Ari Butterfield
asb180007
Dr. Latifur Khan
CS 6350.001
8 May 2023

Final Project – Identifying Hate Speech

Introduction

Hate speech is a common concern in an era of free-flowing communication, and high connectivity via the Internet. Many people avoid public chatrooms, or feel threatened because of the presence of hateful ideas. Additionally, many people are concerned that hyper-censorship could limit the free market of ideas. Counter speech is one such way to address hate speech. Rather than removing it altogether, counter speech can address its faults directly and head-on. Regardless of how it is handled, however, the identification of hate speech is necessary to properly address it. As a lone ranger for my final project, I decided to create a classification model that can ideally be used in conjunction with a counter speech generator to automate the fight against hate speech on public forums.

Data

For this project I used the Multitarget_CONAN.csv dataset [1]. This dataset has over 5000 pairs of hate speech and counter speech. This is an effective dataset to train on for hate speech recognition, as both the hate speech and counter speech are addressing sensitive topics, but only one of them is hateful. In order to use this data set, I separated each hate speech / counter speech pair, and put each as an individual entry, along with a label of 1 for hate speech, and 0 for counter speech. I split the data (10,000 entries) into 80% training set and 20% testing set.

Methodology

I used 2 primary models:

1. Naïve Bayes Classifier from TF-IDF scores
2. Convolutional Neural Network classifier

Naive Bayes Classifier

This classifier was much simpler than the CNN, but it gave a wonderful baseline by which to compare the performance of the more complicated CNN dataset. First, I vectorized each message of the training set into a vector of tokens, each with a TF-IDF score (term frequency-inverse document frequency) in order to measure the significance of each present token. Using these vectors, I trained the model using a multinomial naïve bayes classifier. I then classified the test set and evaluated the results.

Convolutional Neural Network Classifier

For the Convolutional Neural Network, I ran the preprocessing 3 different ways. First, I ran it with no preprocessing. Then I ran the model with stripped punctuation, turning everything lowercase, and removing stop words. Then finally, I ran the model with the addition of lemmatization.

For the neural network itself, I used the keras Sequential() model (CNN), with 3 dense layers. Through suggestions online [2], and slight adjustments of the model, I used the ReLU (Rectified Linear Unit) activation function for the first two layers (sets negative values to 0) and Sigmoid (smooth activation function) for the final layer. My 3 layers had 256, then 128, and of course, 1 unit for the final layer to make a prediction. I used the Adam optimizer (Stochastic Gradient descent) with a log loss function. All of the models stopped training before the full 10 epochs, because I used early stopping and the loss function was already sufficiently diminished.

Results

Model	Accuracy
Naïve Bayes with TF-IDF	90.11%
CNN with no data preprocessing	91.31%
CNN with stripping and stop words	90.26%
CNN with no data preprocessing	90.00%

Discussion

Interestingly enough, the data preprocessing was of no benefit to the overall performance of the models. Even stripping punctuation and removing stop words produced negligible difference. As I doubt stop words produce identifying information, they at least don't harm the model's performance. And it appears that lemmatization removes a lot of significance, likely by changing words unintentionally. Otherwise, the increased training with the CNN did produce better results than Naïve Bayes with TF-IDF, though it performed similarly. Given the difficult nature of the dataset, I believe these models could effectively identify general hate speech and potential hate speech in a public form, and when used in conjunction with a counter speech generator, could effectively counter harmful narratives.

Ari Butterfield
asb180007
Dr. Latifur Khan
CS 6350.001
8 May 2023

References

[1] Guerrini, Marco. "CONAN: COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech." GitHub repository, January 30, 2018. https://github.com/marcoguerini/CONAN.

[2] Keras Team. "Guide to the Keras Sequential Model." Keras documentation. Accessed May 8, 2023. https://keras.io/guides/sequential_model/.