HATE SPEECH DETECTION FOR SOMALI LANGUAGE USING MACHINE LEARNING

ABDULLAHI HERSI AINAB

HASSAN DAHIR ABDULLAHI

AISHA ABSHIR NOR

FARDOWSA MOHAMED MOHAMUD

SUBMISSION OF GRADUATION PROJECT FOR PARTIAL FULFILLMENT OF THE DEGREE OF BACHELOR OF SCIENCE IN COMPUTER APPLICATIONS

JAMHURIYA UNIVERSITY OF SCIENCE AND TECHNOLOGY (JUST) FACULTY OF COMPUTER AND INFORMATION TECHNOLOGY

August 2023

Jamhuriya University of Science and Technology (JUST)

Final Submission for Graduation Project Report

Part I: Candidate's Details

Name of Candidate 1: Abdullahi Hirsi Ainab **ID No:** C119520

Name of Candidate 2: Hassan Dahir Abdullahi **ID No:** C119519

Name of Candidate 3: Aisha Abshir Nor **ID No:** C119484

Name of Candidate 4: Fardowsa Mohamed Mohamud **ID No:** C119542

Research/project title:(in block letters): Hate Speech Detection for Somali Language Using Machine Learning.

Part II - Supervisor's Comments:

(To be filled by the supervisor)

I have checked the candidate's work and hereby certify that the candidate has done all the corrections suggested by the examiners/committee.

Supervisor's Name: _____

First supervisor's signature_____ Date: _____ /_____/_____

Part III: Head of Department's Approval

(To be filled by the dean/head of department)

I have checked the candidate's work and hereby certify that the candidate has done all the corrections suggested by the examiners/committee.

Candidate's Name: _____

Candidate's signature_____ Date: _____ /_____/_____

Part IV: Final Submission of graduation project report/dissertation/ thesis

(To be filled by the candidate, please make sure that part II, III are filled before filling part IV)

To: VP academic

Dear Sir/madam

I hereby submit 3 hardbound copies of my graduation project report/dissertation/thesis and a soft PDF copy in a disk drive which has been approved by the supervisor and the head of department.

Candidate's Name: _____

Candidate's Signature _____ Date: _____ /_____/_____

Jamhuriya University of Science and Technology (JUST)

Original Literary Work Declaration

Name of Candidate 1: Abdullahi Hirsi Ainab **ID No:** C119520

Name of Candidate 2: Hassan Dahir Abdullahi **ID No:** C119519

Name of Candidate 3: Aisha Abshir Nor **ID No:** C119484

Name of Candidate 4: Fardowsa Mohamed Mohamud **ID No:** C119542

Name of Degree: Bachelor Degree of Computer Application

Title of Project/Research Report/Thesis: Hate Speech Detection for Somali Language ML

Field of Study: Machine Learning.

We do solemnly and sincerely declare that:

1. We are the sole author/writer of this Work;
2. This Work is original;
3. Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
4. We do not have any actual knowledge nor do we ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
5. We hereby assign all and every right in the copyright to this Work to Jamhuriya University of Science and technology ("JUST"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of JUST having been first had and obtained;
6. We are fully aware that if in the course of making this Work, we have infringed any copyright whether intentionally or otherwise, we may be subject to legal action or any other action as may be determined by JUST.

Candidate 1's Signature: _____ Date: _____ /_____/_____

Candidate 2's Signature: _____ Date: _____ /_____/_____

Candidate 3's Signature: _____ Date: _____ /_____/_____

Candidate 3's Signature: _____ Date: _____ /_____/_____

Subscribed and solemnly declared before,

Supervisor's Name: _____ Date: _____/_____/_____

Supervisor's Signature: _____ Designation: _____

Dedication

This work stands as a testament to the unwavering support and encouragement we received from our beloved parents. Their belief in us and constant motivation fueled our determination to complete this piece of work. We extend our heartfelt gratitude to our dedicated teachers and supportive supervisor, who not only invested their time and efforts in guiding us through this journey but also provided us with invaluable moral support, always believing in our potential. Our deep respect and appreciation go to our esteemed dean and research coordinator, whose constant encouragement and valuable insights inspired us to strive for excellence in our research. Their wise counsel and protective guidance have been instrumental in shaping our work. We also extend our thanks to the esteemed faculty members and lectures at the department, whose expertise and willingness to assist us whenever needed have been immensely valuable. Their constructive feedback and encouragement played a significant role in refining our ideas. Above all, we acknowledge the blessings and grace of the Almighty God, without whom we would not have achieved this milestone. May His abundant blessings be upon all who have contributed to this work and have been part of our journey. With heartfelt gratitude and humility, we dedicate this work to our parents, teachers, supervisor, dean, research coordinator, faculty members, and all those who have supported us along the way. Your unwavering belief in us has made this accomplishment possible, and for that, we are forever grateful. May this work be a tribute to your unwavering support, and may it contribute positively to the betterment of knowledge and understanding in our field.

Acknowledgement

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project successfully. We really grateful and wish us to Professor **Eng Abdullahi Ahmed Aden**, President and Lecturer, Department of Computer Application at **Jamhuuriya University of Science and Technology,** Deep Knowledge & keen interest of our supervisor in the field of Embedded System and Implementation to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to thank all of our friends and classmates for their timely assistance, ideas, and support until the completion of my thesis. Also, I would like to thank Mr. Sayid Abas Abshir Nor, Mr. Mohamoud Ahmed Mohamed, Mr. Zakariya Mohamed Ali, and Mrs. Najma Abdi Nasir Mohamed for their significant roles as annotators in the dataset annotation process, as well as for everyone who participated in or contributed to this work.

Table of the contents

List of Figures

FIGURE **PAGE**

List of Tables

TABLE **PAGE**

Abbreviations

SMNs	Social media networks
ILGA	International minorities associations
BOW	Bag-of-words
TF-IDF	Term Frequency - Inverse Document Frequency
ML	Machine Learning
CVS	A comma-separated values
Nltk	Natural Language Toolkit
RE	Regular expression operations
NB	Naive Bayes
SVM	Support Vector Machine
NLP	natural language processing
RF	random forest

Abstract

Social media has played a crucial role in the rapid development of internet users, connecting people from all corners of the world. Unfortunately, in recent years, hate speech on social media has become a common phenomenon in the Somalia online community. Hate speech on social media can rapidly spread among internet users and potentially lead to real-world violence and hate crimes. Determining which part of a text contains hate speech is not easy for humans. It takes time and a personal opinion on what constitutes hate speech. This study introduces a method for detecting hate speech through the implementation of machine learning. To establish the dataset, a collection of 9,172 comments from a Somalia language Facebook page was compiled. Among these comments, 3,124 were classified as hate comments, while 6,048 comments were identified as non-hateful. An experimental approach was employed to identify the most effective machine learning algorithms and feature extraction techniques for the model. The entire dataset was utilized to train various models, including Support Vector Machine, Multinomial NB, Random Forest, Logistic Regression, Decision Tree, AdaBoost, ExtraTrees, Bernoulli, GaussianNB, and k-nearest neighbor. The models were then compared based on their classification performance to determine the most suitable one for hate speech detection. the evaluation of various machine learning algorithms for hate speech detection using a train-test split of 80-20 revealed that the Multinomial Naive Bayes algorithm performed remarkably well, achieving an accuracy of 97%. This finding highlights the potential effectiveness of Multinomial NB in identifying hate speech in the Somali language Facebook comments dataset.

Keywords: Hate speech detection, machine learning, social media, Somalia online community and Support Vector

CHAPTER 1: INTRODUCTION

1.1 Background Study

Social media can be defined as computer-mediated technologies that facilitate the creation and sharing of information, ideas, career interests and other expressions through virtual communities and networks, In Somalia, the number of social media users a few years has increased dramatically, and social media communication has grown exponentially.

Social networking sites like Twitter, Facebook, YouTube, and others are being used more often by people to exchange information and express their ideas. Despite the fact that user interactions on these platforms can result in positive discussions, they are increasingly being used to spread hate speech and plan hate-based events. (Mozafari et al., 2019) Hate speech can be for women, religions, countries, cultures.(Ahammed et al., 2019)

It is quite difficult to come up with a single, universally accepted definition of hate speech. The concepts of individual, group, and minority rights, as well as dignity, liberty, and equality, are all closely tied to hate speech. The majority of national and international laws define hate speech as words or actions that cause harm to, promote hatred toward, or call for violence against a certain racial or ethnic group. (Ruwandika & Weerasinghe, 2018) social media is commonly used by people to express themselves freely and interact with others online. Collectively, social media may reveal how the general public feels about specific events. Every time a person interacts online, whether on social media, forums, or blogs, there is regrettably a risk that they may encounter objectionable language or expression.(Pitsilis et al., 2018)

Social media networks (SMNs) are the fastest method of communication since messages are sent and received almost immediately. SMNs are the primary media channels utilized nowadays to disseminate hate speech. As a result, during the past couple of decades, cyber-hate crime has rapidly expanded. To address the growth in hate speech occurrences on social media, more research is being done (SM). There have been several calls for SM providers to review each comment before making it visible to the general audience. (Ahammed et al., 2019a).

As communications are delivered and received virtually instantly, social media networks (SMNs) are the quickest form of communication. Nowadays, SMNs are the main media used to spread hate speech. Accordingly, cyber-hate crime has increased dramatically over the past couple decades. More studies are being done to combat the rise in hate speech incidents on social media (SM). SM providers have received several requests to screen each remark before making it available to the public. (Mullah & Zainon, 2021)

In this study, the detection of hate speech from posts and comments using several feature extraction methods and multiple machine learning algorithms is done, using datasets that collect from public Facebook pages as well as other social media sites, and the most suitable detection is selected.

1.2 Statement of The Problem

Hate speech refers to expression that targets or disparages an individual or a group based on a characteristics like race, religion, ethnicity or gender identity, social media use by Somalis has grown significantly in recent years, which has led to an upsurge in hate speech motivated by clan, culture, and religion. The primary cause is that Somali is not an international language. limit article regarding hate speech has ever been written before. So, there should be fewer individuals to prevent the identification of hate speech in Somalia.

Af-Somali is one of the underfunded languages that lacks language processing tools and methods in comparison to the other languages described above. Therefore, we suggest creating a hate speech detection model that uses machine learning techniques to categorize comments and posts as either hate speech or neutral speech. Due to the lack of existing datasets for the Somali language, we are using Somali dataset utilizing several machine learning methods and feature extraction approaches.

1.3 Objective of study

This study's main goal is to create a model for identifying hate speech on social media in Somalia.

1.3.1 General Objectives

This study examines the application of machine learning algorithms to predict hate speech and detect offensive language in online platforms. By utilizing various natural language processing techniques and training models on labeled datasets.

1.3.2 Specific Objectives

- To collect and annotate the Somali dataset.
- To develop a model detecting Somali hate speech.
- Evaluating the performance of the hate speech detection model.

1.4 Research Question

- How to Collect and Annotate Somali dataset?
- How to develop a model detecting hate speech?
- How to evaluate the performance of the model detecting hate speech?

1.5 Motivation of Study

Speech recognition software aims to identify hateful words and phrases and categorize speech into "hatred" and "non-hate" categories. However, it can be challenging to determine whether a phrase contains hatred or not, especially when the hate speech is concealed by irony or when there aren't any obvious expressions of racism, prejudice, or stereotyping.(Watanabe et al., 2018)

When we observed the young group using language that is unacceptable for society, culture, religion, and a person's reputation, we were inspired to discuss Somali hate speech. We are also pressured to discuss the elders who support or reject one another, which contributes to racial discrimination and social hierarchy. We decided to establish a Somali hate speech for this reason since we lost people who spoke about it and contributed to the development of the Somali language, papers, and articles about it.

Because so many individuals use SM today, occurrences of hate speech have sharply grown. Studies show that hate speech may change the narrative and has a negative effect on political discourse. Monitoring social media networks (SMNs) is essential to halt the spread of hate speech and protect democracy. Furthermore, it is obvious that emerging democracies are more prone to hate speech than more mature democracies. Therefore, developing a system to recognize hate speech can help to uphold world peace. Cyber-hatred may be committed with just a smartphone, an internet connection, and a person with a deranged mindset. If hate speech is posted, it might lead to.(Wu & Bhandary, 2020)

1.6 Significance of The Study

The purpose of this study is to systematically find the existing literature related to hate speech in social media. To achieve this aim, multiple criterions are set, such as identifying the methods used in collecting and analysing the data, normalized the data to remove superfluous

letters, symbols, emoji, and punctuation, then labelled the repeating Somali characters. Hate speech detection system will help for reduction of time and human effort to identify verbal attack on social media. The system will help to filter any hatred comment and post that makes peoples of the local population indirectly or directly participate in the violent activities across the different region of the country. The result of this research also used as input for further research investigation in the area of hate speech detection. And it plays a great role in daily life.

1.7 Scope of The Study

Hate speech is context dependent and language dependent. It is really difficult to identify hate of another culture. So, that a Somalia website is used to identify hate. This study is not focused on (maymay) in Somali language since the volume of maymay reader responses is not sufficient enough to carry on the research was selected for reader response (comments) collection since there are sufficient feedbacks for Somali articles social media.

Content Scope

This study will focus on the role hate speech in social media

Geographical Scope

The study will be conducted in Somalia

Time Scope

This study will be conducted between August 2022 to August 2023

1.8 Organization of The Study

Chapter One: Introduction to study and consists of introduction Background of the study problem statement research objectives, research questions significance of the study motivation of the study organization of the study.

Chapter two: chapter two deals with the literature reviewed so far. It includes social media definition, Hate Speech in Somalia, a concept related to automatic Hate Speech detection, different Machine learning approach, feature extraction, and a performance matrix for the method.

Chapter three: chapter three deals the methodologies and research approaches for collecting datasets so far, dataset building, collection Data , preparation datasets, Annotation Datasets, Development Tools and Techniques and Deployment Environment.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter reviews relevant literature to discuss theoretical foundations used for solving the problem of Somali text social media Hate Speech detection. The first section contains a basic definition of social media, Hate Speech detection, the nature of Hate Speech in Somalia, and current processes to monitor Hate Speech on social media is reviewed. In addition, we review Hate speech identification using machine learning techniques and a standard methodology. In general, panoptic types of literature have been examined not just to problem associated with the area of this thesis but also to bring the appropriate solution(Ababa, 2021)

2.2 Hate Speech and Tools of the social media

Online social media sites today allow users to converse freely with almost no cost. More users utilize these platforms for uses besides only communication between people. also, to spread news. While these systems' open platforms enable users to express themselves, they also have a negative side., these social media platforms in particular have developed into a testing ground for heated debates that frequently pit "us" against "them," leading to numerous instances of disrespectful and derogatory language use.(Mondal et al., 2017)

Hate speech has also been linked to the deterioration of its victims' health, Several research have shown the connection between racial and gender inequality and lower mental health., such as sadness, social isolation, anxiety, and low self-esteem.(Yuan et al., 2019). Identifying if a text has hated speech, is difficult challenge., not even for people. Because of this, it's crucial to provide definitions of hate speech before using machine learning to detect hate speech is due to this.(Biere et al., 2018). It has been found that giving hate speech a clearer, more specific definition can make the annotators' jobs easier and, as a result, raise the annotators' agreement rate. Although, it can be difficult in some nations to discriminate

between appropriate speech and hate speech. Hence, giving a precise and universal definition of hate speech become more difficult and complicated.(Mullah & Zainon, 2021).

Social media companies provide a service, they ease the communication between its users. They benefit from this service and so have obligations to the public about the sent materials.(Fortuna & Nunes, 2018) Abusive messages in social media are a complex phenomenon with a broad range of overlapping modes and goals Cyberbullying and hate speech are typical examples of abusive languages that experts have taken more interest in the past few decades due to their detrimental repercussions in our cultures. Several studies have been done to figure out how to recognize these unwanted communications in social media among others.(Mullah & Zainon, 2021) communities. Below is a list of definitions and definition sources.

