Predicting the Usefulness of Questions in Q&A Communities: A Comparison of Classical Machine Learning and Deep Learning Approaches

Langtao Chen

Department of Business and Information Technology
Missouri University of Science and Technology, Rolla, MO 65409, USA
chenla@mst.edu

Abstract. Questioning and answering (Q&A) communities have become an important platform for online knowledge exchange. With a vast number of questions posted to elicit high-quality solutions as well as a large number of participants engaged in online knowledge sharing, a grand challenge for Q&A communities is thus to effectively and efficiently identify and rank useful questions. The current approach to solving this problem is either through user voting or by community moderators. However, such manual processes are limited in terms of efficiency and scalability, especially for large Q&A communities. Thus, automatically predicting the usefulness of questions has significant implications for the management of online Q&A communities. To provide guidelines for assessing the quality of online questions, this research investigates and compares various classical machine learning and deep learning methods for predicting question usefulness. A dataset collected from a large Q&A community was used to train and test those machine learning methods. The findings of this research provide important implications for both the research and practice of online Q&A communities.

Keywords: Q&A Communities, Question Usefulness, Machine Learning, Deep Learning.

1 Introduction

Users are increasingly participating in online questioning and answering (Q&A) communities such as Yahoo! Answers, Reddit, and forums hosted on Stack Exchange to seek answers to their questions and/or provide solutions to solve others' problems [1, 2]. In 2021 alone, Stack Exchange network had 3.2 million questions posted[1]. That means there were on average 365 questions asked on the platform in every single hour. The efficiency and effectiveness of problem-solving in Q&A communities, however, depends on how quickly the submitted questions are made noticeable to experts with relevant knowledge as well as how potential answer providers perceive the usefulness of the questions. Accordingly, large online Q&A platforms such as Reddit and Stack Exchange have adopted the mechanism of user voting to filter/rank questions submitted to the community. Users can voluntarily and anonymously vote up or vote down questions submitted. Questions with the highest user votes are displayed on the top of the question list or recommended to potential problem solvers with the highest priority.

However, user voting of questions is not efficient especially in large online communities, since it requires a significant amount of cognitive effort spent in assessing various quality aspects of the content submitted. Furthermore, the voluntary nature of user voting in most online communities may lead to a systemic problem due to the error of omission [3]. Studies have shown that the percentage of users participating in content voting is relatively low in various online settings [3, 4]. In addition, user voting may also be seriously biased under certain conditions [5]. Thus, to facilitate effective and efficient knowledge exchange, an imperative task for Q&A communities is to automatically predict the usefulness of questions by using machine learning methods.

Machine learning is to learn patterns from data without explicit programming. There are two broad approaches to machine learning: classical machine learning and the recently developed deep learning methods. Although deep learning methods have shown prospects in various applications especially when large amounts of training data are available, the classical machine learning methods are still popularly applied in numerous scenarios. In the context of online Q&A communities, questions remain as to: (1) how classical machine learning and deep learning methods can be implemented to assess the usefulness of questions, (2) what are the design principles that can guide the implementation of machine learning methods, and (3) under what conditions deep learning methods would perform better than the classical methods.

[1] https://stackexchange.com/about (accessed on February 13, 2022)

To provide guidelines for research and practice, this research investigates the application of a set of classical machine learning and deep learning methods for predicting the usefulness rating of questions in online Q&A communities. A large dataset collected from a Q&A platform was used to train those machine learning methods and compare their predicting performance. The results of this research provide important implications for both the research and practice of online Q&A communities.

This paper is organized as follows. The next section reviews work related to the prediction of question usefulness, machine learning, deep learning, and word embedding methods for machine learning. Then, research method is explained in section 3. Section 4 presents preliminary results. The last section discusses the current work and future directions for improving the performance of predictive models.

