Predicting Question Deletion and Assessing Question Quality in Social Q&A Sites using Weakly Supervised Deep Neural Networks

Souvick Ghosh
School of Information
San José State University
souvick.ghosh@sjsu.edu

Abstract

Community question answering (CQA) sites, which use the power of collective knowledge, have emerged as popular destinations for complex and personalized questions that require human-human interactions and multiple rounds of clarifications between the asker and the answerer. In this paper, we undertook a threefold task: First, we developed a deep neural network model to automatically predict the questions that are likely to be deleted by the moderators. Second, we hypothesized that there exists a relationship between the question quality and its probability of being deleted by the forum moderators. We developed a deep model using deleted questions and used it for predicting question quality. Our contribution is not limited to developing the predictor model; we also created the gold standard data for question quality assessment. Lastly, we explored the efficiency of different input representations, optimization functions, and neural network models for predicting question quality. When assessing question quality, the results highlight that combining natural language features with word embeddings can result in better performance (higher recall and f-scores) than word embeddings alone. Our model predicted deleted-questions with an accuracy of 97.8% and precision and true positive rates (TPR) above 0.95. While assessing question quality, our model obtained a TPR of 0.841 and a precision of 0.514. This research serves as the first step toward automatic content moderation in CQA sites; identifying poor quality questions would allow askers to improve the quality of questions asked and the moderators to handle a large volume of questions during content moderation.

1. Introduction

Research in information science focuses strongly on the information need and knowledge gap of the searcher that motivates him to look for information. Asking questions is an act of actively seeking information and is usually motivated by the lack of knowledge about a topic or domain. When the asker is faced with a real-life problematic situation [1], he becomes aware of his anomalous state of knowledge [2, 3] and tries to remedy the situation by collecting more knowledge. Wilson [4] defined this kind of information behavior as active information seeking. While some information needs are satisfied by querying search engines, there are more complex needs that require human assistance.

Community question-answering (CQA) involves using the online community of users or social media forums for knowledge sharing and management. Such platforms offer huge learning potential for the users, as the askers get easy access to the available knowledge base (questions that have been answered in the past and accepted as helpful by the askers). The users can also ask newer or more specific questions. When the searcher submits a question, in the form of a query, to the search engine, the latter returns either the best answer from one of the many question-answering sites or provides links to several question-answering sites. The design of the system has huge implications in the popularity and authenticity of the site. An open Q&A system attracts users from different backgrounds, but the massive footfall also leads to low-quality questions and therefore, unregulated content which overwhelms the system. To maintain the popularity of the site, it is important to assess the quality of the questions and the posted answers and to eliminate low-quality content. The content regulation and deletion are usually performed by human annotators, but the process is time-consuming and expensive.

In this research, we have primarily focused on how to detect the quality (and the deletion probability) of the questions that are being posted in community Q&A sites. We used the questions posted in a popular educational Q&A site, Brainly, and assessed the question quality using deep neural networks with word-level and natural language features. First, we have developed a supervised deep neural network model to automatically predict the questions that are likely

URI: https://hdl.handle.net/10125/70943
978-0-9981331-4-0

to be deleted by the moderators. The labels were generated using historical data in the forum database (deleted or not) and could be considered as weak-labels which are not gold standard. Second, we hypothesize that the questions deleted by moderators are more likely to be of poor quality (there could be other reasons for deletion like repetitions, wrong sub-forums). As deep neural networks can efficiently process large volumes of data and recognize underlying patterns in the data, we have used our model to automatically predict the quality of the questions posted on CQA sites. Lastly, we have explored the efficiency of different input representations, optimization functions, and neural network models for predicting question quality.

The results highlight that for certain tasks and configurations, input representations combining natural language features with word embeddings perform better than word embeddings alone. Our final model, which uses long short-term memory (LSTM) with Adam as the optimization algorithm and binary crossentropy as the loss function, predicted deleted questions with an accuracy of 97.8%. Another LSTM model, using both word embeddings and NLP features as input vector, predicted the quality of the questions with a true positive rate of 0.841. This research should serve as the first step toward automatic content moderation in CQA sites; by identifying and flagging inappropriate and low-quality questions for the moderators to look at, thereby scaling down millions of posted questions to just a few thousand.

In this paper, we answered the following research questions:
RQ1:*Given a collection of questions, posted by different users in a community question-answering site, can we automatically identify the questions which are likely to be deleted by the moderators?*
RQ2: *Using a gold standard collection of questions, whose quality was assessed by forum moderators, how can we automatically assess the quality of questions in a scalable and accurate manner?*
RQ3:*What are the characteristics of an ideal neural network model – the input representations, the objective function, and the architecture – which can be used for accurately detecting question quality?*

The rest of the paper is arranged as follows: Section 2 discusses the Background and Related work, and Section 3 explains the dataset used in the research. Section 4 highlights the experimental methodology, while Section 5 analyzes the results. Section 6 provides an overall discussion of the paper and results, and the conclusion and future work have been discussed in Section 7.

