Detection of Cyberbullying in Twitter Data Using Machine Learning Techniques

Shahina K M

Abstract*—Analyzing comments in online interactions poses an important role in todays technological world. Although the social media plays a significant role in communication, it spreads cyberbullying among the young generation. Usage of aggressive and distorting words in social media is turn into a trend in nowadays. This will constitute a culture with dishonor and adverse communication in cyber world. so, intelligence systems based on different algorithms are emerged to classify this social media contents. This paper focused on analyzing and experimenting feature extraction and detection of cyber bullying in twitter messages with the help of Natural Language Processing tools and different Machine learning algorithms. Four feature extraction methods including Bag of words, TFIDF, doctovec and wordtovec are applied on the data set to create the feature set and then different classification methods are performed on these features. The classification methods include Logistic Regression, Support Vector Machine, Random Forest, and XGBoost. The result shows that XGBoost model on word2vec features has outperformed all the other methods. Machine learning algorithms for classification is implemented here using anaconda python distribution.*

Keywords — *Bag of words, TFIDF ,doctovec, wordtovec, Random Forest, Logistic Regression*

I. INTRODUCTION

The number of social media users are increasing day by day in an exponential manner. According to google, India has worlds largest number of Facebook Users with over 195 million users, overtaking US by over 4 million subscribers. Where the key Facebook Users from India are aged between 18-24. These online platforms provide better opportunities for communication and discussions. Messages sharing on social medias are spreading across the world within seconds. When we are considering the other side, the usage of aggressive words and insulting comments on cyber world is also increasing day by day and these messages too spreading spontaneously across the world. This will badly affect healthy communications and discussions. As the key users on social medias are teenagers, they are using insulting comments and harassing words without any guilty and they are not bothering about the consequences of this cyberbullying.

In this scenario researchers have given a serious attention in the area of detection and blockage of cyber bullying. So, different natural language techniques along with machine learning algorithms are developed for the detection of aggressive and offensive languages in social media. This paper is focused on analyzing and detecting offensive languages in twitter messages using different feature extraction methods and machine learning algorithms. The same can be applied on any text data set extracted from social media.

II. LITERATURE SURVEY

Over the past few years several techniques have been proposed for the detection and classification of cyber bullying or offensive languages on cyber media including twitter, face book and you tube etc. Majority of these works are based on NLP and machine learning basics using different tools like Weka and RapidMiner.

Cynthia et al. [1] detect cyberbullying on the comments from social media. The data collected from the social networking site Ask.fm is annotated by a fine-grained method and two type of lexical features including bag of word feature and polarity feature are used. The extracted features are classified to two categories as bullied and non-bullied using SVM classifier. The classification yields an accuracy of 53.82% when the model was applied to the original corpus were the distribution of the positive posts was left unchanged.

Despoina et al. [2] proposed a principled and scalable approach to detect bullying and aggressive behavior on Twitter. The robust methodology combines sentiment detection using an unsupervised algorithm. Large feature sets are extracted from the twitter data based on user, text and network features. The whole classifications are performed using the Weka math lab tool for machine learning.

Chris et al. [3] focused on automatic cyberbullying detection in social media text by analyzing posts written by bullies, victims, and bystanders of online bullying. They performed a fine grained annotation of a training corpus for English and Dutch followed by a series of binary classification experiments to determine the feasibility of automatic cyberbullying detection.

Ying Chen et. al. [4] proposed a content and user-based mechanism to detect offensive content and identify potential offensive users. This method is based on Lexical Syntactic Feature (LSF) architecture and were a hand-authoring synctic rule is introduced to incorporate a user's writing style, structure and specific cyber bullying content as features to predict the user's potentiality to send out offensive content.

III. METHODOLOGY

Our approach to detect offensive and bullying languages on Twitter, as summarized in Figure 1, involves the following steps: (1) data collection and labelling, (2) preprocessing and cleaning of tweets, (3) Feature extraction using different techniques (4) Model building and learning (5) prediction and analysis of result.

A. Data collection

Data collection is the first step towards detecting offensive languages in twitter data. Twitter provides one percentage of its total tweets for public using Twitter Streaming API. Here the data set consist of 49159 of rows indicating the no of tweets and three columns representing id, label and tweet content. Each tweet is manually labelled to zero indicating the message is not cyberbullying or labelled to one means bullying. The entire data set is divided into two,70% for training and 30% for testing.

B. Data Preprocessing and Cleaning

Preprocessing of collected tweet data is an essential step as it is directly used for mining. Extracting information from tweet is easier in processed form and also speedup machine learning algorithms. The main contribution of this stage is to eliminate noise and inconsistent data.

- The twitter data contains @user handles and which does not give any sentimental information about the tweet. Preprocessing removes all this twitter handles as it does not add much value.
- Punctuations, numbers, and even special characters does not provide any useful information about the nature of the tweet. So, all this irrelevant information is eliminated from the data set.
- Short words like is, the, all, his and her etc. also removed as it doesnt convey much information.
- The cleaned tweet dataset is then tokenized. Tokens are individual terms or words, and tokenization is the process of splitting a string of text into tokens.
- Stemming is performed over this tokenized data set. Stemming is a rule-based process of stripping the suffixes (ing, ly, es, s etc) from a word.

C. Sentiment Visualization.

Sentiment distribution across the test data set is visualized using word cloud. Word cloud is a visualization tool wherein most frequent words appear in large size and the less frequent words appear in smaller sizes.

