Hate Speech Detection

Akhila Sulgante, Kasi Viswanath Vandanapu, Misha Mody, Poulomi Das

Abstract

As social media platforms allow users to express their views freely, this creates a forum for disagreement and hate directed at someone based on their race, nationality, sexual orientation, and so on. Identifying hate speech online is the need of the hour as cyberbullying takes a toll on one's life. In this project, we have built multi-models to detect hate and offensive speech. We utilized the Davidson dataset which had 24,783 categorized as Offensive, hate, and neither, and the TSA dataset which had around 31,962 categorized as racist, non-racist, sexist, and non-sexist. We built and tested the data on three categories of model such as Transformer model - CNN and MLP, GloVe embedding models - Glove – Bi-LSTM + CNN + MLP and TF-IDF Embedding Models - TF-IDF + MLP + SVM + XGB. Amongst the mentioned model, we tried the TSA dataset on the Transformer model which had the highest accuracy of 0.97

1. Introduction

Hate speech detection is the task of detecting if communication such as text, audio, and so on contains hatred and/or encourages violence towards a person or a group of people. We have trained different models for detecting hate speech using Davidson and TSA Dataset. Out of all the models we trained, we observed that the transformer model with CNN and MLP using ELECTRA and BERT Embedding gave the maximum accuracy which is around 91% as compared to other word embedding models such as Glove – Bi-LSTM + CNN + MLP and TF-IDF + MLP + SVM + XGB. We also developed Visualization for XGB Classifier which takes user input and classifies the sentence into Hate, Offensive, or neither category. For the Highest accuracy, we performed cross-validation, which gave us 97% accuracy.

2. Method

2.1 Dataset and Data preprocessing

The dataset we used for training the models are Davidson and TSA Dataset. The characteristics of the data as follows:

Table 1: Datasets and their key statistics

Dataset	Class and Statistics
Davidson	Offensive – 19,190 (77.4%) Hate - 1,430 (5.8 %) Neither - 4,163 (16.80%) Total $\sim 25k$
TSA	Not Racist/Sexist - 29,720 (92.99%) Racist/Sexist – 2,242 (7.01 %) Total $\sim 32k$

Fig 2.1: Dataset statistics

2.2 Transformer model - CNN and MLP

This approach focuses on the transformer models, including Small BERT, BERT, ELECTRA, and ALBERT, with each in combination with CNN and MLP separately. BERT and other Transformer encoder architectures have been

successful on a variety of tasks in NLP. They compute vector-space representations of natural language that are suitable for use in deep learning models. The BERT family of models uses the Transformer encoder architecture to process each token of input text in the full context of all tokens before and after.

Architecture: For MLP, its architecture consists of one dense layer, a dropout layer with the probability of 0.1, and one classification layer. The CNN architecture, as shown in Figure 1, consists of two CNN layers with filter sizes of 32 and 64, respectively. The output from the last CNN layers is fed to a max-pooling layer, followed by a dense layer with 256 neurons and a dropout layer with the probability of 0.1, before feeding to the last output layer. We have Electra and small Bert as pertaining models. The transformer uses an encoder-decoder architecture. The encoder extracts feature from an input sentence, and the decoder uses the features to produce an output sentence. We use a 512-dimensional vector for such embedding. So, if the maximum length of a sentence is 200, the shape of every sentence will be (200, 512). The output vector from the decoder goes through a linear transformation that changes the dimension of the vector from the embedding vector size (512) into the size of vocabulary (say, 10,000). The SoftMax layer further converts the vector into 10,000 probabilities.

