

# Classifying Forum Questions Using PCA and Machine Learning for Improving Online CQA

Simon Fong<sup>1</sup>, Yan Zhuang<sup>1</sup>, Kexing Liu<sup>1</sup>, and Shu Zhou<sup>2</sup>

<sup>1</sup> Department of Computer Information Science, University of Macau, Macau SAR

<sup>2</sup> Department of Product Marketing, MOZAT Pte Ltd, Singapore

{ccfong, syz, mb45462}@umac.mo,

suzyzhou@mozat.com

**Abstract.** As one of the most popular e-Business models, community question answering (CQA) services increasingly gather large amount of knowledge through the voluntary services of the online community across the globe. While most questions in CQA usually receive an answer posted by the peer users, it is found that the number of unanswered or ignored questions soared up high in the past few years. Understanding the factors that contribute to questions being answered as well as questions remain ignored can help the forum users to improve the quality of their questions and increase their chances of getting answers from the forum. In this study, feature selection method called Principal Component Analysis was used to extract the factors or components of the features. Then data mining techniques was used to identify the relevant features that will help predict the quality of questions.

**Keywords:** Community Question Answering, Principal Component Analysis, Machine Learning, Business Intelligence.

## 1 Introduction

CQA is defined as community services which allow users to post questions for other users to answer or respond [1]. It aims to provide community-based [2] knowledge creation services [3]. Lately, CQA websites specifically in the programming context are gaining momentum among programmers and software developers [4]. CQA can provide them a forum for seeking help and advice from their professional peers about technical difficulties that they face.

One of the most popular programming CQA website currently, Stack Overflow, managed to capture compelling technical knowledge sharing among software developers globally [5]. Registered members in Stack Overflow can vote on questions and also answers. The positive and negative votes show the helpfulness and quality of a question and answer. There is a reputation system in Stack Overflow, the members can increase their reputation in the website by participating in various activities like posting questions, answering, voting, posting comments, etc. With better reputations, they can be upgraded with extra capabilities such as editing question/answers and closing a topic.

This study aims to examine the predictors of ignored questions in a CQA service specifically those posted in Stack Overflow, by using machine learning. Thus, there are two main objectives in this study.

The first objective is the identification of the crucial factors or features that affect the quality of the questions. The quality of the questions is divided into two classes: good and bad questions. In this specific context, good questions are defined as the questions that are solved by the community members. Contrarily, bad questions are defined as the ignored questions, which specifically mean the questions without any answers or comments from the online community for at least three months.

The second objective is to investigate the use of feature selection on classification models by using principal component analysis for improving the accuracy of machine learning algorithms, in classifying between good and bad questions. Feature selection technique is used to infer the importance of the features pertaining to the quality of questions in CQA. In this context, an ignored question is defined as question that is without any answers or comments from the community for at least three months.

## 2 Experiment

The importance of factor analysis, in the context of being a prediction model, is about statistically evaluating how important or significant each model attribute is, pertaining to predicting an outcome from an induced model. In this paper we report an experimentation of a classification model that is built over the historical dataset of a CQA service forum.

The objective of the experimentation is in two-fold. The classification model is potentially being used for testing new question posted to the forum so to estimate its chance of being answered – so called the acceptance rate. This is done by comparing the attributes of the new questions to those that have learned by the model from the historical records of the CQA forum, both that have received replies successfully and otherwise. Feature selection is used in the data pre-processing prior to constructing the model for enhancing the classification accuracy. In a nutshell, feature selection is a computational method that chooses a subset of features or attributes for representing the full feature set. It helps reduce the dimensionality of the classification problem. As a desirable side-effect, feature selection algorithm estimates certain ‘contributing factors’ for the model attributes. Such contributing factors are perceived as the extent of significance for comparing the relative impacts of the attributes on the predicted outcomes.

