Text Classification in Tweets for Hate Speech, Offensive Speech and Sentiment Analysis*

Atta, Saad
Department of Mathematics
University of Essex
Colchester, UK
sa20091@essex.ac.uk

***Abstract*—With the ease of access to internet usage, people from different cultural and educational backgrounds are using social media platforms to express their opinions and feelings on different topics trending around the world. This increases the chances of conflict of interests and argument among people and thus it is necessary to identify hate speech to avoid hate spread in the society. In this paper I have worked on three separate datasets from twitter to identify hate speech/offensive speech and sentiment analysis. I have reviewed some of the related work that has been done in this domain along with detailed analysis of the datasets. Both descriptive and statistical analysis are included in this paper to provide a clear picture of the problem. One of the core tasks when solving an NLP problem is data preprocessing. I have briefly mentioned the preprocessing techniques that I used to clean to the twitter dataset and the feature extraction techniques I applied before feeding the data to train model. I have evaluated various Machine Learning and Deep Learning algorithms to address this problem and I included the performance comparison in this paper as well. I have achieved an f-1 macro averaged score of 0.58, 0.71, and 0.55 for hate speech detection, offensive speech identification and sentiment analysis respectively.**

I. Introduction

Now a days people use social media as a necessary tool to communicate, share memories and express their feelings with each other. Twitter is one of the popular social media tool that allows users to share their thoughts and views on what is happening in the society. Some people misuse it for spreading hatred against an individual or group of people and create divisions in society. These type of activities should be monitored precisely as they harm people interests and tends to begin hate crimes. Analyzing tweets data set is a challenging task because of its abrupt nature. 90% of Twitter users make their profile public [4]. Different people use different language styles to express their opinions. This depends on certain factors including ethnicity, demographics and educational background. Researchers in Ref 1 have shown correlation between number of racial hate crimes and number of racist tweets. They studied hate speech along with hate crimes in 100 cities across United States. They showed that the ratio of discrimination tweets has positive relationship with the number of race and ethnicity based hate crimes in the city. Researchers in Ref 2 showed contribution of social media against minorities in US with respect to Trump Political Campaign. They investigated the role of social media in enabling hate crimes with respect to Twitter usage using difference-in-difference approach. According to [3] 86% of the people in 18-29 age group have witnessed harassment online and 24% of them have experienced emotional illness due to that harassment. 10000 tweets everyday contains racial, ethnic or religious hate terms in English tweets [5].
The automation of detecting hate/offensive speech can be done using Natural Language Processing techniques and Machine Learning models. NLP allows the extraction of features from the raw text that could be useful for sentiment analysis. In this paper I have discussed how NLP can be used along with Machine Learning and Deep Learning models to better predict whether a tweet is hate/offensive or of contains negative sentiment. I have analyzed the top most frequent words used in each of the data set. The task was challenging because the data set is more biased towards non-hate, not offensive and neutral sentiment tweets. The structure of this paper follows by mentioning the related work done in this domain following by the methodology I propose to correctly classify tweets as hate, offensive or negative. I have included the graphs and visualizations for better understanding of data. I have supported my work with experimental evaluations and compared results with other researcher's work in the respective domain.

II. Literature Review

BERTweet [6] and TweetEval [7] have shown a very detailed comparative study of performance of different neural network architectures and machine learning approaches to this problem. BERT (Bidirectional Encoder Representations from Transformers) is a language model that applies bidirectional training of Transformer to language modeling. The paper was published by researchers at Google AI Language. BERTweet uses two training strategies i.e. Masked LM (MLM) and Next Sentence Prediction in parallel to overcome context learning challenges. The bidirectional approach makes the convergence slower as compare to directional approach but it outperforms left-right training after a small number of pretraining steps. TweetEval consists of seven multi-class tweet classification tasks. They selected RoBERTa [8] language model. As

compare to BERT it does not employ the Next Sentence Prediction loss. For evaluation they used macro-averaged F1 over all classes.

TABLE I
CURRENT BENCHMARK

Model	Hate	Offensive	Sentiment	Reference
BERTweet	56.4	79.5	73.4	BERTweet
RoBERTa-Retrained	52.3	80.5	72.6	TweetEval

Researchers in [9] used recurrent neural networks approach to address this problem. The dataset they used for conducting experiments contains three classes i.e. Racism, Sexism and Neutral tweets. The dataset contained 16k short messages from Twitter. The architecture they proposed consists of multiple classifiers based on LSTM (Long Short Term Memory) and it utilizes the tendency of user behavioral characteristics towards racist and sexist tweets. They evaluated model performance on 10-fold cross validation and calculated Precision, Recall and F-1 score. They achieved an F-1 score of 0.9320.

