International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 06 | June-2016 www.irjet.net p-ISSN: 2395-0072

Twitter data mining using Naive Bayes Multi-label classifier

Miss Ankita Manwatkar [1], Prof. R. B. Mapari [2]

[1]*Miss Ankita Manwatkar, M.tech Student at CSE department, MIT Aurangabad, Maharashtra, India*
[2] *Prof. R. B. Mapari, CSE department, MIT, Aurangabad, Maharashtra, India*

Abstract - — *Social media services are important part of today's life. They are the huge information source for users; as a result we can see the increasing use of social media. People share information through different social media sites, among which facebook and twitter have more prevalence. This paper presents systems to collect tweets mostly of engineering students in order to analyze those tweets to focus on their issues and problems in their learning experience. Thus a multi-label Naive Bayes classifier is used to analyze the tweets.*

Key Words: **Social media twitter data analysis, naive bayes multi-label classifier**

1. INTRODUCTION

Social network services have become a huge source of information for users, there is an overload of Information(blogs, photos, videos, bookmarks) and Interaction (friends, taggers, followers, commenters), Data mining of social media can expand researcher's capability of understanding new phenomena to provide better services and develop innovative opportunities. Studying the characteristic of the message is important for many different purposes such as news detection, message recommendation, friends' recommendation, sentiment analysis and others. [1] However, classification of such big amount of data has been a big task. Hence, we need to use algorithms and techniques to analyze them. Social media sites such as twitter, facebook, youtube provides a great platform for users to share their status, especially students' informal conversations on social sites show their emotions, opinions, issues, joy, struggle and feelings, experience in education and concerns about the learning process and gain support for it. Users on Twitter generate over 400 million Tweets every day. Hence in this paper emphasis is given to twitter data Thus Student's tweets or comments provide large amount of knowledge and a new perspective for educational and institutional researchers, and practitioners to understand student's behavior outside the classroom. This understanding can inform institutional decision-making on interventions for at-risk students, improvement of education quality, and thus enhance student recruitment, retention, and success [1].

2. SYSTEM MODEL

2.1. Research goal

It is very difficult to deal with the ever-growing scale of data by Pure manual analysis, while pure automatic algorithms usually cannot capture in-depth meaning within the data [2], [3], therefore some research work is needed to be done to directly mine and analyze student- posted content from the social web with the only goal of understanding students' learning experiences.

The research goals of this study are 1) to demonstrate a work flow of social media data sense-making for education purposes, integrating both qualitative analysis and large-scale data mining techniques and 2) to explore engineering students' informal conversations on Twitter, in order to understand issues and problems students encounter in their learning experiences[6].The focus here is on engineering students' posts on Twitter to see what problems they face in their educational experiences, and that is mainly because: Engineering schools and departments have long been struggling with student recruitment and retention issues [6]. Twitter is a popular social media site with public and very concise contents. Twitter provides free APIs that can be used to stream data. Therefore, it is easier to analyze students' posts on Twitter.

2.2. Research methodology

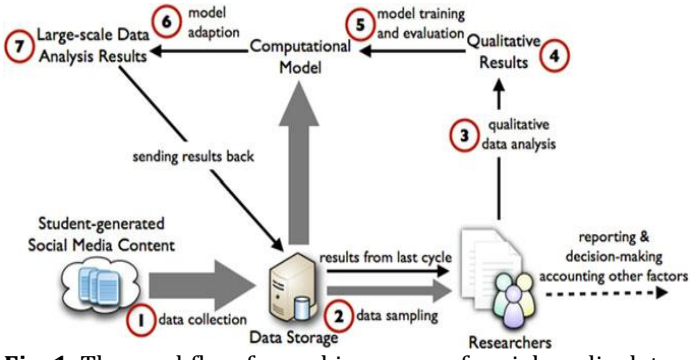

Fig. 1. The workflow for making sense of social media data integrates qualitative analysis and data mining algorithms

To conduct research and content analysis some samples of data set is needed, as in Fig. 1. First there is a need for collecting data related to the research topic, so here researchers collected #engineeringproblems data set and found that major problems engineering students encounter in their learning experiences fall into several prominent categories. Based on these categories, they choose to implement a multi- label Naive Bayes classification algorithm. They used the classification algorithm to train a detector that could assist in the detection of engineering students' problems. The results will help educators identify at-risk students and make decisions on proper interventions to retain them.

