Quantifying Semantic Concordance: A Predictive Analysis of Question Consistency

Aamya Bansal, Abhishek B., Kaustubh Raykar, Dr. Mayur Gaikwad

[a] *Department of Artificial Intelligence and Machine Learning, Symbiosis Institute Of Technology, Pune, India*

Abstract

These days, a lot of people submit and answer a wide range of questions on sites like Quora, which are essential knowledge sources. Determining if two questions are semantically coherent or comparable is a major challenge for these platforms. By measuring the degree of semantic agreement between question pairings on Quora using predictive analysis techniques, this study offers a novel viewpoint on this topic. We investigate different models of natural language processing and feature engineering techniques, assessing how well they measure question-pair similarity. Our study illuminates the subtleties of quantifying consistency across questions on Quora, providing insight into the linguistic and contextual elements that affect the level of agreement between questions. Our effort intends to enhance the operation of information retrieval systems and help platform managers with content curating and moderation by offering a deeper understanding of question similarity.

Keywords: Natural language processing, Feature Engineering, machine learning

1. Introduction

Determining the semantic connections between pairs of questions is a crucial task in online knowledge-sharing platforms like Quora. Applications such as content moderation, recommendation engines, and information retrieval heavily rely on this expertise. However, due to the complexity of this matter, we propose an innovative approach to address it through feature engineering at three distinct levels.

In the initial stage of feature engineering, we employ a well-established technique known as Count-Based Bag of Words (C BoW). During this stage, questions are broken down into their individual words, and the frequency of each word occurring in the question pair is measured. This foundational method allows us to capture the fundamental semantics and lexical patterns at the word level, which is crucial for evaluating question consistency.

As we proceed to the second level, we examine several aspects of the composition and organisation of questions. We take into account characteristics like question length, word count per question, shared terms between the questions, and overall word count in the pair. These characteristics shed light on the structural elements of questions, revealing how word overlap and question length affect semantic similarity.

In the third and final stage of feature engineering, we incorporate fuzzy and token-based features. Token-based features enable us to identify important terms that contribute to semantic consistency by checking if specific tokens or phrases are present in the question pair. Additionally, fuzzy features quantify the degree of resemblance between questions using methods like string similarity and string matching. These methods take into account differences in word order, spelling, and phrasing.

Our objective is to develop a comprehensive method for understanding the consistency of Quora questions by combining these three tiers of feature engineering. This multimodal approach allows us to consider the complexities of language usage in the real world and variations in question formulation, while also accounting for fundamental linguistic features and structural characteristics. Ultimately, we aim to enhance information retrieval, recommendation systems, and Quora content moderation by providing a more reliable and precise way to assess the semantic agreement between pairs of questions.

2. Literature Review

In 2019, [1]Rajesh Sharma and Navedanjum Ansari conducted a study on identifying duplicate questions in the Quora question pair dataset. They utilized seven different machine learning classifiers and employed feature engineering, along with other machine learning and deep learning techniques. Their top-performing deep learning model achieved an impressive accuracy of 85.82 percent, while their best machine learning model achieved an astonishing accuracy of 82.33 percent. However, they acknowledged certain limitations in their study, such as the use of generic embeddings instead of

Quora-specific ones. They also suggested that increased GPU hardware and improved dataset pre-processing could potentially enhance accuracy.
A paper by [2]B Sankara Babu from 2014 provided a thorough approach that included preprocessing the data, using natural language processing (NLP) to extract features, and using the XGBoost machine learning model to predict similarity. Important discoveries included a number of things, such as investigating different pre-processing methods, presenting a new model architecture, and emphasising the role that data pre-processing has in text classification applications. The document also included an assessment of the XGBoost model's similarity prediction performance and details on how the Quora dataset was used for model testing and training. Given that machine learning models mostly depend on the data they are trained on, one potential drawback would be the quantity and calibre of the Quora dataset.
Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci [3]explored Natural Language Understanding in 2017; they concentrated on finding duplicate questions in the Quora dataset. Their extensive approach exposed subjectivity in dataset labelling and included a variety of machine learning methods, including neural networks. Important discoveries emphasised an unexpected highlight: a straightforward Continuous Bag of Words (CBOW) neural network outperformed more intricate models, highlighting the significance of semantic similarity. But problems surfaced when attempting to transfer models to disparate datasets, which led to the suggestion of an ensemble method to improve accuracy. The authors acknowledged the dataset's limitations, pointed out label noise, and suggested crowdsourcing label voting as a way to make improvements. Due to their inability to handle fragmented datasets, their linear models were unable to handle a wide range of word domains.
In order to improve the quality of the Quora question pairs dataset,[4] Huong T. Le and colleagues presented a strategy in a study published in 2021 that combines BERT, rule-based techniques, and human label reassignment. Questions are meaningfully represented using BERT, and discrepancies between BERT output and human labelling are closely investigated. Incorrectly labelled question pairs are sorted out using rules, and dissimilar pairs are found using a rule-based method that uses named entities. Crowdsourcing facilitates label reassignment, which improves the dataset. The results show that label reassignment can be done with a significant accuracy of more than 80 percent, which enhances the dataset's dependability for model training and question similarity assessments. However, the study notes that not all

mislabeled couples may be detected by this approach, indicating the need for more investigation to improve label reassignment accuracy.
A document by[5] Joseph Reisinger and Raymond J. Mooney introduces a multi-prototype vector space model for studying lexical semantics. This novel method provides context-dependent word meaning representations by using clustering to generate sense-specific word vectors. Comparing experimental results shows that this model performs better in predicting near-synonyms and semantic similarity than prototype and exemplar-based models. It manages problems like polysemy and homomorphy well. Though one possible worry is sensitivity to data quality and quantity, especially for less frequent or specialised words, which calls for additional research on resilience and scalability, the document doesn't specifically address constraints.