### 2.1 CQA Dataset and Motivation

Stack Overflow’s data is used for this study due to the popularity of Stack Overflow among programmers globally. The data are rich in metadata such as user’s reputation that are suitable to be used for the study. The topic, Java, is chosen for experimentation in Stack Overflow because this topic has been long-term favored in the past decade. Surely the topic has the highest average ratings in popularity from

2002 to 2014. Java-related questions are the most tagged and also the most ignored, with 456,748 and 8,463 respectively. The data for experimentation are collected from Stack Overflow during October 2014. The Data Explorer service provided by Stack Exchange is applied to obtain the data. SQL queries are written and executed in Data Explorer to crawl the required data from the database of Stack Overflow. For this experimentation, we crawled the data with the tag 'Java' of solved and ignored questions starting from year 2008 onwards. After that, disproportionate stratified sampling is used to sample the data from the two categories. A total of 3,000 data are to be sampled, of which 50% are the good questions and another 50% are the bad questions according to the following criterion.

The prediction labels, good questions and bad questions, are exempted from subjective judgments and only confined in the scope of this research. Nevertheless based on the selection of data attributes for characterizing good questions from the past literature, a selection of features that are used for our experimentation are shown in Table 1. A total of 21 attributes are used in this study.

**Table 1.** Features for characterizing high-quality questions

Category	Sub-category	Features
Meta data features	Asker's user profile	Reputation
		Days since joined
		Upvotes
		Downvotes
		Upvotes/Downvotes
		Questions asked
		Answers posted
Content Features	Question	Answers posted /questions asked
		Time
	Textual features	Day
		Tags
		Title length
		Question length
		Code snippet
	Content appraisal	Wh word
		Completeness
		Complexity
		Language error
		Presentation
		Politeness
		Subjectivity

## 2.2 Classification Algorithms

In this experiment, classification algorithms of machine learning are used to investigate the usefulness of the identified features to predict good and bad question in Stack Overflow. The performances of the popular classification algorithms are investigated: logistic regression, support vector machine (SVM), decision tree, naïve Bayes and k-Nearest Neighbors. These algorithms are implemented in the machine learning platform called Scikit-learn (<http://scikit-learn.org>). Scikit-learn is a free open source Python module integrating a wide range of state-of-the-art machine learning algorithms, with the emphasis on ease of use, performance, documentation completeness and API consistency [6]. The classification algorithms used are briefly explained below.

**Logistic regression:** Logistic regression applies the maximum likelihood estimation after transforming the features into a logistic regression coefficient. In this study, the logistic regression classification algorithm is applied with its L1 regularization parameter enabled in the feature selection process as it was found to be insensitive to the presence of irrelevant features. The logistic regression model is written as [7]:

$$p(y = 1 | x; \theta) = \frac{1}{1 + \exp(-\theta^T x)} \quad (1)$$

where  $\theta$  are the parameters of the model.

In the regularized logistic regression, obtain  $\theta$  that solves the following optimization problem:

$$\arg \max_{\theta} \sum_{i=1}^m p(y^{(i)} | x^{(i)}; \theta) - \alpha R(\theta) \quad (2)$$

where  $R(\theta)$  is a regularization term that is used to penalize large weights/parameters. If  $R(\theta) \equiv 0$ , then this model is the standard, unregularized, logistic regression model with its parameters estimated using the maximum likelihood method. If

$R(\theta) \equiv \|\theta\|_1 = \sum_{i=1}^n |\theta_i|$  then this is the  $L_1$  regularized logistic regression.

**Support Vector Machine (SVM):** SVM is considered as one of the best classification algorithm for many real world tasks, mainly due to its robustness in the presence of noise in the dataset used for training, and also high reported accuracy for many cases.

**Decision Tree:** The classification and regression tree CART algorithm of the decision tree implemented in the Scikit-learn module are used. One of the benefits of using decision tree for classification is the ease of interpretability of the models and results. However, if there are irrelevant features in the training data, decision tree algorithm can create overly complex trees used for classification and causing over fitting (do not generalize the training data to unknown new data).

**Naïve Bayes:** One of the advantages of naïve Bayes is that it requires less training data to perform a classification compared to other algorithms. Its main disadvantage is that it cannot learn the interactions between the features, because a particular feature is assumed to be unrelated to other features, as mentioned earlier. In addition, naïve Bayes can suffer from oversensitivity to redundant or irrelevant features [8].

