Analyzing Social Media Data to Explore Students' Academic Experiences

Priya Lande
Vidyalankar Institute of Technology
Mumbai

Vipul Dalal
Vidyalankar Institute of Technology
Mumbai

ABSTRACT

The casual conversational style used by the students on any front stage environment can educate extensively about their learning process. The collection of data from such an open environment can bring out many important and unknown factors about students' behaviour, their opinions, their feelings their concerns pertaining to their educational system. The inspection of such data can be said to be very provocative. The reflection of students' feelings over the social network, however, has to undergo the human eye to get properly interpreted, which is possible but upto a certain extent, as a result of ever-growing data. In this paper, problems of engineering students have been considered. This has been worked upon by analysing engineering students' tweets from the hashtag #enggproblems on Twitter. Analysis was carried out over 15,000 tweets. These problems were related to heavy study load, negative emotions, sleep problems, lack of social engagement, diversity issues etc. A multi-label classifier was executed to classify and categorize tweets. This technique can dig up into the casual conversations of students and educate about the factors that affect the learning process of students.

General Terms

Multi-label classification using naïve bayes classifier.

Keywords

Naïve bayes multi-label classifier, twitter analysis, education.

1. INTRODUCTION

Social media data has apparently been playing as an integral part of the urban crowd. The internet is exploited due to its ease of access. A simple click can actually copy the views/feelings in one's mind on the internet. People aged in the group of 15-45 are the most active users of the internet. These consists of mainly students, businessmen etc. Students have a lot many reasons to access the internet be it project work, form filling, seeking any study related information, besides all this they also need a resort to entertainment. Eventually, for today's youth entertainment is click, post and share etc. if they like something they will post it and even if they do not like something they will post about it too. Thus, a complete democratic platform for students and everybody else is online social networking sites. The most popularly used social networking websites Facebook, Twitter, Instagram.

Every second on an average, around 6000 tweets are tweeted on twitter, which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and 200 billion tweets per year. Here, the amount of data that is generated has no scale and no vocabulary boundary too. Students post their views spontaneously online, which too has vocabulary overhead and scalability issues. The inspection of such data can give immense scope to understand students' feelings, their concerns and their opinions too. A complete manual analysis may result into incompatibility with the ever-growing data [1]. On the other hand, a complete automatic algorithm cannot provide in depth meaning of data.

2. LITERATURE SURVEY

Earlier offline procedures were carried out to study such problems [2][3]. These problems included surveys, focus groups, interviews and other such classroom programs. Such programs are generally carried out in front stage environment. A front stage environment is a controlled environment, where a person is likely to express superficially and not transparently [4][5], whereas a backstage environment is a relaxed environment, where one has no pressure to answer a question in a particular way. Such a platform can be online social network like Facebook, Twitter which is very frequently used by the students and it is their spontaneous hub too.

Twitter is one of the many popular social networking websites. There is a provision of API which is free of cost, which can be used to stream data. Therefore, the analysis of tweets can be done on twitter. Twitter allows 140 characters per tweet so its conciseness also helps in easy streaming of data.

A hashtag is a word that begins with a '#' which means all the content related to the hashtag name will be tagged or added in that particular hashtag. Analysis was carried out on engineering students because engineers are said to be the future of any nation. Their learning process has to be strong and has to be upgraded for better adaptability to technology [6]. The hashtag #enggproblems was taken into consideration and was examined. Here the students posted more about their problems faced in their learning system. The tweets were worked upon as a large process where the tweets were said to fall under various category such as heavy study load, diversity issues, lack of social engagement, negative emotion and sleep problems. These categories were built by human examination of tweets falling under #enggproblems. The human inspection of such a data is framed as inductive content analysis. [7]

The main goal of this study is to:

1. Categorize and correctly classify students' tweets into the proper category. This helps to understand the problems faced by the students in their learning process.
2. The statistical study of the classification can help the educational system to make necessary improvements into their system so that students' learning experience is a hassle free one.

