University *of* New Haven

DSCI-6003-03: Machine Learning

Final Project

Team Members: Shraddha Shrestha || Smit Patel || Mallikarjun Aitha

MSc Data Science

University of New Haven

Date: April 23, 2024

Table of Contents

1 Abstract

In the dynamic landscape of online knowledge-sharing platforms like Quora, the proliferation of redundant questions presents a significant challenge for users and contributors alike. With a monthly user base exceeding 100 million, Quora encounters numerous queries that share similar wording, leading to inefficiencies in information retrieval and provision. To address this challenge, we present a comprehensive approach centered on feature engineering and advanced modeling techniques for question similarity detection. Through an iterative process, we leverage both basic and advanced feature engineering techniques to enhance the performance of classification models. By incorporating engineered features such as token-based metrics, length-related attributes, and fuzzy matching scores, our models demonstrate improved capability in capturing nuanced similarities between question pairs. Moreover, visualizations play a crucial role in guiding feature selection and model tuning, ensuring the effectiveness of our approach. Overall, our study underscores the effectiveness of thoughtful feature selection and preprocessing in natural language processing tasks, offering insights that contribute to the enhancement of user experiences on knowledge-sharing platforms.

2 Introduction

Quora, a widely-used platform for knowledge-sharing, boasts a substantial monthly user base exceeding 100 million individuals. This vast community contributes to an extensive repository of questions spanning a diverse array of topics. However, the platform's popularity also gives rise to a notable challenge: the proliferation of queries that share strikingly similar wording. This redundancy within the question pool introduces inefficiencies, as users often struggle to locate optimal answers amidst a sea of duplicates, while writers face the burden of addressing repetitive inquiries.

In the realm of question similarity detection on Quora, our study undertakes a systematic approach encompassing various stages of model development and refinement. To ensure the efficacy of our methodology, we commence with an Initial Exploratory Data Analysis (EDA). This foundational step allows us to delve into the intricacies of the dataset, ensuring alignment with model assumptions and paving the way for effective preprocessing strategies.

Following the comprehensive EDA, our journey progresses to the implementation of a Baseline Random Forest classification model. This initial model serves as a benchmark, providing a baseline accuracy level against which subsequent iterations will be evaluated. By deploying a straightforward Random Forest approach without feature engineering, we establish a foundational understanding of the dataset's predictive capabilities.

With a solid foundation laid, we then pivot towards Feature Engineering Implementation, a pivotal phase aimed at enhancing the performance and efficacy of our predictive models. This stage involves a multifaceted approach, including the creation of new features, handling missing values, encoding categorical variables, and scaling numerical features. Through meticulous feature engineering, we endeavor to extract meaningful insights and augment the discriminatory power of our models.

Subsequent to feature engineering, our focus shifts towards Feature Analysis and Model Evaluation. Here, we scrutinize the engineered features and assess model performance based on the derived outputs. By iteratively refining our approach and optimizing feature selection, we strive to maximize accuracy and robustness in question similarity detection.

As we approach the culmination of our study, we embark on the final phase: Advanced Feature Engineering. This stage represents a culmination of our efforts, wherein we delve into sophisticated techniques to further refine the feature set. Special attention is devoted to nuances such as special characters ('$'), numeric patterns (0-9), and grammatical intricacies ('can't'), aiming to capture subtle variations in question semantics. Through this meticulous refinement process, we aim to conduct a comprehensive accuracy assessment, ensuring the readiness of our model for real-world deployment.