**K-Nearest Neighbors (k-NN):** In the testing phase,  $k$  is a user-defined constant, and a data point is classified by a majority vote of its neighbors, the class assigned is the class most common among its  $k$  nearest neighbors. The classification is highly depending on the number of nearest neighbor,  $k$ . Euclidean distance is a commonly used distance metric to find out the  $k$  nearest neighbors of a data point in the data space [9]. For this particular dataset,  $k=55$  is found to give optimal classification results.

In the experiment, the stratified 10-fold cross-validation approach was used. Two evaluation metrics are adopted in validating the performance of the classification, the two evaluation metrics are: accuracy and area under the receiver operating characteristic (ROC) curve. In classification, accuracy is used to measure of the performance of binary classification test, the correctness of the algorithm in identifying or excluding a condition; the accuracy is the fraction of correctly classified results (both true positives and true negatives) of the overall results. The formula for accuracy is shown in equation (3), where TP is true positives: the number of positive classes that are classified as positives and TN is the true negative: the number of negatives classes that are classified as negatives. In the study, both positive class (questions with an accepted answer by the asker) and negative class (questions that are completely ignored) are equally important classes to be considered in the evaluation of the classification performance. Therefore, the evaluation metric: accuracy is used because it considers both true positives and true negatives in the measurement, unlike precision measure that only take the true positives into account.

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (3)$$

In addition to accuracy, the area under the ROC curve (AUC) is also used as an evaluation metric for a reliable comparative study. ROC is a curve that represents the performance of a classification model with the true positive rate (TPR) and false positive rate (FPR) when the discrimination threshold is varied. The TPR is the fraction of correct positive results to all the positive samples, whereas FPR is the fraction of incorrect positive results to all the negative samples. An ROC curve is a two-dimensional representation of the performance of a classification model. Similarly, AUC, the calculated area under the ROC curve, can be used to measure the performance of classification models. The AUC is statistically useful in the sense that the AUC is equivalent to the probability that the classification model will rank a positive class higher than a negative class that is randomly chosen.

### 2.3 Factor Analysis

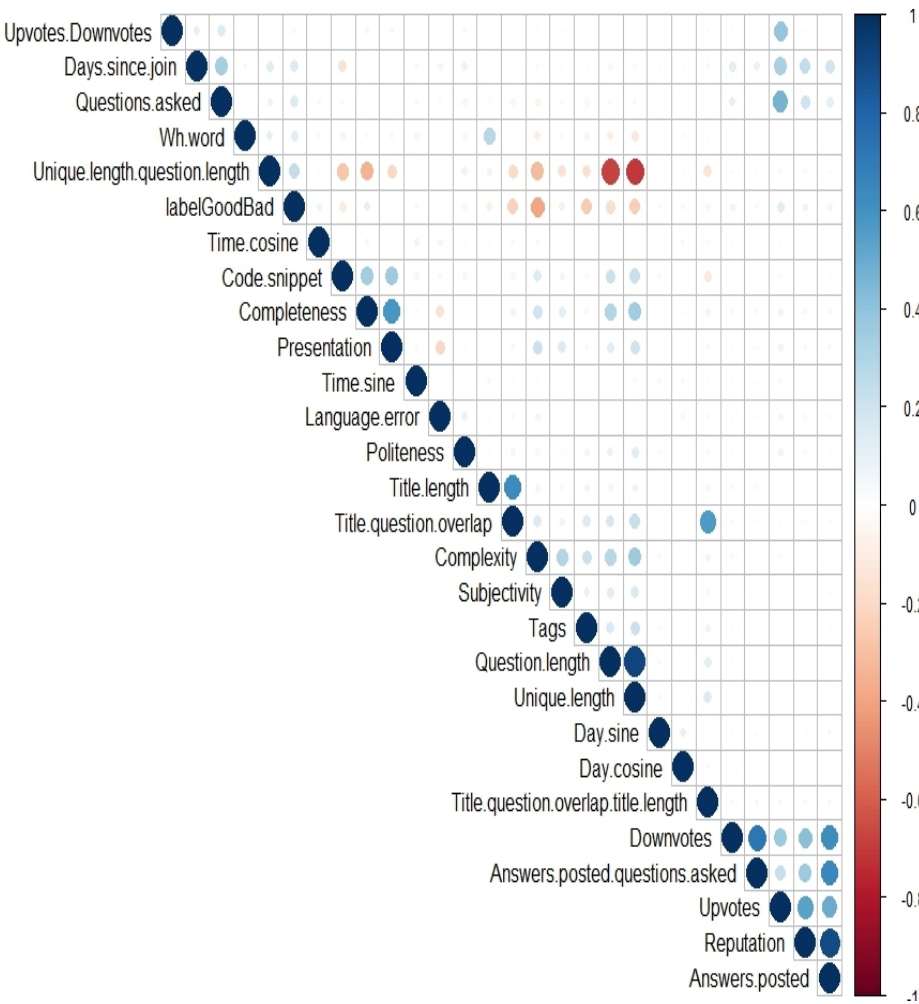
Principal component analysis (PCA) is a popularly used method for extracting factor extraction as an exploratory type of factor analysis. It attempts to identify complex interrelationships among the attributes and the combinations of attributes that contribute to inducing a classified concept.

