Comparative Analysis of Deep Learning Based Afaan Oromo Hate Speech Detection

Gaddisa Olani Ganfure (✉ gaddisaolex@gmail.com)

Dire Dawa University Institute of Technology

Research Article

Keywords: Hate Speech, Hate Speech Detection, Deep Learning, Afaan Oromo

Posted Date: January 31st, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1289454/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

COMPARATIVE ANALYSIS OF DEEP LEARNING BASED AFAAN OROMO HATE SPEECH DETECTION

Gaddisa Olani^{1*}

¹Assistant Professor of Computer Science, Dire Dawa University
Institute of Technology, Sabean, Dire Dawa, 1362, Ethiopia.

Corresponding author(s). E-mail(s): gaddisaolex@gmail.com;

Abstract

The network and openness of online media stages permit individuals to communicate their thoughts and offer encounters without any problem. Nonetheless, the Internet, as a typical stage for correspondence, has been weaponized to spread a blend of misrepresentations, misuse, and inclinations with far-reaching outcomes. Prominently, disdain discourse influences society in numerous angles, for example, influencing the emotional well-being of designated crowds, influencing social association, and prompting viciousness and interruption of properties. Whenever left unchecked, disdain discourse can possibly sabotage confidence in electing processes and undermine popularity-based foundations. Web-based media goliaths, for example, Facebook have been scrutinized for permitting its foundation to be utilized as a vehicle for disdain discourse, with least thought given to non-English pages, which has led to intermittent issues in strongly enraptured nations like Ethiopia and other emerging nations. Subsequently, there is a need to construct a hate speech detection model for Ethiopian languages like Afaan Oromo. Aiming to calm the effect of hate speech this paper examines the viability of deep learning models for Afaan Oromo hate speech recognition. Toward this, we collect and annotate the first and largest Afaan Oromo social media datasets. Variations of profound deep learning models such as CNN, LSTMs, BiLSTMs, LSTM, GRU, and CNN-LSTM are examined to evaluate their viability in identifying Afaan Oromo Hate speeches. The examination result uncovers that the model dependent on CNN and Bi-LSTM somewhat outperforms every one of the models on the test dataset with an average F1-score of 87%. Overall, considering the

nature of the Afaan Oromo language and the prevalence of hate speech, we believe that the finding of this study is promising for future works.

Keywords: Hate Speech, Hate Speech Detection, Deep Learning, Afaan Oromo

1 Introduction

In the last few years, we have witnessed a dramatic increase in the usage of the internet. Notably, the introduction of social media networks such as Facebook, Twitter, Telegram, Skype, and others eases the way we communicate, exchange information with friends and other family members. As of January 2020, in Ethiopia alone, there are more than 21 million internet users of which 6.2 million actively use different social media platforms. Social media technology allows society to easily communicate and express their opinion. However, in developing countries such as Ethiopia, the usage of social media is changing from posting or tweeting social, financial, and political debates to disseminating hate speech [1]. The flourishing of hate speech and disinformation online can disrupt democratic debate and practices, facilitate gross human rights violations, and further marginalize minority groups. Ethiopia, even with one of the lowest internet connectivity penetration rates in the continent, is not immune to this phenomenon.

Given the increase in online hate speech, it has become increasingly important to examine the ways in which harmful speech influences physical world attitudes and behaviors, such as hate crimes. However, there has been difficulty in empirically identifying the connections between online racist speech and physical world consequences and thus the research in this area remains limited.

Social media giant like Facebook has been criticized for allowing its platform to be used as a vehicle for hate speech, with minimum moderation given to non-English pages, which has caused recurrent problems in sharply polarized countries like Ethiopia.

In recent years, the increasing propagation of hate speech on social media and the urgent need for effective countermeasures have drawn significant investment from governments, companies, and researchers. Social media platforms rely on a combination of artificial intelligence, user reporting, and staff known as content moderators to enforce their rules regarding appropriate content publication. However, the problem arises when platforms' artificial intelligence is poorly adapted to under-resourced language and companies have invested little in staff fluent in them.

Aiming to classify textual content into non-hate or hate speech, a large number of methods have been developed for resource rich languages such as English, Chinese, French, and others [2–8] using statistical and machine learning approaches. Some research conducted on those languages [9] depicts

that the detection model built using a deep learning model achieves state of the art accuracy. However, only a few attempts have been made for Ethiopian languages such as Amharic and Afaan Oromo.

