Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6612–6619
Marseille, 20-25 June 2022

Afaan Oromo Hate Speech Detection and Classification on Social Media

Teshome Mulugeta Ababu, Michael Melese Woldeyohannis
School of Computing, Dire Dawa University Institute of Technology, Dire Dawa, Ethiopia
School of Information Science, Addis Ababa University, Addis Ababa, Ethiopia
teshome.mulugeta@ddu.edu.et, michael.melese@aau.edu.et

Abstract

Hate and offensive speech on social media is targeted to attack an individual or group of community based on protected characteristics such as gender, ethnicity, and religion. Hate and offensive speech on social media is a global problem that suffers the community especially, for an under-resourced language like Afaan Oromo language. One of the most widely spoken Cushitic language families is Afaan Oromo. Our objective is to develop and test a model used to detect and classify Afaan Oromo hate speech on social media. We developed numerous models that were used to detect and classify Afaan Oromo hate speech on social media by using different machine learning algorithms (classical, ensemble, and deep learning) with the combination of different feature extraction techniques such as BOW, TF-IDF, word2vec, and Keras Embedding layers. To perform the task, we required Afaan Oromo datasets, but the datasets were unavailable. By concentrating on four thematic areas of hate speech, such as gender, religion, race, and offensive speech, we were able to collect a total of 12,812 posts and comments from Facebook. BiLSTM with pre-trained word2vec feature extraction is an outperformed algorithm that achieves better accuracy of 0.84 and 0.88 for eight classes and two classes, respectively.

Keywords: Afaan Oromo, Hate and Offensive Speech, Under-resourced, Machine Learning, Deep Learning

1. Introduction

Today, Natural language processing (NLP) and Machine learning (ML) are over simplify different natural language applications such as speech recognition, sentiment analysis, hate speech detection, and other related NLP applications. According to the 2020 World Bank statistics report[1], Ethiopia is one of the second largest populated countries in Africa next to Nigeria, which has a more than 114 million population. Oromo ethnic group spoke Afaan Oromo as a mother tongue language that is written using Latin script called Qubee Afaan Oromoo (Degeneh Bijiga, 2015). Ethiopia is a multilingual and multi-ethnic country. In Ethiopia, Oromo is the largest ethnic group that contributes more than 37 million of the Ethiopian population (Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.), 2021).

Afaan Oromo is spoken in Ethiopia special in Oromia regional states and neighboring African countries like Kenya, Tanzania, Sudan, Somalia (Degeneh Bijiga, 2015). Afaan Oromo is the fourth most widely spoken language in Africa next to Arabic, Hausa, and Swahili (Degeneh Bijiga, 2015). Social media is an attractive computer-mediated technology that is used to facilitate communication and sharing of data interactively and effectively for more than one person at a time (FDRE, 2020). Today, people can communicate easily and quickly with each other by using social media like Facebook, YouTube, Twitter, and Snap chat (Alrehili, 2019).

In the world, there are around 3.96 billion social media users, from this 21.14 million users are from Ethiopia this shows the rapid development of social media networks and their users across the world. Especially, Facebook is the most widely used social media platform in the world and Ethiopia too (Tesfaye and Kakeba, 2020). In Ethiopia, Afaan Oromo is spoken by the largest ethnic group Oromo as their first language and it is also spoken by other Ethiopian ethnic groups as a first or second language. In Ethiopia, this largest ethnic group uses Afaan Oromo in social media to express their opinion or feeling regarding to socio-economic and political aspects on social media. In this procedure there is clash of individual or group ideas, agreement and disagreement. There is no problem with agreement but there is the problem whenever there is disagreement, because they are use a tone of hate speech against each other. But there is rare Afaan Oromo hate speech detection and classification study. Hate speech is a speech that deliberately promotes hatred, discrimination, or attack against a person or a discernable group of identity, based on ethnicity, religion, race, gender, or disability (FDRE, 2020) (Alrehili, 2019).

The explosive growth in hate speech and its erosion of democracy, humanity, justice and peace building increases the demand to develop automatic hate speech detection on social media. Technology intervention of hate speech detection is necessary because massive users generated data management on social media is difficult to protect hate and offensive speech (Yonas, 2019). Social Media Company like Facebook hires thousands of employers for content moderator and pays millions of dollars per years for content moderator this is an expensive and need human labour (Facebook, 2018).

The spread of hate speech on social media promotes the content that violence against individuals or groups based on protected identity such as gender, race, religion, disability, age, and veteran status (Abro et al., 2020). Offensive speech can be defined as a language or expression that offend someone, this kids of speech cause someone to feel upset, annoyed, insulting, angry, hurt, and disgusting (Yonas, 2019). There is no formal definition that differentiates offensive speech from hate speech, it is based on subtle linguistic distinction (Davidson et al., 2017).

