li>. throw stones uon</li> </ul> | <ul style="list-style-type: none"> <li>. Most correlated unigrams:</li> <li>. luhya</li> <li>. kihii</li> <li>. ni</li> <li>. kikuyus</li> <li>. uthamaki</li> <li>. uhuru</li> <li>. luo</li> <li>. Most correlated bigrams:</li> <li>. uhuru kill</li> <li>. dont fuck</li> <li>. wacha ujinga</li> <li>. wee kihii</li> <li>. kikuyus think</li> <li>. kill luo</li> <li>. ni kihii</li> <li>. Most correlated Trigrams:</li> <li>. ya mama yenu</li> <li>. kenya si ya</li> <li>. luo luhya kisii</li> <li>. polic kill luo</li> <li>. aliv mourn madiba</li> <li>. wew ni kihii</li> <li>. matiba aliv mourn</li> </ul> | <ul style="list-style-type: none"> <li>. Most correlated unigrams:</li> <li>. peace</li> <li>. great</li> <li>. uhuru</li> <li>. username</li> <li>. kikuyus</li> <li>. mara</li> <li>. maasai</li> <li>. Most correlated bigrams:</li> <li>. peace building</li> <li>. tribal sue</li> <li>. gm masai</li> <li>. maasai tribe</li> <li>. warudi kwao</li> <li>. masai mara</li> <li>. maasai mara</li> <li>. Most correlated Trigrams:</li> <li>. shouldnt deny joy</li> <li>. tweet hardcore kiuk</li> <li>. bump maize flour</li> <li>. points bump maize</li> <li>. hardcore kiuk words</li> <li>. gm masai ujiri</li> <li>. maasai mara university</li> </ul> |

Figure 4.13: Chi-square for correlation

#### 4.4.2 Probabilistic Hierarchical Modelling of Hate Speech

Given the magnitude of the social media codeswitched text corpus, it was necessary to breakdown it in order to lower the high dimensional feature space visible in such text and, as a result, the sparseness of the feature vector. Based on the notions of the study's conceptual framework, this was accomplished by developing a realistic approach for exploring, searching, sorting, and reducing the enormous input feature space to a smaller subset of low dimensional and high-quality features. This information was then used to train a machine to classify future cases as hate speech, offensive speech, or neither. The main goal was to figure out the underlying pattern and statistical links between the words so that the text categorization task could be more accurate. This was accomplished by applying a generative probabilistic model to the text corpus, which can hierarchically organise the corpus into informative topics or word clusters based on likely parallels or relationships to the corpus membership [151].

The Latent Dirichlet Allocation (LDA) model was utilized to discover deep underlying notions of hate in a large corpus of code switched text using topic modeling[92]. LDA, a hierarchical probabilistic model, has previously been used to successfully identify cyberbullying-related subjects [26]. Each word in the corpus is represented by LDA as a finite mixture of underlying Passion, Distancing, and Commitment (PDC) subjects, which are modeled over an unlimited number of topics characteristic of a text document [92]. This aids in the development of a probabilistic model for the codeswitched corpus, which will give high probabilities to messages that are strongly related to the corpus' membership and other messages that are comparable to them. As a result, LDA was utilized specifically to extract a "bag of words" from twenty-three latent subjects closely connected with the hate speech class and bearing the study's conceptual framework's feature characteristics. These are listed in Table 4.5's twenty-three rows. The green cells denote a legally protected trait, in this case the names of Kenyan ethnic groups and nationality. Individual names are included in purple cells, mostly presidential candidates/politicians and one well-known blogger. The blue cells are also groupings, however they do not come into the category of protected characteristics. Police, government, country, and nation are examples of these terms. The yellow cells represent "distancing" or "othering" characteristics that are frequently characterized by the use of pronouns. The "passion" qualities, which are defined by destructive and insulting phrases, are represented by red cells. Table 4.5

depicts a combination of the passion, distancing, and commitment traits, which correspond to the salient feature in the hate speech conceptual framework produced in this study. As a result, the LDA method was useful in swiftly exploring and uncovering the inherent PDC theme structure, which is common in hate speech, in a large corpus of text messages.

The use of LDA, on the other hand, revealed the limits of the bag-of-words technique, which does not keep word order and hence does not preserve word meaning or context. In terms of text classification, relying on LDA as the primary strategy proved insufficient. Regardless, it was highly beneficial in data preparation and as a first-level statistical strategy in automatically detecting and extracting passion, distancing, and discriminative (PDC) features from the huge corpus in our work. The algorithm learned these themes based on the deep underlying concepts in social media big data, which appear to mimic the PDC features explained in the study's conceptual model.

Table 4.5: Topic modeling for hate speech class

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

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

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

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

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

## 4.5 Model Training and Evaluation

An end-to-end pipeline was created to provide consistency and quality assurance in the many machine learning experiments that were undertaken to generate the hate speech models. This included the multiple CRISP-DM phases, which were divided into two parts: data pretreatment (component 1) and dimensionality reduction (component 2). The initial part involved gathering raw data from social media, tokenizing text messages, and filtering out noise by removing Stopwords, punctuation, duplication, and non-alphanumeric characters, among other things. In addition, the tokens were lowercased and stemmed for normalization. The major output of the first component's actions is a manageable feature set free of noisy and redundant features from the enormous raw dataset.

The second part involved feature selection and extraction, which was preceded by the dataset being separated into training and testing sets. The most important psychosocial, linguistic, and app-specific characteristics were chosen. The PDC traits dominated the psycho-social features. The goal was to gradually scale up a set of quality tokens comprised of three subsets of psycho-social variables, namely negative passion features, social distance features, and commitment-to-hate features, by learning the PDC language. In section 2.6.1, the individual PDC properties are thoroughly explained. The linguistic features that were chosen were mostly lexical features from the PDC vocabulary. During the preliminary experiments, other linguistic features such as the part of speech, syntactic features such as capitalization, and punctuations such as exclamation marks were employed solely. However, because their impact on classifier accuracy was minor in compared to the more significant PDC features, these features were eliminated during subsequent studies at the preprocessing step. The value of numerous properties is thoroughly explained in section 4.6.1, based on the feature experiments done. Following that, the features were organized into n-grams and processed at multiple levels, including phrase, word, and character levels. The n-grams were then translated into low-level features, such as TF-IDF vectors, resulting in a reduction in the native feature space's dimensionality. Following that, the dense vector representation of the characteristics was fed into multiple machine learning methods to learn the corresponding classifier models. The learning process was iterative and grid search was used to find the optimum model, which was based on the classifier with the best accuracy performance. Following that, the test set was utilized to evaluate the accuracy of the best-learned classifier model, which had also gone through the second component as indicated by the red arrows in Figure 4.14. This final

classifier was used to predict the class of incoming input messages, each of which had to go through both data preparation and dimensionality reduction components. Figure 4.14 depicts the full procedure.

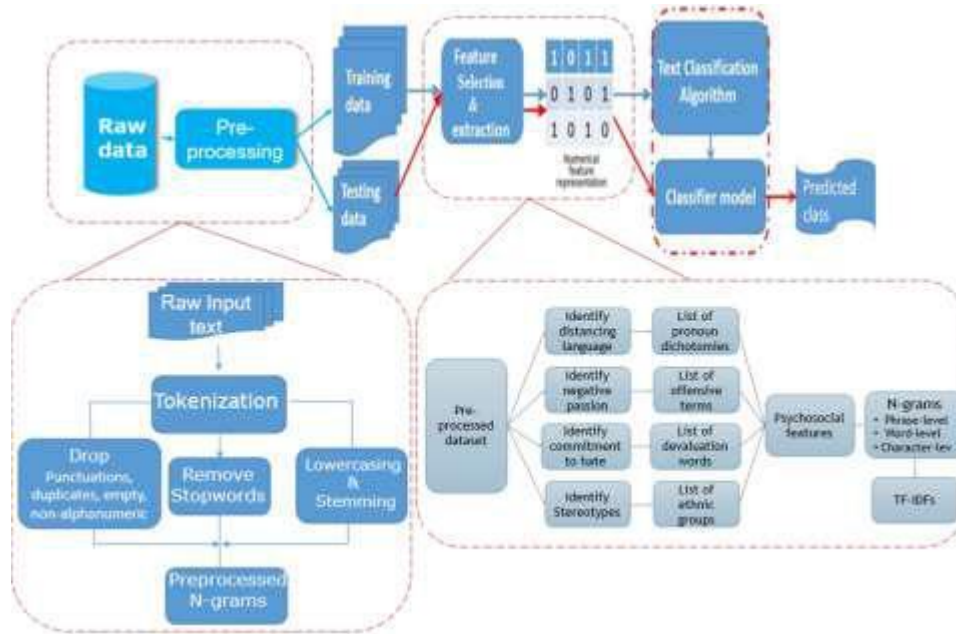


Figure 4.14: Hate Speech Classification Model's End-to-End Pipeline

#### 4.5.1 Experimental Results and Findings

The various experiments were carried out with two basic goals in mind. The first goal was to see if the novel psychosocial traits described in the study helped distinguish hate speech, particularly the nuanced forms that conventional features failed to detect [78]. Second, tests were conducted to assess the efficacy of several machine learning algorithms trained on the psychosocial feature set as well as other traditional features in constructing a more broad hate speech classification model. The performance of earlier studies in automatic hate speech identification guided the selection of machine learning algorithms and conventional features to utilize in the tests. To do so, nine classification models were trained using both traditional and innovative psychological variables (PDC). Following that, the models' accuracy results were evaluated to determine the best model and characteristics for hate speech classification of codeswitched text messages from social media.

### 4.5.1.1 Significance of Features

The accuracies acquired per class using the One-versus-All (OVA) framework are shown in the results from each classifier. Table 4.6 lists the Nave Bayes classifier, Table 4.7 lists the linear Support vector classifier, Table 4.8 lists the Logistic Regression classifier, and Table 4.9 lists the Random Forest classifier. For each classifier, a comparison of accuracy performance across multiple features and feature combinations is done. The basic goal was to determine the feature importance of the various representations that would result in the highest projected accuracies in identifying hate speech across all classifiers.

The LEX-PDC feature was the feature that resulted in the highest prediction accuracies in recognizing hate speech using the Nave Bayes classifier. The accuracy of this feature combination was 76.5 percent. The PDC was the finest single feature. Table 4.6 illustrates this.

Table 4.6: Naïve Bayes classifier performance

|    | Feature Combo   | Mean Accuracy       | Hate                      | Offensive           | Neither |
|----|-----------------|---------------------|---------------------------|---------------------|---------|
| 0  | Lex             | 0.5571546052631579  | 0.6230110159118727        | 0.5312883435582823  | 0.51625 |
| 1  | pdC             | 0.5131578947368421  | <u>0.6242350061199511</u> | 0.645398773006135   | 0.265   |
| 2  | pos             | 0.3852796052631579  | 0.3157894736842105        | 0.4343558282208589  | 0.40625 |
| 3  | App             | 0.39144736842105265 | 0.1481028151774786        | 0.7128834355828221  | 0.3125  |
| 4  | Lex-pdc         | 0.5777138157894737  | <u>0.7649938900489596</u> | 0.4785276073619632  | 0.4875  |
| 5  | Lex-pos         | 0.4798519736842105  | 0.4675642594859241        | 0.4723926380368098  | 0.5     |
| 6  | Lex App         | 0.5444078947368421  | 0.5593635250917993        | 0.5693251533742332  | 0.50375 |
| 7  | pdC-pos         | 0.45394736842105265 | 0.4479804161566707        | 0.4368098159509202  | 0.4775  |
| 8  | pdC App         | 0.5439967105263158  | 0.680538556915544         | 0.49325153374233127 | 0.45625 |
| 9  | pos-App         | 0.3959703947368421  | 0.31701346389228885       | 0.4245398773006135  | 0.4475  |
| 10 | Lex-pdc-pos     | 0.5176809210526315  | 0.5520195838433293        | 0.46748466257668714 | 0.53375 |
| 11 | Lex-pdc-App     | 0.5736019736842105  | <u>0.7086903304773562</u> | 0.5116564417177915  | 0.49875 |
| 12 | Lex-pos-App     | 0.4786184210526316  | 0.4479804161566707        | 0.47116564417177914 | 0.5175  |
| 13 | pdC-pos-App     | 0.46175986842105265 | 0.44430844553243576       | 0.42822085889570555 | 0.51375 |
| 14 | Lex-pdc-pos-App | 0.5189144736842105  | 0.5385556915544676        | 0.47116564417177914 | 0.5475  |

PDC-APP had the best feature combination for the Linear Support Vector classifier, with an accuracy of 78.9%. However, the PDC feature, with an accuracy of 82.9 percent, beat all other features in terms of predicting hate speech as a single feature. Table 4.7 depicts this.

Table 4.7: Performance of the Linear Support Classifier



|    | Feature Combo   | Mean Accuracy       | Hate                      | Offensive           | Neither |
|----|-----------------|---------------------|---------------------------|---------------------|---------|
| 0  | Lex             | 0.5394736842105263  | 0.605875152998776         | 0.4822085889570552  | 0.53    |
| 1  | pdC             | 0.5390625           | <u>0.828641370869033</u>  | 0.1558282208588957  | 0.63375 |
| 2  | pos             | 0.3782894736842105  | 0.3427172582619339        | 0.42208588957055215 | 0.37    |
| 3  | App             | 0.4354440789473684  | 0.30354957160342716       | 0.5239263803680981  | 0.48    |
| 4  | Lex-pdc         | 0.5398848684210527  | 0.6144430844553244        | 0.4822085889570552  | 0.5225  |
| 5  | Lex-pos         | 0.49917763157894735 | 0.5507955936352509        | 0.44785276073619634 | 0.49875 |
| 6  | Lex-App         | 0.5394736842105263  | 0.6009791921664627        | 0.48466257668711654 | 0.5325  |
| 7  | pdC-pos         | 0.4975328947368421  | 0.6340269277845777        | 0.4012269938650307  | 0.45625 |
| 8  | pdC-App         | 0.5616776315789473  | <u>0.7894736842105263</u> | 0.4134969325153374  | 0.48    |
| 9  | pos-App         | 0.41241776315789475 | 0.35862913096695226       | 0.4208588957055215  | 0.45875 |
| 10 | Lex-pdc-pos     | 0.5131578947368421  | 0.572827417380661         | 0.4638036809815951  | 0.5025  |
| 11 | Lex-pdc-App     | 0.5448190789473685  | 0.6119951040391677        | 0.4920245398773006  | 0.53    |
| 12 | Lex-pos-App     | 0.4917763157894737  | 0.5324357405140759        | 0.44785276073619634 | 0.495   |
| 13 | pdC-pos-App     | 0.5069901315789473  | 0.6217870257037944        | 0.4110429447852761  | 0.4875  |
| 14 | Lex-pdc-pos-App | 0.5123355263157895  | 0.5642594859241126        | 0.4736196319018405  | 0.49875 |

The best feature combination, similar to the Linear SVC, was PDC-APP, which had an accuracy of 74.4 percent, with the PDC feature's accuracy of 79.1 percent delivering the highest prediction accuracy in identifying hate speech as a single feature. Table 4.8 illustrates this.

Table 4.8: Performance of the Logistic Regression classifier

|    | Feature Combo   | Mean Accuracy       | Hate                      | Offensive           | Neither |
|----|-----------------|---------------------|---------------------------|---------------------|---------|
| 0  | Lex             | 0.5740131578947368  | 0.6695226438188494        | 0.5042944785276073  | 0.5475  |
| 1  | pdC             | 0.5419407894736842  | <u>0.7906976744186046</u> | 0.1754601226993865  | 0.66125 |
| 2  | pos             | 0.3774671052631579  | 0.3402692778457772        | 0.4294478527607362  | 0.3625  |
| 3  | App             | 0.43133223684210525 | 0.32558139534883723       | 0.47116564417177914 | 0.49875 |
| 4  | Lex-pdc         | 0.5805921052631579  | 0.7209302325581395        | 0.4883435582822086  | 0.53125 |
| 5  | Lex-pos         | 0.5254934210526315  | 0.5801713586291309        | 0.4822085889570552  | 0.51375 |
| 6  | Lex-App         | 0.5744243421052632  | 0.6609547123623011        | 0.49570552147239266 | 0.56625 |
| 7  | pdC-pos         | 0.5037006578947368  | 0.6230110159118727        | 0.42578687116564416 | 0.46125 |
| 8  | pdC-App         | 0.5563322368421053  | <u>0.7441860465116279</u> | 0.4171779141104294  | 0.50625 |
| 9  | pos-App         | 0.4144736842105263  | 0.3623011015911873        | 0.4134969325153374  | 0.46875 |
| 10 | Lex-pdc-pos     | 0.5357730263157895  | 0.6266829865361077        | 0.4662576687116564  | 0.51375 |
| 11 | Lex-pdc-App     | 0.584703947368421   | 0.7086903304773562        | 0.47975460122699387 | 0.565   |
| 12 | Lex-pos-App     | 0.524671052631579   | 0.5891554467564259        | 0.4638036809815951  | 0.54125 |
| 13 | pdC-pos-App     | 0.5148026315789473  | 0.5936352509179926        | 0.4294478527607362  | 0.52125 |
| 14 | Lex-pdc-pos-App | 0.5349506578947368  | 0.6083231334149327        | 0.456441717791411   | 0.54    |

The LEX-PDC characteristics were the best feature combination for the Random Forest classifier, with an accuracy of 62.8 percent. The PDC feature remained the best single feature, with a 72.2 percent accuracy rate in detecting hate speech. Table 4.9 illustrates this.