This paper introduces an automatic classification of social media post that contains Afaan Oromo text into hate speech or not hate by leveraging variants of deep learning approaches. First, we built a corpus of comments and posts retrieved from Facebook and Twitter, and then extract features using word n-grams and word embedding techniques such as Word2Vec [10]. Then, variants of deep learning models such as convolutional neural networks (CNN), long short-term memory networks (LSTM [11]), bidirectional long short-term memory networks (Bi-LSTM [12], GRU, and CNN-LSTMs are used for hate speech detection. The experiment result reveals that the model built using CNN and BiLSTM achieves the highest weighted average F1-score of 87%. Likewise, we also assess the impact of incorporating augmented data into the training dataset. The result shows that the textual data augmentation enhances the model F1-score up to 3%. data. In the light of the gap in this research area, our contributions described in this paper are the following:

- Develop the largest labeled Afaan Oromo hate speech classification dataset of his kind.
- Investigation of state of the art deep learning model for Afaan Oromo hate speech detection
- Assess the impact of adding augmented textual data on the Afaan Oromo Hate Speech classification performance
- Assess the impact of using pre-trained Word2Vec model with the one directly trained with the hate speech classification model
- Build a pre-trained word embedding model, which is useful for other works in this area.

The rest of the paper is organized as follows. We explain related works in Section 2. Section 3 presents the details of the data set. Section 4 reports on approaches we used to classify the hate speech content. In Section 5, we present our results accompanied by detailed performance analysis. Section 6, concludes the paper.

2 Background and Related Work

Online platforms provide a space for discourses that are harmful to certain groups of people. It enables us to provide content that can be bullying, offensive, or hateful. Recently, hate speech can be considered a serious problem by different country authorities. The majority of the research works have been done in resource reached languages like English [2, 3], Spanish [13], and Chinese [7, 8]. The most recent studies on hate speech detection have proposed the use of deep learning techniques for classification [4–6].

A study introduced in [14] developed an Apache spark-based model to classify Afaan Oromo language Facebook posts and comments into hate and

not hate. They employed Random forest and Naïve Bayes as learning algorithms and Word2Vec and TF-IDF for feature extraction using 6,120 (4,882 to train the model and 1,238 for testing). In their experiment, Naïve Bayes and Random Forest outperform with an accuracy of 79.83% and 65.34% with the word2vec feature vector modeling approach respectively. However, they recommended expanding the classification category with different aspects of hate and increasing the corpus size including other sources. So, we increase the corpus size and also used deep learning approaches to improve the performance of the model.

A study in [15] design a system for hate speech text classification in Twitter using, CNN model with 6,655 total datasets. The classifier predicts each Tweet to one of four predefined categories as racism, sexism, both (racism and sexism), and neither. They have created two CNN models based on different input vector sets. using, word2vec, and randomly generated vector baseline. The system based on word2vec word vectors performed best overall, with an F1-score of 78.3%. However, their result is not sufficient and the system wrongly identified some non-hate speech tweets as hate speech. In particular, the system was not able to identify properly the category of both racism and sexism.

Likewise, research introduced in [16] used a Convolutional Neural Network classifier with word embedding as a feature using the Hate Speech Identification dataset distributed via Crowd Flower. They used 24,783 English tweets that have been classified into three classes as hate, offensive language, and neither. And a Publicly available Word2Vec word embedding with 300 dimensions pre-trained on the 3-billion-word from Google News with a skip-gram model. The final model resulted in an F-measure of 90%. However, the model incorrectly identified some non-hate speech as hate speech. Also, the majority of the hated class is misclassified, while the majority of the offensive class is correctly identified.

3 Afaan Oromo Corpus Creation and Annotation

The Afaan Oromo Text dataset for Afaan Oromo hate speech detection (AHSD), which is the main focus of analysis in this paper, is retrieved from comments and posts published in Facebook and Twitter from January 2019 to June 2019 by the authors of this paper.

3.1 Corpus collection

In this work, we targeted Facebook pages and Twitter accounts that are open to suspected hate speech rather than focusing on websites or blogs that already have specific agenda. In Ethiopia, it's common for social network communities are commonly posting on political and religious issues. The summary of Facebook pages and Twitter pages that we use to build the corpus is provided

in Table 1. Those pages listed in Table 1 typically post discussions on political, social, economic, religious, and environmental issues that took place in Ethiopia.