2. Background and Related Works

Research on community question-answering aims to develop better platforms for knowledge sharing and archiving. This is a two-fold task that can be achieved by viewing the CQA sites from the perspective of the system as well as that of the user. The system-focused approach explores the design of the system in terms of content, domain, cost, and responsiveness. Harper et al. [5] categorized question-answering sites into three broad categories - digital reference services, ask-an-expert service, and community Q&A sites, while Srba and Bielikova [6] divided CQA sites into educational and organizational. Harper et al. also explored the design implications of such question-answering platforms. The Q&A site may be paid (expert service) or free (crowd-based service), which influences the type of answers obtained. The authors observed that paid sites led to longer and better answers that had higher probabilities of satisfying the information needs of the asker. Many Q&A sites incorporate the social links between users to provide better recommendations and a more enriching experience. Wang et al. [7] analyzed Quora (a popular Q&A service) and investigated the three different network graphs used by Quora - user graphs, question graphs, and user-topic graphs. The user graphs help in connecting users with similar interests, while the question graphs connect similar questions. The user-topic graph links users to the topics and helps in relevance assessment. Other research works have connected the quality of questions to the quality of the answers [8, 9], arguing that high-quality questions beget high-quality answers, thereby increasing the popularity of the CQA site. Promoting or rewarding high-quality content attracts more users, provides better user experience, and leads to more high-quality content [10].

Gazan [11] and Choi and Shah [12] explored the motivations of the users for asking questions in online Q&A services. They concluded that although cognitive needs are the most significant motivations, other factors also play a key role in motivating the users to ask questions. Different contexts and situations influence the askers to opt for human-human interactions in place of human-computer interactions. Previous research on identifying high and low-quality questions has been limited to using topic modeling approaches [10] or traditional machine learning approaches with feature engineering [13, 14, 15]. Correa and Sureka [16] and Rath et al. [14] tried to identify the reasons that contribute to question deletion in educational Q&A. While the former used profile-, community-, content-,

and syntax-based features, the latter used only textual features for prediction. Such classifiers – which use supervised approach – tend to overfit the small training data and are unsuitable for large-scale application.

Although neural networks have often failed to reach elevated heights of performance using unsupervised learning, yet, supervised learning is not always feasible. Employing human annotators is expensive and time-consuming, and click-through data is not always the most reliable in terms of accuracy. For example, the data which we collected for this study did not come with gold standard labels. The CQA forum moved deleted questions into a separate database table called *deleted*. While it is true that the moderators could have deleted these questions for various reasons, it could be also be erroneous. The decision to delete a question is a subjective assessment made by the moderator. The actions of the moderator are not reviewed and lack sufficient quality control. As such, our labels (we labeled any question obtained from the deleted table as *Delete-Yes*) are logical but not perfect and therefore, we call them weakly supervised labels. By training the neural network using weakly supervised labels [17, 18], and adjusting the parameters, it is possible to solve the problems encountered in supervised (expensive) and unsupervised approaches (no labels) [19].

3. Dataset

In this section, we explain the details of training and test data.

3.1. Training Data

To answer our research questions, we collected data from a popular social Q&A site called Brainly[1]. Initially, we collected around 10,000 questions that were written in English and posted in the CQA forum for a period of twelve months from 2016-2017. Five thousand of these questions were deleted by the site moderators based on various factors – too broad, ambiguous, general statement, poor syntax, socially awkward, redundancy, garbled – related to question quality. The site administrators stored these questions in a separate table (*Deleted*) in the database. The other 5,000 questions collected from active tables survived the moderation phase and were deemed to be above a certain threshold of quality. All the questions belonged to a single subject domain. We created the dataset to have a balanced representation of questions that were deleted and those that were not. Training a machine learning model using human-annotated labels is both expensive and time-consuming. In our work, to evaluate which questions are likely to be deleted by the moderators, we used a weakly supervised approach, where we assigned the labels "delete-no" to the questions that survived moderation, and "delete-yes" to the questions that were deleted by moderators.

[1] https://brainly.co/

3.2. Test Data

For evaluating question deletion and quality, we have used two separate gold standard datasets, each of which contained around 1,000 instances. Both the test datasets were manually labeled by human annotators (unlike the weakly supervised labels of training data).

3.3. Question Deletion

To create this test dataset, we asked three annotators to examine the weak labels for questions deleted (Deleted-Yes and Deleted-No). The annotators assessed if the moderators deleted the questions rightfully. The inter-annotator agreement was high, with Fleiss' kappa score above 0.62.

3.4. Question Quality

This gold standard data was created by forum moderators who annotated each question as either "good" or "poor," based on eight parameters. The parameters used by the moderators to detect poor quality questions were as follows:

Too Broad (TB): If the question was too generic and lacked specific details. In such cases, the asker is not sure of his information need and the lack of specificity confuses the readers.