2 Related Work

2.1 Usefulness of Questions

Rating the usefulness of user-generated content is a common mechanism on online platforms. For example, consumers can rate the usefulness of customer reviews posted by others [6, 7]. In Q&A communities, not all questions posted to the communities have an equal opportunity of being solved. Those questions that are perceived useful are deemed to receive more attentions from potential experts who have sufficient knowledge and experience to solve the problems. Thus, appropriately composing a question can often determine whether and how long the question will be solved. This can be comprehended from the perspective of signaling theory. Signaling theory suggests that people assess the quality of content through a variety of cues or signals that can help reduce information asymmetry between the information signaler and recipient [8]. Thus, knowledge seekers purposively include important information in their questions such that the questions could attract attention and interest from other peers in the community. Guided by the theoretical framework of signaling theory, this research proposes that a set of important cues can signal the usefulness of questions.

Specifically, there is an abundance of basic linguistic cues that can be used to transfer purposive information from one party to another. As presented in Table 1, a set of important cues such as informativeness, diversity, media richness, readability, spelling, and sentiment can be used to explain or predict the usefulness of questions in Q&A communities. In addition, features of Linguistic Inquiry and Word Count (LIWC) can also be used to predict the usefulness of questions. The validity and reliability of LIWC features have been verified by previous studies [9-11].

Table 1. Description of basic linguistic features.

Usefulness Cues	Definitions	Sample Studies
Informativeness	The amount of information embedded in the content	[12-14]
Diversity	The extent to which diverse topics are discussed in the content	[15, 16]
Media richness	The extent to which visual information (e.g., images) is included in the content	[15, 17]
Readability	The ease of reading the content by others	[13, 18]
Spelling	The level of correct spelling in the content	[6, 13]
Sentiment	The strength of opinion expressed in the content	[13]

2.2 Machine Learning and Feature Engineering

Machine learning methods can automatically learn structural patterns from data. In various application scenarios where analytical solutions are not possible and a dataset is accessible, machine learning methods are often preferred to construct empirical solutions such as spam filtering, credit scoring, product recommendation, and image recognition. The well-known no-free-lunch (NFL) theorem proposed by Wolpert [19] suggests that there is not such a single machine learning algorithm that performs best for all learning tasks. In other words, a compari-

son of machine learning methods (both classical and deep learning approaches) is needed for a specific domain task. A typical machine learning process includes data processing, feature extraction, feature selection, model training, model evaluation, and implementation.

A key factor for the success of machine learning projects is feature engineering that generates and prepares a set of important features from the raw data [20]. The process of feature engineering is also the key difference between classical machine learning methods (such as Linear Regression, Decision Trees, Support Vector Machines, Random Forests, and AdaBoost) and the recently developed deep learning methods. Classical machine learning methods rely on a manual process of feature engineering in which a set of important features need to be extracted from the raw data by experts, while deep learning methods have the capability of automatically extracting multiple levels of features from raw data [21].

2.3 Deep Learning

The recent advances in deep learning methods have motivated researchers and practitioners to apply deep neural networks to predict outcomes in numerous applications. Compared to classical machine learning methods, deep learning methods are more computationally expensive. Interestingly, deep learning methods tend to have good performance even when models overfit data [22], a phenomenon generally called benign overfitting [23]. With recent advances in algorithms and hardware, deep learning has emerged as an attractive learning algorithm for various applications including the classification or prediction of user-generated content on social media [24, 25]. Specifically, convolutional neural network (CNN) and recurrent neural network (RNN), the two major types of deep learning algorithms, have been used for various natural language processing and text mining tasks [26]. CNN was originally developed for image recognition by using convolution layers to automatically extract important features. RNN processes sequential data by using a loop structure to connect early state information back to the current state. Long-short term memory (LSTM) is a specific RNN model that was originally developed to learn long-term dependencies in the data [27].

