Intelligent Text Mining to Sentiment Analysis of Online Reviews

P.Asritha[1], P.Prudhvi Raja Reddy [1], C.Pushpitha Sudha[1], Neelima.N[1]

[1]Department of Electronics and Communication Engineering
Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India.
paluruasritha@gmail.com,prudhvirajareddy4512@gmail.com
sudhachatakonda@gmail.com, n_neelima@blr.amrita.edu

Abstract. In the prevailing days social networking sites helps people to connect easily across the world and gain knowledge and also share their interests. But, unfortunately in some cases these sites became a platform for cyber bullying. Cyber bullying is the act which causes emotional and psychological distress leading to depression, anxiety, fear and low selfesteem to the victims. Cyberbullying can be elucidated as usage of digital communication typically by sending messages to threaten, defame, harass or intimidate someone. Common social media platforms like twitter, facebook, instagram are exposed to cyberbullying which has become very common now-a-days. This can be reduced to an extent if such intimidating messages or comments are segregated. The process of classifying a sentence whether it is positive, negative or neutral is known as sentiment analysis. It helps in determining emotional tone behind a sentence. To classify these intimidating messages this paper proposes a hybrid classifier approach which classifies reviews into positive or negative. Experimental results show that the accuracy of the classifier for considered dataset is 89.36%.

Keywords: Cyberbullying, sentiment, classifier, multinomial.

1 Introduction

The act of threatening, forcing or violently dominating others is drafted as bullying. Cyber bullying is bullying by means of electronic media.Cyber bullying involves threats, posting rumours, sexual remarks or vituperative labels. Teenagers are becoming prone to such harassment. According to survey nearly 8 out of 10 individuals have gone through various types of cyber bullying in India[7][9]. Out of these many-faced online abuses and insults as well, and many

ICASISET 2020, May 16-17, Chennai, India

DOI 10.4108/eai.16-5-2020.2303907

were subjected to false rumours and gossips which leads to degradation of their image. Cyber bullying leads to emotional responses which includes frustration, depression and aggressiveness. Sometimes cyber bullying becomes more menacing than the bullying that occurs due to physical proximity. The survey found that large percent of school students in bangladesh are becoming victims of cyber bullying[8][10]. Also, one in six teenagers are being embroiled in the bullying. As number of online users are raising, persons embroiled with it raises multiplying the victim count. Prior, the reviews are labelled based on the content that is known to be text categorization (classification). Text classification is a fundamental and customarily task in a supervised machine learning. This can be performed using python, NLTK and scikit-learn. The detection method allows us to recognize presence of cyber bullying content in reviews. Cyber bullying differs in the varying altitude of strength of bullying, age of the person making bully and other things like social position and educational qualification of the person being bullied.

In this paper, identification of cyber bully and distinguish of twitter comments is proposed. This identification is forwarded by using naïve bayes classifier and support vector machine classifier. Applications include spam filtering, sentiment analysis and email routing.

2 Literature Survey

Text classification is the process of differentiating the dataset into categories depending upon the content present. Sundus Hassan & Muhammed Rafi[4] compared SVM and naïve bayes classifiers for Text Categorization with Wikitology as Knowledge enrichment. This paper implements a hybrid approach for text classification using data mining, and association rule to get the feature set for given text. Naïve bayes classifier is used on feature set and genetic algorithm is used for final classification. The accuracy found for the association rule-based decision tree is 87% and for the genetic algorithm is 68% [4].

There are different methods to increase accuracy of naïve bayes classifier for text classification. It is perceived that systems like n-grams, feature extraction and effectual negation handling with help of naïve bayes using mutual information leads to a raise in accuracy. A fast and accurate classification for sentiment is proposed by Vivek Narayanan, Ishan Arora, Arjun Bhatia in [3]. In this paper, classifier is built using naïve bayes model that contains linear training and testing time complexities. The dataset considered in this paper is taken from internet movie data base (IMDB) which contains movie reviews, the dataset contains

25000 reviews for training and testing each with both positive and negative reviews. The accuracy obtained when considered naïve bayes algorithm with laplacian smoothing for test set is 73.77%, when considered bernoulli naïve bayes accuracy is 83.66%, when considered bigrams and trigrams it is 85.20% and through handling negations it is 82.80% [3].