Ambiguous (AM): If the question was unclear, the moderators considered the questions to be of a poor quality.

General Statement (GS): If the asker presented a statement and not a proper question. Such instances tend to confuse the readers and moderators often delete these questions.

Socially Awkward (SA): If the question asked was socially awkward and risked upsetting a segment of readers.

Poor Syntax (PS): If the question was not framed following the rules of the grammar or syntax.

Missing Information (MI): If the question was missing important information required by the reader to answer it.

Redundancy (RE): If the question was redundant, that is, it had appeared in the forum before, or if the content repeated itself.

Garbled/Non-Formatted (NF): Most Q&A forums require the questions to be syntactically correct, often requiring separate styles for quotes, codes, etc. Non-formatted question or those with garbled characters are often deemed to be of poor quality.

For each parameter, the number of evaluation categories was two (true or false). There were a total of 3 annotators, who assessed around 1,000 questions for content quality. If any of the 8 parameters were true, the question was considered to be of poor quality. For each of the 8 parameters, there was only a slight agreement between the moderators (Fleiss' Kappa scores between 0.1 and 0.18), which further highlights the complexity of this task. The final verdict was reached using majority voting (if 2 out of 3 annotators annotate the question to be of poor quality, we label it as *poor*) Out of the 938 questions contained in the test set, 449 were assessed as "good" quality and 489 as "poor" quality questions. While predicting question deletion is a straightforward task, question quality is a subjective assessment and hence, it is hard to get a strong agreement on a binary scale.

4. Methodology

First, we concentrated on building an ideal and efficient input representation for our model. Next, we explored two deep neural architectures and several configurations of hyperparameters to build the prediction model. We trained our deep models using weakly annotated data and tested the performance on gold standard datasets.

4.1. Input Representation

To answer our research questions, we explored the different types of input layer representations, z_0 which are fed as input to the neural network. The feature representation function ψ maps each question instance into a vector of features: (i) a dense vector representation in which we represent questions (which may contain multiple sentences) using their word embeddings; (ii) a sparse vector representation using only the traditional natural language features like n-grams, dependencies, and part-of-speech tags; and (iii) a higher dimensional sparse representation combining the previous two representations, where we used various natural language features in addition to the word embeddings obtained previously.

4.1.1. Dense representation using word embeddings: In this representation, each question (which may contain multiple sentences) in the dataset was input to the embedding function $\mathcal{E}$, such that: $\mathcal{E} : \mathcal{V} \rightarrow \mathbb{R}^m$ (where $\mathcal{V}$ denotes the vocabulary set and m is the embedding dimension). We used Google's pre-trained Word2Vec model which has a vocabulary of 3 million words obtained from the Google News dataset. The length of the output vector was 300, which was our embedding dimension. The word embeddings helped in capturing the rich linguistic context of the words (as each word was projected onto a 300-dimensional space based on their semantic proximity) [20, 21]. For a question q, which contained a total of n words (w), the feature extraction function ψ concatenates (||) the word embeddings of individual words (obtained using the embedding function) using the merge function M. For the training phase, the weak annotation a_q for question q was appended at the end of the feature vector.

$$\psi(q)_{train} = [\mathcal{M}_{i=1}^{n}(\mathcal{E}(w_i)||a_q], \tag{1}$$

The final representation was a vector of 72,001 dimensions for training and 72,000 for test (without the weak annotation). As the questions differed in the number of words that they contained, we have used zero padding to convert each input vector to a fixed dimension.

4.1.2. Vector representation using NLP features: In this representation, we calculated the different natural language features from our data and used them as part of the feature vector. The different components of the input vector were as follows:

Named Entity vector ($f1$): This seven-dimensional vector represents all the named entity types present in the question. The representation is in one-hot form, where the vector had a 1 for every type present.

Part-of-speech vector ($f2$): The part-of-speech tags for every word in the sentence were annotated using Stanford Part-of-Speech Tagger. We represented the part-of-speech as a 36-dimensional vector by identifying the 36 most frequent parts-of-speech in the data. The presence of each part-of-speech type is represented using a 1 in the vector.

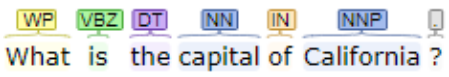

Figure 1: Part-of-speech Relations

Parent ($f3$) *and grandparent relations* ($f4$): A single word can be used in different contexts, and its meaning differs based on the context of use. Parent and grandparent relations in the dependency parse tree

was used for word sense disambiguation [22]. To calculate this feature, we created the dependency parse tree using Stanford CoreNLP ParserFor a given word, we determined its relation to its parent and grandparent. In the example presented in Figure 2, for the word 'California', the parent is 'capital' (relation: nmod) and the grandparent is 'What' (relation: nsubj). The dependency relation vector space is limited to 54 dimensions, and the presence of each relation is represented by a 1 in the one-hot form.

