IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

Investigating Sentiment analysis using Clustering and NLP tools

Ashwini Yerlekar[1*], Devika Deshmukh[2]

[1]Department of CSE, Rajiv Gandhi College of Engineering and research, Nagpur, India
[2]Department of IT, Rajiv Gandhi College of Engineering and research, Nagpur, India

Available online at: www.ijcseonline.org

Accepted: 23/Jan/2019, Published: 31/Jan/2019

Abstract— Twitter is a social media platform, a place where people from all parts of the world can make their opinions heard. Twitter produces around 500 million of tweets daily which amounts to about 8TB of data. The data generated in twitter can be very useful if analyzed as we can extract important information via opinion mining. Opinions about any news or launch of a product or a certain kind of trend can be observed well in twitter data. The main aim of sentiment analysis (or opinion mining) is to discover emotion, opinion, subjectivity and attitude from a natural text. In twitter sentiment analysis, we categorize tweets into positive and negative sentiment. Clustering is a protean procedure in which identically resembled objects are grouped together and form a pack or cluster. We conducted a study and found out that the use of clustering can quickly and efficiently distinguish tweets on the basis of their sentiment scores and can find weekly and strongly positive or negative tweets when clustered with results of different dictionaries. This paper implements the approach of clustering with respect to sentiment analysis and presents a way to find relationships between the tweets on the basis of polarity and subjectivity.

***Keywords*—** Opinion Mining, sentiment analysis, clustering, Twitter

I. Introduction

With the recent growth of mobile information systems and the increased availability of smart phones, social media has become a large part of daily life in most societies[1]. This development has entailed the creation of massive amounts of data which when analysed can be used to extract valuable information about a variety of subjects.

Sentiment analysis is the computational task of automatically determining what feelings a writer is expressing in text. Sentiment is often framed as a binary distinction (positive vs. negative), but it can also be specific emotion an author is expressing (like fear, joy or anger), also known as emotion mining,the process of classifying the emotion conveyed by a text, for example as negative, positive or neutral[2]. The data made available by social media has contributed to a burst of research activity within sentiment analysis in recent times and a shift in the focus of the field towards this type of data. Information gained from applying sentiment analysis to social media data has many potential usages, for instance, to help marketers evaluate the success of an ad campaign, to identify how different demographics have received a product release, to predict user behaviour[3].
Some applications for sentiment analysis include:

- Analysing the social media discussion around a certain topic.
- Evaluating survey responses.
-
- Determining whether product reviews are positive or negative.

As internet is growing bigger, its horizons are becoming wider. Social Media and Micro blogging platforms like Facebook, Twitter, and Tumblr dominate in spreading encapsulated news and trending topics across the globe at a rapid pace. A topic becomes trending if more and more users are contributing their opinion and judgments, thereby making it a valuable source of online perception. These topics generally intended to spread awareness or to promote public figures, political campaigns during elections, product endorsements and entertainment like movies, award shows. Large organizations and firms take advantage of people's feedback to improve their products and services which further help in enhancing marketing strategies. One such example can be leaking the pictures of upcoming iphone to create a hype to extract people's emotions and market the product before its release. Thus, there is a huge potential of discovering and analysing interesting patterns from the infinite social media data for business-driven applications.

Sentiment analysis is the prediction of emotions in a word, sentence or corpus of documents. It is intended to serve as an application to understand the attitudes, opinions and emotions expressed within an online mention. The intention is to gain an overview of the wider public opinion behind certain topics. Precisely, it is a paradigm of categorizing conversations into positive, negative or neutral labels. Many people use social media sites for networking with other people and to stay up-to-date with news and current events[4]. These sites (Twitter, Facebook, Instagram, google+) offer a platform to people to voice their opinions. For example, people quickly post their reviews online as soon as they watch a movie and then start a series of comments to discuss about the acting skills depicted in the movie. This kind of information forms a basis for people to evaluate, rate about the performance of not only any movie but about other products and to know about whether it will be a success or not. This type of vast information on these sites can used for marketing and social studies. Therefore, sentiment analysis has wide applications and includes emotion mining, polarity, and classification and influence analysis.

