IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Twitter Sentiment Analysis

Raju Kumar, Moumita Mog, Shubham Kumar, Sumit Madhukar
Under the guidance of Mr. DILIP JAISWAL
Roorkee College of Engineering

Abstract: - Twitter is a micro-blogging website that allows people to share and express their views about topics, or post messages. There has been a lot of work in the Sentiment Analysis of twitter data. This project involves classification of tweets into two main sentiments: positive and negative. In this project, the use of features such as unigram, bigram, POS tagging, and effects of data pre-processing like stemming is observed. Naive Bayes, Support Vector Machines (SVM) and Maximum Entropy (Maxent) are used as the main classifiers. As we shall see in the sections below, SVM with features of unigrams bigrams and stemming, outperforms Naive Bayes.

I. INTRODUCTION

Microblogging sites, in today's world have become a sea of data for analysts to prey on. This is because most of the individuals today are connected to some kind of microblogging site where they pull out all the hype, they feel regarding anything. It won't be wrong to say that in some way these Microblogging sites have given a right to speech to every individual who can access them. People from diverse parts of the world freely discuss, comment, post their opinions about any topic of their choosing in real time. These blogs are mostly a complain expressing a negative vibe or an appreciation expressing a positive vibe toward any topic of their choosing. The topics people post about could be a product from an organization such as a laptop or a phone. Or it could be a famous entity Or any other thing. Most of the leading organizations in today's era have employed analysts who have a job to derive emotions of people behind these posts. This helps them to get a proper review About their product or company which helps them know public demand and the alterations they Need to make in order to make better product in future. Therefore, from the discussion above it could be concluded that these micro-blogging sites could become an asset to different organizations public or private if analysis of sentiment could be implemented on them. Sentiment analysis also known as analysis of feelings is an useful tool for analysing different sites where people post their opinions regarding a topic of interest .With the help of this kind of analysis organizations can obtain the sentiments of the people which they post as tweets or as comments or even as review regarding a particular entity or product of interest to them .This goes in accordance with[10] who says , almost 87% people having a connection with internet check reviews before purchase. This technique could be used for different purposes such as politicians could use it for analysing what kind of sentiments people from different areas are carrying towards him/her and hence could invest more in those areas. An example of this is recent Trump elections, where he hired a group of analysts for this specific purpose. Sentiment analysis could also be applied in the field of business marketing. With the help of this technology different business organizations capture the feelings of people regarding their products and of that of their competitors. Organizations employ their strategies with accordance to this knowledge only. Leaving market research aside, analysis of sentiments could play a vital part in Service industries as it could analyse a full-fledged customer experience and could reveal customer feeling, which could prove to be very beneficial.

II. LITERATURE SURVEY

These days analysis of feelings from twitter is on constant appraisal within the research community as its applications have a huge influence over the working of different industries today. The main challenge faced by this type of analysis is the variation of speech and complex structure of data when extracted.

Aliza Sarlan, Shuib and Chayanit [2] conducted experiments on twitter data in which they simply extracted the tweets in Jason format and used python lexicon dictionary to assign polarity to the tweets. On the other hand, Mandava Geeta, Bhargavav and Duvvada [3] turned it up a notch and used learning methods for the same purpose and achieved a better accuracy of result. For this they collected data regarding cryptocurrency and applied algorithms like naïve bayes and SVM (Support Vector Machine) on it. These experiments further concluded that naïve bayes classifier has more accuracy then SVM.

Another research was conducted by Agarwal, Xia, Vovshaa, I., Rambow, O., and also Passonneau [4] in which a unigram model was used as a baseline and was compared with other models such as one, model based on features and another model based on kernel tree. The experiments revealed that feature based model out performed the unigram model with a negligible margin where as both unigram as well as feature-based models were outperformed by kernel tree-based model with a significant margin.

We proceeded with an approach which was a combination of corpus based as well as lexicon-based approach. This combination is very rarely found in the work that has being done in this field as machine learning techniques are taking over.

In their experiments they have used adjectives and verbs as their features and have used corpus-based techniques for finding the semantic orientation of various adjectives present in the tweets and as for the verbs they have used lexicon dictionary. A linear equation is used to convey the total sentiment polarity of tweets.

Raju and Moumita gathered data on different aspects of demonetization from twitter. They used R language as a tool for analysing these tweets. Not only were the tweets analysed but the result was visualized using different projections such as word cloud and other different plots. These plots showed that the number of people accepting demonetization is more than the number of people rejecting it.

Shubham and Sumit aimed to predict the emotions behind the audience watching a random tv show as positive or negative. For this purpose, they extracted comments regarding some random tv shows and used these as data set for training and testing the model. The model they choose was naïve bayes classifier for which a result was displayed using a pie chart. This pie chart concluded that the polarity of tweets with respect to negative is more then that of positive.