In [8], automated identification and classification of diverse type of sentiments is carried out on short fragments of text extracted from twitter. The paper proposes a supervised classification framework which exploits twitter smileys and hashtags for providing learning to labels. The twitter processed data allowed for sentiment type identification. Here, the twitter data is classified as smileys where mixed categorization can take place.

In [9], emoticons are exploited as noisy labels to provide learning to the data. Various multi-label classifiers and feature extractors are used to begin this approach. Multi-label classifiers like SVM, Naïve-Bayes and maximum entropy classifier are used and the unigrams and bigrams are used as the feature extractors where both the classifier and the feature extractors are treated as two different components. This eases working with different classifier combinations.

In [10], they have put forward the method to classify the tweets. The tweet classification is done by exploiting information of author and tweet features. Users can also customize their tweet views according to their interest. Short texts do not provide statistical occurrences of words, Bag of words approach has performed inadequately in such situations. Hence, the proposed method says that a set of domain specific features are taken from the profile and text of the author. This fairly categorizes the text into a set of generic classes like private messages, opinions, news, deals and events.

There are a number of classification algorithms out of which Naïve Bayes, Support Vector Machine, are said to be popular. The number of classes defines the type of algorithm; there can be binary classification or multi-class classification for more than two classes. Both of them are for single-label classification, where single label classification means each piece of data will be classified into one class only, whereas multi-label classification can classify a piece of data into more than one class at one point of time.

The data collected from twitter hashtag #enggproblems counted upto 19,799 unique tweets after streaming them for 14 months. All the re-tweets, duplicate tweets were removed. There were no categories presumed, there was a need of manual analysis as purely automated algorithms cannot give quality results. Thus categories were developed such as – heavy study load, diversity issues, sleep problems, lack of social engagement, negative emotion after the manual analysis. Each of these categories is for specific problems faced by the students. It is observed that one tweet is classified into more than one category at the same time. Moreover, there is also one more category i.e. –'others' , if no tweet falls into any other categories then it is said to get classified into 'others' category.

3. NAÏVE BAYES MULTI-LABEL CLASSIFICATION

A multi-label classifier is built according to the developed categories. Naïve Bayes multi-label classifier is said to be more accurate and precise as compared to other classifiers. Text cleaning can be referred to as pre-processing. It is carried out to avoid unnecessary data, before classifier training all the #enggproblems hashtags are removed and hashtag text was kept as it is. Neg-tokens (negative words) can be used for identifying negative emotions. Repetitive same letters are kept as they are, if they are repeated twice, but if the frequency is more than that, then only one could be kept. E.g. engineering can be kept as engineering but 'yesssssssssssssss' can be kept as 'yes'. All repeated tweets and http links are taken out. All common stop words are eliminated using Lemur information retrieval toolkit. Krovertz stemmer is used from the lemur toolkit to unite diverse word forms

3.1 Basic procedure of classification

The notion is to consider all categories as independent. A binary classifier has to be provided with learning specifically for all categories. Any binary classifier can be converted to a multi-label classifier.

In we have a document d and a class c, our goal is to compute the probability of each class of its conditional probability, given a document, we use this probability to pick the best class.

By Bayes rule,

$$P(c \mid d) = \frac{P(d / c)P(c)}{P(d)}$$

P(c|d) is equal to the probability of the document given a class multiplied by the probability of the given class multiplied by the probability of the document, this is used in the classifier.