Twitter is an online networking site driven by tweets which are 140 character limited messages. Thus, the character limit enforces the use of hashtag for text classification. Currently around 6500 tweets are published per second, which results in approximately 561.6 million tweets per day. These streams of tweets are generally noisy reflecting multi topic, changing attitudes information in unfiltered and unstructured format. Twitter sentiment analysis involves the use of natural language processing to extract, identify to characterize the sentiment content. Sentiment Analysis is often carried out at two levels 1) coarse level and 2) fine level. In coarse level, the analysis of entire documents is done while in fine level, the analysis of attributes is done. The sentiments present in the text are of two types: Direct and Comparative. In comparative sentiments, the comparison of objects in the same sentence is involved while in direct sentiments, objects are independent of one another in the same sentence.

However, doing the analysis of tweets expressed in not an easy job. A lot of challenges are involved in terms of tonality, polarity, lexicon and grammar of the tweets. They tend to be highly unstructured and non-grammatical. It gets difficult to interpret their meaning. Moreover, extensive usage of slang words, acronyms and out of vocabulary words are quite common while tweeting online. The categorization of such words per polarity gets tough for natural processors involved. This project uses R studio for fast processing capabilities to analyse sentiment from such high velocity real-time tweets and to represent in graphical representation form.

The rest of this paper report is structured as follows. In Chapter II, we detailed some related work by highlighting important features. Next, Chapter III gives brief details about the technologies used and it Cover details of methodology & implementation of the topic. Also the problems we came across and the challenges resolved during implementation are specified in chapter. Chapter IV gives a brief idea of topic results (Along with detailed description of Result). Chapter V details about the conclusion and future scope.

II RELATED WORK

Sentiment Analysis is an approach of studying people's emotions and sentiments based on a particular topic, service, event and it attributes. Sentiment analysis appears as a part of various business analysis systems to find opinions about their services or products [3]. Both the availability of CPU resources and enormous amount of data generated by the users makes sentiment analysis an active research field in upcoming years. Most of the existing approaches focus on efficient feature extraction, at the same time some approaches focus on extracting semantic features, which makes mush contribution to sentiment analysis. This paper gives a comprehensive overview on Sentiment analysis using NLP (Natural Language Processing) techniques. First, we start with some NLP techniques which are generally needed for preprocessing the input data[5]. Second, we introduce various approaches and some problems related to sentiment analysis and discuss some challenges and problem in sentiment analysis. Finally, we illustrate recent trend in sentiment analysis and its related works .

III METHODOLOGY

3.1Work done in the project is depicted as follows:

STEP 1: GETTING DATA FROM TWITTER:

Getting Twitter API keys

In order to access Twitter Streaming API, we need to get 4 pieces of information from Twitter: API key, API secret, Access token and Access token secret. Follow the steps below to get all 4 elements:

1. Create a twitter account if you do not already have one.
2. Go to https://apps.twitter.com/ and log in with your twitter credentials.
3. Click "Create New App"
4. Fill out the form, agree to the terms, and click "Create your Twitter application"
5. In the next page, click on "API keys" tab, and copy your "API key" and "API secret".
6. Scroll down and click "Create my access token", and copy your "Access token" and "Access token secret".

STEP 2: TWEETS PREPROCESS:

Sentiment analysis helps us gauge sentiment of tweets, however many of the tweets we get from the API might really not be 'classifiable' into some sentiment. In order to fit our model to our dataset we need to clean and process our data

1) *Steps for data cleaning:*

Removal of HTML tags and symbols: When we take data from web pages then some dynamic content is converted into html tags. The symbol @ is used for giving reference to any link or user.

```
tweet = p.clean(tweet)
```

