International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 04 | Apr-2016 www.irjet.net p-ISSN: 2395-0072

Social Media Data Mining For Sentiment Analysis

K. C. Khatib[1] T. D. Kamble[2], B. R. Chendake[3] G. N. Sonavane[4]

[1]*Kousar C. Khatib C. S. E, A/P-Herle, Tal-Hatkangle, Dist-Kolhapur*
[2]*Tanuja D. Kamble C. S. E, A/P-Kavathe Mahankal Tal- K.M, Dist-Sangli*
[3]*Bhagyashri R. Chendake C. S. E, A/P-Rui, Tal-Hatkangle, Dist-Kolhapur*
[4]*Gitam N. Sonavane C. S. E, A/P-Mhaswad, Tal-Man, Dist-Satara*
Professor. P. S. Kulkarni, Department of Computer Science and Engineering
Sou.Sushila Danchand Ghodawat Charitable Trust's Group of Institution (Atigre), Maharashtra, Kolhapur.

Abstract – *Recently, social media is playing a vital role in social networking and sharing of data. Social media is favored by many users as it is available to millions of people without any limitations to share their opinions, educational learning experience and concerns via their status. Twitter API is processed to search for the tweets based on the geo-location. Student's posts on social network gives us a better concern to take decision about the particular education systems learning process of the system. Evaluating such data in social network is quite a challenging process. In the proposed system, there will be a workflow to mine the data which integrates both qualitative analysis and large scale data mining technique. Based on the different prominent themes tweets will be categorized into different groups. Naïve Bayes classifier will be implemented on mined data for qualitative analysis purpose to get the deeper understanding of the data. It uses multi label classification technique as each label falls into different categories and all the attributes are independent to each other. Label based measures will be taken to analyze the results and comparing them with the existing sentiment analysis technique.*

Key Words: **Education, Sentiment Analysis computers and education, social networking, web text analysis word.**

1. INTRODUCTION

The project on Social Media Data Mining for Sentiment Analysis is a web based application for college to their student's posted tweets for Social Media is the use of electronic and Internet tools for the purpose of sharing and discussing information and experiences with other human beings in more efficient ways. Social website like Twitter, WhatsApp and Facebook etc.

I.

The services providing for Student's are Twitter posts to undersatand issues and problems in their educational experiences. Student's encounter problems such as heavy study load, lack of social engagement and sleep deprivation.The complexity of student's experiences reflected from social media content requires human Interpretation. The services provided to administration are Student's posted their tweets in social media to collects data related to student's learning expriences. To determine what student problems a tweet indicates is a more complicated task than to determine the sentiment of a tweet even for a human judge. Therefore, our study requires a qualitative analysis, and is impossible to do in a full unsupervised way. The proposed system has to perform the qualitative analysis using classification algorithm instead of sentiment analysis. Sentiment analysis considers the opinion of the user about a system or product and categorizes it to neutral, negative or positive mood .In the proposed system, searching the information based on the keywords such as engineer, students, campus, class, professor and lab in the twitter data as per the geo location, keyword and search id.

1.1 LITERATURE SURVEY

Several Researchers carried out research work in Social Network Analysis and sentiment analysis. Sentiment analysis is a text processing technique to derive an opinion or mood intention based on the terms used in a real language sentence. The numbers of researchers have concentrated on generating statistical inference from social network data using sentiment analysis models. Bo Pang and Lilliam Lee [2] provided an insightful discussion on sentiment analysis. They considered the ratio of positive words to total words to estimate the opinion. Today's users can easily obtain information but also they can actively generate content. News reports, BBS, forums, blogs, and etc are the main sources of public opinion information. The text contains both facts and opinion which could be extracted using natural language processing. Opinions are usually subjective expressions that describe people's sentiments or feelings toward entities and events; it is a sub-discipline of computational linguistics that focuses on extracting people's opinion from the web. Twitter is a worldwide popular website, which offers a social networking and micro blogging services, enabling its users to update their status in tweets, follow the people they are interested in, retweet other's posts and even communicate with them directly. Micro blogging websites have evolved to become source of varied kind of information. This is due to nature of micros blogs on which people post real time messages about their opinions on a variety of topics, discuss current issues, complain, and express positive sentiment for products they use in daily life. Experimental results on real

Twitter data show that our method can outperform baseline methods and effectively mine desired information behind public sentiment variations.

