JIMMMA UNIVERSITY

INSTITUTE OF TECHNOLOGY

FACULITY OF COMPUTING AND INFORMATICS

DEPARTMENT OF DATA SCIENCE

Research Proposal on Hate Speech Detection Model for Afan Oromo's Texts on Social Media Using Machine Learning

By

NO	NAME	ID
1	Asmellash G/yesus	RM0933/14-0

Submitted to Mr.Geletaw

Jimma, Ethiopia

July , 2022

Table of Contents Page-no

LIST OF TABLE

LIST OF FIGURES

Abstract

Objectives: This study aims to develop a hate speech detection model for Afan Oromo's texts on social networks like Facebook and Twitter using a machine learning algorithm. Methods: we collected comments and posts fromsocial media like Facebook and Twitter pages of BBC Afan Oromo, OBN AfanOromo, Fana Afan Oromo Program, Politicians, Activists, Religious Men, and Oromia Communication Bureau using Face pager tool. The collected data was labelled using Afan Oromo hate speech evaluation system we developed. Text preprocessing tasks applied on data to remove special characters, stop-words,HTML Tags, extra whitespaces, numbers, and lemmatization. The n-gram and TF-IDF and Count vectorizer was applied for feature extraction task to obtain Afan Oromo hate speech detection dataset.Researchers split dataset into train and test set.Finally, we Support Vector Classifier,GuissianNB, kNeighborsClassifier, adaBoostClassifier, BaggingClassifier, GradientBoostingClassifier, ExtraTreesClassifier, XGBClassifier and Random Forest Classifier on 80% of trained data. The performance of proposed model also evaluated using Accuracy-score. We also test the performance of developed model by loading test set into it. Findings: Hate speech on social media violates the welfare of Ethnic groups and citizens for living together. Many researches have been doing for English, Amharic, and other Languages to detect hate content from social media. This study has focused on developing a prototype for Afan Oromo hatespeech detection model using machine learning algorithms and evaluate its performance in which we found Random Forest Classifier scored highest accuracy-score value is 76.9%. In this study, the n-gram and TF-IDF used for feature extraction approach to build model that detect Afan Oromo hate speech on Social media

Keywords: Afan Oromo; Decision tree; Facebook; Hate Speech; Random Forest Classifier; Machine Learning and Social Media

1 Introduction

We are now dealing with many problems in the world and in our country right now. "Hate speech can be cited as a major issue." These hate speech spread at different times and in different ways. From that ways one and the first is social media. In the past, it is an undeniable fact that successive years of these hate have increased.

Hate speech is commonly defined as any communication that disparages a person or a group basedon some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic[1]. It is able to influence the behavior of in-group members through socialization to adopt and understand a particular ideology to recruit members through the construction of a common enemy, which is constructed as evil or a cultural or economic threatto the in-group. Hate speech can therefore serve as an effective tool to intimidate minorities; promote violence and intolerance; and recruit new members – and allow these messages to form part of the permanent visible fabric of society. Hate speech can transmitted through a number of media, and can attack any number of groups.

Due to the massive rise of user-generated web content, in particular on social media networks, the amount of hate speech is also steadily increasing. Over the past years, interest in online hate speech detection and particularly the automation of this task has continuously grown, along with the societal impact of the phenomenon. Natural language processing focusing specifically on this phenomenon is required since basic word filters do not provide a sufficient remedy. In recent years also, the exponential growth of social media such as Facebook, Twitter, linked line, YouTube andothers has been increasingly exploited for the propagation of hate speech and the organization of hate based activities. The anonymity and mobility afforded by such media has made the breeding and spread of hate speech eventually leading to hate crime effortless in a virtual landscape beyond the realms of traditional law enforcement.

Social media has positive and negative impact in social-economy, politics of one country. The positive impacts are it helps peoples to exchange opinion digitally and help peoples to promote product through social media without payment. However, the negative impact is hate speech, which is attacking peoples based on characteristics like race, ethnicity, gender, and sexualorientation, or threats of violence towards others. However, there is No formal definition exists but there is a

consensus that it is speech that targets disadvantaged social groups in a manner that

is potentially harmful to them. A large number of research has been conducted in recent years to develop automatic methods for hate speech detection in the social media domain. These typically employ semantic content analysis techniques built on Natural Language Processing (NLP) and Machine Learning (ML) methods, both of which are core pillars of the Semantic Web research.