Table 4.9: Performance of the Random Forest classifier

| Feature Combo      | Mean Accuracy       | Hate                      | Offensive           | Neither             |
|--------------------|---------------------|---------------------------|---------------------|---------------------|
| 0 Lex              | 0.5254001391788448  | 0.5778210116731517        | 0.6963906581740976  | 0.28761061946902655 |
| 1 pdc              | 0.592205984690327   | <u>0.7217898832684825</u> | 0.22717622080679406 | 0.8252212389380531  |
| 2 pos              | 0.40988169798190677 | 0.21595330739299612       | 0.5753715498938429  | 0.4579646017699115  |
| 3 App              | 0.44258872651356895 | 0.19844357976853895       | 0.4819532908704883  | 0.6792035398230089  |
| 4 Lex-pdc          | 0.5247042446547689  | <u>0.6284046682607004</u> | 0.6242038216560509  | 0.3030973451327434  |
| 5 Lex-pos          | 0.5594989561586639  | 0.4688715953307393        | 0.6518046709129511  | 0.5663716814159292  |
| 6 Lex-App          | 0.5434933890048712  | 0.3521400778210117        | 0.613588110403397   | 0.6890530973451328  |
| 7 pdc-pos          | 0.5581071677105081  | 0.5739299610894941        | 0.5222929936305732  | 0.577433628318584   |
| 8 pdc-App          | 0.5518441196938065  | 0.4708171206225681        | 0.4607218683651805  | 0.7389380530973452  |
| 9 pos-App          | 0.4523312456506611  | 0.11967704280155641       | 0.6050955414012739  | 0.672566371681416   |
| 10 Lex-pdc-pos     | 0.5977731384829506  | 0.8050583657587548        | 0.5902335458475584  | 0.5973451327433629  |
| 11 Lex-pdc-App     | 0.5859429366736256  | 0.5019455252918288        | 0.564756838641189   | 0.7035398230088495  |
| 12 Lex-pos-App     | 0.5288798102992345  | 0.3151750972762846        | 0.6389426751592356  | 0.6592920353982301  |
| 13 pdc-pos-App     | 0.546276965901183   | 0.39883268482490275       | 0.5668789808917197  | 0.6924778761061947  |
| 14 Lex-pdc-pos-App | 0.5768963117606124  | 0.4688715953307393        | 0.5711252653927813  | 0.7057522123893806  |

According to the findings, psychosocial variables, such as PDC, were not only informative but also surpassed most traditional predictors in detecting hate speech.

#### 4.5.1.2 A Model for Machine Classification Based on PDC Features

The PDC psychosocial feature set was used to train seven conventional and two deep learning models. The high-level PDC features were translated into three lower-level representations to determine the most effective feature representation: BoW as count vectors, n-grams as TF-IDF vectors, and word embedding as dense vectors. Figure 4.15 illustrates this.

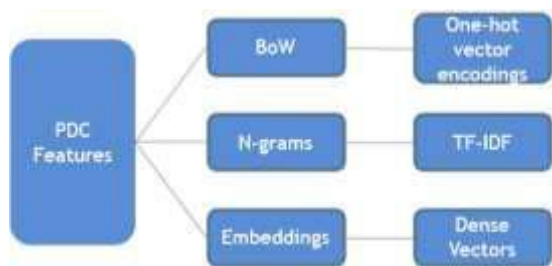


Figure 4.15: Mapping of PDC feature to low-level features

The TF-IDF vectors were also created using three different n-gram levels, namely phrase, word, and character-level n-grams, with n=3 delivering the best results. Table 4.10 shows the psychological feature (PDC) performance of various representations in learning classification models from various machine learning techniques. The names of the features are displayed in the first column of the table, while the names of the individual machine learning methods are displayed in the subsequent table columns. The Naive Bayes, Support Vector Machine, Linear logistic regression, K-Nearest Neighbor, Random Forest bagging technique, Decision Tree, Hierarchical

Attention Network, Convolutional Neural Networks, and Extreme Gradient Boosting algorithm are examples of these algorithms.

Table 4.10: Performance of Features based on nine classification models

| FEATURE               | Machine Learning Algorithms |                         |              |       |       |                 |              |       |       |
|-----------------------|-----------------------------|-------------------------|--------------|-------|-------|-----------------|--------------|-------|-------|
|                       | Naïve Bayes                 | Lin logistic Regression | SVM          | KNN   | DT    | Bagging RForest | Boosting Xgb | HAN   | CNN   |
| Count vectors         | 0.789                       | <b>0.824</b>            | 0.798        | 0.790 | 0.801 | 0.814           | 0.810        |       |       |
| <b>TF-IDF Vectors</b> |                             |                         |              |       |       |                 |              |       |       |
| Word level            | 0.808                       | 0.817                   | <b>0.822</b> | 0.801 | 0.801 | 0.812           | 0.812        |       |       |
| N-gram level          | 0.804                       | 0.804                   | <b>0.806</b> | 0.772 | 0.795 | 0.789           | 0.799        |       |       |
| Character level       | 0.805                       | 0.821                   | <b>0.825</b> | 0.787 | 0.798 | 0.804           | 0.814        |       |       |
| Word Embeddings       |                             |                         |              |       |       |                 |              | 0.600 | 0.672 |

The GloVe vector representations dataset, which has a 1193514-word vector, was used to create the word embeddings features. Only the deep learning techniques, such as Hierarchical Attention Networks (HAN) and Convolutional Neural Network (CNN) models, were referred to as embeddings. With an accuracy of 82.4 percent, the linear logistic regression model performed best for the BoW count vectors. The support vector classifier model had the best performance of 82.2 percent, 80.6 percent, and 82.5 percent for TF-IDF feature vectors at the word, N-gram, and character levels, respectively. Using word embedding as a feature, the accuracy of the HAN and CNN deep learning models was 60% and 67.2 percent, respectively. Overall, the support vector classifier outperformed the others, with TF-IDF feature vectors surpassing the rest at the character level. The most convincing TF-IDF features for a codeswitched dataset, specifically where  $n=3$ , are character-level TF-IDF features, according to this study. Furthermore, in codeswitched datasets, word embeddings do not perform as well. This could be explained by the fact that the current external word embeddings were primarily built using properly structured vocabulary. When applied to a codeswitched dataset, it produces a highly sparse feature vector, which the study set out to solve from the start.

## 4.5.2 Model Parameter Tuning

The Grid Search technique in Scikit-Learn [60] was used to automatically update and discover the best parameters for each model based on a parameter grid supplied in a pipeline. The Support Vector Classifier, Naive Bayes, Logistic Regression, and Random Forest models were the most promising. The alpha, soft margin cost, solver, and estimator parameters were all specified for the corresponding models. Figure 4.16 illustrates this.

```
mnb_params={'clf_alpha': (1e-1, 1e-2, 1e-3, 1),}
svc_params={'clf_c': [0.001, 0.1, 10, 100, 10e5], }
lrc_params={'clf_solver': ('liblinear', 'sag', 'saga', 'newton-cg'),}
rfc_params={'clf_n_estimators': (100, 200, 300), 'clf_max_depth': (3, 5, 8),}
```

Figure 4.16: Model Parameter Grid

The parameter tuning results using grid-search with the Naïve Bayes model are shown in Table 4.11.

Table 4.11: Naïve Bayes with Grid Search

```
mnb_clf = GridSearchCV(mnb_clf, mnb_params, cv=5, iid=False, n_jobs=-1)
mnb_clf=model.train(mnb_clf)
model.get_model_feature_metrics(mnb_clf)
```

|    | Feature Combo   | Mean Accuracy       | Hate                | Offensive           | Neither             |
|----|-----------------|---------------------|---------------------|---------------------|---------------------|
| 0  | Lex             | 0.5859429366736256  | 0.6303501945525292  | 0.5774946921443737  | 0.5442477876106194  |
| 1  | pdv             | 0.44676409185803756 | 0.5642023346303502  | 0.6199575371549894  | 0.13274336283185842 |
| 2  | pos             | 0.35073068893528186 | 0.28599221789883267 | 0.4182590233545648  | 0.35398230088495575 |
| 3  | App             | 0.44328462073764785 | 0.15953307392996108 | 0.5859872611464968  | 0.6172566371681416  |
| 4  | Lex-pdv         | 0.6089074460681977  | 0.7587548638132295  | 0.4968152866242038  | 0.5553097345132744  |
| 5  | Lex-pos         | 0.5163535142658316  | 0.5116731517509727  | 0.4989384288747346  | 0.5398230088495575  |
| 6  | Lex App         | 0.6019485038274183  | 0.5525291828793775  | 0.6220806794055201  | 0.6371681415929203  |
| 7  | pdv-pos         | 0.41892832289492    | 0.40077821011673154 | 0.43524416135881105 | 0.4225663716814159  |
| 8  | pdv-App         | 0.592205984690327   | 0.5739299610894941  | 0.4543524416135881  | 0.7586371681415929  |
| 9  | pos-App         | 0.42171189979123175 | 0.2821011673151751  | 0.416135881104034   | 0.5862831858407079  |
| 10 | Lex-pdv-pos     | 0.5414057063326374  | 0.5544747081712063  | 0.5074309978768577  | 0.5619469026548672  |
| 11 | Lex-pdv-App     | 0.6346555323590815  | 0.6984435797665369  | 0.5477707006369427  | 0.6526548672566371  |
| 12 | Lex-pos-App     | 0.5385344467640919  | 0.47470817120622566 | 0.4819532908704883  | 0.8637168141592921  |
| 13 | pdv-pos-App     | 0.4857341684064022  | 0.4046692607003891  | 0.43736730360934184 | 0.6283185840707964  |
| 14 | Lex-pdv-pos App | 0.5678496868475992  | 0.5428015564202334  | 0.49256900212314225 | 0.6747787610619469  |

Table 4.12 shows the results of parameter adjustment using grid-search in conjunction with the Support Vector Classifier.

Table 4.12: Grid Search plus Support Vector Classifier

```

svc_clf = Pipeline([('clf', LinearSVC(max_iter=1200)),])
svc_clf = GridSearchCV(svc_clf, svc_params, cv=5, iid=False, n_jobs=-1)
model.train(svc_clf)
model.get_model_feature_metrics(svc_clf)

```

|    | Feature Combo   | Mean Accuracy       | Hate                | Offensive           | Neither            |
|----|-----------------|---------------------|---------------------|---------------------|--------------------|
| 0  | Lex             | 0.6353514265831594  | 0.7101167315175098  | 0.5180467091295117  | 0.672566371681416  |
| 1  | pdv             | 0.60473208072373    | 0.7859922178988327  | 0.208067940552017   | 0.8119469026548672 |
| 2  | pos             | 0.42379958246346555 | 0.2723735408560311  | 0.4989384288747346  | 0.5176991150442478 |
| 3  | App             | 0.4405010438413361  | 0.17120622568093385 | 0.45222929936305734 | 0.7345132743362832 |
| 4  | Lex-pdv         | 0.6541405706332637  | 0.7626459143968871  | 0.4182590233545648  | 0.7765486725663717 |
| 5  | Lex-pos         | 0.60473208072373    | 0.6089494163424124  | 0.5031847133757962  | 0.7057522123893806 |
| 6  | Lex App         | 0.6353514265831594  | 0.632295719844358   | 0.46496815286624205 | 0.8163716814159292 |
| 7  | pdv-pos         | 0.5852470424495476  | 0.6770428015564203  | 0.35668789808917195 | 0.7190265486725663 |
| 8  | pdv App         | 0.6346555323590815  | 0.7879377431906615  | 0.3375796178343949  | 0.7699115044247787 |
| 9  | pos App         | 0.4523312456506611  | 0.2237354085603113  | 0.4819532908704883  | 0.6814159292035398 |
| 10 | Lex-pdv-pos     | 0.6290883785664579  | 0.6770428015564203  | 0.4543524416135881  | 0.7566371681415929 |
| 11 | Lex-pdv App     | 0.6437021572720947  | 0.7042801556420234  | 0.40976645435244163 | 0.8185840707964602 |
| 12 | Lexpos App      | 0.6019485038274183  | 0.5680933852140078  | 0.4267515923566879  | 0.8230088495575221 |
| 13 | pdv-pos App     | 0.6005567153792624  | 0.6595330739299611  | 0.38004246284501064 | 0.7632743362831859 |
| 14 | Lex-pdv-pos App | 0.6242171189979123  | 0.6556420233463035  | 0.40552016985138006 | 0.8163716814159292 |

Table 4.13 summarizes the outcomes of parameter adjustment using grid-search and the Support Vector Classifier.

Table 4.13: Grid Search with Logistic Regression

```

lrc_clf = Pipeline([('clf', LogisticRegression(random_state=0,solver='lbfgs',multi_class='auto')),])
lrc_clf = GridSearchCV(lrc_clf, lrc_params, cv=5, iid=False, n_jobs=-1)
model.train(lrc_clf)
model.get_model_feature_metrics(lrc_clf)

```

|    | Feature Combo   | Mean Accuracy      | Hate                | Offensive           | Neither            |
|----|-----------------|--------------------|---------------------|---------------------|--------------------|
| 0  | Lex             | 0.6346555323590815 | 0.7042801556420234  | 0.5307855626326964  | 0.6837168141592921 |
| 1  | pdv             | 0.6082115518441197 | 0.7723735408560312  | 0.23142250530785563 | 0.8141592920353983 |
| 2  | pos             | 0.4286708420320111 | 0.3151750972762646  | 0.416135881104034   | 0.5707964601769911 |
| 3  | App             | 0.453027139874739  | 0.22957198443579765 | 0.445859872611465   | 0.7146017699115044 |
| 4  | Lex-pdv         | 0.6430062630480167 | 0.7120622568093385  | 0.47983014861995754 | 0.7345132743362832 |
| 5  | Lex-pos         | 0.5984690327070286 | 0.5972762645914397  | 0.5138004246284501  | 0.6880530973451328 |
| 6  | Lex App         | 0.6374391092553932 | 0.6147859922178989  | 0.4968152866242038  | 0.8097345132743363 |
| 7  | pdv-pos         | 0.579679886569241  | 0.6614785992217899  | 0.37579617834394907 | 0.6991150442477876 |
| 8  | pdv App         | 0.6102992345163535 | 0.7237354085603113  | 0.3227176220806794  | 0.7809734513274337 |
| 9  | pos App         | 0.4780793319415449 | 0.36770428015564205 | 0.37791932059447986 | 0.7079646017699115 |
| 10 | Lex-pdv-pos     | 0.6263048016701461 | 0.6673151750972762  | 0.46496815286624205 | 0.7477876106194691 |
| 11 | Lex-pdv App     | 0.6464857341684064 | 0.6556420233463035  | 0.4861995753715499  | 0.8030973451327433 |
| 12 | Lexpos App      | 0.5970772442588727 | 0.5642023346303502  | 0.4267515923566879  | 0.8119469026548672 |
| 13 | pdv-pos App     | 0.5977731384829506 | 0.6206225680933852  | 0.4033970276008493  | 0.7743362831858407 |
| 14 | Lex-pdv-pos App | 0.6214335421016005 | 0.6498054474708171  | 0.40552016985138006 | 0.8141592920353983 |

Table 4.14 shows the parameter tuning results using grid-search with the Support Vector Classifier.