As stated in the section above sentiment analysis could be used for politics. Tumasjan et al. [7] came across the field and its benefits in election and used it for predicting the results in 2009 for German federal elections. They extracted approximately 100,000 tweets for this purpose regarding many political parties of that time and area. Then analyzed the tweets in order to gain sentiments for them. For this they used a software popularly known as (Linguistic Inquiry and Word Count) LIWC2007. This software uses textual analysis as a base to derive sentiments. The results obtained by this analysis were very much similar to the actual results of the elections. Another interesting research was carried out by Dr Rajiv along with some of his mates. They have applied the technique of sentiment analysis in a brand-new way, where they have used this technique to better situations in crises situations. They collected the data of 2014 about a deluge which occurred in Kashmir at that time. Data set collected by them consisted of almost 8490 tweets on which naïve bayes classification technique was implemented. Their research showed that applying analysis of feeling in these situations of crises could help the government in saving lives.

III. DATA CHARACTERISTICS

There are way too many social networking sites available these days, but in this paper, we are dealing with just one such site and that is twitter. Twitter is in too much fame in present because of its specific format of writing. Few of the characteristics of tweets is given below:

Tweet length: tweets are short messages consisting of a maximum of 140 characters.

Tweet availability: twitter is in way more fame than any other social networking site till present day. So much so that approximately 1.2 billion tweets are posted on a daily basis.

Topics discussed: tweets are posted by a wide variety of people, thus the topics discussed is also variant and could almost include any topic starting from politics to mobile products etc.

Writing style or technique: tweets are written in total bogus style i.e., there are many spelling mistakes, use of slang is common, use of smileys is common. Thus, pre-processing of these plays a vital role for this.

Real time: tweets are very small and thus are quite often updated.

Emoticons: these represent facial expression of user in a written form. Use of punctuations with other characters is made to create these emoticons.

User mentions: '@' character is used to make a mention of any user as to direct the message towards them.

Hash tagging: '#' character is used to make the mention of the topic relating to which tweet is being written.

Other symbols: the use of 'RT' is done to symbolize the tweet as retweet, meaning posted again.

IV. METHODOLOGY

The proposed method for sentiment analysis in this paper could be represented in 5 stages, each of which are listed below:

A. Data Collection
B. Data Preprocessing
C. Feature Selection
D. Model Selection
E. Model Evaluation

A. Data Collection

Data collection is the first phase for analysis as there needs to be data for us to do analysis on. In our experimentations we have used python programming language as a tool. Being that said, data collection in this particular analysis could be carried out in two ways. First way is to collect preorganized data from different sites such as Kaggle. On these sites this preorganized data is uploaded by the developers of sites themselves or is posted by different researchers for free. All one needs to do to acquire this data is to create a free account on these sites. Second way is to manually extract data from twitter using some API available for twitter. For this we have chosen tweepy as an API for extraction of tweets. Tweepy does not compatible with the new versions of python (python 3.7). So, for using this particular API an older version of python is needed (python 2.7). To access tweets on twitter using API first we need to authenticate the console from which we are trying to access twitter. This could be done by following steps listed below:

Creation of a twitter account.
Logging in at the developer portal of twitter.
Select "New App" at developer portal.

A form for creation of new app appears, fill it out Fill.

After this the app for which the form was filled out will go for review by twitter team.

Once the review is complete and the registered app is authorized then and only then the user is provided with 'API key' and 'API secret' After this "Access token" and "Access token secret" are given.

After this "Access token" and "Access token secret" are given.

These keys and tokens are unique for each user and only with the help of these can one access the tweets directly form twitter. For this paper we have extracted a large data set consisting of almost 3000 tweets. These tweets are taken using #USairlines and thus are about different US Airlines. We have used textblob package of python for pre-data annotation of polarity for these tweets.

Data set	No. of tweets
Training data	2343
Testing data	585

Table 1:- Data Distribution\

B. Data Pre-processing

The pre-processing of data implies the processing of raw data into a more convenient format which could be fed to a classifier in order to better the accuracy of the classifier. Here, in our case the raw data which is being extracted from twitter using an API is initially totally unstructured and bogus as the availability of various useless characters seems very common in it.

For this matter we remove all the unnecessary characters and words from this data using a module in python known as Regular Expressions, are for short. This module adopts symbolic techniques to represent different noise in the data and therefore makes it easy to drop them. Specifically in twitter terminology there are various common useless phrases and spelling mistakes present in the data, which need to be removed to boost the accuracy of our resultant. These could be summoned up as follows:

Hash tags: these are very common in tweets. Hash tags represent a topic of interest about which the tweet is being written. Hashtags look something like #topic.

@Usernames: these represent the user mentions in a tweet. Some times a tweet is written and then is associated with some twitter user, for this purpose these are used.

Retweets (RT): as the name suggests retweets are used when a tweet is posted twice by same or different user.

Emoticons: these are very commonly found in the tweets. Using punctuations facial expressions are formed in order to represent the a smile or other expressions, these are known as emoticons.

Stop words: stop words are those word which are useless when it comes to sentiment analysis. Words such as it, is, the etc are known as stop words.