Generally, hate speech is an unfortunately common occurrence on the Internet and in some cases culminates in severe threats to individuals group of people by verbal attacks and promotion of hated based on race, ethnicity and national origin and other. To overcome the problem on social media many researchers are try to detect hate speech by using different methodology and approach.The aim of this paper also detection Afaan Oromo hate speech by using machine learning Natural language processing.

2 Statement of the problem

Hate and dangerous speech is a serious and growing problem in Ethiopia, both online and offline. It has contributed to the growing ethnic tensions and conflicts across the country thathave created more than 1.4 million new internally displaced people in the first half of 2018 alone. In our country Ethiopia, the hate speeches are spread through social media by using different natural language. From those languages one is Afaan Oromo language takes place alot of role. Afaan Oromo is one of the sub-Saharan language, Even though, users are also using Afan Oromo on social media platforms to express emotions, feelings, and opinion in formof comments and posts that contain hatred ideas which leads to discrimination, social conflict, and even human genocide, yet,no research work attempted to develop hate speech detection prototype for Afan Oromo for any social media platforms. Our input variable (independent variable) is text in our dataset it is the feature post and our output variable (dependent variable) are (label) in our dataset is the feature label, so, itneeds to develop model hate speech detection model for Afan Oromo on social media.

3 Literature Review

Online spaces are often exploited and misused to spread content that can be degrading, abusive, or otherwise harmful to people. Hateful speech has become a major problem for every kind of online platform where user-generated content appears from the comment sections of news websites to real-time chat sessions[2]. There are many researches have done to solve crisis of hate speech onsocial media. Among them, many researches done in abroad countries and some researches now have been doing in Ethiopia. Many researchers interested on classifying text whether they are hateful, offensive and not.

The complexity of the natural language constructs renders the task quite challenging. Irrespectiveof the use of NLP approaches, we can distinguish two major categories in the existing solutions tothe hate-speech problem: The Unsupervised learning and the Supervised learning. The hate speech detection from social media was developed by applying supervised classification Methods. The researcher was applied a linear Support Vector Machine (SVM) classifier and used three groups of features extracted for these experiments surface n-grams, word skip-grams, and Brownclusters. They try to classify their data set in three classes. Those are HATE): contains hate speech,OFFENSIVE): contains offensive language but no hate speech, (OK): no offensive content at all.The dataset features 14,509 English tweets annotated by a minimum of those three annotators.

Effective hate-speech detection in Twitter data using recurrent neural Networks developed in [3]. They proposed a detection scheme that is an ensemble of Recurrent Neural Network (RNN) classifiers, and it incorporates various features associated with user related information, such as the users' tendency towards racism or sexism. This data is fetch as input to the above classifiers along with the word frequency vectors derived from the textual content. They evaluate their approach on a publicly available corpus of 16k tweets, and the results demonstrate its effectivenessin comparison to existing state-of-the-art solutions. They applied before training the neural network with the labeled tweets, it is necessary to apply the proper tokenization to every tweet. Inthis way, the text corpus split into word elements, taking white spaces and the various punctuationsymbols used in the language into account.

Hate speeches can be token by Media like TV channels but most hates occurs on social media. According to literature, most of hate speeches are frequently done on Twitters to differentiating Comments and post whether post is hateful or not. To identify this issues Pinkesh Badjatiya1y and

Shashank Gupta Deep Learning for Hate Speech Detection in Tweets [4]to solve complexity risingin automatic hate speech detection techniques by using semantic word embedding with deep learning algorithms. In addition to this Björn Gambäck and Utpal Kumar Sikdara worked on deeplearning based Twitter hate-speech text classification system. The classifier assigns each tweet toone of four predefined categories: racism, sexism, both (racism and sexism) and non-hate-speech. Four Convolutional Neural Network models were trained on resp. character 4-grams, word vectorsbased on semantic information built using word2vec, randomly generated word vectors, and wordvectors combined with character n-grams[5].