Table 4.14: Grid Search with Random Forest

```

rfc_clf = Pipeline([('clf', RandomForestClassifier(n_estimators=200, max_depth=3, random_state=0)),])
rfc_clf = GridSearchCV(rfc_clf, rfc_params, cv=5, iid=False, n_jobs=-1)
model.train(rfc_clf)
model.get_model_feature_metrics(rfc_clf)

```

|    | Feature Combo   | Mean Accuracy       | Hate                | Offensive           | Neither             |
|----|-----------------|---------------------|---------------------|---------------------|---------------------|
| 0  | Lex             | 0.5254001391788448  | 0.5778210116731517  | 0.6963906581740976  | 0.28761061946902655 |
| 1  | pdv             | 0.592205984690327   | 0.7217898832684825  | 0.22717622080679406 | 0.8252212389380531  |
| 2  | pos             | 0.40988169798190677 | 0.21595330739299612 | 0.5753715498938429  | 0.4579646017699115  |
| 3  | App             | 0.44258872651356995 | 0.19844357976653695 | 0.4819532908704883  | 0.6792035398230089  |
| 4  | Lex-pdv         | 0.5247042449547669  | 0.6284046692607004  | 0.6242038216560509  | 0.3030973451327434  |
| 5  | Lex-pos         | 0.5594989561586639  | 0.4688715953307393  | 0.6518046709129511  | 0.5663716814159292  |
| 6  | Lex-App         | 0.5434933890048712  | 0.3521400778210117  | 0.613588110403397   | 0.6880530973451328  |
| 7  | pdv-pos         | 0.5581071677105081  | 0.5739299610894941  | 0.5222929936305732  | 0.577433628318584   |
| 8  | pdv-App         | 0.5518441196938065  | 0.4708171206225681  | 0.4607218683651805  | 0.7389380530973452  |
| 9  | pos-App         | 0.4523312456506611  | 0.11867704280155641 | 0.6050955414012739  | 0.672566371681416   |
| 10 | Lex-pdv-pos     | 0.5977731384829506  | 0.6050583657587548  | 0.5902335456475584  | 0.5973451327433629  |
| 11 | Lex-pdv-App     | 0.5859429366736256  | 0.5019455252918288  | 0.564755838641189   | 0.7035398230088495  |
| 12 | Lex-pos-App     | 0.5288796102992345  | 0.3151750972762646  | 0.6369426751592356  | 0.6592920353982301  |
| 13 | pdv-pos-App     | 0.546276965901183   | 0.39883268482490275 | 0.5668789808917197  | 0.6924778761061947  |
| 14 | Lex-pdv-pos-App | 0.5768963117606124  | 0.4688715953307393  | 0.5711252653927813  | 0.7057522123893806  |

The nonlinear SVM classifier, which was trained with the psychosocial PDC feature set, had the best performance in particularly recognizing hate speech. With an accuracy of 78.6%, the classifier outperformed all other classifiers. The model's hyper-parameter values were tuned with a soft margin,  $C=0.1$ , and RBF kernel  $\gamma=0.1$  to obtain this performance.

### 4.5.3 Evaluation of the Classification Models

The experimental findings of the classification models evaluation utilizing testing sets from two datasets: the initial unbalanced dataset and a balanced dataset are presented in this section. The performance of the training features and subsequent models were evaluated using the standard evaluation metrics in hate speech classification studies [10] [54] [103], namely classification precision, recall, and f-score. The unbalanced dataset, which was skewed towards the "Neither" class, was used in the initial round of studies. The goal was to determine the best successful model for detecting hate speech in brief text messages from social media by evaluating the performance of the various features, which were divided into training and testing sets. In addition, two approaches were used to solve the categorization problem: The challenge was first structured as a

multiclass classification problem, with a message being classed as hate speech, offensive, or neither. Second, the problem was investigated as a binary classification problem, in which a communication was classified as hate speech or not. These designs were influenced by machine learning theory, which states that the more refined the classifications, the lower the accuracy. As a result, all classification models were submitted to both problem designs in order to find the best results. Before being submitted to the classifier to predict the class of individual messages, the testing dataset was also mapped into tf-idf format, just like the training set. The confusion matrices in figures 4.17 and 4.18 demonstrate the outcomes of the classifier's evaluation based on the two problem designs, respectively. Second, more tests were carried out to train the same set of machine learning algorithms, but this time utilizing a balanced dataset created by under-sampling the imbalanced dataset. The confusion matrix in figure 4.19 shows the evaluation findings from the generated model.

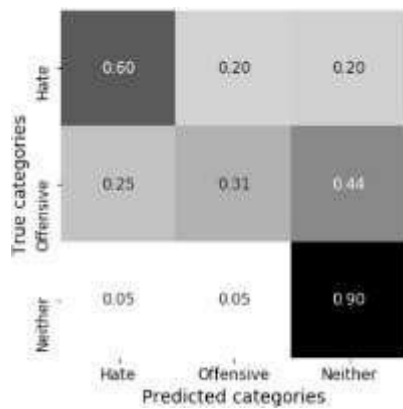


Figure 4.17: 3-Classes Confusion Matrix

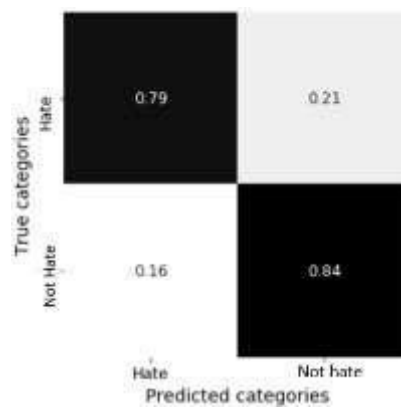


Figure 4.18: 2-Classes Confusion Matrix

#### 4.5.3.1 Experiments with Imbalanced Dataset

According to the first column of the confusion matrix in figure 4.17, the model accurately identified 60 percent of the messages as hate speech (true positive), whereas 30 percent were misclassified as false positives (25 percent offensive messages and 5% of the 'neither' messages as hate speech). The previous 25% false positives highlight the difficulty of machine classifiers when objectionable lexical phrases are present in a message, which is a common feature of many hate speech messages. This, however, reinforces the premise that not all hate speech contains offensive lexical phrases, necessitating the identification of additional criteria, such as the psychosocial factors advocated in this study, to distinguish hate speech from offensive speech. In numerous past

studies, the two classifications have been jumbled, which may not be very useful in real life, particularly for security organizations scanning vast data from social media for precise instances of hate speech. According to the model's predictions of offensive messages in column two and "neither" messages in column three, 20% of real positive messages were incorrectly classified as "Offensive" and another 20% were incorrectly classified as "neither." The annotator's keenness to the annotation scheme in light of the amount of annotations predicted throughout the annotation time could explain the erroneous classification of actual hate speech messages as offensive. This means that the annotators may have mislabeled the communications as hate speech due to their haste in annotating, whereas the messages were actually offensive or neither. For example, two human annotators identified the following message "you are just being real wape kamba wajitie kitanzi" as hate speech, while one labeled it as insulting. The target is not identifiable and cannot be linked to a protected social feature, despite the fact that the statement is most likely an incitement to suicide. As a result, the communication cannot be marked as hate speech indisputably, but rather as offensive in terms of the annotation methodology. True hate speech communications were incorrectly classified as neither because they contained positive sentiments while celebrating the in-group or membership in the in-group. Another example is when sarcasm is directed at a vulnerable population. These statements were the most difficult for the classifier to classify, despite being hate speech as described by the study's annotation system. "Yes, we are proud kikuyus, and we are the government..." is an example from the first occurrence. "Luos are the best individuals to do business with Manze unaeza tajirika haraka niggers dont know what 'Mean' is," says the second instance's sarcasm.

The most widespread misunderstanding appears to be in the 'Offensive' category, where 44 percent of the messages were wrongly labeled as 'Neither.' This could be explained by the lexicon approach's flaw, in which the mere inclusion of a negative word tipped the scales in favor of offensive or hate speech. "You should also apologize for demeaning our Luo males as uncircumcised us," says one comment. "Money does not fool all Luos" contains two foul words: uncircumcised and duped. "You have yet to feel saddened," for example. Keep in mind that even death will die one day when you destroy luos." The presence of the bigram 'kill Luos' could result in the content being flagged as hate speech. All of these communications, however, appear to be a response to and negation of a previous message. Many other messages make reference to past offensive or hate speech-related messages in the context of rejecting or debunking the previous



message in this way. These kinds of communications will be tagged as 'Neither' by a human annotator because the annotator knows that these aren't original claims, but rather disapprovals of earlier offensive comments. Second, the increased confusion may be due to annotator bias infused into the message annotation in accordance with their cultural and linguistic sensitivity. The following message, for example, was identified as 'neither' by all three human annotators despite including a swear word.

*“Fuck this Shit! Kwani @nJORoge's Job is to probe instead of detecting and eradicating money laundering?#BanksInNysSaga”*

Reading the complete communication, not just the initial sentence, which contains words under insults or generally objectionable lists, most likely influenced the decision. "Wtf!! " is another example of a tagged neither. Is that correct? Wow! #RailaTheTribalChief”. The usage of swear words as part of the annotators' everyday conversation lexicon could have been highly tolerable in this case. As a result, the prevalence of swear words is regarded as normal rather than an unusual manifestation of dislike.

The classifier fared best on the 'Neither' category, correctly classifying 90% of the messages. In this aspect, the algorithm was biased in favor of identifying the communications as neither hate speech nor offensive. In the hate speech column, 60 percent of the 90 percent hate speech messages predicted by the algorithm were genuinely hate speech, 25% were offensive, and 5% were neither but were mistakenly identified as hate speech. This misclassification, which accounts for 25% of all offensive messages being misclassified as hate speech, could be explained by the fact that the inclusion of offensive words in a message increases the likelihood of it being classified as hate speech.

When compared to approaching the problem as a multiclass classification, the findings of the binary classification of hate speech show greater performance in terms of higher accuracy in recognizing hate speech. As demonstrated in Figure 4.18's confusion matrix, 79 percent of actual hate speech was projected to be hate speech, whereas 21% was misclassified as not hate speech. In terms of how accurate the model was at identifying hate speech messages, it had an accuracy of 83 percent, whereas the percentage of actual hate speech messages accurately detected was 79

percent in terms of recall. This misclassification might be explained by the use of sarcasm or hate speech messages that did not contain any hateful or insulting terms, just as it was explained for the multiclass confusion matrix. The classifier was particularly bad at detecting subtle hate speech that glorifies in-group membership, especially when it came to out-groups. *"Luos are incredibly nice people, no surprise Kikuyu chics are flooding Nyanza,"* for example. *"Your name betrays you,"* is another example of hate speech misclassified as not hate. When opposed to 'hate,' the likelihood of identifying such messages that are free of offensive material as 'not hate' is often higher.

When compared to when the problem was a multiclass classification, there was a modest drop of 6% in the accuracy of recognizing none hate messages, which is now 84 percent. There were more false positives in the hatred column; that is, 16 percent of the messages projected as hate were incorrectly classified. These might also be explained if the message contains hateful or offensive information but was sent in the context of refuting a previous message. *"Who the hell came up with this nonsense TT >> #KillAllKikuyusToShunTribalism,"* for example. Another example of misclassification is when someone is chastising someone else. *"You're a lovely lady. You only think and speak venomous things. #HateSpeech #sillybitch."*

#### 4.5.3.2 Experiments With the Balanced Dataset

Due to the imbalanced dataset, the classifier from the first round of experiments was skewed towards the 'neither' class. Using a random under-sampling method [82], a balanced dataset was created from the original dataset. In this scenario, the number of records in the minority class serves as the pivot value for balancing the other classes such that they all weigh roughly the same amount. As a result, the f1-score in the hatred and offensive classes improved dramatically. The PDC and TF-IDF feature pairings provided the best performance in terms of features. Table 4.15 illustrates this.

Table 4.15: Balanced dataset performance

| Class        | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 0.62      | 0.72   | 0.66     | 336     |
| 1            | 0.56      | 0.50   | 0.53     | 320     |
| 2            | 0.63      | 0.59   | 0.61     | 317     |
| accuracy     |           |        | 0.60     | 973     |
| macro avg    | 0.60      | 0.60   | 0.60     | 973     |
| Weighted avg | 0.60      | 0.59   | 0.59     | 973     |

Figure 4.19 shows the confusion matrix, which is an array that shows the actual and predicted values for all of the cases in the testing dataset.

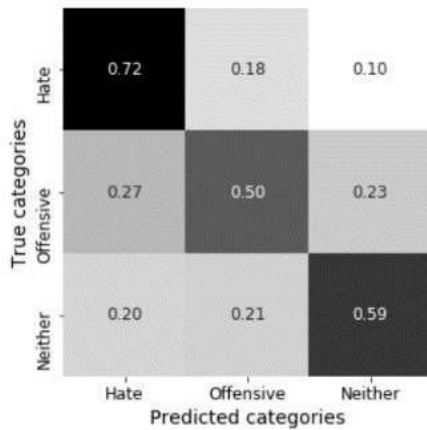


Figure 4.19: Confusion matrix based on the balanced dataset

Figure 4.19 shows that 28% of hate speech is misclassified, with the classifier achieving a precision of 0.62 and a recall of 0.72 for the hate class. The confusion matrix is read row-by-row in terms of actual values and column-by-column in terms of anticipated values. For example, the model correctly predicted 72 percent of actual hate speech messages (true positives) in column 1 (Hate class), whereas 27 percent of offensive and 20 percent of "neither" messages were false positives, meaning they were misclassified as hate speech communications. The program properly identified 50 percent of offensive messages in column 2 (Offensive class), but 18 percent of actual hate speech and 21% of true "neither" messages were misclassified as offensive. The model properly identified 59 percent of the messages in column 3 (Neither class), but 10% of actual hate speech and 23% of offensive messages were misclassified as neither containing hatred nor offensive.

The most common source of misunderstanding for hate speech was in the offensive column, where 18% of occurrences were incorrectly classified as offensive. The greatest substantial ambiguity in the offensive class was in the hate column, where 27 percent of the texts were mistakenly predicted as hate speech. With 20% and 21% of the messages mistakenly predicted as hate and offensive, respectively, the confusion was nearly balanced for neither class. Our algorithm performed the worst at guessing what was objectionable; half of the messages were accurately anticipated, while the other half were mistakenly forecasted. This could be explained by the human annotators' cultural biases and sensitivities, where what was considered objectionable differed based on cultural, religious, and prior experiences, as well as day-to-day linguistic exposure.

## CHAPTER 5: DISCUSSION AND CONCLUSION

This section contains remarks that are primarily centered on the study questions, methodology, and outcomes. The chapter begins with a discussion of the study's major limitations and concludes with a prognosis based on the research findings.

### 5.1 Limitations of the Study

The design of this study, like many others in the field of hate speech classification, is not without flaws. As a result, the empirical findings in this study should be viewed in light of three major limitations that could be addressed in future research. The study's first goal was to build a model using data collected during Kenya's presidential election season when ethnic hate speech is most common. The main question is how generalizable this approach is in detecting other sorts of hate speech besides ethnic hatred, such as religious hatred of Muslims, which has been sparked in the past by terrorist activities. Although the hate speech model in this study was successful in correctly identifying samples of other forms of hate speech, future research might explore the whole range of hate speech types and subtypes, based on the holistic framework in Figure 5.21.

Second, the study concentrated on data from a single social media platform, namely Twitter. This was primarily owing to restrictions on data access and, more importantly, gaining user approval to communications sent to in-group membership on other social media networks, which were generally transmitted privately. As previously stated, unless the user specifies differently in their Twitter settings, all communications posted on Twitter social media are public by default. Taking this into account, the study was able to scrap key public tweets with confidence and ease utilizing the various approaches indicated in section 3.4.2.1, without fear of violating copyright laws. However, the question is if the findings of this study can be applied to text data from other social media platforms such as Facebook and WhatsApp. The classifier was able to correctly identify the types of those cases based on the example text messages retrieved from Facebook. However, to train a better machine learner, the number of examples is critical. Future research, with sufficient data from other social media networks, would be the best way to answer this question.

The lack of publicly available codeswitched datasets for comparison purposes was the third barrier to the generality of these findings. Nonetheless, for comparison purposes, the study employed the human inter-rater reliability score as a baseline. Future research will undoubtedly be more conclusive if the models developed on one dataset can be tested on a similar but code-switched

dataset. Furthermore, despite the lack of resources to recruit professional annotators and a larger annotation team, future studies should look into different approaches to obtain a high-quality and larger annotated dataset for use in improving the classification model's training.

## 5.2 Discussion

The study's main goal was to establish whether a new psychosocial feature set could be used to identify subtle forms of hate speech, particularly in codeswitched text messages exchanged on social media, and test this by training a machine classification model that could generalize to other types of hate speech. To ensure that the core purpose is addressed completely and completely, five objectives were derived from it. These included the creation of a dataset representative of hate speech on social media in Kenya, the development of a machine classification model based on the psychosocial features, and the evaluation of the hate speech classification model. Each goal generated at least one research topic, the method of which and the outcomes of which are detailed in the sections below.

### 5.2.1 Developing A Deep Understanding Of The Hate Speech Phenomenon

We set out to address the question, "What constitutes hate speech in a message?" in order to gain a deeper understanding of hate speech.

To answer this topic, a systematic review of current hate speech definitions from the literature was conducted. The approach began with a review of hate speech definitions in dictionaries and legal terminology available in national policy documents such as constitutions and public acts. Following that, hate speech definitions from international organizations such as the United Nations were examined. Finally, hate speech was gathered and compared as stated in user-content policy documents on websites of major social media networks. In general, the content analysis methodology was used to identify developing themes or commonalities within the various definitions. Hate speech is an expression that often includes a negative attitude, emotion, or sentiments, according to the study's findings. Anger, rage, revenge, fear, or hostility aimed at a person or group are examples of these emotions. Second, hate speech has a target that it wants to distance itself from, whether that target is a single person or a group of people who share a protected trait such as ethnicity, race, or religion. Finally, hate speech has a goal or aim, which is frequently to threaten, offend, degrade, or devalue the target.

