Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Real time Extraction of Sentiment and Analysis for Decision Making

Vishwasa Navada K, Naveen D Chandavarkar
Dept. of Computer Science, NMAM Institute of Technology
Nitte, Karnataka, India

Abstract— Since last decade, Social Media has become the basic need of the internet users around the globe. With more than 3 billion internet users, social media no longer remains as a platform only to share personal activities. So whatever users share in social media websites can have useful information in it, but the data is very huge and picking the right thing out of it is not an easy job for a human being. In this paper, we discuss the ways of extraction and processing the data from social media websites like Facebook, Twitter which can be used by customers to decide whether to buy a product and businesses to make decisions in their sales, marketing, advertising etc. Sentiment analysis is one of the leading concepts in analyzing and categorizing huge and unprocessed data. After categorizing the data into positive, negative, and neutral statements, it is easy to use them as input for business as well as domestic purposes.

Keywords— social media, internet, data, sentiment analysis, opinion mining, business

I. INTRODUCTION

In 2015, number of internet users all over the world crossed 3 billion [1], which is almost 40% of the world population. This popularization of the Internet has radically changed the forms of human interaction through instant messaging, forums, social media, ecommerce etc. In the beginning, social media was just a platform where people used for interacting and content sharing, but now a day's availability of internet and wide range of social networking websites altered people's mind such that they start sharing their personal feelings, opinions etc. As a result of this, every social media service ends up with a huge random data every day. Those data will be consisting of product reviews, discussions about a particular event, complaints, gossips etc.

It is very important for a company to know the opinion of people have about their products. It plays a crucial role in taking decisions related to sales, resource management, etc. For a company with sales worldwide it is very difficult to collect opinions or reviews individually. Growth of social media has made reaching people easy. Even in this process reaching each and every one is time consuming and not an economic process. Because of these problems, many companies decided to spend money on social media monitoring. Social Media monitoring is *"an active monitoring of social media channels for information, generally applied to content sources like blogs, social networking site, forums etc"* [2].

Since the data available in social media is very huge and unprocessed, it is a challenge to pick the required information from it. This challenge can be solved by employing certain algorithms to process them. One of the ways in which it can be made possible is by using the concept of Sentiment Analysis. It is a process that uses natural language processing and computational linguistics to identify and categorize the information. With this concept we can analyze and extract the required information from the unprocessed data and use them for the benefit of a product or a company.

With the help of all above mentioned concepts, we discuss the process of extracting and processing the related information from the social media websites. This paper briefly explains the process of sentiment analysis with respect to the twitter as an example. The reason behind why we chose Twitter is that it is one of the most used social media platform in the current time. More over it is easy to get data from a source which allows you to extract the data easily and in real time. We also cover how this process can be expanded with other social networking websites as well as the methods of using the result for useful purposes.

II. RELATED WORK

Ever expanding social media has been attracting many research scholars since its inception**.** A number of works can be found which explain the application of social media monitoring on various areas like social issues, health and business.

A research paper by Magdalini Eirinaki, Shamita Pisal, Japinder Singh proposes a concept of Opinion Search Engine [3] and it is implemented in the form of a Web Application named "AskUs". They used two algorithms to get their concept working; they are High Adjective Count (HAC) and Max Opinion Score. Their web application uses High Adjective Count (HAC) algorithm to identify the key features of a particular product such as weight, speed, performance. It basically works on the principle that the nouns for which reviewers express a lot of opinions are most likely to be the important and distinguishing features than those for which users don't express such opinions. This algorithm also counts the number of adjectives used to describe a feature and assigns score to respective features. These scores are referred as opinion scores throughout the paper. Once the features are identified and given scores, system uses Max Opinion Score

algorithm to rank them according to the scores assigned. This algorithm carries out sentiment analysis to identify the sentiment behind the adjectives used to describe the features. As the process is applied to each sentences separately, after completion of processing a review scores of each sentence is summed up to get average score for each listed feature. With 87% accuracy, their web application allows users to decide which product to buy easily.

A research by Ye Wu and Fuji Ren [4] shows us the ways to categorizing Twitter tweets into influential and influenced tweets. Here influential tweet refers to a tweet that is said to make a sentimental impression on the readers mind and influenced tweet is nothing but the response tweets that are expressed after being influenced by the influential tweet. They have explained Influence Probability Model based on the assumptions such as a tweet with more number of retweets can be called as an influential tweet and reply tweet can be called as an influenced tweet. Their concept first follows sentiment analysis to categorize the tweets into positive, negative, and neutral sentences. After categorization they found that their dataset had more number of negative tweets indicating that users prefer to tweet when they are in bad mood. Once they have categorized tweets they calculated influence probability and found that the scatter plot of influencing probability vs. influenced probability shows a great correlation between them. Their research inferred that an active user who tries to influence his surroundings more would also get influence from his surroundings keeping the balance between influential probability and influenced probability.