Most researches are mostly focused on identifying comments whether they are hateful not hateful.So identifying comments and text online is challenging. So many researchers used machinelearning and NLP techniques to solve this problem. Among them Dulan S. Dias, Madhushi D. Welikala , Naomal G.J. Dias worked on building a text analytics model with machine learningthat can be used to filter racist comments in Sinhala language. Areej Al-Hassan and Hmood Al- Dossari also made survey on hate speech detection based on multilingual corpus to solve the complexity of Arabic text and to automatically detect hate speech written by Arabic text.

In our country Ethiopia, a lot of researches have done based on automatic hate speech detection by using machine learning approaches. In ASTU, Yonas Kenenisa Defar also made experimental approach to determine the best combination of the machine learning algorithm and features extraction for models to identify hateful speech from none hateful. SVM, NB, and RF models trained using the whole dataset with the extracted feature based on word unigram, bigram,and trigram, combined n-grams, TF-IDF, and combined n-grams weighted by TF-IDF and word2vec for both datasets. The models based on SVM with word2vec achieve slightly better performance than the NB and RF models for both binary and ternary models.

4 Research Question

- What are preprocessing techniques need to be applied to prepare Quality Afan Oromo hate speech data set?
- What are appropriate software tools for data collection from Facebook and Twitter?
- What are appropriate feature extraction techniques need to be applied to obtain important features from Afan Oromohate speech data?

- What is a framework to develop hate speech detection model for Afan Oromo social media?
- Which machine learning algorithms is the most performer to build Afan Oromo hate speech detection model?

5 Objective

5.1 General objective

The general objective is to detect Afaan Oromo hate speech from social media by using supervised machine-learning model for hate speech classification.

5.2 Specific objective

- To review various literature that are undertaking optimal parameters in hate speech detection.
- Develop a corpus for hate speech.
- To design the model suitable for machine learning architecture approach and algorithms.
- To evaluate the performance of the proposed system by using a test dataset.

6 Scope of the study

Hate speech detection is very complex and a huge task that demands a lot of resources and effort. Since there is no formal way to identify offensive language, hate speech or nether of them. The main goal of this study is detecting Afaan Oromo hate speech comment and post on Facebook specifically for Afaan Oromo language only. We select Facebook, comparative studies have shown how in countries with limited Internet penetration, like Ethiopia, Facebook has become almost a Synonym for the Internet, a platform through which users access information, services, and participate in online communications [2]. That means we prepare our data set by using Facebook because of most of hate speech that are posted on Facebook by Afaan Oromo language take a lot of place.

7 Methodology

7.1 Data collection and preparation

In order to classifying the hate level across Facebook for Afaan Oromo language users, we have built a corpus of comments retrieved from Facebook public pages. To prepare that corpus first wetake

retrieve the content of the comments from Facebook posts using Face pager. Facebook is selected to collect data from social media for the following reasons. Facebook is the most important platform for reaching out to online audiences, and especially the youth [3]. After we retrieved the data from the Facebook, we apply the following rules.

- ✓ By using pandas data frame structures and manipulation tools we make data clear and
- ✓ Only kept comments that were in Afaan Oromo and all punctuations were removed by passing to the python translate function.
- ✓ In addition, we prepare our data by formatting pandas. Panda is a software library writtenfor the Python programming language for data manipulation and analysis [11].

7.2 Annotation preparation

To be able to perform experiments on hate speech detection, access to labelled corpora is essential. Despite the differences between the previous studies that analyzed in the related work, the majorityof the described works present instructions for the annotation task. Those annotations are focus on the types of hate speech that are posted through social media and that affect the relationship of citizens. We prepare the annotations in to three parts. Those are Politics, ethnicity and religion. Based on those annotations we categorize the social media posts in to two classes that is HATE and NO HATE (normal). The following diagram shows how we classify posts in to either it is hate or not.