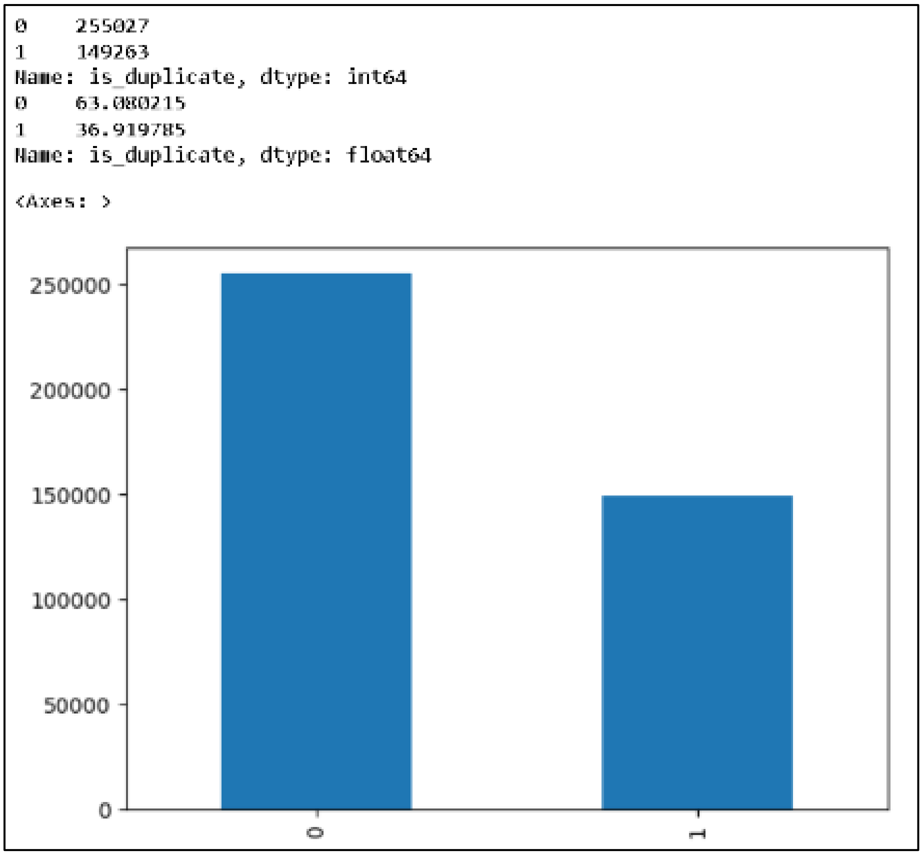
In the subsequent sections, we will delve into the intricacies of each stage, detailing the methodologies employed, the insights gained, and the implications for question similarity detection on Quora. Through this systematic approach, we endeavor to contribute to the advancement of machine learning techniques in enhancing the user experience and efficiency of knowledge-sharing platforms

3 Methods

Our methodology encompasses a comprehensive approach to understanding, enriching, and harnessing the potential of the dataset for effective question similarity detection on Quora. We outline a series of methodological steps, beginning with Exploratory Data Analysis (EDA) and extending to Feature Engineering and Advanced Feature Engineering.

i. Exploratory Data Analysis (EDA):

The initial phase of our methodology involves a thorough exploration of the dataset to uncover key insights and patterns. We start by examining the distribution of values in the 'is_duplicate' column, which indicates whether a question is a duplicate or not. By calculating the frequency of each unique value and visualizing it through a bar chart, we gain a holistic understanding of the dataset's composition.

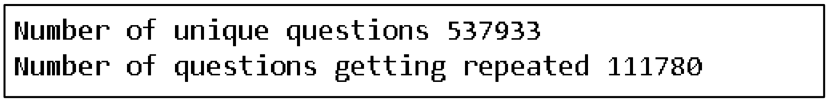


```
0    255027
1    149263
Name: is_duplicate, dtype: int64
0    63.080215
1    36.919785
Name: is_duplicate, dtype: float64

<Axes: >
```

Figure 1: Number of Unique Questions Vs Number of Similar Questions

Next, we determined that there are 537,933 unique questions in the dataset, with 111,780 questions being duplicates or repeated entries.

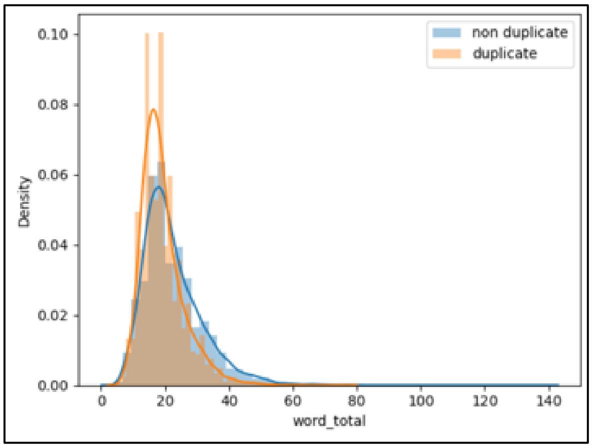


```
Number of unique questions 537933
Number of questions getting repeated 111780
```