2.4 Word Embedding

Machine learning methods applied for text mining usually require a specific type of embedding methods that map the raw data (characters, words, documents, etc.) to vectors that can be further fed into the machine learning models. The word2vec model [28] and doc2vec model [29] are two popular wording embedding methods for text mining such as sentiment analysis [30], online content quality assessment [31], and news classification [32]. Both the word2vec and doc2vec embedding methods can be used as an alternative to the traditional bag-of-words (BOW) approaches such as TF-IDF (term frequency-inverse document frequency) matrices.

Since the word2vec method only supports vector representation for words, the vector representation cannot be directly used for predictive analytics at document level. In practice, word2vec representations need to be aggregated to document level for document classification. Being an extension of the word2vec model, the doc2vec method directly learns the continuous representation of documents. Doc2vec is particularly attractive for various text mining tasks given its capability in capturing semantic meanings from textual data. Thus, this research applies the doc2vec embedding method. Specifically, two variants of doc2vec including distributed memory (DM) and distributed bag-of-words (DBOW) models are used to extract vector representations of online questions.

3 Research Method

An experiment was conducted to implement various classical machine learning methods and deep learning approaches to predict the usefulness of questions. Then those predictive models were compared. The following subsections explain the details of research method used in this study.

3.1 Data

The dataset was collected from a community-based open Q&A website for user experience designers and professionals. In the community, users can ask questions related to the design of user interfaces and answer questions posted by other peers. After a user submits a question to the community, other users can vote up or vote down the usefulness of the question. Those questions with the highest net votes (i.e., positive votes – negative votes) are displayed on the top of the question list so that all community users can first view them when looking at the question list. Fig. 1 shows a sample question with usefulness votes.

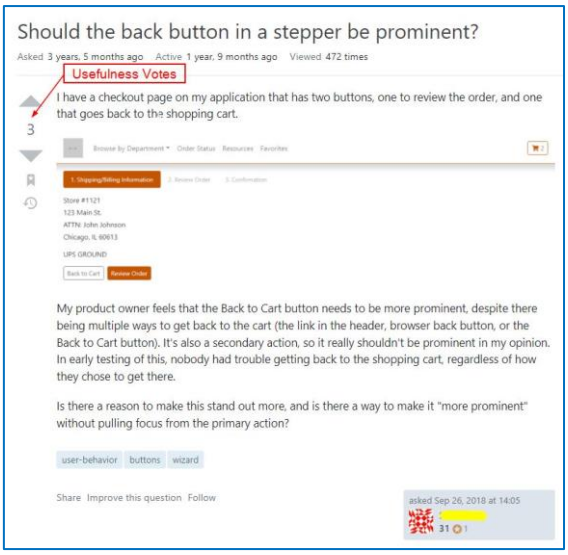

Fig. 1. A sample question with usefulness votes.

The dataset contains 30,718 questions posted from January 2010 to November 2021. The whole dataset was split into a training set of 24,574 questions (80%) and a test set of 6,144 questions (20%). The training set was used to train machine learning models, with the test set used to test the performance of these models.

3.2 Predictive Modeling

Given that a question posted to the community can be voted up and down, usefulness of the question is dichotomized as a binary variable.

$$Usefulness = \begin{cases} 1, \text{if } up\ votes - down\ votes \geq 1 \\ 0, \text{if } up\ votes - down\ votes \leq 0 \end{cases}$$

Fig. 2 presents the overall predictive modeling procedure. After the dataset was collected from the online Q&A community, important features were extracted from the raw data. Specifically, the feature set includes basic linguistic features (explained in Table 1), LIWC features calculated by using the software tool LIWC [10], TF-IDF matrix as BOW features, and doc2vec features (using both DM and DBOW models) trained by utilizing the Gensim package [33]. In total, 1,216 features were extracted. Then, classical machine learning methods including logistic regression, support vector machines, decisions trees, and random forests were applied to classify usefulness based on features extracted. In addition, a CNN deep learning model was directly applied to the textual data to classify usefulness of questions. Finally, all predictive models were compared in terms of their predictive performance.

