THINK INDIA JOURNAL

ISSN: 0971-1260

Vol-22-Issue-17-September-2019

Sentimental Analysis of Twitter Data Set using Machine Learning Techniques

Pooja Rana, Usha Mittal, Sanjay Kumar, Aditya Khamparia

Department of Computer Science and Engineering, Lovely Professional University, Phagwara

Abstract:

Social media is a platform which signifies the viewpoint of people about government, product, and services. Classifying the views in term of positive or negative division from social media text is task of sentimental analysis. Twitter is one such online social networking website where people can post their views about something. This platform help people to share as well as found views on the topics which are going around the world. Twitter is a huge platform with more than 300 million users. So, tweets can be used as a resource for mining views of people. Process of classifying views as positive, negative or neutral is known as sentimental analysis. Saying people tweet about everything and nothing is not an overemphasis. We need a quicker identification method to classify the information that can be used for training in order to construct the systems to twitter sentimental analysis on any given topic. In this project, we progress a method for building such data using Twitter (for example: #excellent #Nature Lover #epic failure) to identify emotional messages that can be used to classify sentiment.

Introduction

Social projects have evolved exponentially since the last few years like Twitter, Facebook. Because people are so interested to discuss the things about the latest news, products. These things made industries, companies, and organizations seeking social projects. Twitter which provides information about what people think, feel about the one's product and services. Some of the organizations that use this twitter as the feedback platform they used to Announce Twitter's sentiment analysis as one of its services. Most obsolete approaches are being used in emotional research the "Bag of words" method. The relationship between word groups is seen as an alternative to the relationship between languages. The feeling of each word is determined and joined using a function while defining the overall feeling. Bag of Words often disregards word order, which contributes to the incorrect identification of phrases with negation in them.

ISSN: 0971-1260

Vol-22-Issue-17-September-2019

Naive Bayes, Maximum Entropy and Support Vector Machines are other methods explored in sentiment analysis.

Sentimental analysis addresses the natural language processing board field which deals with mathematical learning of beliefs, thoughts, and emotions expressed in a message. Sentimental Analysis or Opinion Mining would like to study the opinions, attitudes and emotions of individuals towards an entity. The entity represents the place, item, persons, activities, or subject matter. In the field of emotional analysis, a huge amount of research has been done. Most of them have mastered the identification of larger portions of text such as comments. With the widespread popularity of social media and some blogging, projects provide enormous amounts of data, which act as resources for sentimental analysis. Blogging sites are rising tremendously as they use nostalgic research as a witness. Popular blogging projects such as Twitter had developed into a source of vast knowledge. Twitter maintains the diversity that owes the information on a wide range of topics and also provides the platform where people can post real opinions on a wide range of topics, discuss current affairs and share their experiences on the services and products they use in everyday life. Motivated by the development of microblogging platforms, industries and organizations that explore ways to own Twitter to learn more about how people respond to their services and products. The unbiased amount of research has been granted on how feelings are expressed in structured text trends such as goods, film reviews, tourist locations, and political reviews, but how feelings are articulated when given feelings are given in the informal language and microblogging constraints on the duration of the message have been less traveling.

Twitter is one of the cutting-edge microblogging sites with more than 321,000,000 active users aerated in 2006. In this service, the user-created or posted messages are called tweets. The twitter service's public timeline shows tweets from all subscribers around the world and is a broad source of real-time information. The original idea behind microblogging was to make present status of a person updates accessible. But the situation is amazingly watching tweets covering everything in the world, from political news to personal situations of a person. Reviews of products, travel experience, reviews of movies, places of tourism, etc., add the list. Tweets vary in their basic structure from comments. While reviews are distinguished by formal trends and authors ' condensed opinions, comments are more informal and are restricted to 140 text or email or tweet characters. Tweets offer additional opportunities for companies and