$$C_{MAP} = \arg\max_{\mathbf{c}\in C} P(c \mid d) \tag{1}$$

$$C_{MAP} = \arg\max_{c\in C} \frac{P(d \mid c)P(c)}{P(d)} \tag{2}$$

$$C_{MAP} = \arg\max_{c\in C} P(d \mid c)P(c) \tag{3}$$

In equation (1), the class that we are looking for C_{MAP} i.e. is the maximum aposteriori class is out of all classes, the one that maximizes the probability of that class given a document. So a class has to be found whose probability given a document is greatest. By Bayes Rule whichever class that maximizes P(c) given d also maximizes the equation (2) and as it is traditional in Bayesian classification, whichever class maximizes equation (2) also maximizes equation (3), only the denominator is dropped, because P(d) is how likely the document is. Now, for example if a document is given and it has to be classified among 10 classes, if each of these classes, P(d) is computed given a class then P(c) is computed and then P(d), the P(d) is identical for all 10 classes, for each class, one more time P(d) has to be computed and that means, if 10 things are compared each of which is divided by P(d). The P(d) is a constant and it can be thus eliminated. The most likely class i.e. the C_{MAP} is that class which maximizes the product of two probabilities – the P(d|c) (i.e. likelihood) and the P(c) (i.e. prior).

$P(d \mid c)$ can be represented by a set of features, it means,

$$P(d \mid c) =_{c\in C} P(x1, x2, x3, x4 \ldots\ldots\ldots\ldots\ldots x_n)P(c) \tag{4}$$

P(d|c) is represented by joint probability of x_1,x_2,x_3,x_4 upto x_n , given a class. It is represented by a set of feature vectors. Now computing P(c) means how often this class occurs. This can be done by counting the relative frequencies in a corpus or dataset. The Naïve Bayes classifier can be simplified by making two assumptions. The first simplifying assumption that can be made is the Bag of Words assumption. Here it is assumed that the position of the word in the document does not matter; only the words which occurs or which feature occurs, has to be taken care of. The second assumption that can be made is that the different features x_1,x_2,x_3,x_4 upto x_n , their probabilities are independent given a class. So that whether one feature occurs given a class or another, independently they have to be true.

Conditional independence: Assuming that the feature probabilities P(x_i|c_j)are independent given the class c. It is represented as the joint probability of the whole set of features conditioned on the class, as the product of the whole bunch of independent probabilities.

$$P(x1, x2, ...x_n \mid c) = P(x1 \mid c)P(x2 \mid c)P(x3 \mid c).....P(x_n \mid c) \quad (5)$$

Here, the position of x_1 is not considered. Also the dependencies within x_1 and x_2 are also ignored. In other words, in order to compute a simplifying naïve bayes assumption, the most likely class by multiplying the likelihood, the probability of the whole joint string of features multiplied by the prior probability of the class, simplifying this, it can be said that the best class by the naive bayes assumption is that the class that the maximizes these probabilities of the class, where we multiply for each feature the probability of that feature, given the class.

$$C_{NB} = \overset{\arg\max}{c \in C} \; P(c_j) \prod P(x \mid c) \quad (6)$$

So, now looking specifically at text, first all word positions in the text documents have to be seen. For example, in a text document of 100 words, for word number 1 consider position 1, for number 2 consider position 2. Look at all classes and for each class, the probability of class is watched and for each class every position will be walked through the text and for each position, the word in the position will be looked at, and it has to be seen what its probability is given

$$C_{NB} = \overset{\arg\max}{c_j \in C} \; P(c_j) \prod P(x_j c_j)$$
$$i \in positions \quad (7)$$

So, this will be done for class 1 and P(c1) will be computed, multiplied by the product of all the i's of P(W_i|c_1) and same thing will be done for class 2.

$$C1 = P(c1) \prod P(W1 \mid c1)$$
$$C2 = P(c2) \prod P(W1 \mid c2) \quad (8)$$

And the highest of these two will be selected if C_2 is highest then C_2 is assigned to the document. This, in general is therefore true for any no. of classes.

4. CLASSIFICATION RESULTS

There are a total of 5 categories. Heavy study load, negative emotion, diversity issues, lack of social engagement, sleep problems and one more i.e. others. It is logically set, that the tweet will be classified into that category for which the tweet holds the highest probability value, else it is classified into the others category. The multi-label classification facilitates one tweet to get classified into more than one category. For example, a negative emotion can be caused by sleep problems or lack of social engagement.

