Hate speech detection using SVM and CNN model

Heena Khan

1 Introduction

Hate speech is currently a huge problem and is growing considerably. It particularly affects marginalized communities, like minorities, women, people of color and other vulnerable communities. There are a lot of papers introduced in the last few years that focused on the identification of abuse occurring on social media in the United States and the United Kingdom [6]. India has been equally active and working it's way up to achieve its position in social and racial abuse ranking. Recently, comments and posts that called Bengali Muslims "pigs", "terrorists", "dogs ", "rapists", and "criminals" seemingly in violation of Facebook's standards on hate speech were shared nearly 100,000 times and viewed at least 5.4 million times, showed the Avaaz review, which covered 800 Facebook posts related to the state of Assam alone. [5]. The 2018 Workshop on Trolling, Aggression, and Cyberbully (TRAC) hosted a shared task focused on detecting aggressive text in both English and Hindi [4]. The papers that were introduced during this workshop focused on Facebook and Twitter comments from the same region.

2 Dataset

The dataset from TRAC is available to the public and contains 15,000 Facebook comments labeled as overtly aggressive, covertly aggressive, or non-aggressive. In our paper, we are going to create a model that can classify aggressive text using Multi-class algorithm SVM and a Deep Learning CNN methodology. We are going to use the same dataset used by Kumar, Bhanodai, Pamula and Reddy [4]. Their dataset contains both English and Hindi text set, but we will be focusing only on the English text set. For all their runs they have used Long Short-Term Memory model (LSMT) and used random baseline cross-validation to measure accuracy. For English (Facebook) dataset their best run was (i.e. 0.3572) above the random Baseline (i.e. 0.3535) [4].

3 Methodology

The authors in TRAC [4] mentioned implementing text classification models like SVM, ensemble model, etc and Deep Learning models like CNN and Random forest as a classifier in their future work. We will implement SVM and CNN from their future work and we will be using the English dataset only. The first step in the implementation of any model is text pre-processing. Our text pre-processing involves, changing all the text to lower case, removing stop words, word lemmatization, removing emoji, removing all non-alphabetic text and word tokenization. The processed text is then encoded and implemented using different classification techniques.

3.1 Using Convolutional Neural Network (CNN) for hate speech classification

The most important part is converting an arbitrary length string into a fixed-length vector to use Neural Net API. There are different ways of converting words to vectors like Bag Of Words (BOW), but on using BOW the order of the words is lost which will be too little information for our neural network. LSTM where each character in the word is encoded into one-hot representation, which can be too much information for a neural network. So to find a balance between too little and too much

information, we will be using word embedding were each word is assigned a pre-defined set of values. This can be achieved by using Glove.

CNN Implementation: The first step is to generate feature embedding. Feature embedding for all words will be constructed by using word embedding. We will then use padding to convert each comment into a fixed-length vector, A pooling layer in the network converts each tweet into a fixed-length vector, capturing the information from the entire tweet. A max-pooling layer then captures the most important factors from the comment [2]. In the output, the softmax function will be used to classify comments according to the label.

3.2 Using Multiclass Support vector machines(SVM) text classification to classify hate speech

The n-grams are generated using word tokenization. We will extract the n-gram features from each comment and weight them according to their TF-IDF values. We will experiment on different values of n-grams, where n ranges from one to four. This will then form our feature for the SVM model [3].

4 Testing and Cross Validation

We will follow the proper testing protocol for the classification tasks.

1. Since we have an imbalanced dataset where around 6200 comments are nonaggressive, around 3500 are covertly aggressive and around 4300 are overtly aggressive, SVMs could produce sub-optimal results with imbalanced datasets [1]. So we will be taking care of the imbalance issue before dividing our data into test and train.

2. To estimate the skill of our model on the new data we are going to follow a classic k-fold cross-validation procedure, we'll be using scikit-learn libraries for this. This provides further confidence in our model's ability to generalize.

3. We will consider metrics such as accuracy, false-positive rate (FPR), true positive rate (TPR), and area under the precision-recall curve (PR-AUC), F1 score to quantify the quality of our model using a confusion matrix.

The average k-fold cross-validated results for the multi-class Support Vector Machines (SVM) model will be compared to the Convolutional Neural Network(CNN). A confusion matrix specifying precision, recall and F1 score for both the methods will be created.

References

[1] Rukshan Batuwita and Vasile Palade. Class imbalance learning methods for support vector machines. 2013.

[2] Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90, 2017.

[3] Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. *arXiv preprint arXiv:1809.08651*, 2018.

[4] Ritesh Kumar, Guggilla Bhanodai, Rajendra Pamula, and Maheshwar Reddy Chennuru. Trac-1 shared task on aggression identification: Iit (ism)@ colingâĂŹ18. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 58–65, 2018.

[5] Casey Newton. The verge - hate speech is spreading on facebook in india again.

[6] Marian-Andrei Rizoiu, Tianyu Wang, Gabriela Ferraro, and Hanna Suominen. Transfer learning for hate speech detection in social media. *arXiv preprint arXiv:1906.03829*, 2019.