- **YouTube:** We support free speech and fight for your right to voice unconventional opinions, but we don't allow hate speech. Hate speech refers to content that promotes violence against or has the primary purpose of inciting hatred against individuals or groups based on specific attributes, such as race or ethnic origin, religion, disability, gender, age, veteran status, sexual orientation/gender identity. There is a fine line between what is and what is not considered to be hate speech. In general, it is OK to criticize a nation-state, but if the content's main goal is to stir hatred against a particular group of people based on nothing more than their nationality, on their ethnicity, or if the content promotes violence based on any of these core attributes, like religion, it violates our policy(Zaghi, 2019)
- **Facebook:** 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 covers material that many people could deem offensive (such as jokes, stand-up comedy, well-known song lyrics, etc.)(Fortuna & Nunes, 2018)

- **International minorities associations (ILGA):** Hate speech is public expressions which spread, incite, promote or justify hatred, discrimination or hostility toward a specific group. ILGA They contribute to a general climate of intolerance which in turn makes attacks more probable against those given groups. (Zaghi, 2019)
- **Twitter:** Users 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 serious disease. We also do not allow accounts whose primary goal is to cause harm to others based on these criteria. The consequences for violating the Twitter rules vary depending on the severity of the violation. The sanctions span from asking someone to remove the offending Tweet before they can Tweet again to suspending an account. (Zaghi, 2019)
- **Council of Europe's Committee of Ministers:** Hate speech is defined as any form of expression that spreads, incites, promotes, or justifies racial hatred, xenophobia, antisemitism, or other forms of hatred based on intolerance, including intolerance expressed through aggressive nationalism and ethnocentrism, discrimination and hostility toward minorities, migrants, and people of immigrant origin. In this sense, hate speech covers comments which are necessarily directed against a person or a particular group of people. (Zaghi, 2019).

- **Hate speech has specific targets:** All of these definitions indicate that hate speech has a specific purpose based on the characteristics of groups such as ethnic origin, religion, race, gender, and others.
- **Hate speech contains encouragement to violence or hatred:** These definitions underline the fact that hate speech may cause conflict and violence against people as its fundamental theme.

- Hate Speech is a language or phrase that indicates incitement to violence. Hate can be based on groups or individuals who are directly attacked such as their race, religion, ethnicity, politics, appearance, color, and many other characteristics. Hate speech as a concept refers to a whole spectrum of negative discourse, stretching from expressing, inciting or promoting hatred, to abusive expression and vilification, and arguably also to extreme forms of prejudice, stereotypes, and bias.

2.3 Hate Speech on social media

Hate speech on the internet has been connected to an increase in violence against minorities around the world, including mass shootings, lynchings, and ethnic cleansing. Social media is a very popular way for people to express their opinions publicly and to interact with others online. In aggregate, social media can provide a reflection of popular mood on numerous events. Unfortunately, any user engaging online, either on social media, forums or blogs, will always have the risk(Oriola & Kotze, 2020)

social media platforms are acted by the presence of hate speech. Machine learning and natural language processing technologies have been investigated in recent years to detect risky user content on the web. This thesis deals with the problem of automated hate speech detection. Die rent machine learning and natural language processing algorithms

are combined and investigated. The experiment results are then compared with respect to their usefulness for this task (Laub, 2019).

These online platforms are frequently overutilized, and they are abused to disseminate objectionable or offensive content that incites violence and hatred against groups and individuals. Security and privacy on an anonymous and mobile-friendly social media platform enable users to conceal their true identities behind screens and status updates. Disseminate more repulsive content than other methods online spaces are often exploited and misused to spread content that can be degrading, abusive, or otherwise harmful to people. Hate speech has become a big issue for any online platform where user-generated content occurs, from news website comment sections to real-time chat sessions. Hate speech is widely defined in legal and academic literature as speech or any kind of communication that displays hatred against a person or group of people because of a trait they share, or a group to which they belong But, there is no consensus definition because of prevailing social norms, context, and individual and collective interpretation. According to a recent study, hate speech is defined as speech that either supports violent acts or fosters an environment of prejudice that may eventually result in actual violent acts against others a group of people(Mossie & Wang, 2018)

When hate speech occurs, the various definitions use slightly different phrases. The majority of the definitions point out that hate speech is to incite violence or hate towards a minority (Code of conduct, ILGA, YouTube and Twitter)(Mossie & Wang, 2018)

2.3.1 Hate Speech in Somalia

In Somalia, when it comes to hate speech, it has increased in recent years and the reason is that people are calling each other clans and sometimes in terms of religion and this is the reason for the war in the country. no rules and regulations to ban hate speech on Somali social media, according to Somali parliamentarians and the Somali constitution.

2.4 Hate Speech Detection Techniques

Several researchers have addressed the issue of hate speech detection, employing various techniques to detect the spread of hate speech on social media and other online networking platforms. There is no agreement on how to most effectively identify hate speech because there is no clear consensus on how to define hate speech. Most automated hate speech detection methods begin with binary classification campaigns, in which researchers are interested in labeling posts, tweets, or comments as "hate speech or not." To detect hateful language, it also uses a multiclass approach that flags posts, tweets, or comments as "hateful," "offensive," or "clean." Furthermore, the most popular categories with multiple categories of hate speech, racism, sexism, most harmful expressions based on religion, antisemitism, and citizen and political research seek English content. even though some other languages have developed methods for detecting hate speech.(Mossie & Wang, 2018)

2.4.1 Feature Extraction Used in Hate Speech Detection

The science of hate speech detection has just recently grown in popularity. However, a few research have previously been undertaken in a few languages. Papers concentrating on hate speech detection algorithms, as well as other research focusing on similar ideas, can provide us with insight into which features to employ in this classification assignment. Therefore, authors allocate this specific section to describe the features already employed in previous works dividing into two categories: general features used in text mining and specific hate speech detection features. (Mossie & Wang, 2018)

Dictionaries and lexicons: Most of the paper's authors attempted to adapt existing text mining algorithms to the unique challenge of hate speech identification. The paper classifies the characteristics as dictionaries and lexicons, which are frequently utilized

in text mining. This method entails creating a list of terms that are searched for and tallied in the text. In the instance of hate speech identification, content terms such as insult and swear words, response words, and personal pronouns, as well as a number of disrespectful phrases, were used in the text, with a dictionary that consists of words for English language including acronyms and abbreviations , label specific features which consisted in using frequently used forms of verbal The Ortony lexicon was also used for negative affect detection (a collection of terms expressing a negative connotation that might be beneficial because not every harsh statement necessarily involves improper language and can be equally detrimental).(Mossie & Wang, 2018)

Bag-of-words (BOW): Another model similar to dictionaries is the use of bag-of-words in this case, a corpus is created based on the words that are in the training data, instead as in dictionaries, of a predefined collection of words. The disadvantages of this kind of approaches are that the word sequence is ignored, and also, it's syntactic and semantic content. Therefore, it can lead to misclassification if the words are used in different contexts. To overcome this limitation n-grams were implemented. N-grams are one of the most used techniques in hate speech automatic detection and related tasks In a study character N-gram features proved to be more predictive than to n-gram features, for the specific problem of abusive language detection (Mossie & Wang, 2018)

N-gram features: are counts of sets of sequential N words per comment, where N is the number of words in the comment, which may range from 1 to N. In this study, we evaluated unigram (n = 1) and bigram (n = 2) word features because of their good performances in previous works in order to optimize their performance, the n-grams were weighted by term frequency-inverse document frequency (TF-IDF), which offset

the number of words in a document by the frequency of the word (term) in a corpus.(Oriola & Kotze, 2020) .

word embeddings: method that converts each word in a language's lexicon into a vector of real numbers, are yet another popular way to represent text. Using this method, word semantics may be incorporated and related words can be expressed using related vectors. Compared to n-grams, word embeddings are far more useful. or character n-gram features because they take word semantics into account and can have comparable representations for related words.(Themeli, 2018)

TDIDF: TF-IDF was also used in this kind of classification problem. It is a measure of the importance of a word in a document within a corpus and increases in proportion to the number of times that word appears in the document. However, It differs from an n-gram because the frequency of the word in the corpus is offset by the frequency of the term, compensating for the fact that some words appear more frequently overall's(Agarwal & Sureka, 2017)

2.5 Existing Hate Speech Detection approaches

Recent years have seen the development of studies on the detection of hate speech on social media. However, a number of research have previously been carried out in a number of languages. Researchers concentrating on the methodology used in this assignment from earlier research of hate speech identification on social media

2.6 Machine Learning

Machine learning is a growing field of computing algorithms that aim to mimic human intelligence by learning from their surroundings. They are regarded as the workhorse in the new era of so-called big data. Techniques based on machine learning have been applied successfully in diverse fi elds ranging from pattern recognition, computer vision,

spacecraft engineering, finance, entertainment, and computational biology to biomedical and medical applications. Ionizing radiation (radiotherapy) is used to treat more than half of all cancer patients, and it is the primary therapeutic technique in advanced stages of local disease. Radiotherapy entails a complex series of activities that not only cover the time between consultation and treatment, but also go beyond that to ensure that patients receive the best possible care. The prescribed radiation dose and are responding well. (El Naqa & Murphy, 2015)

Machine learning algorithms can learn from data and their own potential by drawing on prior knowledge. Identifying latent structures in unlabeled data that is kept in memory and translating input to output are just a few examples of jobs that the algorithms may learn.(De Smedt et al., 2018). Machine learning is the process of creating algorithms that enable a machine to learn. Learning is Finding statistical regularities or other patterns in the data is the process of learning, which does not always include consciousness(Ayodele, 2010)

Supervised machine learning: This approach is domain-specific because it is based on the manual labeling of large amounts of text. when the algorithm creates a function that translates inputs to desired outputs. The classification issue is a common formulation of the supervised learning task: the learner is supposed to learn (to approximate the behavior of) a function that maps a vector into one of several classes by looking at numerous input-output samples of the function Is fairly common in classification problems because the goal is often to get the computer to learn a classification system that we have created(El Naqa & Murphy, 2015)

Un-Supervised machine learning: Unsupervised learning analyzes unlabeled datasets without the need for human interference, i.e., a data-driven process This is widely used

for extracting generative features, identifying meaningful trends and structures, groupings in results, and exploratory purposes. The most common unsupervised learning tasks are clustering, density estimation, feature learning, dimensionality reduction, finding association rules, anomaly detection,) used a priming approach to construct a lexicon by starting with the seed of a small hateful verb and extending it repeatedly. Their model gives the best results by combining semantic hatred and themed characteristics. (Sarker, 2021)

Unsupervised learning seems much harder: the aim is for the computer to learn how to perform something that we do not instruct it how to do! Unsupervised learning can be approached in two ways. The first strategy is to train the agent by employing a reward system to signify success rather than by providing explicit categorizations. It is important to note that this form of training will typically fit into the decision problem framework since the objective is to make judgments that maximize rewards(El Naqa & Murphy, 2015)

Semi-supervised machine learning: is a hybridization of the supervised and unsupervised approaches stated above, as it operates on both labeled and unlabeled data. As a result, it lies in between learning and teaching “without supervision” and learning “with supervision”. In the real world, labeled data could be rare in several contexts, and unlabeled data are numerous, where semi-supervised learning is useful the ultimate goal of a semi-supervised learning The model's goal is to generate a better prediction result than that produced by the labeled data alone. Semi-supervised learning is utilized in a variety of applications, including machine translation, fraud detection, data labeling, and text categorization. (Sarker, 2021)

Reinforcement learning: is a form of machine learning technique that allows software agents and computers to automatically analyze the ideal behavior in a certain context or environment in order to increase its efficiency, i.e., an environment-driven approach. This sort of learning is focused on reward or punishment, and its ultimate purpose is to use environmental activists' insights to take action to raise the reward or limit the danger. It is a powerful tool for training AI models that can help increase automation or optimize the operational efficiency of sophisticated systems such as robotics, autonomous driving tasks, manufacturing, and supply chain logistics; however, it is not recommended for solving basic or straightforward problems. (Sarker, 2021)

2.7 Related work

This section offers a comprehensive examination of basic related works to the area of automatic hate speech detection on social media, this topic to revealed, clearly understand on the general techniques, methods, and results of existing analyzes.

(Omar et al., 2020) This paper presented a dataset for Arabic hate speech detection in OSNs. The dataset containing 20,000 posts, comments, and tweets collected from Facebook, Twitter, Instagram, and YouTube and manually labeled by three Arabic annotators into two balanced classes: Hate, and not hate. To the best of our knowledge, this is the first Arabic hate speech collected from more than one platform. Twelve machine learning algorithms (e.g., are MultinomialNB, Complement NB, BernoulliNB, SVC, NuSVC, LinearSVC, LogisticRegression, Decision Tree, SGD, Ridge, Perceptron, and Nearest Centroid) and two deep learning architectures (e.g. CNN, and RNN) were applied to evaluate the performance of the dataset. In machine learning algorithms, Complement NB yielded the best performance achieving an accuracy score

of 97.59%. While in deep learning architectures, RNN gave the highest performance achieving an accuracy score of 98.70.

(Romim et al., 2021)all the models achieved good accuracy. SVM achieved the overall best result with accuracy and an F-1 score of 87.5% and 0.911, respectively. But BengFastText with LSTM and Bi-LSTM had relatively the worst accuracy and F-1 score. Their low F-1 score indicates that deep learning models with BengFastText embedding were overfitted the most. BengFastText is not trained on any YouTube data but our dataset has a huge amount of YouTube comment. This might be a reason for its drop on performance. Then we looked at the performance of Word2Vec and FastText embedding. We can see that FastText performed better in terms of accuracy and had a lower F-1 score than Word2Vec. Word2Vec was more overfitted than FastText. FastText has one distinct advantage over Word2Vec: it learns from the words of a corpus and its substrings. Thus, FastText can tell ’love’ and ’beloved’ are similar words.

(Mossie & Wang, 2018)The model developed using Naïve Bayes and Random Forest utilizing a dataset of 6,120 Amharic posts and comments out of this 4,882 to train the model and 1,238 for testing after passing different steps as stated in the experiment section. The model was tested to classify whether the post and comments are hated or not and able to detect and classify in an accuracy of 79.83 % and 65.34% for Naïve Bayes with word2vec feature vector and Random Forest with TF-IDF feature modeling approach respectively. The work shows that word2vec feature model is better in maintaining the semantics of the posts and comments as proved in other works. The result is promising for such work in social network big data which can be extended to compute large volumes data since the work used the distributed platform of Apache spark. Reviewed Indonesia's hate speech on social media. The authors collect tweets and

create binary data sets with "hate" and "non-hate" voice expressions. How to characterize the negative feeling using BOW models, n-gram words, n-gram characters, and Naive Bayes, SVM, BLR, and RFDT models using different combinations of feature classifiers and machine learning algorithms. Use speech for classification. Second, we obtained the best measurement performance of 93.5% on n-gram word combinations using RFDT over other combination models.

(Alfina et al., 2017) We manually annotated the tweets into two classes, tweets containing hate speech and not. The resulting dataset had a size of 520, consists of 260 tweets for each “hate-speech” and “non-hate-speech” class. Based on the experimental results, the superior F-measure was achieved when using word n-gram, especially when combined with RFDT (93.5%), BLR (91.5%) and NB (90.2%). We found that word n-gram feature was superior to character n-gram. The results also showed that instead of using word unigram alone, it was better to union word unigram and word bigram. We also found that adding character n-gram and negative sentiment to the feature sets was not needed. We also had different results with two previous works in hate speech detection in English. While reported that character n-gram was better than word n-gram, we found the opposite. While said that RFDT, BLR, and SVM had the same performance in detecting hate speech, we found out that SVM performance was much below RFDT and BLR.

(Sigurbergsson & Derczynski, 2019) We construct a Danish dataset containing user-generated comments from Reddit and Facebook. It contains user generated comments from various social media platforms, and to our knowledge, it is the first of its kind. Our dataset is annotated to capture various types and target of offensive language. We develop four automatic classification systems, each designed to work for both the

English and the Danish language. In the detection of offensive language in English, the best performing system achieves a macro averaged F1-score of 0.74, and the best performing system for Danish achieves a macro averaged F1-score of 0.70. In the detection of whether or not an offensive post is targeted, the best performing system for English achieves a macro averaged F1-score of 0.62, while the best performing system for Danish achieves a macro averaged F1-score of 0.73. Finally, in the detection of the target type in a targeted offensive post, the best performing system for English achieves a macro averaged F1-score of 0.56, and the best performing system for Danish achieves a macro averaged F1-score of 0.63.