Table 1 Summary of public pages to retrieve comments and posts from Facebook and twit from Twitter

Page/Account Names	Page/Account Names
FBC Afaan Oromoo TV	Ethiopian Press Agency /Bariisaa
BBC Afaan Oromo TV	Oromia Democratic Part/ODP
OMN TV	Kush Media
Fanabc Afaan Oromo	OBS
Jawar Mohammed	Taye Dendea Aredo

In total, 35,200 posts and comments were collected. In order to remove the noise from the data set, rigorous preprocessing was carried out, which resulted in the removal of HTML, URLs, tags, and emoticons, and other language scripts. Using the data augmentation approach outlined in [17], we increase the total dataset to 42,100. Detailed statistics of the balanced Afaan Oromo dataset categorized into 4 classes (i.e., NEUTRAL, HATE, OFFENSIVE, and BOTH) are provided in Table 2.

Table 2 Summary of Balanced data distribution in four classes

Class	Class Label	No of Texts
Neutral	0	10,525
Hate	1	10,525
Offensive	2	10,525
Both	3	10,525

3.2 Annotation Guideline

Annotation is an integral part of the development of Text classification. Annotated data provides useful quantitative information about the occurrence of certain contextual features. For the Afaan Oromo, there is no standardized and labeled corpus for hate speech detection. In hate speech detection, dataset annotation can be performed either manually or crowd-sourcing. A research work in [18] shows annotations generated by expert annotators outperform crowd-sourcing strategy. The reason is that cultural norms play an important role in how hate is expressed and whether or not people perceive something as hate speech. Thus, we also believe that testing for inter-annotator reliability may help alleviate these biases in some cases but will not necessarily neutralize the impact of views on what is hateful or not, which may be shared by the majority of annotators. Consequently, we engaged Afaan Oromo language speakers from different corners of Ethiopia to participate in the annotation process. Each annotator labeled a text into one of the four classes as shown

in Table 2. To ensure consistency among annotators we provide the following guidelines (or rules) for each annotator: A post has been marked as hate:

- If a post/comment uses references to the alleged inferiority or superiority of some target groups.
- If a post/comment affects different characteristics of the person and motivates audiences to take action or make violation.
- If a post/comment contains stereotype which means over-generalized belief about a given target.
- If a post/comment Accusing or Condemning people based on their target groups.

A post has been marked as Offensive:

- If a post or comment contains violent or insulting words but not possible to explicitly identify a target group in the post/comment.
- If a post or comment contains defamation, which is a false accusation a person or attack on a person's character.
- If a post or comment contains insulting, dirty, disgusting, or upsetting words but does not motivate the people to take action.

A post has been marked as Both (i.e, Hate and Offensive):

- If there is a combination of hateful expression, and use an insult, threats, or derogatory terms toward a target's groups.

The structure of the Afaan Oromo Hate Speech Detection system is depicted in Figure 1. As shown in Figure 1, the system has three main components, i.e., Data preprocessor, Feature extractor, and deep learning models. First, the system accepts a sequence of the token as the input, and then it goes through a data preprocessor module which discards URLs, tags and other unnecessary inputs. Using a pre-trained Word2Vec or a Word2Vec trained with the model, the feature engineering component will extract the relevant feature vectors that is used for classification. Finally, the deep learning model will take the vector that corresponds to the input to output the corresponding class label after repetitive training.

3.3 Data Preprocessor

As shown in Figure 1, the data preprocessor module is responsible to remove Twitter profile tags, hashtags, URLs, emojis, and stop words (words that do not add any meaning or value to the sentence) in addition to the tokenization task.

3.4 Feature Engineering

Once the raw data undergoes preprocessing, the Feature engineering part will transform each token into their corresponding embedding vector (a vector that

Fig. 1 Overview of The Investigated Deep Learning Based Afaan Oromo Hate Speech Detection.

summarizes a given token into a new representation that captures the contextual similarity). Deep neural networks can capture features automatically without any human intervention but they are designed to learn from numerical data or word vectors. So, in this study, we have used word embedding techniques called Word2Vec, which is an algorithm that uses a neural network to learn word embedding. Its goal is to estimate each token's position in a multi-dimension vector space based on the similarity of different tokens. Embedded representation can be implemented in two ways: (1) pre-training and using it as an input layer at the beginning of the model architecture, or (2) training it with the model itself. In this study, we assess the impact of both approaches in detecting Afaan Oromo hate speech using Python together with Keras and Tensorflow. Overall, the utilization of embedded representation has two implications to the proposed hate detection models (1) it reduces the dimension of the input into a k -size vector, and (2) it is a more expressive representation as it captures contextual resemblance and semantic sequence of data.

3.5 Feature Learning and Classification Using Deep Learning Models

Once the feature engineering module produces the embedding vectors, the next step is feature learning and classification. In this study, we select five deep learning models for comparison.

