International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 7, ISSN No. 2455-2143, Pages 46-51
Published Online May - June 2016 in IJEAST (http://www.ijeast.com)

WEB CRAWLING TO EXTRACT AND ANALYZE THE DATA FROM TWITTER AND COMPUTE THE PROBABILITY OF TWEETS

Manjunath Patil
Department of CSE
Reva Institute of Technology & Management
Bangalore, Karnataka, India

Prof. Pavithra P
Department of CSE
Reva Institute of Technology & Management
Bangalore, Karnataka, India

***Abstract* -In recent years, Social Networking Sites or Micro Blogging sites such as Twitter has become an exclusive tool for every update. Huge amount of data available on these social networking sites has created tremendous interest in researchers to extract and analyze the day to day sentiments of people. Twitter is one such place where people share their interests and gather information on the same. In the present era it is difficult to extract such data and analyze. Many organizations and institutions are under pressure to implement such social networking and micro blogging sites with their business. The application gives scope especially for extraction and analysis of data like students information, their tweets on various ongoing issues, their tweets on the problems they face during their academic career etc. It gives an expansive process of extracting, storing and analyzing the formless and vibrant data from the micro blogging or social networking sites. The application concentrates on analysis of the various problems faced by the students and thus reach out to them through this application and thus improve the marketing tactics.**

I. INTRODUCTION

Micro Blogging sites or Social Networking sites[1] such as Facebook and Twitter give a greater platform for students to share vast amount of diverse information like their joy and struggle, information on the books and the authors, information on different colleges, information on various subjects, their problems faced in the colleges etc. On many of the social networking sites, the data which is shared and discussed by students in a casual manner will not be in the right format and is a vibrant data. These tweets or the information shared on such sites by the students help many educational researchers with vast amount of knowledge to understand their issues, problems and their experiences experienced outside the institutions. Which in turn help institutions intervene in the issues and problems faced by students and thus help in improving the quality of education. This vast amount of data available on such networking sites may help the researchers and institutions to understand the student's issues and their problems, but it's really difficult and a complex task to extract and analyze such data shared by them. Imagine the volume of data exchanged between the students, the language and the slangs they used to tweet, the locations of their tweets, the time of their tweets, there may be attachments, there may be pictographically tweeted messages. With such huge amount of data, a casual analysis or by using any available algorithms cannot provide in-depth meaning of the data shared.

This vast amount of extracted data is of no use until it is converted into some readable and sensible format which is understandable by a layman. The process does not include just the extraction and analysis; but it also involves other processes such as Mining of the required data, Cleaning the extracted data, Integration, Transformation, Evaluating the data Pattern and Presentation. Only after all these processes are through, we will be able to use the extracted data for further analysis purpose as per the requirement and can be used in many other applications such as Fraud Detection, Current Market Analysis on various products, Sales of recently launched products, Reviews on various products, movies etc.

II. LITERATURE SURVEY

The paper[2] titled "*Academic pathways study: Processes and realities*" - the authors describe various methods and procedures such as activities inside classrooms, activities outside the classrooms, various interviews and surveys to gather the information related to students experiences and their day to day activities as part of their learning. These methods required lot of manual work and require lot of time, due to which they may have become obscured over time.

The paper[4] titled "*The state of learning analytics in 2012: A review and future challenges*" - the author namely R. Ferguson, who describes the data mining related to students and their education, is focused on extracting and analyzing the formatted data based on the technologies used in the classroom activities, based on the course management systems

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 7, ISSN No. 2455-2143, Pages 46-51
Published Online May - June 2016 in IJEAST (http://www.ijeast.com)

(CMS), or the environments used online for learning activities. But there is no research found to directly extract and analyze the student's education related data that is being posted on the networking sites with the aim of understanding the students' issues and their problems.

The paper[6] titled "*The Presentation of Self in Everyday Life*" - the author Goffman describes his theory, i.e the Goffmanan's theory of social performance to extract the most unstructured, vibrant and informal data from the social networking sites. This theory is mostly used to explain and describe the interactions of people on social sites. The most important and fundamental factor of this Goffman's theory is the notion of back_stage and front_stage of students interactions on social sites. With this approach, the information gathered from online interactions may require more authentication process and data may not be filtered as compared to other methods and theories used in the market today.