According to (Teh and Cheng, 2020), offensive speech is also called bad language, the wrong choice of words, expletives, curse words, swear words, vulgar language, profanity

[1] https://data.worldbank.org/indicator/SP.POP.TOTL?locations=ET

and is used to reflect behavior that is offensive or lacking in respect to others. The content of hateful speech spread over social media is disturbing the normal life of society by raising violations of human rights; motivating genocide, massacre, ethnic cleansing, and civil war. Facebook and Twitter is criticized for not well-combating hate and offensive speech in their platform(Yonas, 2019). Hate speech is drastically increased on social media because it connects billions of users across the world, those who have different cultural backgrounds, norms, values, and beliefs (Alrehili, 2019).

In this study, we focused on the Hate speech concept given by Facebook social media platforms and the Ethiopian government Proclamation (Facebook, 2021)(FDRE, 2020). According to the Ethiopian house of people representative, hate speech and disinformation prevention, and suppression define hate speech as the speech that deliberately promotes hatred, discrimination, the attack against a person or group of identity-based based on gender, race or ethnicity, religion and disability (FDRE, 2020). Similarly, Facebook describes hate speech as a direct attack against people based on what we call protected characteristics such as race, national origin, ethnicity, sexual orientation, religious affiliation, sex, serious disease or disability, gender, gender identity, and caste (Facebook, 2021).

Thus, there is a need to overcome the above stated problem of Afaan Oromo hate and offensive speech on social media in automated way using different machine learning techniques.

Accordingly, the rest of the paper is organized as follows. In section 2, we presents a brief discussion of related work attempted in the hate and offensive speech detection. Section 3 presents overview of Afaan Oromo Language. Section 4 explains the challenge of hate and offensive speech detections. Section 5 presents the data collection and preparation procedure to achieve the goal of this study. Section 6 presents the architecture of proposed model. Similarly, Section 7 presents the experiment results and discussion. Finally, In Section 8 we Presents about conclusion for this study and recommendation for further study.

2. Related Work

Several scholars have studied hate speech detection and classification for foreign languages like Bangla (Ahammed et al., 2019) , Arabic (Alsafari et al., 2020),Turkish (Polat et al., 2018) English (Alrehili, 2019), (Abro et al., 2020) and local language like the Amharic language (Yonas, 2019), (Mossie and Wang, 2018), (Mossie and Wang, 2020) but we found a few studies for Afaan Oromo Language (Defersha and Tune, 2021) (Kanessa and Tulu, 2021).

According to (Abro et al., 2020), the author developed a model used to detect hate speech and conducts experiments using three feature extraction techniques (bigram, word2vec, doc2vec), and they are compared to eight machine learning algorithms. The author collected 4,509 datasets in total; annotate the dataset into three classes (hate speech, not offensive, offensive but not hate speech). Finally, the researcher achieved 79 % accuracy by using the combination of SVM with bigram feature extraction. But here the researcher provides three class labels of hate speech which is inadequate classification. According to Zewdie Mossie and Jenq-Haur Wang's (Mossie and Wang, 2018) study, they have developed an apache-spark-based model to classify Amharic Facebook comments and posts into not hate and hate classes. The author used a total of 6,120 datasets from this around 80 % (4,882) for training and 20% (1,238) for testing the model; they experimented by using Navies Bayes, ML algorithm with word2vec feature extraction, and Random forest ML algorithm with TF-IDF feature extraction and finally achieved 79.83 % and 65.34 % model accuracy. The author recommends expanding datasets and multi-class classification. Based on the author's recommendation we expanded the dataset and providing a multi-classification of hate speeches.

Author (Tesfaye and Kakeba, 2020), proposed an automated Amharic hate speech post and comment detection model by RNN. The author used 30,000 datasets from this 80 % used for training, 10% validation, and 10% testing. The author used two deep learning algorithms (LSTM, GRU) with word n-gram and word2vec feature extraction. Finally, the author achieved the highest 97.9 % of accuracy by LSTM algorithm with word2vec feature extraction techniques. But author provides with binary classification classes this is an inadequate classification. Additionally, the author has used only a limited Algorithm (LSTM and GRU).

Recently, we found one paper on Afaan Oromo hate speech detection and classification proposed by the author (Defersha and Tune, 2021), the author collect 13,600 datasets from Facebook and Twitter social media platforms. The researcher has used classical and ensemble machine learning algorithms with N-gram and TF-IDF feature extraction techniques. The researcher split the data set into 67:33 ratios. Finally, the researcher achieved a 64 % F1 score by a linear support vector classifier. However, the author attempted for binary classification and also the dataset used for experiment were not accessible.

In addition, other attempts have been made toward Afaan Oromo hate speech detection model using a machine learning algorithm by Author (Kanessa and Tulu, 2021). The researcher proposed Afaan Oromo Hate speech detection framework for Afaan Oromo language. The attempts were made by collecting a total of 6,120 dataset from Facebook, and the researcher used three different feature extraction techniques such as TF-DF, word2vec and N-gram feature extraction techniques. Finally, the researcher achieved 96% of accuracy by combining of SVM with ngram, and TF-IDF. The researcher, on the other hand, is only able to identify hate speech without any attempt to classify the hate type.

