IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 3, Ver. VI (May - June 2017), PP 27-32
www.iosrjournals.org

Improved Students' Social Media Content Analysis Using Machine Learning Algorithm

Bushra SarwatAra Syed, Harshali Patil[2], Mohammad Atique[3]
[1](Computer Engineering, Thakur College of Engineering, India)
[2](Computer Engineering, Thakur College of Engineering, India)
[3](Computer Science and Engineering, Amravati University, India)

Abstract: *Sentimental Analysis has become most profound research areas for prediction and classification. Student's discussion on social media contains sentiwords that are a word or set of words expressing some thought or judgment or idea about something which provides us with some idea about their experiences in learning and views about the particular field [5]. Data from social media site would be raw and difficult to understand but by analyzing it through supervised learning approach we can find out the exact views of students. In this Paper MultiLabel Text Classification is done with Label Correlational Model will give us desired result which wasn't possible with conventional Single Label Classification. The proposed work is to extract the features of text in the form of labels and Correlational Model can find the relation between the labels and understanding among Labels.*

Keywords: *Correlational model, Machine Learning, Multi Label Classification, Multinomial Naïve Bayes, sentimental Analysis*

I. Introduction

Most work on classification has been done on learning from a set of instances that are associated with single or MultiLabel classification but not on label association and combination which would reveal the correlation among labels. Textual data from Social Media like Twitter, Facebook and LinkedIn will be analyzed and different categories can be formed from labels inside them. Generated labels will be further explored using the label correlation technique which will reveal the dependency of the particular label with other label and correlation among them.

Paper further discusses as in Section I. Students Twitter Content, Section II.Related Work, and Section III. Students Twitter Content, Section IV. Tweets cleaning and pre-processing, Section V. Student data filtering, formation of categories and labels classification, Section VI. Proposed method for correlation among labels, Section VII. Conclusion and Future work.

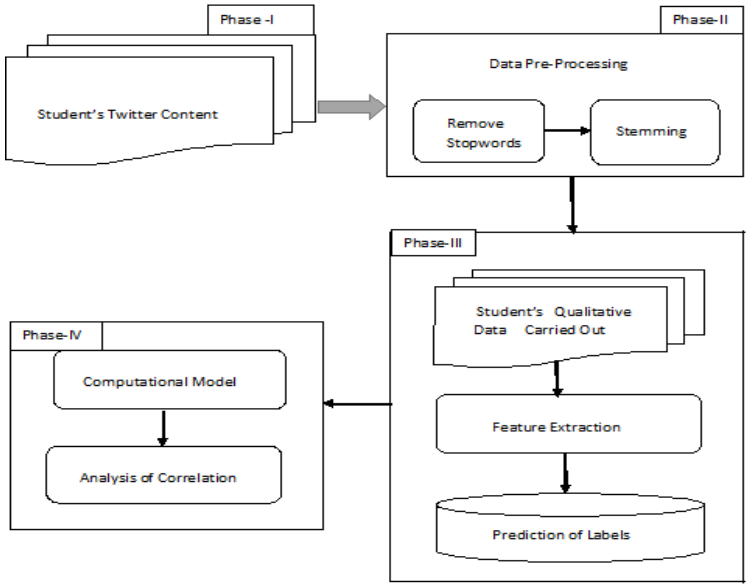
II. Related Work

Text analysis not only provides the negative or positive sentiments of people but also gives many directions to it. Different studies on twitter dataset such as in [6] twitter sentiments were linked with public opinion polls, in [7] twitter sentiments was also applied to predict elections result, in [8] a method was reported for predicting comment volumes of political blogs, in [9] blogs and news sentiments was used to study trading strategies. Different studies on twitter content such as composing a more relevant and complete twitter dataset [1], tweets classification using data compression [10], Enhanced sentiment learning using twitter Hashtag and Smileys [11], Sentiment Analysis and Opinion Mining [15] were useful for present work.

Text classification involves many algorithms with different approaches such as Problem Transformation methods and Algorithm Adaptation methods. Problem Transformation methods include Label PowerSet (LP), Binary Pairwise (BP), Random k-Label Sets Method (RAKEL), Ranking by Pairwise Comparison (RPC), Calibrated Ranking [13]. These algorithms are easy to implement but the complexity increases as the number of labels increases and does not handle the unlabeled tweets.

