ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

A Novel Approach for the Classification of Social Media Data using Decision Tree

Divya S Pillai, Jebin R Oliver

Pursuing M.Tech, Dept. of CSE, Caarmel Engineering College, MG University, Kerala, India

Assistant Professor, Dept of CSE, Caarmel Engineering College, MG University, Kerala, India

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 provide free API is processed to search for the tweets based on the students informal conversation. Student's posts on social network gives us a better concern to take decision about the particular education system's learning process of the system. Evaluating the informal conversations in social media site is challenging. The system proposes a workflow to mine the data which integrates both qualitative analysis and large scale data mining technique. The tweets will be categorized into different groups based on certain prominent themes. 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. The classifier will be taken to analyze the results and comparing them with the existing sentiment analysis technique.. In existing system the Naive Bayes classifier is used to classify the twitter dataset thus increases the mean squared error. The Decision Tree classification technique will prove the proposed works efficiency.

KEYWORDS: Social media, naïve bayes classifier, Decision Tree.

I. INTRODUCTION

Social Networking Internet services [1] are changing the way to communicate with others, entertain and actually live. Social Networking is one of the primary reasons that many people have become avid Internet users; people who until the emergence of social networks not provide users interests in the web and a robust indicator of what is really happening online. The rapid growth in popularity of social networks has enabled large numbers of users to share their content, give and receive recommendations, but it opened new challenging problems. The unbounded growth of content and users pushes the Internet technologies to its limits and demands for new solutions.

The proposed system is different from the existing system in terms of analysis technique of data. The proposed system performs the qualitative analysis of data 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 [5] [6]. 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. The system proposes the categorization of tweets based on the twitter data.

The rest of the paper is organized as follows. Section 2 formulates the problem definition. Section 3 shows the system architecture and module description. Section 4 describes the implementation. Section 5 describes the experimental setup on online anomaly detection task. Section 6 concludes the work.

II. PROBLEM DEFINITION

Methods used to analyze 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

DOI: 10.15680/ijircce.2015.0306157

volume with different Internet slang and the timing of the student posting on the web. The sentiment analysis [5][6] 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.

The objective of the system is to design a flow for the analysis of social media data for understanding student learning problems and experience and applying the mining classification algorithm to get the results which will be useful for policy makers to take proper decisions to improve the education system of the institution.

This system provides user friendly functionalities and attractive user interface,
some of the functionalities are listed below

□ Connectivity of social media,
□ Collecting informal conversations from social media,
□ Recognizing keywords
□ multi label classification
□ Data clustering

III. SYSTEM ARCHITECTURE

The system architecture includes different components such as the twitter content is very concise and provides free APIs it will be easy to mine using twitter4j [6] and analyze the data. The proposed system is suggested to implementing each module of the architecture presented below. At first the data will be collected from large volumes of dataset , later Inductive content analysis will be performed on it once the Data sampling and data analysis takes place in the data flow. In model training and evaluation the multi level classifier [7] is implemented upon some prominent theme categories. The different categories helps to understand the learning problems of students also to take proper decisions to overcome them. Once the data is collected first pre-processing of the text should be done by removing the consecutive letters, non lettered symbols, hash tags etc. Once the data is preprocessed Naive Bayes Multi-Label Classifier is implemented to analyze the data and later on evaluating the measures for classifier in terms of performance. The words here are mutually exclusive that is independent of each other’s category. System architecture defines the structure and behavior of a system. Authorized users after the authorization process, collects the data and performing qualitative analysis that is naïve bayes classifier after pre processing and removing the duplicates the data is supposed to evaluate the label based measures. The system specifies a naive bayes classifier for the classification of social media data and the results can be compared with the naïve bayes classifier. The student emotions or their learning experiences obtained from the informal conversations in the twitter. The collected tweets contain certain nonstop words, negative words and it can be removed by the text pre processing before classification. The naïve bayes multi label classifier proves the best classification results than the sentiment analysis techniques. The naïve bayes classifier result in certain mean squared error and can be minimized using the Reduced Error Pruning Tree.

The figure 3.1 specifies the overall system architecture and the working flow of the system and the analysis of the results based on the multilabel classification and its efficiency in classification based on throughput.

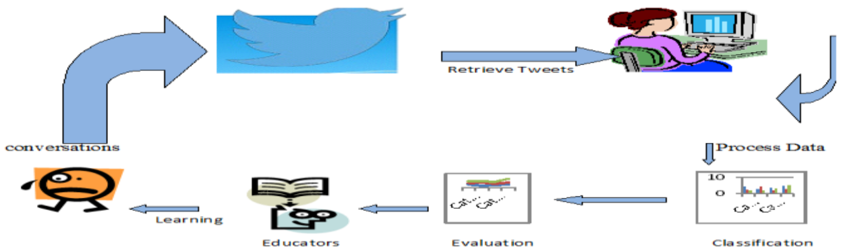


Fig 3.1: System Architecture

3.1.1 System Authentication

Registered user should have to use their username and the password to get login into the system. If the user is unregistered, user can create an account to get login into the system. After passing the system authentication, then only user can see the main application of the system. The figure 3.2 specifies the twitter application.

