International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

Collective Intelligence Based Opinion Mining of Social Data

Hemavati*[1], Lakshmi K V[2], Punyashree K[3], SindhuPriya M A[4]
[1]Assistant Professor, Department of ISE, SIT, Tumakuru, Karnataka, India
[2-4]Department of ISE, SIT, Tumakuru, Karnataka, India

ABSTRACT

The most critical factors in today's world in formulating our views and influencing the success of a brand, product or service are opinions and reviews that are accessible to us. With the advent and growth of social media, stakeholders often express their opinions on popular social media, namely twitter where the data is extremely informative and it presents a challenge for analysis because of its immense volume, disorganized nature, difficulty in verifying the data authenticity and data security. In this system, the tweets are fetched from Twitter via Twitter API which uses open standard for authorization OAuth which also provides security for the users. This massive volume of data fetched are then stored into Hadoop Distributed File System (HDFS). The system can also take the input from local Excel file for processing. Efficient pre-processing techniques are applied to get sentiment for the user's opinion. The results are depicted in the form of graphs.

I. INTRODUCTION

Today, the generation of sheer volume data is enormous and tedious task is to make out sense from that data. The resources for such data are social sites like portals, e-commerce websites, blogs and online discussions. The new buzzword for many business strategies is analysing data from these social networking sites. The attitude, taste, opinion, interest, mood, reviews of the public regarding many ongoing issues in the society which are spread across various social networking sites can be effectively analysed and evaluated using Opinion mining and is also called as Sentimental analysis, Opinion extraction, Subjectivity analysis, mining of reviews, Emotion analysis, Sentiment mining or Affect analysis.

The process of computationally deriving the view of a speaker by determining the polarity of the text which may be positive, negative or intermediate is popularly known as Sentimental analysis. Emotion detection is involved in Sentimental analysis and polarity detection is focused by Opinion mining which is often one of the steps in Sentimental analysis. Hence these two fields are combined together.

Data mining uses a collection of homogenous data for experimentation task. The information that has been gathered during a survey is collectively known as dataset which contains data regarding products, movies, hospitals, celebrities, images, politics etc. Veracity of the data i.e. the quality of the data is the main concern. The authenticity of the data is difficult to verify. Therefore in the proposed work, the data is collected from Twitter using Twitter API [4]. Twitter is a significant platform where people express their views on various ongoing issues in the world and humongous amount of data in the order of billions of kilobytes are produced daily. Open Standard for Authorization OAuth [4] is used by the Twitter to provide authorized access to its API and also provides security for users. It is used by the users to login to any third party websites without

disclosing their passwords and by using social networking sites account. To understand the opinions of people, such huge amount of data needs to be analysed. Hadoop Distributed File System (HDFS) [3] is used for storing large amount of data. The stored data in Excel file can also be given as a input for analysis using an Excel file uploader. Data thus collected may contain many misspelled words, slang words, numerical, hyperlinks, emoji which needs to be corrected during pre-processing stage to make analysis easier. Then the polarity of the opinion is classified as positive, negative or neutral and also undecided reviews are handled properly. Results are then represented using graphs.

Collective Intelligence [11] is a group or shared intelligence that emerges from the collaboration, collective efforts and competition of many individuals and appears in consensus decision making.

This approach is used by many business organizations to determine various marketing strategies [7], technical decisions, improvement of product quality [8], customer services, market trends, customers buying pattern, to improve customer relationship [9] and also to improve campaign success to name a few.