Removal of Punctuations :For example: “.”, “,”,”?” are need to be removed to create text only file.

```
tweet = re.sub(r".", "", tweet)

tweet = re.sub(r",", "", tweet)

tweet = re.sub(r"?", "", tweet)
tweet = re.sub(r"\\\S+", "", tweet)

tweet = re.sub(r":", "", tweet)

tweet = re.sub(r"\"", "", tweet)

tweet = re.sub("-", "", tweet)

tweet = re.sub(r";", "", tweet)
```

Removal of Retweets: We will remove retweets from the list to provide a "pure" set of tweets.

```
tweet = p.clean(tweet)
```

Removal of URLs: URLs and hyperlinks in text data should be removed

```
tweet = re.sub(r"http\S+", "", tweet)
```

STEP 3: SENTIMENT EXTRTACTIONS

The package comes with four sentiment dictionaries and provides a method for accessing the robust, but computationally expensive, sentiment extraction so that you can quickly extract plot and sentiment data from your own text files.

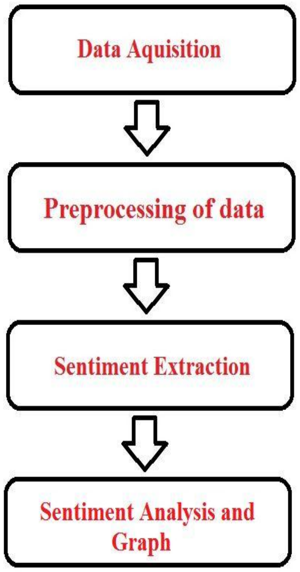
2) *Using NLP(NLTK Tool) and Clustering*

According to comparison based on word that these tokens combine with emotion-strengthening components to form sentiment then the system calculates sentiment value (positive/negative) and polarity score for tweets. This is how whole process of the extraction of text file in the sentiments is done.

STEP 4: SENTIMENT ANALYSIS AND GRAPH

The count of sentiment evaluated from the tweets text is represented using graphical visualization[6].

3.2 Proposed Architecture



3.3 Basic Step

Step -1. Data acquisition
Collecting the reviews and comments of social media sites.
Step -2. Processing of data
This step involves filtration and removal of irrelevant content comments from the collected dataset.
Step -3. Sentiment Extraction
Now from the filtered dataset extraction of user sentiments and emotion using NLP(NLTK Tool) And Clustering
Step -4. Sentiment analysis and Visualization

Final from extracted result classifying the feedback as positive or negative and representing in graphical form.

DATA COLLECTION:
Sentiment analysis will be performed on user tweets oftwitter. So, data required for analysis is tweets, it will be collected from twitter using API (Application Programming Interface). This API provides gateway for accessing data.

DATA PREPROCESSING:
Data collected from tweets cannot be used directly. It contains Emojis, special characters, retweets, user details and timestamps which is unnecessary. We need only text data for analysis. Text data will be separated is CSV (Comma Separated value) file.

NATURAL LANGUAGE PREPROCESSING:
Applying NLP algorithms on text data will interpret the tweets as positive and negative based on text expressed in tweets.

CLUSTERING:

The keywords expressed helping to manipulate string as positive or negative will be clustered into groups. Various clustering algorithms are available for clustering.

GRAPH VISUALIZATION:

Using the tweets positive and negative value, the graph will be plotted using graphic library. the graph will help to understand the summarized effect of of post on general public about that particular event.

IV. RESULTS AND DISCUSSION

A live Twitter comment is collected under the keywords entered by the user. Approximately over 2000 tweets are then stored as a csv file for analysis[12]. The chosen classifier for this work is a Naive Bayes Classifier utilizing the text processing tools in NLTK and their capacity to work with human language data. It is trained on tagged tweets and then used to analyses the sentiment in the tweets about the searched topic. we came to know about clusters that our sentiment scores belong to both polarity wise and subjectivity wise. Pre-defined dictionaries or sentiment tools can’t contain the proper score of every word in the context to a sentence, hence forming a cluster of the results from the tool score, we are able to group ‘definitely’ positive and ‘definitely’ negative tweets. The result is represented in the form of a graph which shows the count of comment have positive opinion on the searched topic as compared to the ones have negative opinion or are neutral[14].