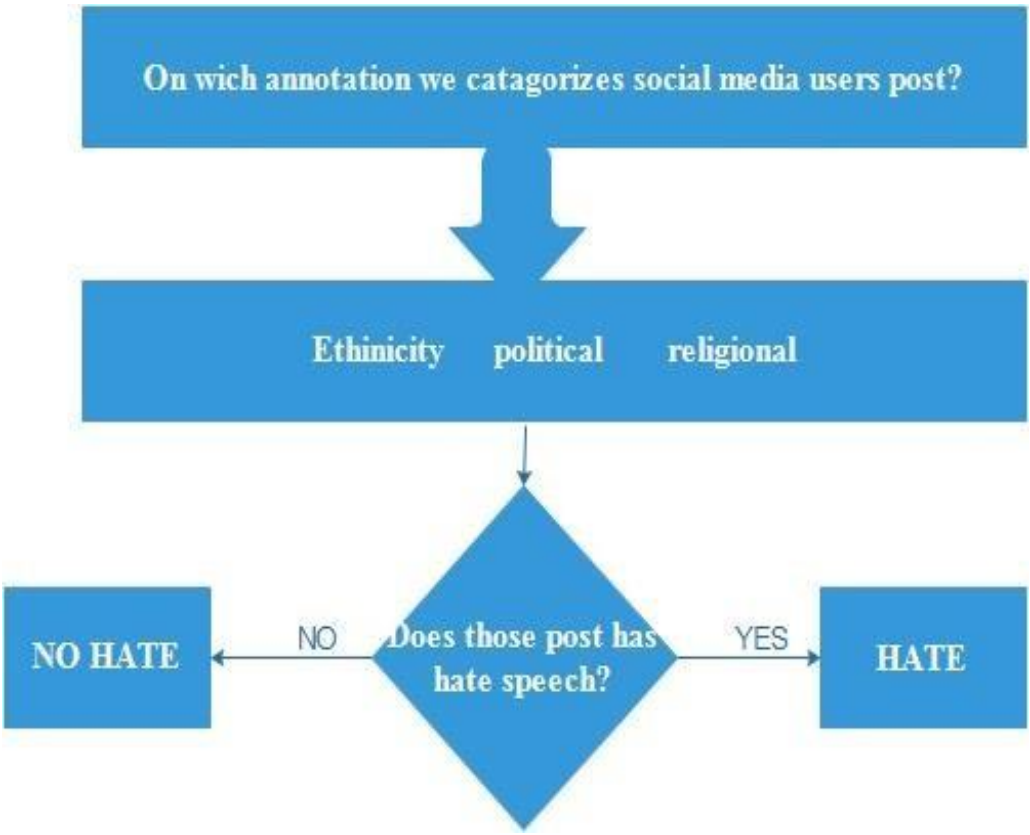

Figure 1: Annotation preparation

Table 1: Sample Annotated Afan Oromo Text document

Sno	Content/qabiyyee	class/garee
1	Oromo is enemy of Ethiopia "Oromoon diina Itiyoophiyaati"	hate "jibba"
2	Selfish"Abbaa garaa"	hate "jibba"
3	struggle you contributed is unforgettable "qabsoon ati giite hin dagatamu	normal "fayyaalessa"

The meaning of the text document in Oromo is enemy of Ethiopia "Oromoon diina Itiyoophiyaati" is against Oromo Ethnicgroup and has to labelled as **hate "jibba"**. Like that selfish "abbaa garaa" is insulting somebody, therefore annotators annotateit as **hate "jibba"**. In oromo culture, selfish "abbaa garaa" anybody who is not worrying about others even for his/her brother ifthey achieve what they want. The contents of "struggle you contributed is unforgettable "qabsoon ati giite hin dagatamu" freeand it has to be labelled as normal "fayyaalessa" by annotators.

7.3 Feature selection for hate speech detection

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction .Since hate speech detection system for Afaan Oromo is based on classification, It requires features used for classification. There are many simple feature selection methods. The first ones is Surface-level or bag of word feature with n-gram models are used to predict words. The second one is character level n-gram feature for spelling correction in which most comments can have spelling mistakes. The third one is word generalization to carry out word clustering by which it cluster hateful and none hateful words by using brown clustering algorithm [1].The fourth feature selection methodsare embedding such as word and paragraph embedding which are based on neural network and used to class similar word, paragraphs in similar vector respectively. For hate speech detection paragraph-embedding is better one since most hate speeches written by paragraph form. In our proposal we used combination of paragraph-embedding with n-gram model and Tf-idf and countvectorizer for better classificationand word prediction.