II. RELATED WORK

There has been significant research in the field of Sentimental analysis and opinion mining. The work proposed by Pooja Kherwa [1], categorizes the data into positive and negative. Data about which sentiment analysis is to be done is automatically extracted from the web like online forums, popular portals, e-commerce websites and analysis is done to summarize the sentiments for each feature. This analysis is broadly classified into parsing and scoring. The result is then visualised using Google Chart API. The system fails to function effectively for the modern informal writings, misspelled or slang words used by the users. The work proposed by Ajit Danti [2] has the highest efficiency rate as measured by using Kappa statistics of 77.7% and also has lower error rates where the system generates the count of score words into seven categories such as strong-positive, positive, negative, weak-negative, weak-positive, neutral and strong-negative and are evaluated using machine learning algorithms like Naïve Bayes, Support vector machine and Multilayer Perceptron. Sentiment analysis has been done on the movie reviews where the reviews were collected from the web using web crawler. Pre-processing techniques were applied such as stop words removal, stemming and POS tagging. Efficacy was calculated for each of the machine learning technique by calculating the Precision, Recall and F-measure. The work that has been proposed by Sunny Kumar [3] deals with the efficiency of the storage and retrieval of huge amount of data for analysis purpose. In this system, Rhadoop connector has been used which stores whole data into Hadoop (HDFS) and R is used for fetching data from HDFS (Hadoop) and performs analysis on that data. This system has combined the processing (data analysis) power of R Language and the efficient storage technology provided by Hadoop that is, HDFS (Hadoop Distributed File System). Thus this has become an efficient approach to handle large volume of data from various networking sites. The work that has been proposed by Hase Sudeep Kisan [4] evaluates the sentiments of Twitter data by using StandfordNLP Libraries implemented in SaaS (cloud) which handles all current affairs in the world. The hash tag notation '#' has been used to categorise the tweets. The tweets related to various topics have been fetched from the Twitter using the hash tag notation. The occurrence of the keyword has also been counted to calculate the popularity of that keyword. This system has used REST API which gives access to fetch, read and write data from the Twitter. OAuth is used by Twitter to given authorised access to its API. OAuth also provides security for the users. The work that has been proposed by Sumit Kumar Yadav [5] has used Hadoop framework for handling large amounts of data. Bloom filter technique was used by building a

log using database which searches results faster and accurately. Tree data structure has been used for analysis. Advantages of Bloom filter is space efficiency, faster construction, efficient membership testing. Hashing technique with locality sensitive has been used to remove errors of the bloom filter results. Negating positive sentences and negating negative sentences analysis has also been performed. In this approach, Hadoop provides large amount of dataset storage platform and also provides faster processing in limited time.

III. SYSTEM ARCHITECTURE AND WORKING

The architecture of the proposed system is as shown in the Fig. 1. The proposed system performs the Sentimental analysis for dynamic tweets as well as stored local data by using efficient pre-processing techniques and based on the threshold concept. The system implementation is based on MVC (Model View Controller architecture. Any user, who wants to perform Sentimental analysis, needs to register to the system. As soon as the user logs into the system, he can obtain the data from Twitter and that obtained data is stored in HDFS (Hadoop Distributed File System) or the user can also select the stored local data by using Excel file uploader which is then pre-processed and threshold value is calculated as follows: for each word in a tweet which is neither stop word nor a noun and not present in the training set, sentiment type is detected and its corresponding positive, negative and neutral count is incremented by one. If the sentiment cannot be detected for any tweet then, it not considered as an undecided tweet. The word for which sentiment type is not detected, its sentiment is decided based on the positive, negative and neutral count. If the positive count is more than the negative and neutral count then that word is considered as positive. If that word crosses the threshold value, then that word's sentiment is considered. Classification is performed based on the training set and the results are displayed using graphs.

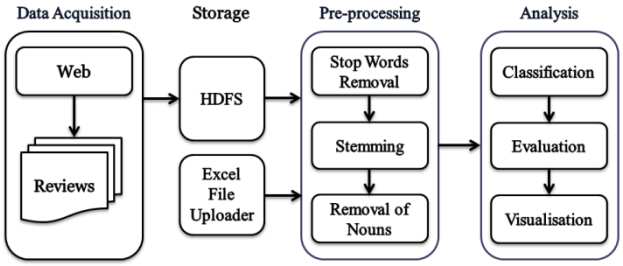

Fig.1 Architecture of the proposed system

The various phases of the proposed system are as follows:

A. Registration and Login

If any user wants to perform Sentimental analysis on any topic, first the user need to register to the system by entering the Name, E-mail, Password, Date of birth and Gender. Each field is validated by using regular expression and the appropriate error message is displayed for the user if the user doesn't enter valid information. Logging in to this system can be done by entering the registered user name and password.

