Stack Overflow Questions Quality Classification

Group 11

Parv Thakkar
thakkar.pa@northeastern.edu

Abstract

Stack Overflow is a popular online platform where developers ask and answer programming-related questions. The quality of questions on this platform plays a critical role in the effectiveness of the website in assisting developers in finding solutions to their problems. In this paper, we present a machine learning approach to predict the quality of Stack Overflow questions and categorize them into High Quality, Low Quality requiring Edits, and Low Quality which were closed. I trained the model on dataset of 60,000 stack overflow questions using various features such as title, body, and tags to predict the question quality. A variety of techniques, including Vectorization, LSTM, Logistic Regression, Random Forest, Decision Tree, Confusion Matrix have been developed to train and evaluate the model. The model achieved an accuracy of 83% in predicting the question quality.

1. Introduction

Online platforms such as Stack Overflow play a critical role in the software development community. They provide a vast pool of knowledge, enabling developers to learn from each other's experiences and find solutions to their problems. However, the quality of questions on these platforms is paramount to their effectiveness in assisting developers. Low-quality questions with poor grammar, inadequate information, or vague problem statements make it harder for developers to find relevant answers, which can be frustrating and time-consuming. Therefore, it is crucial to develop an automated mechanism that can evaluate the quality of questions on online platforms such as Stack Overflow.

In this paper, I present an approach to predict the quality of Stack Overflow questions using machine learning and deep learning models. I used a dataset of 60,000 Stack Overflow questions to train our models and categorized them as High Quality, Low Quality with Edits, or Low-Quality Closed. I applied various pre-processing techniques and performed exploratory data analysis to gain insights into the dataset. I then used multiple machine learning algorithms such as Logistic Regression, Multinomial Naive Bayes, Random Forest, Decision Tree, K Neighbors Classifier, and Gradient Boosting Classifier. In addition, I also developed deep learning models using Long Short-Term Memory (LSTM) with different epochs and batch sizes. The goal of our research is to provide a scalable and efficient way of identifying low-quality questions on

Stack Overflow, thus enabling moderators to take appropriate actions to improve the overall user experience of the platform.

2. Literature Survey

Stack Overflow is a popular question-answering platform used by developers worldwide to share their knowledge and seek help from others. However, the quality of questions posted on the platform varies widely, and many low-quality questions receive poor responses or remain unanswered. This has led researchers to focus on developing models that can automatically classify the quality of Stack Overflow questions to improve the platform's overall quality.

Antoaneta Baltadzhieva and Grzegorz Chrupala have worked on predicting the quality of questions on Stack Overflow using various natural language processing (NLP) techniques. In their paper titled "Predicting question quality in question answering forums," they proposed a supervised machine learning approach to classify Stack Overflow questions into three quality categories: low-quality, medium-quality, and high-quality. They used a dataset of 12,000 Stack Overflow questions annotated by human judges to train and evaluate their models. They experimented with different feature sets, including lexical, syntactic, and semantic features, and found that the combination of all three sets achieved the best performance.

In their follow-up paper titled "Predicting the Quality of Questions on Stack Overflow," published in the Proceedings of Recent Advances in Natural Language Processing, Baltadzhieva, and Chrupala improved upon their previous work by using a larger dataset of 60,000 Stack Overflow questions annotated by human judges. They also used a deep neural network model called a Convolutional Neural Network (CNN) to classify the questions' quality. The model was trained on a combination of lexical and semantic features, including word embeddings and part-of-speech tags. They achieved promising results with an F1-score of 0.66 for the low-quality category and 0.78 for the high-quality category.

Overall, the research done by Baltadzhieva and Chrupala shows that machine learning models, especially those using deep learning techniques, can effectively classify the quality of Stack Overflow questions. The models' accuracy can be further improved by using more complex features and larger annotated datasets. These models can be useful for improving the platform's overall quality by automatically identifying and flagging low-quality questions, which can then be moderated or improved by the community.

3. Dataset

The dataset used in our research is the Stack Overflow-Questions-Quality-Dataset, which contains 60,000 questions from Stack Overflow posted between 2016-2020. Each question in the dataset has been classified into one of three categories: HQ, LQ_EDIT, and LQ_CLOSE, based on their quality. The dataset has an equal distribution of all the classes, ensuring that each category has an equal representation. The columns in the dataset include the question Id, Title,

Body, Tags, Creation date, and the target variable Y. The question body is in HTML format, and all dates are in UTC format. The availability of this dataset allowed me to train my models on a large and diverse set of questions, enabling me to accurately predict the quality of Stack Overflow questions.

