International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

DEPRESSION DETECTION WITH SENTIMENT ANALYSIS OF TWEETS

Hemanthkumar M[1], Latha A[2]

[1]Student, Department of Computer Science and Engineering, Sapthagiri College of Engineering, Bengaluru
[2]Professor, Department of Computer Science and Engineering, Sapthagiri College of Engineering, Bengaluru

---

Abstract - *Nowadays risk of early death is increasing due to mental illness which is mostly caused due to depression. Depression creates suicidal thoughts causing serious impairments in daily life. Sentiment analysis is a hot topic that's been on research for decades, which intends to find the nature of text and classifies into positive, negative and neutral. In today's digital world lot of data is available for sentiment analysis of both image and text. The aim of this paper is to apply natural language processing on Twitter feeds for conducting emotion analysis focusing on depression. Based on a curated word list, the tweets are classified as positive or negative. Naive-Bayes and support vector machine algorithms are used for the classification of the text data obtained from twitter feed. The results are presented using the classification metrics like accuracy, precision, recall and confusion matrix.*

Key Words: Sentiment analysis, Naïve-Bayes, SVM, Lemmatization, confusion matrix.

1. INTRODUCTION

Modern society has made human life so busy, making it vulnerable to mental disorders like depression, anxiety etc. Depression can impair many facets of human life. The detection and prevention of depression is so difficult and has been the hottest topic of research since the past decade. The stresses, work schedules and many other things affects the normal life and can cause depression. Depression is regarded as the mental cancer by the researchers. Some of the feelings that can arise due to depression are loss of interest in activities, Suicidal thoughts, feeling of worthlessness or hopelessness and Worsened ability to think and concentrate.

Sentiment analysis is a hot topic that's been on research for decades, which intends to find the nature of text and classifies into positive, negative and neutral. Today's advanced technologies provide lot of tools for efficient retrieval, analysis and classification of text data into different classes. In today's digital world lot of data is available for sentiment analysis of both image and text through social media networks like Facebook, twitter etc., but using data from some social media networks raises concern of privacy, thus, twitter is the most suitable social media network for getting enough data and also to protect us from privacy laws.

This proposition aims to apply natural language processing on Twitter feeds for conducting emotion analysis focusing on depression. Individual tweets are classified as positive or negative, based on a curated word-list to detect depression tendencies. Numerous machine learning algorithms are available for the classification of the tweets, some of them are Naïve Bayes, Support vector machine algorithms, Random forest algorithms etc. But the documented experiments show better result for Naïve Bayes and support vector machine algorithms, the study is limited to these algorithms [1]. The performance of the algorithms can be measured over different metrics like accuracy, precision, recall and confusion matrix.

The Determination of the sentiment of an entire document is called as coarse level and fine level deals with attribute level sentiment analysis [2]. Since Twitter provides text that is at the sentence level, this study will be between these two levels of determination. In this paper we mostly deal with Tweets which are short message which are bound by a 140-character limit [3]. In this concise format, users express their emotions and feelings about ongoing happenings in their life and the world around them.

2. BACKGROUND

A. Dataset

The dataset is comprised of 43,000 tweets, collected from kaggle website. A ratio of 70:30 has been adopted for splitting the data collected into training and test dataset. The dataset consists of 2 columns namely sentiment and text. The text part contains the tweets and the sentiment part contains 0 or 1 where 0 represents negative meaning the tweet is linked to depression and 1 represents positive.

B. Naive Bayes Classifier

Naive Bayes Classifier implements Bayes theorem with a solid independence assumption [4], particularly, independent feature model. Bayes Theorem works on conditional probability which finds out the probability of an event given that some other event has already occurred. It predicts the conditional probability of a class given the set of evidences and finds the most likely class based on the highest one. A naive Bayes classifier is a famous and popular technique because it is very fast approach and gives a high accuracy [5]. The equation of the conditional probability is defined as follows

$$P(xi|c) = \frac{\text{Count of Xi in document of class C}}{\text{Total number of words in document of class c}}$$

Where,

xi is a given term
c is a predefined class label
In this proposition, Multinomial Naïve Bayes algorithm is being used since it is most efficient for multi-class prediction [6].