Algorithm Adaptation methods can change or adapt themselves to meet the results. There are many algorithms which are categorized according to machine learning technique like SVM, Decision Tree, Nearest Neighbors, Information Theoretic[13].Hierarchical algorithms break down the data into small pieces where multiple leaves can be the labels complexity is less but correlation is not possible. Lazy Learning algorithms work on KNN's approach i.e. retrieving the k-nearest examples, each approach in it differ in way of aggregation or clustering the labels from instances.

Fig. 1. Proposed Architecture for Correlational Model for Multiple Labels



III. Students' Tweeter Content

Sentiment analysis can be done on the dataset from twitter which contains sentiwords or expressive words. The analysis is done on each tweet basis, tweets from different hashtags such as #engineeringprobs, #engineeringstudentprob, #engrstudpob, #engrstudindia are taken through twitter API 1.1 [14]. These tweets provide the idea about what students' feel about the particular subject, course, learning methods, exam patterns and recruitment in their engineering majors.

IV. Data Pre-Processing

Social Media users use some different words. For example in twitter # is used to indicate a hashtag on basis of it different users can comment their view on particular trains, words such as wowwwww, grtttt, non-letter symbols, and punctuation mark also brings some noise to data and stopword's which are common words with high frequency like "a", "an", "the", "of", "and", "he", removal of them enhances the performance of features extraction algorithms.

Following are the steps for data pre-processing:

1. Remove # hashtag sign only.
2. Remove @ symbols, Http links, non-letter symbols and punctuation marks.
3. Replace the abbreviated words with their full forms.
4. For repeated words, if one word is coming more than once then replace it one and words which have some definite meaning are kept as it is.
5. Negative words or words with n't is treated as a negative sentiment.
6. Common stop words are removed and Porter stemmer is used to reduce each word to its stem removing any suffixes or prefixes.

V. Prediction Of Labels

To identify the opinion expressed by the tweets, here in we used SentiwordNet. SentowordNet a lexical resource in which each WORDNET synset s is associated with three numerical scores Obj(s), P os(s) and Neg(s), describing how objective, positive, and negative the terms contained in the synset are [2]. For each tweet, positive sentivalue and negative sentivalue will be calculated which is used to find whether particular tweet from the student is positive or negative. Tweets with +ve senti value will be considered as a positive tweet and –ve senti value will be treated as negative emotions or tweets having some issues. Machine learning explores the study and construction of algorithms can learn from and make predictions on data. A supervise machine learning base Multinomial Naïve Bayes algorithm is used for classification and prediction of labels.

Multinomial Naïve Bayes: Multinomial Naïve Bayes [3] is a specialized version of Naïve Bayes that is designed more for text document whereas simple Naïve Bayes would represent a document as the presence and absence of particular word (i.e. Bags of words representation), Multinomial Naïve Bayes explicitly models the calculation's to deal with in [3][4]. Suppose d is for document here each tweet is considered as a document, c is

for a category where C= {c_1, c_2, c_3…c_j} and there are i number of words in training data set. To calculate particular word falls in which class

1. Create a vocabulary from the training dataset
2. Calculate the maximum likelihood of a word and posterior probability of a class.

MAPis "maximum a posteriori" i.e. most likely class for a document [3] [4].

$$C_{MAP} = argmax_{c\in C} P(c/d)$$

According to Bayes rule

$$C_{MAP} = argmax_{c\in C} \frac{P(d/c)P(\mathrm{c})}{P(\mathrm{d})} - - - - - - - - - - - - (1)$$

Probability of the document for each class will be same for all so after dropping the denominator

$$C_{MAP} = argmax_{c\in C} P(d/c)P(c) - - - - - - - - - - - - (2)$$

Document d represented as feature x

$$C_{MAP} = argmax_{c\in C} P(x_1, x_2, x_3, \ldots . x_n/c\ )P(c) - - - - - - - - - - - - (3)$$

$$C_{NB} = argmax_{c\in c} P(c_j) \prod_{i\in positions} P(x_i/c_j) - - - - - - - - - - - - (4)$$

Posterior probability of a class:

$$P(c_j) = \frac{doccount(C = c_j)}{N_{doc}} - - - - - - - - - - - - (5)$$

Maximum Likelihood Estimate:

$$P(w_i/c_j) = \frac{count(w_i, c_j)}{\sum_{w\in V} count(w, c_j)} - - - - - - - - - - - - (6)$$

Where *v* = vocabulary

This would give the fraction of times word w_i appears among all words in documents of class c_j[4]. Create a mega document for c_j by concatenating all the documents of class c_j. If we don't have any word present in the training data but i the documents then maximum likelihood of the document will be zero so to avoid this condition Laplace (add 1) smoothing for Naïve Bayes can be used so