(Badjatiya et al., 2017) We looked at the use of deep neural network architectures to identify hate speech in Twitter. Cyberbullying, gender, religion, and LGBT are targeted by abusive communication that is categorized and filtered as racist, sexiest, or neither. Various DNN methods are used (CNN, LSTM, Fast Text) to identify which one is the best. They have experimented with different type classifiers like logistic regression, SVM, random forest classifier and deep neural networks. They used CNN [Convolutional Neural Network] for sentiment classification, LSTM [Long Short-Term Memory] for capturing long range dependencies, Fast-Text for averaging word vectors and backpropagation to fine tune the representation. They used 16k annotated tweets with respect to Glove pre trained word embedding. They have used ‘Adam’ algorithm in CNN and LSTM on the other hand ‘RMS-props’ in fast-text for optimization purpose. For baseline approach, the TF-IDF implementation is performing better than those of the character n-gram algorithm. Similarly, random embeddings work perfectly with the help of GBDT. They did many tests with many methods but LSTM with random embedding working with deep neural network models and gradient boosted decision

trees gave the best accuracy values. Prec-0.930, Recall-0.930, F1- 0.930. On the other hand, Combinations of CNN, LSTM, Fast Text embeddings as features for GBDTs was not that much better to be appreciated.

(Warner & Hirschberg, 2012) SVM [Support Vector Machine] was used to detect hate speech from websites that contained derogatory terms, phrases, and anti-racial material. It had an accuracy of 94% in categorizing hate speech, 68% in precision, 60% in recall, and F1 at a measure of.6375. The American Jewish Congress and Yahoo! were used to get the data. The data had an average length of 31, however some were longer. Dates containing one word or more than 64 words, incomplete sentences, contained two or more Unicode characters was ignored. An amount of 9000 paragraphs was matched by using a general regular expression of words related to Judaism and Israel and another 1000 paragraphs were labeled by annotators. The paragraphs then were labeled into 7 different categories- Anti Semitic, anti-feminism, Anti-Asian, Anti-black, anti-immigrants, anti-Muslim or other hate. Another set called gold was created for removing error. For classification approach they used template-based strategy and each template was centered around single words and this process produced 4379 features. Using an SVM classifier model the data was fed and by eliminating features they got 3537 features. Two additional sets containing 272 features were found using the unigram set. The set contained 13 features and the most significant is “television”. A baseline was set (N-Np)/N which yielded a baseline accuracy of 0.910. Six classifies was created to classify. The research successfully determined the hateful content of different websites (Mossie & Wang, 2020) To develop an Apache Spark-based model combining Naive Bayes and the Random Forest method to categorize social network Hate Speech detection for the Amharic language. The goal of this project was to create a system for

detecting hate speech in Amharic. The author proposes employing Random Forest and Naïve Bayes for Machine learning algorithm, Word2Vec, and TF-IDF for feature selection.

Since the Amharic language, have not public available dataset authors have built a corpus of comments retrieved from Facebook public pages of Ethiopian newspapers, individual politicians, activist, TV and radio broadcast and groups. By doing so, authors could capture both casual conversations and politically Hated posts and comments. Then preprocess the posts and comments. remaining 10% of the dataset was used for test purposes. At the last LSTM based RNN of batch size 128 and learning rate 0.001 with RMSProp optimizer and 0.5 dropout archives an accuracy of 97.9% to detect post as Hate and free by training with 100 epochs.

(Tesfaye & Kakeba, 2020) Using collecting posts and comments from specific Facebook pages of activists who took an active part in the study, researchers hope to create a massive Amharic dataset that has been categorized. The authors propose to use recurrent neural network models for automated Hate Speech posts detection from Amharic posts on Facebook is developed by using Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) with word n-grams for feature extraction and word2vec to represent each unique word by vector representation. The experiment was conducted on those two models by using 80% data for training and 10% for validation. The remaining 10% of the dataset was used for test purposes. At the last LSTM based RNN of batch size 128 and learning rate 0.001 with RMSProp optimizer and 0.5 dropout archives an accuracy of 97.9% to detect post as Hate and free by training with 100 epochs.

(Oljira, 2020) focuses on sentiment analysis of Afaan Oromo social media material due to the ease with which it can automatically recognize and categorize views from postings

on social media. By employing Multinomial Naïve Bayes Machine learning algorithm and different n-grams such as unigram, bigram, trigram and their combinations as features. The Authors propose to use the Naive Bayes algorithm for the classification of Afaan Oromo sentiment using n-gram techniques and use precision, recall, f-measure and accuracy to evaluate performance. The proposed MNB approaches achieved According to the experiment, the result shows that accuracy of 90.7%, 71.1%, 54.6%, 92.7%, 92.4%, and 75% for unigram, bigram, trigram, unigram-bigram, unigram-trigram, and bigram-trigram respectively.

Table 2.1: Comparison of Hate Speech Detection Model Accuracy

Reference	Methodology	Model accuracy
Omar et al., 2020	MultinomialNB, Complement NB, BernoulliNB, SVC, NuSVC, LinearSVC, LogisticRegression	97.59%, 98.70%
Romim et al., 2021	SVM, LSTM and Bi-LSTM	87.5%
Mossie & Wang, 2018	Naïve Bayes and Random Forest utilizing	65.34% 79.83 % & 93.5%
Alfina et al., 2017	RFDT, BLR, SVM, n-gram	(93.5%) BLR (91.5%) NB (90.2%).
Sigurbergsson& Derczynski, 2019	TF-IDF scores for a range, n-grams.	F1-score of 0.74
Badjatiya et al., 2017	LSTM, Glove, SVM	Prec-0.930, Recall-0.930, F1-0.930
Warner & Hirschberg, 2012	SVM, Random Forest	94%.68%,60%
Mossie & Wang, 2020	n-gram TF-IDF and word2vec embedding	92.5%
Tesfaye & Kakeba, 2020	Word2Vec, n-grams, LSTM, GRU	97.9%
Oljira, 2020	Naïve Bayes, n-grams	90.7%,71.1%, 54.6%,92.7%, 92.4%, and 75%

However, A different study was conducted on the problem of hate speech detection, and they used different algorithms to detect hate speech propagated on social media and other online web platforms. most of researches used hate words to labeled pots and comments as hate and hate-free speech, therefore, for this study, we are creating machine learning methods to identify hate speech on Somali social media platforms and building on new Somali dataset, the research focus on preparing hate speech annotation guideline and identify good algorithms used to build detection model for Somali language.

CHAPTER 3: METHODOLOGY

3.1 Introduction

In this chapter, we discuss the methodologies and research approaches for collecting datasets to complete the goals of this study and provide insights into the clear research topics. The following sections of this chapter describe and illustrate the method used in the study of hate speech in the Somali language.

3.2 Dataset Building

The objectives of this study are to detect Hate Speech in the Somali language. Therefore, it needs to build a new dataset nothing has been recorded or annotated for Somali hate speech detection, so this new dataset is important. In the dataset-building process for Somali hate speech detection, we described the three most popular techniques as well as the most crucial steps in establishing a dataset.

1. Somali post and comment textual data from public Facebook pages
2. Preparing, filtering, or merging gathered data into one file dataset.
3. Dataset annotation.

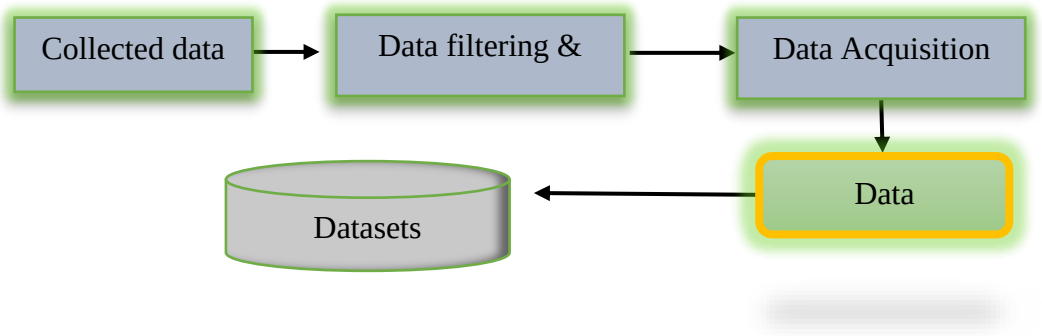

Figure 3. 1: dataset building

3.3 Collection Data

This study uses social media platforms to collects Somali textual data. Somali textual data were collected from different social media platforms in order to compile posts, comments, and a list of published Facebook sites. was compiled based on the content of these pages. We made the decision to use Facebook rather than another social media site.

In addition to the posts and comments we collected, the study also collected keywords to filter on Facebook text data is collected and used when annotating post Comments. These keywords are rude words or terms that provide negative indicators. or, hate speech text and phrases that describe target populations based on politics, gender, culture, and religion.

3.4 Preparation Dataset

Once the data is gathered a preparation process follows, which are collecting, cleaning, filtering, and merging data into one file or data table. Microsoft Excel was used as a preparation tool in this process. The following step is able us to build a dataset using the partitioned CSV data. Since the Hate Speech data records are stored in a CSV file, we cannot use them directly in our system. At a preliminary stage, the constructor will read the CSV file using Panda's library, which is data manipulation and analysis tool built on top of the Python programming language, and then passes it to NumPy arrays. NumPy is a Python library used as an array processor for numbers, strings, and records. Finally, after obtaining the array values from NumPy, the dataset constructor performs the following two tasks. That will then allow us to create a dataset using the CSV data that has been partitioned. During dataset preparation below tasks are performed.

- Filtering using keywords that are an indicator of Hate and normal.
- Removing all null values.
- Removing all non-Somalia and non-textual posts and comments.
- Removing duplication to ensure the uniqueness of each text in a dataset

All of the above data preparation guidelines point to the nature and practice of the Somali language to maintain the essence of the context of the written data collection process.

3.5 Annotation Dataset

Annotation is a method for carefully adding information to the acquired data. In this situation, the posts or comments should be marked as spam during the moderation process to collect information on the expression. To choose texts and reviews for interpretation in this study, a straightforward random sampling method is used.

3.5.1 The Purpose of Annotations Is

- To organize important material.
- Monitor your learning as you read
- Identify key concepts.
- Systematic summary of the text that you create within the document.
- A key tool for close reading that helps people uncover patterns, notice important words and identify the main point. An active learning strategy that improves compression and retention of information

3.5.2 Annotation Procedure

The annotation process for the dataset mostly carried out by the researcher, annotation is also done by two extra annotators. The number of annotators is limited because of a lack of

resources, mainly budget and time need more annotators to participate in the process. The annotators were selected based on their willingness to perform the task and Somalia language skills. Af-Somalia instructors make up two of the annotators. The dataset's annotators were given instructions to annotate a randomly selected subset of an equal number of posts and comments. The annotations were chosen based on their willingness to complete the task and Somali language skills. Annotators were asked to annotate a random subset with the same number of posts and comments from the dataset. The challenging manual annotation process of the hate speech dataset allows annotators to annotate their time independently.

3.6 Development Tools and Techniques

This study uses several development tools and packages to implement the proposed solution, Af-somali hate speech detection, this study uses python programming language for applying and experimenting. We considered every feedback, from model construction to data preprocessing. We used Python language for evaluating the proposed classifier model implemented. Python is used because researchers prefer the language Python is chosen because researchers like language Several libraries in the Python programming language allow natural language processing. we used to google Collaboration (google collab) All of this is web-based interactive computing notebook environment for edit and run python codes in each code separately. Easy to identify errors is one of the purposes of a collaborative notebook. since each cell displays its own output.

Tools	Version	Description
Pandas	1.5.3	Pandas-profiling is used to display sample, correlation, and duplicated data easily
Microsoft Excel	2019	Used data preparation tasks during data are crawled from Facebook pages and sorting the gather data. Also, used to manage the annotation task.
Anaconda Navigator	2.3.1	Allows us to launch development applications and easily manage condominium packages, environments, and channels without the need to use command-line commands.
Python	3.8.9	Powerful programming language to develop a Machine learning application. It is also easy to process natural language.
Nltk	3.8.1	Natural Language Toolkit is a suite of libraries and programs for symbolic and statistical for English written in the Python programming language. This study uses it for data reading, manipulation, writing, and handling the data frame.
NumPy	1.24.2	Python library for topic modeling document indexing and similarity retrieval with large corpora. this Study uses it to handle text-to-number conversions Functionality and training and test data model
RegEx (Re)	___	that indicates the set of strings to match This form can be used to match, remove, replace, etc. This study's text preparation package was utilized to handle the Somali text.
Google Collaboratory	3.9	is a product from Google Research. Collab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Collab is a hosted Jupiter notebook service that requires
Matplotlib	3.3.2	is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible This study uses it for data and results visualization.

Table 3. 1:packages with their Version and description

3.7 Deployment Environment

The tools described in section 3.5 above have been deployed on a personal computer with processor Intel(R) Core (TM) i5-7200U CPU @ 2.50GHz 2.71 GHz, 12.00 GB, 500 Gigabyte hard disk (GB HDD) Storage. The operating system is Windows 10 Pro 64 bit.

CHAPTER 4: SYSTEM ANALYSIS AND DESIGN

4.1 Architecture Overview

This chapter describes the proposed solution to detect hate speech in the Somali language by using Appropriate machine learning techniques to solve the problem under study. This study Defines new harmful expression data for posts and comments on Facebook pages and is used as an important part of the proposed solution for the detection of harmful expressions in social media.

4.2 Architecture of Hate Speech Detection and Classification Models

The method used to detect Somali hateful speech for posts and comments. The proposed solution is based on the architecture shown in Figure 4.2. It takes the Somali dataset and processes it according to the properties of the language. Punctuation, tokenization, and other basic pre-processing are needed. In the end, after pre-processing, feature extraction, TF-IDF, and n-gram, this activity's result is the dataset's important feature to be trained Model. After feature extraction, we use models like. RF, LR, SVM, DT, KNN, LSVC, and other machine learning to develop models Characteristics of the data set Algorithm, and training set using data frames.

This study aims to detect hate speech in Somali using machine learning. Algorithms for labeling and classifying datasets. Using Supervised Machine Learning Algorithms for comparing and verifying accuracy. The reasons for choosing an algorithm They have good ranking performance and are popular for combating hate speech detection.

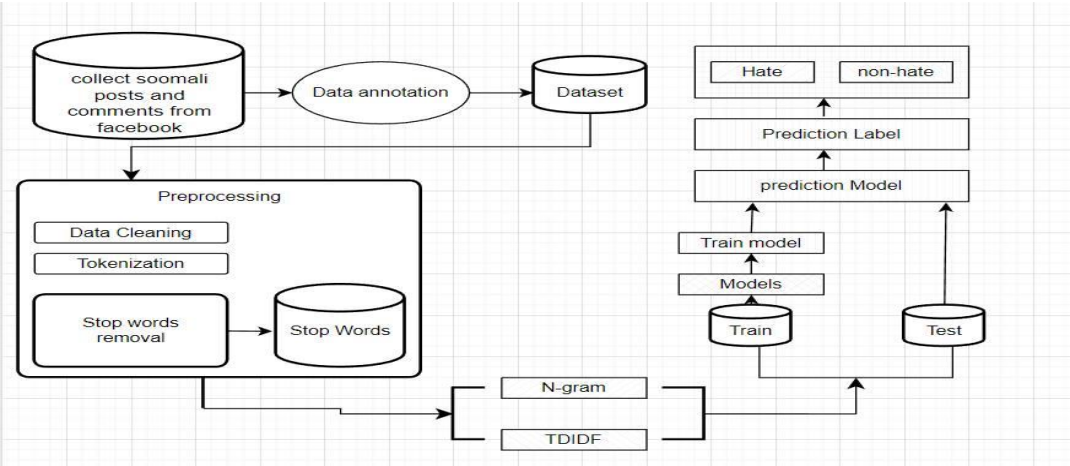

Figure 4. 1:Somali Hate Speech Detection Architecture

Random Forest (RF): is a machine learning algorithm that is used for classification, regression, and other tasks. It is an ensemble learning method that combines multiple decision trees to make more accurate predictions. A decision tree is one component of an RF model (DT). In RF, a large number of decision trees are created, each using a randomly selected subset of the training data and a randomly selected subset of the input features. The output of the model is the average prediction of all the trees in the forest, which helps to reduce the variance and improve the accuracy of the model.