For pages other than the first page, start at the top of the page, and continue in double-column format. The two columns on the last page should be as close to equal length as possible.It has been observed that most of the tweets get classified into the 'others' category, which does not give any information regarding the students' problem. In the fig. 1 beginning from the leftmost bar categories are: 1. Others, 2. Sleep problems, 3. Negative Emotion 4.Diversity Issues 5.Lack of Social engagement. 6. Heavy study load.

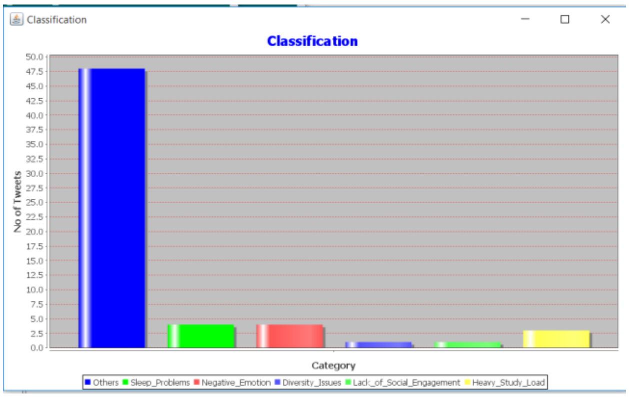

Fig 1: Classification result

5. CONCLUSION

Naïve bayes Multi-label classifier is used to classify the tweets from #enggproblems into the categories-Heavy study load, negative emotion, diversity issues, lack of social engagement, sleep problems and one more i.e. others. This helps to understand the major problems faced by engineering students. It is observed that most of the tweets get classified into the 'others' category which implies that classifier is not giving desired results, which can be expected from any automatic algorithm Considering that this is an automatic algorithm it is not expected that it gives 100% effective and perfect classification, a need of atleast a manual intervention is felt, but this again would consume time, and give late results, it is thus a tedious job. One notion that constantly stands by this problem's solution is that if we understand the sentiment behind each tweet, we can say, that particular tweet is positive, negative or neutral and ultimately we can identify the students who are tweeting those specific tweets and can conclude that more of which type of tweets have been twitted. One way to achieve this is to utilize the Sentiwordnet dictionary, to classify the 'others' category into positive, negative or neutral. This helps us to understand the 'others' category further. It will contribute towards mining of 'others' category. The obtained results are not sufficient for proper understanding of the students' problems. Categorization is one level but accurate categorization is another higher level.

6. REFERENCES

[1] M. Rost, L. Barkhuus, H. Cramer and B. Brown, "Representation and communication: Challenges in interpreting large social media datasets," in *Proceedings of the 2013 conference on Computer supported co-operative work*, 2013, pp. 357-362

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

[3] C.J Atman, S.D Sheppard, J. Turns, R.S Adams, L. Fleming, R. Stevens, R. A. Streveler. K. Smith, R. Miller, L. Leifer, K. Yasuhara and D. Lund, "Enabling Engineering Students' Success: the final report for the Center for the Advancement of Engineering Education," Morgan and Claypool Publishers, Center for the Advancement of Engineering Education, 2008.

[4] E. Goffman, *The Presentation of Self in Everyday Life*. Lightning Source Inc, 1959.

[5] 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.

[6] R. Fergusson, the state of learning analytics in 2012: “A review and future challenges,” *Knowledge Media institutes, Technical report, KMI-2012-01*, 2012.

[7] Xin chen, Krishna Madhvan, Mihaela Vorvoreanu: Mining Social Media Data for Understanding Students’ Learning Experiences in *IEEE Transactions on Learning Technologies,* 2014.

[8] D. Davidov, O. Tsur and A. Rappoport, “Enhanced Sentiment learning using Twitter Hashtags and smileys,” in *Proceedings of 23rd International Conference on Computational Linguistics: Posters*, 2010, pp. 241-249.

[9] A. Go, R. Bhayani and L. Huang, “Twitter sentiment classification using distant supervision,” *CS224N Project Report, Stanford,* pp.1-12, 2009.

[10] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu and M. Demirbas, “Short text classification in twitter to improve information filtering,” in *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval,* New York, NY, USA, 2010, pp. 241-842.