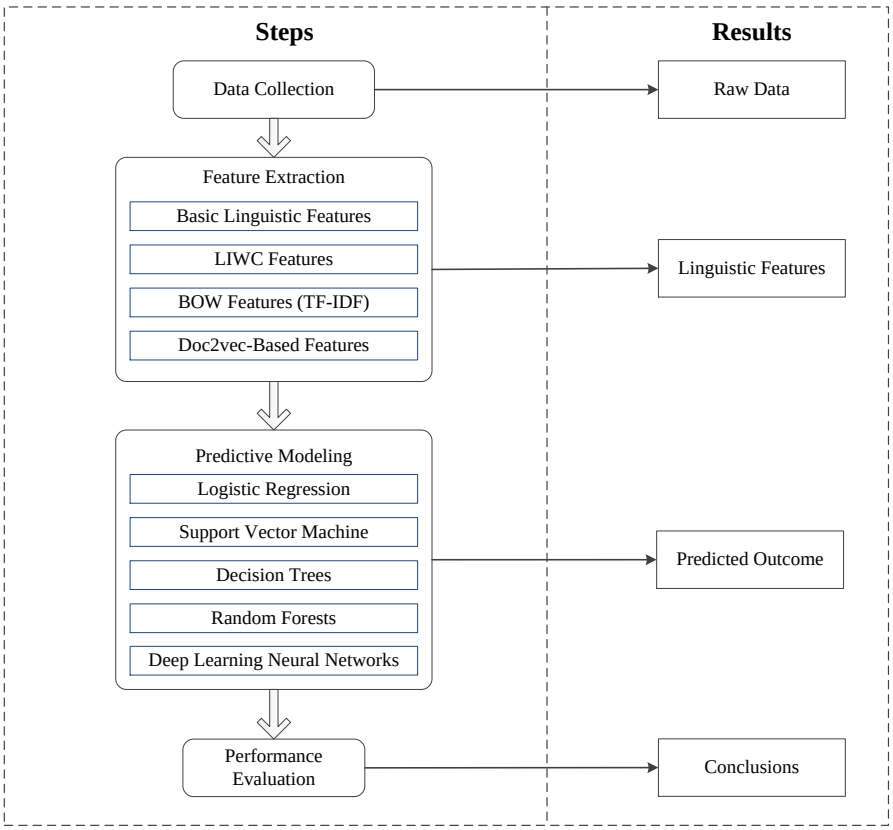

Fig. 2. Predictive modeling procedure.

3.3 Feature Selection

The importance of all features was evaluated by applying a random forests algorithm. Fig. 3 presents the importance scores of all 1,216 features.

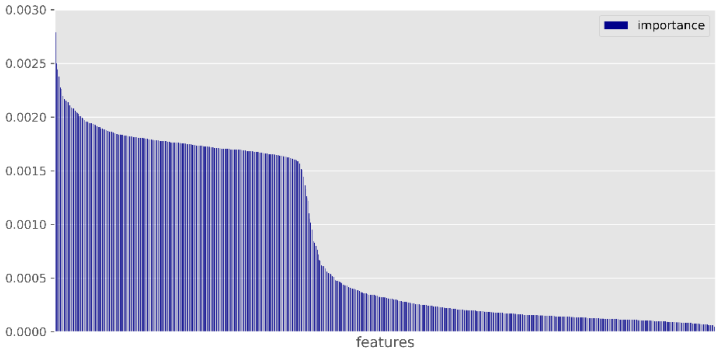

Fig. 3. Feature importance.

To reduce the dimensionality of predictive models, only the 600 most important features were selected for classical machine learning modeling. Table 2 presents a summary of those most important features with their average importance scores. Among all 600 important features, 400 features are trained from doc2vec models (i.e., 200 features from doc2vec DBOW, and 200 features from doc2vec DM). This clearly shows the capability of doc2vec models in deriving important features.

Table 2. Summary of top 600 most important features.