Furthermore, the NCIC definition of hate speech was scrutinized in order to verify that the phenomena was contextualized to the Kenyan situation. Furthermore, the researcher's attendance at NCIC's hate speech training for human monitors prior to the 2017 elections proved the Kenyan government's commitment in dealing with the phenomena, particularly in light of its potential to spread during the general election campaign season. This also validated the period as the most appropriate trigger event for gathering larger volumes of hate speech data from social media in the country.

Hate speech did not have a universal definition. Furthermore, most hate speech categories were found to be derived from legal perspectives as embodied in particular country policy papers. For example, the major American-owned social media networks shared many commonalities and adhered to the First Amendment of the United States constitution. As a result, the study's working definition of hate speech encapsulated the NCIC definition: "Hate speech" is defined as "any communication that expresses distancing language (prejudice, discrimination, or hatred) directed at an individual or a group based on their membership in a protected social category" (including ethnicity, religion, race, gender, disability, nationality, and sexual orientation).

### **5.2.2 Developing a Hate Speech Conceptual Framework**

The focus of this research objective was to develop a conceptual framework that encompasses key psychological characteristics in order to detect subtle types of hate speech in text messages. The research question was: How useful are psychosocial characteristics in identifying hate speech in text messages?

Various hate speech theories were explored in order to qualitatively identify essential psychological themes in hateful language exchanges. Social identity theory, self-categorization theory, speech act theory, communication theory, critical race theory, Baumeister's theory, integrated threat theory, sociological theory of homophile, and triangular theory of hatred were among them. The notions derived from these hate theories, as indicated in Table 2.8, were utilised to develop a strong theoretical foundation and, as a result, a conceptual framework for hate speech. This methodology was successful in building a psychosocial feature set that was then utilized to train several machine learning algorithms and, as a result, learn the most effective classifier model for nuanced types of hate speech in coded text messages from social media.

In general, the procedure began with each theory's key dimensions of hatred being extracted. These variables were qualitatively investigated, and three key notions, psychosocial distancing, negative passion, and commitment to hate, were identified. These ideas subsequently constituted the cornerstones of the original hate speech conceptual framework. The seed features under each hate speech idea in the framework were then identified through brainstorming. Furthermore, the ideas and specific traits that were painstakingly introduced were assessed for their informativeness in capturing hate speech in sample text messages by qualitatively assessing them. The iterative procedure yielded a psychosocial feature set, the significance of which was empirically assessed by a series of machine learning experiments, the findings of which are described in Table 5.1.

### 5.2.3 Building a Hate Speech Dataset

The aim of this research question was to create a dataset that captured hate speech on social media in Kenya. The fundamental topic was how to extract hate speech-containing text messages from online social media in order to create a high-quality dataset representative of hate speech in Kenya.

Hate speech on the internet has been observed to surge immediately after a major event, such as a terrorist attack or presidential elections [152]. As a result, the months leading up to Kenya's 2017 presidential election became a perfect data collection period for hate speech. First, throughout previous presidential elections, the country has a history of perpetuating bad ethnicities. Furthermore, due to the repeat elections in October of that year, the August 2017 elections featured a unique and extended data collection time. Second, existing research shows that hate speech is often more prevalent during trigger events, such as presidential campaigns, and then declines afterward [16]. During the 2017 general election, a sizable dataset of roughly 400k SMS was gathered.

Hate speech could be efficiently crawled online utilizing a combination of problematic hashtags, pro-hate user accounts, inflammatory terms, and phrase patterns, according to the study.

The use of problematic hashtags, offensive terms, pro-hate user accounts, and phrase patterns in scraping the 400k tweets from the January to December 2017 electioneering period revealed the effectiveness of using problematic hashtags, offensive terms, pro-hate user accounts, and phrase patterns in scraping the hate content. Tweets were chosen because they are frequently topically structured, publicly available, and programmatically accessible via Twitter APIs, python tweet

collecting utilities, and even custom-built crawlers, unlike text messages from other social media networks. First, unlike most other social media, it was possible to collect and compile a large dataset of text from publicly available tweets. By this, we mean that we didn't need a Twitter account to access public tweets, but we did need to register an account in order to access data on the other social media networks, which was quite restrictive. Second, utilizing a tweet crawler and an application created with Twitter's API, data from Twitter could be accessed programmatically. Finally, hashtags allowed for the grouping of all linked tweets on a specific topic. The hashtag #killallkikuyus, for example, drew a lot of angry reactions. Furthermore, the platform's design allowed for broad involvement from all demographics, including those who would normally be excluded from traditional platforms. Furthermore, Twitter data has been used in multiple prior comparable research in automatic hate speech identification [12], [10], [23], [54], [103].

Cleaning tweets was part of the data pretreatment process detailed in section 3.2. All other portions of a tweet, such as the dates, URL, and user account name, were removed, leaving only the tweet ID and the message section wrapped in double quotes. The rationale was that the other elements frequently do not contribute much to the information needed to classify a tweet [153]. Despite the fact that the tweet ID adds no useful information, it was kept so that the dataset may be released publicly as tweet IDs, in accordance with Twitter's data sharing policy[147].

A team of 27 human annotators chose 60k messages at random from the 400k communications for annotation. Each tweet was annotated by a group of three people, with the class chosen by the majority vote. Approximately 50k tweets were annotated, with 18% consisting only of hate speech messages, while the bulk were classified as neither hate nor offensive. Hate speech, which constitutes the minority class, was expected from such a large social media dataset and is consistent with prior similar study [9]. One of the conclusions was that ethnic hate speech is the most common form of hate speech in Kenya during election campaigns. As a result, ethnic hate speech vocabulary might be used as a domain in the development of a classifier model for the Kenyan environment. Second, unlike binary classification systems, the addition of the "offensive" class made it easier to discern between hate and offensive communications, lowering the risk of mislabeling tweets as hate speech, which is a typical error in annotation activities [54].

The initial annotation's inter-coder reliability score was poor, indicating that the annotation team was only half of the time in agreement with each other. This is in line with other similar studies,



one of which had an inter-rater score of 0.17 [154]. The low dependability score was attributed, as in prior studies, to the employment of less expensive but inexperienced annotators in comparison to SME annotators, as well as their personal sensitivities and social prejudices [32]. When annotators incorrectly categorize a communication as hate speech when it is not, or vice versa, noise signals are created, often known as instructor noise[80]. This was notably noticeable with some of the staff annotators who did not complete the full annotation course due to "work constraints." Colleague participation in research might be difficult, especially if they are primarily driven by monetary incentives associated with research activities. Despite the training, teacher noise, along with the tacit information and biases they bring to annotation, may constitute part of the latent qualities that are modeled as random components in the noise signal. Another reason could be because the original gateway design assigned each tweet to any three random annotators, whereas Krippendorff's Alpha assumed that each annotation was completed by the entire team of annotators, in this example, the group of 27.

Because the K-alpha computation implies that all of the annotations were done by the same team of annotators, with no room for missing data, this could lower the value of Alpha. In the case of twenty-seven annotators rating ten messages, the total number of annotations is estimated to be 270, or  $27 \times 10$ . However, if the annotation is constructed in such a way that the sole requirement is for a team of any three random annotators to rate each message, rather than a team of twenty-seven annotators, the expected outcome would be 30 ratings, or  $3 \times 10$ . This would account for about 89 percent of missing data, lowering the value of alpha substantially. This, however, does not necessarily imply a lack of inter-rater dependability. Requiring all annotators to rate all messages increases the annotators' workload and reduces the amount of messages that may be rated without necessarily enhancing the outcomes. On this basis, instead of having all of the annotators assess each message in the dataset, the design of allowing any three annotators annotate each message was adopted. Furthermore, having a large number of annotated messages was more desirable as a criterion for efficiently training a supervised machine learning model.

A tougher annotator recruitment criterion, an extended training session, and a monitored annotation could be considered in the future. Furthermore, the annotation activity could take an iterative approach, with the selected messages being replayed at random in different cycles to see if the human annotators are adhering to the defined annotation strategy. This will be critical for

finding and removing outliers in order to improve inter-coder reliability and, as a result, the performance of machine classifiers trained on the labeled dataset. Nonetheless, rather than imposing an annotation scheme based solely on existing literature or methods, and the researcher's assumptions, it might be worth establishing, considering, and accommodating the beliefs, values, and theories already held by the human annotators about the phenomenon under study at the outset. This may result in a more realistic inter-coder reliability result. Furthermore, the annotation tool [21] was created to counter the drawbacks of Krippendorff's [141] inter-coder reliability technique, which results in greater costs and slower annotations when used to a large dataset, such as the over 25k tweets in this case. Future research will look into a better dependability metric that overcomes the alpha's flaws, such as its failure to account for chance agreement [155]. Even for human annotators, the annotation task reveals how difficult classification is.

In general, the amount of the annotated dataset in this work, which consisted of about 48k usable tweets, was not only appropriate, but also exceeded the size of prior hate speech datasets, which comprised 13k [32], 16k [137], and 21k [54] tweets, respectively.

#### 5.2.4 Training a Model for Hate Speech Classification

The focus of this research question was to develop a machine classification model to recognize nuanced kinds of hate speech in social media codeswitched text communications. The main question was: How informative is the psychosocial feature set in training machine learning models to accurately categorize hate speech in comparison to conventional features?

In training several machine learning models, the unique psychosocial feature set was compared to traditional features, and their performance was used to determine the best classification model for subtle types of hate speech in codeswitched messages. Lexical features (LEX), Part of Speech linguistic features (POS), and Application-specific features were among the standard features (APP). To determine the best features and the highest-performing machine classification model, these features were examined alone and in combination using a feature combo. The broad lexicon from the input corpus makes up the lexical characteristics.

The performance of the Nave Bayes classifier improves when PDC is paired with the Lex feature, as seen by the feature significance in Table 4.6. This is due to the strong likelihood that the Lex

feature sample contains certain informative characteristics for hate speech that are not present in the PDC feature set. Fundamentally, the architecture of the PDC feature set merely reveals the primary psychological categories of hate speech that must be filled in with specific features. The PDC feature set began with only a few features under the relevant psychosocial categories, as inspired by the LIWC psychological word list [70]. The categories elements were added over time when additional specific features in texts previously reported as hate speech were uncovered, as well as translation equivalents to account for codeswitched occurrences. Furthermore, when compared to the dense and informative PDC characteristics, the Lex features were rather sparse. This can be explained by the vectorizer's use of a random feature sampling approach to extract Lex features from the input dataset, which includes a parameter for setting the amount of features. The more features there are in a text, the more complicated the computation becomes, especially when it comes to the quantity of memory and compute time necessary to process the sparse input vector. Unlike the typical broad and "diluted" Lex feature set, the PDC feature set consists of fewer but selected high informative features, i.e. "concentrated features," to identify hate speech. The addition of PDC to the usual Lex feature set always resulted in greater performance, as seen in each of the classifiers in section 4.5.1. The addition of Lex or other features, on the other hand, resulted in a decrease in performance. This is due to the noise provided by these characteristics, as well as the sparseness of the new input vector used to train the classifiers.

Previous hate speech detection research has relied on lexical and other NLP-based features. These types of capabilities, on their own, will not be able to capture hate speech in codeswitched messages effectively. As a result, classifier models that explicitly use these traditional traits would underperform, resulting in a high number of false negatives, contrary to how hate is actually expressed in social media postings.

Psychosocial features (PDC) from the study, as well as other high-level features like 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 was compared to that of traditional text classification algorithms like Nave Bayes, Linear Logistic Regression, Random Forest, and Support Vector Machine. The models were created using 5-fold cross-validation and a dataset with an 80/20 ratio of training and testing characteristics. Across the array of machine learning techniques, a grid search algorithm was utilized to discover the feature or feature combination that

yielded the highest accuracy performance. Given that the primary goal was to identify hate speech, the attention moved to the models' performance in the hate speech category. As a result, only the accuracy performance for the hate speech class was removed from the overall experimental findings, which also included the offensive and neither classes' performance. The experimental results, which are shown in Figure 5.1 and documented in Table 5.1, were based on a balanced dataset.

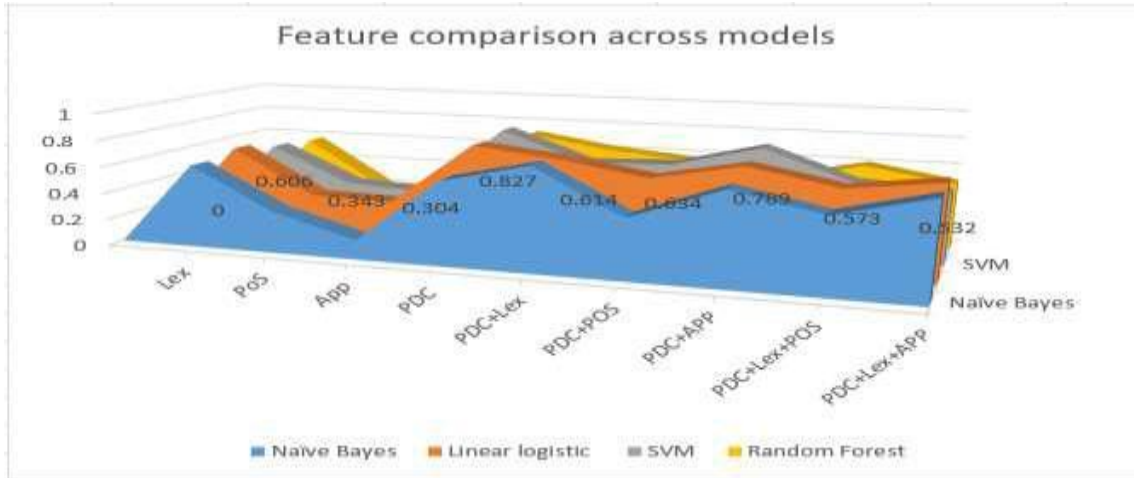


Figure 20: Comparison of Feature Across Models

The feature name appears in the first column of Table 5.1, followed by the model names and the appropriate feature performance in terms of accuracies.

Table 5.1: Comparison of Features across machine classifiers

| High-Level   | Classification Model |                            |              |               |
|--------------|----------------------|----------------------------|--------------|---------------|
| Feature Name | Naive Bayes          | Linear logistic Regression | SVM          | Random Forest |
| LEX          | 0.623                | 0.669                      | 0.606        | 0.578         |
| POS          | 0.316                | 0.340                      | 0.343        | 0.216         |
| APP          | 0.148                | 0.326                      | 0.304        | 0.199         |
| PDC          | 0.624                | <b>0.791</b>               | <b>0.827</b> | <b>0.722</b>  |
| PDC+LEX      | <b>0.765</b>         | 0.721                      | 0.614        | 0.628         |
| PDC+POS      | 0.448                | 0.623                      | 0.634        | 0.574         |
| PDC+APP      | 0.681                | 0.744                      | 0.789        | 0.471         |
| PDC+LEX+POS  | 0.552                | 0.627                      | 0.573        | 0.605         |
| PDC+LEX+APP  | 0.709                | 0.707                      | 0.532        | 0.502         |

The multidimensional framework in Figure 5.2 defines the co-occurrence of psychological variables from the three dimensions as a robust technique to recognizing hate speech in text messages. This strategy overcomes the limitations of lexicon-based solutions, which rely on the capacity to identify hate by looking for specific domain terms in a message, frequently without taking into account hate's syntactic patterns, especially in cases where codeswitching might be utilized to avoid the domain keywords.

Distancing words, negative passion, and a devotion to hate words are all aspects of the PDC feature set. Hate speech can be distinguished from other messages by the presence of these words in a message. However, it is unlikely to apply to hate speech that does not utilize words explicitly or in its vocabulary.

Hate speech detection is reliant on the presence of specific traits, which is a typical disadvantage of dictionary-based techniques because the model may not generalize without them.

The main drawback of the generic lexicon feature is its sparse vector representation, which results in many zeros in the vectors. When modeling, this necessitates greater computer resources, particularly memory, which is a hurdle, especially for traditional machine learning algorithms.

Based on accuracy performance, the best features and classifiers for identifying hate speech are PDC features trained with linear SVC classifiers. Another finding indicates that, when comparing character-level n-grams to word or phrase-level n-grams, PDC characteristics had the greatest significant impact on accuracy performance. This conclusion is consistent with past hate speech research [10].