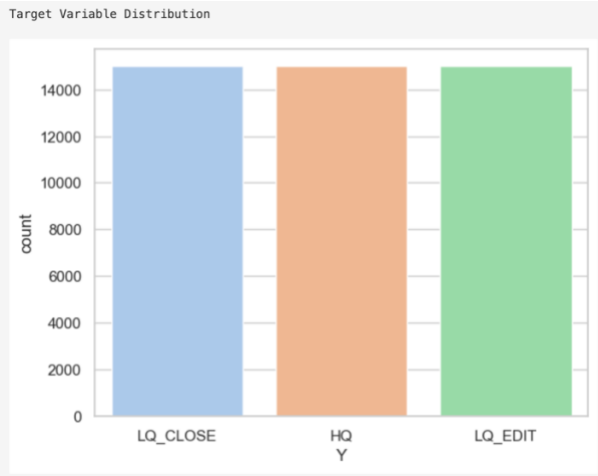


Figure 1: Target Variable Distribution

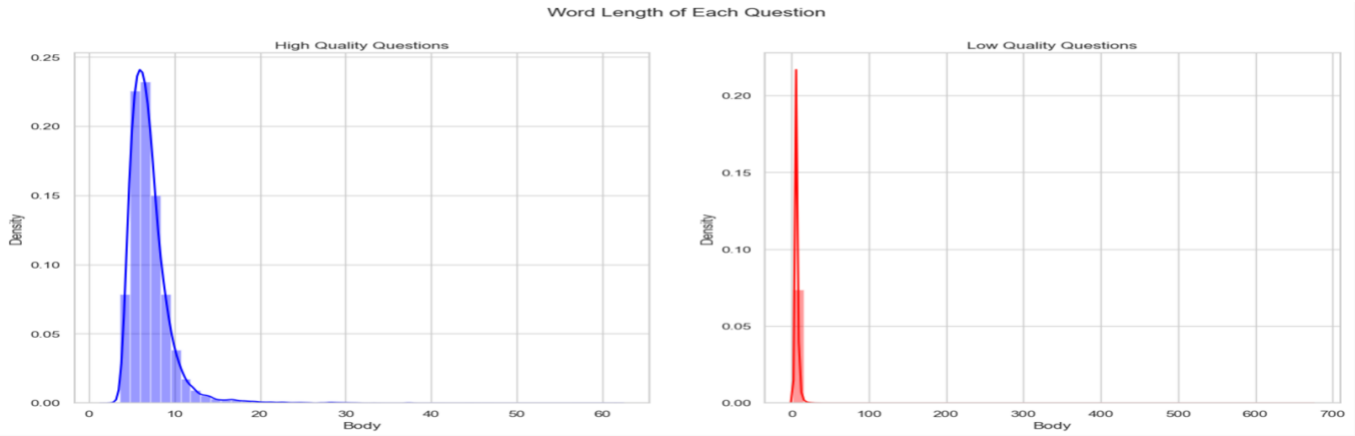


Figure 2: Density Chart for Word Length of Questions

4. Method

In this project, I divided my methodology into three main sections. The first section involved performing exploratory data analysis to identify the most common programming languages used in Stack Overflow questions. I also generated density and bar charts for visualization purposes. In the second section, I performed data preprocessing by dropping unnecessary columns, merging important columns, and removing HTML tags from the dataset. Additionally, I carried out text cleaning to prepare the data for modeling. The final section of the methodology involved model training, where I used a range of classification models such as Logistic Regression, Multinomial Naive Bayes, Random Forest, Decision Tree, K Neighbors Classifier, and Gradient Boosting Classifier. I also built deep-learning models using LSTM. By using a range of

classification models, I was able to identify the most effective method for predicting the quality of Stack Overflow questions. Overall, my methodology involved a rigorous process of data analysis, preprocessing, and model training to achieve accurate results.