Researchers in [10] evaluated several Machine Learning Models on Twitter dataset in predicting hateful, offensive or normal tweets. They used dataset from three different sources (two from crowdflower and one from github) and merged them. The dataset contains tweets with 3 classes Hateful, Offensive and Clean. They used ngram features from the text and trained models on the tfidf values from those features. N gram is the sequence of n words appearing in the corpus of documents. They applied general NLP techniques to clean the data i.e. lower casing, stop words, URLs, twitter handles and retweet symbols removal. They also applied porter stemmer for stemming of words. They achieved an accuracy of 93.4% using Naïve Bayes with alpha equals to 0.1.

Researchers in [11] worked on 16K tweets dataset annotated with sexist, racist and normal. They used deep learning approach to predict the class of a tweet. They proposed that tweet embeddings initialized with random vectors are trained on LSTM using back propagation and then the learned embeddings were fed to train a Gradient Boosted Decision Tree Classifier. Using this approach they achieved an F1-score of 0.930. The embedding size used was set to 200.

Researchers in [12] proposed a two step approach in identifying hateful tweets using convolutional neural networks. In first step they classify whether a tweet is abusive or not. In the second step they classify the whether the abusive tweet is sexist or racist. They showed that using a HybridCNN which takes both word and characters as input could perform very well on large datasets. HybridCNN is a variation of WordCNN that has two input channels. HybridCNN can capture features from both character and word inputs. They achieved an F-1 score of 0.734 for abusive language detection and 0.950 for the classification of Sexist/Racist.

Researchers in [13] proposed a Convolutional Neural Approach to classify hate speech. They trained four CNN models using different word embedding techniques and evaluated their performance using 10 fold cross validation. The four word embedding methods are Random Vectors, Word2Vec, Character n-grams, and Word2Vec + Character n-gram. They achieved an F-1 score of 0.7829 using Word2Vec approach.

III. METHODOLOGY

A. Dataset Preprocessing

I worked on three datasets from twitter. Each dataset comprises of three text files for train, validation and test sets. For each of the set a labels text file is also provided. The files contains data separated by new lines. The number of observations for each dataset can be shown in the table below

TABLE II
DATASET BREAKDOWN

Dataset	# of Tweets	Classes
Hate Speech	(9000 train),(1000 val),(2970 test)	Not-Hate/ Hate
Offensive Speech	(11916 train),(1324 val),(860 test)	Not -Offensive/ Offensive
Sentiment Analysis	(45615 train),(2000 val),(12284 test)	Negative/ Neutral /Positive

The average number of words in tweets for each training dataset can be shown in the table below:

TABLE III
DATASET ANALYSIS

Dataset (Train	Average # of Words in a Tweet
Hate Speech	21
Offensive Speech	22
Sentiment Analysis	19

For preprocessing of data I used the following procedure:

- I expanded the contraction of words i.e. converted “are’nt” -¿ “are not”, “can’t ‘ve”-¿”cannot have” etc.
- Converted all the tweets text to lower case.
- Upon analyzing the dataset, I found out that some tweets contain emoticons and emojis so replaced them with their textual representations.
- Removed twitter handles
- Removed hash from hashtags in tweets.
- I removed punctuations, special characters and stop words from the dataset using spacy library.
- Removed numbers and digits from the tweets
- Removed URLs and emails.
- Removed Named Entities from the tweets to avoid biasness.
- Performed lemmatization to words to convert them to their root form.
- Removed extra spaces from the tweets.

Fig. 1. Wordcloud for Hate Tweets

Fig. 2. Wordcloud for Offensive Tweets

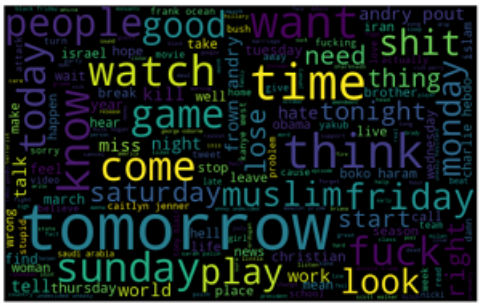

Fig. 3. Wordcloud for Negative Tweets

B. Machine Learning Models

The dataset is labelled with classes so that we can apply supervised machine learning algorithms for classification task. However the data needs to be in some numerical representation in order to train a model. This process is called Feature Extraction . I evaluated the performance of machine learning models using several feature extraction techniques and compared the results.

1) Feature Extraction: Bag of Words is the simplest feature extraction technique that store the word count in text or word frequency. It creates a matrix with all the words in the corpus as columns and each row as a document. For each row it stores the frequency of that word in the document.

TF-IDF is an important feature extraction technique that assigns weight to words based on the whole corpus. That means that a word commonly occurring in all the documents will have lesser weight. It shows the importance of word to a document.

Word2Vec is built on neural network architecture and used to learn word embeddings. Word embedding allows capturing of context of word in a document. Words having similar context would lie close to each other in the vector space of word embeddings. In our use case I calculated the word vector for all the words in the tweet and then took a mean of them as a single tweet vector.

2) Feature Selection for BOW and TF-IDF: For Bag of Words and TF-IDF approaches, I also performed feature selection to reduce the matrix dimensionality. I performed Chi-square test to determine whether a feature (word) is important in determining the class of tweet. I kept only the features with p-value greater than 0.95 from Chi-Square test.