The existence of terms or concepts in the communication that tried to distance the target or object of hatred influenced the psychological characteristics the most. The use of pronoun words like 'us,' 'them,' and other pronoun dichotomies like 'we,' 'they,' were salient in hate speech identification. For example,

*"The Merus are betraying **us**. We will defrock **them** from GEMA."* (1)

*"Jubilee party is another nusu mkate govt for Kikuyus & Kalenjins. **We** will punish **them**."* (2)

Negative stereotypes involving negative attitudes and generalizations directed towards specific ethnic groups also had a social distancing factor. 3 and 4 are examples of actual messages.

*"**Kaos** don't make good leaders...**they** are Cowards"* (3)

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

*"Even if all **Luos** are circumcised they will never change their hooliganism behavior!" (3)*

Offensive and passionate remarks reflecting emotions of rage, hate, fear, or hostility toward a target group were also identified as psychosocial traits. 5 and 6 are examples of actual messages.

*"Arrest everyone mpaka their grand kids Kikuyus are Mungikis Luos are Hooligans*

*Kambas are witches and Somalis are Terrorists.Twende kazi" (5)*

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

Some of the texts featured threats and incitement to violence directed at a certain social group. The use of uppercase letters, such as message 7, denoted strong feelings and emphasis. This was also the case with message 8's codeswitching. 9 and 10 are two more examples of messages.

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

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

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

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

Words that undervalued or demeaned the target were identified as psychosocial traits indicative of a commitment to hate. Words that referred to the target as immature or compared them to insects, animals, or objects were common.

11 and 12 are two examples of messages from the dataset.

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

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

Furthermore, some of these doubled as coded language intended to dehumanize the target by employing terms or phrases whose meaning was evident to in-group members but not to out-group members.

These high-level psychosocial characteristics were crucial in constructing the study's initial conceptual framework. Throughout the investigation, the framework was updated to incorporate empirical data from the different experiments that were carried out. The awareness that hate speech is multifaceted was one of the most significant results in this regard. 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., as evidenced by the multiple examples of annotated and automatically identified hate speech messages. Any communication that lacked these characteristics, especially the identification of the target based on race, was deemed offensive or ineffective. This is nicely captured in the multidimensional framework of hate speech depicted in Figure 5.2 by the Venn diagram. It comprehensively grasped the five key principles that depict the multifaceted nature of hate speech.

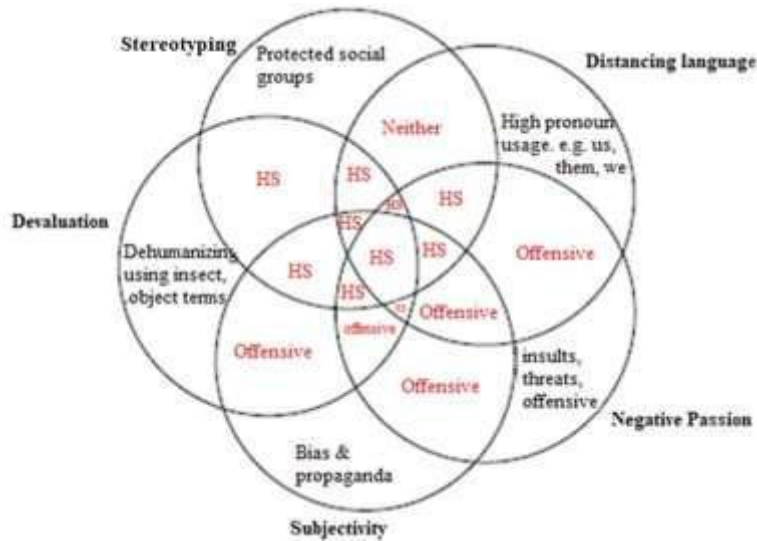


Figure 5.21: Multi-dimensional Conceptual framework for hate speech Identification.

The existence and frequency of pronouns in messages has previously been demonstrated to indicate interpersonal relationship [70]. The use of first-person pronouns such as 'we,' 'us,' and 'our,' for example, denotes closeness and a high-quality relationship between in-group members and the overall group identity. The use of the second-person pronoun 'you,' and particularly the third-person pronoun 'them,' on the other hand, is suggestive of social distance and low-quality relationships. Hate speech was found when these pronouns were employed in conjunction with the other ideas of devaluation, negative passion, or subjectivity in connection to a protected trait.

The study's main goal was to find positive cases in a codeswitched text sample by learning the class "hate speech." Hate speech, offensive, and neither were among the 50k examples of tweets that had already been classified. The comments were based on the three psychological aspects of negative Passion, Distance, and Commitment, as mentioned in the conceptual framework section

(PDC). The human annotator looked for indicators of distance (D) and passion (P) or commitment (C) in a tweet (C). Hate speech was based on D+P, D+C, or D+P+C combinations, in which psychosocial distancing was used to target an individual or a group based on a protected trait such as ethnicity. "Kenyarra is a bumbling Kikuyu president," for example. The message would be classified as a real positive if it mentioned the president's ethnicity, which is Kikuyu.

Offensive speech, like hate speech, can be based on any of the three combinations but cannot be based on a protected social feature, either directly or indirectly. "Kenyarra is a silly drunk," for example. Although the notion is offensive, it will not be considered hate speech.

Any message that falls outside of these parameters will be regarded as "neither." In theory, class learning is best when each class has its own characteristics. In essence, a feature description is shared by all instances of a class, but not by rival classes[80]. However, using the Chi-square to investigate variances in the distributions of class features within the same class and the relationship between class features, a different pattern emerged than previously anticipated. Ethnic names appear regularly across the three classes, with Kikuyu, Luo, and Kalenjin (with their Swahili language variants) being the most common. Figure 5.3 depicts this. As a result, ethnic names aren't a particularly useful feature for training a classifier to distinguish between the three classes. This goes against our original assumptions; nonetheless, if this is to be believed, the use of ethnic labels and unpleasant emotion frequently crosses the line into hate speech.

### Hate speech

```
# '0':
. Most correlated unigrams:
. fuck
. uhuru
. wajaluo
. wakikuyu
. kalenjin
. luo
. kikuyus
. Most correlated bigrams:
. ni kihii
. kikuyus think
. kill kikuyus
. njeri tamara
. luo kill
. kill luo
. kikuyus kalenjin
. Most correlated Trigrams:
. ruto join raila
. kenya kikuyus kalenjin
. presid kikuyus kalenjin
. tribal kikuyus sue
. luo tribal kikuyus
. njeri tamara luo
. tamara luo tribal
```

### Offensive

```
# '1':
. Most correlated unigrams:
. luhya
. kihii
. ni
. kikuyus
. fuck
. uhuru
. luo
. Most correlated bigrams:
. uhuru kill
. leav alon
. wew ni
. kenya si
. kikuyus think
. kill luo
. ni kihii
. Most correlated Trigrams:
. ya mama yenu
. kenya si ya
. luo luhya kisii
. polic kill luo
. aliv mourn madiba
. wew ni kihii
. matiba aliv mourn
```

### Neither

```
# '2':
. Most correlated unigrams:
. kill
. thiev
. wakikuyu
. wajaluo
. luo
. kalenjin
. kikuyus
. Most correlated bigrams:
. njeri tamara
. wako na
. kalenjin stupid
. stupid kikuyus
. kikuyus steal
. kikuyus thiev
. kikuyus kalenjin
. Most correlated Trigrams:
. tamara luo tribal
. njeri tamara luo
. tribal kikuyus sue
. luo tribal kikuyus
. ruto join raila
. join raila uhuru
. ask ruto join
```



Figure 5.22: n-grams covariance using chi-square.

After qualitatively assessing example hate messages from the dataset, it's clear that a message must have evidence of negative passion (P) or commitment (C), not only ethnic names or pronouns, to be classified as hate speech. Is there a comprehensive list of indicators relating to sets P, D, and C? Is it true that the elements in these sets evolve with time? Is a popular term in one election campaign carried over to the next, given the ambiguous nature of language use, especially in codeswitched texts? If not, how does the classifier handle new phrases from a different election campaign? These are critical problems that must be addressed, if not resolved, at the very least a fresh topic for future effort.

### 5.2.5 Generalizability of the Model

The goal of this research question was to assess the classification model. The main concern was whether our algorithm might be used to predict additional sorts of hate speech in text messages from social media.

The subject of model generalizability was central to this study, and it was utilized to underpin all of the other goals and experiments. As a result, the study's goal from the start was to gain a thorough understanding of the hate speech phenomenon and its key characteristics, informed by relevant psychological and sociological theories. As a result, a multidimensional hate speech conceptual framework was developed, which is clearly explained in section 2.8.2. The data collecting and annotation efforts were guided by this framework. Although the machine classifier in this work was trained using data collected during Kenya's 2017 presidential elections, which primarily contained ethnic hate speech, it is not limited to classifying ethnic hatred. First, as mentioned in section 4.5.3.2, the highest performing classifier in terms of generalizability was trained on a balanced dataset, based on the results of the numerous tests. When compared to a classifier trained on a dataset that is substantially skewed towards the majority class, a balanced dataset with an equal or nearly equal number of annotated class instances will not be biased towards any specific class [82]. Second, our approach is based on a universal multidimensional hate speech conceptual framework that can be applied to every type of hate speech after being retrained with positive examples of that type of hate speech. Examples of unseen sample messages that were positively classified as hate speech by the model include:

“ *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”* [Race hate];

*“Women are some of the most corrupt individual s when placed in position's of power. ”* [Gender hate]

Three major features of hate speech were present in these texts, as specified in the conceptual framework. Negative passions, such as Kill and Wtf; distancing language, such as those and hawa; and stereotyping by stating a protected attribute, such as Muslims, Wasomali, Wazungu, and Women, are examples. The combination of these characteristics in a single communication sets off the hate speech alarm.

Fundamentally, the conceptual framework aids in the identification of the hypothesis class  $H$ , i.e., Hate Speech, to which hate speech cases can be mapped. As a result, the machine learning algorithm's job is to find the specific hypothesis,  $h \in H$ , that most closely resembles hate speech. The researcher characterized the hypothesis classes, i.e., Hate speech, Offensive, and Neither, as a multi-class classification job in the supervised learning approach utilized in this work. Furthermore, comparative experiments were conducted with two hypothesis classes, namely, Hate speech or Neither (non-hate speech), with the problem being handled as a binary task. The results of the two, as shown in the confusion matrices in Figures 4.17 and 4.18, highlight the impact of fine-grained categories on classification performance, i.e., moving beyond binary classification. The major discovery was that the more data categories or labels involved, the worse the classifier's accuracy performance. This is because finding unique feature descriptions to distinguish class instances that inherit shared traits in a classification hierarchy becomes more difficult. A basic hyperplane can be used to classify a new message from social media, for example, to detect whether it contains positive or negative terms. If the answer is no, the next step is to figure out if it's hate speech or objectionable, as defined by the classification framework. This classification can be broken down further in the hierarchy, with finer-grained lower levels involving specific sorts of hate speech. In general, regardless of the number of categories involved in a multi-class classification problem, a simple yet successful strategy is to dig further into the fine categories using a hierarchical binary structure[156]. As a result, despite the fact that the hypothesis class,  $H$ , has been defined, the final parameter values are unknown. This implies that you should look for  $h \in H$  that is as close to the hate speech class as possible. The technique is repeated as many times as

there are classes accessible in the task. Learning hate speech as a hypothesis becomes easier when considering the binary classification challenge since the focus is simply on finding the most discriminant qualities that isolate instances of hate from the wider hypothesis class  $C$ . The purpose is to determine the specific  $xPDC$  that is approximate to Class Hate speech as identified in an instance  $x$ , utilizing the PDC as the bigger hypothesis class.

$$0 \quad h \quad h \quad h^{h(x)=1} \quad h \quad h$$

Given that  $C(x)$  is unknown in real life, determining how effectively  $h(x)$  translates to  $C(x)$  becomes difficult. This is conceivable, however, due to the existence of a training set  $X$  that is a subset of the set of all possible  $x$ .

The challenge of generalization refers to how to deal with the dynamic character of language and how to handle future terms that were not part of the training set  $X$ . When the test or validation data is provided via cross-validation, the question is whether the hypothesis will hold true for future unseen examples that were not part of the training set. This can be solved by creating a class  $S$  with the property  $h = S$ . This means that  $S$  must only contain examples of hate speech that are positive. Alternatively, a general hypothesis,  $G$ , can be employed, which encompasses all good examples of hate speech while excluding any incorrect examples. The algorithm can be retrained using the  $G$ -set, which includes instances of the new terms, and the margin can be increased, resulting in a greater gap between the boundary and the nearest instances [80].

### 5.2.6 Evaluating and Tuning the Model

Following the creation of the classifier, the next concerns usually revolve around two important questions. To begin, how can we tell if the classifier is working? Second, how can the classifier be improved?

To determine whether the trained classifier was performing well, i.e., using a count of true positives and true negatives, or not, i.e., using a count of false positives, i.e., Type I error, and false negatives, i.e., Type II error, standard machine learning performance metrics such as F1 score, precision, and recall as expressed in confusion matrices were used in this study.

Figure 4.17 shows that our model had the most difficulty classifying 'Offensive' messages since it was slanted toward classifying messages as 'Neither.' 69 percent of the communications identified as "offensive" by human annotators were misclassified as "Neither" and "Hate speech," or 44

percent and 25%, respectively. This could be explained by the human annotator's inherent bias and subjective nature in this activity. Furthermore, the presence of potentially offensive lexical phrases would cause the classifier to categorize it as hate speech, despite the fact that the human annotator would consider it non-offensive. This, too, is dependent on the sensitivity of the human-rater, which is often influenced by their day-to-day language use, as well as long-held cultural, religious, and other societal belief systems that are inherent, notwithstanding intensive instruction on the annotation scheme. [54], [70] are two examples. Given the high level of ambiguity in the 'Offensive' class, may binary be the most appropriate categorization for codeswitched datasets? Could it be that human annotators perceive this task as 'black and white,' rather than the 'gray' introduced by the 'Offensive' class, especially when dealing with codeswitched messages? Further experiments confounded the terms 'hate speech' and 'offensive class,' resulting in an improved performance of 83 percent accuracy. The confusion matrix in Figure 4.18 shows that the 'Not Hate Speech' category had the highest misunderstanding, with 21% of actual hate speech messages misclassified as 'Neither.' The frequency of communications misclassified as "hate speech" in the "neither" category grew to 16 percent. This was 6% higher than when there were three classes participating. Given that the primary goal was to discriminate hate speech messages in a codeswitching environment, the choice of whether to treat the task as binary or multi-class is determined by the alternative that has the highest accuracy in determining hate speech class category in comparison to the ground truth of human annotation.

Machine learning algorithms have assumptions about the structure and form of the model, as well as means of optimizing their functions to get the best approximations of the model feasible [82]. Furthermore, different machine algorithms have different processing rates and data handling capacities [87]. As a result, it was advisable to test a variety of machine learning methods to see how well they performed in detecting hate speech in codeswitched text messages. This is the triangulation method, in which different approaches are used to validate the results of the same phenomenon.

The trained model is frequently fine-tuned to improve its performance by iteratively altering various model hyperparameters in relation to their impact on a specific goal, in this instance classification accuracy.

The SVM was the best model in this study out of the nine that were used in the experiments. Based on TF-IDF character-level features as the training features, it attained the greatest classification

accuracy of 82.5 percent. With the soft margin,  $C=0.1$ ,  $\text{probability}=\text{true}$ , and a Gaussian Radial Basis Function (RBF) kernel,  $\gamma=0.1$ , as the ideal hyper-parameter values, the nonlinear SVM classifier outperformed all other classifiers. The necessity for a categorization model that might generalize effectively to various types of hate speech prompted the decision to use the  $C$  value. As a result, the model had to be trained to have more tolerance when setting the decision boundary, which is accomplished in machine learning by lowering the penalty for model misclassification [157]. The  $C$  hyper-parameter value represents this penalty. The kernel used in SVM influences how the model develops a nonlinear decision boundary depending on the characteristics it generates. When given additional characteristics, the  $\gamma$  hyper-parameter is crucial in establishing the sensitivity of the decision border. A higher  $\gamma$  value indicates that new features will have a greater influence on the decision boundary, twisting it. As a result, the lower values for the soft margin and kernel hyperparameters were the best for configuring the SVM classifier to deal with the otherwise non-linearly separable situation of text data from social media. Furthermore, SVM classifiers are quite reliable and produce amazing predictions as models.

### 5.3 Classification Model based on PDC Features

The study proposes a new text classification framework that employs a combination of psychosocial features (PDC) based on the language connoting negative passion, psychosocial distancing, and commitment to hate as primary informative concepts for detecting subtle forms of hate speech that are missed by traditional methods that rely solely on lexicon. These qualitative ideas, which are well-established in hate theories such as the duplex theory of hate, provide a rich mechanism for catching these elusive hostile sentiments, especially when hidden in codeswitching, which prior methods failed to capture.