Additionally, we assess the number of unique questions and identify duplicate entries, providing crucial insights into the prevalence of redundancy within the dataset. Furthermore, we analyze the distribution of unique values in the 'qid' column using a logarithmically scaled histogram, offering insights into the distribution of question identifiers.

ii. Feature Engineering:

Subsequently, we perform feature engineering to enrich the dataset and enhance the predictive power of our models. For this purpose, we sample 30,000 random instances and introduce seven new columns based on various properties of the text data. These include the lengths of the question strings, word counts in each question, the total number of common words between questions, the total number of unique words, and the ratio of common words to unique words. These engineered features are then visualized using graphs to gain deeper insights into their distributions and relationships.



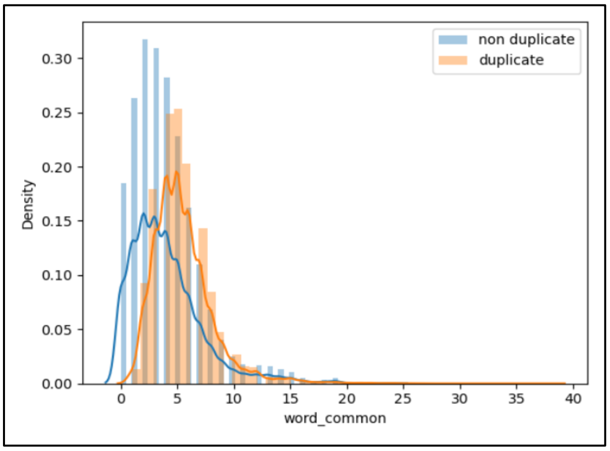


Figure 2: Number of Unique Words

Figure 3: Number of Common Words

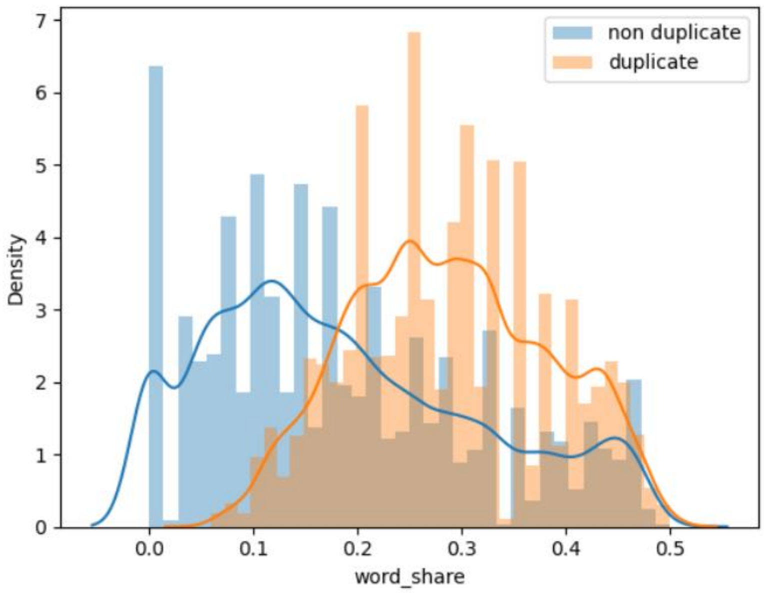


Figure 4: The ratio of Common Words to Unique Words

iii. Advanced Feature Engineering:

Advanced Feature Engineering represents a more intricate phase aimed at preprocessing and transforming text data, extracting meaningful features, and visualizing these features for better understanding. This stage involves several steps:

Symbol Conversion and Normalization: Symbols are converted to words (e.g., '%' to 'percent'), numeric values are converted to letters (e.g., 'billions' to 'b'), and words are normalized into their standard forms (e.g., "can't" to "can not").

Token-Based Feature Extraction: Pairs of text questions are processed to extract token-based features, including common words, stopwords, and tokens between question pairs. Ratios and binary indicators based on token intersections are computed.

Length-Related Feature Calculation: Various length-related features are computed for pairs of text strings, such as absolute length differences between questions and the average token length of both questions. The length of the longest common substring is normalized by the minimum question length.