In our research, we adopted several techniques and approaches used by other scholars. Most of the research done on hate speech detection is done by selecting a single method learning approach of machine learning method without comparing it with other methods learning approaches of machine learning but our method is compared with the different machine learning approaches (classical, ensemble and deep learning). Additionally, most research is focused on hate speech detection rather than identifying the class of hate speech.

3. Afaan Oromo Language

Afaan Oromo is the widely spoken language in the Horn of Africa (Tegegne, 2016). Oromo people Spoke Afaan Oromo, which is their native language, it is the most widely spoken Cushitic family of Afro-Asiatic language, Afaan Oromo is one of the under-resourced African language widely spoken in Ethiopia and neighbor countries like Kenya, Somali, Tanzania, and Sudan (Degeneh Bijiga, 2015). Afaan Oromo is the most extensively spoken and utilized language in Ethiopia, with the biggest number of speakers (Tegegne, 2016).

The first Oromo newspaper, *Bariisaa* is published in 1975 and nationalized in 1976 by the Ministry of Information and National Guidance, and between the 1970s to 1980s most of Oromo material is written in Gee'ez script (Tegegne, 2016). However, on November 3, 1991, OLF and other Oromo scholars call a national meeting for Oromo speakers in heat of Oromia to discuss and decide the destiny of their language, after the hour of discussion and deliberation, over 1000 men and women have attended the meeting, unanimously decide Afaan Oromo writing system is based on Latin Script with modification (Degeneh Bijiga, 2015).

3.1. Afaan Oromo Writing System

In 1991 Qubee was adopted as an official and formal orthography for Afaan Oromo writing script (Degeneh Bijiga, 2015). Qubee orthographic writing system has a total have 33 alphabets with 3 classifications; consonant, compound consonant and vowels. Table 1 presents the three classes of Afaan Oromo language alphabets.

Types	***Alphabets***	***Total***
Consonants	b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, w, x, y, z	21
Compound consonants	dh, sh, ny, ph, ch, ts, zh	7
Vowels	a, e, i, o, u	5
	Total	***33***

Table 1: Afaan Oromo Alphabets

The first class is five vowels in Afaan Oromo is similarly available in the English language (a, e, i, o, u). the second class has 7 double consonants (dh, sh, ny, ph, ch, ts, zh). the third class has 21 single consonants likewise the English language. Apostrophe mark (') in Afaan Oromo is considered as a unique consonant in Afaan Oromo writing system also called *Hudhaa* in Afaan Oromo. In Afaan Oromo all single consonants can be germination except h and all double consonants (Degeneh Bijiga, 2015).

4. Challenge of Hate and Offensive Speech Detection

The significant challenge in hate speech detection on a social network is the separation of hate speech from other offensive speech (Davidson et al., 2017). The border between offensive speech and hate speech is somewhat blurry. In hate speech detection different languages have different difficulties based on the nature of language. In social media, people are sometimes writing whatever they went carelessly without keeping language rules (spelling, grammar and punctuation mark) this is the great challenge next to the definition given to hate and offensive speech on social media (Bawoke, 2020).

One of the main challenges of hate speech detection is data annotation since there is no defined rule for labeling hate speech text and how many classes it can be classified (Bawoke, 2020). The other challenge in hate speech detection is social media ambiguity, a language developed among young people, and the culture of society (Alrehili, 2019). For a local language like Afaan Oromo, even though there are some attempts made towards hate speech detection, in one way, the data collected are not in accordance with this research work and their datasets are not published for hate speech detection. So, there is a need for collection of the dataset from Social media for Afaan Oromo hate and offensive speech detection. Similarly, hate speech detection may have complicated factors such as Sarcasm, intentionally slang/misspelling words, socio-political context, idioms, and slurs.

5. Data Collection and Preparation

Due to unavailability of Afaan Oromo hate and offensive speech detection dataset, we opted to collect and prepare a data from social media in the four thematic areas. In the data collection process firstly, we have searched the data on Facebook by using the respective thematic areas keywords. The keywords is selected by the domain expert based on the four thematic areas such as gender class (related), religion class (related), race class (related), and offensive class (related).

The compiled keywords have contained both hate and free lexicons. This is followed by checking the relevance of the dataset. If the data is related to our thematic area then check the account has a minimum of five hundred followers or members. If the data is achieved by both criteria we are scraping the posts or comments by using Facepager and/or Data minor tools.