C. Support Vector Machine

Support Vector Machine(SVM) is a supervised learning algorithm that analyses the data and recognizes patterns [6]. Given an input set, SVM classifies them as one or the other of two categories. SVM can deal with both linear and non-linear classification.

With kernel trick, it can efficiently perform non-linear classification. It does so by mapping the input set to high dimensional feature space. The types of kernel include polynomial, Gaussian radial basis function (RBF), Laplace RBF kernel, Hyperbolic tangent kernel and Sigmoidal kernel. Construction of hyper plane is employed by SVM for the classification.

The goal of a text classification system in this paper is to determine whether a given tweet belongs to a set of predefined categories [7]. An optimal SVM algorithm for text classification does this via multiple optimal strategies.

3. METHODOLOGY

The workflow in this paper is divided into three phases

i. Pre-processing
ii. Training
iii. Testing

3.1. Pre-processing

The majority of the tweets usually can be divided into 3 parts, not specifically in the same order.

- The first part contains the people whom the tweet is intended to or who is it referencing. This is usually denoted by "@".
  ex: @Hickman.
- The second part contains the actual message. This is what the user actually wants to convey.
- The third part contains the hashtag denoted by "#". This is usually to categorize the tweets. This can also be used by others to find tweets related to the particular content.
  ex: #SupportWHO.

The tweets are pre-processed to filter the first part and third part since they do not hold very less to zero significance in sentiment analysis, later, the remaining message part of the tweet is pre-processed further for obtaining useful keywords which accounts much significance in identifying the emotions.

The following steps are carried out for further pre-processing.

Emoji Extraction: Since twitter users express their feelings along with emoticons, emoticons play vital role in identifying the sentiment of the tweet.

Hyperlink Removal: Hyperlinks can be considered as a noise in the tweets whose presence degrades the quality of the data, thus, they are removed.

Slang substitution: Efficiency of the model can be increased with substitution of full forms of abbreviations like LOL, BRB etc., which provides more key words for the model.

Timestamp removal: Timestamps can be in various forms like 10:30 AM, 10:30:22 etc., identifying and removing them is an important task.

Digits removal: Even digits do not hold much of the significance; thus they need to be removed.

Symbols removal: Unwanted symbols which do not form any meaningful emoticons need to be removed to increase the quality of the data.

Spelling correction: Most of the tweets contains lot of misspelled words, hence to create an efficient model spelling correction is very important. Spelling correction is divided into 2 processes,

a. Shortening
b. Correction

a. Shortening

Sometimes, to emphasize feelings twitter users repeats characters, which results in misspelling of the word, these words are shortened by reducing repetition of every characters to maximum of 2 letters.
Ex: Haaaappppyyyyy is reduced to Haappyy

b. Correction

This process substitutes correct spellings for misspelled words. It is not only effective in correcting misspelled words by users but also to correct words which were obtained in shortening process. Thus, Haappyy is substituted with Happy.

Proper nouns removal: Proper nouns are the name of specific person, place, things or ideas like shobha, Kanchi, Pacific Ocean etc., These are removed since the names do not contribute much to the sentiment analysis.

Lemmatization: Lemmatization is the process of mapping words to its lemma, i.e., the root word based on its meaning. It is better than stemming since stemming simply removes the suffixes of the words. Ex: The word better is substituted with word good after lemmatization thus process of extracting keywords becomes efficient.

Stop words removal: Stop words are most commonly used words which are intended to be ignored since their presence does not increase efficiency but has the potential to decrease it. Words like "is", "are", "want" etc., are considered as stop words. Stop words are not universal, the context decides the stop words thus identifying set of effective stop words is an important task.

The below diagram shows the pre-processing for the tweet "@raki it was 2 boring until I met you at http://www.friendslist.com at 10.30 :) LOL. I jussssttttt wish we meet at Paris this week #LOVELIESTDAY !!!!!!!!!!!!!!!!!!!!!!!". Assuming the stop words list contains at, I, it, this, we, was, you.

SENTENCE EXTRACTION

(Returns "it was 2 boring until I met you at http://www.friendslist.com at 10.30 :) LOL. I jussssttttt wish we meet at Paris this week !!!!!!!!!!!!!!!!!!!!!!!".)