7.4 Model selection

The main feature of interest for this work is comments and posts in one as post of users and the output variable is label towards hate speech in social media. The classification is supervised learning task because, the objective is to use machinelearning to automatically classify comments/posts into categories based on previously labelled comments and posts [2]. In supervised learning, there are many classifier algorithms. I applied different algorithm such as Support Vector Classifier, GuissianNB, kNeighborsClassifier, and adaBoostClassifier, BaggingClassifier, GradientBoostingClassifier, ExtraTreesClassifier, XGBClassifier and Random Forest Classifier. Random Forest Classifier are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. The Random Forest Classifier algorithm is a popular machine learning tool that offers solutions for both classification and regression problems. I select Random Forest Classifier because it perform the highest accuracy-score on classification of hate speech is when we compare to the other.

7.5 Software Tools

The Afaan Oromo hate speech detection will implemented by software specification python 3.10 with full nltk (natural language toolkit library) for natural language processing purpose, Pandas library for data preparation and analysis, Jupiyter note book for code implementation, Face pager for retrieving data from Facebook page, Numpy library for some mathematical operation and plotlibrary for system analysis purpose .

8 Programming Language and Tool Used

As indicated in the previous section, we collected data from channels of BBC Afan Oromo, OBN Afan Oromo, Fana Afan Oromo Program, Politicians, Activists, Religious Men, and Oromia Communication Bureau. Face pager is a tool that retrieves data from Facebook and Twitter pages and saves retrieved data in csv format on a local machine. Face pager: In this study, we used face pager to collect posts and comments from Facebook and Twitters Pages. MySQL database: MySQL database server used to develop the Afan Oromo hate speech dataset annotator system alongside using php.Python programming language: Python programming language is a powerful programming language currently used in various disciplines.

In this particular work, python programming language is used for data preprocessing, dataset splitting andmodel development.

9 Natural Language Processing Tasks in Hate Speech Detection

In this work, we divided hate speech identification tasks into subtasks that were finally combined to support proposed hate speech detection and classification. Each subtask purposely designed to aim of handling all about hate speech. We divided hatespeech detection tasks into five tasks. Task 1. Hate speech identification tasks: At hate speech identification tasks, researchers identified whether the given posts and comments are either hating speech or normal speech. Annotations of data carried out based on the specific labels. Task 2. Automatic Hate speech detection tasks: the aim of task in step is to check whether the posts and comments are hating speech or normal based on the label in task 1. Task 3. Automatic Hate speech classification task: in this step the target of the hate speech identified. Task 4. Identifying target: the class identified in tasks 3, the aim of target identification is identifying the target of hate speech based on labeled posts and comments. Task 5, Target of Speech Identification: Analyzing the contents of the text either text content is hating or normal is essential. Therefore, target of speechidentification level, the researcher analyzed content, text document content with the help of experts.

9.1 **Document preprocessing**

Combine collected data into one File: Data collected from various pages of Facebook and Twitter pages. To make collected data ready for text preprocessing, data annotation and data splitting, researchers merged all collected data into single file name with excel file extension “hate1 dataset.xls”.

Spell correction: most people type Afan Oromo words with the correct spelling, whereas few people type Afaan Oromo words incorrectly. An Afan Oromo word with incorrect spelling changes the meaning of a sentence, paragraph and entire document written in Afan Oromo. To overcome challenges of the misspelled word in Afan Oromo, we identify words with misspelling and try to replace them with correctly spelled words by writing using python scripting. Removing Irrelevant Contents: The text preprocessing tasks are essential to achieve relevant dataset. In this step, we identified punctuation marks, special symbols, emoji, number, URL and stop words thoroughly. As indicated in the work by irrelevant data has to be eliminated at the text preprocessing phase. To clean punctuation marks, special symbols, emoji, number,URL and stop words, first the plain texts are

tokenized into tokens by tokenization process. A second list of stop words in Afan Oromo Languages, punctuation mark, special symbols, html, and URL removed from the data. Finally, researchers wrote python scripts to carry out text preprocessing tasks and removed stop words, punctuation mark, special symbols, convert upper case to lower case, html, and URL. Among all punctuation marks, pseudo code internationally prepared to remove all function marks except, apostrophe “hudhaa” “ ’ ’ that helped for word formation in Afan Oromo.