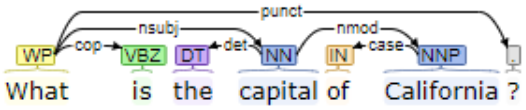

Figure 2: Dependency Relations

Punctuation marks ($f5$): The punctuation marks in the question are useful in determining the underlying structure of the question. A properly framed question should have adequate punctuation. For example, a long sentence should have commas, a question with multiple sentences should have periods, and every question should end with the question mark. We have used a 12-dimensional vector to represent the counts of different punctuation symbols present in the question, where each dimension belongs to a specific punctuation symbol.

Total number of sentences ($f6$): A single question may comprise one or more sentences. The number of sentences present in the question can be influential in deciding the wordiness, or the clarity of the question. More sentences may help in clarifying the question asked, thereby making the question more detailed and less ambiguous. It could also make the question wordier and unnecessarily complex, making it hard for the answerer to understand it properly.

Number of Words per Sentence ($f7$): A single dimensional vector which counts the number of words in the asked question and averages (normalizes) it by the number of sentences present.

Total number of characters ($f8$): The number of characters present in the question is indicative of the length of the sentence. While the number of words shows how informative the question is, the number of characters can also include non-dictionary words, special characters, garbled letters, and white spaces.

Total number of misspelled words ($f9$): If a question contains many misspelled words, it is likely to be a poor-quality question. With the availability of word processors that automatically perform spell checks, it is not difficult for the asker to eliminate any misspelled words. However, many questions in community question-answering forums contain misspelled words that reflect the lack of effort from the asker, and hence, poor quality of the question.

Total number of words ($f10$): The number of words in the question can be indicative of how detailed or wordy the question is. Descriptive and detailed questions are usually of higher quality, although the wordiness can make them more difficult to read and understand.

Total number of interrogative words ($f11$): If a question contains more interrogative words, it is likely that the asker has combined several questions in a single post, which may adversely affect the quality of the question asked.

If the question contains any URL ($f12$): A question that contains a URL is likely to be referring to some other Webpage and may be influential to how the answerers and readers perceive the quality of the question.

If the question ends with interrogation mark ($f13$): A question should always end with an interrogation mark. The absence of interrogation mark indicates that the post may be a statement and not a question.

If the question contains numbers ($f14$): Presence of numbers in the question is a powerful indicator of the topic of the question. While the numbers themselves do not reflect the quality of the question, they provide insight into the topic or domain of the question, which may be useful in assessing quality.

Presence of taboo or profane words ($f15$): Taboo or profane words may offend the readers and experts answering the question, which in turn may reduce the response rate to the question, or may lead to off-the-topic discussions in the forum. A question that contains taboo or profane words is usually of lower quality.

Readability scores (Flesh Kinkaid scores) ($f16$): The readability scores help in assessing the difficulty level of understanding a question or text. A higher readability score is better for any posts in the community question-answering sites that are frequented by non-native speakers and school students. More complicated texts with jargons and complex structures limit the audience of the question and can, therefore, be considered lower quality.

If question begins with interrogative word ($f17$): Presence of interrogative words (what, where, when, who, whom, why, how) at the beginning of a sentence is a good practice when asking a question. This increases the specificity and focus of the question.

If question starts with a lower-case letter ($f18$): Any text which begins with a lower-case letter is non-

conformational, and hence can be considered of lower quality.

All the features from $f6$-$f18$ are single dimensional. For any question q, containing a total of n words (w), the feature representation function ψ concatenates (||) all the NLP features. The weak annotation a_q for question q was appended at the end of the feature vector for training data. The input representation thus obtained contained 177 dimensions for training and 176 for test instances.

$$\psi(q)_{train} = [f1||f2||f3||f4||....||f18||a_q], \quad (2)$$

4.1.3. Vector representation using all features: In this representation, we have combined all features from the previous two representations, i.e., we concatenated the 72,000-dimensional dense vector of word embeddings with the 176-dimensional NLP features. The feature representation function concatenated the weak annotation at the end to obtain a final representation of 72,177-dimensions for training data.

4.2. Architecture of the Deep Neural Network

We have used two neural network architectures – Multilayer Perceptron and Long Short-Term Memory (LSTM) Networks. Multilayer Perceptron is a type of feedforward artificial neural network that has been widely used in image recognition, speech recognition, and machine translation problems. LSTM, however, is better at capturing the context of a word or sentence, and hence is more suitable for tasks involving sequential information. For building the final model, we have explored architectures with different number of hidden layers, and layers with different number of neurons [23]. The final models – one MLP and the other LSTM – are 6-layered with four hidden layers and one input and one output layer (Figure 3). Following are the details of the hyperparameters used:

Activation Function: For any artificial neural network, in the absence of an activation function, the output signal will be a linear combination of inputs, and the model will be reduced to a linear transformation model. Most real world problems are non-linear, and modeling of complex non-linear functions requires activation, which introduces non-linearity into the model and maps complex input signals to outputs. In our model, we have explored and used four different activation functions in different layers of our model: Rectified Linear Units (relu), Hyperbolic Tangent (tanh), and Sigmoid. tanh is symmetric around the origin which results in larger derivatives and faster convergence. On the other hand, relu is more effective computationally as only certain neurons get activated at a time. Therefore, we used tanh and relu activations for hidden layers. As we were dealing with binary classification, we used sigmoid in the final output layer (as it transforms values between 0 and 1).