↓

EMOJI EXTRACTION

(Returns "**:)**")

↓

HYPERLINK REMOVAL

(Returns "it was 2 boring until I met you at at 10.30 LOL. I jussssttttt wish we meet at Paris this week !!!!!!!!!!!!!!!!!!!!!!!".)

↓

SLANG SUBSTITUTION

(Returns "it was 2 boring until I met you at at 10.30 Laughing out loud. I jussssttttt wish we meet at Paris this week !!!!!!!!!!!!!!!!!!!!!!!".)

↓

TIMESTAMP REMOVAL

(Returns "it was 2 boring until I met you at at Laughing out loud. I jussssttttt wish we meet at Paris this week !!!!!!!!!!!!!!!!!!!!!!!".)

↓

DIGITS REMOVAL

(Returns "it was boring until I met you at at Laughing out loud. I jussssttttt wish we meet at Paris this week !!!!!!!!!!!!!!!!!!!!!!!".)

↓

SYMBOLS REMOVAL

(Returns "it was boring until I met you at at Laughing out loud I jussssttttt wish we meet at Paris this week ".)

↓

SPELLING CORRECTION

(Returns "it was boring until I met you at at Laughing out loud I just wish we meet at Paris this week ".)

↓

PROPER NOUNS REMOVAL

(Returns "it was boring until I met you at at Laughing out loud I just wish we meet at this week".)

↓

LEMMATIZATION

(Returns "it was bore until I meet you at at Laugh out loud I just wish we meet at this week".)

↓

STOP WORDS REMOVAL

(Returns "bore until meet Laugh out loud just wish meet week".)

Thus final useful keywords extracted are bore, until, meet, laugh, out, just, wish, meet, week. Similar process is applied for all the tweets to obtain useful set of keywords.

3.2. Training

The training process has two tasks,

a. Creation of bag of words model
b. Creation of predictive model

a. Creation of bag of words model

The bag-of-words method is a simplifying representation used in natural language processing [8] and information retrieval. In this model, a text like tweets is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity. The bag-of-words model is commonly used in methods of document classification where

the (frequency of) occurrence of each word is used as a feature for training a classifier.
Ex. Joey likes food. Joey likes games.
The bag of words will be
["Joey":2," likes":2," food":1," games":1].

The Bag-of-words model is mainly used as a tool of feature generation. After transforming the text into a "bag of words", we can calculate various measures to characterize the text. The most common type of characteristics, or features calculated from the Bag-of-words model is term frequency, namely, the number of times a term appears in the text.

b. Creation of Predictive model

Two Predictive models are created using naïve Bayes and support vector machine algorithms which were discussed afore. The data file will have two columns with one column representing either the tweet is positive or negative and other will contain the actual tweet. The column with tweets is pre-processed. The pre-processed data is used to create bag words model which will be used as training data to build the predictive models.

3.3. Testing

In testing phase, the 30% of data which was split randomly from the dataset is tested on the predictive model. The test data is pre-processed and classified either positive or negative.

4. RESULTS

The analysis on most occurring words in different classes can give an idea about the nature of tweets and methods to improve the accuracy as they can be used as curated word lists for quick checking of tweets.

The below figures shows top 50 words in different classes in the dataset that we used.

Fig. 1. Top 50 words in the negative class

Fig. 2. Top 50 words in the positive class

The results are evaluated on the basis of accuracy, precision and recall.

TABLE I. CONFUSION MATRIX PARAMETERS

		Predicted class	
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.

True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no

False Positives (FP) – When actual class is no and predicted class is yes.

False Negatives (FN) – When actual class is yes but predicted class in no.