V. CONCLUSION AND FUTURE SCOPE

In this paper, we are using twitter data and applying NLP on it and performing clustering approach to classify the tweets. Using these techniques, the accuracy of previous result will be enhance d. Then all similar type of data will be separated, to provide decision making results and finally this result can represented in graphical form. The task of sentiment analysis is still in the developing stage and far from completion, then NLP tool will will be used so we can train data and apply algorithms to get the accurate feedback.. So, we propose a couple of ideas which we feel are worth exploring in the future and may result in further improved performance.

REFERENCES

[1] Peng, Zhichao, Qinghua Hu, and Jianwu Dang. ”Multi-kernel SVM based depression recognition using social media data.” International Journal of Machine Learning and Cybernetics (2017): 1-15.

[2] Banitaan, Shadi, and Kevin Daimi. ”Using data mining to predict possible future depression cases.” International Journal of Public Health Science (IJPHS) 3.4 (2014): 231-240.

[3] Abhyankar, Anjali. ”Social networking sites.” SAMVAD 2 (2011): 18-21.

[4] Braithwaite, Scott R., et al. ”Validating machine learning algorithms for twitter data against established measures of suicidality.” JMIR mental health 3.2 (2016).

[5] Tripathy, Abinash, Abhishek Anand, and Santanu Kumar Rath” Document-level sentiment classification using hybrid machine learning approach.” Knowledge and Information Systems (2017):1-27.

[6] Yousefpour, Alireza, Roliana Ibrahim, and Haza Nuzly Abdel amed. “Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis” Expert Systems with Applications 75 (2017): 80-93.

[7] Hussain, Jamil, Maqbool Ali, Hafiz Syed Muhammad Bilal, Muhammad Afzal, Hafiz Farooq Ahmad, Oresti Banos, and Sungyoung Lee. ”SNS based predictive model for depression.” In International Conference on Smart Homes and Health Telematics, pp. 349-354. Springer, Cham,2015.

[8]R. Joshi and R. Tekchandani, "Comparative analysis of Twitter data using supervised classifiers," 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, 2016, pp. 1-6. doi:10.1109/INVENTIVE.2016.7830089

[9] I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[10] M. Kumar and A. Bala, "Analyzing Twitter sentiments through big data," 2016 3rd International Conference on Computing for Sustainable Global evelopment (INDIACom), New Delhi, 2016, pp. 2628-2631.

[11] R. A. Ramadhani, F. Indriani and D. T. Nugrahadi, "Comparison of Naïve Bayes smoothing methods for Twitter sentiment analysis," 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Malang, 2016, pp. 287-292.

[12] Deng, Li and Yu, Dong.Deep Learning: Methods and Applications.2014. NOW Publishers,United State of America.

[13] Miachel, Ray. 2012. 3 steps of text mining [Online] Available at: http://www2.cs.man.ac.uk/~raym8/comp38212/main/node203.html [Accessed 20 May 2017]

[14] Tomar, Shubham Simar.2017.Text Mining in R: A Tutorial [Online] Available at : https://www.springboard.com/blog/text-mining-in-r/[Accessed 20 May 2017]

Authors Profile

MS. Ashwini S. Yerlekar Bachelor of Engineering from Nagpur Universityin 2010 and Masters in 2014.Currently she is working as Assistant Professor in Department of Computer Science and Engineering.She is a member of ‘computer society of India. Her area of interest is Data Mining and Machine Learning.she has 5 years of teaching experince and 2 years of Industry experience.

MS. Devika Deshmukh Bachelor of Engineering from Nagpur Universityin 2007 and Masters in 2013 from Bhopal university. .Currently she is working as Assistant Professor in Department of Information technology.She is a member of ‘IEEE society’ Her area of interest is Data Mining and Machine Learning.she has 5 years of teaching experince and 2 years of Industry experience.