Precise Tweet Classification and Sentiment Analysis paper by Rabia Batool, Asad Masood Khattak, Jahanzeb Maqbool and Sungyoung Lee [5] tells us about their enhanced sentimental analysis approach for extracting keywords, entities from the tweets. The proposed architecture is divided into various modules such as preprocessor, knowledge generator, knowledge enhancer, synonym binder and filter engine. In each module the information gets processed more accurately and categorized into domain specific and sentiment categories. Preprocessor module acts as a translator which converts slang words and short forms such as *plz, gud* etc to correct format which makes sentimental analysis easy. Knowledge Generator gets the preprocessed data from the Preprocessor module and applies natural language processing and machine learning techniques to it. After this process of Knowledge Generation we get outputs like Sentiment involved in the sentence and topic to which that tweet is related to. This output is not accurate because it only extracts only one or two keywords that a tweet has. So Knowledge Enhancer further processes the tweet and extracts more relevant keywords from it. Synonym binder is used to identify synonyms from various tweets and consider them as one. Binding them as one makes the process efficient and time saving. Filter engine is used to filter out the domain specific tweets from whole dataset. They said that this multi module approach allows us to extract domain specific and keyword based information from a large amount of tweets with more precision.

A case study named *Business Intelligence From Twitter For The Television Media* [6] by Jayanth Marasanapalle, Vignesh T.S and others contains a study on twitter tweets about a TV show. It covers 3 major areas in which twitter is mined; they are trend detection, sentiment mining and user characterization. They considered one episode of an English comedy show which featured a celebrity interview. They extracted tweets before, during and after the show was aired. When they plotted a line graph on number of tweets vs. time they found that there were a small number of tweets before the show started and increased suddenly when show was started. Even though number of tweets gradually decreased when show was going on, they saw a rapid increase when celebrity interview was started. This result shows us how people's interest about the show decreased gradually but increased due to the celebrity interview which can be called as a main attraction of the show. They also extracted personal information of the viewers such as gender, location etc to study nature of the viewers of that show. This data can be used to predict viewers of other similar shows. They also employed Sentiment analysis to do a complete analysis on nature of tweets about that TV show they collected. This analysis fetched them the positive and negative reviews of the show. At the end of the case study they have mentioned how this process can be utilized to start new TV shows, know public interest and popularity prediction.

Predicting the Future With Social Media [7] by Sitaram Asur and Bernardo A. Huberman adapts sentimental analysis to predict success rate of movies. This paper proposes a movie popularity and box office prediction model which is very useful for movie makers for predicting their revenue. The authors conducted the test on twitter tweets in 3 time periods, first one being the pre-release period, followed by the first weekend after the release and finally one week after the release. They stated that most of the tweets from the pre-release period will be having urls as there will tweets about posters, trailers, photos etc. Hence we can also predict orally that most of them will be positive in nature. This prediction was found to be matching with the result obtained from performing sentiment analysis on pre-release tweets. Success of the movie can be predicted from the revenue collected on first weekend. Popularity of the movie is calculated from the ratio of tweets with positive sentiment to tweets with negative sentiment. A movie that has far more positive tweets that negative tweets is considered to be popular and successful. This paper also proposes regression models to estimating the revenue of movies. After conducting several experiments for different movies they came to know that their way of predicting the success rate of a movie gives the result which is comparable with the theoretical predictions based on newspaper reviews and other standard methods followed so far.

III. METHODOLOGY

In this section, we will discuss general method of extracting, processing, and categorizing of unprocessed data from social media websites. In order to make it simple, we divided the whole process into 3 modules; they are Extractor, Preprocessor, and Analyzer.