Cyber bullying activities cause mental health problems to people who are facing it. multinomial naïve bayes classifier is used to determine type of bullying which has types of harassing, rasicm and sexual harassments. Fuzzy logic is used to determine the strength of bully. The data set used is collected from the approach used by Akhter, Arnisha from facebook data [2]. It is shown that naïve bayes classifier has more accuracy than support vector machine model (SVM) and less run time. The accuracy for the SVM model is 76.38% and for naïve bayes is 88.89% [2]. Concept maps (CM) isanother approach that gives relationship among concepts. The concept maps approach is proposed by Zubrinic and Krunoslav [1],is used to compare results of naïve bayes and SVM classifiers. SVM is a hyperplane which separates the positive and negative examples from the dataset with maximum margin. To find the better classifier the parameter Fα is calculated along with precession and recall. The one with more Fα is chosen as better classifier [1].

3Methodology

This paper proposes a hybrid approach to find the sentiment involved in the given tweet (or) review. This helps the user to identify easily the negative reviews. The First step in text classification is collection of dataset. Dataset is collection of data that contains related or discrete items of data that is managed as whole entity. The model is edified with labelled dataset which contains both input, output parameters. In this paper, the dataset used has more than 1600 comments, consisting of both the categories bullying and non-bullying. The proposed methodology is shown in figure (1). The entire process is divided into multiple stages which are explained in further sections.

3.1 Pre-processing:

In general, the tweets (or) comments contains many comments which include special characters and punctuation marks as well. These special characters and punctuation marks do not contribute any role in text classification. So, the cleaning of the tweets is the first step which helps in removing the special characters, symbols and punctuation marks. Pre-processing allude to modifications applied to the dataset before passing it through an algorithm. Pre-

processing of data is done which helps to increase the ability of model to learn and achieve better results. Lowercasing is one of the most effective and simplest form of data pre-processing which is relevant to most of the text mining problems and also helps with consistency of predicted output. The most common task to reduce problem space is to change entire text into lower case.

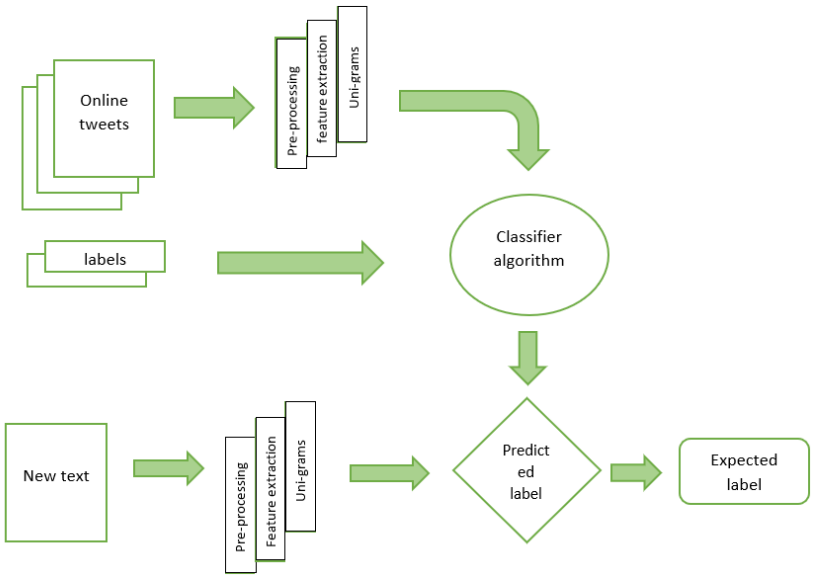

Fig. 1. Block diagram of the proposed approach

3.2 Tokenization and Lemmatization:

In tokenization a parser is incorporated in pipeline which breaks down text flow into words, symbols, phrases or other useful elements known as tokens. Extraction of individual words from a sentence is a main goal in tokenization. Tokenization converts text into tokens before transfiguring into vectors and also percolates unessential tokens. Then, lemmatization is done to trim the words and map it to its root form. It removes inflectional endings and return to dictionary form of words known as lemma. Words which have same core meaning are mapped to centre word or symbol. It helps to accelerate user's task and lower the noise. Further the stop words are removed. In this process words which have a less predictive power are removed. It helps in minimizing usage of storage space and saves computational time.