Logistic Regression: is a common statistical method used for binary classification problems. It is a type of generalized linear model that uses a logistic function to model the probability of a binary response variable (i.e., a variable that takes on one of two values) based on one or more predictor variables. The logistic function is an S-shaped curve that maps any input value to a value between 0 and 1. In logistic regression, the output of the logistic function is used to represent the probability of the binary response variable taking on a particular value given the values of the predictor variables.

BernoulliNB: is a probabilistic machine learning model that is used for classification tasks. It is a variant of the Naive Bayes algorithm, which is based on Bayes' theorem and the assumption of conditional independence among the features. In BernoulliNB, the input variables are assumed to be binary, meaning that they take on a value of either 0 or 1. The model calculates the probability of a given binary output variable based on the presence or absence of each binary input variable.

GaussianNB: model is particularly useful for classification tasks where the input variables are continuous and have a normal distribution, such as in medical diagnosis or financial analysis. It is also commonly used for text classification tasks, where input variables are transformed into a continuous representation, such as word frequency or tf-idf. To train the GaussianNB model, the algorithm estimates the mean and variance of each input variable for each class, as well as the prior probability of each class. These estimates are then used to make predictions about the class of new data points.

Support Vector Machine (SVM): is a machine learning algorithm used for classification, regression, and outlier detection. It works by finding the hyperplane that best separates the data into different classes. In SVM, the hyperplane that best separates the two classes the goal is to find a hyperplane in a high-dimensional space that between the classes. Or to defined as the distance between the hyperplane and the closest data points from each class. SVM can handle both linearly separable and non-linearly separable data by using different types of kernels, such as linear, polynomial, and radial basis function (RBF) kernels. The kernel function is used to transform the input features into a higher-dimensional space, where the data may be more easily separable.

Multinomial NB: is a probabilistic machine learning model used for classification tasks. It is a variant of the Naive Bayes algorithm, which is based on Bayes' theorem and the assumption of conditional independence among the features. In Multinomial NB, the input variables are assumed to be discrete, such as in the case of word counts or frequency of occurrence of certain events. The model calculates the probability of a given output variable based on the frequency of occurrence of the input variables. To train the Multinomial NB model, the algorithm estimates the probability of each input variable (i.e., the frequency of occurrence of a particular word or token in a document)

A decision tree: A decision tree is a machine learning algorithm used for both classification and regression tasks. It works by recursively partitioning the data into subsets based on the values of the input features, and then assigning a label or value to each subset.

A decision tree consists of several nodes. The root node is the start of a decision tree, usually the machine learning dataset. Leaf nodes are the endpoints of a branch or the result of a result set. Decision trees do not branch beyond leaf nodes. In a machine learning decision tree, the data function is an internal node and the result is a leaf node. Decision trees are a method used in supervised machine learning. It is a technique for training a model using labeled input and output datasets. This method is mainly used to solve classification problems. that is, using a model to classify or categorize objects. Machine learning decision trees are also used in regression problems, which are methods used in predictive analytics to predict the output of unseen data

K-nearest Neighbor: is a machine learning algorithm used for both classification and regression tasks. It works by finding the K closest data points in the training set to a given input data point, and then making a prediction based on the labels or values of those neighbors. KNN is a data classification method that is used to estimate the probability that a data point is a member of one group or another based on which group the data point is

closest to. K-nearest Neighbor is an acronym for "k-nearest neighbor," which stands for "k-nearest neighbor to the data point." An example of supervised machine learning is the closest neighbor algorithm, which is used to solve classification and regression problems. On the other hand, its primary application is in the realm of classification. Because it does not undergo training even when it is presented with KNN training data, this type of learning algorithm is referred to as a lazy learner. Instead, it merely stores the information while it is being trained and does not carry out any calculations at all. After the dataset has been queried, only then will the model be constructed.

4.3 Somali Text Preprocessing

This subcomponent performs pre-processing training, modifies comments and posts, and tests the detection model. This pre-processing is based on the Somali language and the basic text pre-processing methods, such as punctuation and special character removal (cleanup), and tokenization

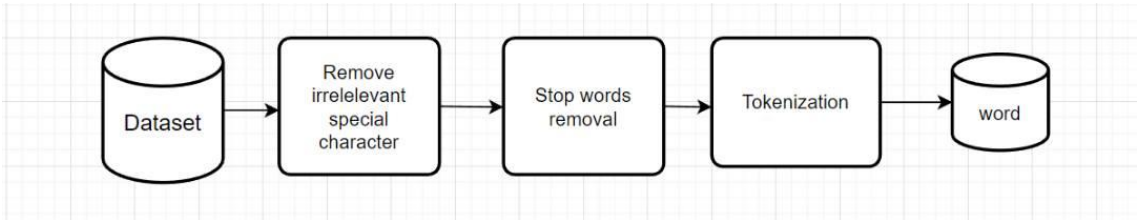

Figure 4. 2: Somali Dataset Pre-Processing

4.3.1 Removing (Cleaning) Irrelevant Character, Punctuations Symbol

Removing irrelevant characters and punctuation symbols is an important step in text pre-processing. These characters do not carry any meaningful information and can affect the accuracy of text analysis and machine learning models. In Somali language, irrelevant characters and punctuation symbols may include various diacritics, special characters, and punctuation marks. The process of removing these characters can be done using regular expressions or predefined lists of characters and symbols. Here are some common characters and symbols that can be removed during the cleaning process:

- Punctuation marks such as commas, exclamation marks, and question marks
- Special characters such as asterisks, ampersands, and dollar signs
- Diacritics and other accent marks that are specific to Somali language
- Non-letter characters such as numbers and mathematical symbols

Once these characters have been removed, the text data can be tokenized into individual words, which can then be used for further text analysis or machine learning tasks.

4.3.2 Tokenization

After cleaning, Tokenization is the process of breaking down a piece of text into individual words or tokens. This step is a fundamental part of text pre-processing in natural language processing (NLP). Tokenization is important because it allows for the analysis of text at the individual word level, which is necessary for many NLP tasks such as text classification, sentiment analysis, and machine translation. The tokenization method is as follows and splits posts and comments. text with single words or tokens using spaces between words or punctuation Marking is important because the meaning of the text often depends on the relationships between the words. Text and feature extraction methods help you extract the right features from your dataset.

4.3.3 Removal Stop word

Stop words are commonly used words in a language that are often removed during text pre-processing because they do not carry much meaning or contribute to the overall understanding of the text. since they don't make additional meaning, you can easily remove the word that does not affect the scanning process. Removing them reduces the number of features considered. Therefore, you can improve the performance of your classifier. The same Somali stop words are {"aad","ay","ayaa","ayee","ayuu","dhan","hadana","in","inuu","isku","ka","kale","kasoo", "ku","kuu","laakiin","markii","oo","si","soo","uga","ugus"uu","waa", "waxa", waxuu", etc.} These keywords explain important information. Therefore, the removed stop words are filtered through an automated process associated with the classification process.

4.4 Extraction Features Methods

Feature extraction is the process of transforming raw data into a set of features that can be used for machine learning algorithms to make predictions or classifications. In the context of natural language processing (NLP), feature extraction involves converting text data into numerical representations that can be used as inputs to machine learning models. Methods for reducing the number of redundant structures and dimensions. Approaches involving aversion, aggression, and the selection of relevant train subsets aid in the identification of data sets that are not available. in the modeling of detection problems Using research on hate speech detection and hate speech problems, N-gram and TF-IDF feature extractions were used as the most popular feature extraction methods in this study of Somali hate speech.

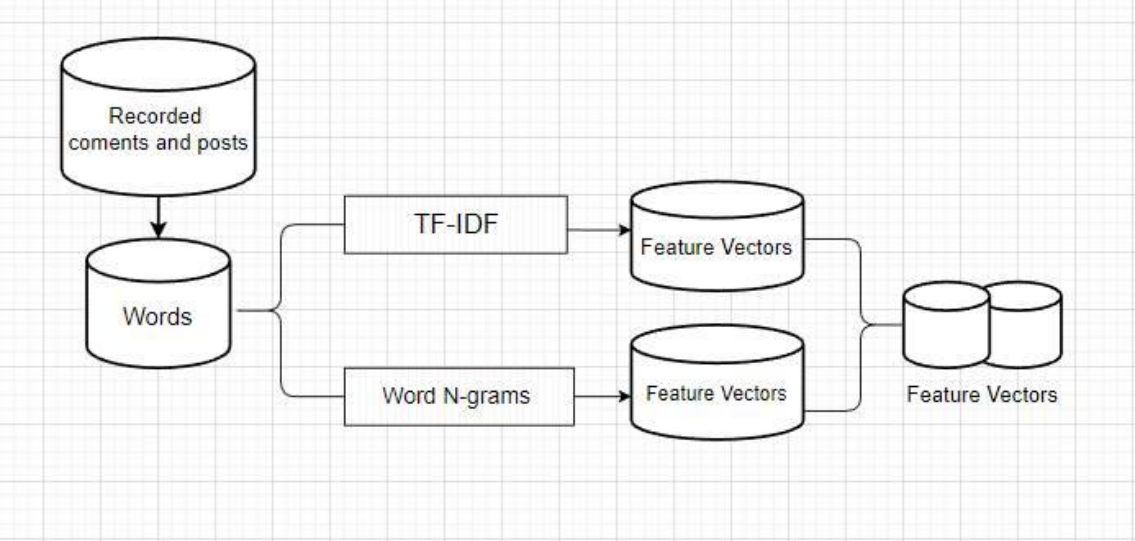

Figure 4. 3:Feature extraction model

4.4.1 N-gram:

N-gram is a technique in natural language processing (NLP) for extracting features from text data. It involves splitting text into contiguous sequences of n words (or characters) and counting their occurrences. The N-gram word prediction model anticipates the following word using probabilistic techniques. Note the use of N-1. The most popular N-gram techniques combine different sequences. Where N is the total number of words used in the probability sequence, add a word to a list of size N. For instance, they are referred to as unary, dual, and triple, respectively, if N=1, N=2, and N=3. East In this study, we develop N-gram functions for posts and comments using the N-gram word approach. These three n-grams are employed since the model's performance is independent of the number of N.

4.4.2 TFIDF:

One of the most popular feature modeling methods for detecting hate speech also uses this method. work team a gauge of the rise in the prominence and percentage of words in dataset documents A word's frequency of occurrence in a document. Chapter two contains more

discussion. The survey model is trained using the feature vectors that were extracted from each of the above techniques. Next, all feature extraction techniques are integrated, including TF-IDF and N-grams.

4.5 Machine Learning Models Building

Hate speech detection works on a sub-component of this architecture, proposed for Somali Hate. Train machine learning classifier on all feature vectors constructed from feature extraction component method. In this study, we mainly used machine learning to build classification models.

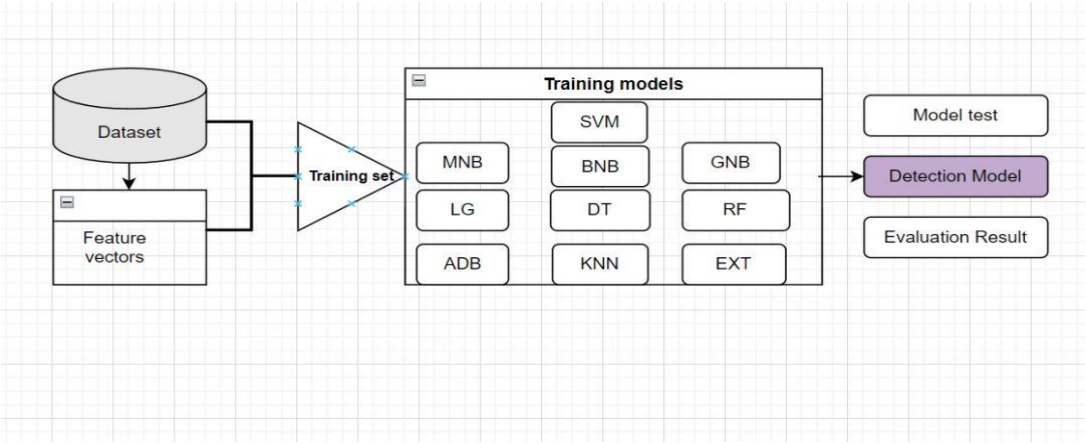

Figure 4. 4: Model Building Flow Diagram

RF, LR, SVM, DT, MNB, GNB, BNB, EXT, ADB and KNN classification algorithms for dataset features and labels. which you can select Detect hateful and normal speech in Somali. In search of the modeling process A dataset training set model that maps posts and comments based on functionality Use learning algorithms to identify classes. The result of this modeling is the trained model that can be used for Detecting hate speech in Somali and predictions as new results posts and comments.

4.6 Models Evaluation and Testing

Analyze and assess a dataset that includes a machine learning model that has been trained using hatred. Machine learning models' main goal is to find predictions that are as accurate as possible. generalization error caused by a model that was trained 28 on just a small amount of data. Use a variety of performance data to complete your research. The Confusion matrix, Accuracy, Precision, Recall, and F-Measure are the matrices that have been applied to the model.

- **True Positives (TP):** the outcomes of actual data points were true and the predicted is also true.
- **True Negatives (TN):** the outcomes when the actual point int false and predicted also False.
- **False Positives (FP):** The result of the data point's actual class is False, while the prediction is True.
- **False Negatives (FN):** The actual class of the data point's result is True, while the prediction is False.

4.6.1 Accuracy:

Accuracy is the problem of classifying correct predictions and calculating them as a sum. The number of correctly predicted models is divided by the number of all data events used in the model Calculate Accuracy Using Equations:

$$Accuracy = \frac{TN+TP}{\text{Total number of instance}}$$

Figure 4. 5:Equation of Accuracy

4.6.2 Precision:

precision measures the value of the true prediction and indicates how many times the model predicted correctly. Answer how many parts of the introduction are correct. precision calculation using the equation:

$$Precision = \frac{TP}{TP + FP}$$

Figure 4. 6:Equation of Precision

4.6.3 Recall:

Recall the answer for the number of correctly identified real positive parts. Recall to use calculation equation:

$$Recall = \frac{TP}{TP + FN}$$

Figure 4. 7:Equation of Recall

4.6.4 F-Measure:

The F-measure, often known as the F1-score or F-score, is a weighted indicator of test accuracy. Harmonized method of test accuracy and recovery. F1 is more practical than precise. The data set contains an uneven distribution of classes. This score is calculated according to the following formula:

$$F1 - score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)}$$

Figure 4. 8:Equation of F1-score

4.6.5 Confusion Matrix:

A confusion matrix is a technique for summarizing the performance of a classification algorithm. The number of correct and incorrect predictions is aggregated and broken down by count value. It varies according to each class. Provides additional information and errors caused by the categorizer. What matters is the type of error. Therefore, in this study, we used confusion matrix. Hate and OK-only two-class models and two-class models (Hate, Non-hate). Table 4.1 below shows an example format of a confusion matrix with three categories

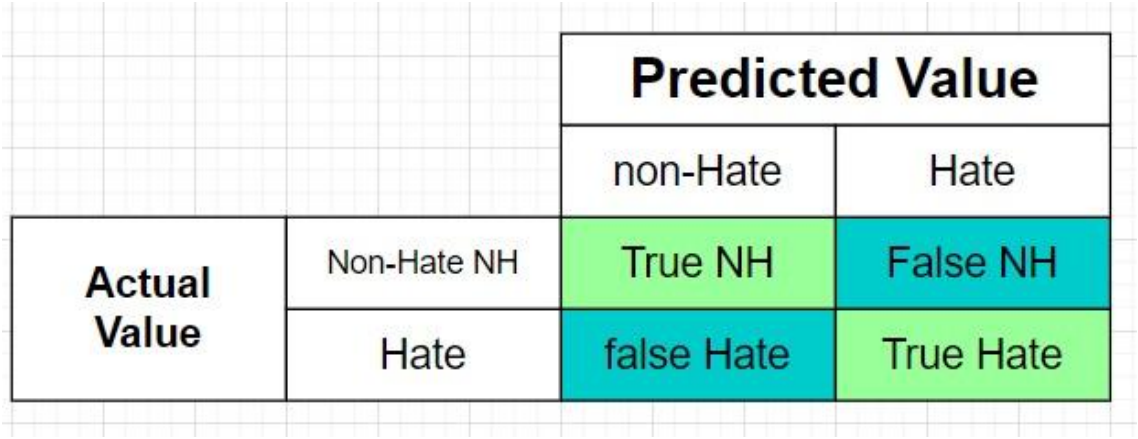

Figure 4. 9: Confusion Matrix

CHAPTER 5: IMPLEMENTATION AND TESTING

5.1 Introduction

In this chapter, the authors provide a detailed account of the implementation process, including data preprocessing, feature extraction, and model training. The performance of various machine learning algorithms, such as logistic regression, support vector machines, and decision trees, is evaluated to determine the most effective approach for detecting hate speech in Somali.