Once the relevant data is collected, it is broken down to the posts or comments at a sentence level as some posts and comments seem documents so they need to break down into lower levels. Consequently, the data is further preprocessed by re-correcting misspelled words in our sentences using the language expert. Normally, In social media, some people have misspelled words intentionally or unintentionally. In Afaan Oromo research such as sentiment analysis, many scholars are re-correcting or removing the misspelled word (Defersha and Tune, 2021), (Rase, 2020). We have followed the same procedure as other researchers. Additionally, annotating the dataset based on guidelines. After data annotation we are calculate inter annotators agreement between annotators by Fleiss Kappa. Furthermore, merging each individually annotated thematic area into a single file.

5.1. Dataset Description

The data is collected from the Facebook account, which is most frequently posts and comments in Afaan Oromo languages. Similarly, the collected dataset is from different broadcasting media services, personal accounts, figurative people, journalist, activist, and another minor account

which has a minimum of 500 followers on Facebook.
To scrape the data from any Facebook account the researcher selects the account/page which has a minimum of 500 followers or group members because of three reasons, the first it is prove the validity of the dataset (our data is a valid dataset), and the second reason is the account needs more attention because the posts published by those accounts is received by many social media users, and the third reason is Ethiopian constitutional orientation on (Proclamation) concern on any media, person, or organization which have a high number of followers.
Ethiopian constitutional Proclamation state that if the user or media having more numbers of followers and transmitted hate speeches it is considerable for imprisonment or fine penalty as stated on Ethiopian hate speech and disinformation proclamation (FDRE, 2020). For this study, the researcher has collected a total of 12,812 posts and comments from Facebook. We have collected the datasets from the 20th, of January 2021 to the 20th of May, 2021.

5.2. Dataset Annotation Guidelines

So in our study, we prepare the general guideine to ignore ambiguity during annotating a dataset. We define and classify hate speech as shown below from the perspective of Law, freedom of speech, and literature review. The dataset is collected from four thematic areas (gender, religion, race, offensive) each area has hate and free speech.

- **Gender, Race and Religion related speeches:**
  If at least one of the following criteria (1-5) is achieved, the annotators classify sentence (speech) into gender-class, religion-class, race-class hate speech respectively but if the sentence talks about gender, religion, race but it does not have any criteria list below the sentence is classified as free gender, religion and race class speeches respectively.

  1. If the post/comment promotes hatred or encourage violence by discriminating a gender, religion and race.
  2. If the post/comment promotes hatred directly or indirectly discriminating by gender, religion and race.
  3. If the post/comment promotes violence or injury action is to be taken on someone discriminating by gender, religion and race.
  4. If the post/comment is mimic some to the unnecessary thing/device/animal and others to discourage psychological moral by discriminating by gender, religion and race.
  5. If the post/comment is insulting by discriminating a gender, religion and race

  Finally, the class dataset is labeled as (gender, religion and race) class hate speech. If the above criteria are fulfilled otherwise labeled as gender, religion, race class free speech respectively.

- **Offense class Speech:** If at least one of the following criteria (1-4) is achieved, the annotators can classify sentence (speech) into offensive-class hate speech else it is the offensive class free or neutral speech.

  1. If the post or comment contains a common insult without specifying the other 3 classes above.
  2. If the post or comment contains insult but it does may or may not promote to take the violent attack on individual or group.
  3. If the post or comment contains an upsetting word but the particular subject of the sentence is unknown (sentence having hidden subject/noun/pronoun).
  4. If the comment contains imprecation words.

  Finally, an offensive class dataset is labeled as offensive if the above criteria fulfilled otherwise labeled as free or neutral offensive class.

The following Table 2 illustrated Sample annotated data accepted from our dataset.

Language	Post or Comments	Classes
Afaan Oromo	Dhiirri gabaabduun farroodha irraa of eega warri Shamarranii	Gender class , Hate speech
English	The shortest boys are not good, dear girls stay away from them	
Afaan Oromo	Haati tokko altokkotti daa'imman afur oofkalan	Gender class, Free speech
English	A mother gave birth to four children at a time.	
Afaan Oromo	Yaa uummata pheenxee saroota of kabajaa.	Religion class, Hate speech
English	You protestant people, you are a dog respect yourself	
Afaan Oromo	Jiini Soomanaa ji'a hordoftoota amantaa Islaamaa hunda biratti kabaja qabuudha.	Religion class, Free speech
English	The fasting month is respected by all Muslim society.	
Afaan Oromo	Amaa**i summiidha duruu	Race class, Hate speech
English	Amh**a has always been poison	
Afaan Oromo	Qeerroo fi qarreen Oromoo dhugaa nama boonsitu.	Race class, Free speech
English	Oromo's 'Qerroo' and 'qarree' really makes us proud	
Afaan Oromo	Gantuu haadha***wu kan akka kee lammata hin dhalatin	Offensive class, Offensive.
English	Traitor, f**k your mother your like don't reborn	
Afaan Oromo	Jabaadhu Abbichuu keenna, abbaa afrikaa!!	Offensive class, Free speech
English	Be strong, our Abichu the father of Africa !!	