III. PROPOSED SYSTEM

The proposed system can be described by using the following diagram

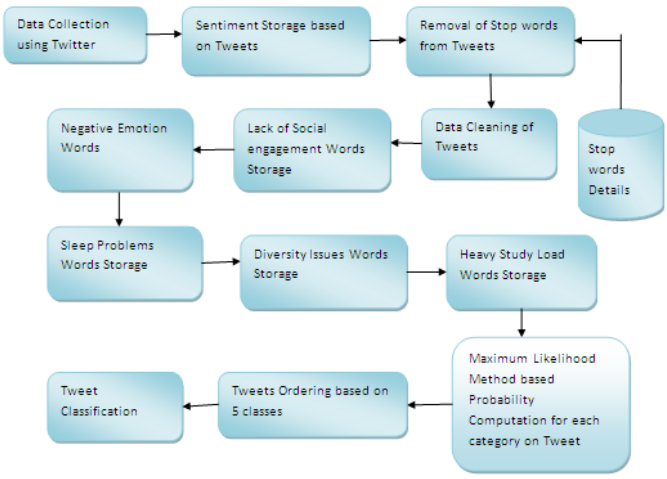

Fig. 1. Naïve Bayes Multi-Label Classifier

A. Collection of data using Twitter -

The application considers 20 accounts; the tweets are extracted from these 20 accounts and collected. Twitter Application Interface (API) is used to extract these tweets with the help of OAuth [3] API which is used to authenticate the login process to Twitter application.

B. Storage of collected Tweets/Sentiments -

This is the process of storing the collected tweets/sentiments in a database structure in terms of (Twitter_Id, Twitter_Desc, User_Id). Where Twitter_Id is an unique identifier associated with the tweet, Twitter_Desc is the description what the tweet is about, and User_Id is the unique identifier of the tweeted user.

C. Stopwords -

Stopwords are those words in the tweet descriptions which have very less meaning or the words which are almost meaningless. These words are also called as set of keywords defined by the data mining forum. These meaningless words are being filtered out from the tweet descriptions to reduce the size of tweets as much as possible. The list of stop words used in the application is as follows

a,able,about,across,after,all,almost,also,am,among,an,and,any,are,as,at,be,because,been,but,by,can,cannot,could,dear,did,do,does,either,else,ever,every,for,from,get,got,had,has,have,he,her,hers,him,his,how,however,i,if,in,into,is,it,its,just,least,let,like,likely,may,me,might,most,must,my,neither,no,nor,not,of,off,often,on,only,or,other,our,own,rather,said,say,says,she,should,since,so,some,than,that,the,their,them,then,there,these,they,this,tis,to,too,twas,us,wants,was,we,were,what,when,where,which,while,who,whom,why,will,with,would,yet,you,your

D. Data Cleaning -

This is the process to clean the tweets from stopwords or meaningless words. Once the tweets are cleaned from stopwords, the tweets are represented as (*Clean_Id ,Clean_Data ,User_Id*). *Clean_Id* is the unique identifier associated with the Tweet, *Clean_Data* is the clean data after removal of stopwords and *User_Id* is the unique identifier associated with the user.

E. Maximum Likelihood Method -

The cleaned tweets are categorized and the probability of each of the tweets belonging to a particular category is computed by following steps:

1. Tweets are categorized into 5 categories as shown below and the set of categories is represented as
   $C = \{c_1, c_2, c_3, c_4, c_5\}$
   where
   $c_1 = heavy\ study\ load$
   $c_2 = Lack\ of\ Social\ Engagement$
   $c_3 = Negative\ Emotion$
   $c_4 = Sleep\ Problems$
   $c_5 = Diversity\ Issues$
2. Denote N as the total number of words in a tweet.
3. A positive and negative probability is computed for each of the words in a tweet as given below:

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 7, ISSN No. 2455-2143, Pages 46-51
Published Online May - June 2016 in IJEAST (http://www.ijeast.com)

a) Positive probability - The probability of a word in a particular category *c* is computed as:

$$p(w_n \mid c) = \frac{m_{w_n,c}}{\sum_{i=1}^{N} m_{w_c,n}}$$

Where,

$m_{w_n,c}$ = *Number of times a word belongs to category*

b) Negative probability - The probability of a word in the categories other than *c* is computed as:

$$p(w_n \mid c^1) = \frac{m_{w_n,c^1}}{\sum_{i=1}^{N} m_{w_n,c^1}}$$

Where,

m_{w_n,c^1} = *Number of times a word does not belongs to category*

IV. HIGH LEVEL DESIGN

A. Data Flow Level 1 -

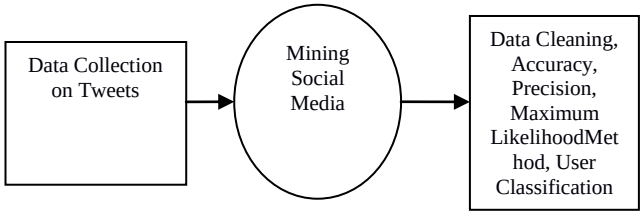

Fig. 2. Data Flow Diagram Level 1

Fig. 2 shows the Data Flow Diagram Level 0. As shown in the figure:

1. Data Collection from Tweets acts as an input where we read the data on the topics from twitter using OAuth API.
2. Mining Social Media using Naïve bias algorithm is responsible for finding the best tweets and to find the user classification
3. Data Cleaning, Accuracy, Precision, Maximum Likelihood Method, User Classification are the parameters which acts as an output of our algorithm

B. Data Flow Level 2 -

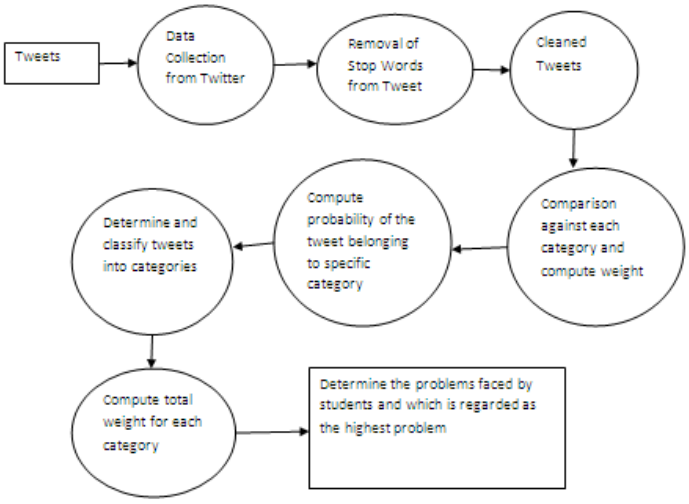

Fig. 3. Data Flow Diagram Level 1

Fig. 3 shows the Data Flow Diagram Level 1. As shown in the figure:

1. Tweets for various cases acts as input
2. The tweets are collected from the twitter using [11]OAuth API.
3. For each of the tweets the stop words are removed and clean tweets are also obtained.
4. Each tweet is then compared against the words belonging to 5 different categories.
5. Compute the probability using nave byes formula.
6. The sums of all probabilities are performed per tweet and across all tweets.
7. Compute the weight for each of the tweet.
8. The tweet is then associated with a category based on maximum probability.
9. Determine the problems faced by the students

C. Data Flow Level 3 -

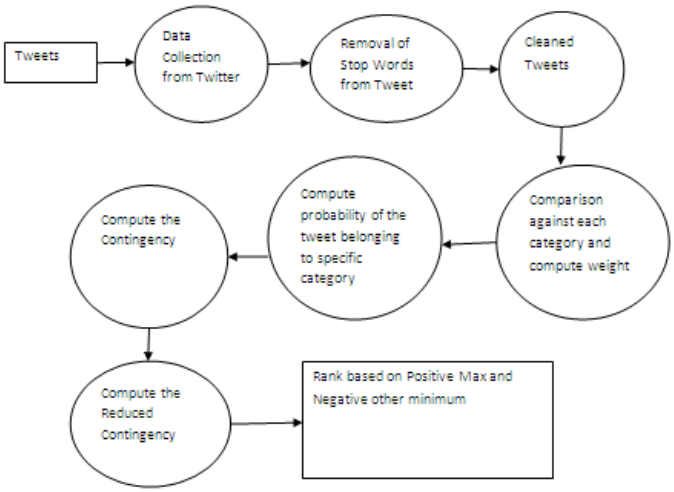

Fig. 4. Data Flow Diagram Level 2

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 7, ISSN No. 2455-2143, Pages 46-51
Published Online May - June 2016 in IJEAST (http://www.ijeast.com)

Fig. 4 shows the Data Flow Diagram Level 2. As shown in the figure:

1. Tweets for various cases acts as input
2. The tweets are collected from the twitter using OAuth API.
3. For each of the tweets the stop words are removed and clean tweets are also obtained.
4. Each tweet is then compared against the words belonging to 5 different categories.
5. Compute the probability using nave byes formula.
6. Compute the contingency.
7. Compute the Reduced contingency.
8. Rank the Tweets based on maximum positive and negative minimum.