Fuzzy Matching Score Calculation: Fuzzy matching algorithms are utilized to compute similarity scores between question pairs, including simple ratio, partial ratio, token sort ratio, and token set ratio.

Graphical Representation: Extracted features are visualized using graphical plots, leveraging tools like seaborn to create pairplots and other visualizations. Scatter plots and histograms are employed to explore relationships and distributions, providing further insights into the dataset's characteristics.

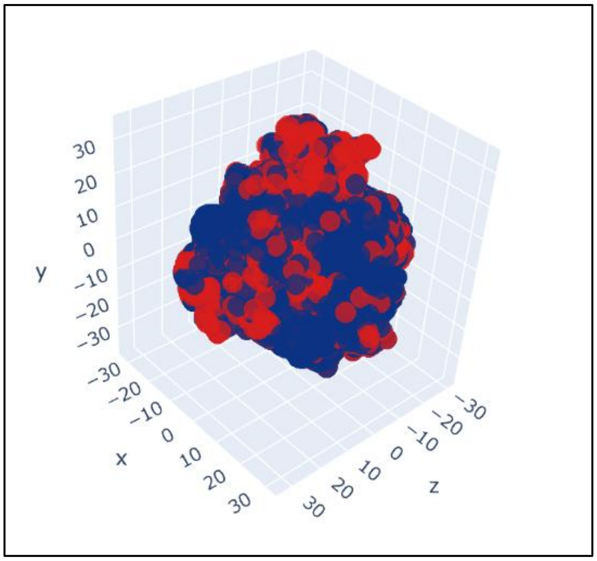


Figure 5: 3D Representation of Dimensionality Reduction

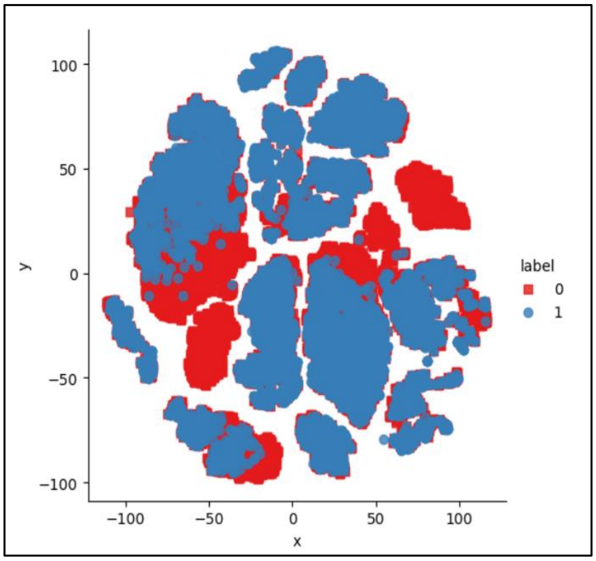


Figure 6: 2D Representation of Dimensionality Reduction

Through this iterative and systematic approach, we aim to unlock the full potential of the dataset and develop robust models for question similarity detection on Quora. By combining rigorous analysis with innovative feature engineering techniques, we strive to enhance the efficiency and accuracy of knowledge-sharing platforms, ultimately enriching the user experience for millions of users worldwide.

4 Results and Accuracy

Initially, we executed the Random Forest classifier without any feature engineering. The obtained results were as follows:

- Random Forest Classification achieved an accuracy of approximately 73.32%.
- Extreme Gradient Boosting achieved an accuracy of approximately 72.22%.