Optimization Functions: Optimization algorithms used in training deep learning models are different from the traditional optimization algorithms [24]. These algorithms do not directly influence the performance measure P. Instead, they aim at reducing a cost function J which is expected to reduce P. In this research, we have worked with two adaptive methods of optimization - RMSProp [25] and Adam [26]. Both of these methods use minibatches and automatically adjust the learning rates of model parameters. In our deep model, RMSProp optimizes the loss function in shorter number of epochs compared to Adam.

Dropout: Dropout attempts to regularize the model so that it could learn diverse parameters. By masking certain parameters in the hidden units, it forces the model to learn more efficiently, using different patterns every time. In our model, we have varied dropout rates between 1% to 5%.

Number of Epochs: The number of epochs or iterations were varied in steps between 50 and 200. Although all the reported results are for 200 epochs used with early stopping technique, the loss function and accuracy measures for different epoch sizes were obtained. The curves for accuracy and loss function smooths after 150th epoch, therefore, setting the number of epochs below 150 could affect the performance of our prediction model.

Loss Function: As we were dealing with binary classification for the first two research questions – deleted or survived (not-deleted) questions, and good or poor (not good) quality questions – therefore, we used binary crossentropy as the loss function for our models.

5. Results and Analysis

For both the prediction tasks, the deep neural models were trained using weakly supervised annotations. For the first task (question deletion prediction), although the labels were not gold standard, they were deleted by forum moderators and therefore, could be considered more reliable. For the second task (question quality prediction), the training instances were labeled as "good" if they survived moderation and "poor" if they were deleted. Based on the insights provided by the forum moderators, we hypothesized that there is a strong relation between deleted and poor quality questions. Although questions are deleted for various

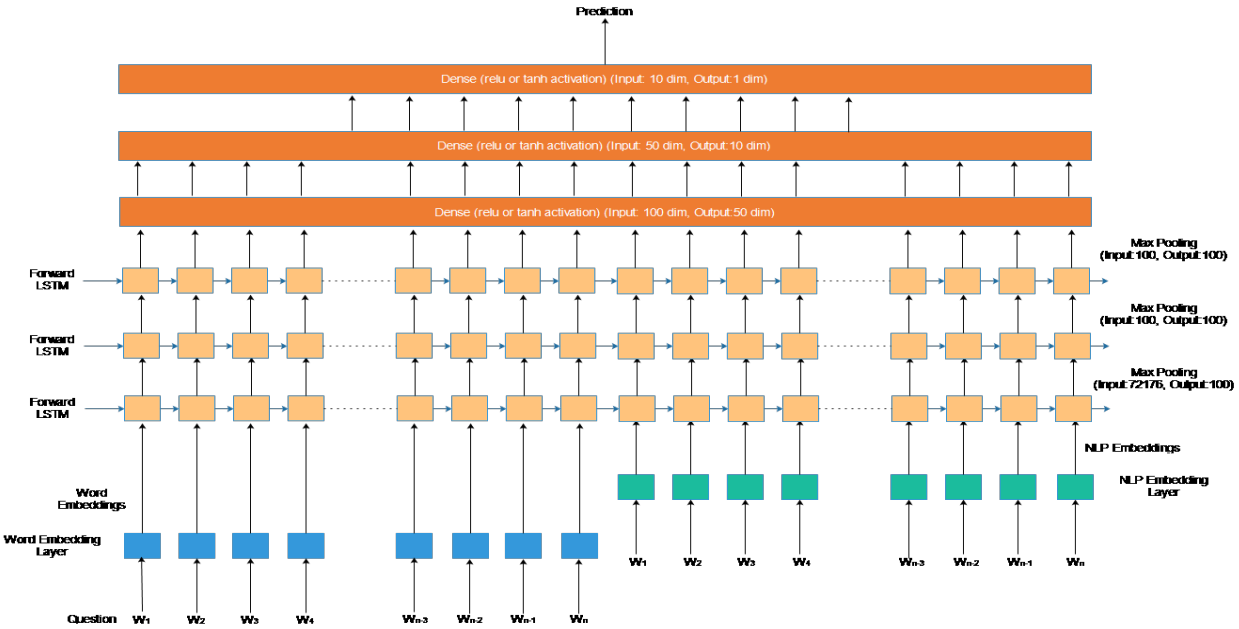

Figure 3: LSTM Model Architecture.