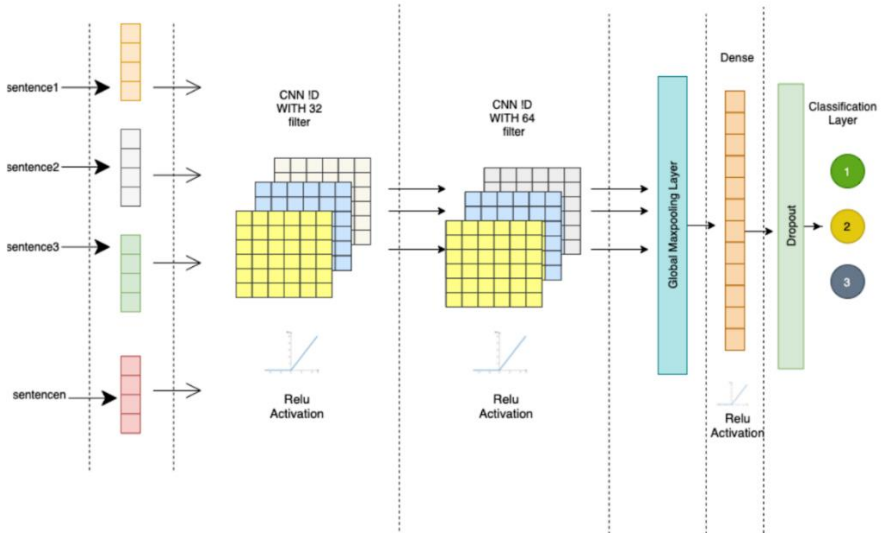

Fig 2.2: CNN Network Architecture

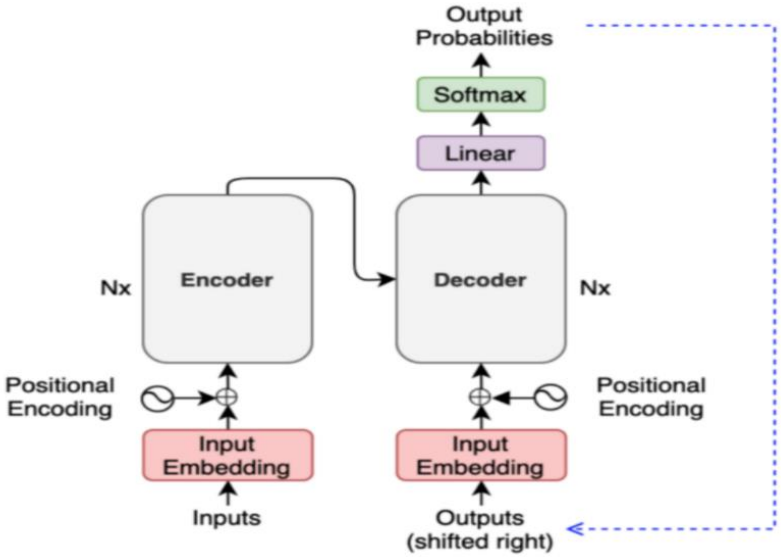

Fig 2.3: Encoder-Decoder stacks in Transformer model

Optimization: As input for transformer embedding, we tokenize each tweet with the BERT tokenizer. It contains invalid characters removal, punctuation splitting, and lowercasing the words. Based on the original BERT, we split words to sub word units using Word Piece tokenization. As tweets are short texts, we set the maximum sequence length to 64 and in any shorter or longer length case it is padded with zero values or truncated to the maximum length. We use the cross-entropy loss function and the Adam optimizer to optimize the models.

2.3 GloVe embedding models – BiLSTM, CNN and MLP

This approach uses Global vector embedding with Convolutional Neural Network, Bi-Long Short-Term Memory, and Multilayer perceptron. The text is classified using a deep learning model to analyze hate speech. GloVe is used to perform word embedding. Bi-LSTM is used to model the sequential dependencies between words and phrases in both directions of the sequence. CNN and MLP are used for multi-class classification tasks.

Architecture: In this model, the word embedding is done using GloVe. It encodes the corpus of tweet text into pre-trained weights. Then, the above-mentioned deep neural networks use this embedded layer as input. For deep neural networks, the architecture flow is building Bi-LSTM model and then trying it with CNN and MLP layers for the same to obtain maximum accuracy. First, Bi-LSTM model is built with a global max-pooling layer, batch normalization, 3 dropout, 2 dense, and one classification layer. This is followed by two convolutional layers, one max-pooling, dense, dropout, and classification layer each for CNN. This model was followed by flattened and dense layers for MLP

Optimization: We used "**Adam**" optimizer for obtaining accurate results. Internal and dense layer activation is done using "**ReLU**"

2.4 TF-IDF Embedding Models - MLP, SVM and XGB

This approach tests data on MLP deep learning model with a linear model such as SVM and a powerful distributed gradient-boosted decision tree (GBDT) machine learning library, XGB. Word embedding for each model is done using TF-IDF, which assigns scores for each word. MLP is used for label generation using non-linear activation. SVM is used for classification and regression issues in the model. XGB is used to turn weak learning into a strong one.

Architecture: In this model, after transforming and assigning weights for each class, the importance of the term is analyzed using TF-IDF. This model uses the TF-IDF score instead of frequency for each word. MLP layer considers the vectorized data and includes layers such as dense, dropout, and classification layers. SVM classifier from sklearn is used for locating decision borders between classes. XGBoost classifier is used on the model for predicting accurate targets.

Optimization: Back-propagation is used in MLP to update the weights of all the nodes. Addiction of the XGB classifier increased the accuracy of the model.