Factor weights are calculated in order to select the maximum possible variance, followed by further factoring continuing until there is no more meaningful variance remains. This approach helps summarize the variances in the dataset characterized by

many attributes. Each attribute represents a different dimension. It could be difficult for a human to visualize a multi-dimensional hyperspace of attributes greater than three.

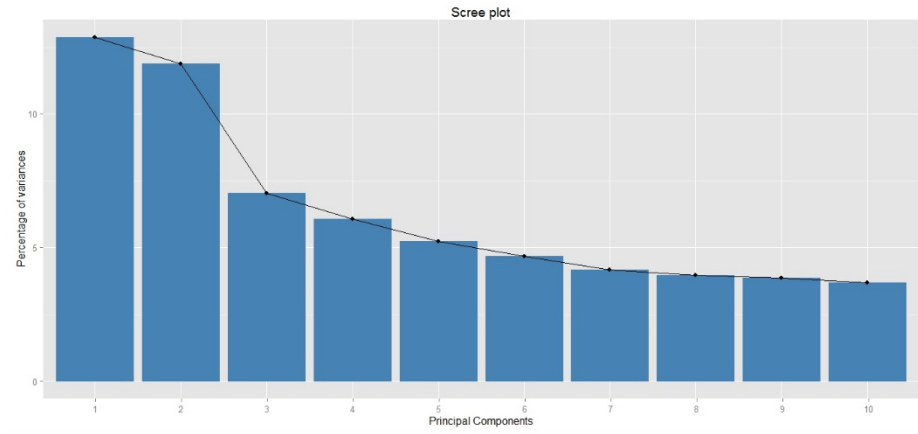
In essence the goal of PCA is to transform the original attributes into a new set of variables which reflect the variation in the data. These new variables that are formed corresponding to a linear combination of the initial attributes and are called principal components.

In this experiment, PCA is used to first reduce the dimensionality of the attributes of the data for constructing a classification model; and then the selected attributed are visualized graphically with minimal loss of information. Users can visually inspect the importance of each of these selected attributes pertaining to the predicated class. The correlation between the classification model attributes are calculated, and visualized as a correlation matrix via a correlogram in Figure 1.



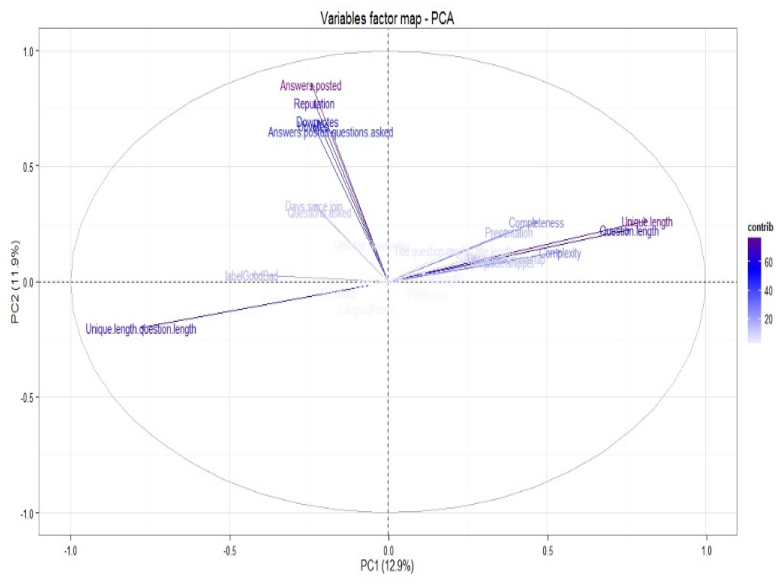
**Fig. 1.** The correlation matrix among the model attributes

