Identifying Offensive Language in English Tweets using Machine Learning and Deep Learning Methods

Arif Shahriar
ashahri1@ualberta.ca

Alvina Awwal
awwal@ualberta.ca

1 New Literature Reviewed

After reading and researching papers on our exact tasks for the previous proposal, this time we dug into the history of offensive language on social media. The term "hate speech" was first established by Warner and Hirschberg (2012) and the paper defined it as "abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation". Another paper by Schmidt and Wiegand (2017) did a thorough research on this phenomenon and identified a number of features for hate speech such as words, sentiment, knowledge-based features, linguistics, etc. They also provide insight on the usage of these terms and mention how recently, "hate speech" is being replaced by the term "cyberbullying".

These backgrounds on the topic helped us understand the core of our problem and motivated us to work on it more. A SemEval (SE) 2020 paper based on our task helped us understand how the BERT (Bidirectional Encoder Representations from Transformers) model worked and how it came to be. They got their best results using BERT and therefore describe the process in an informative way (Thenmozhi et al., 2020). Another SE paper by Ahn et al. (2020) gave a thorough description of their pre-processing steps which helped us immensely while implementing our initial SVM architecture. Our train dataset was processed and normalized similarly to theirs through emoji substitution, letter casing normalization, URL replacement, etc.

2 Draft of Methods

The main objective of our first task is to classify posts of the dataset into two categories- offensive (OFF) and not offensive (NOT) posts. Offensive tweets include unacceptable language, and not offensive tweets do not. The second task further classifies the identified offensive tweets as targeted (TIN) or untargeted insult (UNT). Supervised machine learning and deep learning approaches will be implemented on both the tasks which were taken from SE 2020 Task 12. We have successfully scraped 14,687 posts from the Twitter database using the Twitter IDs given in the SOLID dataset for our training set (Rosenthal et al., 2021). The test set, which was provided contains 3887 posts, and will be used for evaluation purposes.

Initially, the training dataset had been divided into train (70%) and dev (30%) sets using sklearn train test split function. During data splitting, stratify parameter was used to keep the proportion of classes similar in train and dev distribution. Parikh et al. (2020) did not explicitly mention applying similar approaches but that is necessary to address uneven class distribution in the dataset.

For task 1, we have already performed data preprocessing and trained an SVM model (VanderPlas, 2016). The raw social media data contained emoji, punctuation, URL tags, hashtag, stop words, user tags, etc. and thus, data pre-processing was required to facilitate feature extraction, remove inconsistency from the data, and make it more understandable for the NLP models. Firstly, since some of the offensive tweets had already been identified and removed by the twitter authority, the non-retrievable rows were removed from the dataset. This resulted in a final dataset comprising 9,768 posts. Secondly, all the text data were converted to lowercase, and URL tags, emoji, usernames were removed using Python's regular expressions module. In addition to that, each tweet text was broken into meaningful smaller tokens like word, phrases, symbols with a tokenization module from nltk library. This library also helped to remove punctuations and stop words from the data. Finally, POS tags of each word were identified to perform word lemmatization, which identified the common

lemma of different forms of verbs in the provided text data. Inspired by Parikh et al. (2020), a similar hypothesis was followed to label data where posts with average confidence value greater than 0.5 were mapped as 'OFF' class, else mapped to 'NOT' class.

We incorporated a statistical method called TF-IDF for word vectorization, which keeps track of total occurrences of certain words to recognize their importance in the given context. TF-IDF representation of each tweet data was generated with 10,000 top feature values. Unlike other SE papers, which had around 9 million data, our significantly smaller dataset could have resulted in overfitting. To avoid that problem and ensure generalization of the model, SVM algorithm with K fold cross-validation has been applied where data was further divided into 5 smaller sets and training was performed on 4 sets. Finally, the resulting model was validated on the remaining set. The optimal model which performed better on the dev set was trained with the following parameter values (C=1.0, kernel='linear', degree=3). We evaluated performance of the optimal model on the test set.

In future, a logistic regression (LR) model will be trained on the derived feature values from word vectorization method for both the tasks. Furthermore, a pre-trained bi-directional model (BERT) will also be used for the tasks because it is computationally less expensive (Devlin et al., 2018). We have observed from multiple papers of the SE 2020 competition that BERT has outperformed other ML and DL models by achieving the best macro F1 score (Thenmozhi et al., 2020).

3 Draft of Evaluation Protocol

We will be using three metrics for evaluation protocol – macro F1 score, precision, and recall. Most of the SemEval 2020 papers have used just the macro F1 score to evaluate the performance of their models for these tasks. Precision indicates how many positive observations were predicted accurately to the total number of positive observations. On the other hand, recall estimates the ratio of actual positives to total correctly classified observations. For task 1, offensive tweets are considered as positive class, and non-offensive tweets are considered negative class. For task 2, targeted insults and untargeted insults indicate positive and negative samples. However, the F1 score combines both precision and recall in a single metric, which computes their harmonic mean to summarize both of their relative performance. Finally, We consider simple arithmetic mean of per class F1-score, also named as macro averaged F1-score.

Methods	F1-macro	Comments
TF-IDF with SVM (5 fold CV)	0.85000 -	average score with SD of 0.02
TF-IDF with SVM	0.87712 -	Score from original training and test data

Table 1: Experimental Results

4 Draft of Results

Initially, we had 9,768 training data samples after pre-processing, of which 7,416 samples were from 'OFF' class and 2,352 samples from 'NOT' class. This uneven distribution was due to fewer offensive tweets being posted on social media than non-offensive ones. Additionally, Twitter occasionally filters out some offensive tweets and removes them from their database. We did a few experiments using SVM with TF-IDF and achieved results which are shown in Table 1. Here, CV stands for cross validation and SD is standard deviation. However, the test set (3,887 samples) used for evaluation purposes contained a 2:1 ratio of not offensive to offensive tweets. Next, in order to experiment with a similar proportion of classes in both train and test data, we temporarily removed 2500 tweets from our train data. As a result, train data demonstrated the same 2:1 ratio of non-offensive to offensive tweets and yielded a F1-macro of 0.88197 on the test set. In our opinion, collecting more data might increase the classification performance of our proposed models as our results were slightly lower than some other SE papers.

5 Plan for completing the project

Activities	Completion time	Estimated dates
Training models	2 weeks	Until 08.11.2021
Hyperparameter tuning on dev set	1 week	Until 15.11.2021
Evaluating metrics on test set	1 week	Until 22.11.2021
Final report	2 weeks	Until 06.11.2021

Table 2: Project Completion Plan

6 Repository URL

https://github.com/UOFA-INTRO-NLP-F21/f2021 - proj-arif-shahriar-anik

References

Hwijeen Ahn, Jimin Sun, Chan Young Park, and Jungyun Seo. 2020. Nlpdove at semeval-2020 task 12: Improving offensive language detection with cross-lingual transfer. *arXiv preprint arXiv:2008.01354*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Apurva Parikh, Abhimanyu Singh Bisht, and Prasenjit Majumder. 2020. Irlab_daiict at semeval-2020 task 12: Machine learning and deep learning methods for offensive language identification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2006–2011.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. Solid: A large-scale semi-supervised dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 915–928.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

D Thenmozhi, Nandhinee Pr, S Arunima, and Amlan Sengupta. 2020. Ssn_nlp at semeval 2020 task 12: Offense target identification in social media using traditional and deep machine learning approaches. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2155–2160.

Jake VanderPlas. 2016. *Python Data Science Handbook*. Oreilly Media Inc.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.