Afaan Oromo Hate Speech Detection and Classification on Social Media

Teshome Mulugeta Ababu
Dire Dawa University Institute of Technology
School of Computing
Dire Dawa, Ethiopia
teshome.mulugeta@ddu.edu.et

Michael Melese Woldeyohannis
Addis Ababa University
School of Information Science
Addis Ababa, Ethiopia
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. Bi-LSTM 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.

1. Build Afaan Oromo baseline dataset from scratch for hate speech detection and classification.
2. We developed a custom pre-trained word2vec model for Afaan Oromo Language, which is a multipurpose model that can be used for other Afaan Oromo NLP applications.
3. We developed the model that is used to detect and classify Afaan Oromo hate speech on social media.
4. We introduced new data annotation techniques; provides clear and concise dataset annotation guidelines.

1 Introduction

According to the 2020 World Bank statistics report, Ethiopia is the second largest populated countries in Africa next to Nigeria, which has a more than 114 million population. In Ethiopia, Oromo is the largest ethnic group that contributes more than 37 million of the Ethiopian population and Afaan Oromo is their Native Language 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). 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 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). In this study, we focused on the Hate speech definition given by Facebook social media platforms and the Ethiopian government Proclamation FDRE (2020) Facebook (2018).

The spread of hate speech on social media is provoke crime, genocide, massacre, murder, violence, and terrorism, especially for low-resourced languages like Afaan Oromo. So this motivates this research to overcome this challenge because the hate speech detection model developed for other languages is not more functional for the Afaan Oromo. The main contribution of this paper are the followings:

1. Build Afaan Oromo baseline dataset from scratch for hate speech detection and classification.
2. We developed a custom pre-trained word2vec model for Afaan Oromo Language, which is a multipurpose model that can be used for other Afaan Oromo NLP applications.
3. We developed the model that is used to detect and classify Afaan Oromo hate speech on social media.
4. We introduced new data annotation techniques; provides clear and concise dataset annotation guidelines.

2 Data Collection and Preparation

We collected a dataset from four thematic areas (gender, religion, race, and offensive). The dataset
is sampled from Facebook based on our predefined keywords for each areas. The data is collected from the Facebook account, which is most frequently posts and comments in Afaan Oromo. Similarly, the collected dataset is from different broadcasting media services, personal accounts, figurative people and another minor account which has a minimum of 500 followers on Facebook. For this study, we collected a total of 12,812 posts and comments from Facebook. We have collected the dataset from the January 2021 to May, 2021. One challenging task in hate speech detection is that speech that is considered hate speech for one group may or may not be considered hate speech for others. To overcome this problem, we prepare a general guideline to ignore ambiguity while annotating the dataset.

2.1 Data Annotation

If the post/comment promotes hatred or encourage violence by discriminating in a gender, religion and race it is annotated as hate classes for each respective classes. Similarly, if the post or comment contains insult, upsetting word, imprecation words it is annotated as offensive class of offensive. Every four categories of the dataset are annotated individually with three-persons by domain expert from target language based on their willingness and skill to perform the task, then applying equation 1 and select the annotation in which two-persons are agreed upon it. To accept the correct annotation of each post or comment we applied the following equation.

\[
\text{Label} = \text{Mode}(Ex1, Ex2, Ex3)
\]  

(1)

Where: Ex is the annotator expert of Afaan Oromo.

We used Fleiss kappa to compute inter-annotator agreement between annotators. In our study, We obtained 0.801 total average of Fleiss Kappa’s which is a substantial class level of agreement.

3 Experiment Result and Discussion

We have followed Pareto Principle (80:20) during the train test split of datasets. 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 hyperparameters of the classical, ensemble, and deep learning algorithms are tuned by grid search (classical, ensemble) and Keras Tuner for the deep learning algorithm. The experiment result of classical, ensemble and deep learning classifier is presented in Table 1

| 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 1: Eight classes experiment result with classical, ensemble, Deep ML classifier

As we observed in the Table 1 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. According to our experiment result, Bi-LSTM algo-

| 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           | -   | -    | 0.86 | 0.88 |

Table 2: Two classes experiment result with classical, ensemble, Deep ML classifier

rithm is the best algorithm for Afaan Oromo hate speech detection and classification for both in 8 and 2 classes respectively.

4 Conclusion

In this study, we collected 12,812 posts or comments from the suspicious Facebook account oriented on four thematic. We applied different machine learning algorithms from classical, ensemble, and deep learning with different feature extraction techniques such as BOW, TF-IDF, word2vec, and Keras embedding layers.
References

Sindhu Abro, ZA Sarang Shaikh, Sajid Khan, Ghulam Mujtaba, and Zahid Hussain Khand. 2020. Automatic hate speech detection using machine learning: A comparative study. *Machine Learning*, 10(6).

Ahlam Alrehili. 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.

Teferi Degeneh Bijiga. 2015. The development of oromo writing system.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2021. *Ethnologue: Languages of the world*.

Facebook. 2018. *Facts About Content Review on Facebook*.

FDRE. 2020. *Hate Speech and Disinformation Prevention and Suppression Proclamation*.