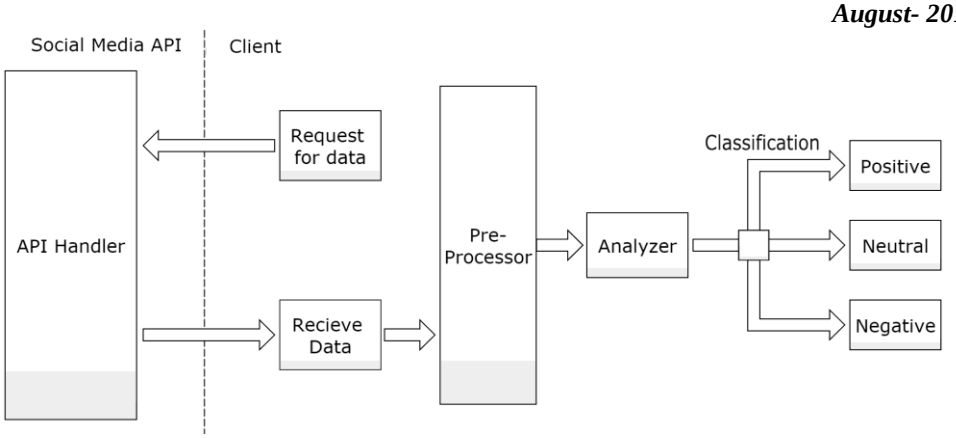

Figure 1 Block diagram representing the methodology

a. Extractor Module

As the name suggests, this module acts as an intermediate layer between the social media website from where you need to extract data and the web application that we developed for conducting sentiment analysis. Most of the social media website will be having their own Application Program Interface (API) [8]. API is a set of routines, protocols, and tools provided by a particular software or Web Service to allow others to interact and manage their data. So, by using respective API of social media website we can easily access data available from it. We just have to configure the API and then use suitable procedures to request for the data. API handler at the server processes it and provides you with the requested data.

The process of extraction can be done in two ways; 1) Real time extraction, where we extract particular data whenever necessary. 2) Building a dataset by extracting huge data belonging to different time periods. Either way, analyzing process remains same. Each one of them has its own field of application. We can choose real time extraction when we need to conduct sentiment analysis on particular event or similar things. By building and analyzing huge datasets, we can get more accurate results and can be used on the topics like predicting movie revenue, political campaign etc. Once the data is ready for further processing, it is sent to preprocessor to do some alterations to the raw data.

b. Preprocessor

Raw data from social media websites can have following elements,

- Website URLS
- Image links
- Usernames, hashtags
- Unnecessary punctuation marks
- Extra spaces and line breaks

Analyzing the raw data as it is can decrease efficiency and accuracy. So this module removes the above listed elements and sanitizes the data to clean text format using regular expressions. For higher precision, we can use external dictionaries to convert short forms and abbreviations to correct English. Preprocessing the data increases the time efficiency of the analyzer and helps to perform accurate categorization of the data. Once the preprocessing is complete, data can be sent to next module directly or store it in database to analyze later.

c. Analyzer and Classifier

This module plays an important role in the whole process. We came across many classifiers like Naive Bayes, Support Vector Machine (SVM), Logistic Regression, Decision trees, K Nearest Neighbors, etc. Every classifier has its own advantage over the other and also some classifiers depend on the type and quantity of data that need to be processed. We found in a blog article [9] that Naive Bayes Classifier shows greater accuracy for data that are independent of each other and data of manageable size.

Naive Bayes Classifier is a simple yet efficient classifier that is used to classify uncertain data into positive, negative, and neutral statements. Since there is no relation between each of the social media posts we extracted, the condition of independence for the Naive Bayes Classifier is satisfied. So we decided to employ this classifier model for our analysis.

It works on the principle of Bayes Probability theorem. That means, it will be having a set of known results for some words and it compares those results with the words in a particular sentence. Hence the overall sentiment of the sentence will be calculated. If we need to know overall sentiment of set of sentences same principle will be applied again. The equation for calculating the probability of a sentence to have a particular sentiment can be given by,

P (sentence to have a sentiment) = product of (P (words having the same sentiment))

i.e. $P(\text{sentence}|\text{sentiment}_i) = \Pi\, P(\text{word}_k|\ \text{sentiment}_i)$

Where,

Sentiment_i – any one of the sentiment in the list of sentiments (positive, negative, neutral)

$word_k$ – each word in the given sentence

Thus we can obtain the probability of sentiments of a particular sentence. Sentiment with greater probability is considered as the sentiment of the whole sentence. So in our case, once we give extracted data to the analyzer, each sentence in the data will go through the same equation and gets a sentiment label. By this way, we can categorize random data into positive, negative, or neutral sentences.