4.1 Exploratory Data Analysis

In this project, I carried out an extensive exploratory data analysis to gain insights into the Stack Overflow dataset. The first step was to explore the tags column to identify the most used technologies in the Stack Overflow questions. I created a bar plot to understand the frequency distribution of these technologies and found that JavaScript and Python were the most used technologies. Next, I analyzed the average word length of high-quality versus low-quality questions using a density plot. The plot showed that the average word length for high-quality questions was mainly in the range of 0-20, which suggests that concise and to-the-point questions tend to be of high quality.

I also created an area plot to show the number of questions asked on Stack Overflow over the years. The plot revealed a decreasing trend in the number of questions asked on Stack Overflow over the years. I observed that in the year 2020, there were very few high-quality questions being asked, which could indicate a shift in the way developers are seeking help on Stack Overflow. Overall, my exploratory data analysis helped me gain valuable insights into the Stack Overflow dataset and provided a foundation for the subsequent stages of data preprocessing and model training.

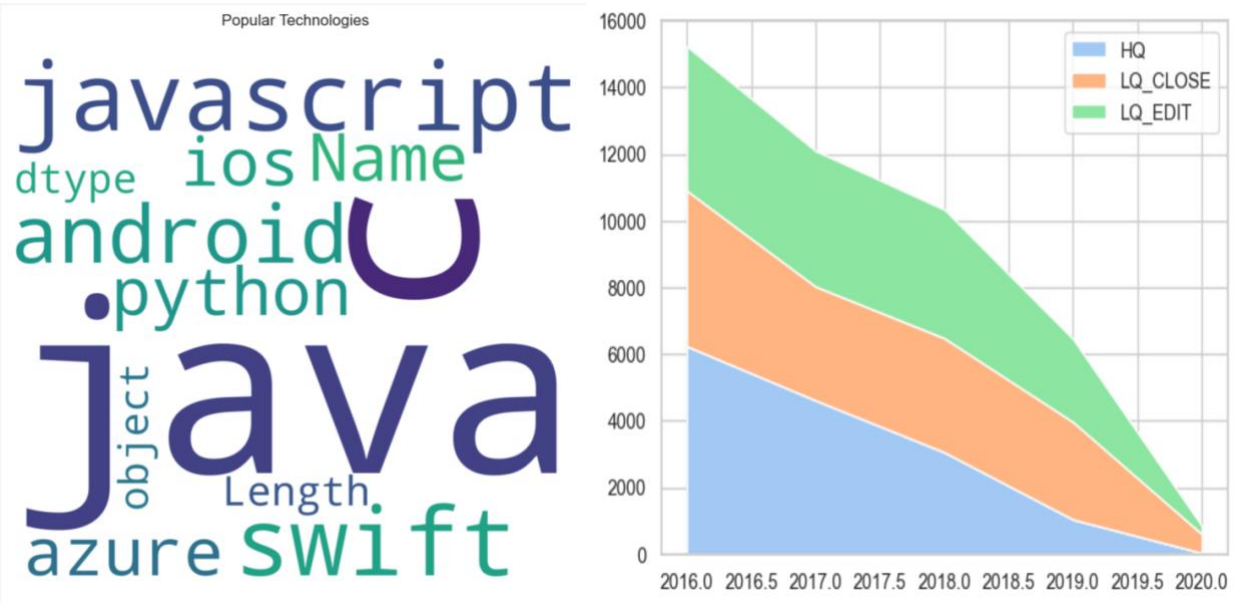


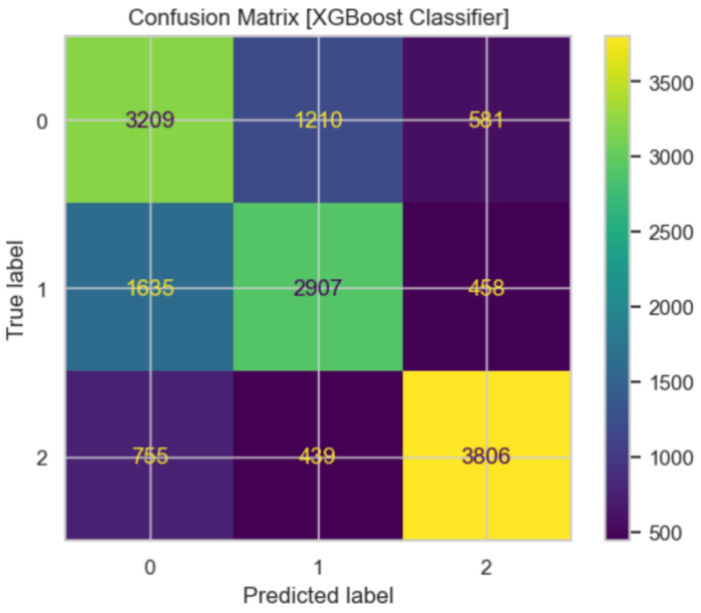
Figure 3: Area Plot

Figure 4: Displaying Popular technologies.

