Institute of Information Science,
Academia Sinica

Hate Speech Detection for Amharic Language

NLPIR Term Project

Link to code on GitHub

Mequanent Argaw
1-23-2022

1. Introduction:

1.1. Background:

The spread of hate speech through the existing social medias is becoming critical concern these days. Some of social media platforms such as Facebook and Twitter are working on managing the spread of hate speeches on their services.

Amharic language is the official language of the Ethiopian federal government and a mother tongue language for most of Ethiopians. It has its own scripting alphabets called 'Fidel' and there is no capitalization aspect in the language. Similar to other languages in the world, Amharic is also one of the languages being used to spread hate speeches in the social media platforms these days.

There are some approaches previously used to detect hate speeches for Amharic and other languages. Many of the approaches used for it includes mostly the supervised learning approaches like support vector machine (SVM), random forest (RF), Naïve Bayes (NB), logistic regression (LR) and deep learning methods (Tesfaye & Tune, 2020). Convolutional neural networks (CNNs) are also being used for sentiment analysis in different languages and are achieving promising results.

1.2. Problem Statement:

Due to the improvements in mobile computing and internet services, spreading hate speech is becoming easier for those who cannot overcome their anger against other individuals or groups. The same thing is going on in Ethiopia due to the existence of bad governance regarding to the management of diversities in different aspects of humanity. Sometimes, authoritarian governments want to profit from the over-intensified differences among different communities they govern. Another issue is, officials in charge of managing hate speeches may misbehave by judging speeches only based on their own egos.

The motivation to do on hate speech detection is to contribute to automating the detection of hate speeches in Amharic language too. While there are many researches done on hate speech detection for some other languages, very little is done for the morphologically rich and under-resourced

Amharic language, the second in Semitic languages next to Arabic. Hate speech detection in the language is not matured yet and requires more works to be done.

1.3. Objectives:

While the general objective of this project is to develop a hate speech detection for Amharic language, it has the following specific objectives.

- To preprocess the dataset
- To develop a hate speech detection model for Amharic language using CNNs.
- To evaluate the model by predicting the labels of testing sentences.

1.4. Related Works:

There are some research works done so far on hate speech detection as one of sentiment analysis task for Amharic language. Some of them can be generalized by the following table.

Authors	Dataset size	Approach	Result (accuracy)
Zewdie Mossie et al., 2018	6,120	Naïve Bayes with word embeddings	79.83%
Surafel Getachew et al., 2020	30,000	RNN with word embeddings	97.9 %
Zewdie Mossie et al., 2020	491,424	RNN-GRU with word embeddings.	92.56

2. Methodology:

The flow chart of this project can be represented by the following simplified block diagram.