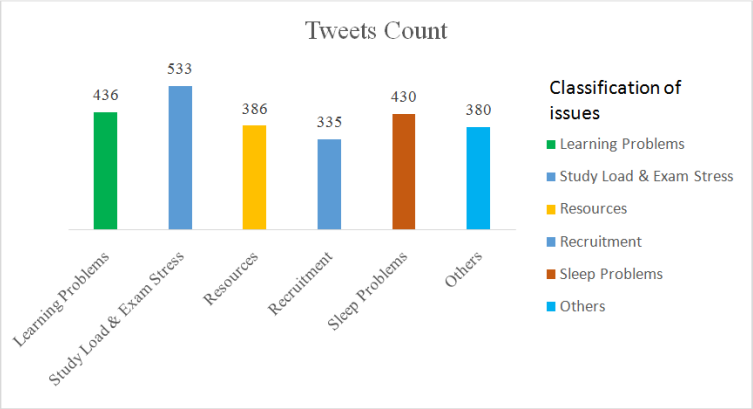
$$P(w_i/c) = \frac{count\ (w_i, c) + 1}{\sum_{w\in V}(count(w, c) + 1)}$$

$$= \frac{count(w_i, c) + 1}{(\sum_{w\in V} count(w, c)) + |v|} - - - - - - - - - - - - (7)$$

VI. Result Of Tweets Classification In Different Cateories By Multinomial Naïve Bayes

Total tweets collected from different hashtags as explained in Students twitter content section are 2500 from this data 30% of it(750tweets) is taken as test data and 70% (1750 tweets) is taken as training data. The class probability of each word is calculated and C_{NB} is calculated for each tweet which classifies it into different categories (Learning problems, Study Load & Exam Stress, Resources, Recruitment, Sleep Problem, Others) which are specified before. The result of classification of tweets using Multinomial Naive Bayes is shown from the following graph.

Fig. 2. The Classification of Tweets Using Multinomial Naïve Bayes



VII. Evaluation Measures For Multi-Label Classifier

There are two types of evaluation measures—Example-Based Measures and Label-Based Measures, Here each tweet will be considered as a document or called as example on each document Example-Based Measure will and then averaged over all documents in the data set, whereas Label-Based measures are calculated based on each label or category and then averaged over all labels (categories)[17].

1. Example-Based Evaluation Measure:

Suppose for a document d the label is Y and predicted set of classifier is Z then

Accuracy: Is the correctly predicted number of labels divided by the number of labels in the union of Y and Z.

$$Accuracy\ a = .\frac{1}{m}\sum_{i=1}^{m}\frac{Y_i \cap Z_i}{Y_i \cup Z_i}$$

Precision: Is correctly predicted number of labels divided by total number of labels in Z.

$$Precision\ p = \frac{1}{m}\sum_{i=1}^{m}\frac{Y_i \cap Z_i}{Z_i}$$

Recall: Is correctly predicted number of labels divided by the number of true labels.

$$Recall\ r = \frac{1}{m}\sum_{i=1}^{m}\frac{Y_i \cap Z_i}{Y_i}$$

F_1 Measure: Is interpreted as a weighted average of precision and recall.

$$f1 = \frac{1}{m}\sum_{i=1}^{m}\frac{2.p_i.r_i}{p_i + r_i}$$

2. Label-Based Evaluation Measure: Labels based measures are calculated and averaged over each category

Table 1. Contingency Table per Category

	True c	True not c
Predicted c	true positive tp	false positive fp
Predicted not c	false negative fn	true negative tn

For each of the one-versus-all binary classification we can create matrix in contingency table for corresponding category c

Then for category c

$$Accuracy\ a = \frac{tp + tn}{tp + tn + fp + fn}$$

$$Precision\ p = \frac{t_p}{tp + fp}$$

$$Recall\ r = tp/tp + fn$$

$$f1 = \frac{2.p.r}{p + r} = \frac{2tp}{2tp + fp + fn}$$

For categories C_1, C_2, C3…$C_L$there are two measures F_1Micro-averaged and F_1Macro-Averaged

$$Micro - Averaged\ F_{1\ score} = \frac{2.\sum_{j=1}^{L} tp_{cj}}{2.\sum_{j=1}^{L} tp_{cj} + \sum_{j=1}^{L} fp_{cj} + \sum_{j=1}^{L} fn_{cj}}$$

$$Macro - Averaged\ F_{1\ score} = \frac{1}{L}\sum_{i=1}^{L} \frac{2.tp_{cj}}{2.tp_{cj} + fp_{cj} + fn_{cj}}$$

Micro-averaging gives equal weight to each per-document classification decision, while macro-averaging gives equal weight to each category. Thus micro-averaging score is dominated by categories that have a larger number of documents, while macro-averaged F_1 is closer to the algorithm effectiveness on smaller categories. So micro-averaged F_1 is higher for classifiers work well on large categories, while macro-averaged F_1 is higher for classifiers work better on smaller categories [16].