The supervised machine learning underpins the PDC-based categorization model. It has three basic components: data pre-processing, feature processing, and model construction and evaluation. Figure 5.4 depicts these elements. Data annotation and data preprocessing are two subcomponents of the data preprocessing component. The PDC-based model's input is labeled data that could be constructed for binary or multi-class classification tasks, as is typical of supervised machine learning. By this, we imply that the data could be annotated with only two labels, such as positive or negative in binary classification, or more than two labels, such as high, medium, and low in

multi-class classification. Human annotators label the raw data input according to some annotation strategy. For example, the PDC-based multi-dimensional framework can capture the usage of devaluation in a message like "Do not make music with those insects, spray wote wote to hush them," as stated in section 2.8 and depicted in the first zoomed-out component in Figure 5.4. In Kenya, certain ethnic devaluation names are well-known and used by in-group members to refer to out-groups. For example, the term "foreskins" or "fish" is often used to refer to and disparage the Luo ethnic group, which does not perform circumcision. The usage of stereotypes translates into the use of subtle harsh language that isn't necessarily accompanied by nasty lexicons. For example, the Kikuyu, Kamba, and Kisii ethnic groups are referred to as "money lovers," "tire thieves," and "night runners," respectively. These nuanced types of hate speech, especially when codeswitching is used, can go undetected by conventional schemes' filters.

Tokenization and data cleaning of the annotated text, which is typically noisy, are part of the data pre-processing sub-unit. Dropping punctuations, duplicates, empty strings, non-alphanumeric letters, lowercasing, stemming, and removing Stopwords are among the usual data cleaning methods. Unlike traditional models, which drop all pronouns indiscriminately throughout the preprocessing step, the PDC-based model keeps the pronouns while removing the Stopwords. This is because pronoun dichotomies in a message are informative elements in suggesting "othering" language [78], which is a hate speech concept under psychosocial distancing. As an example, “*We shall not allow **them** to cross river Tana. Punda hao!*”

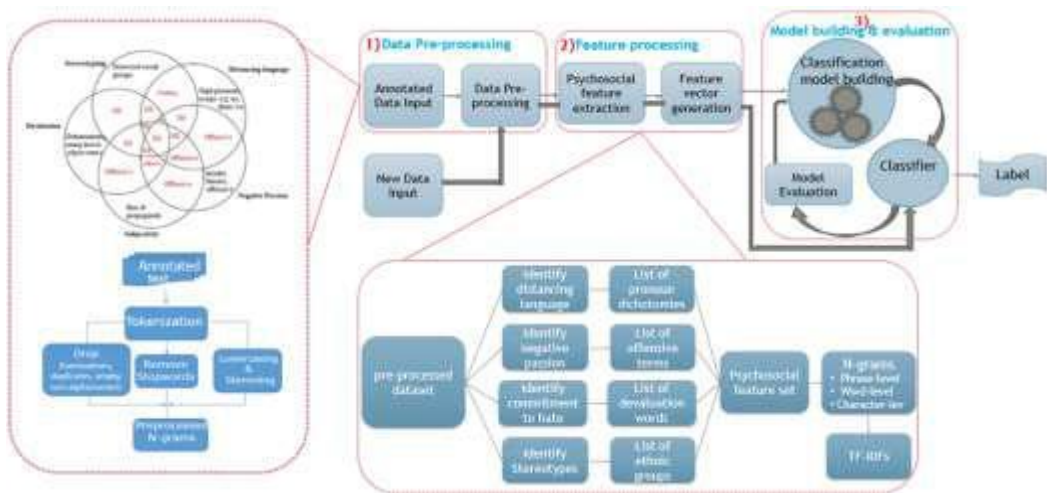


Figure 5.23: A Text Classification Model based on PDC features

The pre-processed dataset from the first component has been stripped of the regular noise signals and normalized with lowercasing and stemming. In comparison to the initial raw annotated input, this results in a considerable reduction in dimensionality. Textual data, on the other hand, poses a challenge because the words or tokens are frequently the most important attributes. As a result, this feature set becomes a high-dimensional feature space, with a sparse input vector to the machine learning algorithm in component 3 and numerous zeros, necessitating additional processing effort and memory. Component 2 of the PDC-based approach, which includes the PDC vocabulary learning subcomponent and feature vector generation subcomponent, solves this problem. To generate their respective lists, the first subcomponent filters the pre-processed dataset by extracting terminology suggestive of psychosocial detachment, negative passion, dedication to hate, and stereotyping. The seed features for each list were primarily influenced by features that had previously been found to be useful in similar problems in the literature, as summarized in Table 2.5, as well as features derived from psychological word category features using the Linguistic Inquiry Word Count analyzer[70]. In addition, as indicated in table 4.5, the respective lists were populated with co-related terms that were generated automatically by topic modeling using the Latent Dirichlet Allocation technique. These lists were grouped structurally into a table based on word families, with the first column indicating the word-family, as suggested by the conceptual framework in figure 2.26, and succeeding columns containing the word forms or features, and the rows storing the meanings. New words or words with similar meanings in other languages, i.e. codeswitched terms, could be quickly added to the feature columns via bootstrapping. As seen in table 5.2, the structural form is as follows.

Table 5.2: PDC Conceptual lookup table

| Word-Family                 | Word Form (Features) |     |     |             |     |           |
|-----------------------------|----------------------|-----|-----|-------------|-----|-----------|
|                             | Feature1             |     |     | Feature ... | ... | Feature n |
| Negative Passion            | FL1                  | FL2 | FLn |             |     |           |
| Distancing                  |                      |     |     |             |     |           |
| Commitment<br>(Devaluation) |                      |     |     |             |     |           |
| (Stereotyping)              |                      |     |     |             |     |           |
| (Subjectivity)              |                      |     |     |             |     |           |

This psychosocial feature set, PDC, might then be analyzed at various levels, such as phrase, word, or character level, and turned into numerical feature vectors, such as TF-IDFs in this case. TF-IDF

is a feature selection and representation method that ranks tokens based on their significance to the entire corpus, penalizing tokens at both extremes, i.e., words that are exceedingly common or infrequent across documents, because they are considered unimportant or outliers. In section 2.7, the formula for the TF-IDF feature is thoroughly explored. As a result, the TF-IDF feature vector based on the high-level PDC feature set is dense and a better input for classification model construction in component 3. To create their classifiers, a set of machine learning algorithms is trained on the TF-IDF input vector, informed by their performance in identifying hate speech in prior similar studies. It's often difficult to predict which machine learning algorithm will be best for a classification task ahead of time. As a result, it is typical practice in machine learning to explore a variety of algorithms, starting with the simplest, to determine which is best for a given machine learning problem [82]. Based on its results and performance, the best classifier model is reviewed and tested. The evaluation can be done in two ways. The correlation between the features and the class, i.e., the text vector and the label column value, is first computed using the Chi-Square feature scoring method. Second, using the testing dataset, the confusion matrix is utilized to determine the precision, recall, and accuracy of the trained model. Finally, the pre-processing sub-component receives a new text message as input. It isn't required to be annotated. It must, however, pass through the feature processing component and be turned into the TF-IDF vector representation in the same way. The vector is then passed straight to the classifier, with the predicted class label as the output. Figure 5.23 illustrates this.

In conclusion, tests were done to confirm our strategy of using psycho-social ideas derived from current hate theories in psychology and sociology to construct a novel psycho-social feature set, which we call PDC. The PDC feature set was then converted to tf-idf vectors to train a classification model for detecting subtle kinds of hate speech in codeswitched data. When compared to the baseline, which was the human inter-rater reliability score for the identical annotated dataset, our classifier outperformed it by over 32% in classification accuracy. The classifier's generalization was tested on an unknown dataset for racist, religious, and nationality-based abusive comments. The results were comparable to state-of-the-art baseline classifiers for similar hate speech classification. However, it would be unrealistic to compare directly with publically accessible monolingual datasets due to the usage of various datasets, especially as the focus of this study was on codeswitched data. Furthermore, this indicated the psycho-social features' robustness in generalizing to other types of hate speech, such as racial comments.



## 5.4 CONCLUSION

The widespread use of codeswitching among Kenya's multilingual community, as well as across Africa, prompted this investigation into the suitability of traditional features for collecting hate speech posted on social media. The classification of hate speech in short text messages generated on social media platforms is a difficult task. The lack of a common definition of hate speech exacerbates the situation, making it an ill-defined phenomenon. The amount of discussion around the right to free speech or freedom of expression demonstrates this. To determine if a text message contains hate speech, social media platforms now rely on users to flag such messages, which are then manually examined by human reviewers. This strategy is impractical for evaluating and categorizing the massive amounts of material created on social media, some of which border on hate speech. Furthermore, the process of human annotators annotating communications is not without prejudice and subjectivity, making it difficult to formalize. The goal of this study was to see how far data-driven approaches combining psychosocial factors and machine learning techniques may help researchers gain a better understanding of hate speech on social media. The work intended to objectively uncover the most prominent aspects of hate speech in such material, inspired by the ability of a human annotator to decipher hate speech in unstructured and loud text messages published on social media. A team of human annotators combed through a dataset of hate-related remarks on social media and annotated a sample. Following that, in a codeswitching environment, this annotated dataset was preprocessed and utilized to train a multiclass classification model to distinguish hate speech from offensive and other messages.

It is possible to create a gold-standard dataset for code-switched text. The annotation task requiring human annotators, on the other hand, is expensive. Second, despite having a consistent annotation system, the initial poor inter-rater reliability score demonstrates how much bias and subjectivity are incorporated into the annotation process. This emphasizes how emotionally charged hate speech is, as well as the difficulty it poses for human annotators who already have some inherent knowledge based on their ethnic and political prejudices. Non-Kenyans with no ethnic or political views may appear to be a stronger annotation team, but they will be limited by a lack of intrinsic expertise in deciphering the semantic content of the code-switched text messages. Is this a problem that can't be solved? It may appear so, but the study has already demonstrated critical methodological approaches that will undoubtedly be useful in augmenting human judgment in

classifying codeswitched hate speech related messages from big data generated from social media; a challenge that would otherwise be unfeasible with human annotators.

Topic modeling proved to be an effective strategy for identifying the latent semantic representations underlying social media data. Furthermore, it allowed for the automatic exploration and identification of the hate concepts defined in section 2.8.2 of the study's conceptual framework. Furthermore, the topic modeling technique contributed to a better understanding of the underlying latent factors underpinning the numerous subjects or clusters of hate words, which would have been missed by traditional methods. The researcher was able to uncover additional salient features to the PDC feature set that were utilized to train the classification models using qualitative text analysis and theme modeling.

The PDC-based tf-idf feature surpassed all other features in terms of overall performance, with the character-level features being the most superior. The performance of deep learning algorithms was lower than that of traditional models. For deep learning algorithms, this could be explained by the comparatively limited data size. Furthermore, fine-tuning the hyper-parameter values could improve the results of deep learning systems.

Existing feature representation learning methods and techniques have been concentrated on one language, with English being the most prominent. Existing pre-trained embeddings, for example, are primarily in English and other European languages. When creating classifiers to handle hate speech in African and other non-European codeswitched datasets, this becomes a hurdle. It is clear from the comprehensive tests conducted in this paper, utilizing both conventional, shallow, and deep learning algorithms, that conventional algorithms perform better on smaller datasets than deep learning algorithms. Furthermore, the presence of codeswitched text degrades the effectiveness of most traditional classifiers. Traditionally, classifiers have been trained to handle text messages, which are generally limited to a single language. As a result, the classifier is predicted to drop unintelligible phrases, just as it did most unseen words during training. In this context, a better method for dealing with codeswitched language datasets is required.

### **5.5 Recommendation for Future Research**

Future research could look at the study's applicability to other relevant topics like cyberbullying and fake news detection, both of which are spreading at an alarming rate in public online places.

In addition, based on the holistic conceptual model for hate speech produced in this work, future research could look into the model's generality in identifying hate speech in other multimedia, such as speech data. This would necessitate retraining the model with appropriate features using datasets of images, visuals, voice recordings, and music that are frequently broadcast on television, radio stations, and the internet, particularly during trigger events like the presidential elections. Furthermore, future study might be undertaken using deep learning or an entirely statistical technique to identify hate speech in streaming media such as radio stations that broadcast content in native languages that is completely language-independent and wholly automatic. These have previously been shown to be potential platforms for hate speech dissemination, particularly given that a radio presenter from a native language broadcasting radio station was one of four people charged with crimes against humanity by the International Criminal Court (ICC) during Kenya's post-election violence in 2007/2008 [158]. Fundamentally, the hate notions advocated in this study should be ubiquitous enough to recognize hate speech in any language and in any media format in any future study.

## 5.6 The Research Contributions

Theoretical, methodological, dataset, empirical, and artifact contributions were among the five critical contributions made to knowledge and general discourse in computing.

### 5.6.1 Theoretical contribution

According to Whitten's [159] analysis of what constitutes a contribution to theoretical knowledge, this study contributed significantly to the four elements to consider when answering questions about the social phenomenon of interest, namely the concepts relating to hate speech (What), the relationships between the concepts (how), the justification for the selection of the concepts relating to identifying hate speech, and the constraints imposed on the theoretical model (who, where, when). Figure 5.5 succinctly summarizes this.

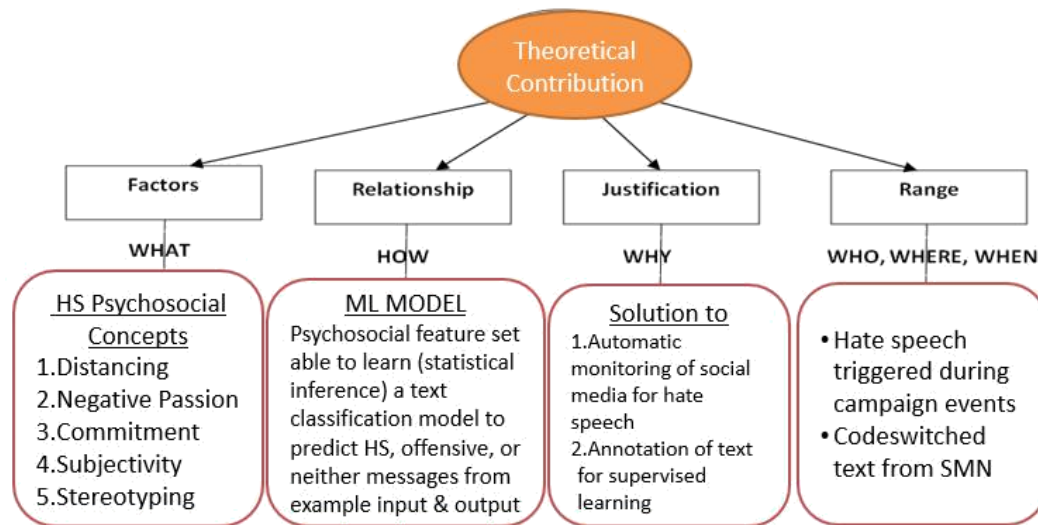


Figure 24 Theoretical Contribution to Knowledge

The key theoretical contribution was the development of a multidimensional conceptual framework for identifying hate speech. The framework was built with a strong theoretical foundation derived from psychological and sociological theories of hate. Five major psychosocial themes, including distancing, negative emotion, commitment to hatred, subjectivity, and stereotyping, were identified empirically as significant in identifying hate in short text messages obtained on social media. There was no published framework based on psychological ideas of hate that could be used to inform supervised machine learning of a classification model before this study. What was available were disconnected annotation techniques that were restricted to particular sorts of hate speech and lacked theoretical support. This generalized other forms of hate speech is more difficult. Thus, the empirical creation of a conceptual framework that captures the multidimensionality of hate speech is critical. It assists in qualitatively analyzing and processing data to uncover subtle kinds of hate speech in machine learning text classification.

### 5.6.2 A Classification Model for Hate Speech

Another significant contribution is that the hate speech classification model created in this study not only identifies ethnic hate speech but also generalizes well to other categories of hate speech such as religion, gender, and nationality. For the first time, the study employs a novel psychosocial feature set to train a classification model that makes use of statistical inference to connect features that are uniquely suited to discriminate hate speech in text messages, particularly in codeswitched text messages from social media, which previous classifiers were incapable of capturing.

### **5.6.3 Methodological Contribution**

The research contributed to the development of a methodology for an end-to-end machine learning pipeline comprising a novel preprocessing technique that reduces the dimensionality of the input feature space via automatic vocabulary learning of high-level psychosocial features. The PDC features are extremely useful in identifying hate speech across multiple classifier models. The technique describes the steps involved in collecting data, annotating it, preparing it, processing features, developing models, and evaluating it. These points are succinctly represented in Figure 5.23.

### **5.6.4 Dataset Contribution**

The Dataset contribution included a 48k-word annotated dataset of hate speech in Swahili, Luo, Kikuyu, Kisii, and other indigenous Kenyan languages. The text messages in this dataset were collected during Kenya's 2017 presidential campaign period, which included the August elections and repeat presidential elections in October 2017. As is the case elsewhere in the world, presidential campaign seasons serve as ideal trigger events for the propagation of hate speech. As a result, this new dataset will be extremely beneficial for performing comparison research with other academics undertaking similar studies on hate speech.