D. Features Extraction

Accuracy of classification algorithms are heavily dependent on feature set. So, feature set is constructed from the preprocessed data by the use of different feature extraction technique. Here four different extraction methods, Bag-of-Words, TF-IDF, Doctovec and Wordtovec are used and all of these methods return different feature sets.

a) Bag-of-Words Features

Bag-of-Words is a method to represent text into numerical features. Consider a corpus called C of D documents d1, d2..dD and N unique tokens extracted out of the corpus C. The N tokens will form a list, and the size of the bag-of-words matrix M will be given by D X N. Each row in the matrix M contains the frequency of tokens in document D(i).

D1: He is a lazy boy. She is also beautiful.

D2: Smith is a lazy person.

The list created would consist of all the unique tokens in the corpus C=[He,She,Boy,Smith,Person]

IV. TABLE 1
B AG - OF -WORDS EXAMPLE FEATURE SET

	She	*She*	*Lazzy*	*Boy*	*Smith*	*Person*
D1	1	1	2	1	0	0
D2	0	0	1	0	1	1

Now the columns in the matrix can be used as features to build the classification model. Bag-of-Words features can be easily created using sklearns Count Vectorizer function.

b) TF-IDF Features

TF-IDF is based on frequency of tokens in the entire corpus instead of a single tweet. TF-IDF works by assigning lower weights to common words while giving importance to words which are rare in the entire corpus but appear in good numbers in few documents.

- TF = (Number of times term t appears in a document) / (Number of terms in the document).
- IDF = log(N/n), where, N is the number of documents and n is the number of documents a term t has appeared in.
- TF-IDF = TF*IDF

c) Word2Vec

Word2Vec is a computationally efficient method for word representation from raw text. It has two approaches namely Continuous Bag of Words

(CBOW) model and Skip-gram model. Where CBOWŁmethod predicts target words from source context words and Skip-gramŁmethod predicts source context words from target words. Basic implementation of Word2Vec is done with genism library. The similarity property of Word2Vec states that similar words tend to have similar vectors. In detail the similarity between two words correlates with the cosine similarity between those words vectors also.

d) Doc2Vec

Doc2vec is an unsupervised algorithm intended to generate vectors for sentence/paragraphs/documents. This algorithm is an extension of genisms original Word2Vec method which will generate vectors for words. The vectors generated by doc2vec can be used for tasks like finding similarity between sentences/paragraphs/documents. And Doc2Vec is based on two algorithms known as distributed memory (dm) and distributed bag of words (dbow).

E. MODEL BUILDING

Four kinds of classification algorithms are used for building machine learning predictive model with four feature sets. They are 1) Logistic regression: -predicts the probability of occurrence of an event by fitting data to a logit function. 2) SVM:-based on finding an upper plane that divides the feature set into two categories using kernel function. 3) Random forest:-builds multiple decision trees and merges them together to get a more accurate and stable prediction. 4) XG boost: - stands for Extreme Gradient Boosting. All the work mentioned here are carried out in python scikit learn. Accuracy is calculated for all models and which shows that XGBoost model on word2vec features has outperformed all the previous models.

V. TABLE 2
CLASSIFICATION RESULT

	Bag-of -words		**TF-IDF**		**Word2Vec**		**Doc2Vec**	
	precision	F1 score	precision	F1 score	precision	F1 score	precision	F1 score
SVM	.50	.55	.51	.54	.61	.65	.20	.21
Logistic regression	.55	.59	.56	.59	.50	.55	.05	.07
Random forest	.55	.60	.56	.59	.50	.55	.05	.07
XG boost	.51	.55	.51	.59	.68	.70	.35	.37

VI. CONCLUSIONS

Ultimate goal of this work is to find out an automatic prediction system for cyberbullying in twitter data. Data setis collected from twitter streaming API and the collected data is preprocessed using natural language techniques. Then the data set is tokenized and stemming is performed over this data. Here four different data sets are prepared from the actual data set using Bag-of -words, TFIDF, word tovec and doctovec. Then different machine learning prediction algorithms like SVM, Logistic Regression, Random Forest and XGboost are applied over these feature sets and the tweets are classified into bullied or non bullied. The result shows that XGBoost model on word2vec features has outperformed all the previous models.

REFERENCES

[1] Cynthia Van Hee, Els Lefever, Ben Verhoeveny, Automatic Detection and Prevention of Cyberbullying. HUSO 2015- The First International Conference on Human and Social Analytics.

[2] Despoina Chatzakouy, Nicolas Kourtellisz, Jeremy Blackburnz, Mean Birds: Detecting Aggression and Bullying on Twitter, arXiv:1702.06877 [cs.CY].

[3] Cynthia Van Hee1, Gilles Jacobs1, Chris Emmery2, Bart Desmet1, Els Lefever1, Ben Verhoeven2, Guy De Pauw2, Walter Daelemans2, and Veronique Host, Automatic Detection of Cyberbullying in Social Media Text.

[4] Ying Chen, Sencun Zhu, Yilu Zhou, Heng Xu, Detecting Offensive Language in Social Media to Protect Adolescent Online Safety, ACM, 2012

[5] K. Van Royen, K. Poels,W. Daelemans, and H. Vandebosch, Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability, Telematics and Informatics, vol. 32, 2015, pp. 8997, ISSN: 0736-5853