Asian Journal of Engineering and Technology Innovation 03 (06) 2015 (15-17)

Asian Journal of Engineering and Technology Innovation

ISSN : 2347-7385

REVIEW ARTICLE

Received on: 03-03-2014
Accepted on: 15-03-2015
Published on: 20-03-2015

Rajeshri Thorat, Ekta Pawar, Pranali Shendekar, Kajal Lokhande, Apeksha Mengade
Indira College of Engineering, Pune
rajeshri.thorat@indiraicem.ac.in
ekta.pawar00@gmail.com
pranalishendekar1@gmail.com
lokhande.kajal006@gmail.com
mengade.apeksha@gmail.com

QR Code for Mobile users

Conflict of Interest: None Declared

Mining social media data using Naïve Bayes algorithm

Rajeshri Thorat, Ekta Pawar, Pranali Shendekar, Kajal Lokhande, Apeksha Mengade
Indira College of Engineering Pune.

ABSTRACT
Social network services have become a big source of information for users. Studying the characteristics of message is important for a number of tasks such as, breaking news detection, user's recommendation, etc. However, classification of such big amount of data has been a big task. Hence, we need to use algorithms and techniques to categorize them. In this paper, we are developing a workflow to collect both qualitative analysis and large-scale data mining analysis. We are focusing on engineering student's problem, which will be helpful for understanding their issues and educational resources. For doing analysis we are using the multiple label classification algorithms.

Keywords: Social media data classification, education, web test analysis

Cite this article as:
Rajeshri Thorat, Ekta Pawar, Pranali Shendekar, Kajal Lokhande, Apeksha Mengade, Mining social media data using Naïve Bayes algorithm. Asian Journal of Engineering and Technology Innovation 03 (06); 2015; 15-17.

INTRODUCTION

Background:

Social network service is very important medium for communication for the online users, where users can share their joys and problems. Traditionally, people were using methods as, survey, feedback, interviews, group discussions etc. which were time consuming and not generating specified results. Such methods were limited for large amount of data. We are implementing automated design technique which will automatically categorize the data depending upon their characteristics.

Related work:

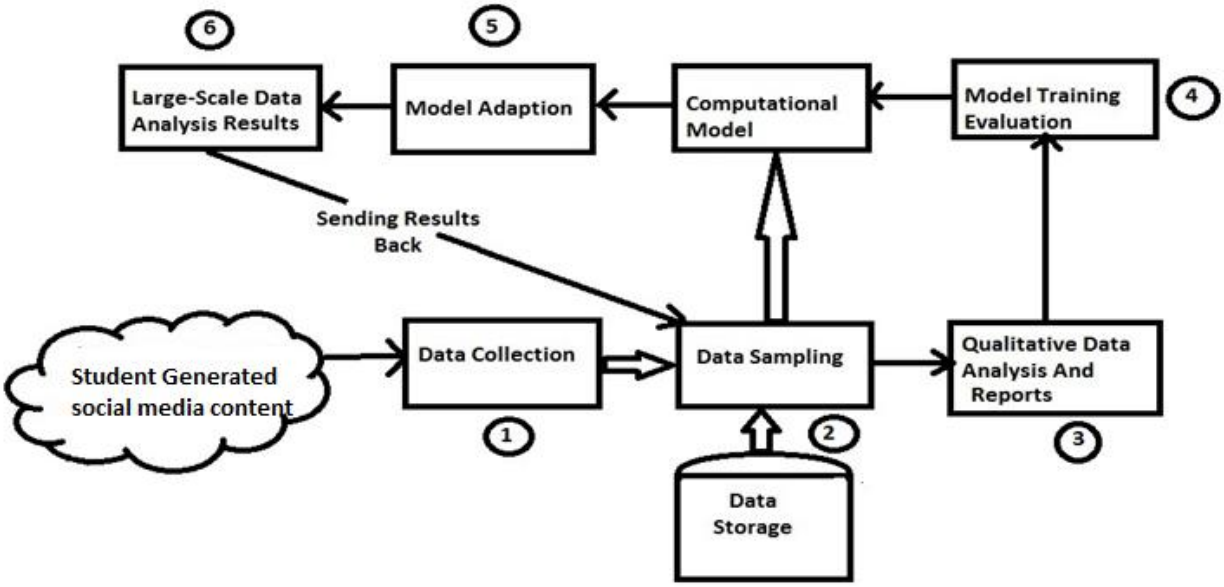
The theoretically value of the data can be evaluated by Goffman's theory of social performance [1]. It has been developed to explain face-to face interactions, Goffman's theory is widely used everywhere to explain mediated interactions on the web today [2]. Many studies conclude that social media users manage their online identity to "look better" than in real life [3], [4].