Table 2: Sample Annotated dataset with their respective classes

Table 2 illustrate annotated dataset accepted from the

datasets, with eight class of classification. Later for two-class classification hate classes with offensive classes are merged into one class and all other free speech is merged into the other class.

5.3. Data Annotation

As described in Table 4 our dataset is collected from four thematic areas. Every four categories of the dataset are annotated individually by three-persons who are voluntary to do the task then applying mode to the annotation of three annotators and select the annotation in which two-persons are agreed upon it. The annotators are selected purposely by the researcher based on their willingness and skill to perform the task. From our annotators, four of them are from university professors whereas the remaining are from graduate program students. All the annotators are native speaker of the language whose age ranges from twenty-four to thirty-five of male annotators. To accept the correct annotation of each post or comment we applied the following equation.

$$Label = Mode(Ex1, Ex2, Ex3) \quad (1)$$

Where: Ex is the Annotator expert of Afaan Oromo Language. Inter-Annotator Agreement is a metric that determines how well many annotators may agree on an annotation choice for a given category. Individual annotator agreements can be used to analyses qualitative and quantitative data, the inter annotators reliability statistic is used to evaluate the level of agreement between individual annotators (Sreedhara and Mocko, 2015). To measure the inter annotator's agreement between annotators different scholars are used Cohen's Kappa (Mossie and Wang, 2020) However, we used Fleiss kappa to compute inter-annotator agreement between annotators. Fleiss kappa is computed inter-annotator agreement between more than two annotators (Sreedhara and Mocko, 2015). Kappa is given by the following equation.

$$K = \frac{P - P_e}{1 - P_e} \quad (2)$$

where P is means proportion of agreement by chance, Pe is the mean proportion of agreement by analytical reasoning (Sreedhara and Mocko, 2015). Kappa varies from 0 to 1, Table 3 shows the detail Fleiss' kappa inter annotator agreement adopted from (Sreedhara and Mocko, 2015).

Fleiss' Kappa	Interpretation
<0.00	Poor agreement
0.00 to 0.20	Slight agreement
0.21 to 0.40	Fair agreement
0.41 to 0.60	Moderate agreement
0.61 to 0.80	Substantial agreement
0.81 to 1.00	Almost perfect

Table 3: Fleiss kappa's inter annotator agreement

In our study, we obtain **0.781, 0.784, 0.740, and 0.90** Fleiss Kappa inter-annotator overall agreement for Gender, Religion, Race, and offensive class respectively. The total average of Fleiss Kappa's inter-annotator agreement is **0.801**, which is a substantial class of agreement this agreement indicates that the level of inter-annotator agreement is approximately more related to each other.

5.4. Dataset Distribution

In our study, we collected data from four different thematic areas. Total for our study we have collected 12,812 posts and comments as illustrated in the Table 4 **where:**

Data categories	Sentences	Posts and comments	Classes	count
Gender	3,297	2,685 P	GCHS	1,556
		612 C	GCFS	1,741
Religion	3,027	1,453 P	RCHS	1,468
		1,574 C	RCFS	1,559
Race	3,319	2,764 P	RaCHS	1,898
		555 C	RaCFS	1,421
Offensive	3,169	3,169 C	OCOS	1,714
			OCFS	1,455
Total	12,812	P(6,902) / C(5,910)	H (6,636)	12,812
			F (6,176)	

Table 4: Dataset Distribution

GCHS: Gender Class Hates Speech
GCFS: Gender Class Free Speech
RCHS: Religion Class Hates Speech
RCFS: Religion Class Free Speech
RaCHS: Race Class Hates Speech
RaCFS: Race Class Free Speech
OCOS: Offensive Class Offensive Speech
OCFS: Offensive Class Free Speech
C: Comments
P: Posts

All three thematic areas of the dataset have both posts and comments except offensive class. This is because of the nature of datasets (it is interaction with ideas). Additionally, during the collection of our data, we have obtained enough comments that can balance with other thematic areas. Our dataset is almost balanced because we have 6,636 hate posts or comments and 6,176 free posts or comments. As part of this research work, one of the main contributions of the study is a standard and annotated Afaan Oromoo dataset for hate and offensive speech detection concerned with four different thematic areas. Additionally, We developed pre-trained word2vec which is a multipurpose model that can be useful for other Afaan Oromo natural language processing applications.

