Opinion Sentiment Analysis on Twitter and Wikipedia

[1] KONDAVEETI SRI VENKATESH & [2] SREEJA SRIRAM
[1-2] Graduates from Gokaraju Rangaraju Institute of Engineering and Tech, Hyderabad, India

Abstract

This is a Project based on Text Mining of Sentiment Analysis which provides a novel approach to identify and classify opinions called sentiments from source text. It is performed by analysing tweets from Twitter and information from Wikipedia. The User Interface allows users to enter a name of any famous Personality with a valid twitter account or a page in Wikipedia .The input for twitter textbox is used to retrieve the recent hundred tweets, which can be up to a maximum of 280 characters in length from twitter using Authentication Keys. The input for wikipedia textbox fetches data from Wikipedia and conducts pre-processing. Pre-processing is executed on the raw data through Regular Expressions to remove noisy data like URL's, special characters etc., to make them clean and well organised. The challenge is to gather all such relevant data, perform predictions and calculate the polarity which is critical for decision making. Polarity refers to be the emotion expressed in a sentence, which can be categorised as positive, negative or neutral ranging from -1 to +1. Finally, the output of polarity is exhibited in the form of a pie chart and graph. This sort of analysis can mainly be advantageous for organisations, where status of the organisation can be known by simply having a look at the graph/pie chart saving a lot of valuable time rather than going through all the reviews in the form of text.

Keywords: Tweet, Sentiment Classification, Wikipedia, Common Gateway Interface, Sentiment Analysis, Text Mining

1. INTRODUCTION

In the past decade, there has been no limit to the range of information that is being conveyed using tweets and text messages, which are often short messages used for sharing opinions called sentiments about things happening around them. The language used is mostly informal with creative spellings, new words, URLs, abbreviations, punctuations and hashtags, which is a way of tagging. Similarly, Wikipedia often called as the wisdom of mankind, has become one of the growing trends in today's world where we can get detailed information about a person, product, trend, institution etc. It contains more than 43 million pages in total and about 510 million unique visitors till date. There are about 80000 active contributors working on more than 50,000,000 articles in 295 languages. Every day, hundreds of thousands of visitors all around the world make tens of thousands of edits altogether and create new articles to enlarge the knowledge held by Wikipedia. Sentiment Analysis provides an approach to identify and classify opinions called sentiments

from source text. It uses Pre-Processing, text analysis and computational linguistics for identifying and analysing the raw data and predicting opinions. It is also referred as opinion mining. The internet has established a platform for people to express their views, emotions on products, people and things around them. The objective of sentiment analysis is to extract important insights from large amounts of data by removing unnecessary data that would help organisations, better decision making and quality product consumption. In this paper, we look at an approach, where a model is built for classifying tweets, which are retrieved using consumer keys, access tokens from Twitter Search API and data from Wikipedia into positive, negative or neutral. The noisy data like URLs, stop words, hash tags etc., are removed using regular expressions in order to get an organized data. This process is for calculating polarity. The polarity is then depicted in the form of a pie chart and graph using matp lotlib and pandas modules. The status of the organisation can be known by simply having a look at the graph/pie chart saving a lot of valuable time rather than going through all the reviews in the form of text.

2. RELATED WORK

Applying Sentiment Analysis on a microblogging site like Twitter and Wikipedia has been a trend to many of the scientists scrutinizing the scientific trials and its applications. Pak and Paroubek [1] use distant learning where twitter has been used as a corpus for sentiment analysis. They cited:

- Microblogging sites are used by different people for the purpose of expressing their opinion on different topics around them.
- Twitter tweets are enormous in size and grow larger every day due to increase in usage by people collectively.

Parikh and Movassate [2] implemented two models, a Naive Bayes bigram model and a Maximum Entropy model to classify tweets. They found that Naive Bayes classifiers are more efficient than Maximum Entropy model. Barbosa and Feng [3] used noisy labels to train a model, 1000 manually labelled tweets for tuning and another 1000 manually labelled tweets for testing. They proposed the use of features like retweets, punctuations, hashtags, URLs with polarity of words which helped but only marginally. Agarwal et al. [4] approached with 3 types of models: unigram model, feature-based model and a tree kernel based model. They concluded that tree kernel model outweighed the other 2 models in accuracy and efficiency. Go et al. [5] introduced a solution by using distant supervision, which included trained data consisting of tweets with emoticons. This approach was initially proposed by Read [6]. In this approach, emoticons were treated as noisy labels using Naive Bayes, Support Vector Machines and MaxEnt. They reported that Support Vector Machines overtook other models in performance.