Start:

1. Open the file
2. Read text in dataset;
3. While (! end of text in dataset):
   If the text contains symbol [= <> << >> +! ~] then Remove symbol and add space
   If text contain special_char [,!#$@%^*] then Remove special_char
   If text contains number= [0 9] then Remove number
   If text contains emoji= [EMOJI] thenRemove emoji
   If text contains white space, thenTrim text
4. Return corpus;

END:

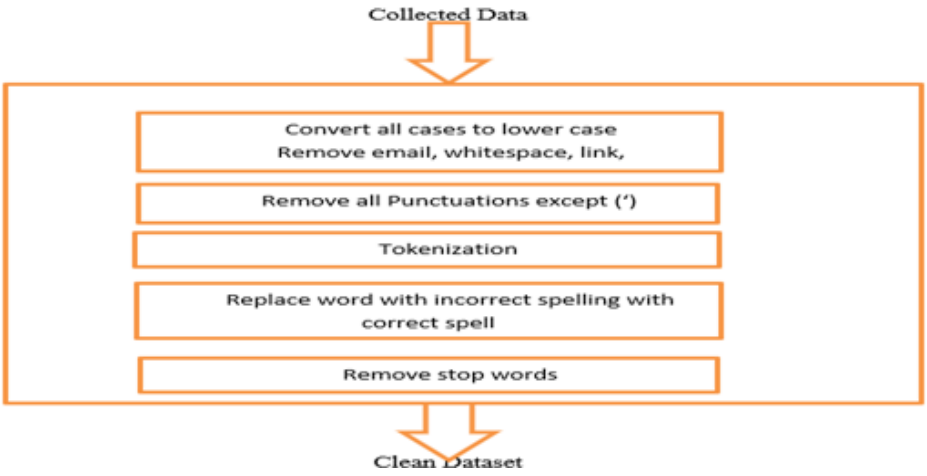

Figure 2: Afan Oromo Text document preprocessing Architecture

10 Afan Oromo Hate Speech Detection

Machine learning algorithms applied to develop Afan Oromo hate speech detection model. The machine learning algorithms, particularly, supervised machine learning require properly annotated dataset to obtain models with highest performance. We annotated a dataset for Afan Oromo hate speech detection depending on the annotation procedure prepared.

10.1 Machine learning algorithms

Several activities such as text classification, text categorization, pattern recognition, pattern discovery, decision making and the like, those that need human intelligence are automating by Machine learning. Machine learning is a branch of Artificial Intelligence, which is categorized into supervised, unsupervised. In the machine learning approach for predefined classes, a document that will be classified manually by the user always exists. Therefore, the predefined data sets are used for automatically learning the meaning that the user assigned attributes to the classes due to the existence of available data. It contains two main learning approaches: unsupervised learning and unsupervised learning approaches. Supervised learning approach needs predefined class and deals with classification techniques; whereas unsupervised learning approach does not predefined data and deals with clustering techniques. Supervised machine learning approach requires human involvement partially for a labelling class of data, to divide a dataset into train and test dataset. Decision tree, support Vector machine and Naïve Bayes are the most known supervised machine learning algorithms. As we understand from literature review, currently, supervised machine Learning algorithms are also utilized for hate speech detection and classification. In our work, we also used machine learning algorithms listed under for conducting experiments then compared their performance.

11 Evaluation System

11.1 Dataset Labelling Evaluation System

Researchers strategically identified the classes of a hate speech detection dataset into hate and normal. Afan Oromo hates speech detection dataset classes become the name of two radio buttons for row data displayed from the database that holds hate speech detection dataset which was created in the MySQL database we used as back-end software for evaluating Afan Oromo hate

speech detection system. Depending on the numbers of experts assigned either hate or normal label the systemselects classes and assigns them to each in the database. To study, since we used five experts to annotate data, the class of threeor more than three experts will assign as a class to the data.