The dataset is then processed by R with the built-in function of PCA. The generated eigenvalues are corresponding to the extent of the variation reflected by each principal component (PC). Eigenvalues should be strong for the first PC and their sizes shrink for the subsequent PCs. As shown in Figure 2 the first two eigenvalues constitute to almost 25% that indicates the first few PC account for sufficient variance.



**Fig. 2.** Plot of eigenvalues vs Principal components

A R-based visualization package called FactoMineR (<http://factominer.free.fr/>) is used to chart up the cos2 on a Factor Map which is shown in Figure 3. The most



**Fig. 3.** Variable Factor Map

important variables in the determination of the principal components are highlighted in the hue of red, gradually down to blue.

From the correlation matrix, which is shown in Figure 1, there are two negatively correlated attributes, Unique questions and Question lengths. That shows questions that are unique but they are not long in sentences and vice-versa. Concise questions are mostly unique; long questions are duplicated or cross-posted in different places. When it comes to judging whether a question is a good question, by the variable called labelGoodBad, it is most negatively correlated with complexity. That means questions that are too complex nobody would want to reply; simple questions attract prompt replies. The same negative correlation goes for questions that are overlapped, having inappropriate tags, and overly lengths. These are undesirable features in a CQA which may lead to unanswered dead posts. With regards to favourable questions/posts, as reflected by the variable called Upvotes, they are positively correlated to reputation of the asker, the number of questions asked (that translate to the experience of the posters), and the days since joined.

From the Variable Factor Map, as shown in Figure 3, several most contributing factors stand-out; they are, in order of importance: Answers-posted, Reputation, Unique question length, Upvotes and Downvotes. In terms of machine learning, these are the attributes that contribute most to the mapping of the attributes values to the predicted classes. They contribute to the prediction power by their relative significances. The results can be interpreted that usually experienced posters who have a reputation, who post unique and concise questions would likely invite replies from the peer users. And their questions are likely to be voted, either up or down, implying certain popularity is attracted by these questions.

## 2.4 Classification Model Performance

The validation of the performance of classification is used to verify the usefulness and reliability of the attributes to predict good and bad questions in Stack Overflow. Table 2 shows the average accuracy and AUC from the stratified 10-fold cross-validation for all the classification algorithms used, containing both the average and AUC without feature selection (using the whole set of attributes) and with feature selection by PCA.

The value of the accuracy and AUC has a maximum value of 1 if the prediction of good and bad questions from the classification model is 100% correct, and a minimum value of zero if all the predictions are wrong. For a binary classification task in the study, the value of accuracy or AUC for a random guess is 0.5, which means that the classification model is useless if either the value accuracy or AUC falls under 0.5.

There are three important information that can be depicted from the results of the classification performance in Table 2. First, it is found that all the classification algorithms perform reasonably well in the prediction of good and bad questions. Without the feature selection, the overall accuracy is found to be ranging from 0.574 (naïve Bayes) to 0.735 (logistic regression). The AUC ranges from 0.759 (naïve Bayes) to 0.816 (logistic regression and SVM). This basically means that it is a



feasible option to make use of the identified features to determine whether the questions are good or bad in Stack Overflow.