Table 1: Results: Predicting question deletion.

Neural Network	Input Features	Optimization Function	Accuracy	Precision	TPR (Recall)	F-measure
MLP	All features	RMSProp	62.6%	0.568	1.000	0.724
MLP	NLP features	RMSProp	54.9%	0.565	0.359	0.439
MLP	Word embeddings	RMSProp	97.0%	0.989	0.949	0.968
MLP	All features	Adam	95.6%	0.927	0.987	0.956
MLP	NLP features	Adam	55.3%	0.562	0.417	0.478
MLP	**Word embeddings**	**Adam**	**97.8%**	**0.966**	**0.989**	**0.977**
LSTM	All features	RMSProp	94.4%	0.966	0.919	0.942
LSTM	NLP features	RMSProp	55.3%	0.538	0.632	0.582
LSTM	Word embeddings	RMSProp	97.5%	0.976	0.974	0.976
LSTM	All features	Adam	94.5%	0.974	0.912	0.942
LSTM	NLP features	Adam	56.8%	0.569	0.500	0.532
LSTM	**Word embeddings**	**Adam**	**97.8%**	**0.985**	**0.969**	**0.977**

Table 2: Results: Predicting question quality.

Neural Network	Input Features	Optimization Function	Accuracy	Precision	TPR (Recall)	F-measure
MLP	All features	RMSProp	47.8%	0.500	0.002	0.004
MLP	NLP features	RMSProp	53.3%	0.538	0.736	0.622
MLP	Word embeddings	RMSProp	50.0%	0.515	0.680	0.587
MLP	All features	Adam	49.7%	0.542	0.237	0.330
MLP	NLP features	Adam	52.0%	0.534	0.628	0.577
MLP	Word embeddings	Adam	49.9%	0.524	0.421	0.467
LSTM	All features	RMSProp	48.7%	0.507	0.614	0.555
LSTM	NLP features	RMSProp	53.5%	0.569	0.442	0.498
LSTM	Word embeddings	RMSProp	51.5%	0.535	0.526	0.530
LSTM	**All features**	**Adam**	**50.3%**	**0.514**	**0.841**	**0.639**
LSTM	NLP features	Adam	49.0%	0.514	0.409	0.456
LSTM	Word embeddings	Adam	49.8%	0.517	0.566	0.541

reasons (repetitions, wrong sub-forum, etc.), poor content quality is one of the most frequent causes. Therefore, the weakly annotated labels were expected to be good enough to train the deep neural models. Traditional machine learning approaches would have a hard time dealing with the amount of noise present

in such labeling. However, we expected to design the deep neural architecture such that the resulting objective function would help the model learn the underlying relationship between reasons of deletion and measures of quality. The weak labels helped the model perform better than unsupervised approaches. Weakly supervised approach can learn from massive amounts of data without incurring expenses in terms of manual annotation (as is common in supervised learning).

To test our models, we used two smaller gold standard collections – one for predicting question deletion (RQ1), and the other for assessing question quality (RQ2) – both of which contained labels generated by human annotators. For the reported results in Table 1 and 2, we had set the number of epochs to 200 and binary crossentropy as the loss function. We randomized the order in which the questions were input to the algorithm. We have also repeated our experiment multiple times using different order of input and reported the average of the result.

5.1. Predicting question deletion

For predicting question deletion, we have used the original training dataset, which contained 5,000 questions that were deleted by the moderators and 5,000 questions that survived the moderation phase. For testing, we used a separate set of 1,000 questions, of which 508 were deleted and 492 survived. All the prediction models were tested using this test dataset and the different performance metrics – precision, true positive rate (or recall), accuracy, and f-measure – are reported in Table 1. The results indicate that our deep neural network models have performed exceptionally well while detecting questions that were deleted by the moderators. Looking at Table 1, we can state that the best performing model used word embeddings as input representation, with Adam as optimization algorithm, and binary crossentropy as the loss function. Both the LSTM and MLP models showed equivalent results with reported accuracy being as high as 97.8%, with both precision and true positive rate (TPR) above 0.95. It must be noted that TPR is an important metric while detecting questions that are likely to be deleted. This allows the user to modify his question before posting, so as to avoid deletion by the moderators. The confusion matrix for the best performing model has been presented in Table 3. Only 22 instances were misclassified, with 978 correctly classified instances.

5.2. Assessing question quality

To assess question quality, we trained the deep model using the weakly supervised collection of 10,000 questions. The deletion of a question may involve a number of possible reasons related to context as well as content. As poor quality content is a possible scenario which warrants deleting a posted question, we attempted to detect the underlying patterns of poor content quality by training our neural network on deleted questions. Although the forum moderators identified eight factors that contributed to poor content quality (as reported in Section 3.2), we classified each question instance as “good” or “poor” in quality (binary classification). The performance of the model was evaluated on a test dataset of 938 questions, of which 449 were of “good” quality, and 489 of “poor” quality. Different metrics like precision, true positive rate (or recall), accuracy, and f-measure were calculated and are reported in Table 2.