Moreover, the authors also experiment with different feature sets, such as bag-of-words, n-grams, and word embeddings, to assess their impact on the performance of the models. Through extensive experimentation and analysis, the study aims to provide insights into the most effective approaches for detecting hate speech in Somali and contribute to the development of automated systems for hate speech detection in low-resource languages.

Overall, this chapter provides a valuable contribution to the field of hate speech detection and showcases the potential of using machine learning techniques to address this critical social issue.

5.2 Dataset Description

We manually collected comments and posts from various sources, such as Facebook pages belonging to news organizations like BBC Somalia and Jowhar, as well as individual and group pages of politicians, activists, TV and radio broadcasts, and group associations. The dataset was pre-processed to remove all non-textual posts and comments, resulting in a total of 18,548 Somali posts and comments. Of these, 6048 were considered non-hate or normal data, while 3124 were classified as hate speech.

Table 5. 1: The Total Two-Class Dataset

Name of the class	**The number of comments**	**Percentage %**
Non-Hate	6048	65.94%
Hate	3124	34.06%
Total	9172	100%

The two-class dataset by Considering all Hate to Non-Hate to build the binary class dataset

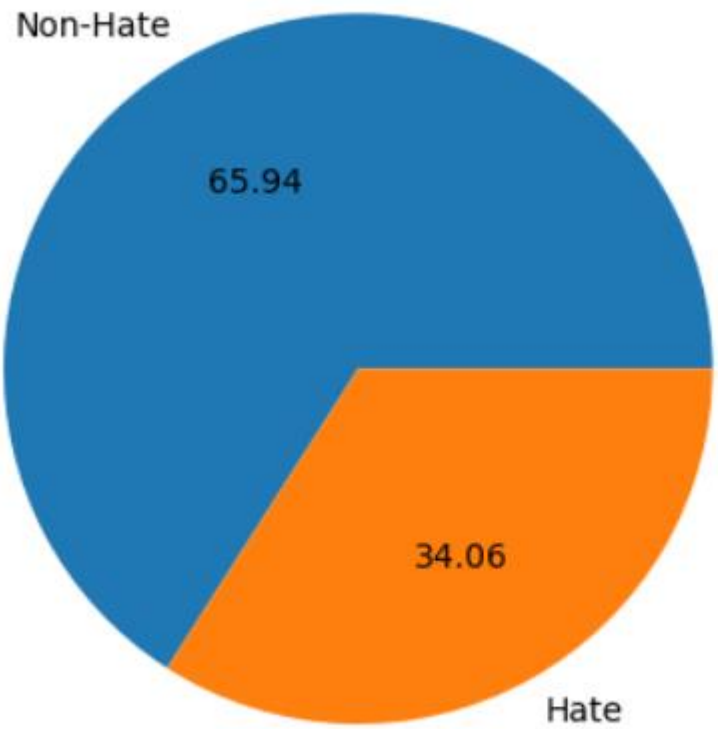

Figure 5. 1 The average Two-Class Dataset

5.3 Pre-Processing Process

For data pre-processing, the authors utilized Python programming language and relevant modules. After importing and loading essential library packages, we used the Pandas library to load the dataset for this study. To prepare the text data for analysis, we applied some modules in Python to remove stopwords and punctuations from the Somali text. This step is crucial as it helps to eliminate irrelevant or redundant words that do not contribute to the meaning of the text, and we used tokenization techniques to break down the pre-processed text into individual words or tokens, which are then used to train the machine learning models. The authors also employed data visualization techniques to gain insights into the distribution of the dataset and identify any patterns or trends.

5.3.1 Removing Stop Words and Punctuations

In the pre-processing stage, we removed stop words, special characters, and impolite characters from the imported dataset using Python programming language and relevant modules. The code used to perform this removal was implemented and tested to ensure the accuracy and effectiveness of the pre-processing techniques.

The following graph that shows the word frequency of the normal speech and Hate speech data. This graph highlights the most frequently occurring words in the dataset after removing the stop words

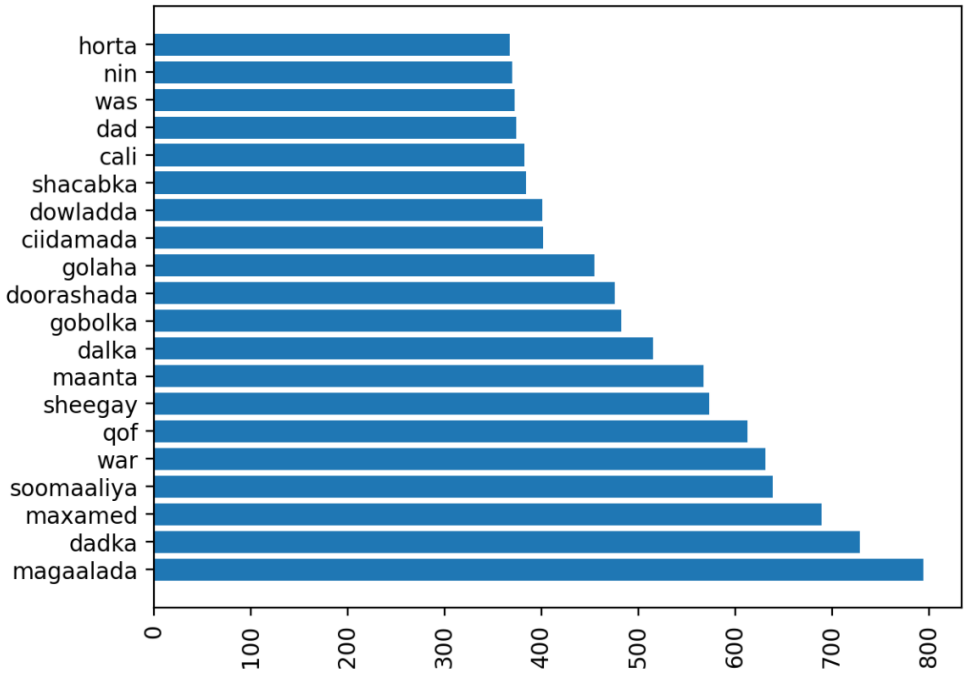

Figure 5. 2: both hate & Normal frequency

The following graph that shows the word frequency of the normal speech data. This graph highlights the most frequently occurring words in the dataset after removing the stop words

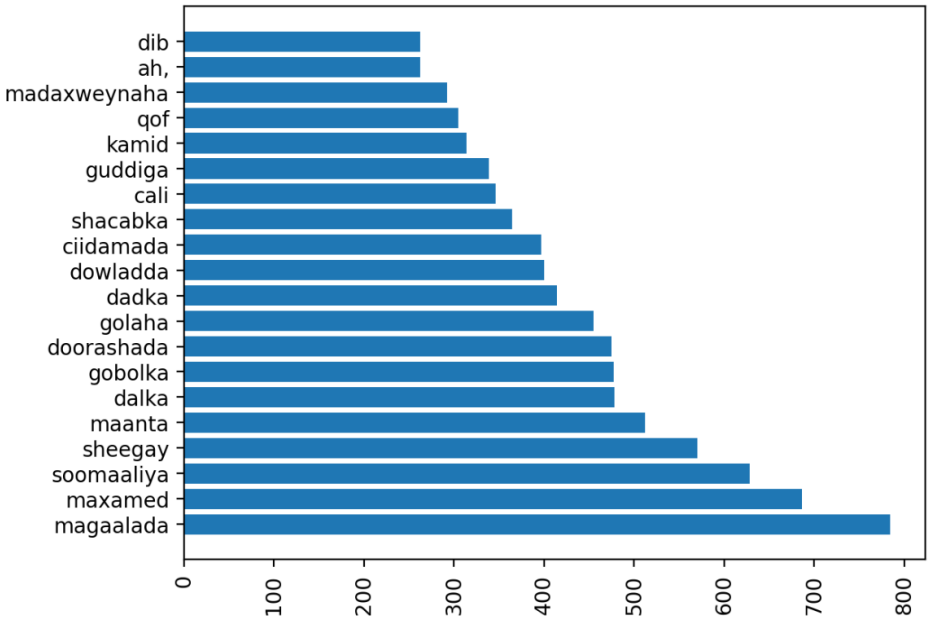

Figure 5. 3: non-hate speech frequency words

The following graph that shows the word frequency of the Hate speech data. This graph highlights the most frequently occurring words in the dataset after removing the stop words

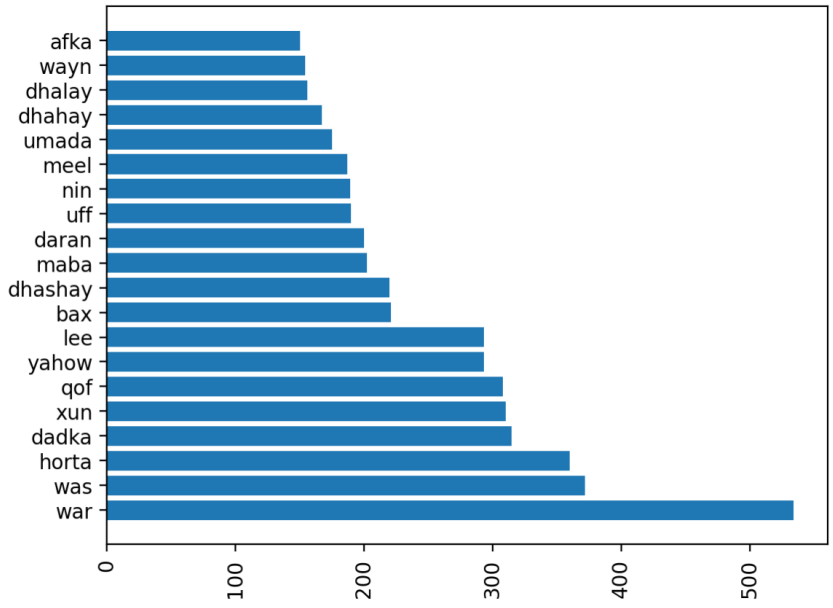

Figure 5. 4 : hate speech frequency words

5.3.2 Tokenization

Tokenization is a widely used technique in natural language processing (NLP) that involves dividing a piece of text into smaller units, known as tokens. These tokens can be individual words, sentences, or even smaller components like punctuation marks and special characters. In Python, the NLTK library offers a range of tokenization methods and functions, allowing researchers and developers to efficiently tokenize text data for analysis and processing. The following table shows how to split our dataset.

5.4 Feature Extraction

Feature extraction is a crucial step in machine learning and data analysis, where the goal is to extract relevant information or features from raw data that can be used to build models or make predictions. To train a machine learning model for hate speech detection, the dataset

needs to undergo a feature extraction process. This involves cleaning up the dataset and applying techniques such as tokenization, TF-IDF, and N-gram analysis using Scikit-Lean's Python modules. The goal is to transform the textual data into numerical representations, such as vectors, that can be used by the machine learning algorithms. This conversion allows the models to work with numbers instead of raw text, enabling them to learn patterns and make predictions based on the extracted features. The vectorizer program plays a crucial role in converting the text data into numerical features that can be fed into the machine learning model for training. By utilizing these feature extraction techniques, the machine learning model can effectively learn from the dataset and make predictions on new instances of hate speech.

5.5 Machine Learning Models Evaluation Results

Chapter 5 presents the experimental results of the hate speech detection models developed in this study. The models were evaluated based on their validation outcomes, and the results were analyzed using a confusion matrix and a classification report. These reports were generated using the sklearn package, which offers useful tools for evaluating machine learning models. we used these evaluation metrics to assess the performance of different machine learning algorithms and feature sets in detecting hate speech in the Somali language. The results provide insights into the effectiveness of different approaches and highlight the potential of using machine learning techniques to address the issue of hate speech.

The confusion matrix logistic regression model is very good when looking at the accuracy of 0.9657 which corresponds percentage as 96.57% and recall 0.9869 which is 98.69%, f1-score 0.9745 which is 97.45% and precision 0.9624 which is 96.24%. on the other side Hate and Normal according side of Classification report Its-Hate 0.97 0.95 0.96 615 non-Hate 0.98 0.99 0.98 1220 Accuracy 0.97 1835

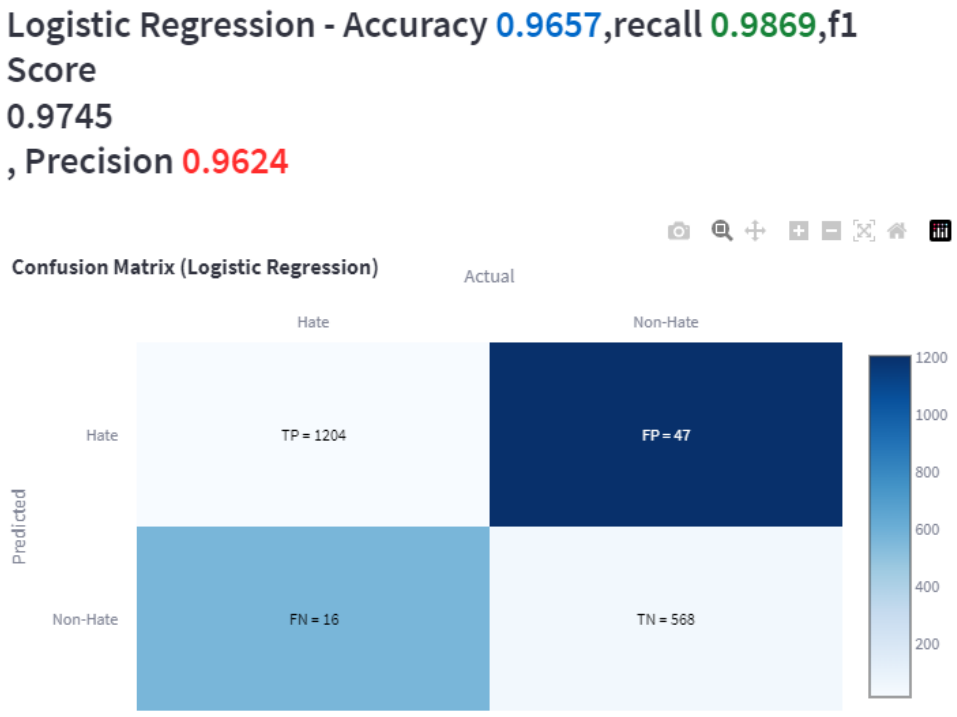

Figure 5. 5 : Logistic Regression Confusion Matrix

Figure 5. 6 : Logistic Regression Classification report

The confusion matrix Support Vector Machine model is Excellent when looking at the accuracy of 0.9744 which corresponds percentage as 97.44% and recall 0.9885 which is 98.85%, f1-score 0.9809 which is 98.09% and precision 0.9734 which is 97.34%. on the other side Hate and Normal according side of Classification report Its-Hate 0.97 0.95 0.96 615 non-Hate 0.98 0.99 0.98 1220 Accuracy 0.97 1835