- CNN which is a class of Deep Learning model that use convolutional layers and maximum pooling or max-over-time pooling layers to extract higher-level features.
- LSTM is a powerful kind of RNN used for processing sequential data such as sound, time series (sensor) data or written natural language.

- BiLSTM is a hybrid bidirectional LSTM and CNN architecture.
- GRU is similar to long short-term memory (LSTM) with a forget gate, but has fewer parameters than LSTM, as it lacks an output gate.
- CONV-LSTM is a type of recurrent neural network for spatio-temporal prediction that has convolutional structures in both the input-to-state and state-to-state transitions.

After grid search, the optimal hyperparameters selected for those models are summarized in Table 3.

Table 3 Architectural Hyperparameters

Hyperparameter Name	Hyperparameter Value
Number of Convolution Layer	3
Number of Filters in Convolution Layer	250
Filter Size	3×3
Dropout Rate	0.5
Batch Size	128
Embedding Dimesion	300
Hidden Layer Activation Function	Relu
Output Layer Activation Function	SoftMax
Optimizer	AdaGrad
Learning Rate	0.001

4 Experiment and Discussion

4.1 Evaluation Setup

To provide the evaluation results, three evaluation metrics are adopted in this work. These are “Precision”, “Recall”, and “F1-score”. The equation of these three performance metrics is shown in Table 4.

Table 4 Evaluation Metrics

Metrics	Formula
Precision	$TP/(TP+FP)$
Recall	$TP/(TP + FN)$
F1-Score	$2 \times ((\text{precision} \times \text{recall}) / (\text{precision} + \text{recall}))$

In Table 4, “TP”, “FP”, “TN”, and “FN” represents True Positive, False Positive, True Negative and False Negative, respectively. In this work, 10-fold cross-validation is used, where 9 of the fold are used for training the model, and 1-fold will be used for testing the model (i.e., at a time. Finally, the average test result after 10-fold cross-validation is reported in Table 8, and 7.

4.2 Results and Discussion

In this section, we present the results of the experiments we attained using different models. As our primary goal is to compare the performance of models

in different circumstances, the results for each model are presented in separate Tables. We conduct three series of experiments using five different deep learning models. The first one involves the case where the word embedding is pre-trained (see Figure), whereas the second one is the case where word embedding is trained together with the model itself (see Figure). The third and fourth experiments are conducted to assess the impact of data augmentation on classification performance.

Table 5 shows the weighted average results of the experiments carried out with five deep learning methods when the pre-trained Word embedding is used for feature extraction. As can be seen in Table 5, the BiLSTM and CNN accomplished the best performance (with a weighted average F1-score of 87%). The average F1-score of CNN-LSTM, GRU, and LSTM is 85%, 86%, and 82%, respectively.

Table 5 Weighted Average Score of Deep Learning Models With Pre-trained Word Embedding

Model	Precision	Recall	F1-Score
CNN	0.88	0.87	0.87
BiLSTM	0.87	0.87	0.87
CNN-LSTM	0.85	0.85	0.85
LSTM	0.83	0.82	0.82
GRU	0.87	0.86	0.86

We have repeated the experiment revealed in Table 5 using a word embedding structure that was trained with the model as well. The results of these experiments are listed in Table 6. As it can be seen in Table 6, training the model with embedded representation slightly increases the classification performance of the investigated model by 1.5% on average. We believe that training the model with embedding representation will allow the model to optimize the trainable parameters, and hence improve the classification performance.

Table 6 Weighted Average Score of Deep Learning Models Trained With Embedding

Model	Precision	Recall	F1-Score
CNN	0.88	0.88	0.88
BiLSTM	0.89	0.89	0.89
CNN-LSTM	0.86	0.86	0.86
LSTM	0.84	0.84	0.84
GRU	0.88	0.88	0.88

Another question we examined was whether the inclusion of an augmented dataset will affect the classification performance. Results for our experiments carried out with data augmentation are listed in Table 7 and 8. The experiment result shows that the performance of all the investigated models improved after adding augmented data to the training dataset, where the F1-score of CNN, BiLSTM, and CNN-LSTM increase by 1%, whereas that of the remaining models increase by 2%. This is expected that the inclusion of more training

data allows the model to decrease the classification error by repetitive training. In this experiment, we also observe that BiLSTM outperform all the other model with a weighted F1-score of 91%.