Table 3: Confusion matrix for best model (deletion).

	Deleted	Survived
Deleted	477	15
Survived	7	501

The results suggest that it is hard for a deep neural network to learn higher levels of abstraction (from small training data), as low accuracies were reported for all the models. However, it must be understood that the assessments of quality are very subjective and vary from one person to another. We obtained the performance of human annotators against the labels obtained using majority voting. The results, presented in Table 4, show highest of 78.4% accuracy for Annotator 3, with precision and true positive rate of 0.75 and 0.89 respectively. The best performing model for assessing question quality – a long short term memory network with one input layer, four hidden layers, and one output layer – used Adam optimization algorithm with binary crossentropy as the loss function, and had a true positive rate of 0.841 and precision of 0.51, which is comparable to the worst performing human annotator. Instead of relying solely on human annotation, which is both expensive and time-consuming, automatic classification provides a cheaper and faster alternative by flagging possible low-quality content. True Positive Rate is a reliable measure for any automatic anomaly detection system. By flagging possible cases for anomaly (such as poor quality questions in our case), it would allow the users to improve the content, and the moderators to delete or moderate the question. The confusion matrix for quality assessment, obtained using the best performing model, has been presented in Table 5. The table highlights that a number of "good" questions were wrongly classified as “poor”. While the model is still premature to automatically moderate content based on quality (without any human involvement), it is certainly

the first step towards automatic content moderation.

Table 4: Performance Measure of Human Annotators.

Annotator	Precision	TPR	Accuracy	F-measure
Annotator 1	0.699	0.896	74.0%	0.785
Annotator 2	0.747	0.850	76.9%	0.796
Annotator 3	**0.750**	**0.889**	**78.4%**	**0.813**

Table 5: Confusion matrix for best model (quality).

	Good	Poor
Good	61	**388**
Poor	78	411

6. Discussion

Users tend to frequent question-answering sites when faced with problems and unclear information needs. Community question-answering sites, which use the power of collective knowledge, have emerged as popular destinations for complex and specific questions that require human-human interactions and multiple rounds of clarifications between the asker and the answerer. As the number of visitors increases, the CQA sites are flooded with low-quality questions that are difficult to understand and answer. Even if these questions are answered, the unclear question could corrupt the quality of the answer. The decline in question quality creates a vicious cycle that reduces the overall content quality and the reputation of the site, which in turn repels new users from visiting the site.

In this research, we attempted to answer three research questions. First, we explored if we can automatically identify the questions which are likely to be deleted by the moderators. Our results are encouraging, with our deep neural network model reporting accuracy of 97.8% for predicting deleted questions, with both precision and true positive rates above 0.95. The model has the potential to be used in online Q&A content moderation. An automated deletion detection system can alert the user that the posted question may not meet certain standards and hence, risk getting deleted. However, more work is required to develop a universal model for cross-domain deletion prediction.

Second, we investigated if we can automatically predict the quality of questions, in a scalable and accurate manner. The results suggest that it is hard to obtain higher degrees of abstraction when using deep neural network with only 10,000 instances of training data. While our model reported low accuracy when assessing content quality, it has a high true positive rate of 0.841 and precision of 0.514. Thus, it can be argued that the proposed deep neural model is a step towards automatic content moderation; by automatically flagging possible low quality content, it can precede human moderation, thus, saving both time and effort.

Lastly, we explored the different configurations of the deep neural models – the input representations, the objective functions, and the architecture – that can be used for accurately detecting question quality. Both our neural networks – MLP and LSTM – were 6-layered with four hidden layers and one input and one output layer. The results do not highlight any clear advantage of one model over the other. For predicting question quality, input representations combining natural language features with word embeddings performed better than those using word embeddings alone. However, the inclusion of NLP features slightly decreased the accuracy of prediction for deleted questions. We also tested the performance of two optimization algorithms – Adam and RMSProp. While RMSProp showed faster convergence during training, Adam optimized the loss function better and led to higher accuracy for both the research questions.

7. Conclusions and Future Work

In this paper, we developed a deep neural network model to automatically predict the questions that are likely to be deleted by the moderators. Both the MLP and LSTM report accuracy as high as 97.8% for predicting deleted questions, with precision and true positive rates of 0.966 and 0.989 for MLP and 0.985 and 0.969 for LSTM respectively. However, there are two limitations pertaining to the dataset. First, the training data is too small to obtain the desired levels of abstraction. Second, the training and test sets contained questions from the same subject domain, which could have led to a higher accuracy. In future, we would like to collect more data from a mixed collection of subject domains. This would help us to better tune the hyperparameters and create more generalizable models. We would also like to evaluate the performance of our system for cross-domain prediction.