4.2 Data Pre-processing

Data preprocessing is an important step in any NLP task. To obtain cleaned data, the raw data collected through various means is passed through a pipeline of NLP preprocessing. This pipeline includes the following tasks:

- Dropping unnecessary columns.
- Label Encoding the Target Variable.
- Combining Title and Body into one column.
- Removing HTML tags from the text.



- Removing newlines and tabs/whitespaces/links/special characters/stopwords.
- Removing punctuations.
- Correcting misspelt words.
- Lemmatization/Stemming.

4.3 Model Training

4.3.1 Machine Learning Models

Logistic Regression is a popular linear classification model that works well with large datasets. Multinomial Naive Bayes is a probabilistic classifier that is commonly used for text classification tasks. Random Forest Classifier is an ensemble learning method that combines multiple decision trees to improve the accuracy of the model. KNN Classifier is a non-parametric classification model that determines the class of a new data point by finding the most similar training examples. XGBoost is a powerful gradient-boosting algorithm that is widely used for supervised learning tasks.

In the model training section, we experimented with various machine learning models to predict the quality of Stack Overflow questions. Logistic Regression, with a value of C=1, achieved the highest accuracy of 66.91%. Multinomial Naive Bayes achieved an accuracy of 64.31%, followed by XGBoost with an accuracy of 66.15%. However, Random Forest Classifier and KNN Classifier did not perform well, achieving accuracies of 63.92% and 53.94%, respectively. The Decision Tree Classifier had the lowest accuracy of 50.89%.

4.3.2 Deep Learning Model

For the deep learning model, I started by tokenizing the data. The model architecture consists of an input layer that takes in the tokenized sequences of questions with variable lengths. The next layer is an embedding layer with 2.5 million parameters which maps the input sequences to dense vectors of fixed size. The output of the embedding layer is fed into a bidirectional LSTM layer with 98,816 parameters. The bidirectional LSTM layer processes the input sequences in both forward and backward directions to capture the contextual information. The output of the first LSTM layer is then fed into a second bidirectional LSTM layer with the same number of parameters to further refine the contextual information. A dense layer with 32 units and a dropout layer is added to prevent overfitting. The final output layer has 3 units for the 3 classes: High Quality, Low Quality with Edits, and Low-Quality Closed. The model uses the Adam optimizer for efficient optimization of the model parameters. Overall, this model architecture is well-suited for sequence classification tasks and is expected to provide competitive results compared to traditional machine learning models.

Figure 5: Confusion Matrix of XGBoost classifier

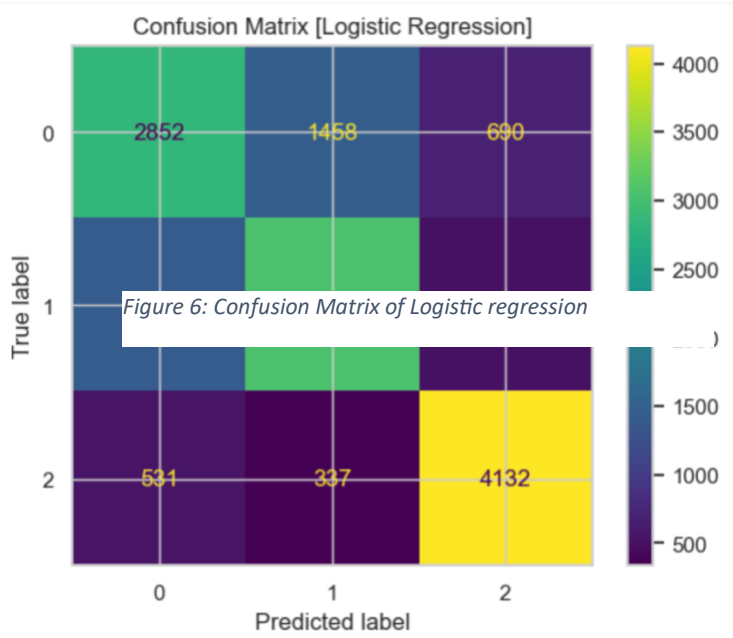


Figure 6: Confusion Matrix of Logistic regression

Hyper Parameters	Training AUC	Test AUC
EPOCH = 15, BATCH SIZE = 32	93%	83%
EPOCH = 5, BATCH SIZE = 64	95%	82%
EPOCH = 10, BATCH SIZE = 64	96%	81%