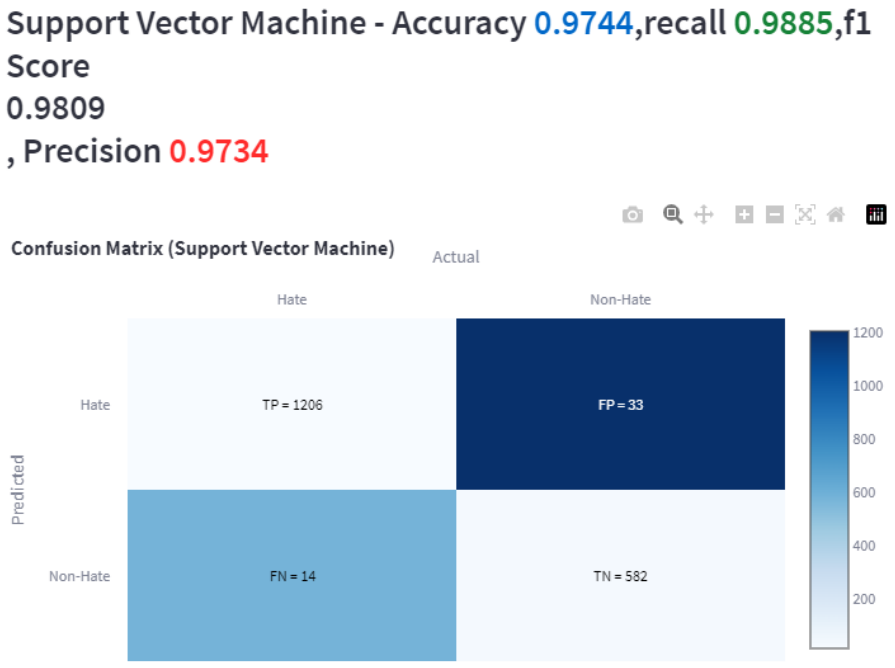

Figure 5. 7 : SVM Confusion Matrix

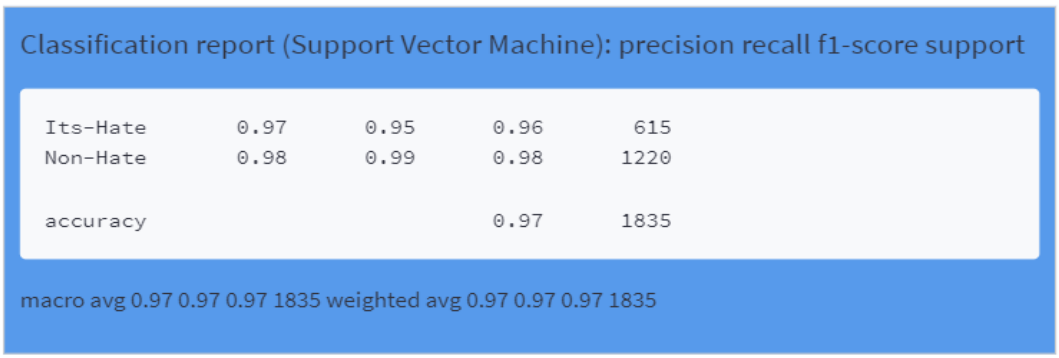

Figure 5. 8 : SVM Classification report

The confusion matrix Decision Tree model is a good when looking at the accuracy of 0. 9003 which corresponds percentage as 90.03% and recall 0.9303 which is 93.03%, f1-score 0.9254 which is 92.54% and precision 0. 9205 which is 92.05%. on the other side Hate and Normal according side of Classification report Its-Hate 0.97 0.95 0.96 615 non-Hate 0.98 0.99 0.98 1220 Accuracy 0.97 1835

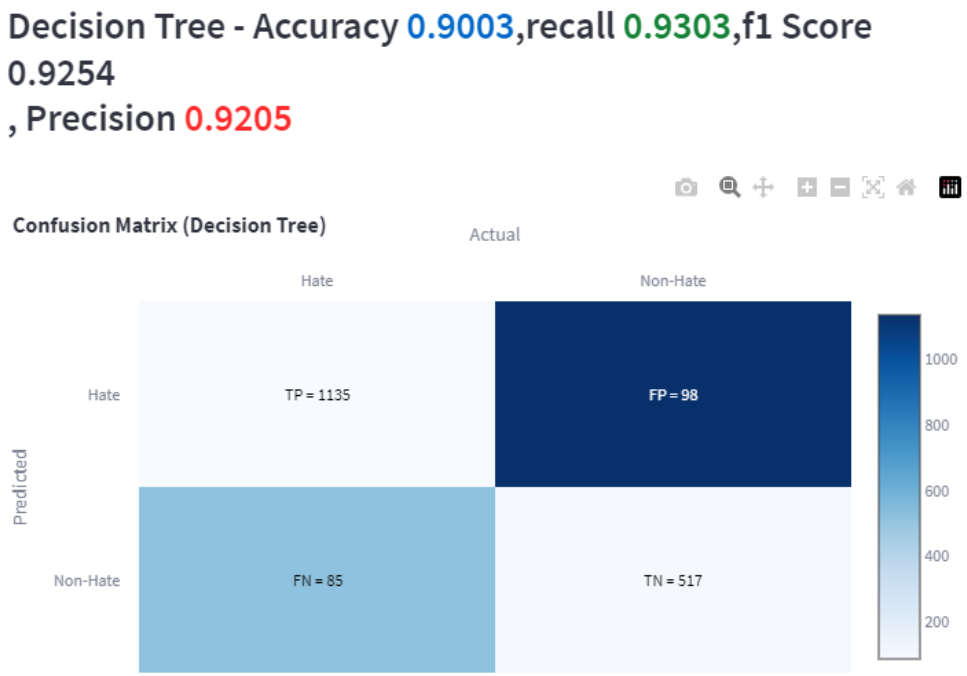

Figure 5. 9 : Decision Tree Confusion Matrix

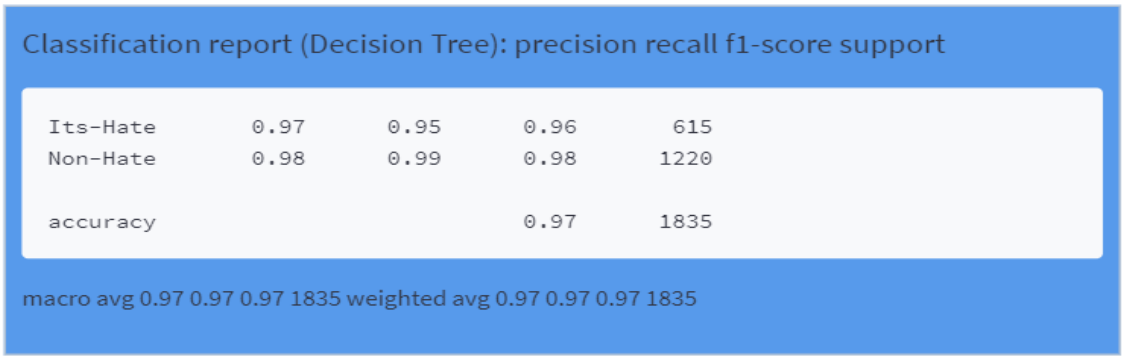

Figure 5. 10 : Decision Tree Classification report

The confusion matrix Multinomial Naive Bayes model is Avery good when looking at the accuracy of 0. 9749 which corresponds percentage as 97.49% and recall 0. 9812 is 98.12%, f1-score 0.9812 which is 98.12% and precision 0. 9772 is 97.72%. on the other side Hate and Normal according side of Classification report Its-Hate 0.97 0.95 0.96 615 non-Hate 0.98 0.99 0.98 1220 Accuracy 0.97 1835

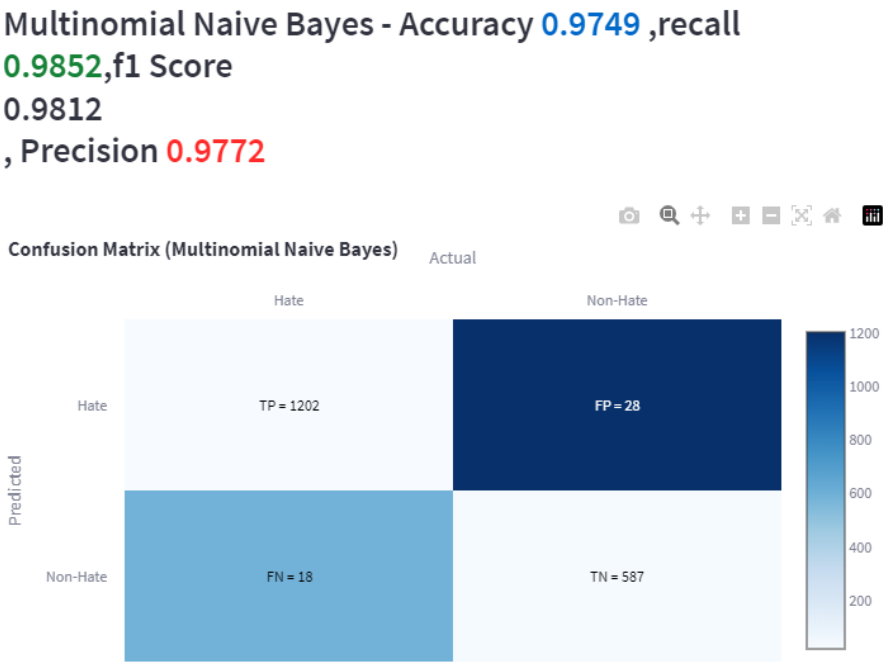

Figure 5. 11 : Multinomial Naive Bayes Confusion Matrix & Classification report

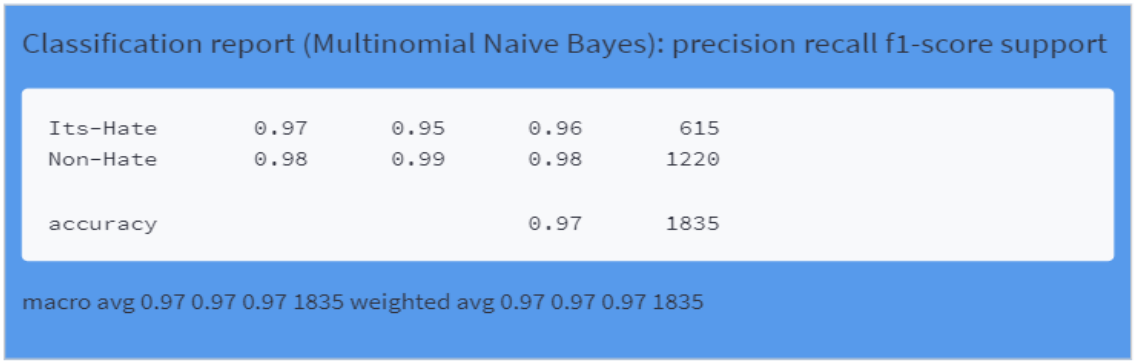

Figure 5. 12 : Multinomial Naive Bayes Confusion Matrix & Classification report

The confusion matrix GaussianNB model is good when looking at the accuracy of 0.8959 which corresponds percentage as 89.59% and recall 0.8607 is 86.07%, f1-score 0.9166 which is 91.66% and precision 0.9804 is 98.04%. on the other side Hate and Normal according side of Classification report Its-Hate 0.97 0.95 0.96 615 non-Hate 0.98 0.99 0.98 1220 Accuracy 0.97 1835

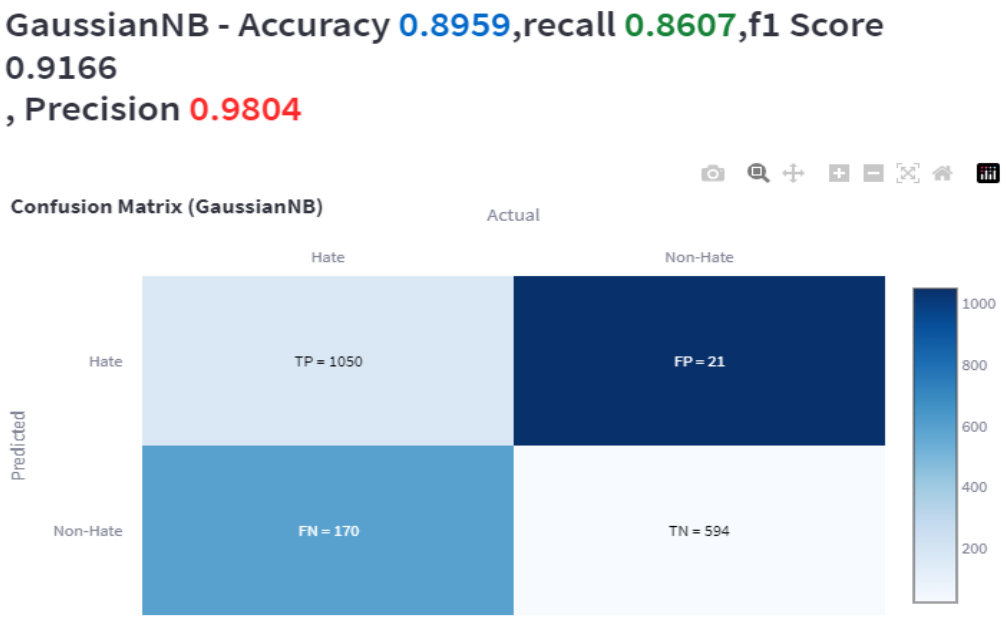

Figure 5. 13 : GaussianNB Confusion Matrix

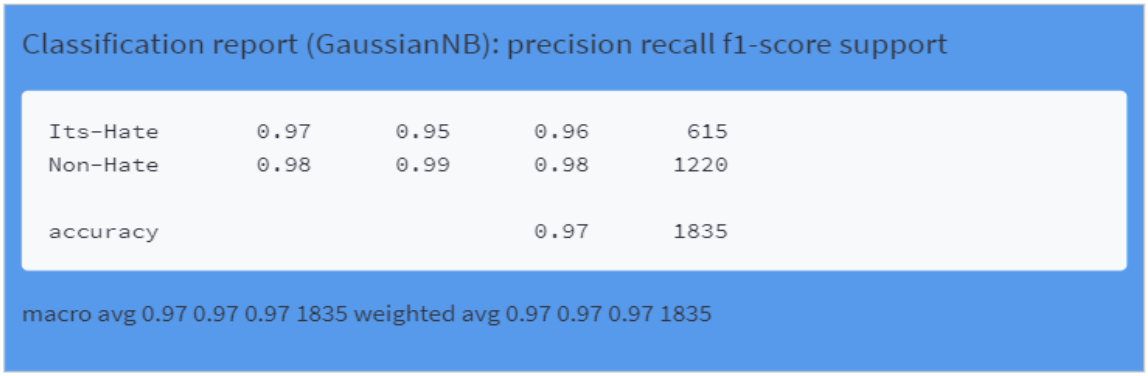

Figure 5. 14 : GaussianNB Classification report

The confusion matrix BernoulliNB model is very good when looking at the accuracy of 0.9428 which corresponds percentage as 94.28% and recall 0.9287 which 92.87%, f1-score 0.9557 which is 95.57% and precision 0. 9844 is 98.44%. on the other side Hate and Normal according side of Classification report Its-Hate 0.97 0.95 0.96 615 non-Hate 0.98 0.99 0.98 1220 Accuracy 0.97 1835

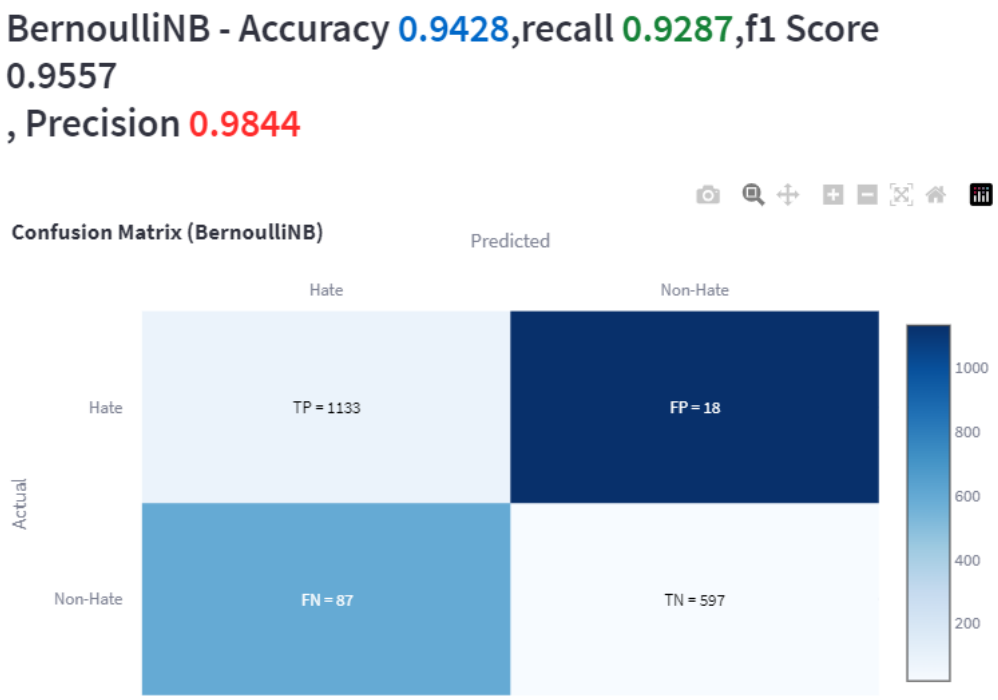

Figure 5. 15 : BernoulliNB Confusion Matrix

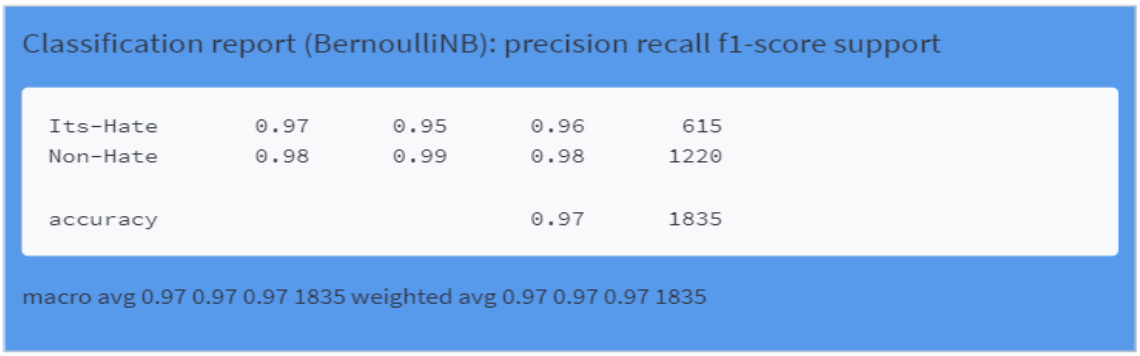

Figure 5. 16 : BernoulliNB Classification report

The confusion matrix Random Forest model is a good when looking at the accuracy of 0.9526 which corresponds percentage as 95.26% and recall 0. 9703 which is 97.03%, f1-score 0.9649 which is 96.49% and precision 0. 9500 which is 95%. on the other side Hate and Normal according side of Classification report Its-Hate 0.97 0.95 0.96 615 non-Hate 0.98 0.99 0.98 1220 Accuracy 0.97 1835