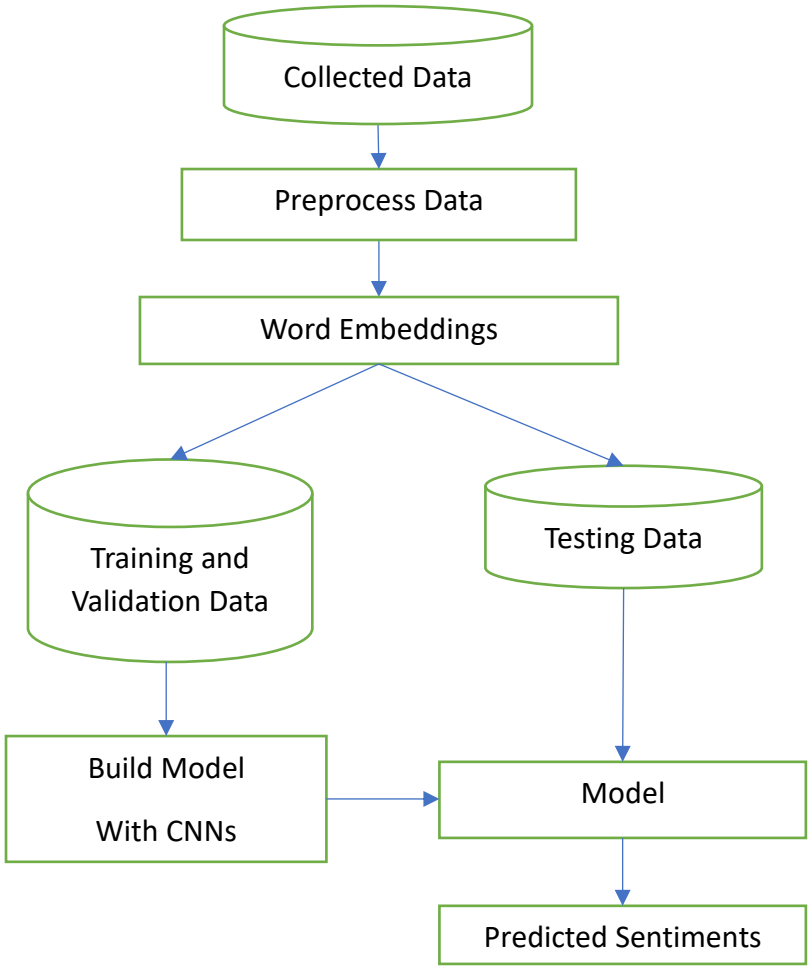

Figure 1. Flow diagram of hate speech detection

Data Preprocessing

Dataset: obtained from: Getachew, Surafel (2020), "Amharic Facebook Dataset for Hate Speech detection", Mendeley Data, V1, doi: 10.17632/ymtmxx385m.1. It contains about 30,000 posts and their corresponding labels.

There are some sentences that have higher number of words while there are others having a single word. Some of single-word sentences have many words with no space used and hence counted as a single word. Because of this, single word sentences have been removed since their total number

is not significant, 421 such sentences found. Sentences having more than 55 words are also removed to avoid padding more zeros onto sequence of many sentences.

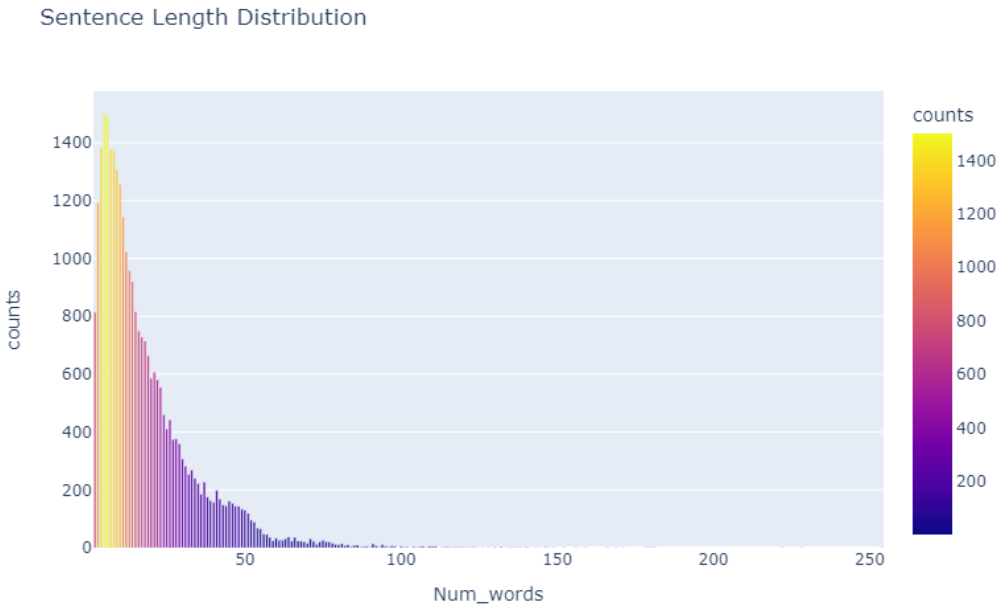

Figure 2: The data distribution after removing single word sentences.

The is a long-tail in the distribution of sentence length counts after removing single word sentences. It is tried to remove the long tail by ignoring sentences having more than 55 words. The maximum sentence length is set to 55 after some trail and errors.