2.5 Visualization on Hate Speech Detection

We have created Diverging Stacked Bar Chart visualization for detecting the input sentence as either Hate, Offensive or Neither speech. This visualization is deployed in the Observable Platform [6] which takes user input sentence, generates the following graph dynamically. XGBoost classifier with TF-IDF embedding is used to classify the sentences.

Fig 2.4: Diverging Stacked Bar Chart visualization

3. Results

In terms of macro F1 score, it can be observed that Glove is better than TF-IDF and BERT Embedding. TF-IDF in predicting hate speech more accurately. The results obtained by Glove embedding with CNN and Bi-LSTM are like the highest score obtained by TF-IDF embeddings with XGB. with just a percent difference. This high score was almost touched twice by the glove embedding. The TF-IDF based MLP and SVM models obtain similarly superior performance as the Glove-based MLP model.

In terms of weighted average F1 score, it can be found that XGB with TF-IDF produced incredibly competitive results, which is more effective than all Glove-based models. Compared to the TF-IDF and Glove-based models, the transformers-based models achieve a significant increase in both metrics, achieving the best macro F1 score of 0.76 and the best weighted average F1 score of 0.91.

More specifically, in terms of macro F1, all the transformers' models outperform every single model in combination with TF-IDF or Glove by a large margin. These results demonstrate that the transformers can perform better in both large and small classes, especially the small classes (i.e., the hate class), when compared to Glove and TF-IDF. Small BERT is less effective than TF-IDF-based XGB in weighted average F1 score, indicating that Small BERT underperforms XGB on the large class (i.e., the offensive class). Overall, BERT-based CNN and Electra-based MLP turns out to be the best performers for the Davidson and TSA datasets.

4. Discussion

- Paralinguistic signals like emoticons and hashtags are frequently used in social media posts, and there is a lot of poorly written text in their linguistic content.
- The task's context-dependence and the lack of agreement on what constitutes hate speech provide additional challenges, making it challenging even for humans to complete.
- Unbalanced data is another issue to consider. Even though the propagation of unpleasant and abusive content on social media is a big issue, it is still true that this type of content makes up a small portion of all content.

5. Conclusion

In this project, we experimented with various machine learning models (Bidirectional LSTM, MLP, CNN etc.) for online hate detection and found the best performance with **Transformer model** (transformer based embedding model ELECTRA and BERT) as the most impactful representation of hateful social media comments. Although there are subtle differences between the sites, the methodology may be used on many different social media platforms with strong generalizability. Our results lend credence to the need for more generalized online hate classifiers across various social media sites. Online hate researchers can apply the model we make openly available to real-world applications and build upon it.

Online hate is a rampant problem, with the negative consequence of prohibiting user participation in online discussions and causing cognitive harm to individuals. Since hate is prevalent across social media platforms, our goal was to develop a classifier that performs satisfactorily in multiple platforms. The results show that it is possible to train classifiers that can detect hateful comments in multiple social media platforms with solid performance and with the share of false positives and negatives remaining within reasonable boundaries.

Future Work

In order to enhance the amount of training data available, we also intend to experiment with data augmentation techniques, such as adding errors or filtering to terms or using synonyms in tweets.

Future trials will involve the adoption of more explicable models, and a more in-depth analysis of the working principles of our current models will also be conducted (transformer models for example have been successfully examined using the Captum tool).

Extending the model to different languages would also be a great way to detect more hateful speech.

Creating a dataset which consists of online posts from various social media sources would also give a good generality to the model.

6.Acknowledgment

This project is completed for CS 6120 summer course of 2022 taught by Professor Uzair Ahmad

7. References

1. Kovács, G., Alonso, P. & Saini, R. Challenges of Hate Speech Detection in Social Media. *SN COMPUT. SCI.* **2,** 95 (2021). https://doi.org/10.1007/s42979-021-00457-3
2. Salminen, J., Hopf, M., Chowdhury, S.A. et al. Developing an online hate classifier for multiple social media platforms. Hum. Cent. Comput. Inf. Sci. 10, 1 (2020). https://doi.org/10.1186/s13673-019-0205-6
3. Al-Hassan, Areej & Al-Dossari, Hmood. (2019). DETECTION OF HATE SPEECH IN SOCIAL NETWORKS: A SURVEY ON MULTILINGUAL CORPUS. 83-100. 10.5121/csit.2019.90208.
4. Mullah, Nanlir & Zainon, Wan Mohd Nazmee. (2021). Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3089515.
5. Malik, Jitendra & Pang, Guansong & Hengel, Anton. (2022). Deep Learning for Hate Speech Detection: A Comparative Study.
6. https://observablehq.com/@kasivisu4/hsd-visualization
7. https://github.com/kasivisu4/hate-speech-detection