Figure 5. 17: Random Forest Confusion Matrix

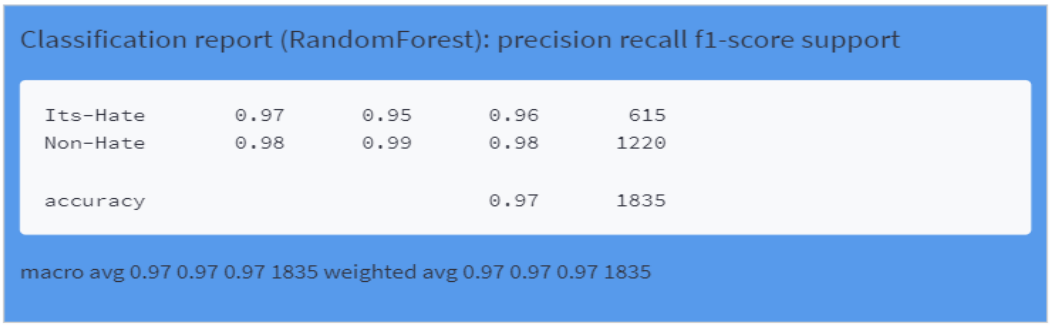

Figure 5. 18*:* Random Forest Classification report

The confusion matrix AdaBoost model is a good when looking at the accuracy of 0. 9210 which corresponds percentage as 92.10% and recall 0. 9484 which is 94.84%, f1-score 0. 9410 which is 94.10% and precision 0. 9338 which is 93.38%. on the other side Hate and Normal according side of Classification report Its-Hate 0.97 0.95 0.96 615 non-Hate 0.98 0.99 0.98 1220 Accuracy 0.97 1835

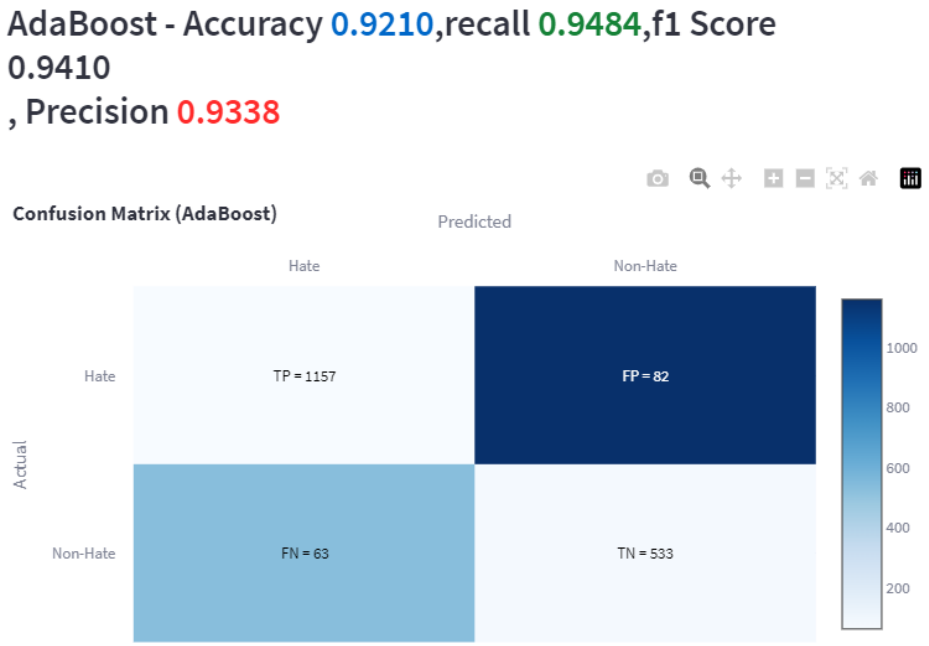

Figure 5. 19 : AdaBoost Confusion Matrix

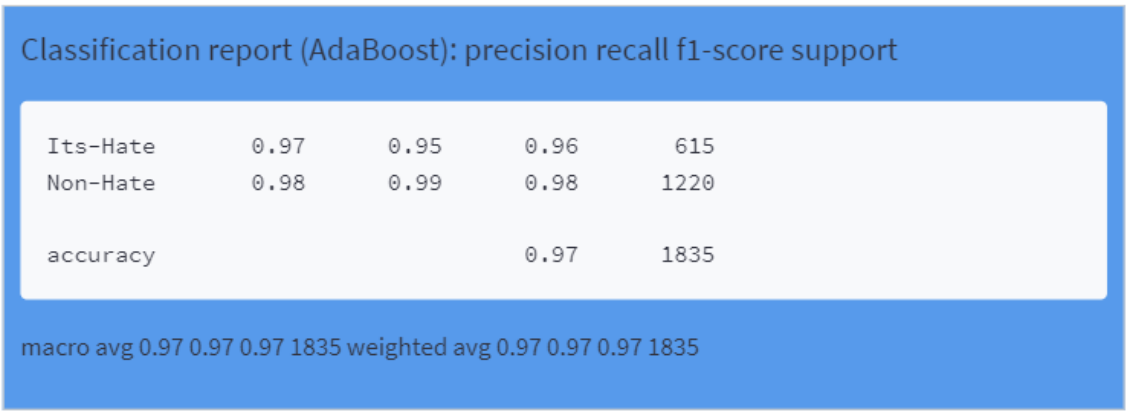

Figure 5. 20 AdaBoost Classification report

The confusion matrix ExtraTrees model is a good when looking at the accuracy of 0. 9700 which corresponds percentage as 97 % and recall 0. 9885 which is 98.85%, f1-score 0. 9777 which is 97.77% and precision 0. 9671 which is 96.71%. on the other side Hate and Normal according side of Classification report Its-Hate 0.97 0.95 0.96 615 non-Hate 0.98 0.99 0.98 1220 Accuracy 0.97 1835

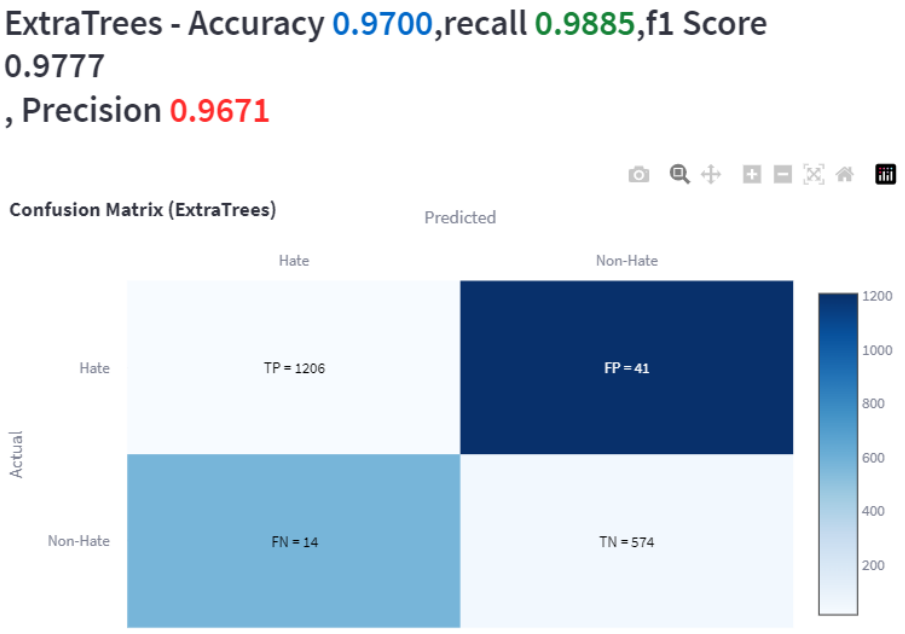

Figure 5. 21: ExtraTrees Confusion Matrix

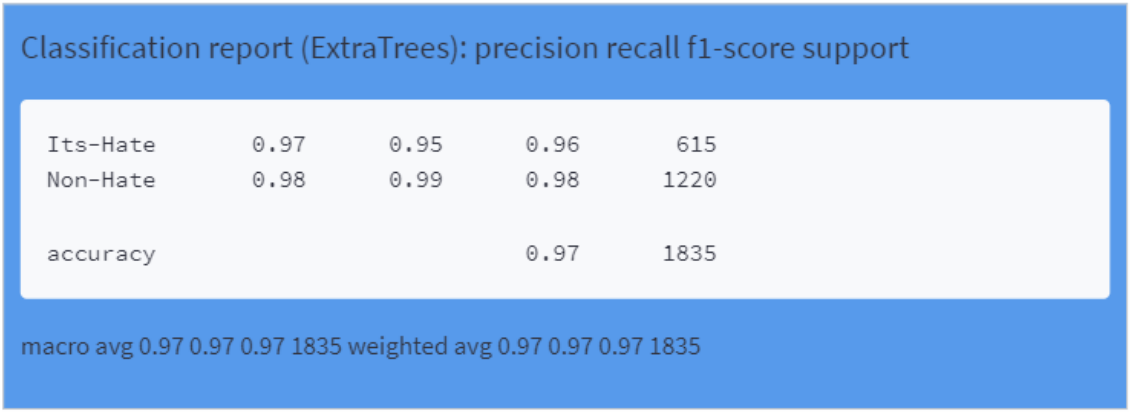

Figure 5. 22 : ExtraTrees Classification report

The confusion matrix KNeighbors model is a good when looking at the accuracy of 0.9700 which corresponds percentage as 97% and recall 0. 9885 which is 98.85%, f1-score 0. 9777 which is 97.77% and precision 0.9671 which is 96.71%. on the other side Hate and Normal according side of Classification report Its-Hate 0.97 0.95 0.96 615 Non-Hate 0.98 0.99 0.98 1220 Accuracy 0.97 1835

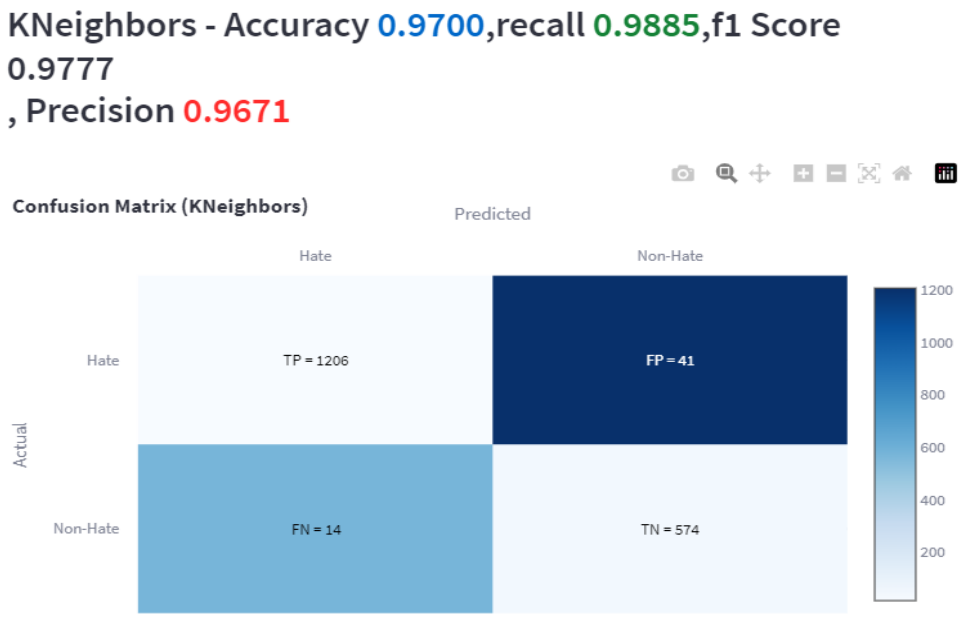

Figure 5. 23 : KNeighbors Confusion Matrix

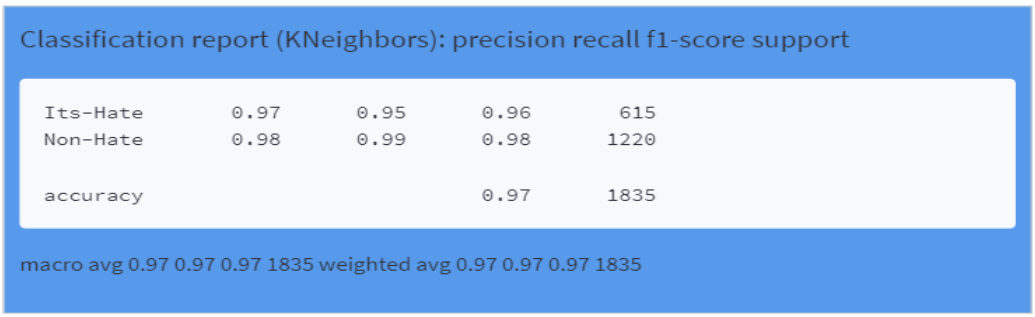

Figure 5. *24* KNeighbors Classification report

Table 5. 2 Summary for Classification Performance of The Models

Model	Accuracy	Recall	F1-score	Precision
logistic regression	0.96	0.98	0.97	0.96
Support Vector Machine	0.97	0.98	0.98	0.97
Decision Tree	0.89	0.92	0.91	0.91
Multinomial Naive Bayes	0.97	0.98	0.98	0.97
GaussianNB	0.89	0.86	0.91	0.98
BernoulliNB	0.94	0.92	0.95	0.98
Random Forest	0.95	0.98	0.96	0.95
AdaBoost	0.92	0.94	0.94	0.93
ExtraTrees	0.96	0.98	0.97	0.96
KNeighbors	0.96	0.98	0.97	0.96

5.6 Input Test Model

We created a method to determine whether the text is hated speech or normal speech after training many different models and discussing their results such as accuracy, precision, recall, and f1_score. We passed the method "Normal Speech" as an argument, so the model checked the text we passed and displayed the result shown below.