3) Model Selection: I trained Logistic Regression, Random Forest, SVC and XGBoost Classifiers on the datasets and evaluated their performance using F-1 score (macro averaged). I used the hyperparameters using GridSearchCV. The final list of hyperparameters for mentioned machine learning algorithms can be shown in fig below:

Model	Hyperparameters
Logistic Regression	penalty='l2', C=1.0, solver='lbfgs'
Random Forest	n_estimators=100, max_depth=None, criterion='gini'
SVC	Default
XGBoost	Default

IV. Results

TABLE IV
HATE SPEECH MODEL EVALUATION

Model	BOW	TF-IDF	Spacy
Logistic Regression	0.51	0.52	0.56
Random Forest	0.45	0.45	0.58
SVC	0.54	0.54	0.56
XGBoost	0.49	0.50	0.55

TABLE V
OFFENSIVE SPEECH MODEL EVALUATION

Model	BOW	TF-IDF	Spacy
Logistic Regression	0.68	0.66	0.65
Random Forest	0.69	0.71	0.65
SVC	0.67	0.69	0.65
XGBoost	0.66	0.68	0.68

TABLE VI
SENTIMENT ANALYSIS MODEL EVALUATION

Model	BOW	TF-IDF	Spacy
Logistic Regression	0.53	0.52	0.55
Random Forest	0.46	0.47	0.47
SVC	0.52	0.52	0.51
XGBoost	0.47	0.47	0.53

V. Discussion

From the results that i showed in the previous section, it is evident that some feature extraction method changes the performance of model drastically. Moreover for offensive

speech identification and sentiment analysis, instead of Machine Learning Model, a deep learning model would be more appropriate because the datasets have enough observations to train a deep neural model. Bi-directional LSTM, CNN or transformer based models like BERT could perform much better for sentiment analysis.

VI. Conclusion

After working on this project, I have realized that preprocessing of text is a necessary and core part of solving a NLP classification problem. I have done a detailed preprocessing of tweets however it seems that sentiment analysis dataset require more of preprocessing because the words I am using to create embedding vector are common in different classes and thus provide less meaning. A lot of work has been done in twitter dataset analysis and still researchers are working on improving the classification accuracy. With the introduction of transformers and use of deep learning techniques in language processing domain, i believe the classification accuracy will improve more in future.

References

[1] Kunal Relia, Zhengyi Li, Stephanie H. Cook, Rumi Chunara, "Race, Ethnicity and National Origin-based Discrimination in Social Media and Hate Crimes Across 100 U.S. Cities" New York University

[2] Karsten Muller and Carlo Schwarz, " From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment "

[3] MAEVE DUGGAN, "Online Harassment 2017" PEW Research Center

[4] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, J Niels Rosenquist "Understanding the Demographics of Twitter Users", Northeastern University, Technical University of Denmark, Harvard Medical School

[5] Jamie Bartlett, Jeremy Reffin, Noelle Rumball, Sarah Williamson, "ANTI-SOCIAL MEDIA", DEMOS UK

[6] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen, "BERTweet: A pre-trained language model for English Tweets" VinAI Research, Vietnam; Oracle Digital Assistant, Oracle, Australia; NVIDIA, USA

[7] Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves and Luis Espinosa-Anke, " TWEETEVAL: Unified Benchmark and Comparative Evaluation for Tweet Classification " Snap Inc., Santa Monica, CA 90405, USA

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach" Paul G. Allen School of Computer Science Engineering, University of Washington, Seattle, WA

[9] Pitsilis, Georgios & Ramampiaro, Heri & Langseth, Helge, "Effective hate-speech detection in Twitter data using recurrent neural networks. Applied Intelligence.", in press

[10] 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", Department of Computer Engineering, Maharashtra Institute of Technology, Pune, Pune, India

[11] Pinkesh Badjatiya1, Shashank Gupta1, Manish Gupta1, Vasudeva Varma1, "Deep Learning for Hate Speech Detection in Tweets", IIIT-H, Hyderabad, India, Microsoft, India

[12] Ji Ho Park and Pascale Fung, "One-step and Two-step Classification for Abusive Language Detection on Twitter", Human Language Technology Center, Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology

[13] Björn Gambäck and Utpal Kumar Sikdar, "Using Convolutional Neural Networks to Classify Hate-Speech", Department of Computer Science, Norwegian University of Science and Technology, NO–7491 Trondheim, Norway

[14] Thomas Davidson, Dana Warmsley, Michael Macy, Ingmar Weber, "Automated Hate Speech Detection and the Problem of Offensive Language", Department of Sociology, Cornell University, Ithaca, NY, USA, Department of Applied Mathematics, Cornell University, Ithaca, NY, USA, Department of Information Science, Cornell University, Ithaca, NY, USA, Qatar Computing Research Institute, HBKU, Doha, Qatar