11.2 Performance evaluation parameters

In machine learning techniques, accuracy, precision, recall and f-measure are used as the main performance evaluation techniques. Among those performance evaluation parameters, f-measure is the average of Precision and recall. Therefore, in this research work, f-measure/score was used to evaluate the performance of Afan Oromo hate speech detection system. Among those performance evaluation parameters, f-measure is the average of precision and recall. In the confusion matrix, the performance of each machine learning algorithm is evaluated using comparatives of Accuracy, Recall, Precision and F-score. Inour study, we evaluated the performance of algorithms using the accuracy-score only. Therefore, in this research work, f accuracy-score was used to evaluate the performance of Afan Oromo hate speech detection system.

12 Result and Discussion

12.1 Results

Afan Oromo hate speech detection data collected from Facebook and Twitter social media platforms using Face pager. Thesystem we developed using php and MySQL database assigned labels for the loaded data into the database. Generating accounts for experts of the developed system able to annotate the dataset. From Annotated Afan Oromo hate speech dataset, train and test data set obtained after the annotated dataset divided into using python programming language. The important feature selected from the prepared dataset helped to result in a Benchmark Afan Oromo hate speech dataset that contains the train and test set.We conducted experiments by loading machine learning algorithms turn by turn on the dataset and the performance of each applied algorithm. The Developed Afan Oromo hate speech detection was able to be tested with the test dataset accuracy scored performance of 76.9 %.

12.2 Discussion

The study is centered on developing hate speech detection models for Afan Oromo social media platforms, specifically fromFacebook and Twitter. For successful development of the proposed model, we performed a series of activities. First, data collected from selected sources and annotated according to prepared procedures. Then, text preprocessing applied on gathereddata to select relevant data and remove irrelevant data. At text preprocessing phase, Afan Oromo stop words, punctuations except', numbers, all none Afan Oromo text document, row with empty space, image, video, audio, link, emoji and email removed. All typos errors tried to replace by word correct spelling. We also applied data normalization. Next to those, feature selection techniques such Bigram and TF-IDF applied for data vectorization. On vector zed Afan Oromo hate speech data, supervised machine learning algorithms were applied. To conduct the experiment, we used Support Vector Classifier, GuissianNB, KNeighborsClassifier, AdaBoostClassifier, BaggingClassifier, GradientBoostingClassifier, ExtraTreesClassifier, XGBClassifier and Random Forest Classifier. From all machine learning algorithms applied to build models, Random Forest Classifier achieved higher accuracy than others and the Random Forest Classifier selected as the highest performer.

Finally, we also tested the performance of developed Afan Oromo hate speech detection using a test data set and model accuracy-scored 77% .Since, the Random Forest Classifier shows good results to detect hate contents on both training and testing depicts that Random Forest Classifier has trained from training data and can also apply the knowledge to new text document withunknown class. Finally, Afan Oromo hate speech text model from Afan Oromo posts and comments can identify hate speech contentsby training by training using dataset collected from Facebook and Twitter in Afan Oromo. This model can be challenged by detecting and alerting the hate contents from Facebook and Twitter. The output of this developed Afan Oromo hate speech detection model can overcome the problems that may the country face due to hate speech if properly implemented by the Ministry of peace in Ethiopia and social media companies.

13 Related Work

The current literature shown that understanding and analyzing social media become a main concern. Today, one of the mainconcerns about social media is positive and negative impacts that comments and posts in social media platform have either onindividual, groups or society. Hence, sentiment analysis, hate speech detection and classification, abusive language detection and/or classification, offensive language detection and/or classification and cyberbullying detection and/or classification become topic of research interest for researchers, Government and Social Media Company. Hate speech detection techniquesthat used to identify content is displayed on social media platform in the form of comments or posts irrespective of its nature whether the content is hate or normal. It used approaches such as machine learning, natural language processing, statistics and the like to design a model that detects hate speech. Using hate speech detection, natural language processing tasks and machine learning algorithms, comments and posts in social media platforms like Facebook, YouTube and Twitter can analyze and identify either as it is hate or normal. In this section, we review related work from the perspective of Machine Learning algorithms, hate speech detection and classification, sentiment analysis, and natural language processing.