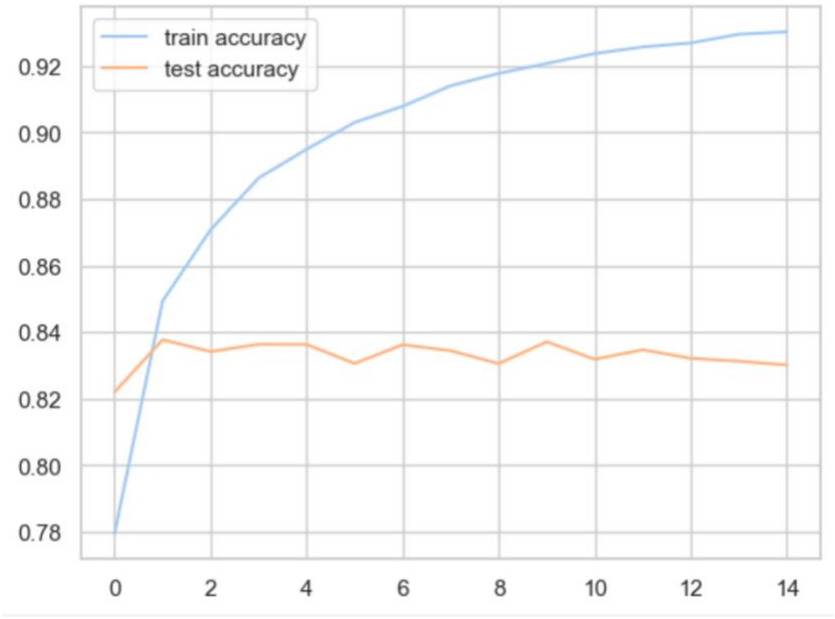


Figure 7: Accuracy Graph for Deep Learning Model

5. Conclusion

In conclusion, I have developed a model to predict the quality of Stack Overflow questions based on their content. I explored the data, performed data preprocessing, and trained various machine learning models such as Logistic Regression, Multinomial Naive Bayes, Random Forest Classifier, Decision Tree Classifier, K-Nearest Neighbors Classifier, and XGBoost Classifier. I also developed a deep learning model using LSTM and achieved the highest accuracy of 83%, although the model seemed too overfit as the training accuracy increased while the test accuracy only varied slightly.

This project is significant as Stack Overflow is a widely used platform for developers to get their queries resolved. The model can assist in identifying and categorizing low-quality questions and thus improve the overall quality of Stack Overflow. Furthermore, the study can help researchers understand the characteristics of high-quality questions and the factors that contribute to their success.

6. Future Work

There are several avenues for future work that could be explored in this project. One potential area of improvement would be to collect a larger dataset with more recent Stack Overflow questions. This could improve the performance of the machine learning and deep learning models, especially given the rapidly changing landscape of programming languages and technologies. Another potential area for improvement is to explore more advanced deep learning architectures, such as attention-based models or transformer networks. These models

have shown promising results in natural language processing tasks and could potentially improve the performance of the current LSTM-based model.

7. Acknowledgment

This project was completed for CS6120 under the instruction of Professor Uzair Ahmad.

References

[1] Dataset – https://www.kaggle.com/datasets/imoore/60k-stack-overflow-questions-with-quality-rate?datasetId=850380&sortBy=voteCount
[2] Python Notebooks – https://github.com/parv212/60K-Stack-Overflow-Classification
[3] Antoaneta Baltadzhieva, Grzegorz Chrupała, “Predicting the Quality of Questions on Stack overflow” Proceedings of Recent Advances in Natural Language Processing
[4] Antoaneta Baltadzhieva and Grzegorz Chrupala. 2015. “Predicting question quality in question answering forums”.