ISSN: 0971-1260

Vol-22-Issue-17-September-2019

organizations to gather feedback. Sentimental research to evaluate criticism of movies, goods, etc., and help decision-makers before buying a product or planning a film. Enterprises found that this Twitter microblogging platform is more useful in researching their company's public opinion and products that they release, or in analyzing customer satisfaction. Organizations use this information to gather feedback about their newly released product if they take this review as a source of further implementation or improvement for any faults or improvements. Different methods incorporating ML techniques such as natural language processing, supporting vector machine, k means clustering, etc., feeling lexicons, hybrid approaches have been sponsored useful for structured text sentimental research. But it will need to investigate their efficacy in collecting tweets, emails, emotions (negative or positive or neutral) from information from microblogging projects. In one of the tweet review reviews, it shows that the size of the 140 character limits users who use more jargon than imparts the feelings. Some of the users use their tweet analyzer to place hyperlinks in their tweets to limit the hyperlinks that are often present in these tweets, in order to restrict the size of the vocabulary. It would definitely enforce sprints for learning on many different domains discussed. Some of the users don't know the spellings while tweeting and some people use their own slang in a tweet. The rate in tweets with misspellings and slang words is much higher than in other language tools, which is another sprint to be conquered. On the other hand, the enormous amount of data available on wide-ranging domains from various microblogging projects is incomparable with other available data resources. Project language for microblogging is characterized by quoted text that delivery a large number of tweets. Exclamations boldly lettered words, question marks, single quoted, double quoted messages, etc., leaving room for extraction of sentiments. The proposed project attempts an innovative approach to twitter data by dividing a modified polarity lexicon that has learned from considered domain reviews, tweet features, and unigrams to figure a model classifier using artificial ML techniques.

Literature review

The social networking sites is a platform from which data can be retrieved easily. People express their emotions and viewpoints on the social media [6]. Based upon the availability of data on social media platform many researchers get lured to the field of sentimental analysis.The business organizations hire researchers to explore the hidden facts about their

ISSN: 0971-1260

Vol-22-Issue-17-September-2019

products and services. Key concern of multinational compines is the review of natural and involuntary determination of sentiments from reviews [7–10].

Sentimental Analysis is a prosperous topic in the field of research. Various text mass like newspaper articles, movie reviews and product reviews were available to study the sentimental analysis. In the beginning research on this topic is conducted using support vector machine and maximum entropy. The claimed maximum optimum result achieved was 83 %.Neural sentiment was not available in the old research. To overcome this limitation, Pak and Paroubek then used 3-class Naïve Bayes Classifier which was able to detect neutral messages along with the polar ones [1].

Twisent, a sentimental system was developed by Researchers from IIT Bombay, India [2], for Twitter.This system collects the tweets and classify them in different categories like positive, negative or neutral[3]. Researchers Barbosa and Feng used Subjective versus Objective classes and Positive versus negative classes as classifiers. Separate evaluation on both model is presented but combining them was not explored [4]. Jiang et al., 2011 present results on building a 3-way classifier for Objective, Positive and Negative tweets. However, they do not explore the cascaded design and do not detect Neutral tweets [5].

In this system [6], machine learning based approach is used for for sentiment analysis. For this, dataset of tweets has been constructed, which was acquired from Twitter using Tweepy API. After obtaining tweets, noise removal was performed by pre-processing of the data. The tweets are categorized as either positive, negative or neutral. The machine learning classifiers are applied on the training dataset.

Proposed Technique:

There are four types of implementation models used in our proposed technique: Logistic Regression Model, Support Vector Machine Model Random Forest Classification Model, Ensemble Learning Classification Model

Now we require the training of Logistic Regression model for classification of the tweets into racist versus non-racist comments. This is a good regression test that is conducted if the variables that are dependent are dividing into branches. Similar to all other regression analyses,

ISSN: 0971-1260

Vol-22-Issue-17-September-2019

this regression model falls under predictive analysis. It is required to explain data and to demonstrate the relation between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables

Support vector machine **(**SVM**),** is one of the most beneficial and extensively used learning algorithm that can be seen as a branch of the perceptron. Misclassification errors can be reduced using the perceptron algorithm. While using SVM the goal is to make the margin large. The distance that separates the hyper plane (decision boundary) and the train sample data is which have least distance to the hyper plane, is defined as margin. And such training samples lying on margin are termed as support vectors.

Random forests achieved large public attention in applications of machine learning in recent years because of the good performance of classification they possess, alongside scaling, and they are easy for using. Frankly, a random forest is a group or *ensemble* of decision trees. The intuition beneath ensemble learning is to group learners that are weaker and make a robust model, a learner that is stronger, that gives less error in generalization, which has less susceptibility for over fit.

The idea beneath ensemble learning models is to group different classification model to make a combined-classification model with overall good performance of generalizing as compared to single classification model individually.