Reference language	Feature extraction Techniques	Social media	Algorithm	Dataset	Availability	F1-score
[6]Danish and English	-	Facebook Reddit and Twitter	-	Original	No	74%
[7]Indonesian	Textual acoustic and their combination	Facebook line today YouTube	LSTM	Original	No	70.4%
[8]Arabic	Bow,TF and TF-IDF	Twitter	SVM ,NB,RF and DT	Original	No	91.2%

Table 2: A brief summary of the related work

14 Conclusion

We have outlined that developing hate speech detection for Afan Oromo social media is essential to eradicate the risk of hatespeech on social welfare. Our work has led to the conclusion that machine learning is applicable for the development of hate speech detection models for Afan Oromo on Facebook and Twitter. We conducted experiments by applying machine learning algorithms such as Support Vector Classifier, GuissianNB, KNeighborsClassifier, AdaBoostClassifier, BaggingClassifier, GradientBoostingClassifier, ExtraTreesClassifier, XGBClassifier and Random Forest Classifier to build hate speech detection prototypes for Facebook and Twitter. To evaluate the performance of each algorithm, researchers used performancemetrics Accuracy. The feature selection techniques for machine learning, bigram and TF-IDF applied. The result of the study indicated that Support Vector Classifiers Performance accuracy 69%.The GuissianNB achieved performance accuracy of 54.9%. The Random forest classifier achieved performance accuracy 76.9%. The KNeighborsClassifier achieved performance accuracy 62.8%. The AdaBoostClassifier achieved performance accuracy 74.2%. The BaggingClassifier achieved performance accuracy 75%. The GradientBoostingClassifier achieved performance accuracy 70.5%. The ExtraTreesClassifier achieved performance accuracy 75.7%. XGBClassifier achieved performance accuracy 70.5%.The result of the experiment shows that the performance of Random forest classifier accuracy-score value is 76.9% and we have confirmed that Random forest classifier scored highest performance compared with others. Therefore, the researchers agreed to use Random forest classifier to deploy Afan Oromo hate speech detection model. Even though we have developed the Afan Oromo hate speech detection model using machine learning algorithms by collecting data from Facebook and Twitter, this study only investigated posts and comments in text documents. The posts and comments in mode of image/photo, audio and video data have not been considered. The most important limitation of this study also lies in applying conventional machine learning algorithms that need manual labelling of dataset. In this study, experiments conducted on data were of small in size.

15 Future work

In future study can also be conducted by collecting datafrom other Social Media platforms. In addition to collecting data from other social media platforms, the researchers can considerother modes of data for further research to be investigated. Applying beyond conventional machine learning algorithms for experiments can also be the next study we try to develop the hate speech detection for afan Oromo on social media such as twitter and Facebook, so will recommend to further study for tringual Hate speech detection on social media for Afaan Oromo, Amharic and Tigrigna Languages.

Reference

[1] A. Schmidt, D.- Saarbrücken, and D.- Saarbrücken, “A Survey on Hate Speech Detection using Natural Language Processing,” no. 2012, pp. 1–10, 2017.

[2] Z. Mossie and J. Wang, “S OCIAL N ETWORK H ATE S PEECH,” pp. 41–55, 2018.

[3] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, “Effective hate-speech detection in Twitter data using recurrent neural networks,” 2018.

[4] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” *15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf.*, vol. 2, no. 2, pp. 427–431, 2017, doi: 10.18653/v1/e17-2068.

[5] B. Gambäck and U. K. Sikdar, “Using Convolutional Neural Networks to Classify Hate-Speech,” no. 7491, pp. 85–90, 2017.

[6] G. I. Sigurbergsson and L. Derczynski, “Offensive language and hate speech detection for danish,” *Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc.*, no. May, pp. 3498–3508, 2020.

[7] G. B. Herwanto, A. M. Ningtyas, I. G. Mujiyatna, K. E. Nugraha, and I. N. Prayana Trisna, “Hate Speech Detection in Indonesian Twitter using Contextual Embedding Approach,” *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 2, p. 177, 2021, doi: 10.22146/ijccs.64916.

[8] I. Aljarah *et al.*, “Intelligent detection of hate speech in Arabic social network: A machine learning approach,” *J. Inf. Sci.*, vol. 47, no. 4, pp. 483–501, 2021, doi: 10.1177/0165551520917651.