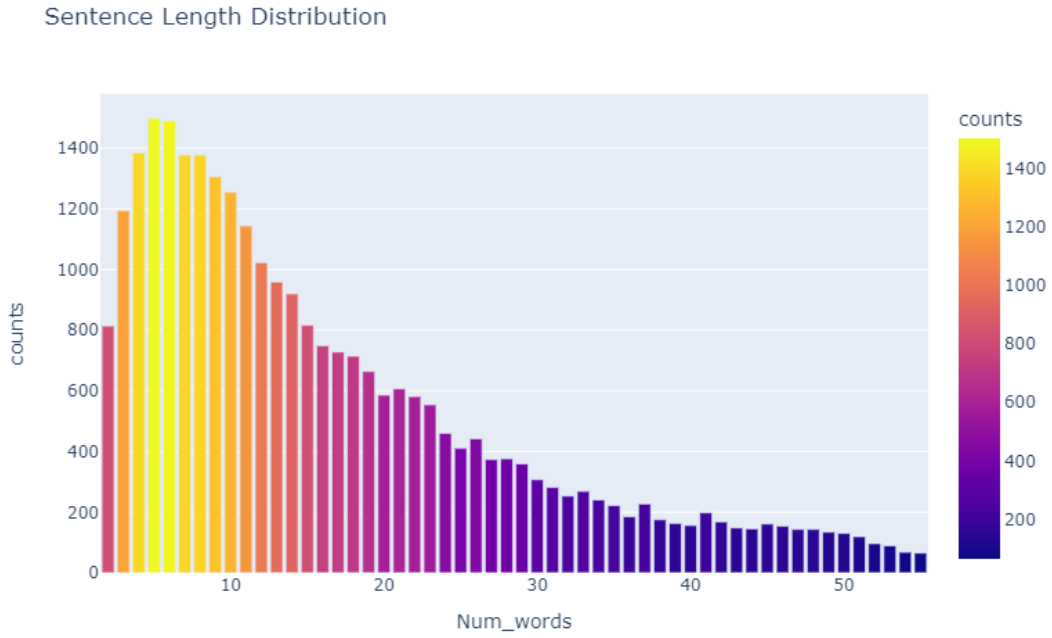

Figure 3: Removing long sentences of smaller count are removed and the distribution looks the following.

After such preprocessing is done, the distribution of the hate and hate-free sentences is somehow balanced each other as we can see from Table 1.

Table 1: Distribution of hate and hate-free sentences

Number of posts in the two sentiments	
Hate	14450
Free	14158
Total	**28,608**

There is no capitalization in Amharic language to deal with as a preprocessing task and there are no punctuation marks in the dataset to be cleaned even if the language has many punctuation marks. From the dataset, 80% is used for training, 10% for validation and 10% for testing.

Convolutional Neural Networks

The approach used to develop the hate speech detection model from Facebook posts is the deep learning approach, convolutional neural networks (CNNs). It has been done using other deep learning approaches like recurrent neural networks and the intention of this project is to examine the performance of CNNs.

The features used for classifying the sentences' sentiments are word embeddings. Currently, word embeddings are promising features for different natural language tasks and hence for sentiment analysis too. The word embeddings are generated by the embedding layer having 100 dimensions of vectors to represent each word in the vocabulary of the corpus.

The convolution layer has 128 filters of size 3 each. One convolution layer with relu activation, a max pooling layer, a dropout layer, and a dense layer with sigmoid activation are used as parts of hidden layers.