3.3 Feature Extraction:

Vectorization:

Vectorization is a process of transforming text into numbers. Counting number of times, a token has occurred is known as term frequency and countvectorizer works on term frequency. Normalization of occurrence of word with length of document is known as Term Frequency.

$$\text{TF(x)} = \frac{number\ of\ times\ a\ term\ x\ appears\ in\ a\ document}{number\ of\ terms\ in\ a\ document} \quad (1)$$

TF-IDF Vectorizer:

Term Frequency-Inverse Document Frequency implies the measure of importance of a word in a document. A quick run of this will sort all entries in dataset and help in eliminating great deal of inconvenience. Simplest way of representing text in numeric for is count of occurrence of word in entire document. TF-IDF is one of the familiar algorithms used to mutate text into meaningful delineation of numbers. Representation of text in numeric form helps in depicting significant traits and extracting features of text.

Inverse Document Frequency:

Term Frequency of certain words ('a', 'the', 'of') that are regular in documents might put down weights of meaningful words. To overcome this problem, term frequency is discounted by factor called inverse document frequency (IDF). It is measure of how isolated a term is. If the IDF score is more, then the term is considered to be more isolated.

$$\text{IDF(x)} = \log_e \frac{total\ number\ of\ documents}{number\ of\ documents\ with\ term\ x\ in\ it} \quad (2)$$

The product of Term Frequency with Inverse Document Frequency is known as TFIDF.

$$(TF\_IDF)score = TF * IDF (3)$$

3.4 Topic Models:

Unigram:

Single word or element in list of tokens is considered as Unigram. Unigram model is also known as bag of words model.

Bigrams:

Sequence of two elements that are adjacent in string of tokens, these elements can be either words, letters, syllables. Use of bigrams in text classification is more effective than usage of bag of words (unigrams).

3.5 Classification:

The prediction of the classifier is based on the models chosen. The experimental results show that multinomial with unigrams is more efficient than others. The extracted features are used in classifying the tweets into either positive tweet or negative tweet sentiment. In this paper two classifiers namely naïve bayes and support vector machine are used. The results provide the comparison analysis between the two classifiers used.

Naïve Bayes Classifier:

In machine learning, naïve bayes classifier is associated to family of probabilistic classifiers that applies bayes theorem with strong independent hypothesis between features.This classifier requires slight number of training set to deem features required for classification. Naïve bayes classifier can be used for both multiclass and binary classification. Performance of naïve bayes classifier is well in case of categorical input variables when compared to that of numerical variables. Naïve bayes classifier predicts probabilities of every class. The most likely class is the one with highest probability. This is called as maximum a posteriori.

Gaussian Naïve Bayes:

Gaussian naïve bayes deals with data which is continuous in which hypothesis is continuous values that are affiliated with each class are dispensed according to gaussian (or normal) distribution.

$$P(x_i/y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \qquad (4)$$

Where, x is the continuous data, μ_y is the mean of each class and σ_y^2 is the variance of each class.

SVM Classifier:

SVMs are supervised learning models which are related with learning algorithms which analyse data that is required for classification analysis. SVM is a classifier which is formally defined by separating hyperplane. Examples in this model are

represented in form of points in space which helps in separating different categories by a gap. SVM not only performs linear classification but also non-linear classification. Set of examples which are marked belonging to either of the category helps in building a model that categorizes a new example.

4. Results

The proposed hybrid approach involves selecting the unigrams with multinomial naïve bayes classifier for classification. The accuracy of the processed approach is shown in figure (2). The confusion matrix of the classifier shows that the sentiment is classified correctly as positive/negative. The proposed method also provides a provision to the user to enter a random tweet to test the classifier. Some of the comments/tweets are shown in figure (3). The output of the classifier for these random tweets which has different combinations of the sentiment involved is shown in figure (4). The comparison analysis of the sentiment classification with the existed method is tabulated in table 1. From table 1 it is clear that the proposed hybrid approach is more efficient for sentiment analysis. The accuracies of different classifiers are plotted as shown in figure (5).

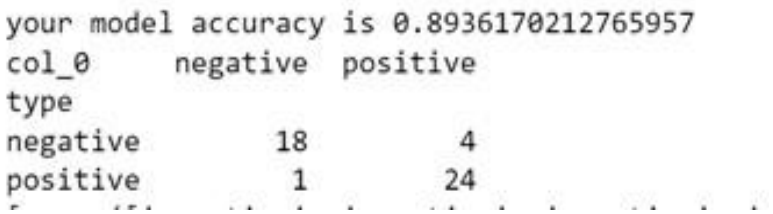

Fig. 2. Accuracy of the proposed classifier

Table 1. Comparison analysis of accuracy with existing approaches

Algorithm	Accuracy%
SVM(existing)	85.6%
Naive Bayes(Gaussian)	81.26%
Proposed Approach	89.36%

```
tweet1 = pd.Series(["she is not good"])
print(tweet1)
tweet2 = pd.Series(["she is not bad"])
print(tweet2)
tweet3 = pd.Series(["she is bad at studying"])
print(tweet3)
tweet4 = pd.Series(["she is good at singing"])
print(tweet4)
```

```
0    she is not good
dtype: object
0    she is not bad
dtype: object
0    she is bad at studying
dtype: object
0    she is good at singing
dtype: object
```