Bing Liu [7] has described several challenges in sentiment analysis some of them being feature extraction, opinion orientation classification, synonyms grouping. Jalai S Modha et al. [8] described an approach to handle both subjective and objective sentences and used classification techniques to classify sentences as opinionated or non-opinionated for objective analysis. G. Vinodhini and R. M. Chandrashekaran [9] stated

that an opinion word can have different orientations in different sentences and also differentiated opinion summarization task which defers from text summarization that features mining of products alone.

3. EXISTING SYSTEM

It is common that people make opinions on brands or on famous personalities, which may vary from person to person. For an organization, as long as they are growing, it indeed becomes very difficult to know how people feel about their brand. For large scale companies, it is extremely difficult to manually handle thousands of references made on social media or on e-news sites, as it effects their growth in the business they are associated with. In present generation, this is how things work. One believes in and buys a product by mostly looking at the user reviews. So, every organization has to think over how good their product/service is reached out to a customer. They have to know where they are bad at and have to work over it to make it better every day. Such a problem can be resolved using a software called sentiment analysis which analyses people's sentiments on a particular product or personality.

4. PROPOSED SYSTEM

CGI is used as an interface between the user and server which generates web pages dynamically. An interactive screen is developed with which user can find the sentiment of a particular Twitter user's tweets or a sentiment of any topic in Wikipedia. Results are shown in the form of graphs and pie charts.

Advantages of proposed system

- It provides a User Interface where user can enter input in either Twitter or Wikipedia.
- It uses TwitterSearch API to access the tweets based on the input given.
- It removes stop words from the raw data which don't contribute to any of the sentiment.
- It uses a pie chart and graph to represent the sentiment.

5. DATA DESCRIPTION

- **Trained Data:** We define some data in a Comma Separated File (CSV) file and classify them as positive, negative or neutral manually.
- **Test Data:** The data on which the classifier is applied. In this system, it is either Wikipedia data or the tweets from Twitter.
- **Message Length:** The maximum length of tweets in twitter is 280 characters. Initially, only a maximum of 140 characters was allowed, which has now been increased.
- **Writing Method:** As the length of tweet is less, people often prefer using acronyms and cyber slang which is more often used when compared to any other microblogging site.
- **Availability:** There is a plethora of data available. The Twitter API allows people to retrieve only the tweets which can be further filtered using regular expressions.
- **Topics:** There is a wide range of topics on which users of twitter post messages on unlike any other microblogging site which confines to a specific topic only. It varies from movie reviews to products.

- **Processing:** To process the data, we make use of Naive Bayes classifier. This classifier goes through the trained data and builds a model. This model is then applied on test data to classify it into positive, negative or neutral. When classifier encounters a text in test data, it tries to match it with the similar text in trained data. It then processes and classifies the text.

6. DATA COLLECTION

Data is collected from Twitter and Wikipedia. Twitter is a microblogging and social networking site which provides a feature of allowing users to post short messages called tweets. Twitter Search API can be used to retrieve 100 recent tweets of an account and they are processed. Pre-requisite: To get an access to twitter API, we need to have a twitter account with which we have to create an application in developer's site and get consumer keys and access tokens. There is no separate authorization required for Wikipedia to be accessed.

7. URL'S, STOP WORDS REMOVAL

The data obtained from the above method is noisy data, which has to be cleaned to make the process efficient. Check the data which might consist of URLs, user names, several stop words which do not signify any sentiment. We make use of regular expressions to remove usernames and URLs. Further we can add steps to remove stop words, emoticons, special symbols etc.

Hashtags: A feature that allows users to mark topics in order to increase the visibility reach to others.

Targets: The users of twitter use the "@" symbol to mention other users on the tweet to alert them about the topic.

Stop words: The words like "is", "a", "an", do not possess any sentiment. Hence removed.

Special Symbols: The special symbols refer to shortcuts that are used in the microblog like "RT" for re-tweet which indicates posting a repeat of some other's earlier tweet.

8. SYSTEM ARCHITECTURE

The words people use to express their opinions are used to calculate the strength or polarity. In this system, we proposed a novel hybrid approach which involves corpus-based techniques.