After completion of the whole process of analyzing the data from social media websites, we can use the generated information for many purposes. For example, if the search query is about a particular product then after the analysis we get a clear picture about public response for that product. Using that response companies can make decisions in their sales and marketing. Result from the analysis can also be used by common people before buying any product. Sentiment Analysis will also help us in recognizing major issues in particular products.

IV. IMPLEMENTATION

To implement our proposed model, we selected Twitter as a source of data. Twitter's API provides real time as well as old data. This makes us easy to extract and analyze the data. Twitter being one of the most populated social networking and micro blogging site, features Hashtags. Hashtags are nothing but words prefixed with '#' symbol that social media posts contain. Using hashtags make us easily identify group of data that is related to same thing or person

We developed a web application that allows you to enter a word or username on which you desire to conduct the analysis. Once the user submits the keyword in our web application, it makes a call to twitter requesting data related to that keyword. Twitter Developer Documentation [10] for fetching tweets from twitter shows us various parameters for specifying the type of data we need. For example, if we include 'result_type' as 'recent' in our call then twitter will send us the recent tweets about the requested keyword. We can also use 'popular' instead of 'recent' to get the most popular tweets. Other parameters are 'lang' for specifying the language of the tweets, count for number of tweets required etc. With the help of these parameters, it is very easy to narrow our search based on the topic. Our query is as follows,

```
$query = array(
"q" => "$keyword -filter:retweets",
"count" => "100",
"result_type" => "recent",
"lang" => "en",
);
```

Where, 'q' contains the keyword that should be searched and we used the filter to prevent twitter from sending retweets that makes us analyze same thing again and again. Since we are only analyzing English tweets, we included a parameter specifying the language.

Twitter servers receive the call and send the requested data. After receiving the data, we used pattern matching techniques to remove unwanted elements like URLs, punctuation marks etc from the data. Once the data is sanitized, we store it in the database. If we need to analyze the data from different time periods then we just need to repeat the process of requesting twitter. This completes the process of extraction and preprocessing. For analyzing, Naive Bayes classifier needs a set of known results, so we used an open source sentiment dictionary [11] that has almost all the English words along with their sentiment scores. Also this dictionary gives you the ability to identify and prioritize some of the simple emotions.

Working: To get overall sentiment of a particular query, we should calculate probability of sentiments for each of the tweets separately. So, we take each tweet from the database where we have stored extracted tweets and split them into bag of words [12]. Some words will not contribute to the sentiment of the sentence, so we ignore them during the analyzing. Computer cannot recognize the difference between a positive word and same word with a negative prefix, so whenever we find a negative prefix we remove the whitespace between the prefix and the word, which makes us easy to do the calculations. Working procedure of the analyzer is depicted as,

Step 1. For each tweet in the database
Step 2. Read a single tweet from database.
Step 3. Split the sentence into bag of words
Step 4. Consider each word from bag of words
Step 5. If this word is contributing to sentiment of sentence
Step 6. If it has a negative prefix
Step 7. Remove the whitespace after the prefix [End if]
Step 8. Search for word's score in the dictionary [End if]
Step 9. Add the score to the previous word's score. [end for]
Step 10. Finalize the sentiment of the tweet. [end for]

At the end of step 10 we will be having sentiment scores of each of the tweets present in our database. To make the user easy to understand the results obtained from the test conducted, we used pie charts to represent all the sentiment scores and overall sentiment of the analysis.

V. EXPERIMENTAL RESULTS

In this section, we show you how the analyzer module decides the results by taking some example datasets from twitter. You will get to know how exactly a sentence splits into bag of words model from the table below,

Table 1. Table showing analysis of simple sentences

Sentence	**Bag of words**	**ignored words**	**Considered Words**	**Sentiment**
This works perfectly.	{ “this”:1, “works”:1, “perfectly”:1 }	this, works,	perfectly	Positive
This mobile is bad.	{ “this”:1, “mobile”:1, “is”:1, “bad”:1 }	this, mobile, is	Bad	Negative
This phone has a front camera.	{ “this”:1, “phone”:1, “has”:1, “a”:1, “camera”:1 }	this, phone, has, a, camera,	NILL	Neutral
This is not good	{ “this”:1, “is”:1, “notgood”:1 }	this, is	notgood	Negative

As you can see from the table that sentence “This phone has a front camera” is shown as neutral since it is not describing anything. Figure 2 shows the output of our web application if the same sentences from the table are given as input.
We carried our analysis on 100 tweets during Motorola’s new mobile launch event which was trending in twitter with the hashtag #motolaunch (on 28th July 2015). We noticed that there were a small number of Negative responses from the public as the event started a bit late. But we got a significant number of positive as well as neutral responses which clearly shows us how much expectations people have about the new phone. Below is the pie chart of the analysis we did on the hashtag #motolaunch.