To predict question quality, we developed a gold standard data using three human annotators who classified each question to be of poor or good quality based on 8 categories of content quality. Our model reported low accuracy for predicting content quality; the LSTM network showed best results with a high true positive rate of 0.841 and precision of 0.514. In future, we would like to train separate networks on each of the categories of quality assessment and combine the classifications using another network.

While the performance of MLP and LSTM were comparable, Adam outperformed RMSProp while optimizing the loss function. The results also highlighted that the choice of input representation should be made by considering the problem and the level of abstraction required. The inclusion of NLP features did not improve results when predicting question deletion, but it had a significant influence when predicting question quality. We performed ablation analysis on word embeddings and NLP features as a collection and not individually. Therefore, in future work, we plan to look at the NLP features individually for better feature selection. Using more explainable architectures and using other categories of data would be other options.

References

[1] G. Wersig, *Information-Komunikation-Dokumentation*, vol. 5. Verlag Dokumentation, 1971.

[2] N. J. Belkin, "Anomalous states of knowledge as a basis for information retrieval," *Canadian journal of information science*, vol. 5, no. 1, pp. 133–143, 1980.

[3] N. J. Belkin, R. N. Oddy, and H. M. Brooks, "Ask for information retrieval: Part i. background and theory," *Journal of documentation*, vol. 38, no. 2, pp. 61–71, 1982.

[4] T. D. Wilson, "Human information behavior," *Informing science*, vol. 3, no. 2, pp. 49–56, 2000.

[5] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan, "Predictors of answer quality in online q&a sites," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 865–874, ACM, 2008.

[6] I. Srba and M. Bielikova, "Askalot: community question answering as a means for knowledge sharing in an educational organization," in *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, pp. 179–182, ACM, 2015.

[7] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao, "Wisdom in the social crowd: an analysis of quora," in *Proceedings of the 22nd international conference on World Wide Web*, pp. 1341–1352, ACM, 2013.

[8] C. Aritajati and N. H. Narayanan, "Facilitating students' collaboration and learning in a question and answer system," in *Proceedings of the 2013 conference on Computer supported cooperative work companion*, pp. 101–106, ACM, 2013.

[9] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proceedings of the 2008 international conference on web search and data mining*, pp. 183–194, ACM, 2008.

[10] S. Ravi, B. Pang, V. Rastogi, and R. Kumar, "Great question! question quality in community q&a.," in *ICWSM*, 2014.

[11] R. Gazan, "Social q&a," *Journal of the Association for Information Science and Technology*, vol. 62, no. 12, pp. 2301–2312, 2011.

[12] E. Choi and C. Shah, "User motivations for asking questions in online q&a services," *Journal of the Association for Information Science and Technology*, vol. 67, no. 5, pp. 1182–1197, 2016.

[13] B. Li, T. Jin, M. R. Lyu, I. King, and B. Mak, "Analyzing and predicting question quality in community question answering services," in *Proceedings of the 21st International Conference on World Wide Web*, pp. 775–782, ACM, 2012.

[14] M. Rath, L. T. Le, and C. Shah, "Discerning the quality of questions in educational q&ausing textual features," in *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pp. 329–332, ACM, 2017.

[15] J. Bian, Y. Liu, D. Zhou, E. Agichtein, and H. Zha, "Learning to recognize reliable users and content in social media with coupled mutual reinforcement," in *Proceedings of the 18th international conference on World wide web*, pp. 51–60, ACM, 2009.

[16] D. Correa and A. Sureka, "Chaff from the wheat: characterization and modeling of deleted questions on stack overflow," in *Proceedings of the 23rd international conference on World wide web*, pp. 631–642, ACM, 2014.

[17] A. Ratner, B. Hancock, J. Dunnmon, F. Sala, S. Pandey, and C. Ré, "Training complex models with multi-task weak supervision," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4763–4771, 2019.

[18] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 6, pp. 3660–3671, 2016.

[19] M. Dehghani, H. Zamani, A. Severyn, J. Kamps, and W. B. Croft, "Neural ranking models with weak supervision," *arXiv preprint arXiv:1704.08803*, 2017.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[22] J. Hale, "The information conveyed by words in sentences," *Journal of Psycholinguistic Research*, vol. 32, no. 2, pp. 101–123, 2003.

[23] S. Ghosh and S. Ghosh, "Exploring the ideal depth of neural network when predicting question deletion on community question answering," in *Proceedings of the 11th Forum for Information Retrieval Evaluation*, pp. 52–55, 2019.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[25] G. Hinton, N. Srivastava, and K. Swersky, "Rmsprop: Divide the gradient by a running average of its recent magnitude," *Neural networks for machine learning, Coursera lecture 6e*, 2012.

[26] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.