1.2 PROBLEM DEFINITION

In existing system the data include surveys, interviews, questionnaires, classroom activities about the student educational experiences and problems they are facing. But these traditional methods are time consuming and very limited in scale. The manual analysis does not make sense of analyzing student learning experiences which are huge in volume with different Internet slang and the timing of the student posting on the web.
The sentiment analysis of the tweets does not cover much relevant experience because even for a human judge to determine what student problems a tweet indicates is a more complicated task than to determine just the sentiment of a tweet. These traditional methods are very time consuming and very limited in scale.

2. MODULE IN SYSTEM

Tweets Extraction and Preprocessing

To extract tweets related to the target, we go through the whole dataset and extract all the tweets which contain the keywords of the target. Compared with regular text documents, tweets are generally less formal and often written in an adhoc manner. Sentiment analysis tools applied on raw tweets often achieve very poor performance in most cases. Therefore, preprocessing techniques on tweets are necessary for obtaining satisfactory results on sentiment analysis:

Slang words translation:

Tweets often contain a lot of slang words (e.g. Th, omg). These words are usually important for sentiment analysis, but may not be included in sentiment lexicons. Since the sentiment analysis tool we are going to use is based on sentiment lexicon, we convert these slang words into their standard forms using the Internet Slang Word Dictionary1 and then add them to the tweets.

Non-English tweets filtering:

Since the sentiment analysis tools to be used only work for English texts, we remove all non-English tweets in advance. A tweet is considered as non-English if more than 20 percent of its words (after slang words translation) do not appear in the GNU Aspell English Dictionary.

URL removal
A lot of users include URLs in their tweets. These URLs complicate the sentiment analysis process. We decide to remove them from.

Sentiment Label Assignment

To assign sentiment labels for each tweet more confidently, we resort to two state-of-the-art sentiment analysis tools. One is the SentiStrength3 tool [8]. This tool is based on the LIWC [10] sentiment lexicon. It works in the following way: first assign a sentiment score to each word in the text according to the sentiment lexicon; then choose the maximum positive score and the maximum negative score among those of all individual words in the text; compute the sum of the maximum positive score and the maximum negative score, denoted as Final Score; finally, use the sign of Final Score to indicate whether a tweet is positive, neutral or negative.

Administration

Admin ControlPanel to visit then related student learning experience to available in twitter API query fetch then this collects the tweets from the database to extract the tweets into the dataset.

NAÏVE BAYES MULTI-LABEL CLASSIFIER

Naive Bayes is a simple probabilistic model based on the Bayes rule with independent feature selection, which worked well on text categorization. Naive Bayes does not restrict the number of classes or attributes to deal with. Asymptotically Naive Bayes is the fastest learning algorithm for the training phase. In this paper, we make use of multinomial Naive Bayes model. Class c* is assigned to tweet d, where

$$C^* = \arg Max_c P_{NB}(C \mid D)$$

$$P_{NB}(C/D) = \frac{(P(c)\sum_{i=1}^{m} P(f \mid c)^{n_i(d)})}{P(d)}$$

In this formula, f represents a feature and ni (d) represents the count of feature f; found in tweet d, m represents the number of total features taken into consideration. Parameters P(c) and P (*) are obtained through maximum likelihood estimates [11].