### **5.6.5 Artifact Contribution**

Contributions to the Artifact included the construction of a public portal (a CGI back-end application launched via an ASP.NET front-end application) for the annotation and classification of hate speech in text documents. Appendices D and E illustrate this. The annotation portal is easily customizable and may be used by other academics to provide their annotations, whereas the classifier is open and freely accessible to the public to detect hate speech in any text message.

### 5.6.6 Empirical Contributions

Numerous empirical contributions were made in light of the research aims and associated activities and experiments. These included the following strong study findings.

- i. Psychosocial features aid in the identification of hate speech. They have the potential to expose subtle types of hate speech embedded in social media data.
- ii. The compressed PDC feature set significantly compresses the original feature space and converts it to a dense feature vector suitable for machine learning.
- iii. Character-level characteristics are optimal for categorizing codeswitched text messages, especially when  $n=3$ . Additionally, existing pre-trained word embeddings are frequently based on a grammatical and monolingual corpus, making them less useful for categorizing ungrammatical and codeswitched text messages than character-level  $n$ -gram features.
- iv. Conventional machine learning techniques, such as SVM, outperform deep learning algorithms for small datasets of fewer than 100k documents. Additionally, the former is more straightforward and straightforward in terms of understanding how the feature weights equate to the feature relevance.
- v. Hate speech features can be classified into two types: high-level features that are intelligible by humans and low-level traits that are understandable by machines.
- vi. Hate speech is multifaceted. This means that a single feature is frequently less informative than when combined with multiple features. For instance, when identifying hate speech in documents, a single token will be insufficient to convey context in comparison to  $n$ -grams.
- vii. Improve the reliability of data annotation for supervised machine learning by better training amateur annotators, who might be utilized to perform the initial round of bulk annotations and then refined by subject matter experts.
- viii. The finer the classes or categories, the greater the classifier's categorization confusion. This implies that binary classification tasks are more likely to achieve higher accuracy than multi-class classification tasks.
- ix. Hate speech has no universal definition that could be used for machine learning. However, hate speech can be characterized by psychosocial distancing, and negative passion directed towards a target, be it an individual or group, to devalue or demean.

In summary, this study's findings are significant because they advance thinking and knowledge regarding the classification of subtle types of hate speech that are abundant in codeswitched data from social media. The work departs from past research in that it focuses on a novel psychosocial feature set, grounded in sound theory, for qualitatively evaluating and capturing nuanced types of hate speech that earlier methods were unable of finding. Additionally, these features were used to construct a hate speech framework composed of psychosocial concepts, which was then used to construct a feature set, dubbed the PDC feature set, to learn a classifier capable of effectively identifying hate speech in codeswitched text messages. Hate speech is a global issue that degrades user experience and has the potential to grow into actual hate crimes if left unchecked. Thus, our work contributes concisely to ongoing efforts to expand our understanding of online hate speech to better address present and future concerns through the use of technology.

## REFERENCES

- [1] BBC, “DR Congo election: Internet shut down after presidential vote,” *BBC Africa*, 31-Dec-2018.
- [2] “Ethiopia shuts Internet amid growing protests,” *The East African*, 05-Oct-2016.
- [3] B. Duggan, “Uganda shuts down social media; candidates arrested on election day,” *CNN*, 19-Feb-2016.
- [4] “Kenya to monitor social media during elections,” *The EastAfrican*, 12-Jan-2017.
- [5] J. Grygiel, “Hate speech is still easy to find on social media,” *The Conversation*, 31-Oct-2018.
- [6] BBC, “Twitter bans religious insults calling groups rats or maggots,” *BBC Technology*, 09-Jul-2019.
- [7] C. Chew and G. Eysenbach, “Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak,” *PLoS One*, vol. 5, no. 11, 2010.
- [8] E. Orehek and L. J. Human, “Self-Expression on Social Media: Do Tweets Present Accurate and Positive Portraits of Impulsivity, Self-Esteem, and Attachment Style?,” *Personal. Soc. Psychol. Bull.*, vol. 43, no. 1, pp. 60–70, 2017.
- [9] A. Schmidt and M. Wiegand, “A Survey on Hate Speech Detection using Natural Language Processing,” *SocialNLP@EACL*, 2017.
- [10] 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.
- [11] J. Waldron, *The harm in hate speech*. Harvard University Press, 2014.
- [12] 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.
- [13] Twitter, “Hateful conduct policy,” *Twitter, Inc.*, 2019. [Online]. Available: <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>. [Accessed: 12-Dec-2018].
- [14] YouTube, “Hate Speech Policy,” *YouTube Policies*, 2018. [Online]. Available: <https://support.google.com/youtube/answer/2801939?hl=en>. [Accessed: 12-Dec-2018].
- [15] J. Cheng, “Report: 80 percent of blogs contain ”offensive” content.,” 2007.
- [16] 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.
- [17] R. Ajulu, “Politicised Ethnicity, Competitive Politics and Conflict in Kenya: A Historical Perspective,” *Afr. Stud.*, vol. 61, no. 2, pp. 251–268, 2002.
- [18] P. Makori, “Whatsapp admins face jail in crackdown to curb hate-speech,” *Business Today*, 17-Jul-2017.



- [19] 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.
- [20] 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.
- [21] M. Karani, E. Ombui, and A. Gichamba, “The Design and Development of a Custom Text Annotator,” in *IEEE Africon*, 2019.
- [22] W. Warner and J. Hirschberg, “Detecting hate speech on the world wide web,” in *Language in Social Media (LSM 2012)*, 2012.
- [23] I. Kwok and Y. Wang, “Locate the hate: Detecting tweets against blacks,” *AAAI*, 2013.
- [24] 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.
- [25] E. Spertus, “Smokey: Automatic recognition of hostile Messages,” in *IAAI*, 1997.
- [26] 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.
- [27] 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.
- [28] C. Van Hee and G. De Pauw, “Automatic Detection and Prevention of Cyberbullying,” in *The First International Conference on Human and Social Analytics*, 2015.
- [29] 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.
- [30] 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.
- [31] P. Fortuna, “Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes,” University of Porto, 2017.
- [32] 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.
- [33] 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.
- [34] P. Burnap and M. L. Williams, “Us and them: identifying cyber hate on twitter across multiple protected characteristics,” *EPJ Data Sci.*, 2016.
- [35] 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.
- [36] S. Benesch, “Dangerous Speech: A Proposal to Prevent Group Violence,” 2012.
- [37] Ö. Çetinogl, S. Schulz, and N. Thang Vu, “Challenges of Computational Processing of Code-Switching,” in *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, 2016, pp. 1–11.
- [38] Z. Zhang, *Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter*, vol. 2. 2018.
- [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 International Conference on World Wide Web (WWW2015)*, 2015, pp. 29–30.
- [40] 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.
- [41] 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.
- [42] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” *Found. Trends Inf. Retr.*, vol. Vol. 2, no. No 1-2, pp. 1–135, 2008.
- [43] V. Landeiro and A. Culotta, “Robust Text Classification in the Presence of Confounding Bias,” *Assoc. Adv. Artificial Intell.*, 2016.
- [44] M. A. Campbell, “Cyber Bullying: An Old Problem in a New Guise?” *J. Psychol. Couns. Sch.*, vol. 15, no. 1, pp. 66–76, 2005.
- [45] J. P. Kaminski, L. W. Levy, and K. L. Karst, “Encyclopedia of the American Constitution,” *The Journal of American History*, vol. 74, no. 4. Macmillan Library Reference USA, USA, p. 1409, 1988.
- [46] Oxford, *The Oxford Dictionary*. Oxford University Press, 1992.
- [47] Oxford, *The Oxford English dictionary*. Oxford University Press, 2004.
- [48] *Merriam Webster dictionary*. Zane Publishers, 1995.
- [49] CERD, “Combating Racist Hate Speech,” *Int. Conv. Elimin. A ll Forms Racial Discrim.*, 2013.
- [50] COE, “Freedom of expression and information,” *Art. 10 European Convention on Human Rights*. [Online]. Available: <https://www.coe.int/en/web/freedom-expression/freedom-of-expression-and-information-explanatory-memo>. [Accessed: 24-Oct-2018].
- [51] N. C. for Law, *NATIONAL COHESION AND INTEGRATION ACT*, vol. 12. the National Council for Law, 2008.
- [52] FaceBook, “Hate Speech,” *FaceBook Inc.*, 2018. [Online]. Available: [https://www.facebook.com/communitystandards/hate\\_speech](https://www.facebook.com/communitystandards/hate_speech). [Accessed: 12-Dec-2018].

- [53] LinkedIn, “LinkedIn Professional Community Policies,” *LinkedIn Inc.*, 2018. [Online]. Available: <https://www.linkedin.com/help/linkedin/suggested/34593/linkedin-professional-community-policies?lang=en%0D>. [Accessed: 12-Dec-2018].
- [54] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated Hate Speech Detection and the Problem of Offensive Language,” in *ICWSM*, 2017.
- [55] N. Sambuli, F. Morara, and C. Mahihu, “Monitoring online dangerous speech in Kenya. Nairobi: UMATI.” 2013.
- [56] 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.
- [57] KLR, *NATIONAL COHESION AND INTEGRATION ACT NO.12 of 2008*. Kenya: National Council for Law, 2012.
- [58] D. Busolo and S. Ngigi, “Understanding Hate Speech in Kenya,” *New Media Mass Commun.*, vol. Vol.70, 2018.
- [59] A. Des Forges, “Leave None To Tell The Story: Genocide in Rwanda,” *New York Hum. Rights Watch*, 1999.
- [60] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [61] A. Reynolds, “Elections, Electoral Systems, and Conflict in Africa,” *Brown J. World Aff.*, vol. 16, no. 1, pp. 75–83, 2009.
- [62] A. Tsesis, *Destructive Messages: How Hate Speech Paves the Way for Harmful Social Movements*. NEW YORK UNIVERSITY PRESS, 2002.
- [63] R. K. Whillock and D. Slayden, *Hate Speech*. Sage Publications, 1995.
- [64] M. O. Makoloo, “Kenya: Minorities, Indigenous Peoples and Ethnic Diversity,” 2005.
- [65] UN OHCHR, “Report from OHCHR Fact-finding Mission to Kenya,” 2008.
- [66] K. Constitution, *THE CONSTITUTION OF KENYA, 2010*. Kenya: LAWS OF KENYA, 2010, p. 26.
- [67] *KENYA INFORMATION AND COMMUNICATIONS ACT*. Kenya, 2012.
- [68] M. Muendo, “Kenya targets WhatsApp administrators in its fight against hate speech,” *The Conversation*, 2017.
- [69] J. Vollhardt, M. Coutin, E. Staub, G. Weiss, and J. Deflander, “Deconstructing Hate Speech in the DRC: A Psychological Media Sensitization Campaign,” *J. Hate Stud.*, vol. 5, no. 15, 2017.
- [70] 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.
- [71] W. Collins, *Collins Dictionary of Sociology: Sociology Defined and Explained*, 3rd ed.

- HarperCollins, 2000.
- [72] S. Treppe and L. Loy, “Social Identity Theory and Self-Categorization Theory,” *The International Encyclopedia of Media Effects*. John Wiley & Sons, Inc., 2017.
- [73] W. G. Stephan and O. Ybarra, “Intergroup Threat Theory,” *Handbook of Prejudice*, 2016. [Online]. Available: [https://oscarybarra.psych.lsa.umich.edu/wp/wp-content/uploads/2016/03/1Stephan-Ybarra-\\_RiosMorrisonInPressHandbookCh.pdf](https://oscarybarra.psych.lsa.umich.edu/wp/wp-content/uploads/2016/03/1Stephan-Ybarra-_RiosMorrisonInPressHandbookCh.pdf).
- [74] 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.
- [75] UMATI, “UMATI Project Final Report,” 2013. [Online]. Available: <https://bit.ly/2rc6t0D>. [Accessed: 03-Nov-2018].
- [76] N. Coupland, “‘Other’ representation, Society and Language.” John Benjamins Publishing, 2010.
- [77] V. Dijk and A. Teun, “Discourse and racism, The Blackwell companion to racial and ethnic studies,” pp. 145–159, 2002.
- [78] 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.
- [79] M. M. Mirończuk and J. Protasiewicz, “A recent overview of the state-of-the-art elements of text classification,” in *Expert Systems With Applications*, vol. 106, Elsevier, 2018, pp. 36–54.
- [80] E. Alpaydin, *Introduction to Machine Learning*, 2nd Editio. London: The MIT Press, 2010.
- [81] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other Kernel-Based Learning Methods*. Cambridge University Press, 2000.
- [82] J. Brownlee, *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*. 2016.
- [83] A. Dey, “Machine Learning Algorithms: A Review,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 3, pp. 1174–1179, 2016.
- [84] T. Ayodele, “Types of Machine Learning Algorithms,” in *New Advances in Machine Learning*, Intech Open Access, 2010.
- [85] J. Burrell, “How the machine ‘thinks’: Understanding opacity in machine learning algorithms,” *Sage*, 2016.
- [86] T. Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” *Mach. Learn.*, 1998.
- [87] P. Flach, *MACHINE LEARNING The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 2012.

- [88] Saxena Rahul, “How the Naive Bayes Classifier works in Machine Learning,” *Data Aspirant*, 2017. [Online]. Available: <https://dataaspirant.com/2017/02/06/naive-bayes-classifier-machine-learning/>. [Accessed: 29-Jan-2020].
- [89] F. Sebastiani, “Machine Learning in Automated Text Categorization,” *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, 2002.
- [90] L. Bottou, “Large-Scale Machine Learning with Stochastic Gradient Descent,” in *Proceedings of COMPSTAT2010*, 2010.
- [91] R. Garg, “7 Types of Classification Algorithms,” 2018. [Online]. Available: <https://analyticsindiamag.com/7-types-classification-algorithms/>. [Accessed: 21-Oct-2019].
- [92] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
- [93] S. H. Lee, M. J. Maenner, and C. M. Heilig, “A comparison of machine learning algorithms for the surveillance of autism spectrum disorder,” *PLoS One*, 2019.
- [94] V. Spruyt, “The Curse of Dimensionality in classification,” 2014. [Online]. Available: <https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/>. [Accessed: 31-Oct-2019].
- [95] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer, 2013.
- [96] R. A. Fisher, “The Use of Multiple Measurements in Taxonomic Problems”,” *Ann. Eugen.*, vol. 7, no. 2, pp. 179–188, 1936.
- [97] S. J. D. Prince and J. H. Elder, “Probabilistic Linear Discriminant Analysis for Inferences About Identity,” in *2007 IEEE 11th International Conference on Computer Vision*, 2007.
- [98] K. Pearson, “On Lines and Planes of Closest Fit to Systems of Points in Space,” *Philos. Mag.*, vol. 2, no. 11, pp. 559–572, 1901.
- [99] P. Lovie and A. . Lovie, “Charles Edward Spearman, F.R.S. (1863-1945),” *Notes Rec. R. Soc.*, vol. 50, pp. 75–88, 1996.
- [100] A. Mead, “Review of the Development of Multidimensional Scaling Methods,” *J. R. Stat. Soc. Ser. D (The Stat.*, vol. 41, no. 1, pp. 27–39, 1992.
- [101] J. Tang, S. Alelyani, and H. Liu, “Feature Selection for Classification: A Review,” in *In Data Classification: Algorithms and Applications*, CRC Press, 2014, pp. 37–64.
- [102] D. Yarowsky, “Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French.,” in *32nd Annual Meeting of the Association for Computational Linguistics*, 1994, pp. 88–95.
- [103] 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.
- [104] J. Brownlee, *Deep Learning for Natural Language Processing*, V1.2. 2018.