Feature category	Number of features	Mean importance
Doc2vec DBOW	200	0.0020
Doc2vec DM	200	0.0016
LIWC	84	0.0016
Basic linguistic feature	12	0.0014
BOW (TF-IDF)	104	0.0005

4 Preliminary Results

Table 3 summarizes the preliminary comparison of both classical and deep learning models. Among all machine learning models compared, random forest has the highest level of accuracy (0.6918), F1 score (0.8139) and recall (0.9544), whereas logistic regression has the highest level of AUC (area under the curve of ROC, 0.6286). The CNN model that directly learns word embeddings from the textual data achieves a mediate performance. This result indeed shows the need for theoretical guidance for classical machine learning modeling. With strong theoretical bases (such as signaling theory in this study) guiding feature engineering, classical machine learning methods could outperform deep learning methods. The result also shows the prospect of deep learning methods in automatically extracting important features for textual content classification. In application situations where strong theoretical guidelines are not possible, deep learning approaches still can reach a good performance, thanks to their capabilities of automatically extracting important features.

Table 3. Comparison of predictive models.

Method	Accuracy	AUC	F1 score	Precision	Recall
Logistic regression	0.5838	0.6286	0.6629	0.7743	0.5795
SVM	0.5911	0.5452	0.7074	0.7150	0.6999
Decision tree	0.5953	0.5139	0.7129	0.7144	0.7114
Random forest	0.6918	0.5382	0.8139	0.7095	0.9544
CNN	0.6234	0.5330	0.7420	0.7211	0.7641

5 Discussion

Online Q&A communities have offered an excellent opportunity for people to solve their problems without temporal and spatial constraints. To effectively seek answers, questions need to be composed in a way that can reduce the information asymmetry between knowledge seekers and potential knowledge providers. Informed by signaling theory, this research suggests that a variety of linguistic features can be used to predict the usefulness of questions submitted to Q&A communities. Specifically, this research has explored various classical machine learning and deep learning methods for predicting question usefulness.

As demonstrated in the preliminary results in section 4, this study has evaluated a set of classical machine learning methods in classifying usefulness of questions. However, only a specific CNN model was evaluated in this study. For the future work, more deep learning neural network structures (such as a simple RNN and an LSTM) will be thoroughly evaluated. Features manually extracted from textual content can also be fed to deep learning structures to test how the deep learning methods perform with those manual features. An ensemble of both classical machine learning and deep learning methods can also be further evaluated. Importantly, grid search strategy will be used to tune numerous hyper-parameters in deep learning models.

Future work can also model the prediction of question usefulness as a regression problem by applying a variety of regression models to predict the natural count of usefulness votes. Findings of this research will provide practical and theoretical implications for improving the effectiveness and efficiency of knowledge exchange in online Q&A communities. Machine learning algorithms provide a technical approach to automatically fil-

ter/rank questions submitted to online Q&A communities, without the need for usefulness voting by users. This brings rich opportunities for designing new online community features or mechanisms that can address the grand challenge of supporting effective online knowledge exchange.