B. Data acquisition and storage

In this system data is retrieved from Twitter API dynamically [10] based on the keyword that has been given by the user or stored local data can also be selected as dataset by using an Excel file uploader. At a time only 100 tweets can be fetched from the twitter API. The obtained tweets from Twitter are then stored into Hadoop Distributed File System (HDFS) which is an efficient way to store large amount of data.

C. Pre-processing

Pre-processing is an important phase in Sentimental analysis which includes the following steps:

1) Removal of Special characters, hyperlinks, numerals: Dataset may contain numerals, whitespaces, emoji, hyperlinks and many special characters. So it is necessary to remove them in pre-processing step. This can be achieved by writing the appropriate regular expressions.

2) Stop words removal: Stop words are the words that do not have much importance in the analysis process. While searching for phrases that include these, stop words may cause problem. Few of them are 'and', 'is', 'was', 'were' etc. Removal of stop words improves the efficiency as it reduces the number of words for analysis. Database containing these stop words are used in this system.

3) Stemming: Stemming is referred to as a process that chops off the ends of the words to achieve the goal correctly most of the time, and it includes the removal of derivational affixes. Porter Stemmer algorithm is the most widely used algorithm for stemming purpose. But this algorithm has two limitations. They are over-stemming and under-stemming.

Over-stemming: This situation occurs when two words with different stems are stemmed to the same root.

Under-stemming: This situation occurs when two words that should be stemmed to the same root remains distinct even after stemming.

Our proposed system can overcome these limitations by using a new approach that uses a database containing the derivational forms of words along with their corresponding root words as a pair. This database is used for stemming purpose.

4) Removal of nouns: As the tweets are used for Sentimental analysis process, most of the tweets contain the keyword about which we are fetching the data from Twitter and those words are nouns. Nouns do not have any importance in the analysis process. So it is necessary to exclude them from sentimental analysis process. A database containing the few possible nouns is used for this purpose.

D. Analysis

This phase mainly concentrates on sentiment detection of the given dataset which involves the following sequence of steps:

1) Classification: Classification of tweets into positive, negative, neutral and undecided can be done considering the database which contain list of positive, negative and neutral words. The classification of the dataset is done based on the training set.

2) Comparison technique based on threshold value: Few words in the dataset are neither stop words nor nouns which are not removed and also, sentiment for these words cannot be detected by using the training set. So, this technique is used which works as follows: for each word of this type in the dataset, sentiment type is detected and its corresponding positive, negative or neutral count is incremented by one. The word for which sentiment type is not detected, its sentiment is decided based on the positive, negative and neutral count. If the positive count is more than the negative and neutral count then that word is considered as positive. The same technique is used for deciding a word as negative or neutral.

3) Evaluation: The dataset thus classified is then evaluated based on the sentiment for each tweet and is stored in the database. Total count of the positive, negative, neutral and undecided tweets are determined and are displayed along with the tweets and their corresponding sentiment type.

4) Visualization: The obtained results are displayed in the form of bar graph. Graph represents the number of tweets that are positive, negative, neutral and undecided.

IV. EXPERIMENTAL RESULTS

The result which is depicted in the graph shown in Fig. 2 represents the total number of tweets fetched, total number of positive, negative, neutral and

undecided tweets. The data has been fetched from Twitter via Twitter API by the using the keyword "bahubali". Maximum 100 tweets can be fetched at a time. The most recent 100 tweets are fetched from the Twitter for one fetch operation.

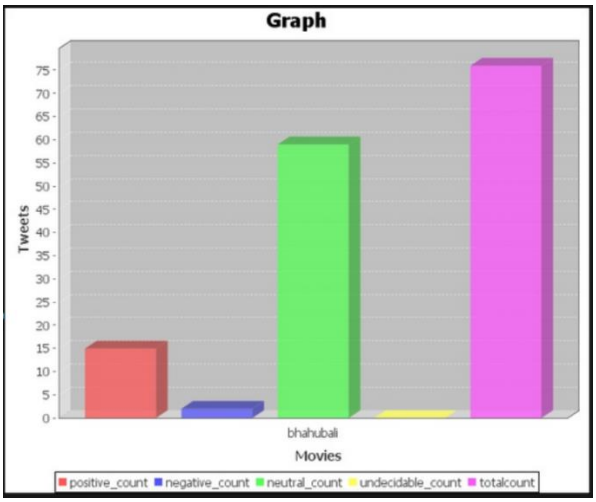

Figure 2. Graph shows the total number of tweets fetched along with total positive, negative, neutral and undecided tweets for "bahubali" keyword.

