# Relevancy Prediction of Micro-blog Questions in an Educational Setting

Mariheida Córdova Sánchez[*]
Information Technology
Purdue University
115 S. Grant Street
West Lafayette, IN 47907
cordovas@purdue.edu

Parameswaran Raman
Department of Computer Science
Purdue University
305 N. University Street
West Lafayette, IN 47907
params@purdue.edu

Luo Si
Department of Computer Science
Purdue University
305 N. University Street
West Lafayette, IN 47907
lsi@purdue.edu

Jason Fish
Information Technology
Purdue University
115 S. Grant Street
West Lafayette, IN 47907
jfish@purdue.edu

## ABSTRACT

Micro-blogging has become increasingly popular in recent years. Using micro-blogging in a large classroom could be beneficial for learning. However, sometimes addressing the large number of posts could be cumbersome to a reader who has only limited time in a classroom. We propose a novel solution for predicting the relevancy of a question asked in a class by looking at the questions asked in previous semesters, the similarity of the question to the lecture material, as well as a set of question features such as the number of students' votes, number of replies, the length of the question, and whether it was asked anonymously. To identify similar questions asked previously, topic modeling and feature selection are used. Empirical results show that topic modeling leads to better prediction performance score as compared to feature selection. The similarity of the question and its corresponding lecture material further improves the relevancy prediction of the questions.

## Keywords

Text Categorization, Micro-blogging, Topic Modeling, Feature Selection

## 1. INTRODUCTION

Using micro-blogging, students are able to ask questions about the material without interrupting the class, which increases student participation. However, answering these questions could be a cumbersome activity as the posts pile up and there is only limited time during the lecture.

In this paper, we propose a novel solution for predicting the relevancy of a question asked in a class by looking at the questions asked in previous semesters and the course lecture material. We also use a set of features such as the number of replies and votes the question received, the length of the question, and whether the question was asked anonymously. To identify similar questions asked previously, topic modeling and feature selection are used. The data consists eight semesters of a Personal Finance course offered at Purdue University.

Cetintas et al. [2] propose a few approaches to this by using the correlation between questions to identify the most relevant and irrelevant questions. In [1], Cetintas et al. propose a text categorization approach that uses personalization, correlation between questions themselves, and students' votes on questions. However, Cetintas et al. do not explore using topic modeling, nor did they explore using feature selection for their classification task. The use of topic modeling for microblog content has been explored by Remage et al. [3].

Empirical results show that topic modeling leads to a better prediction score as compared to feature selection. The similarity of the question and its corresponding lecture material further improves the relevancy prediction of the questions.

## 2. MODELS

For purposes of training and testing the models, the data were divided into two parts in time, which means that the train data corresponds to previous semesters, while the test data belongs to future semesters. Cross validation and regularized were used in all models.

### 2.1 Model using Post Features

[*] This author is also a student in the Department of Computer Science at Purdue University.

A model was built using features from the posts. These features are: the length of the post, the number of votes the post received, the number of replies the post received, and whether or not the post was posted anonymously. These features are referred to as *Post Features* and the model that only uses these features is called *LR_Post*.

## 2.2 Topic Modeling

In order to find which posts are relevant and which ones are not, we first find what topics the students are talking about. The intuition behind this is that we might find that some posts are about topics directly related to the course, while other topics are regarding projects, assignments, exams, etc. We used Latent Dirichlet Allocation, or LDA, to find a set of topics from the posts.

### *2.2.1 Model using LDA*

The output of the Latent Dirichlet Allocation algorithm is a set of topics with the probability distribution of each post belonging to them. These probability distributions are used as prediction features for this model, along with the *Post Features* discussed in section 2.1. We call this model *LR_LDA_Post*. We experimented using different number of topics and terms and chose 10 topics with 15 terms in each since we observed the best performance with this combination.