Results

This proposed system is all about providing a platform to the users to know about the particular product, current affairs like leading political party during elections, best tourist place. It is an integrated platform where users can register and monitor their own progress and users be informed about the things that there are searching for. The user has some accessibilities like viewing positive, negative and neutral sentiments on the products, places, political issues. Then the user can get some idea whether he or she can choose the product, he or she can select the person, go to that place or not.

It would become easy to divide the group of users on what they are searching based on their on searching they will get some suggestions like alternate solutions to meet their requirements. So as to provide extra support to the users who are using our technique. The proposed technique

ISSN: 0971-1260

Vol-22-Issue-17-September-2019

description includes the basic structure of the whole system as the primary part of the system is extracting data from the twitter and load the dataset into the platform. After loading the dataset with the help of some inbuilt functionalities we used to clean the data like non phrases, bold letters, punctuations, Numbers, and special characters and also removing short forms. After cleaning the data. We normalize the data which helps the data to use easily in predictive modals. And then extract the hashtags from racist and non-racist data. Taking this training data and loading into predictive modals and finding out the best modal by which is giving the highest accuracy. Which help the users to find out the best and worst things of the products. So many projects are done on twitter data analysis but our system is one step forward to find out best accuracy model from this system we can find which the best predictive modal. Following tables 1,2,3 show the result of proposed model.

Table 1: Results of individual techniques on the twitter data

Test Case No.	Action	Objective	Result	Remarks
01	Use Logistic Regression Model to test the accuracy on dataset.	Test the model on test data and attain maximum accuracy possible.	Accuracy achieved 56%.	Tested OK, Passed. Need some reconsiderations.
02	Use Support Vector Machine Model to test for its accuracy.	Test the model on test data and attain maximum accuracy possible.	Accuracy achieved 54%.	Tested OK, Passed. Need some reconsiderations.
03	Use Random Forest Classification Model to test for its accuracy.	Test the model on test data and attain maximum accuracy possible.	Accuracy achieved 56%.	Tested OK, Passed. Need some reconsiderations.

ISSN: 0971-1260

Vol-22-Issue-17-September-2019

Table 2: Result of Ensemble learning Model

Test Case No.	Action	Objective	Result	Remark
01	Use Ensemble Learning Model to test for its accuracy.	Test the model on test data and attain maximum accuracy possible.	Accuracy achieved 95%.	Tested OK, Passed

Table 3: Performance Testing

Test Case No.	Action	Objective	Result	Remark
01	Run the each model of the code.	Reduction in the mean time required to execute the models.	The mean time required to execute the models and check the result is less than 10 seconds.	Tested OK, Passed

Conclusion

At the time of writing this report, the system was able to achieve accuracy of 95% through Ensemble Learning Model, but individual models have accuracy less than that. The project is able to classify the tweets into racist and non- racist categories. It is also able to tell the words in the tweet that are racist and non- racist.

References :

[1] David Ahn& Balder ten Cate. Simple language models and spam filtering with Naive Bayes, 2005.

[2] TwiSent: A Multistage System for AnalyzingSentiment in Twitter by Subhabrata Mukherjee, AkshatMalu, A.R. Balamurali, Pushpak Bhattacharyya.

ISSN: 0971-1260

Vol-22-Issue-17-September-2019

[3] Sentimenatal Analysis of Twitter Data by Apoorv Agarwal BoyiXie Ilia Vovsha Owen Rambow Rebecca Passonneau Department of Computer Science Columbia University New York, NY 10027 USA.

[4] Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. Technical report, Stanford.

[5] Kim, S. M. and Hovy, E. (2004). Determining the sentiment of opinions.

[6] Gound R.S, V. Priyanka, S. Shivani.(2018). Twitter Data Sentiment Analysis and Visualization

[7]. Cernian A, Sgarciu V, Martin B (2015) Sentiment analysis from product reviews using SentiWordNet as lexical resource. In: 2015 7th international conference on electronics, computers and artificial intelligence (ECAI). doi:10.1109/ecai.2015.7301224

[8]. Hammer HL, Solberg PE, Øvrelid L (2014) Sentiment classification of online political discussions: a comparison of a word-based and dependency-based method. In: Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis. doi:10.3115/v1/w14-2616

[9]. Zadeh L (2006) Toward human-level machine intelligence. In: 2006 18th IEEE international conference on tools with

artificial intelligence (ICTAI'06). doi:10.1109/ictai.2006.114

[10]. Joachims T (2002) Text classification. Learning to classify text using support vector machines. p 7-33. doi:10.1007/978-1-4615-0907-3_2