The sentiment for two or more keywords can also be compared by using the proposed system. Fig.3 shows the comparison of sentiments for the two keywords "bahubali" and "Tagaru".

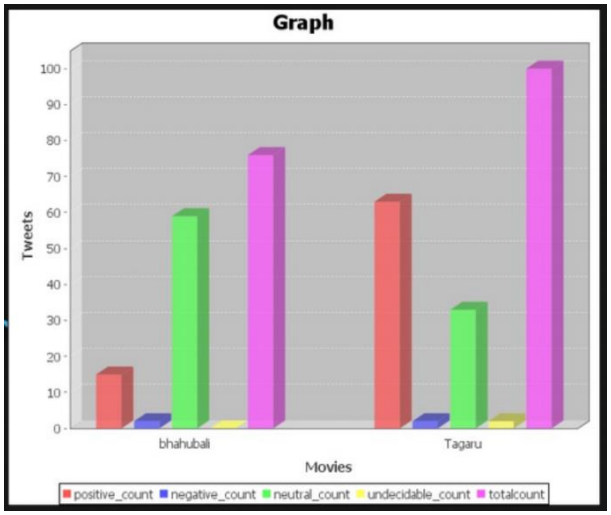

Figure 3. Comparison of sentiments for the keywords "bahubali" and "Tagaru"

IV. CONCLUSION AND FUTURE ENHANCEMENT

The proposed system handles the huge volume of data which is needed for analysis by using the Hadoop Distributed File System and uses efficient pre-processing techniques for performing Sentimental Analysis of the given dataset. The classification of the result is done by using training set. The results are then visualised using easy to analyse representations like graphs. The system can also be used to compare the popularity of the two aspects based on their reviews in Twitter.

The system can be made efficient by using still larger dataset which ensures the certainty of the collected data. By using a more matured training set, the classification can be done effectively [6]. The system can also be improved by using efficient classification and evaluation techniques like Naïve Bayes classifier, SVM.

V. REFERENCES

[1]. Prof. Pooja Kherwa, et al. "An approach towards comprehensive sentimental data analysis and opinion mining." 2014 International Advance Computing Conference (IACC). IEEE, 2014.

[2]. Shoiab Ahmed and Ajit Danti, "A novel approach for Sentimental Analysis and Opinion Mining based on SentiWordNet using web data." 2015 International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15). IEEE, 2015.

[3]. Sunny Kumar, Paramjeet Singh, and Shaveta Rani. "Sentimental analysis of social media using R language and Hadoop: Rhadoop." 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). IEEE, 2016.

[4]. Hase Sudeep Kisan, Hase Anand Kisan, and Aher Priyanka Suresh. "Collective intelligence & sentimental analysis of twitter data by using StandfordNLP libraries with software as a service (SaaS)." 2016 International Conference on Computational Intelligence and Computing Research (ICCIC). IEEE, 2016.

[5]. Devendra K Tayal and Sumit Kumar Yadav."Fast retrieval approach of sentimental analysis with implementation of bloom filter on Hadoop." 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT). IEEE, 2016.

[6]. Shoiab Ahmed, and Ajit Danti, "Effective sentimental analysis and opinion mining of web reviews using rule based classifiers." Computational Intelligence in Data Mining-Volume 1, pp 171-179. Springer, New Delhi, 2016.

[7]. Sunil Kumar Khatri, and Ayush Srivastava."Using sentimental analysis in prediction of stock market investment." 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). IEEE, 2016.

[8]. Hui Song, et al. "Extracting product features from online reviews for sentimental analysis." 2011 6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT). IEEE, 2011.

[9]. Walter Kasper, and Mihaela Vela. "Sentiment analysis for hotel reviews." Computational linguistics-applications conference. Vol. 231527. 2011.

[10]. https://apps.twitter.com/

[11]. https://en.wikipedia.org/wiki/Collective_intelligence/