Accuracy - Accuracy is simply a ratio of correctly predicted observation to the total observations. For our model, we have got 0.7299 and 0.7200 for Naïve Bayes and SVM algorithms respectively which means our models are approximately 73% and 72% accurate.
The formula to find accuracy is given by,

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Higher precision means lower false positives thus this parameter is important as we want our model to predict negative tweets correctly and has less false positives.
Precision is given by,

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.
Recall is given by,

$$\text{Recall} = \frac{TP}{TP+FN}$$

Multinomial Naïve Bayes algorithm gave an output of about 0.7297 accuracy, 0.7504 recall, 0.7458 precision and the confusion matrix is as shown below

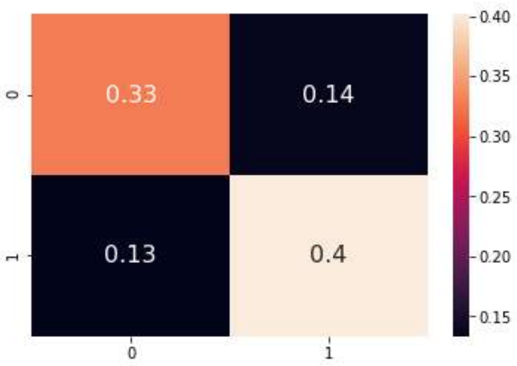

Fig. 3. Confusion Matrix for Multinomial Naïve Bayes algorithm

Support vector machine algorithm gave an output of about 0.7200 accuracy, 0.7675 recall, 0.7251 precision and the confusion matrix is as shown below

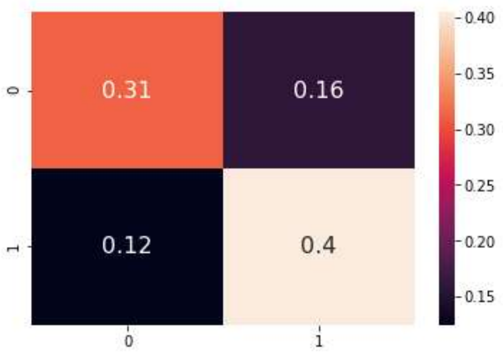

Fig. 4. Confusion Matrix for support vector machine algorithm

TABLE II. RESULTS

PERFORMANCE METRICS			
ALGORITHM	ACCURACY	PRECISION	RECALL
MULTINOMIAL NAÏVE BAYES	0.7297	0.7458	0.7504
SUPPORT VECTOR MACHINE	0.7200	0.7251	0.7675

5. CONCLUSION

Based on the above observations, we can conclude that for the context of depression detection Multinomial Naïve Bayes algorithm has performed well since the precision score is higher for Multinomial Naïve Bayes. We also observed anomalies like conjoined words, separation of which will increase the efficiency of the model. Though better results have been achieved in the previous proposals, those results are over a limited number of tweets and since data is dynamic, it requires dynamic methods to tackle all the obstacles that arise in the sentiment analysis in the future.

ACKNOWLEDGEMENT

A special thanks to Mr. Santhosh for motivating us to take up this project and guiding throughout. Thanks to Dr. Kamalakshi Naganna for their valuable suggestions and thanks to everyone who has given valuable information about technologies through their research papers, journals and conference papers and finally thanks to all my friends and family for their moral support.

REFERENCES

[1] M. Deshpande and V. Rao, “Depression detection using emotion artificial intelligence”, IEEE ICISS Palladam, 2017, ISBN:978-1-5386-1959-9

[2] Y. Mejova, Sentiment analysis: An overview, http://www.cs.uiowa.edu/ymejova/publications/CompsYelenaMejova.pdf, 2009.

[3] A.K.Jose, N.Bhatia, and S.Krishna, Twitter Sentiment Analysis,National Institute of Technology,Calicut, 2010.

[4] A. N Hasan, B. Twala, and T. Marwala,Moving Towards Accurate Monitoring and Prediction of Gold Mine Underground Dam Levels, IEEE IJCNN (WCCI) proceedings, Beijing, China, 2014.

[5] T.Mitchell, H.McGraw, ”Machine Learning, Second Edition, Chapter One”, January 2010.

[6] Xu, Shuo & Li, Yan & Zheng, Wang. (2017). Bayesian Multinomial Naïve Bayes Classifier to Text Classification. 347-352. 10.1007/978-981-10-5041-1_57.

[7] C.D.Manning, P.Raghavan, H.Schutze,Introduction to Information Retrieval,Cambridge UP, 2008

[8] Zi-qiang Wang,Xia Sun, De-xian Zhang, Xin Li, ”An Optimal SVMBased Text Classification Algorithm”, International Conference on MachineLearning and Cybernetics, 2006

[9] Chowdhury, G., ”Natural language processing”, Annual Review of Information Science and Technology, 2003