References

1. Chen, L., Baird, A., Straub, D.: Why do participants continue to contribute? Evaluation of usefulness voting and commenting motivational affordances within an online knowledge community. Decision Support Systems 118, 21-32 (2019).
2. Chen, L., Baird, A., Straub, D.: The impact of hierarchical privilege levels and non-hierarchical incentives on continued contribution in online Q&A communities: A motivational model of gamification goals. Decision Support Systems 153, 113667 (2022).
3. Liu, X., Wang, G.A., Fan, W., Zhang, Z.: Finding useful solutions in online knowledge communities: A theory-driven design and multilevel analysis. Information Systems Research 31, 731-752 (2020).
4. Cao, Q., Duan, W., Gan, Q.: Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. Decision Support Systems 50, 511-521 (2011).
5. Ochi, M., Matsuo, Y., Okabe, M., Onai, R.: Rating prediction by correcting user rating bias. In: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 452-456. (2012).
6. Ghose, A., Ipeirotis, P.G.: Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. IEEE Transactions on Knowledge and Data Engineering 23, 1498-1512 (2011).
7. Chen, L.: The impact of the content of online customer reviews on customer satisfaction: Evidence from yelp reviews. Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing 2019, Austin, TX, USA (2019).
8. Spence, M.: Job market signaling. Quarterly Journal of Economics 87, 355-374 (1973).
9. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology 29, 24-54 (2010).
10. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and psychometric properties of LIWC2015. University of Texas at Austin, Austin, TX (2015).
11. Pennebaker, J.W., Francis, M.E.: Cognitive, emotional, and language processes in disclosure. Cognition & Emotion 10, 601-626 (1996).
12. Huang, K.-Y., Long, Y.: Fighting together: Discovering the antecedents of social support and helpful discussion threads in online support forums for cannabis quitters. In: Proceedings of the 52nd Hawaii International Conference on System Sciences, pp. 4319-4328. (2019).
13. Chen, L., Baird, A., Straub, D.: A linguistic signaling model of social support exchange in online health communities. Decision Support Systems 130, 113233 (2020).
14. Mudambi, S.M., Schuff, D.: Research note: What makes a helpful online review? A study of customer reviews on Amazon.com. MIS Quarterly 34, 185-200 (2010).
15. Wu, L.: Social network effects on productivity and job security: Evidence from the adoption of a social networking tool. Information Systems Research 24, 30-51 (2013).
16. Bechmann, A., Nielbo, K.L.: Are we exposed to the same "news" in the news feed? An empirical analysis of filter bubbles as information similarity for danish facebook users. Digital Journalism 6, 990-1002 (2018).
17. Hlee, S., Lee, J., Yang, S.-B., Koo, C.: The moderating effect of restaurant type on hedonic versus utilitarian review evaluations. International Journal of Hospitality Management 77, 195-206 (2019).
18. Yin, D., Bond, S., Zhang, H.: Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. MIS Quarterly 38, 539-560 (2014).
19. Wolpert, D.H.: The lack of a priori distinctions between learning algorithms. Neural Computation 8, 1341-1390 (1996).
20. Domingos, P.: A few useful things to know about machine learning. Communications of the ACM 55, 78-87 (2012).
21. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521, 436-444 (2015).
22. Belkin, M., Hsu, D., Ma, S., Mandal, S.: Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences 116, 15849-15854 (2019).
23. Bartlett, P.L., Long, P.M., Lugosi, G., Tsigler, A.: Benign overfitting in linear regression. Proceedings of the National Academy of Sciences 117, 30063-30070 (2020).

24. Chen, L.: A classification framework for online social support using deep learning. Lecture Notes in Computer Science 11589, 178-188 (2019).
25. Haralabopoulos, G., Anagnostopoulos, I., McAuley, D.: Ensemble deep learning for multilabel binary classification of user-generated content. Algorithms 13, 83 (2020).
26. Chai, J., Li, A.: Deep learning in natural language processing: A state-of-the-art survey. In: 2019 International Conference on Machine Learning and Cybernetics (ICMLC), pp. 1-6. (2019).
27. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9, 1735-1780 (1997).
28. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at the International Conference on Learning Representations. (2013).
29. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning, pp. 1188-1196. (2014).
30. Liang, H., Fothergill, R., Baldwin, T.: Rosemerry: A baseline message-level sentiment classification system. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 551-555. (2015).
31. Dang, Q.V., Ignat, C.-L.: Quality assessment of wikipedia articles without feature engineering. In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, pp. 27-30. ACM, (2016).
32. Trieu, L.Q., Tran, H.Q., Tran, M.-T.: News classification from social media using twitter-based doc2vec model and automatic query expansion. In: Proceedings of the Eighth International Symposium on Information and Communication Technology, pp. 460-467. ACM, (2017).
33. Rehurek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Citeseer, (2010).