### *2.2.2 Model using feature selection*

Another approach to topic modeling used was to take the most popular terms of each topic and only consider those terms, disregarding all other terms belonging to the topics. For the top terms of each topic, we find the term frequency in the post. We then have a set of features, which are the frequencies of these terms in the posts. We call this model *LR_FeatSel_Post*.

## 2.3 Model using the lecture material

An important factor when considering the relevancy of a post is what was actually being discussed during that particular lecture. It could happen that the post is relevant to the overall course, but not relevant to the current lecture. For this matter, the similarity of the post to the lecture was calculated. The Kullback-Leibler divergence, or KL divergence, was used. For this, the lecture was divided into smaller overlapping chunks and the similarity between these chunks and the post was then calculated. We explore using different sizes of chunks and chose a size of 100 characters since we observed the best performance. This feature is referred to as *KLD*. *KLD*, together with *Post Features*, forms another model called *LR_KLD_Post*.

## 2.4 Model Comparison

The different models were evaluated using the F1 score. Figure 1 shows a comparison of the models. All the different models shown in this figure include the *Post Features* described in section 2.1. The *Basic* model in the figure contains only the *Post features*, i.e. model *LR_Post*. Following this model, we show the models *LR_FeatSel_Post*, *LR_LDA_Post*, and *LR_KLD_Post*. Then we show the models which include the post features together with topic modeling features and KLD features, i.e. *LR_FeatSel_KLD_Post* and *LR_LDA_KLD_Post*. From this figure we can see that having the *Post Features* alone yeilds the lowest performance score of 0.72. Adding feature selection to it gives us a performance of 0.782, while adding the LDA features to it achieves a performance of 0.795.

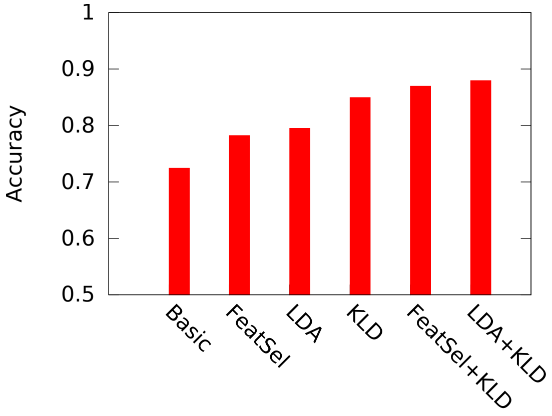

**Figure 1: Model comparison**

Comparing the two topic modeling approaches to the KL divergence approach, we can see that the KL divergence performs better. With the *LR_KLD_Post* model we obtain a performance score of 0.850. Including the *KLD Feature* to both topic models achieves a better performance. When the *KLD Feature* is added to the Feature Selection model, the performance goes up to 0.870. Similarly, when the *KLD Feature* is added to the LDA model, the performance of the model goes up to 0.880.

## 3. CONCLUSIONS

The experimental results show that all the features used in our models are helpful in predicting the relevancy of questions. LDA performs slightly better than feature selection for our application. We also show that adding the similarity of the posts to the lecture material further improves the performance of the techniques.

## 4. ACKNOWLEDGMENTS

This work was partially supported by NSF research grants IIS-0746830, CNS- 1012208, IIS-1017837 and CNS-1314688. This work was also partially supported by the Center for Science of Information, an NSF Science and Technology Center, under grant agreement CCF-0939370.

## 5. REFERENCES

[1] S. Cetintas, L. Si, H. Aagard, K. Bowen, and M. Cordova-Sanchez. Microblogging in a classroom: Classifying students' relevant and irrelevant questions in a microblogging-supported classroom. *IEEE TLT*, 4(4), 2011.

[2] S. Cetintas, L. Si, S. Chakravarty, H. Aagard, and K. Bowen. Learning to identify students' relevant and irrelevant questions in a micro-blogging supported classroom. In *ITS*, volume 2, 2010.

[3] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.