3. LITERATURE REVIEW

3.1. Mining educational social media data

Two Prominent reasons to use data mining are, there is too much data and too little information and thus there is need to extract useful information from the data and to interpret it. Thus one can automate the process of finding relationships and patterns in raw data and the results can be either utilized in an automated decision support system or assessed by a human analyst, especially in science and business areas to analyze large amounts of data, and to discover trends. If the valuable knowledge hidden in raw data can be revealed, then this data might be one of most valuable assets. The persistent growth of data in education continues. More institutes now store terabytes and even petabytes of educational data. Data complexity in education is increasing learning developers, universities, and other educational sectors confirm that tremendous amount of data captured is in unstructured or semi-structured format. Educators, students, instructors, tutors, research developers and people who deal with educational data are also challenged by the velocity of different data types, organizations as well as institutes that process streaming data such as click streams from web sites; need to update data in real time to present the right offers to their customers. This analytical study is helps analyze the challenges with big educational data involved with extracting knowledge from large data sets by using different educational data mining approaches and techniques.

Two specific fields that significantly exploit big data in education are educational data mining and learning analytics. In general, educational data mining tries to uncover new patterns in acquired data, building new algorithms or new models, while learning analytics looks for identified predictive models in educational systems. As in the figure 2, educational data such as log files, user interaction data, and social network data types can grow in the near future. This research study is oriented to the challenges and analysis the big educational data involved with uncovering knowledge from large data sets by using different educational techniques. [7]

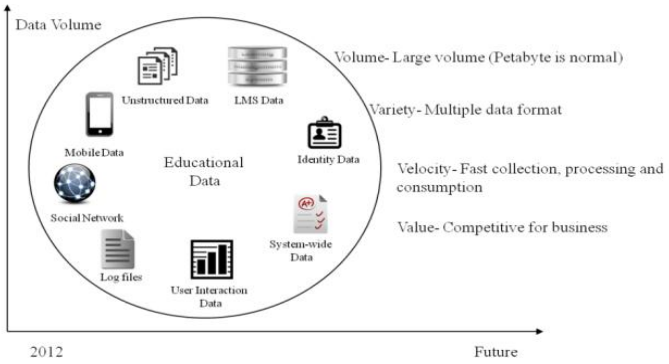

Fig. 2. Growth of different educational data

3.2. Sentiment analysis using machine learning algorithms

Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases. A learner can take advantage of examples to analyze characteristics of their interest or to illustrate relations between observed variables. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data.

3.2.1. Naive Bayes

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. A probabilistic model can allow us to determining probabilities of the outcomes. It can solve diagnostic and predictive problems. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data. [8]

3.2.2. Maximum Entropy

Maximum entropy (ME) models provide general purpose machine learning technique for classification and prediction. A leading advantage of ME models is their flexibility. The parameter estimation for ME models is conceptually straightforward, in practice ME models for typical natural language tasks are usually quite large, and frequently contain hundreds of thousands of free parameters. Estimation of such large models is expensive, highly efficient, accurate, scalable methods are required for estimating the parameters of practical models. [8]

4. SYSTEM DEVELOPMENT

4.1. Data collection and Preprocessing

There's a model which collects tweets from social networking sites and provide a view of business intelligence. Here, there are two layers in the sentiment analysis tool first the Data processing layer deals with data collection and data mining, while sentiment analysis layer use an application to present the result of data mining. The list of tweets has been set up manually. Now go through the website of social network sites to collect tweets. All data collected will be stored in a database for further analysis. During the analysis process, words and their polarities are taken into considerations. Combining with social semantic analysis and natural language processing, unrelated contents will be discarded, and thus relative contents are accurately extracted.