C. Feature Selection

As mentioned earlier in this paper different researchers have used different features for the classification of the tweets, in our experimentations similar feature are being used. These features include Unigram, Bigram, N-gram, POS tagging, Subjective, objective features and so on. NLTK short for Natural Language Tool Kit is another module available in python which also open source and could be used for extraction of these features.

D. Model Selection

Once the data is being pre-processed, this data is to be fed to a classification model for further processing. There are different classification algorithms on which these models are built on. In this paper, we have chosen k-nearest neighbour model to perform the classification.

KNN or k-Nearest Neighbour algorithm represents a machine learning technique used for classifying a set of data into its given target values (in our case positive, neutral or negative). KNN could also be used for regression problems but is widely used for classification problems.

Now, any classification model needs a target set on which we train the model for its further use. As for mentions in the literature survey section most of them have manually set these target values to positive, negative or null. For this paper we have used a library in python known as Textblob to automatically set the target for each tweet.

The data set then is divided into two halfs training set and testing set. The data set used by us in our experimentations consisted of 2928 tweets so we segregated it into training and testing data. Training portion consisted of 2343 tweets whereas the test set consisted of 585 tweets. Now this training as well as test set needs to be transformed into binary values so as to be fed to the model. The models don't understand any values other than the binary.

For this we have used another module of the python known as sklearn which contains many classifications model as well as different encoders in it. For this paper this library is being for model selection, label encoding as well as model evaluation which would be mentioned in next section.

E. Model Evaluation

One of the most common and appropriate technique used for evaluation of a classifier is through confusion matrix. A generalized form of confusion matrix is given in table 4.5 below:

	Predicted class1	Predicted class 2
Actual class 1	True positive(tp)	False negative(fn)
Actual class 2	False positive(fp)	True negative(tn)

Table 2:- General Confusion Matrix

By applying this technique, we can derive the generalized evaluation parameters. These parameters include:

Accuracy: accuracy of a classifier indicates how accurately the classifier has predicted the result. It can be calculated using the formula:

$$\textbf{(a)}\quad \frac{tp + tn}{tp + tn + fp + fn}$$

Precision: precision shows how often the result that is being predicted by the classifier when it indicates true is correct. The formula for precision is:

$$precision(p) = \frac{tp}{(tp + fp)}$$

Recall: it indicates the true positive rate of the classifier. Formula for recall is:

$$recall(r) = \frac{tp}{(tp + fp)}$$

F1 score: it indicates the weighted average of recall and precision. Formula for recall is:

$$F1\ score = \frac{2p.r}{(p - r)}$$

V. RESULTS OF PROPOSED SYSTEM

The experiments conducted by us showed a model accuracy of 65.33% for KNN (k nearest neighbor) classifier. The confusion matrix guised after completion of testing of classifier is given in table 3 below:

	Negative	Neutral	Positive
Negative	1612	161	80
Neutral	418	161	34
Positive	276	46	140

Table 3:- Output Confusion Matrix

Other important model evaluation parameters as mentioned in section before, for this experimentation are given in the table 4 presented below:

	Precision	Recall	F1-Score
Positive	0.70	0.87	0.78
Neutral	0.44	0.26	0.33
Negative	0.55	0.30	0.39

Table 4:- Evaluation Parameters

The positive, null and negative could also be represented using wordclouds. wordclouds for our data set are given below:

Fig 1:- Negative Wordcloud

Fig 2:- Positive and Null Wordcloud

VI. CONCLUSION AND FUTURE WORK

In this paper, we have demonstrated a system for the analysis of textual twitter data in particular for the emerging field of sentiment analysis. For this we built a model for the analysis of feeling using KNN algorithm with unigram, bigram and ngram features. We then performed a training and testing of this model on #USairline data set, for which we attained an accuracy of 65.33 % . In near future we aim perform similar analysis in order to increase the accuracy of the model and doing so without extracting any features, by using different deep learning techniques such as neural networks.

REFERENCES

[1] H. Zang, "The optimality of Naïve-Bayes", Proc. FLAIR 2004.

[2]. "Aliza sarlan, chayanit nadam, shuib basri ," twitter sentiment analysis 2014 international conference on information technology and multimedia (ICIMU), november 18 – 20, 2014, putrajaya, malaysia 978-1-4799-5423-0/14/$31.00 ©2014 ieee 212"

[3]. "Mandava geetha bhargava , duvvada rajeswara rao , " sentiment analysis on social media data using R" , international journal of engineering & technology, 7 (2.31) (2018) 80-84"

[4]. "Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. "Sentimentanalysis of twitter data." In Proceedings of the ACL 2011,Workshop on Languages in Social Media, pp. 30–38, 2011.

[5]. "IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012 ISSN (Online): 1694-0814"

[6]. "International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 6, Issue 7 July 2017 "

[7]. "Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I."Predictingelectionswith twitter: What 140 characters reveal about political sentiment.." In Proceedings of theFourth International AAAI Conference onWeblogs and Social Media (2010), pp.178–185, 2010