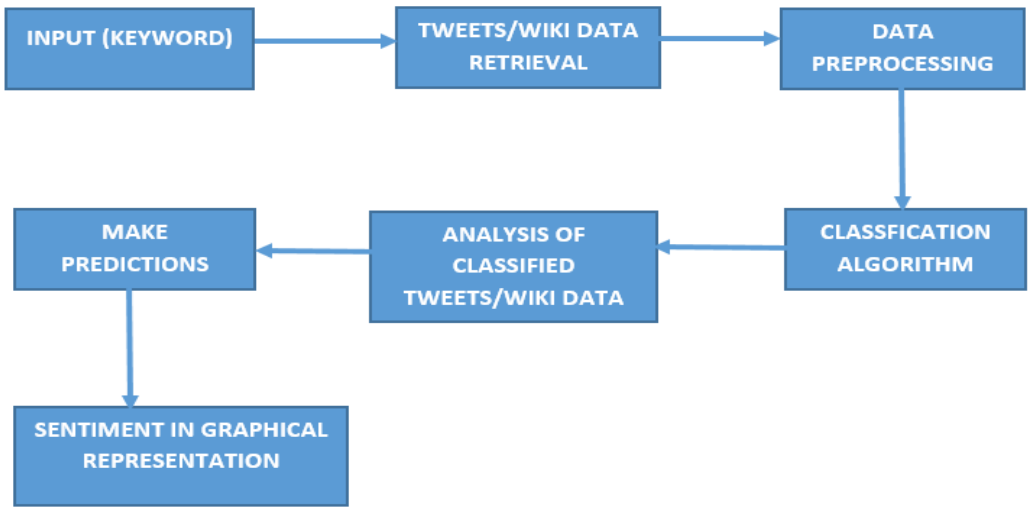

Fig:- System Architecture

To calculate the polarity, we first start by giving an input to the twitter textbox. This input is used to retrieve tweets using TwitterSearch API with the help of keys provided. The tweets which is raw are then pre-processed as there is noisy data. The classification algorithm called Naive Bayes is then applied to classify the data which is organised and then analysis is performed on classified tweets. The predictions are made to find the sentiment of a particular input.

9. EXPERIMENTS AND RESULTS

TWITTER (Accounts of different categories taken)	AVERAGE POLARITY
1) Singers	Positive: 84% Negative: 12% Neutral: 4%
2) Actors	Positive: 64% Negative: 24% Neutral: 12%
3) Politicians	Positive: 61% Negative: 34% Neutral: 5%

Table 1. Table showing 3 different categories taken from twitter to calculate polarity

WIKIPEDIA (Pages of different categories taken)	AVERAGE POLARITY
1) Singers	Positive: 86% Negative: 10% Neutral: 4%
2) Actors	Positive: 72% Negative: 18% Neutral: 10%
3) Politicians	Positive: 70% Negative: 14% Neutral: 16%

Table 2. Table showing 3 different categories taken from wikipedia to calculate polarity
Note: 1) 5 examples have been taken and experimented for each category.

2) The examples taken to calculate polarity for different categories are not being disclosed in order to avoid issues.

The following are the results obtained by conducting analysis and calculating their corresponding polarities –

Singers

- For both Twitter and Wikipedia, singers dominated other 2 categories by carrying the highest positive polarity.
- They possess the least percentage of neutral polarity in both Twitter and Wikipedia.

Actors

- For Twitter, actors carry the highest percentage of neutral polarity when compared with other 2 categories.
- For Wikipedia, they own the highest proportion of negative polarity.

Politicians

- For both Twitter and Wikipedia, politicians maintain the least percentage of positive polarity when compared to other 2 categories.
- For Wikipedia, they hold the highest percentage of neutral polarity and for Twitter, they have the highest proportion of negative polarity.

10.1 User Interface

Fig 10.1.1 User Interface which allows user to select from which domain he wants to know the sentiment about his desired search

10.Graphical Representation of the Sentiment

The result is represented in the form of a pie chart and a graph. It shows the positive, negative and neutral tweets/articles count in percentages.

Pie Chart:

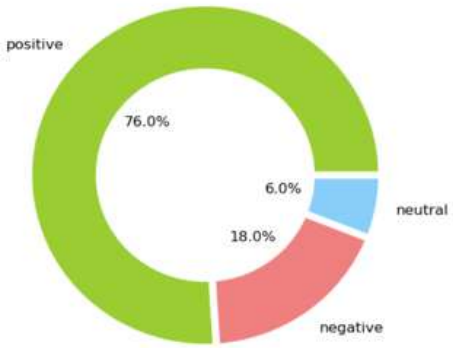

Fig 10.2.1 Wikipedia

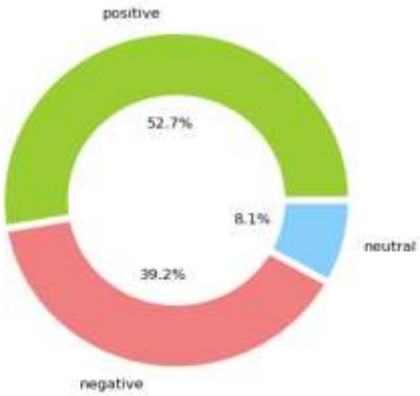

Fig 10.2.2 Twitter

Graph:

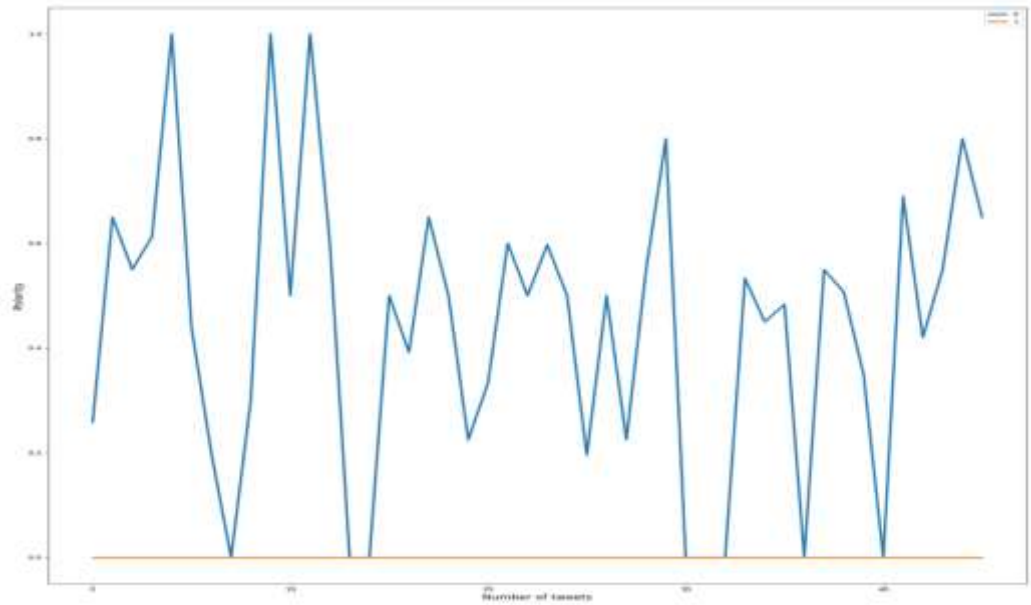

Fig 10.2.3 Graph showing the number of tweets on x-axis and polarity on y-axis

11.CONCLUSION

With an increase in usage of technology an abundant amount of data is being generated every day from smartphones, blogs, social networking sites etc. In a survey, researchers have stated that the amount of data generated in the past 2 years is almost equal to the amount of data generated till 2012 since its inception. All that raw data is not useful. In such a case, sentiment analysis comes to the rescue of generating useful insights that can be advantageous to organizations for better decision making and better quality consumption. It can be used to know the opinion of the crowd. Despite all the challenges of implementing sentiment analysis, one cannot disregard the value it adds to the industry. Improved consistency and accuracy in opinion mining, may reflect as a solution to the problems faced in sentiment analysis. Keeping this in mind, we have worked to put forward a novel approach to obtain data, followed by performing an analysis, overcoming the limitations and representing them in a user-friendly graph and pie chart.

REFERENCES

[1] A. Pak and P. Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320–1326.

[2] R. Parikh and M. Movassate, "Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques", CS224N Final Report, 2009.

[3] L. Barbosa, J. Feng. "Robust Sentiment Detection on Twitter from Biased and Noisy Data". COLING 2010: Poster Volume, pp. 36-44.

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

[5] A. Go, R. Bhayani, L.Huang. "Twitter Sentiment Classification Using Distant Supervision". Stanford University, Technical Paper, 2009.

[6] J. Read. "Using emoticons to reduce dependency in machine learning techniques for sentiment classification". In Proceedings of ACL-05, 43nd Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2005.

[7] Bing Liu, "Sentiment Analysis: A Multi-Faceted Problem", IEEE Intelligent Systems, 2010.

[8] Jalaj S. Modha, Gayatri S. Pandi, Sandip J. Modha, "Automatic Sentiment Analysis for Unstructured Data", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol 3, Issue 12, Dec 2013, pp. 91-97.

[9] G.Vinodhini, R.M.Chandrashekaran, "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol. 2, Issue 6, June 2012, pp. 282-292.