Secondly, the overall performance of the classification improves by replacing the features with a smaller set of features obtained from the feature selection step. With the inclusion of feature selection, the lowest accuracy actually improves to a 0.698 (k-NN) and the highest accuracy stays the same at 0.735, whereas the lowest AUC increases to 0.763 (k-NN). Among all the classification algorithms, the performance of naïve Bayes improves significantly with feature selection. This is because as mentioned earlier, naïve Bayes is sensitive to redundant or irrelevant features [8]. Therefore, this essentially means that the feature selection step successfully determines a subset of highly relevant and significant features, which serves a better representation of the dataset compared to using all the original features.

Thirdly, two classification algorithms, namely: logistic regression and SVM are found to have the best overall performance both in terms of accuracy and AUC, when compared to other algorithms. This is expected because these classification algorithms represent some of the best performing supervised learning methods in the current age [10].

**Table 2.** Average accuracy and AUC from 10-fold cross-validation.

Algorithms	Without feature selection		With feature selection	
	ACC	AUC	ACC	AUC
Logistic regression	0.735	0.816	0.735	0.813
SVM	0.734	0.816	0.735	0.813
Decision tree	0.728	0.780	0.732	0.784
Naïve Bayes	0.574	0.759	0.712	0.777
k-NN	0.694	0.763	0.698	0.763

### 3 Conclusions

Factor analysis was used to extract component of attributes and these components are used to build a classification model for classifying questions between the good (questions that contain at least an accepted answer by the askers) and bad questions (questions that are completely ignored by the community) in a CQA, with a case of topic Java in Stack Overflow.

The outcome of this study covers computational techniques for quantitatively finding the important features that are useful in the classification of good and bad questions. The features that are revealed from this study were extracted from the metadata of the askers' user profile and questions, as well as from the contents of the questions, which include textual features and content appraisal features.

From the analysis, it was found that when predicting the quality of a question, a user does not consider the previously asked similar or exactly same questions that are solved or unsolved. Information about the previously asked similar questions can be

useful in classifying whether the newly asked question whether it is good or bad. In this way, the forum will be improved by providing guidelines on how to post questions that are likely to be answered. Likewise forum rules can be established for preemptively discouraging questions of poor styles to be posted to the forum. This improvement can be possible only when the insights between good and bad questions in CQA is known. Machine learning and feature selection methods offer such possibility.

**Acknowledgement.** The authors of this paper would like to thank Research and Development Administrative Office of the University of Macau, for the funding support of this project which is called “Building Sustainable Knowledge Networks through Online Communities” with the project code MYRG2015-00024-FST.

## References

1. Li, B., Jin, T., Lyu, M.R., King, I., Mak, B.: Analyzing and predicting question quality in community question answering services. In: Proceedings of the 21st International Conference Companion on World Wide Web, pp. 775–782. ACM, April 2012
2. Chen, L., Zhang, D., Mark, L.: Understanding user intent in community question answering. In: Proceedings of the 21st International Conference Companion on World Wide Web, pp. 823–828, April 2012
3. Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J.: Discovering value from community activity on focused question answering sites: a case study of stack overflow. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 850–858 (2012)
4. Barua, A., Thomas, S.W., Hassan, A.E.: What are developers talking about? An analysis of topics and trends in stack overflow. *Empirical Software Engineering*, 1–36 (2012)
5. Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G., Hartmann, B.: Design lessons from the fastest q&a site in the west. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2857–2866, May 2011
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E.: Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12, 2825–2830 (2011)
7. Ng, A.Y.: Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In: Proceedings of the Twenty-First International Conference on Machine Learning. ACM, July 2004
8. Ratanamahatana, C.A., Gunopulos, D.: Scaling up the naive bayesian classifier: using decision trees for feature selection. In: Proc. Workshop Data Cleaning and Preprocessing (DCAP 2002), at IEEE Int’l Conf. Data Mining, ICDM 2002 (2002)
9. Weinberger, K., Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbor classification. *Advances in Neural Information Processing Systems* 18, 1473 (2006)
10. Caruana, R., Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 161–168. ACM, June 2006

Soft Computing in Data Science

First International Conference, SCDS 2015, Putrajaya,  
Malaysia, September 2-3, 2015, Proceedings

Berry, M.W.; Mohamed, A.; Yap, B.W. (Eds.)

2015, XV, 276 p. 94 illus. in color., Softcover

ISBN: 978-981-287-935-6