VIII. Correlation Of Labels

The proposed method is found to be able to cluster a given set of a text document into a number of classes depending on their contents where the number of classes is known a prior. Correlation is a measure for finding out the association between two labels. This can be achieved using Association mining. Thus it indicates the direction of a linear relationship between two variables i.e. change in value of one variable will affect the value of other variable in the same or opposite direction. This would specifically address the correlation among students'problems. Educators and researchers using this can focus on actual data analysis and investigate the types of educational learning issues that they perceive as critical to their institutions and students.

IX. Conclusion And Future Work

In this paper we proposed the correlation among different problems of engineering students using tweeter data, it includes three phases which are an extraction of positive or negative or positive sentiments next is the classification of tweets using Multinomial Naïve Bayes algorithm and third is the correlation among classified tweets. This paper is useful to address the reasons in the learning process, this can be used in a different learning environment and work environment to address the problems and reasons for them. Cross-language sentiments on different topics can also be analyzed in future. Different lexicon and images are used to express the sentiments which can be analyzed in future work.

References

[1]. Han van der Veen "Composing a more relevant and complete twitter dataset" Master Thesis [online].Available: http://essay.utwente.nl/67800/1/vanderveen_MA_EEMCS.pdf

[2]. Andrea Esuli∗ and Fabrizio Sebastiani "SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining"[online]. Available: http://nmis.isti.cnr.it/sebastiani/Publications/LREC06.pdf

[3]. Prof. Dan Jurafsky "Text Classification and Naïve Bayes"[online]. Available: https://web.stanford.edu/class/cs124/lec/naivebayes.pdf

[4]. Andrew McCallum and Kamal Nigam" A Comparison of Event Models for Naive Bayes Text Classification"[online]. Available: http://www.cs.cmu.edu/~knigam/papers/multinomial-aaaiws98.pdf

[5]. "sentiword public"[online].Available: https://www.npmjs.com/package/sentiword

[6]. Brendan O'Connor, Ramnath Balasubramanyan and Bryan R. Routledge, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series" in Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 2010, pp.122-129.

[7]. Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment", in Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, Leopoldstraße 139, 80804 Munich, Germany, 2010, pp.178-185.

[8]. Tae Yano and Noah A. Smith, "What's Worthy of Comment? Content and Comment Volume in Political Blogs", Available: http://www.cs.cmu.edu/~taey/pub/yano+smith.icwsm10_submit.pdf

[9]. Wenbin Zhang and Steven Skiena," Trading Strategies to Exploit Blog and News Sentiment", in Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, Department of Computer Science, Stony Brook University Stony Brook, NY 11794-4400 USA, 2010, pp.375-378

[10]. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques", Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing, vol. 10,pp. 79-86, 2002.
[11]. Dmitry Davidov, Oren Tsur, Ari Rappoport," Enhanced Sentiment Learning Using Twitter Hashtags and Smileys", Available: https://www.aclweb.org/anthology/C/C10/C10-2028.pdf
[12]. G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining Multi-Label Data," Data Mining and Knowledge Discovery Handbook, Springer, 2010, pp. 667- 685.
[13]. RajasekarVenkatesan, MengJooEr, "Multi –label classification method based on extreme learning machines," in IEEE international conference control, Automation, Robotics & Vision, Singapore, IEEE, Dec 2014, pp.619-624
[14]. Using the Twitter API, Available: https://api.twitter.com/1.1
[15]. Bing Liu "Sentiment Analysis and Opinion Mining" Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.
[16]. V. Van Asch, "Macro- and Micro-Averaged Evaluation Measures," http://www.cnts.ua.ac.be/_vincent/pdf/microaverage. Pdf, 2012.
[17]. Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan," Mining Social Media Data for Understanding
[18]. Students' Learning Experiences" in IEEE Transactions for Learning Technologies, VOL. 7, NO. 3, JULY-SEPTEMBER 2014