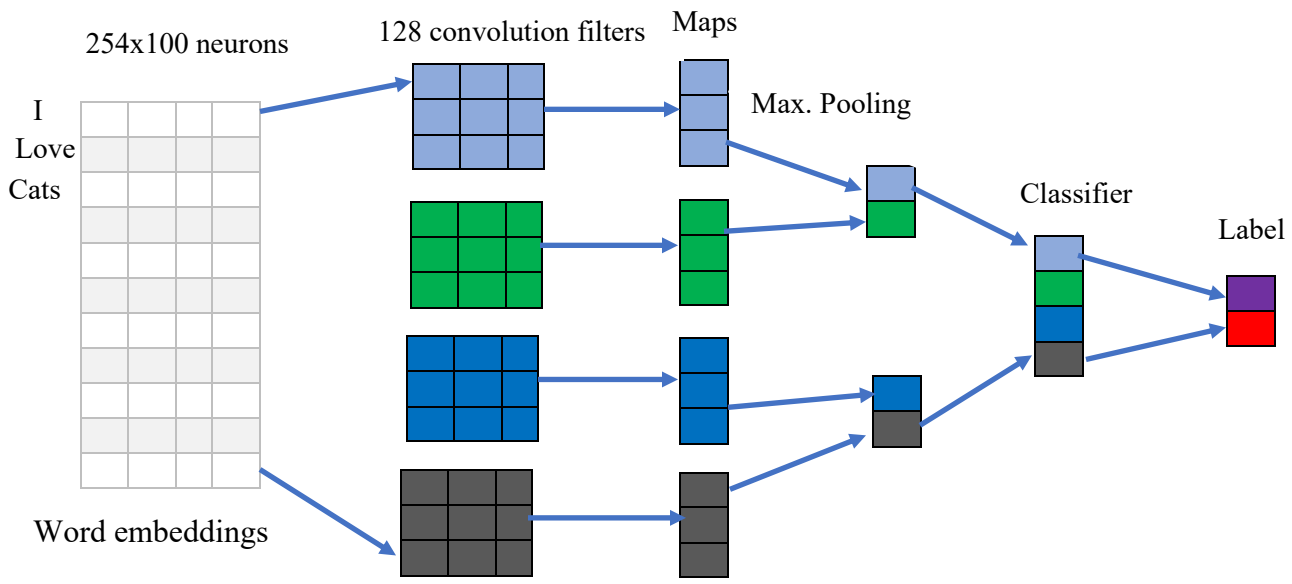

Figure 2. Simplified representation of CNN used for the hate speech detection

3. Results:

The model built using CNNs was evaluated by 10% of the dataset. The maximum result obtained from the model is 66% accuracy. Part a in the following subsection show the screenshot of the predicted labels vs the actual labels.

a. Predicted Sentiments

```
Generate predictions for test sets.
['Free' 'Free' 'Free' ... 'Hate' 'Free' 'Hate']
```

```
1 y_test
```

```
array(['Hate', 'Free', 'Hate', ..., 'Hate', 'Free', 'Hate']
```

b. Classification Report

From the classification report, we can observe the model's precision and recall values for hate and hate-free sentences. The accuracy of the model is limited to 66% which may be due to some limitations like inconsistences in the dataset or the setup limitations in the model.

	precision	recall	f1-score	support
Free	0.65	0.62	0.64	1388
Hate	0.66	0.69	0.67	1473
accuracy			0.66	2861
macro avg	0.66	0.65	0.65	2861
weighted avg	0.66	0.66	0.66	2861

c. Confusion Matrix

It is also possible to observe the true negative, false positive, false negative and true positive distributions from the confusion matrix.

	Hate	Free
Hate	858	530
Free	454	1019

The plots in Figure 3 and Figure 4 show the training and validation loss and accuracy curves, respectively. The figures show a kind of overfitting even if a dropout layer was used to damp values that lead to the case.