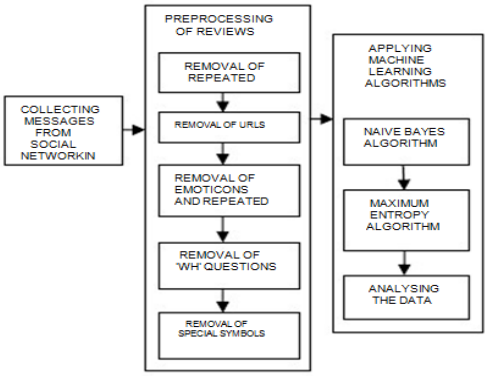

Fig. 3. Architecture for SAT using Machine learning algorithms

Twitter users use special symbols to convey certain meaning. As, # indicates a hashtag, @ indicate a user account, and RT to indicate re-tweet. Twitter users sometimes repeat letters in words so that to emphasize the words, example, "cuttee", "and soo muuchh", "Monnndayyy". Also, common stopwords such as "a, an, and, of, it", non-letter symbols, and punctuation produce noise in the text. So there is a need to preprocess this data.

4.2. Naive Bayes Multi-label classifier

One popular way to implement multi-label classifier is to transform the multi-label classification problem into multiple single-label classification problems [10]. The basic concept is to assume independence among categories, and train a binary classifier for each category. The following are the basic procedures of the multi-label Naive Bayes classifier. Suppose there are a total number of N words in the training document collection (in our case, each tweet is a document) W =$\{w_1, w_2, \ldots, w_n\}$, and a total number of L categories C = $\{c_1, c_2, \ldots, c_n\}$. If a word w_n, appears in a category c for $m_{w_n c}$ times, and appear in categories other than c for $m_{w_n c'}$ times, then based on the maximum likelihood estimation, the probability of this word in a specific category c is

$$p(\omega_n|c) = \frac{m_{w_n c}}{\sum_{n=1}^{N} m_{w_n c'}} \quad (1)$$

Similarly, the probability of this word in categories other than c is

$$p(\omega_n|c') = \frac{m_{w_n c'}}{\sum_{n=1}^{N} m_{w_n c'}} \quad (2)$$

Suppose there are a total number of M documents in the training set and C of them are in category c. then the probability of category c is

$$p(c) = \frac{C}{M} \quad (3)$$

And the probability of other categories c' is

$$p(c') = \frac{M-C}{M} \quad (4)$$

5. CONCLUSION

Various classification techniques and algorithms can be used. Machine learning can help to obtain high accuracy. The study can inform educational administrators, practitioners and other relevant decision makers to gain further understanding of engineering student's college experiences. It is providing a workflow for analyzing social media data for educational purposes that will overcome the major limitations of both manual qualitative analysis and large scale computational analysis of user-generated textual content further understanding of engineering students' college experiences.

REFERENCES

[1] Liangjie Hong and Brian D. Davison, "Empirical Study of Topic Modeling in Twitter", Workshop on Social Media Analytics (SOMA '10), July 25, 2010.
[2] Pooja R. Takle, *"Identification of students Behaviour in Higher Education from Social Media by using Opinion based Memetic Classifier"*, IJRITCC, March 2015.

[3] R. Ferguson, "The State of Learning Analytics in 2012: A Review and Future Challenges," Technical Report KMI-2012-01, Knowledge Media Inst. 2012.
[4] R. Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," J. Educational Data Mining, vol. 1, no. 1, pp. 3-17, 2009.
[5] 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,"
[6] Xin chen, Mihaela Vorvoreanu, Krishna Madhvan,"Mining Social Media Data for Understanding Student's Learning Experiences", IEEE Transactions on Learning Technologies, vol. 7, no. 3, Jul-Sept 2014
[7] Saeed Aghabozorgi, Hamidreza Mahroeian, "An Approachable Analytical Study on Big Educational Data Mining", Department of Information System, University of Malaya 50603 Pantai Valley, Kuala Lumpur, Malaysia.
[8] Hemalatha, Dr. G. P Saradhi Varma, Dr. A.Govardhan, "Sentiment Analysis Tool using Machine Learning Algorithms",IJETTCS, Vol 2, Issue 2, Mar – Apr 2013 .
[9] Rajeshri Thorat, Ekta Pawar, "Mining social media data using Naïve Bayes algorithm ", Asian Journal of Engineering and Technology Innovation 03 (06); 2015;
[10] G. Tsoumakas, I. Katakis, and I. Vlahavas,"Mining Multi-Label Data", Data Mining and Knowledge Discovery Handbook, pp. 667-685, Springer, 2010