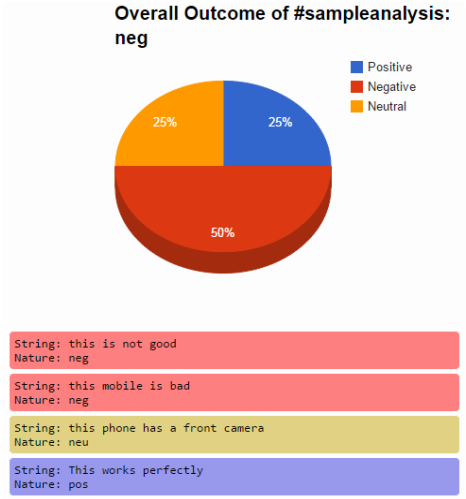

Figure 2 Screenshot of the Output for sample analysis

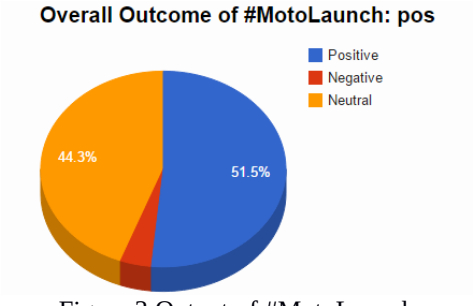

Figure 3 Output of #MotoLaunch

VI. FUTURE WORK

Even though our web application extracts limited amount of data from twitter, Twitter's new Streaming API [13] allows us to retrieve large amount of data in real-time if we build an enterprise server. For big data we can build systems to analyze huge data and get results time to time. Using multiple databases to store real time data and high precision algorithms to analyze, if we build a Data Warehouse to extract, analyze and report the public responses automatically.

Many companies are already spending to build social data warehouses to report public sentiment towards their products. It plays a crucial role in converting random quantity into quality. Companies with fewer infrastructures have started to use cloud storage to build their own data warehouse so that they can manage and control their resources according to customer's sentiments.

VII. CONCLUSION

Hence social media proves itself that it plays very important role in modern day business intelligence. Our web application allows you to check the public response of a particular topic instantly. It is a simple model which can further be developed to analyze big data using the technique of Data Warehousing. This method minimizes the human effort involved in manual extraction, analysis of data from social media websites and generating reports. Having a report on the public response of any products gives ability for a company and individual to make decisions related to that product. This process also supports a company to identify the issue related to their products.

REFERENCES

[1] List of countries by number of Internet users. (2015, July 23). In Wikipedia.[Online] https://wikipedia.org/wiki/List_of_countries_by_number_of_Internet_users

[2] Social media measurement. (2015, July 18). In Wikipedia. [online] https://en.wikipedia.org/wiki/Social_media_measurement

[3] Pisal. S, Singh. J, Eirinaki. M. "AskUs: An Opinion Search Engine", 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW) p.1243 – 1246

[4] Ye Wu, Ren F. "Learning Sentimental Influence in Twitter." 2011 International Conference on Future Computer Sciences and Application (ICFCSA) p.119 - 122

[5] Batool R, Khattak A.M, Maqbool J, Sungyoung Lee. "Precise tweet classification and sentiment analysis". 2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS), p.461 – 466

[6] Marasanapalle J, Vignesh T.S, Srinivasan P.K, Saha, A. "Business intelligence from Twitter for the television media: A case study", 2010 IEEE International Workshop on Business Applications of Social Network Analysis (BASNA), p.1 – 6

[7] Asur S, Huberman, B.A. "Predicting the Future with Social Media", 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (Volume:1) p.492 – 499

[8] Vangie Beal, "API-application program interface", [online] http://www.webopedia.com/TERM/A/API.html

[9] Edwin Chen, "Choosing a Machine Learning Classifier", [Online] http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/

[10] "Twitter Developer Documentation". [Online] https://dev.twitter.com/rest/reference/get/search/tweets

[11] PHPInsight. Available at GitHub. https://github.com/JWHennessey/phpInsight/

[12] Bag-of-words model. (2015, April 9). In Wikipedia. [Online] https://en.wikipedia.org/wiki/Bag-of-words_model

[13] The Streaming API, Twitter Developers. [Online] https://dev.twitter.com/streaming/overview