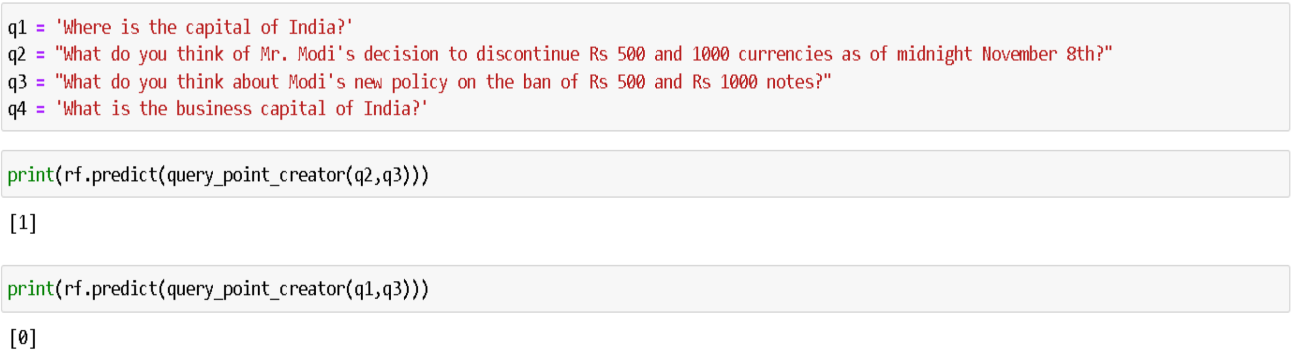
Upon integrating feature engineering into the Random Forest classifier, we observed improvements in accuracy:

- Random Forest Classification achieved an accuracy of approximately 76.78%.
- Extreme Gradient Boosting achieved an accuracy of approximately 76.61%.

Finally, by incorporating advanced feature engineering techniques into the Random Forest classifier, we attained further enhancements in accuracy:

- Random Forest Classification achieved an accuracy of approximately 78.5%.
- Extreme Gradient Boosting achieved an accuracy of approximately 79.46%.

In our evaluation of model performance, we conducted a test using four questions, with two pairs having different questions yet sharing the same meaning. The objective was to assess whether the model could accurately discern between questions with similar intents. A result of [1] indicates that the questions have the same meaning, while [0] denotes differing meanings. For instance, questions q2 and q3 are deemed identical, resulting in [1], whereas q1 and q3 are discerned as different, resulting in [0].



```
q1 = 'Where is the capital of India?'
q2 = "What do you think of Mr. Modi's decision to discontinue Rs 500 and 1000 currencies as of midnight November 8th?"
q3 = "What do you think about Modi's new policy on the ban of Rs 500 and Rs 1000 notes?"
q4 = 'What is the business capital of India?'
```

```
print(rf.predict(query_point_creator(q2,q3)))
```

```
[1]
```

```
print(rf.predict(query_point_creator(q1,q3)))
```

```
[0]
```

Figure 7: Testing the Model

For example:

- Question 2 (q2) and Question 3 (q3) are identical, resulting in [1].
- Question 1 (q1) and Question 3 (q3) are different, resulting in [0].

These results underscore the effectiveness of feature engineering, especially advanced techniques, in augmenting the performance of the Random Forest classifier and Extreme Gradient Boosting model. The incremental improvements in accuracy demonstrate the importance of thoughtful feature selection and engineering in refining models for question similarity detection, ultimately enhancing the utility and reliability of knowledge-sharing platforms like Quora.

5 Conclusion

Through a systematic and iterative process of feature engineering, encompassing both basic and advanced techniques, we observed a significant enhancement in the performance of our classification models. The incorporation of engineered features, such as token-based metrics, length-related attributes, and fuzzy matching scores, notably enriched the models' capability to discern nuanced similarities between question pairs.

Moreover, visualizations played a pivotal role in elucidating the dataset's characteristics, facilitating informed feature selection, and guiding model tuning. By leveraging visual insights, we were able to make informed decisions regarding feature importance and model optimization.

Furthermore, our study underscores the importance of adaptability and continuous refinement in the realm of natural language processing. As language evolves and user behaviors shift, the efficacy of classification models heavily depends on their ability to adapt to new patterns and nuances. By embracing an iterative approach to feature engineering and model optimization, we remain agile in responding to changing dynamics within the dataset and user interactions. This adaptability ensures that our models remain relevant and effective in accurately capturing question similarities, thereby contributing to the seamless functioning of knowledge-sharing platforms and enhancing user experiences over time.

Overall, the synergistic combination of feature engineering and advanced modeling methodologies resulted in substantial improvements in accuracy. This underscores the efficacy of meticulous feature selection and preprocessing in bolstering the effectiveness of natural language processing tasks, particularly question similarity detection. Moving forward, the continued refinement and innovation in feature engineering techniques hold promise for further enhancing the efficiency and efficacy of knowledge-sharing platforms.