6. Architecture of the Proposed System

The proposed architecture of Afaan Oromo Hate speech (Hate speech) detection and classification is illustrated in Figure 1 As we observed from the architecture the system begins with data set collection with Face pager and data miner. The collected data is labeled with different annotators and an inter-annotator is computed. In the pre-processing phase, we removed special characters, emojis, punctuation marks, HTML tags, and stop words. After pre-processing we perform train test split of a dataset, the training data set is used to train the model whereas the test

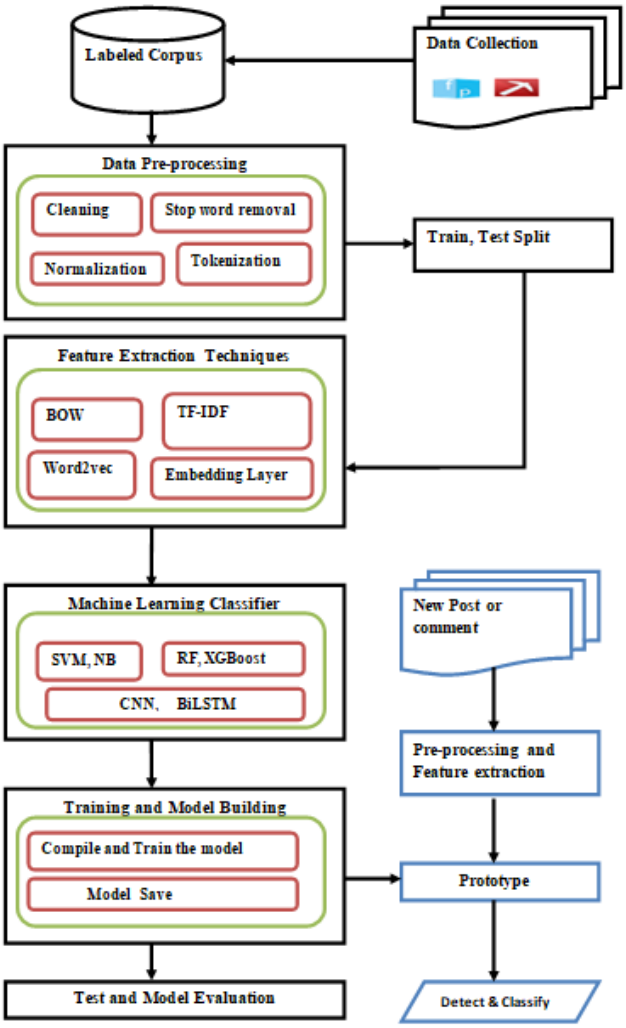

Figure 1: Afaan Oromo hate speech detection and classification architecture

data set is used to test the model performance. In addition to this, the data pre-processing activity is an essential technique used to increase the performance of the machine learning model.

In this study, we applied different data pre-processing including data cleaning, normalization, stop word removal, tokenization, Encode text to sequence, and Pad sequence are done. The last two steps (Encode text to sequence, Pad sequence) are only used with a deep learning algorithm.

Feature extraction is the process of extracting necessary features from the text that is used by a machine learning algorithm. In the feature extraction phase, we have used four types of feature extraction those are BOW, TF-IDF, word2vec, and Embedding Layers with classical, ensemble and ensemble machine learning algorithms. In Feature extraction phase, we used different feature extraction techniques such as BOW, TF-IDF, Word2vec, and Embedding Layer

- **Bag of Word (BOW):** is one of the simplest forms of text representation in numbers. The drawback of this approach is word sequence, the syntactic and semantic meaning of a sentence is ignored (Yonas, 2019).

- **Term frequency-inverse document frequency (TF-IDF):** This technique of feature extraction is measures the importance of a word in a document within a corpus and increases in proportion to the number of times that word appears in a document (Mossie and Wang, 2018).
  TF-IDF is computed by dot product TF and IDF.

$$TFIDF = FT(t,d).IDF(t,d) \quad (3)$$

  where d is documents and t is term or word in a documents

- **Word to vector (Word2vec):** is another popular text feature extraction; it is an extended vision of TF-DF. Word2vec is the valuations of word representations in vector space developed by Tomas Mikolov (Mikolov et al., 2013). Since there is no Afaan Oromo pertained word2vec model we trained our model from the scratch by using our collected corpus plus other digital Afaan Oromo textbooks (e.g. Fiction). Totally, our custom word2vec is developed with 45,592 unique words in 100 dimensions with skip-gram training algorithm. Because Skip-gram works well with a minimal amount of training data and accurately depicts even uncommon words or phrases.

- **Embedding Layer:** Keras offers an embedding layer that is frequently used for neural networks on textual data. The Embedding layer starts with random weights and learns an embedding for each word in the training dataset.

Additionally, train and model building, that can be interconnected with the developed prototype. Finally, test and evaluate the model accuracy.

7. Experiment Result and Discussion

In our first experiment, we have 8 classes of classification that is implemented with different machine learning algorithm (classical, ensemble and deep) and with different feature extraction techniques such as BOW, TF-IDF, Word2vec and embedding layer.