3.1.2 Creating Twitter Application

Millions of opinions will be tweeted daily on twitter. These tweets can be collected using API (Application program interface) provided by twitter. Using the Twitter APIs [16][6] twitter4j the integration of the project to twitter is done and the data in the form of tweets are collected as shown in Figure 3.2.

Twitter Apps

Mining Twitter Students Problems

Tweets will be fetched and processed to get information about student life

FIGURE 3.2 TWITTER APPLICATION

IV. IMPLEMENTATION

Tweets will be collected from the account by using the secret tokens of the twitter application. Twitter authenticates the secret tokens and allows the user to access the twitter to collects the tweets. From this text mining , the list of keywords that have to be used to collect the tweets will be saved in the database and that keywords will be retrieved from the database to query data from the twitter. Tweets can be retrieved based on the hash tag, keyword hash tag and geo specific location.

Inductive Content Analysis

Because social media content like tweets contain a large amount of informal languages that cannot be used for analysis meaning is often ambiguous and subject to human interpretation. The unsupervised algorithms does not detect the informal conversations it need proper training for required analysis of data.

The data is not specified in any category so we needed to explore what students were revealing in the tweets. Thus, first need to conduct an inductive content analysis on the #engineeringProblems data set. The Inductive content analysis is one popular qualitative research method for manually analyzing text content.

Development of Categories

The purpose of conducting the inductive content analysis was to identify what the major problems, concerns, and issues that engineering students encountered in their study and life. Researcher A read a random sample of 2,000 tweets from the 19,799 unique #engineering Problems tweets, and developed initial categories including: curriculum based problems, heavy study load, study difficulty, imbalanced life, future and carrier problems, lack of diversity, sleep problems, stress, lack of motivation, physical health problems, nerdy culture, identity crisis, and others. These were developed to identify as many issues as possible, without accounting for their relative significances. Based on the twitter API the tweets are collected based on major engineering problems and different categories can be created. The seven prominent themes are: heavy study load, lack of social engagement, negative emotion, sleep problems, diversity issues, positive emotions and others. Each category reflects one issue or problem that engineering students encounter in their learning experiences.

Finally there were seven categories (heavy study load, lack of social engagement, negative emotion, sleep problems, diversity issues, positive emotions and others). Also create a category to understand how many students is happy with their classroom learning. If a tweet does not convey any of the six prominent problems, it is categorized as “others”. A tweet in “others” can be an engineering student problem other than the five prominent ones, or a tweet that does not have clear meaning. Unlike the five prominent themes, “others” is an exclusive category. The positive emotions category specifies those who are happy with the learning process.

Naive Bayes Multi-Label Classifier

Naive Bayes classifier is very effective compared with other state-of-the-art multi-label classifiers. The Bayes Naïve classifier selects the most likely classification Vnb given the attribute values b1,b2,...bn. This results in:

$$Vnb = argmaxvj\ \Pi\ V\ P(vj)YP(bi/vj)$$

Estimate P (bi|vj) using m-estimates:

$$P\ (bi|vj) = nc + mp\ /n + m$$

where: n = the number of training examples for which v = vj nc = number of examples for which v = vj and b= bi p = b priori estimate for P (bi|vj) m = the equivalent sample size.

Text Pre-Processing

Twitter users use some special symbols to convey certain meaning. For example, # is used to indicate a hashtag, @ is used to indicate a user account, and RT is used to indicate a re-tweets. Twitter users sometimes repeat letters in words so that to emphasize the words, for example, “huuungryyy”, “sooo muuchh”, and “Monnndayyy”. Besides, common stopwords such as “a, an, and, of, he, she, it”, non-letter symbols, and punctuation also bring noise to the text. So we pre-processed the texts before training the classifier.

1) We removed all the #engineering Problems hashtags. For other co-occurring hashtags, we only removed the # sign, and kept the hashtag texts.
2) Negative words are useful for detecting negative emotion and issues. So we substituted words ending with “n’t” and other common negative words (e.g., nothing, never, none, cannot) with “negtoken”.

3) We removed all words that contain non-letter symbols and punctuation. This included the removal of @ and http links. We also removed all the RTs.