Table 7 Weighted Average Score of Deep Learning Models With Pre-trained Word Embedding and Data Augmentation

Model	Precision	Recall	F1-Score
CNN	0.88	0.88	0.88
BiLSTM	0.89	0.89	0.89
CNN-LSTM	0.86	0.86	0.86
LSTM	0.86	0.84	0.85
GRU	0.87	0.86	0.86

Table 8 Weighted Average Score of Deep Learning Models Trained With Embedded Representation and Data Augmentation

Model	Precision	Recall	F1-Score
CNN	0.89	0.89	0.89
BiLSTM	0.91	0.91	0.91
CNN-LSTM	0.87	0.87	0.87
LSTM	0.88	0.87	0.87
GRU	0.89	0.89	0.89

Overall, three main observations can be made from comparing experimental results of the neural network. First, a model trained with embedding representation is able to capture syntactic and semantic relations of Afaan Oromo words. Secondly, the data augmentation mechanism improves the performance of the hate detection models. Finally, Bidirectional Long Short-Term Memory (BiLSTM) achieved the highest F-score than all classifiers used in our experiments.

5 Conclusion

In this paper, we build the largest Afaan Oromo Corpus for Hate Speech Detection and compared five state-of-the-art deep learning models for hate speech detection. Moreover, we have compared the effect of using a pre-trained embedded representation with the one training with the model. Likewise, we examined the effect of leveraging additional data (i.e., augmented data) on the model performance. Through our experiments, we have concluded that training an embedded representation with a model and incorporating augmented samples will enhance the classification performance of all the investigated models. The best performance was showcased by the BiLSTM with a weighted classification F1-score of 91%.

6 Abbreviations

CNN: Convolutional Neural Networks

Long Short-Term Memory: LSTM

Bi-LSTM: Bi-directional Long short-term Memory

GRU: Gated Recurrent Unit

AHSD: Afaan Oromo hate speech detection

CONV-LSTM: Convolutional Long Short-Term Memory

7 Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

Gaddisa Olani Ganfure contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

Acknowledgements

Not applicable.

Authors' information

Affiliations

Assistant Professor of Computer Science at Dire Dawa Univesity

Institute of Technology, Dire Dawa, Ethiopia.
Gaddisa Olani Ganfure

References

- [1] Negussie, N., Ketema, G.: The relationship between facebook practice and academic performance of university students. *Asian Journal of Humanities and Social Sciences (AJHSS)* **2**(2), 1–7 (2014)
- [2] Zhang, Z., Robinson, D., Tepper, J.: Detecting hate speech on twitter using a convolution-gru based deep neural network. In: *European Semantic Web Conference*, pp. 745–760 (2018). Springer
- [3] Pereira-Kohatsu, J.C., Quijano-Sánchez, L., Liberatore, F., Camacho-Collados, M.: Detecting and monitoring hate speech in twitter. *Sensors* **19**(21), 4654 (2019)
- [4] Aluru, S.S., Mathew, B., Saha, P., Mukherjee, A.: Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465* (2020)
- [5] Alshaalan, R., Al-Khalifa, H.: Hate speech detection in saudi twitter-sphere: A deep learning approach. In: *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pp. 12–23 (2020)
- [6] Zimmerman, S., Kruschwitz, U., Fox, C.: Improving hate speech detection with deep learning ensembles. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018)
- [7] Davidson, T., Warmley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11 (2017)
- [8] Das, M., Mathew, B., Saha, P., Goyal, P., Mukherjee, A.: Hate speech in online social media. *ACM SIGWEB Newsletter (Autumn)*, 1–8 (2020)
- [9] Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 759–760 (2017)
- [10] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems*, pp. 3111–3119 (2013)

- [11] Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation* **9**(8), 1735–1780 (1997)
- [12] Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* **45**(11), 2673–2681 (1997)
- [13] Rodriguez, A., Argueta, C., Chen, Y.-L.: Automatic detection of hate speech on facebook using sentiment and emotion analysis. In: 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC), pp. 169–174 (2019). IEEE
- [14] Mossie, Z., Wang, J.-H.: Social network hate speech detection for amharic language. *Computer Science & Information Technology*, 41–55 (2018)
- [15] Gambäck, B., Sikdar, U.K.: Using convolutional neural networks to classify hate-speech. In: *Proceedings of the First Workshop on Abusive Language Online*, pp. 85–90 (2017)
- [16] Biere, S., Bhulai, S., Analytics, M.B.: Hate speech detection using natural language processing techniques. Master Business Analytics Department of Mathematics Faculty of Science (2018)
- [17] Aroyehun, S.T., Gelbukh, A.: Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In: *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 90–97 (2018)
- [18] Mubarak, H., Darwish, K., Magdy, W.: Abusive language detection on arabic social media. In: *Proceedings of the First Workshop on Abusive Language Online*, pp. 52–56 (2017)