V. IMPLEMENTATION

The Detailed Design can be described as follows

A. Data Collection using Twitter -

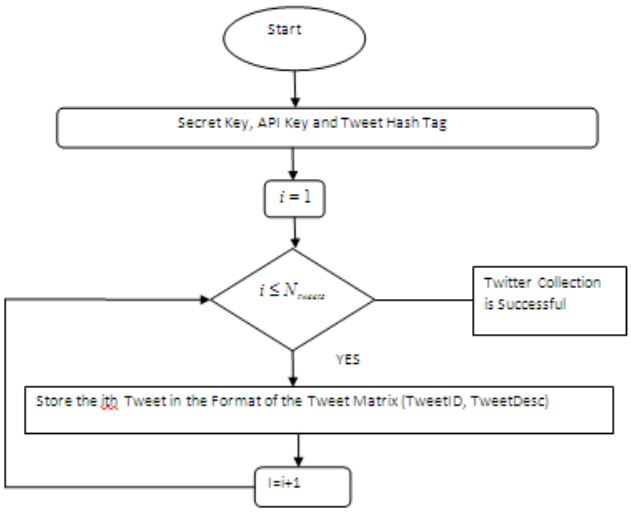

Fig. 5. Data Collection using Twitter

Fig. 5 shows the Data collection using Twitter. Secret Key, API Key and Tweet Hash Tag acts as an input. All the tweets for the specific Hash Tag are collected and then they are stored in the format of the Tweet Matrix

B. Noise Reduction -

This process is responsible for removal of noise the stop words that are present in the given tweet.

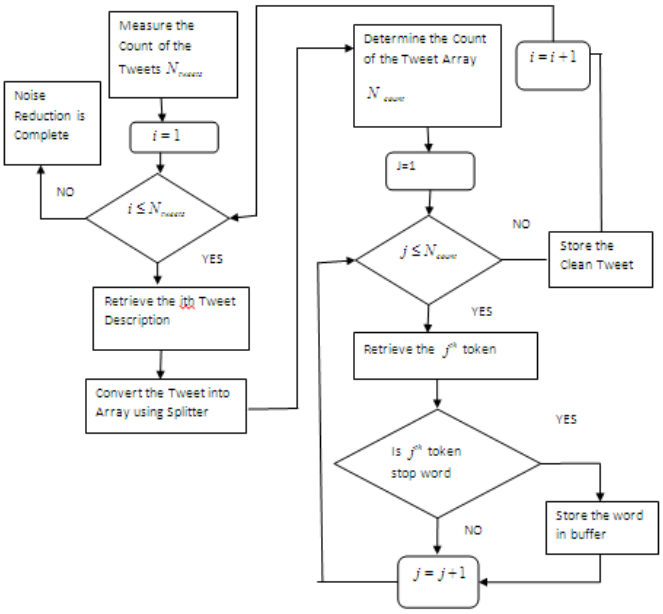

Fig. 6. Noise Reduction Process

Fig. 6 shows the Noise Reduction process

1. Determine the count of the number of tweets.
2. For each of the tweet the cleaning is performed and the stop words are removed and one can obtain the clean tweet.

C. Probability Computation -

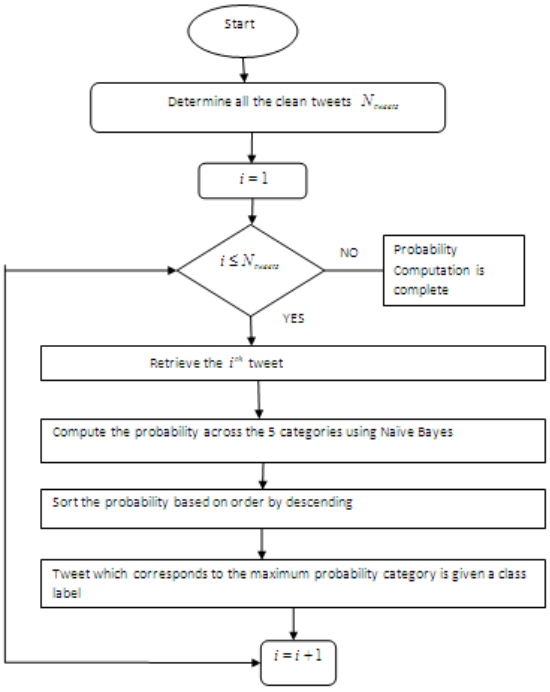

Fig. 7. Probability Computation

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 7, ISSN No. 2455-2143, Pages 46-51
Published Online May - June 2016 in IJEAST (http://www.ijeast.com)

Fig. 7 shows the probability computation

1. Retrieve the List of Tweets
2. For each of the tweets compute the probability for all the categories.
3. The maximum probability is found.
4. Tweet is regarded as belonging to that specific category

VI. RESULT

A. Authentication Process

Application to Twitter is registered using two keys (secret key and consumer key) Using the Stream or Rest API Twitter accesses the user token. Thus every Social Networking Site or a Micro Blogging site has standard authentication process and its own API. Below figures Fig 8 and 9 show the creation of OAuth application for getting the required keys (consumer key and secret key) with which we can connect to Twitter database.