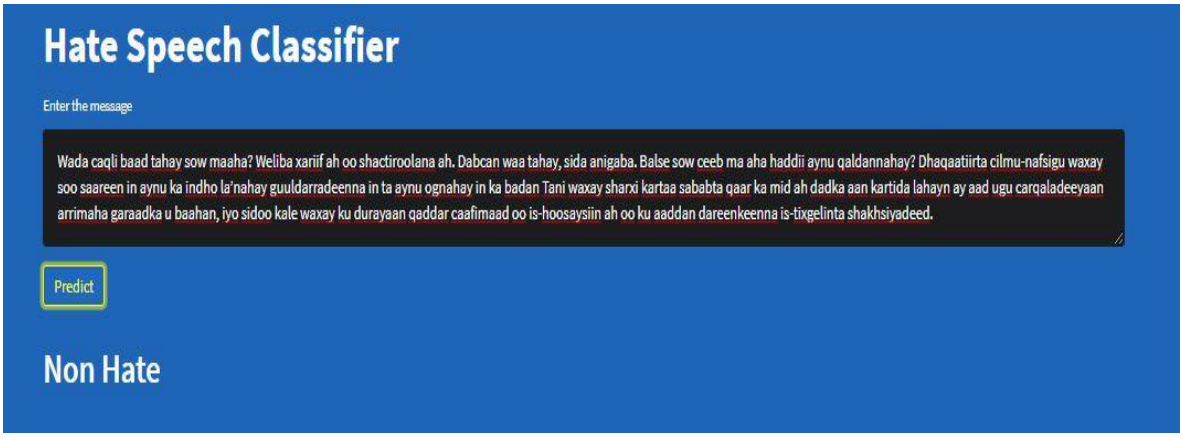

Figure 5. 25: normal speech output

Figure 5. 26: Hate speech output

Finally, we provided the method "Hate Speech" as an argument once more; as a result, the model checked the text we entered and displayed the following result.

CHAPTER 6: DISCUSSION OF RESULTS

In the realm of hate speech detection on social media, several research works have been conducted. However, prior to this study, there was a significant gap in the literature when it came to hate speech detection in the Somali language specifically for Facebook posts and comments. As natural language processing (NLP) applications are language-dependent, the absence of data and research in Somali necessitated manual data collection. The researcher meticulously gathered Facebook postings, comments, and selected sites to create a dataset for analysis. Feature extraction methods are crucial in capturing patterns and enabling machine learning techniques to transform human-readable text into machine-readable text or machine code.

In this study, the feature extraction techniques employed were n-grams and TF*IDF. While n-grams generally performed well, TF*IDF yielded the best results across all classification algorithms. The study utilized Support Vector Machine, Multinomial NB, Random Forest, Logistic Regression, Decision Tree, AdaBoost, ExtraTrees, BernoulliNB, GaussianNB, and k-nearest neighbor classifiers. Overall, the Multinomial NB consistently exhibited higher prediction power for all feature extraction algorithms, while GaussianNB showed the lowest performance. However, the performance varied depending on the type of feature extractor, with TF*IDF being the most effective.

The evaluation of the models was based on the test dataset, and the results indicated varying levels of prediction power among the categorization models. The Multinomial NB stood out as the most accurate classifier with an accuracy value of 97%. Consequently, it was chosen as the primary model for this study due to its superior accuracy and predictive potential in hate speech detection. Despite the success of the hate detection model, it is crucial to acknowledge the study's limitations. Misclassifications of hate and hate-free posts and

comments were observed in the classification model. These limitations stemmed from inconsistencies in manually labeling the collected dataset and the inadequacy of training datasets, among other factors. The nature of hate speech expression makes it challenging to label content as hate-free, as some instances may use explicit hate words, while others might employ more subtle and indirect methods to convey hateful messages. Nonetheless, the study provided valuable insights and demonstrated the importance of undertaking research in unexplored territories, such as hate speech detection in the Somali language. It revealed the significance of building and executing a project that has not been previously conducted in the specific language context. Achieving the objectives outlined in the research and contributing to related work solidified the study's impact and relevance.

Moving forward, future research should continue to address the limitations observed in this study. Efforts should be directed towards improving consistency in dataset labeling, increasing the size of training datasets, and exploring more advanced NLP techniques for hate speech detection in the Somali language. By addressing these aspects, hate speech detection models can be further refined, fostering a safer and more respectful online environment for Somali-speaking users on social media platforms like Facebook.

Finally, we learned something new while working on this project, such as having the courage to do research on this topic that has not been done before in Somalia. We learned a lot about building and doing a project that no one has done before like our language, we also achieved the goal we mentioned in the research objectives and also related work.

CHAPTER 7: CONCLUSION AND FUTURE WORK

7.1 Introduction

This chapter includes the conclusion and future works of the study of hate speech detection from Facebook post and comments in Somalia language. The conclusion shows basic process of implementation and analysis as well as summarized basic finding of analysis then clearly show the best findings. Result of the research should be recommended as a solution for the real-world problems of hate speech detection and future work is expected to improve Somalia hate speech detection.

7.2 Conclusion

In recent years, the proliferation of hate speech on social media platforms has become a significant concern, and Facebook, being one of the largest and most influential platforms, is no exception. To combat this issue, researchers have turned to machine learning techniques as a potential solution. One such study delved into the problem of hate speech specifically in the Somali language on Facebook. In order to develop an effective solution, the study required a comprehensive understanding of hate speech, non-hate speech, and the nuances of the Somali language, all of which were thoroughly covered in the preceding chapter.

The study's methodology involved several crucial steps to build an accurate hate speech identification model. Firstly, to create the dataset, a substantial amount of posts and comments from Facebook pages were gathered. These posts were then manually classified into two categories: hate speech and non-hate speech, a process that involved the researcher and another annotator. The resulting dataset consisted of 3,124 hate speech samples and 6,048 non-hate speech samples, totaling 9,172 posts and comments. Next, the researchers developed annotation guidelines to ensure consistency and reliability in the classification process. Proper pre-processing techniques were employed to clean and prepare the text data

for analysis. To extract relevant features from the text, the study utilized n-grams and TF*IDF (Term Frequency-Inverse Document Frequency) techniques.

After meticulously preparing the dataset and extracting relevant features, we embarked on a comprehensive experimentation phase involving ten distinct classification algorithms. These algorithms encompassed a wide range, including Support Vector Machine, Multinomial Naive Bayes, Random Forest, Logistic Regression, Decision Tree, AdaBoost, ExtraTrees, Bernoulli Naive Bayes, Gaussian Naive Bayes, and k-Nearest Neighbor. Each of these algorithms was tested with different feature extraction techniques to identify the most effective combination. The evaluation of the models was done using accuracy, a confusion matrix, and a classification report. Among the ten algorithms, heights algorithm the Multinomial Naive Bayes classification algorithm with TF*IDF feature extraction emerged as the most promising. It achieved an impressive accuracy of 97.49 %, an F1-score of 98%, precision of 97.72 %, and recall of 98.52 %. Based on the results, the study's researcher concluded that the Multinomial NB model with TF*IDF feature extraction outperformed the other nine models in detecting hate speech in posts and comments written in the Somali language on Facebook. The implications of this study are significant as it offers a viable approach to tackle hate speech in a specific language context on a major social media platform. The findings highlight the importance of incorporating machine learning techniques and language-specific considerations when developing hate speech detection systems.

However, it's essential to acknowledge that the fight against hate speech is an ongoing and complex challenge, and there is no one-size-fits-all solution. Further research and refinement of these models will be necessary to adapt them to different languages and social media platforms effectively. Nonetheless, this study represents a commendable step towards

mitigating the harmful effects of hate speech and fostering a safer and more inclusive online environment for all users.

7.3 Future Work

In our preliminary study, we utilized a supervised learning algorithm along with text mining features to detect hate speech in Somali text. However, to enhance the performance of our system, we are eager to explore the effectiveness of unsupervised machine learning models. Our current research is limited to the Somali language, but future endeavors will aim to create datasets and models that can comprehensively detect hate speech on various social media platforms.

Moreover, we are aware that online content often includes non-text elements like images, emojis, and videos, which may carry hateful expressions. Thus, we need to extend our detection capabilities to encompass these multimedia components.

Furthermore, we plan to expand our dataset to include diverse Somali dialects, such as May-May, and other regional variations spoken in the country. To keep our system up-to-date and effective, we are committed to implementing regular updates that allow it to adapt to evolving trends in online hate speech

Reference

Ababa, A. (2021). *Department of Computer Science Hate Speech Detection Framework from Social Media Content : The Case of Afaan Oromoo Language Lata Guta kanessaa A Thesis Submitted to the Department of Computer Science in Partial Fulfilment for the Degree of Master of Scie*.

Agarwal, S., & Sureka, A. (2017). Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website. *ArXiv Preprint ArXiv:1701.04931*.

Ahammed, S., Rahman, M., Niloy, M. H., & Chowdhury, S. M. M. H. (2019). Implementation of machine learning to detect hate speech in Bangla language. *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 317–320.

Alfina, I., Mulia, R., Fanany, M. I., & Ekanata, Y. (2017). Hate speech detection in the Indonesian language: A dataset and preliminary study. *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 233–238.

Ayodele, T. O. (2010). Types of machine learning algorithms. *New Advances in Machine Learning*, *3*, 19–48.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. *Proceedings of the 26th International Conference on World Wide Web Companion*, 759–760.

Biere, S., Bhulai, S., & Analytics, M. B. (2018). Hate speech detection using natural language processing techniques. *Master Business AnalyticsDepartment of Mathematics Faculty of Science*.

De Smedt, T., Jaki, S., Kotzé, E., Saoud, L., Gwóźdź, M., De Pauw, G., & Daelemans, W. (2018). Multilingual cross-domain perspectives on online hate speech. *ArXiv Preprint ArXiv:1809.03944*.

El Naqa, I., & Murphy, M. J. (2015). What is machine learning? In *machine learning in radiation oncology* (pp. 3–11). Springer.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*(4), 1–30.

Laub, Z. (2019). Hate speech on social media: Global comparisons. *Council on Foreign Relations*, *7*.

Mondal, M., Silva, L. A., & Benevenuto, F. (2017). A measurement study of hate speech in social media. *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, 85–94.

Mossie, Z., & Wang, J.-H. (2018). Social network hate speech detection for Amharic language. *Computer Science & Information Technology*, 41–55.

Mossie, Z., & Wang, J.-H. (2020). Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, *57*(3), 102087.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). A BERT-based transfer learning approach for hate speech detection in online social media. *International Conference on Complex Networks and Their Applications*, 928–940.

Mullah, N. S., & Zainon, W. M. N. W. (2021). Advances in machine learning algorithms for hate speech detection in social media: a review. *IEEE Access*.

Oljira, M. (2020). Sentiment Analysis of Afaan Oromo using Machine learning Approach. *International Journal of Research Studies in Science, Engineering and Technology*, *7*(9), 7–15.

Omar, A., Mahmoud, T. M., & Abd-El-Hafeez, T. (2020). Comparative performance of machine learning and deep learning algorithms for Arabic hate speech detection in osns. *The International Conference on Artificial Intelligence and Computer Vision*, 247–257.

Oriola, O., & Kotze, E. (2020). Evaluating Machine Learning Techniques for Detecting

Offensive and Hate Speech in South African Tweets. *IEEE Access*, *8*, 21496–21509. https://doi.org/10.1109/ACCESS.2020.2968173

Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Effective hate-speech detection in Twitter data using recurrent neural networks. *Applied Intelligence*, *48*(12), 4730–4742.

Romim, N., Ahmed, M., Talukder, H., & Islam, S. (2021). Hate speech detection in the bengali language: A dataset and its baseline evaluation. *Proceedings of International Joint Conference on Advances in Computational Intelligence*, 457–468.

Ruwandika, N. D. T., & Weerasinghe, A. R. (2018). Identification of hate speech in social media. *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 273–278.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 1–21.

Sigurbergsson, G. I., & Derczynski, L. (2019). Offensive language and hate speech detection for Danish. *ArXiv Preprint ArXiv:1908.04531*.

Tesfaye, S. G., & Kakeba, K. (2020). *Automated amharic hate speech posts and comments detection model using recurrent neural network*.

Themeli, C. K. (2018). Hate Speech Detection using different text representations in online user comments. *No. October*, *2018*.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. *Proceedings of the Second Workshop on Language in Social Media*, 19–26.

Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, *6*, 13825–13835.

Wu, C. S., & Bhandary, U. (2020). Detection of hate speech in videos using machine learning. *2020 International Conference on Computational Science and Computational*

Intelligence (CSCI), 585–590.

Yuan, L., Wang, T., Ferraro, G., Suominen, H., & Rizoiu, M.-A. (2019). Transfer learning for hate speech detection in social media. *ArXiv Preprint ArXiv:1906.03829*.

Zaghi, C. (2019). *Automatic detection of hate speech in social media*. *June*.

APPENDIX A: IMPORTING LIBRARIES

```
import numpy as np

import pandas as pd

import nltk

import string

from nltk.corpus import stopwords

from nltk.stem import PorterStemmer

from wordcloud import WordCloud

from collections import Counter

from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer

from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier, ExtraTreesClassifier

from sklearn.naive_bayes import GaussianNB, MultinomialNB, BernoulliNB

from sklearn.tree import DecisionTreeClassifier

from sklearn.metrics import accuracy_score, confusion_matrix, precision_score

from sklearn.linear_model import LogisticRegression

from sklearn.neighbors import KNeighborsClassifier

from sklearn.svm import SVC

from sklearn.ensemble import StackingClassifier

import pickle

from sklearn.metrics import accuracy_score, confusion_matrix, precision_score
```

APPENDIX B: MODEL EVALUATION

```
# Define a list of classifiers
classifiers = {
    "Gaussian Naive Bayes": GaussianNB(),
    "Multinomial Naive Bayes": MultinomialNB(),
    "Bernoulli Naive Bayes": BernoulliNB(),
    "Random Forest": RandomForestClassifier(n_estimators=600),
    "Decision Tree": DecisionTreeClassifier(criterion='entropy', random_state=42),
    "AdaBoost": AdaBoostClassifier(),
    "Extra Trees": ExtraTreesClassifier(),
    "Logistic Regression": LogisticRegression(),
    "K-Nearest Neighbors": KNeighborsClassifier(),
    "Support Vector Classifier": SVC(kernel='sigmoid', gamma=1.0)
}

# Iterate through each classifier
for clf_name, clf in classifiers.items():
    clf.fit(X_train, y_train)
    y_pred = clf.predict(X_test)

    # Calculate and explain accuracy
    accuracy = accuracy_score(y_test, y_pred)
    print(f"{clf_name} Accuracy: {accuracy:.4f}")
```

```
# Calculate and explain confusion matrix
confusion = confusion_matrix(y_test, y_pred)
print(f"{clf_name} Confusion Matrix:")
print(confusion)

# Calculate and explain precision
precision = precision_score(y_test, y_pred)
print(f"{clf_name} Precision: {precision:.4f}")

print("\n")
```

APPENDIX C: DATA PREPROCESSING AND SAMPLE MANUAL TEST

```
def transform_text(text):
    text = text.lower()
    text = nltk.word_tokenize(text)

    y = []
    for i in text:
        if i.isalnum():
            y.append(i)

    text = y[:]
    y.clear()

    for i in text:
        if i not in stopwords.words('Somali') and i not in string.punctuation:
            y.append(i)

    text = y[:]
    y.clear()

    for i in text:
        y.append(ps.stem(i))

    return " ".join(y)
```

```
tfidf = pickle.load(open('vectorizer.pkl','rb'))
model = pickle.load(open('model.pkl','rb'))

st.title("Somali Hate Speech Detection in Machine Learning")

input_sms = st.text_area("Enter the message")
if st.button('Predict'):
    # 1. preprocess
    transformed_sms = transform_text(input_sms)
    # 2. vectorize
    vector_input = tfidf.transform([transformed_sms])
    vector_input = vector_input.toarray()
    # print(default_text.find(i))
    # print(default_text)
    if not input_sms:
        st.header(" please enter a data")
    elif input_sms in default_text:
        st.header("Non-Hate")
        rain(
        emoji="😍🥰😜😝😆😛❤️💖💋",
        font_size=20,  # the size of emoji
        falling_speed=3,  # speed of raining
        animation_length=0.7,  # for how much time the animation will happen
```

```
    )
else:

# 3. predict
 result = model.predict(vector_input)[0]

 if result == 0:
    st.header("hate")
    rain(
    emoji="😡🤬😈🤬😡😝🤬😡😈",
    font_size=20,  # the size of emoji
    falling_speed=3,  # speed of raining
    animation_length=0.7,  # for how much time the animation will happen
    )

 else:
    st.header("Non Hate")
    rain(
    emoji="😍🥰😜😝😆😛❤️💖💞",
    font_size=20,  # the size of emoji
    falling_speed=3,  # speed of raining
    animation_length=0.7,  # for how much time the animation will happen
    )
```