Bayes Rule Influence of one event's occurrence on the probability of another event is known as conditional probability. From the probability theory, Bayes theorem, allows for the calculation of conditional probability. Usually in data mining Bayes theorem will be used to decide among alternate hypothesis. Bayes' theorem formula for the conditional probability of A given B, is as follows, Where P (A) is prior probability P (A/) is the posterior probability of A given B.

$$\text{Polarity}=\frac{P(\text{positive Words}) / P(\text{Total Words})}{P(\text{Negative\_ Words}) / P(\text{Total\_ Words})}$$

However, this method works only for independent features based on Standard English dictionary and failing to capture query specific sentiments. Table- l provides example of positive, negative and neutral tweets.

Table -1: Example of tweet categorization

Sentiment	Tweet
Positive	I feel like I'm hidden from the world—life of an Engineering student.
Negative	I feel myself dying, #nervous.
Neutral	My problem is that I can never actually fall asleep without lying in bed for several hours.

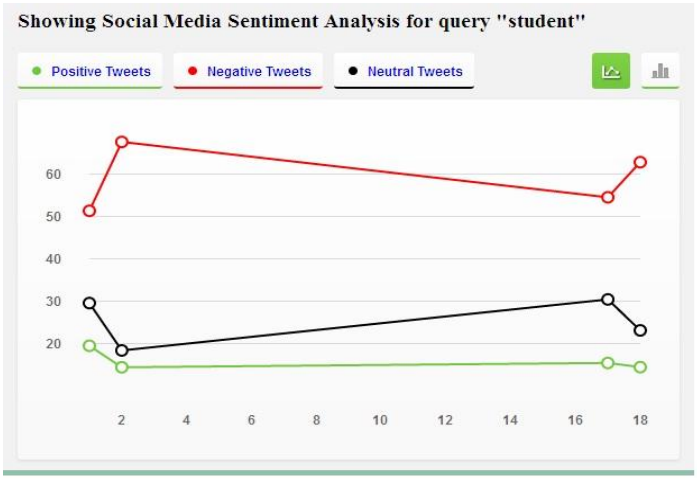

Chart -1: Outcome Result

This window showing the result of sentiment analysis.

SYSTEM ARCHITECTURE

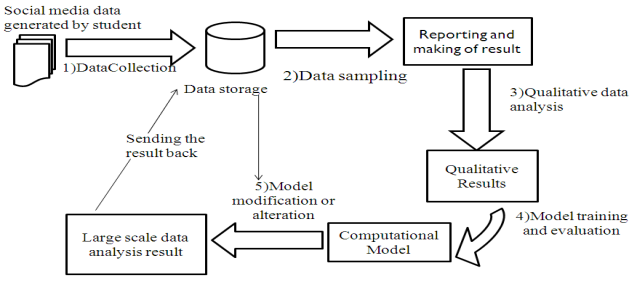

Fig -1: System Architecture

Fig. shows an example of High level system flow. To analyze public sentiment analysis, There are two Latent Dirichlet Allocation (LDA) based models: (1) Foregroundand Background LDA (FB-LDA) and (2) Reason Candidate and Background LDA (RCB-LDA). NaïveBayes, SVM, MaxEnt, ANN classifiers with features extracted from Twitter data using feature extraction methods such asUnigram, Bigram and Hybrid (Unigram + Bigrams) for sentiment analysis. In order to remove stop words and extract features, we perform data cleaning and normalization. We extract the target based extended features model [7] by modifying it and twitter user data from the normalized data. We extract tweets related to our interested targets (e.g. "Student's"), and preprocess the extracted tweets to make them more appropriate for sentiment analysis.

3. CONCLUSIONS

Overall, we conclude that social network based behavioral analysis parameters can increase the prediction accuracy It is beneficial to researchers in learning analytics, educational data mining, and learning technologies. It provides a workflow for analyzing social media data for educational purposes that overcomes the major limitations of both manual qualitative analysis and large scale computational analysis of user-generated textual content. Our study can inform educational administrators, practitioners and other relevant decision makers to gain further understanding of student's college experiences. However, presence of all the entities in unbiased and equal manner is necessary to provide accurate results. To understand the influential parameters that effect the results, semantic features are also very useful from point view of the entity itself. Twitter based social networks provides a great platform in measuring the public opinion with the reasonable accuracy.

