Topical classification of comments in MOOC discussion forums

Bita Akram, Devarshi Pratap Singh, Pranjal Deka, Simerdeep Singh Jolly
Department of Computer Science
North Carolina State University
Raleigh, NC 27606
{bakram,dsingh4,pdeka,sjolly}@ncsu.edu

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction and Background

With the emergence of technology and more specifically Internet, a new era in education is beginning. Taking advantage of accessibility and tremendous resources available on the Internet, open online courses are becoming more and more prevalent. Despite the great opportunity that massive open online courses (MOOCs) provide for learners, many issues arises by the autonomous nature of MOOCs and the lack of social interactions compared to the conventional classrooms. One of the main tools for interaction among learners are discussion forums. Hence, forums are a rich resource for evaluating students' progress, satisfaction and learning. In this project, we aim to classify forum discussions based on their categorical types such as questioning, reflections, scaffolding, and also levels of critical thinking such as not applicable, undeveloped thinking, and critical thinking.

To aim for this purpose, we first conduct feature extraction using bag of words. Since the coding scheme uses certain words to distinguish between different topics, bag of words seems like a suitable candidate to classify between different topics. On the other hand bag of words produce a large number of features, many of which might not be deterministic of the type of our project. Hence, we reduce the dimension of our features using PCA, to capture the significant variations in our data-set. Next, we apply a couple of different classification methods including Naive Bayes and SVM and compare them through cross-validation to pick the best classification approach. Naive Bayes classification method is a fast and easy method to apply. However, having high dimensions SVM might be able to provide more accurate results and less overfitting.

2 Related Work

Many studies have attempted to underestand students' interactions with MOOCs through pattern recognition and data-mining. For example, a study demonstrated in [7] used trace-data to build behavior patterns of low and high achieving students. In another study [8], the relevance of the course threades are ranked using linear regression and generative models.

In this study we are going to classify forum discussions based on their content. Past studies like [10] have shown that supervised models perform better when clustering short noisy data. We hence plan to use supervised models trained on hand-coded forum discussions. An unsupervised clustring method, k-medoids with greedy seed selection approach, is used to cluseter discussion forum comments

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

based on their content in one study [6]. In this study we apply Naive Bayes and SVM to classify the discussion forum comments. SVM has also been used for sentiment analysis in [9], which combines data from multiple resources through unigram models to produce more accurate results.

3 Proposed Method

In this section, we describe our approach of classifying topic category based on discussion text of MOOC discussion forums. Our dataset[1] consists of 487 samples of conversation data in a MOOC discussion forum. In particular, it consists of comments shared between learners in the forum and other attributes like timestamp of comment, title of discussion, category of discussion, type of comment (eg. question, statement, reflection etc) and level of critical thinking the comment represents.

3.1 Data Cleaning

We cleaned the dataset to render it free from all HTML tags. We then created a training dataset that contained the Comment Text and Discussion Category attributes. We further removed those Discussion Category attributes that contained less than 10 samples. This left us with 5 Discussion Categories containing 399 training samples of Comment Text.

3.2 Feature Extraction and Dimension Reduction

We used a bag of words approach to perform feature extraction of all words in our dataset. For the 399 samples in our dataset and using document frequency as our feature parameter, we extracted 4023 features. This was represented in the form of a matrix having 399 rows and 4023 columns, each element representing the weight of a feature. Here we experimented with different feature extraction techniques like removing stop words, bigram technique and unigram technique. Since the unigram technique performed better than others, we moved forward with the unigram matrix. Next we performed dimension reduction using PCA to reduce the dimensions of our feature vector. We observed that 33 features had their standard deviation above 1, hence we reduced our dimensions from 4023 to 33. The plot of our PCA analysis is shown in Figure 1.

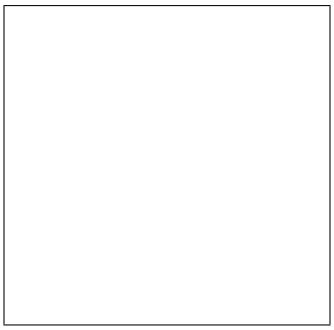

Figure 1: PCA variance plot.

4 Experiments and Results

We divided our dataset randomly into training set and testing set with a ratio of 70% in training set and 30% in testing set. Next we applied SVM tuned on the best set of cost and gamma values and used it on 5 kernel types namely polynomial, radial basic function (gaussian kernel), sigmoid and quadratic. We deduced the confusion matrices for each technique representing our 5 label classes. We recorded the overall accuracy of each of the 5 kernel types. This is represented in Table 1.

Table 1: Accuracy of SVM kernels

Method	Accuracy
Linear	$\sim 73.33\%$
Polynomial	$\sim 58.33\%$
Radial	$\sim 58.33\%$
Sigmoid	$\sim 58.33\%$
Quadratic	$\sim 58.33\%$

Conclusion

References

[1] Kellogg, Shaun, Sherry Booth, and Kevin Oliver. "A social network perspective on peer supported learning in MOOCs for educators." The International Review of Research in Open and Distributed Learning 15.5 (2014).

[2] Zadeh, Reza Bosagh, and Ashish Goel. "Dimension independent similarity computation." Journal of Machine Learning Research 14.1 (2013): 1605-1626.

[3] Ezen-Can, Aysu, et al. "Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach." Proceedings of the fifth international conference on learning analytics and knowledge. ACM, 2015.

[4] Kellogg, Shaun; Edelmann, Achim, 2015, "Massively Open Online Course for Educators (MOOC-Ed) network dataset",doi:10.7910/DVN/ZZH3UB, Harvard Dataverse, V1

[5] FAQ of SVM in R and Matlab http://www.csie.ntu.edu.tw/ cjlin/libsvm/faq.html

[6] Ezen-Can, Aysu, et al. "Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach." Proceedings of the fifth international conference on learning analytics and knowledge. ACM, (2015).

[7] Brinton, Christopher G., et al. "Learning about social learning in MOOCs: From statistical analysis to generative model." IEEE transactions on Learning Technologies 7.4 (2014): 346-359.

[8] Anderson, Ashton, et al. "Engaging with massive online courses." Proceedings of the 23rd international conference on World wide web. ACM, (2014).

[9] Mullen, Tony, and Nigel Collier. "Sentiment Analysis using Support Vector Machines with Diverse Information Sources." EMNLP. Vol. 4. 2004.

[10] Rosa, Kevin Dela, et al. "Topical clustering of tweets." Proceedings of the ACM SIGIR: SWSM (2011).

Appendix

Revised proposal

In the new project, we aim to topically cluster discussions obtained from forums of a course offered for K-12 teacher development. To aim for this purpose we use a data-set that has been hand-coded based on dicussions' content. We train our classifier based on different approaches such as SVM and Naive Bayes models. We then compared the results of applying different techniques on our data-set to find the one that produces the highest accuracy.

The complete proposal can be found here: https://www.sharelatex.com/project/57fbf4cb179e38c4552fc1e8

4.1 Previous Proposal

In our previous proposal, we aimed to exploit efficient data mining techniques to predict categories (eg. social, judicial, political, personal, health, sports) of tweets on Twitter.

Divison of work