The experiment result of classical, ensemble and deep learning classifier is presented in Table 5

Experiment result	Accuracy in percentage			
	Feature Extraction Techniques			
Algorithms	BOW	TF-IDF	Pre-trained Word2vec	Embedding Layer
SVM	0.78	0.80	0.82	-
NB	0.80	0.80	0.74	-
RF	0.79	0.79	0.81	-
XGBoost	0.80	0.77	0.81	-
CNN	-	-	0.81	0.82
BI-LSTM	-	-	**0.84**	0.81

Table 5: Eight classes experiment result with classical, ensemble, Deep ML classifier

Firstly, from the classical classifier the highest accuracy which is 0.82 is recorded by SVM with word2vec feature

extraction. This is because word2vec feature extraction is capture more semantic and syntactic of text data than BOW and TFIDF feature extraction techniques.
However, with Naïve Bayes algorithm word2vec feature extraction techniques is achieved low accuracy, this is because Naive Bayes is fail when we have negative values, therefore it needs to be normalization but during normalization the data is missing of some features that is why it record low accuracy.
Secondly, From the ensemble classifier the highest accuracy which is 0.81 is recorded by both RF and XGboost with word2vec feature extraction. In ensemble also like classical algorithm the experiment illustrated that word2vec feature extraction is better than both BOW and TFIDF.
Thirdly, From the deep learning classifier the highest accuracy which is 0.84 is achieved by BiLSTM with pre-trained word2vec feature extraction techniques.
Graphically, the accuracy of eight classes of classification with different ML algorithms and different feature extraction techniques is summarized in the 2D column chart is presented in Figure 2

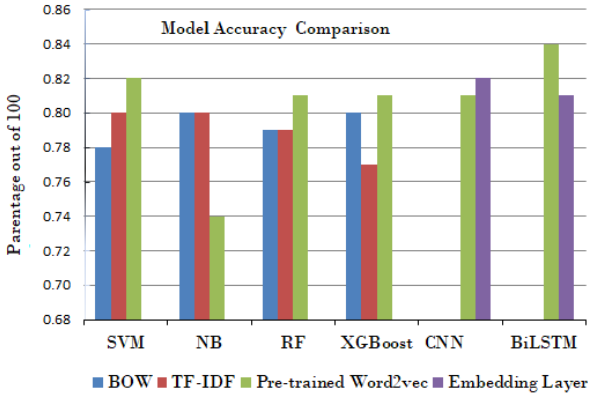

Figure 2: Eight classes column chart accuracy visualization

As we observed in the above figure Bidirectional Long short-term memory algorithm with pre-trained word2vec is achieved the highest accuracy which is 0.84 percent.

In this second experiment, we have only two classes of classification by consolidating all hate classes and offensive classes as hate speech and all free speech (FS) classes into another class. The same algorithm and feature extraction techniques are applied as experiment one.
Firstly, From classical ML classifier like experiment one the highest accuracy which is 0.88 is recorded by SVM with TF-IDF and word2vec feature extraction. However, with Naïve Bayes algorithm word2vec feature extraction technique is achieved low accuracy likewise eight class of classification. Secondly, From the ensemble ML classifier the highest accuracy which is 0.87 is recorded by RF with TFIDF and word2vec feature extraction. However, the low accuracy is recorded which score 0.85 is recorded by Xgboost algorithm with TF-IDF and word2vec feature extraction.
Thirdly, From the deep learning classifier, the highest accuracy which is 0.88 is recorded by BiLSTM and CNN

Experiment result	Accuracy in percentage			
	Feature Extraction Techniques			
Algorithms	BOW	TF-IDF	Pre-trained Word2vec	Embedding Layer
SVM	0.86	**0.88**	**0.88**	-
NB	0.87	0.87	0.79	-
RF	0.86	0.87	0.87	-
XGBoost	0.86	0.85	0.85	-
CNN	-	-	0.82	**0.88**
BI-LSTM	-	-	**0.88**	**0.88**

Table 6: Two classes experiment result with classical, ensemble, Deep ML classifier

with direct embedding feature extraction of text. However, the lowest accuracy 0.82 is recorded by CNN algorithm by combination with word2vec feature extraction techniques.
Graphically, the accuracy of two classes of classification with classical, ensemble and deep learning algorithms and different feature extraction techniques are summarized in the 2D column chart at Figure 3.

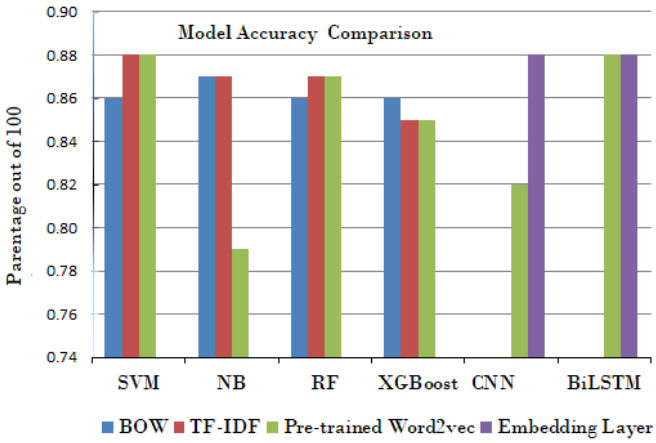

Figure 3: Two classes column chart accuracy visualization

As we observed in above Figure 3 SVM, CNN and BilSTM algorithm is slightly achieved the same accuracy than another algorithm. According to our experiment result, also most BiLSTM algorithm is the best algorithm for Afaan Oromo hate speech detection and classification for both in 8 and 2 classes respectively. In generally, likewise experiment one in experiment two also BiLSTM model with pre-trained word2vec feature extraction is achieved the highest accuracy which is 0.88 percent.