In case of different studies related to identities, there is a lack of awareness about managing online identity among college students [5], and that young people usually regard social media as their personal space to hang out with outside the sight of parents and teachers [6]. Researchers from different fields have analyzed Twitter content to generate specific knowledge for their respective subject domains. For example, Gaffney [7] seems tweets with hashtag #iranElection using histograms, user networks, and frequencies of top keywords to quantify online activism. Similar studies have been conducted in other fields including healthcare [8], marketing [9], athletics [10], just to name a few.

System architecture:

The student will post their comments or the views on the social media site. Then the data is collected in database, extraction is done on the data. Result of last cycle is collected and noisy data is removed. Refining the data from the store gives the model training evaluation. Finally by the model adaption, large scale data analysis result is generated.

In the step 1, we are collecting posts by engineering students on the social networking site. The inductive content analysis then performed on the database in step 2 and 3.In step 4, we will categorize their several problems in prominent categorizes. Based on these categories, we are implementing a multi-label Naïve Bayes classification algorithm. We will evaluate the performance of the classifier by comparing it with other state-of-the-art multi-label classifiers in step 5. We are using the classification algorithm to train a detector that could assist detection of engineering student's problems in step 6. The results are seen in the step 7 that could help educators to identify at-risk students and make decisions on proper interventions to retain them.



Algorithm:

Naïve Bayes Multi-label Classifier

It is a popular method of classification, 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=$\{c_1, c_{2,\ldots,} c_L\}$. 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

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

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

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

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

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

For a document d_i in the testing set, there are K words $W_{d_i} = \{w_{i1,} w_{i2,} w \dots, w_{ik}\}$ and W_{d_i} is a subset of W. The purpose is to classify this document into category c or not c. We assume independence among each word in this document, and any word w_{ik} conditioned on c or c' follows multinomial distribution. Therefore, according to Bayes' Theorem, the probability that d_i belongs to category c is

$$p(c/d_i)=\frac{p\left(\frac{d_i}{c}\right).p(c')}{p(d_i)} \prod_{k=}^{k} p\left(\frac{w_{ik}}{c}\right).p(c) \quad (5)$$

And the probability that d_i belongs to categories other than c is

$$p(c'/d_i)=\frac{p\left(\frac{d_i}{c'}\right).p(c')}{p(d_i)} \alpha \prod_{k=}^{k} p\left(\frac{w_{ik}}{c'}\right).p(c') \quad (6)$$

Because $p(c/d_i) + p(c'/d_i)$=1, we normalize the latter two items which are proportional to $p(c/d_i)$ and $p(c'/d_i)$ to get the real values of $p(c/d_i)$ is larger than the probability threshold T, then d_i belongs to category c, otherwise, d_i does belong to category c. Then repeat this procedure for each category. In our implementation, if for a certain document, there is no category with a positive probability larger than T, we assign the one category with the largest probability to this document. In addition, "others" is an exclusive category. A tweet is only assigned to "others" when "others" is the only category with probability larger than T.

CONCLUSION

Our 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. Our study can inform educational administrators, practitioners and other relevant decision makers to gain further understanding of engineering students' college experiences.

REFERENCES

1. E. Goffman, *The Presentation of Self in Everyday Life*. Lightning Source Inc, 1959.
2. E. Pearson, "All the World Wide Web's a Stage: The performance of identity in online social networks," *First Monday*, vol. 14, no. 3, pp. 1–7, 2009.
3. J. M. DiMicco and D. R. Millen, "Identity management: multiple presentations of self in facebook," in *Proceedings of the 2007 international ACM conference on Supporting group work*, 2007, pp. 383–386.
4. M. Vorvoreanu and Q. Clark, "Managing identity across socialnetworks," in *Poster session at the 2010 ACM Conference on ComputerSupported Cooperative Work*, 2010. M. Vorvoreanu, Q. M. Clark, and G. A. Boisvenue, "Online Identity Management Literacy for Engineering and Technology Students," *Journal of Online Engineering Education*, vol. 3, no. 1, 2012.
5. M. Ito, H. Horst, M. Bittanti, danah boyd, B. Herr-Stephenson, P. G. Lange, S. Baumer, R. Cody, D. Mahendran, K. Martinez, D. Perkel, C. Sims, and L. Tripp, "Living and Learning with New Media: Summary of Findings from the Digital Youth Project," The John D. and Catherine T. MacAuthur Foundation, Nov. 2008.
6. D. Gaffney, "#iranElection: Quantifying Online Activism," in *WebSci10: Extending the Frontier of Society On-Line*, Raleigh, NC, 2010.
7. S. Jamison-Powell, C. Linehan, L. Daley, A. Garbett, and S. Lawson, "'I can't get no sleep': Discussing #insomnia on Twitter," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, 2012, pp. 1501–1510.
8. M. J. Culnan, P. J. McHugh, and J. I. Zubillaga, "How large US companies can use Twitter and other social media to gain business value," *MIS Quarterly Executive*, vol. 9, no. 4, pp. 243–259, 2010.
9. 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," *International Journal of Sport Communication*, vol. 3, no. 4, pp. 454–471, 2010.