- [105] H. J. . Palacios, R. Toledo, G. Pantoja, and A. Martínez, “A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change,” *Adv. Sci. Technol. Eng. Syst.*, vol. 2, no. 3, pp. 598–604, 2017.
- [106] 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.
- [107] A. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, “Offensive Language Detection Using Multi-level Classification,” *Springer*, p. 1627, 2010.
- [108] N. Haslam, “Dehumanization: An integrative review,” *Personal. Soc. Psychol. Rev.*, vol. 10, pp. 252–64, 2006.
- [109] P. . O’Sullivan and A. . Flanagan, “Reconceptualizing ‘flaming’ and other problematic messages,” *New Media Soc.*, vol. 5, pp. 69–94, 2003.
- [110] 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.
- [111] I. Chaudhry, “Hashtagging hate: Using Twitter to track racism online,” *First Monday* 20(2), 2015.
- [112] 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.
- [113] 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.
- [114] E. Greevy and A. . Smeaton, “Classifying racist texts using a support vector machine,” in *27th annual international conference on research and development in information retrieval*, 2004.
- [115] B. Stecanella, “What is TF-IDF?,” *Machine Learning*, 2019. [Online]. Available: [https://monkeylearn.com/blog/what-is-tf-idf/?utm\\_source=Email&utm\\_medium=Newsletter&utm\\_campaign=what-is-tf-idf](https://monkeylearn.com/blog/what-is-tf-idf/?utm_source=Email&utm_medium=Newsletter&utm_campaign=what-is-tf-idf). [Accessed: 22-Oct-2019].
- [116] “Research,” *Merriam-Webster, Inc.* [Online]. Available: <https://www.merriam-webster.com/dictionary/research>. [Accessed: 07-Nov-2019].
- [117] K. Mahoney, “Methodologiae,” *Latdict Group*, 2002. [Online]. Available: <https://latin-dictionary.net/definition/26845/methodologia-methodologiae>. [Accessed: 07-Nov-2019].
- [118] A. Bryman, “Integrating quantitative and qualitative research: how is it done?,” *Qual. Res.*, vol. 6, pp. 97–113, 2006.
- [119] B. Johnson, A. Onwuegbuzie, and L. Turner, “Toward a definition of mixed methods research. *Journal of Mixed Methods Research.*,” *J. Mix. Methods Res.*, vol. 1, pp. 112–133, 2007.
- [120] B. R. Schoonenboom, Judith Johnson, “How to Construct a Mixed Methods Research

- Design,” *Springer Open Choice*, vol. 69, no. 2, pp. 107–131, 2017.
- [121] M. Sunders, P. Lewis, and A. Thornhill, “Research Methods for Business Students,” 2007.
- [122] S. Dasgupta, *It Began with Babbage: The Genesis of Computer Science*. Oxford University Press, 2014.
- [123] B. Steitz, “A BRIEF COMPUTER HISTORY,” 2006. [Online]. Available: [http://people.bu.edu/baws/brief computer history.html](http://people.bu.edu/baws/brief%20computer%20history.html). [Accessed: 18-Oct-2019].
- [124] G. Marczyk, D. DeMatteo, and D. Festinger, *Essentials of Research Design and Methodology*. New Jersey: John Wiley & Sons, Inc., 2005.
- [125] E. Ombui, M. Karani, and L. Muchemi, “Annotation Framework for Hate Speech Identification in Tweets: Case Study of Tweets during Kenyan Elections,” in *IST-2019*, 2019.
- [126] S.-S. Shai and B.-D. Shai, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.
- [127] A. Azevedo and M. F. Santos, “KDD, SEMMA, AND CRISP-DM: A PARALLEL OVERVIEW,” *IADIS*, 2008.
- [128] P. Chapman *et al.*, “CRISP-DM 1.0 Step-by-step data mining guide,” 2000.
- [129] 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.
- [130] NCIC, “Functions of the Commission,” 2019. [Online]. Available: <https://cohesion.or.ke/index.php/about-us/functions-of-the-commission>. [Accessed: 16-Sep-2019].
- [131] Kenet, “Kenya Education Network,” 2018. [Online]. Available: <https://cert.kenet.or.ke/node/4>. [Accessed: 16-Sep-2016].
- [132] K. N. B. of Statistics, “2019 Kenya Population and Housing Census Volume I: Population by County and Sub-County,” 2019.
- [133] H. Kim, S. Jang, Mo, S.-H. Kim, and A. Wan, “Evaluating Sampling Methods for Content Analysis of Twitter Data,” *Sage*, 2018.
- [134] 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.
- [135] P. . Cavazos-Rehg *et al.*, “A content analysis of depression-related tweets,” *Comput. Hum. Behav.*, vol. 54, pp. 351–357, 2016.
- [136] A. Fink, *Conducting Research Literature Reviews: From the Internet to Paper*, Second Edi. Sage Publications, 2005.
- [137] 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.

- [138] W. Warner and J. Hirschberg, “Detecting Hate Speech on the World Wide Web,” in *Language in Social Media (LSM 2012)*, 2012.
- [139] “Three Kenyan politicians arrested over ‘hate speech,’” *The Telegraph*, 15-Jun-2010.
- [140] “Kenyan authorities arrest blogger after posts on alleged official corruption,” *CPJ*, 30-May-2018.
- [141] K. Krippendorff, “Computing Krippendorff’s Alpha-Reliability,” *University of Pennsylvania ScholarlyCommons*, 2011. [Online]. Available: [mhttp://repository.upenn.edu/asc\\_papers/43](http://repository.upenn.edu/asc_papers/43).
- [142] G. R. Semin, “Linguistic Markers of Social Distance and Proximity.” 2009.
- [143] 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.
- [144] F. Middleton, “The four types of validity,” *Scribbr*, 2019. [Online]. Available: <https://www.scribbr.com/methodology/types-of-validity/>. [Accessed: 07-Nov-2019].
- [145] 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.
- [146] W. Ahmed, P. Bath, and G. Demartini, “Using Twitter as a data source: An overview of ethical, legal and methodological challenges,” in *The Ethics of Online Research. Advances in Research Ethics and Integrity*, Second, Ed. Emerald, 2017, pp. 79–107.
- [147] “Twitter Privacy Policy,” *Twitter, Inc.*, 2018. [Online]. Available: <https://twitter.com/en/privacy>. [Accessed: 26-Oct-2019].
- [148] R. Damary, “NCIC deploys peace monitors to arrest triggers of election chaos,” *The Star*, Nairobi, 13-Apr-2017.
- [149] A. Brown, “What Is Hate Speech? Part 1: The Myth Of Hate,” *Law Philos.*, vol. 36, pp. 419–468, 2017.
- [150] M. Makinen and M. W. Kuira, “Social Media and Post-Election Crisis in Kenya,” *Inf. Commun. Technol. - Africa*, vol. 13, 2008.
- [151] N. Péladeau and E. Davoodi, “Comparison of Latent Dirichlet Modeling and Factor Analysis for Topic Extraction: A Lesson of History,” in *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.
- [152] M. L. Williams and P. Burnap, “Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data,” *Br. J. Criminol.*, vol. 56, no. 2, pp. 211–238, 2016.
- [153] S. Joshi and D. Deshpande, “Twitter Sentiment Analysis System,” *Int. J. Comput. Appl.*, vol. 180, no. 47, 2018.
- [154] P. Fortuna, L. da Silva, João 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.
- [155] M. L. McHugh, “Interrater reliability: The kappa statistic,” *Biochem. Medica*, vol. 22, no. 3, pp. 276–282, 2012.
- [156] S. Voloshynovskiy, O. Koval, F. Beekhof, and T. Holotyak, “Multiclass classification based on binary classifiers: On coding matrix design, reliability and maximum number of classes,” in *2009 IEEE International Workshop on Machine Learning for Signal Processing*, 2009.
- [157] 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].
- [158] M. Arseneault, “Kenyan radio host faces ICC trial on hate speech charges,” *RFI*, 2013. [Online]. Available: <http://www.rfi.fr/en/africa/20130909-kenyan-radio-journalist-trial-icc>.
- [159] D. A. Whetten, “What Constitutes a Theoretical Contribution?,” *Acad. Manag. Rev.*, vol. 14, no. 4, 1989.

## APPENDICES

### Appendix A: List of Publications

|   | <b>Paper/Book Title</b>  | <b>Conference/ Journal/Publisher</b>  |
|---|--|---|
| 1 | Hate Speech Classification of Codeswitched Data<br><i>Leveraging Psycho-Social Features to Classify Hate Speech: Case of Kenyan Tweets During the 2017 Elections</i> | Eliva Press<br><a href="https://www.elivapress.com/en/book/book-6275129957/">https://www.elivapress.com/en/book/book-6275129957/</a>  |
| 2 | Psychosocial Features For Identifying Hate Speech In Social Media Text   | <a href="#"><u>Journal of Education, Society and Behavioural Science</u></a><br>Manuscript Number. 2021/JESBS/77760   |
| 3 | Building and Annotating a Codeswitched Hate Speech Corpora   | MECS Press : IJITCS Vol.13, No.3, pp.33-52, Jun. 2021<br><a href="http://www.mecs-press.org/ijitcs/ijitcs-v13-n3/IJITCS-V13-N3-3.pdf">http://www.mecs-press.org/ijitcs/ijitcs-v13-n3/IJITCS-V13-N3-3.pdf</a>                                      |
| 4 | Hate Speech Detection for Codeswitched Messages  | ISMSIT 2019 Ankara, Turkey<br><a href="https://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&amp;queryText=Edward%20ombui">https://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&amp;queryText=Edward%20ombui</a>       |
| 5 | Leveraging Hierarchical Features for Hate Speech Identification in Short Text Messages   | IEEE AFRICON 2019 – Accra, Ghana<br><a href="https://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&amp;queryText=Edward%20ombui">https://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&amp;queryText=Edward%20ombui</a> |
| 6 | Annotation Framework for Hate Speech Identification in Tweets: Case Study of Tweets during Kenyan Elections  | IST Africa 2019, Nairobi, Kenya<br><a href="https://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&amp;queryText=Edward%20ombui">https://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&amp;queryText=Edward%20ombui</a>  |
| 7 | The Design and Development of a Custom Text Annotator  | IEEE AFRICON 2019 – Accra, Ghana<br><a href="https://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&amp;queryText=Edward%20ombui">https://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&amp;queryText=Edward%20ombui</a> |
| 8 | Bootstrapping Language Technology Development for Under-Resourced African language   | ACALAN's journal, KUWALA. 2016  |
| 9 | Wiring Kenyan Languages for the Global Virtual Age: An audit of the Human Language Technology Resources  | The IJSRIT, 2015<br><a href="https://www.ijsr.it.com/uploaded_all_files/2495116127_15.pdf">https://www.ijsr.it.com/uploaded_all_files/2495116127_15.pdf</a>   |

## Appendix B: Research Budget

| Expenses                             | Units | Cost per Unit (Ksh) | Total Ksh        | Total (USD)   |
|--------------------------------------|-------|---------------------|------------------|---------------|
| Ph.D. Tuition                        | 4     | 222                 | 838,600          |               |
| <b>Research Equipment</b>            |       |                     |                  |               |
| <b>Laptop Computer</b>               | 3     | 87000               | 261000           | 2691          |
| <b>Computer Server</b>               | 1     | 290000              | 290000           | 2990          |
| <b>Toner for printer</b>             | 1     | 18000               | 18000            | 186           |
| <b>Smart cell phone</b>              | 2     | 69000               | 138000           | 1423          |
| <b>Photocopy paper</b>               | 1     | 600                 | 600              | 6             |
| <b>External Hard drive</b>           | 1     | 8000                | 8000             | 82            |
| <b>Research Materials</b>            |       |                     |                  |               |
| <b>Anti-virus software</b>           | 1     | 12000               | 12000            | 124           |
| <b>Internet bundles</b>              | 12    | 3000                | 36000            | 371           |
| <b>Other</b>                         |       |                     |                  |               |
| <b>Participant honorariums</b>       | 12    | 30000               | 360000           | 3711          |
| <b>Refreshments</b>                  | 12    | 8000                | 96000            | 990           |
| <b>Administration</b>                | 12    | 5000                | 60000            | 619           |
| <b>Publication and Dissemination</b> |       |                     | 128400           |               |
| <b>Journal Application fees</b>      | 2     | 28000               | 56000            | 577           |
| <b>Binding of reports</b>            | 4     | 9000                | 36000            | 371           |
| <b>Miscellaneous</b>                 | 4     |                     | 132200           |               |
|                                      |       | <b>TOTAL</b>        | <b>2,463,800</b> | <b>25,400</b> |

## Appendix C: Annotation Scheme

### Annotation Scheme

#### Definition of Terms

Hate Speech: Any communication that expresses hatred towards an individual or group on the basis of belonging to a protected characteristic, for example, ethnicity.

Ethnic Group: means a group of persons defined by reference to color, race, religion, or ethnic or national origins, and references to a person's ethnic group refers to any ethnic group to which the person belongs.

| Feature Category | Feature             | Description and Indicators  | Examples  |
|------------------|---------------------|---|---|
| Sociological     | Distance / Othering | <p>Any message that expresses distance or places a dividing line amongst particular social groups.</p> <p>Indicated by elements of exclusion or otherness. This includes psychological distancing where you perceive others as objects or as non-existent.<br/>Indicators include: -</p> <ul style="list-style-type: none"> <li>Perceived superiority, morality and purity of the in-group as compared to the out-group</li> <li>Us vs. them (tribe, gender, religion, etc.) – in-group vs. out-group, insider vs. outsider, pure vs. impure, etc.</li> <li>Discrimination</li> </ul> | <p>Send community x home<br/><i>Tutatoa madoadoa yote</i></p> <p>Merus are letting us down, let us defrock them from GEMA</p>   |
|                  | Passion             | <p>The use of negative sentiments or the presence of emotions of anger, fear, disgust, and contempt targeted towards a specific group of people.</p> <p>Indicators include: -</p> <ul style="list-style-type: none"> <li>Abusive language</li> <li>Insults</li> <li>Defamatory statements</li> <li>Threats</li> <li>Incitements</li> </ul>  | <p>Kisiis are a DANGEROUS THREAT to our businesses; they MUST be STOPPED.</p> <p><i>Luos</i> and their cultures are generally stupid<br/><i>Kikuyus</i> are <i>mungikis</i>, <i>luos</i> are hooligans, <i>kambas</i> are witches, Somalis are terrorists</p> <p>We shall beat the uncircumcised hands down, <i>luos</i> will never rule Kenya.</p> |

|            |                                 |   |  |
|------------|---------------------------------|---|--|
|            | Devaluation                     | Perceiving others less human.<br><br>Indicators include: -<br>Intolerance of others<br>Referring to others using negative coded language, e.g., use of animal or insect terms.  | Kikuyus are enemies of <i>luos</i> , stop making music with these cockroaches. |
| Linguistic | Lexical                         | Presence of given words or phrases of given hate speech keywords.<br>Indicators include: -<br>Known hate keywords<br>Use of n-grams (unigram, bi-gram, etc.)  |  |
|            | Syntactic                       | The order of words or patterns that are associated with hate speech.<br>Indicators include: -<br>Part of speech<br>Parse structures<br>Frequency of punctuations  |  |
|            | Semantic                        | The meaning of the words. Clarity of meaning of words in the context in which they are used in the message.<br><br>Indicators include: -<br>Metaphors or idioms<br>Negative polarity  |  |
|            | Stylistic                       | The styles inherent in the sentence/message.<br>Indicators include: -<br>The word length<br>Average sentence length in words and characters<br>Emoticons<br>Short-form words, e.g., SMH, NKT<br>Use of exclamation marks<br>Capitalization<br>Flooding  |  |
|            | Subjectivity (Tone of Language) | The use of subjective sentences, heavily subjective expressions, e.g., use of negative polarity, hate verbs, etc. A conversation or messages that elicit negative emotions, opinions, or arguments.<br>Indicators include: -<br>Back / forth insults<br>Negative polarity<br>Hate and curse words |  |

## Appendix D: Annotation Portal

**HS Project**   Home   Classifier   About   Annotation ▾   Administration ▾   Contact

Message Count: 1694

Hate Speech - is a message that expresses, promotes or rationalizes any form of hatred towards a given group or individual.

This alone should make kenyans mad..all kenyans even kikuyus . This is irresponsible and a sign of

1. How would you classify the above message?

- Hate Speech
- Offensive but not hate speech
- Neither
- None

2. How do you feel about your choice in 1. above?

- Not very strongly
- Average
- Very strongly

3. What category(s) below best describes the message/tweet above?

Ethnic    Gender    Disability    Nationality    Sexual Orientation    Religion

4. Please select most applicable feature below that you identified in the message above

- Distance:**the use of "othering" language or us vs them; in-group vs out-group
- Passion:**the presence of strong emotions, offensive language, curse words , incitement
- Commitment:**Stereotyping, obvious prejudice or devaluation of a particular person or group
- Other**

## Appendix E: Hate Speech Classifier Portal

The screenshot shows a web browser window with the address bar displaying "nlp.anu.ac.ke/Classifier". The page has a dark navigation bar with links for "HS Project", "Home", "Classifier", "About", and "Contact". The main content area is titled "Tweet Classifier" and features a text input field containing the text "we shall not accept stupid politicians to si". Below the input field is a "Classify" button. The output shows the full tweet: "Tweet: we shall not accept stupid politicians to steal our votes" and a classification result: "Classification: Hate Speech!". At the bottom, there is a copyright notice "© 2019 - HS Project Team." and logos for "kenet Kenya Education Network" and "AFRICA NAZARENE UNIVERSITY".

## Appendix F: Turnitin Report

### A Model for Classifying Hate Speech Text from Social Media Leveraging on Psycho-social features and Machine Learning

ORIGINALITY REPORT

**10%**

SIMILARITY INDEX

**5%**

INTERNET SOURCES

**5%**

PUBLICATIONS

**5%**

STUDENT PAPERS