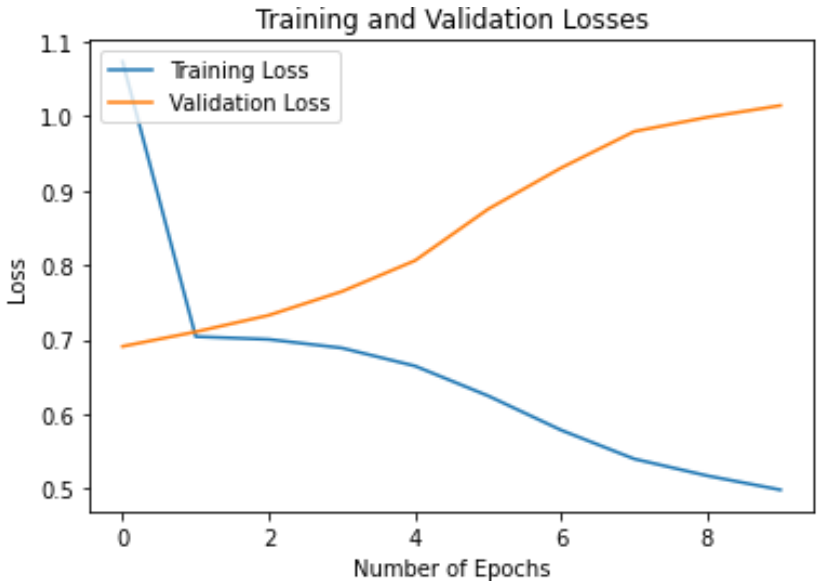

Figure 3: Training and validation losses

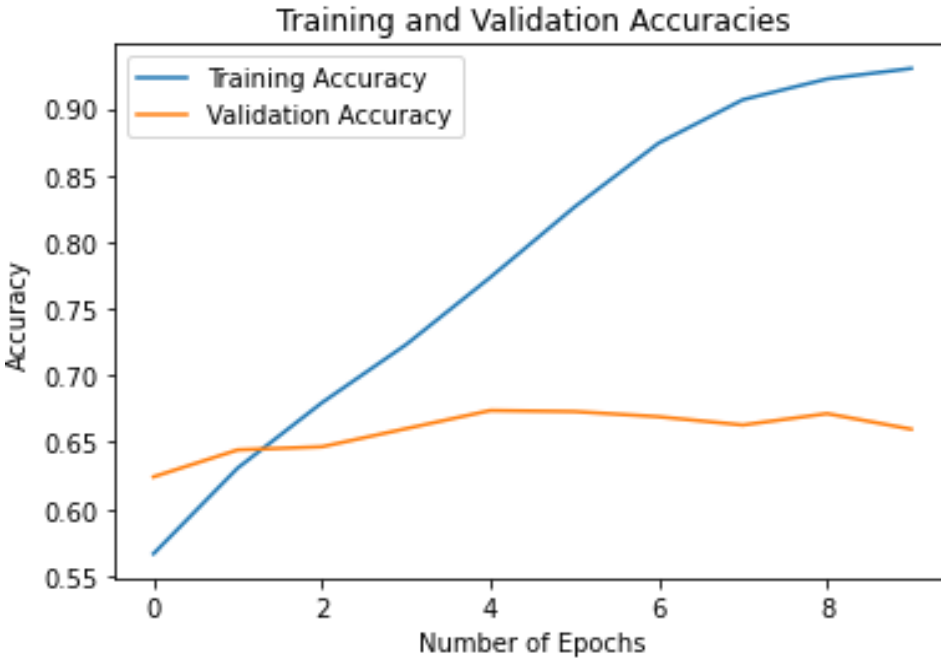

Figure 4: Training and validation accuracies

4. Conclusion:

In this project, it is tried to develop a hate speech detection for Amharic language using convolutional neural networks and word embeddings as features. The result observed so far is 66% accuracy which may be because of the dataset and/or the parameters of the model. The dataset has too many misspelled words and words written together with no white spaces in between. The other issue may be because the project is done with no prior experience on convolutional neural networks which may lead to misconfigure some model parameters.

5. References:

1. Zewdie Mossie and Jenq-Haur Wang, Social Network Hate Speech Detection For Amharic Language, Conference on Computer Science & Information Technology, 2018
2. Surafel Getachew Tesfaye and Kula Kekeba Tune, Automated Amharic Hate Speech Posts and Comments Detection Model using Recurrent Neural Network, preprint https://doi.org/10.21203/rs.3.rs-114533/v1, 2020
3. Zewdie Mossie and Jenq-Haur Wang, Vulnerable community identification using hate speech detection on social media, *Information Processing and Management, Elsevier, 2020*
4. Safa Alsafari, Samira Sadaoui and Malek Mouhoub, Effect of Word-Embedding Models on Hate and Offensive Speech Detection, arXiv:2012.0753, 2020
5. Nadia Nedjah et al., Sentiment Analysis using CNN via Word Embeddings, Evolutionary Intelligence, 2019