8. Concluding Remark and Future work

Social media like Facebook is staying connected the people from different backgrounds such as race, gender, religion, culture, norms, values, and beliefs. Similarly, it's grants full anonymity and confidentiality to its platform users consequently, it is enabled to spread hate and offensive speech on social media, which disturbs normal life society. Hate and offensive speech detection is a challenging task in almost all social media and it is also a crime in the entire world that is why social media platform is criticized for not

enough fight against hate and offensive speech in their platform. So in our study, we developed an art technique used to overcome the problem Afaan Oromo hate speech detection and classification on social media.

In this study, we collected dataset from scratch by using Face pager and data miner open source tools. we collected twelve thousand eight hundred twelve posts or comments from the suspicious Facebook account oriented on four thematic areas such as race, gender, religion, and offensive.

In our study, we applied different machine learning algorithms from classical (SVM and NB), ensemble (RF and XGBoost), and deep learning (CNN and BiLSTM) with different feature extraction techniques such as BOW, TF-IDF, word2vec, and Keras embedding layers.

From classical and ensemble machine learning algorithm SVM is outperformed machine learning algorithm with word2vec feature extraction techniques which is achieved 0.82 percentage of accuracy for eight classes of classification. From the deep learning algorithm, BiLSTM algorithm is achieved better accuracy which is 0.84 with pre-trained word2vec feature extraction techniques for eight classes of classification. Likewise to octal classification in binary classification also BiLSTM has achieved a better performance result which is 0.88 percent of accuracy with pre-trained word2vec.

Further we are working toward extending hate speech detection for audio, video, emoji, and memes. In addition to this, Hate speech on social media is not only written in the Afaan Oromo language, therefore we are also recommend to developing the model that detects and classifies hate speech from social media with multilingual language.

9. Bibliographical References

Abro, S., Shaikh, Z. S., Khan, S., Mujtaba, G., and Khand, Z. H. (2020). Automatic hate speech detection using machine learning: A comparative study. *Machine Learning*, 10(6).

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

Alrehili, A. (2019). Automatic hate speech detection on social media: A brief survey. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–6. IEEE.

Alsafari, S., Sadaoui, S., and Mouhoub, M. (2020). Hate and offensive speech detection on arabic social media. *Online Social Networks and Media*, 19:100096.

Bawoke, E. (2020). Amharic text hate speech detection in social media using deep learning approach. Master's thesis, Faculty of Computing, Bahirdar.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Defersha, N. and Tune, K. (2021). Detection of hate speech text in afan oromo social media using machine learning approach. *Indian Journal of Science and Technology*, 14(31):2567–2578.

Facebook, (2018). *Facts About Content Review on Facebook*.

Facebook. (2021). Facebook community standard. `shorturl.at/oIJ01`. Accessed: 2021-04-25.

FDRE, (2020). *Hate Speech and Disinformation Prevention and Suppression Proclamation*, 23 March.

Kanessa, L. G. and Tulu, S. G. (2021). Automatic hate and offensive speech detection framework from social media: the case of afaan oromoo language. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 42–47.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

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

Polat, F., Subay, Ö. Ö., and ULUTÜRK, A. S. (2018). Hate speech in turkish media: The example of charlie hebdo attack's. *Journal of Advanced Research in Social Sciences and Humanities*, 3(2):68–75.

Rase, M. O. (2020). Sentiment analysis of afaan oromoo facebook media using deep learning approach. *New Media Mass Commun*, 90:7.

Sreedhara, V. S. M. and Mocko, G. (2015). Control of thermoforming process parameters to increase quality of surfaces using pin-based tooling. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 57113, page V004T05A016. American Society of Mechanical Engineers.

Teh, P. L. and Cheng, C.-B. (2020). Profanity and hate speech detection. *International Journal of Information and Management Sciences*, 31(3):227–246.

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

Yonas, K. (2019). Hate speech detection for amharic language on social media using machine learning techniques. Master's thesis, ASTU.

10. Language Resource References

Degeneh Bijiga, Teferi. (2015). *The Development of Oromo Writing System*.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). (2021). *Ethnologue: Languages of the World.* SIL International, Twenty-fourth edition.

Tegegne, Wondimu. (2016). *The Development of Written Afan Oromo and the Appropriateness of Qubee, Latin Script, for Afan Oromo Writing*.