Fig. 3. Different tweets given at run time for testing the classifier

```
print("the tweet1 is:",model.predict(tweet_transformed1))
print("the tweet2 is:",model.predict(tweet_transformed2))
print("the tweet3 is:",model.predict(tweet_transformed3))
print("the tweet4 is:",model.predict(tweet_transformed4))
```

```
the tweet1 is: ['negative']
the tweet2 is: ['positive']
the tweet3 is: ['negative']
the tweet4 is: ['positive']
```

Fig. 4. The classifier output of the classifier for the tweets given at run time

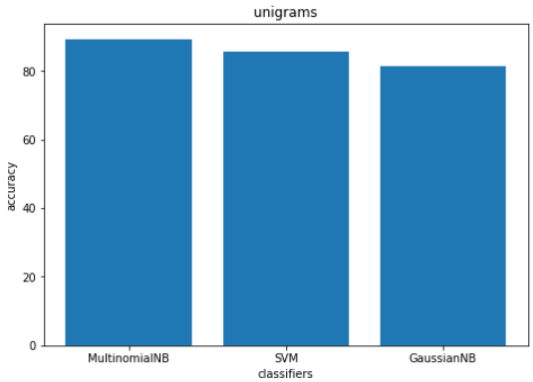

Fig.5. Plot of accuracy of different classifiers

5. Conclusion

With the advent of many innovations through technology, there is a gradual decrease in direct communication among people. At the same time, a surge is found in the communication through digital medium. Though digital medium connects people from various locations, the same is taken as advantage by others leading to cyberbullying. Disturbance created and executed by means of digital devices is cyberbullying. It effects mental stability of a person increasing stress, pressure leading to extreme activities. So by detecting the comments which are cyberbullying helps an individual overcome the mental stress. This paper implements a hybrid approach to determine the sentiment involved in the tweets posted by the individuals. From the results it is evident that the proposed hybrid approach which is a combination of unigrams with multinomial naïve bayes is more efficient in sentiment analysis.

References:

1. Zubrinic, Krunoslav, Mario Miličević, and IvonaZakarija. "Comparison of Naive Bayes and SVM classifiers in categorization of concept maps." International journal of computers, 2013, pp:109-116.
2. Akhter, Arnisha&Acharjee, Uzzal&Polash, Md.“Cyber Bullying Detection and Classification using Multinomial Naïve Bayes and Fuzzy Logic” International Journal of Mathematical Sciences and Computing, 2019, vol.4, pp:1-12.
3. Narayanan, Vivek, Ishan Arora, and Arjun Bhatia. “Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model.” Lecture Notes in Computer Science 2013, pp:194–201.
4. Hassan, Sundus& Rafi, Muhammad & Shaikh, M, “Comparing SVM and Naive Bayes classifiers for text categorization with wikitology as knowledge enrichment”, International Multi Topic Conference, 2012.
5. Madan.k, Bhanu Anusha.K, Pavan Kalyan.P, Neelima.N, “Research on different classifiers for early detection of lung nodules”, International Journal of Recent Technology and Engineering, July 2019, pp: 1037-1040.
6. J.V.V.S.N. Raju, P.Rakesh, Neelima.N, “Driver Drowsiness Monitoring System”, Smart Innovation, Systems and Technologies, (SIST vol.169), 2019, pp: 675-683.
7. İfflab.org/how-to-prevent-cyber-bullying-anti-cyber-bullying-laws-in-india/
8. Sarker, Supriya& Shahid, Abdur. (2018). “Cyberbullying of High School Students in Bangladesh: An Exploratory Study.”

K. Pradeep Mohan Kumar, M. Saravanan, M. Thenmozhi ,K. Vijayakumar, " Intrusion detection system based on GA-fuzzy classifier for detecting malicious attacks", wiley, Feb 2019.

9. J. Dafni Rose, K. Vijayakumar and S. Sakthivel, "Students' performance analysis system using cumulative predictor algorithm", Int. J. Reasoning-based Intelligent Systems, Vol. 11, No. 2, 2019.