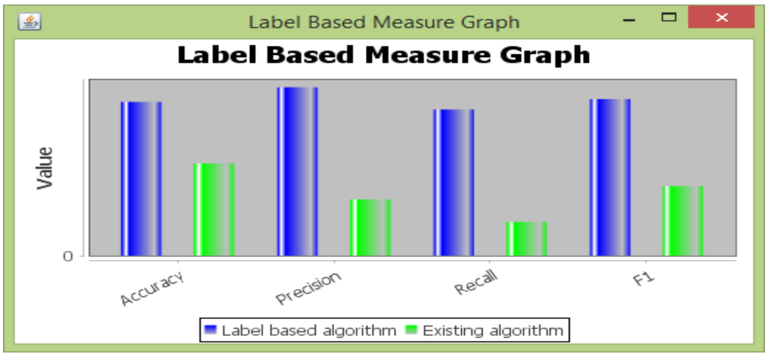
4) For repeating letters in words, our strategy was that when we detected two identical letters repeating, we kept both of them. If we detected more than two identical letters repeating, we replaced them with one letter. Therefore, “huuungryyy” and “sooo” were corrected to “hungry” and “so”. “muuchh” was kept as “muuchh”. Originally correct words such as “too” and “sleep” were kept as they were.

5) We used the Lemur information retrieval toolkit to remove the common stopwords. We kept words like “much, more, all, always, still, only”, because the tweets frequently use these words to express extent. The porter stemmer in the Lemur toolkit was used to perform stemming in order to unify different forms of a word, such as plurals and different forms of a verb.

Stemming is the term used in linguistic morphology and information retrieval to describe the process for reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. The stem needs not to be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.

V. Experimental Results

With eight multi-label categories and one "others" category, there are (25-1)+1=32 possible label sets for a tweet. Table 5.1 provide all the evaluation measures under random guessing. The random guessing program first guessed whether a tweet belongs to "others" based on the proportion this category takes in the training dataset.



5.1 Comparison Result

If this tweet did not belong to "others", it then proceeded to guess whether it fell into the rest of the categories also based the proportion each category takes in the rest categories. We repeated the random guessing program 100 times, and obtained the average measures.

Example based accuracy	Precision	Recall	F1
0.5518	0.6058	0.5830	0.5800

VI. Conclusion

On various social media sites students discuss and share their everyday encounters in an informal and casual manner. Analyzing such data ,however can be challenging and the complexity of atudent's experiences reflect from social media content requires human interpretation .However the growing scale of data demands automatic data analysis techniques this work focus on to demonstrate a workflow of social media data sense making for educational purposes ,integrating both qualitative analysis and large scale data mining techniques and to explore student's problems encounter in learning process and this also gives the protection of students. In existing system the naive bayes classifier is used to clarify the twitter dataset thus increases the mean squared error .The Decision tree classification technique will prove the proposed works efficiency.

Acknowledgement

We would like to thank reviewers for their helpful and constructive comments. We would also like to thank the faculties, the resource providers and the system administrator for the useful feedback and constant support.

References

1. S. Cetintas, L. Si, H. Aagard, K. Bowen, and M. Cordova- Sanchez, "Microblogging in Classroom: Classifying Students' Relevant and Irrelevant Questions in a Microblogging- Supported Classroom," Learning Technologies, IEEE Transactions on, vol. 4, no. 4, pp. 292–300, 2011.
2. S. Chakrabarti. Data mining for hypertext: a tutorial survey. SIGKDD Explor. Newsl., 1(2):1–11, 2000.

3. S. Cetintas, L. Si, H. Aagard, K. Bowen, and M. Cordova- Sanchez, "Microblogging in Classroom: Classifying Students' Relevant and Irrelevant Questions in a Microblogging- Supported Classroom," Learning Technologies, IEEE Transactions on, vol. 4, no. 4, pp. 292–300, 2011.
4. 2. S. Chakrabarti. Data mining for hypertext: a tutorial survey. SIGKDD Explor. Newsl., 1(2):1–11, 2000.
5. 3. J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explor. Newsl., 1(2):12–23, 2000.
6. 4. Hussain, T.; Asghar, S.; Masood, N., "Web usage mining: A survey on preprocessing of web log file," Information and Emerging Technologies (ICIET), 2010 International Conference on , vol., no., pp.1,6, 14-16 June 2010.
7. 5. J. Yang and S. Counts, "Predicting the speed, scale, and range of information diffusion in twitter," Proc. ICWSM, 2010.

BIOGRAPHY

Divya S Pillai received her Bachelor of Engineering in Computer Science and Engineering from Cochin University, kerala, 2012. At present, she is pursuing M.Tech in Computer Science and Engineering at Caarmel Engineering College, Kerala, and Affiliated to MG University. Her research interests include Data Mining, Big Data processing, Cloud Computing.

Jebin R Oliver is an Assistant Professor in the Computer Science and Engineering Department, Believers Church Caarmel Engineering College, Pathanamthitta, Affiliated to MG University. He received Master of Engineering degree in 2012 from Anna University, Chennai. His research interests are Data Mining, Cloud Computing and Big Data Analytics.