4 FUTUREWORK

The following section discusses the work that will be implemented with future releases of the Web Application.

1. First, not all students are active on Twitter, so we may only find the ones who are more active and more likely to expose their thoughts and feelings. Also, students' awareness of identity management online may increase overtime. The "manipulation" of personal image online may need to be taken into considerations in future work.

2. Second, we only identified the prominent themes with relatively large number of tweets in the data. There are a variety of other issues hidden in the "long tail". Several of these issues may be of great interest to education researchers and practitioners. Future work can be done to

design more sophisticated algorithms in order to reveal the hidden information in the "long tail".

3. Often times, manual analysis is time consuming not only because of the time spent on analyzing the actual data, but also the time spent on cleaning, organizing the data, and adapting the format to fit the algorithms. We plan to build a tool based on the workflow proposed here combining social media data and possibly student academic performance data.

4. Possible future work could analyze student's generated content other than texts (e.g. images and videos), on social media sites other than Twitter (e.g. Facebook, Tumbler, and YouTube). Future work can also extend to students in other majors and other institutions.

ACKNOWLEDGEMENT

We take this golden opportunity too we our deep sense of gratitude to my project guide Mrs. P. S. Kulkarni,[1] for her instinct help and valuable guidance with a lot of encouragement throughout this project work, right from selection of topic work up to its completion. My sincere thanks to Head of the Department of Computer Science & Engineering Mr. A. S. Kamble [2], who continuously motivated and guided us for completion of this project. I am also thankful to our Project Coordinator, all teaching and non-teaching staff members, for their valuable suggestions and valuable co-operation for completion of this Project. I specially thank to those who helped us directly-indirectly in completion of this work successfully.

REFERENCES

[1] M. Rost,L.Barkhuus, H. Cramer, and B. Brown, "Representation and Communication: Challenges in Interpreting Large Social Media Datasets," Proc. Conf. Computer Supported Cooperative Workpp. 357-362, 2013.

[2] [4] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," J. Comput. Sci., vol. 2, no. 1, pp. 1–8, Mar. 2011.

[3] [1] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," Educause Review, vol. 46, no. 5, pp. 30–32, 2011.

[4] R. Ferguson, "The state of learning analytics in 2012: A review and future challenges," Knowledge Media Institute, Technical Report KMI-2012-01, 2012

[5] G. Siemens and P. Long, "Penetrating the Fog: Analytics in Learning and Education," Educause Rev., vol. 46, no. 5, pp. 30-32, 2011.

[6] M. Vorvoreanu, Q.M. Clark, and G.A. Boisvenue, "Online Identity Management Literacy for Engineering and Technology Students,"J. Online Eng. Education, vol. 3, article 1, 2012.

[7] M.E. Hambrick, J.M. Simmons, G.P. Greenhalgh, and T.C.Greenwell,"Understanding Professional Athletes' Use of Twitter: A Content Analysis of Athlete Tweets," Int'l J. Sport Comm., vol. 3, no. 4, pp. 454-471, 2010.

[8] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in Proc. 4th Int. AAAI Conf. Weblogs SocialMedia, Washington, DC, USA, 2010.

[9] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena",inProc.5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.

[10] A. Abrahams, F. Hathout, A. Staubli and B. Padmanabhan, "Profit-Optimal Model and Target Size Selection with Variable Marginal Costs," 2013.

[11] R. Ferguson, "The state of learning analytics in 2012: A review and future challenges," Knowledge Media Institute, Technical Report KMI-2012-01, 2012.

[12] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," Educause Review, vol. 46, no. 5, pp. 30–32, 2011.