Fig. 8. Creation of Outh application

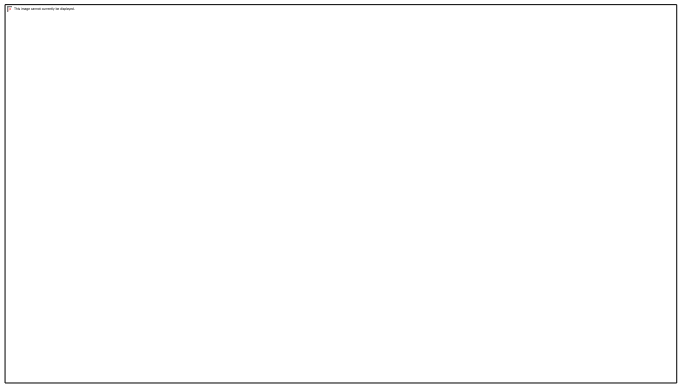

Fig. 9. Consumer and security keys

B. Hashtags

HashTag is nothing but a tag or a topic based on which we are going to extract the Twitter information. Here the main aim of application is to extract the sentiments, issues or problems tweeted by the students which are nothing but the HashTags. The figures Fig. 10, 11 and 12 show adding, validating and viewing the submitted HashTags.

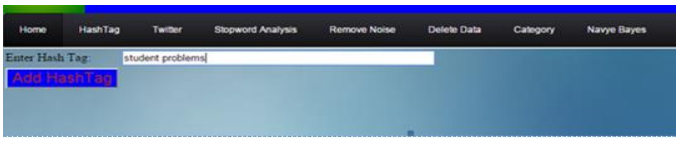

Fig. 10. Addition of Hashtags

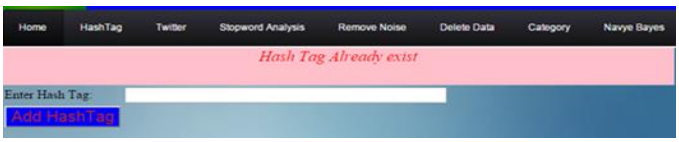

Fig. 11. Checking the Hashtags existence

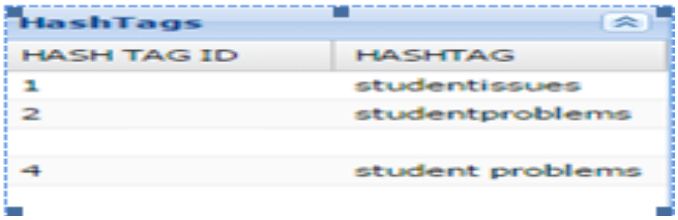

Fig. 12. Sample of Hashtags stored in database

VII. CONCLUSION

These days' social networking sites and micro blogging sites are trends, which may have information in the form of videos, audios, pictures, peoples conversation, updates on latest technologies, movies, latest products etc. Using these sites such as Twitter we can access to, tweets, friends and their credentials. The data on these sites is highly unstructured and not in the format required by the user. Due to which the process of Extracting and Analyzing is the most difficult and challenging task. The development of this application will help in understanding the steps of extracting; analyzing and computing the probability of tweets based on a given subject and thus help develop a new advanced application by using the extracted and analyzed data.

VIII. REFERENCES

[1] Brad Dinerman, "Social Networking and Security Risks", in GFI White Paper, 2011, page(s) 1-8. IEEE © 2011

[2] M. Clark, S. Sheppard, C. Atman, L. Fleming, R. Miller, R. Stevens, R. Streveler, and K. Smith, "Academic pathways study: Processes and realities," in *Proceedings of the American Society for Engineering Education Annual Conference and Exposition*, 2008.

[3] Shamanth Kumar, Fred Morstatter, Huan Liu, "*Twitter Data Analytics*", Springer, Aug 19, 2013. Page(s) 35-48

[4] R. Ferguson, "The state of learning analytics in 2012: A review and future challenges," *Knowledge Media Institute, Technical ReportKMI-2012-01*, 2012. pp. 57–62

[5] R. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

[6] E. Goffman, *The Presentation of Self in Everyday Life*. Lightning Source Inc, 1959. pp. 35–41