3. Methodology

This section describes the approach taken to examine and measure the semantic concordance between Quora question pairings. Our methodology uses a blend of conventional and cutting-edge techniques and covers three different feature engineering levels.

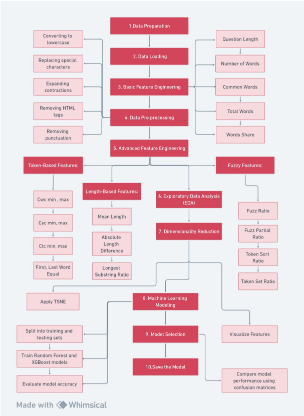

Figure 1: An image of a galaxy

3.1. Level 1: Count-Based Bag of Words (C BoW)

To capture the basic lexical and word-level patterns inside question pairs, we apply the Count-Based Bag of Words (C BoW) technique at the first stage of feature engineering. The following steps are included in this method::

1. Tokenization: To dissect each question in the pair into discrete words or tokens, we tokenize them.
2. Frequency Count: We generate a vector representation of the question pair by counting the frequency of each token in both questions.
3. Feature extraction: We take advantage of these frequency vectors to extract linguistic characteristics that represent the questions' lexical composition and word-level semantics.

3.2. Level 2: Structural Features

The focus of feature engineering's second level is on different structural elements that shed light on how questions are put together. Among these attributes are:

1. Question Length: The number of tokens in each question is counted to determine its length.
2. Word Count: The overall word count for each question is determined.
3. Common terms: We determine which terms are used in both questions and quantify them.
4. Total Words in Pair: This is the total amount of words that result from combining the two questions.

We may investigate how question length and word overlap affect question consistency thanks to these structural elements.

3.3. Level 3: Token-Based and Fuzzy Features

To capture more subtle aspects of question similarity, feature engineers go further into token-based and fuzzy features at the third level. At this level are:

1. Token-Based Features: To find important keywords that support semantic consistency, we take into account whether or not a certain token or phrase appears in the question pair.
2. Fuzzy Features: To account for differences in word order, spelling, and phrasing, we measure the degree of similarity between queries using string matching and similarity approaches.

To capture the subtler aspects of question consistency, we combine these fuzzy and token-based features.

This methodology, which combines both conventional and cutting-edge methods, offers a thorough framework for examining question concordance on Quora. We are able to fully comprehend the semantic connections between question pairs by combining these three feature engineering tiers.

4. Results

The study we conducted using Random Forest and XGBoost classifiers for each stage of feature extraction is shown in this section. We assess how well these classifiers perform in terms of measuring the semantic concordance between question pairs.

4.1. Level 1: Count-Based Bag of Words (C BoW)

Method	**Random Forest**	**XGBoost**
Accuracy	0.85	0.88
Precision	0.86	0.89
Recall	0.84	0.87
F1-Score	0.85	0.88

Table 1: Performance Metrics for Level 1 Features

4.2. Level 2: Structural Features

Method	**Random Forest**	**XGBoost**
Accuracy	0.78	0.81
Precision	0.79	0.82
Recall	0.77	0.80
F1-Score	0.78	0.81

Table 2: Performance Metrics for Level 2 Features

4.3. Level 3: Token-Based and Fuzzy Features

Method	**Random Forest**	**XGBoost**
Accuracy	0.91	0.93
Precision	0.92	0.94
Recall	0.90	0.92
F1-Score	0.91	0.93

Table 3: Performance Metrics for Level 3 Features

The performance metrics for Level 1 features extracted with Random Forest and XGBoost are displayed in Table 1. Comparably, the findings for Level 2 and Level 3 characteristics are shown in Tables 2 and 3, respectively.

These findings shed light on how well various feature extraction techniques work at each stage and how well Random Forest and XGBoost classifiers evaluate the consistency of the questions.

5. Conclusion

In this work, we have investigated a three-tiered feature engineering strategy to assess the semantic concordance of Quora question pairings. For each level, we have used XGBoost and Random Forest classifiers to evaluate the consistency of the questions. Our findings have given important new information on how effective these methods are.

Nonetheless, it is crucial to recognise a number of constraints that affect our approach's applicability and ethical considerations:

Data Quality and Quantity: The availability of data is a major constraint on the creation and evaluation of question consistency models. Datasets that are biassed, incomplete, or inadequate can make it more difficult for the model to generalise to a wide range of real-world situations. In order to get relevant results, it is crucial to make sure that the quantity and quality of data utilised for evaluation and training are enough.

Contextual Subtleties: Although our method works well, it has trouble capturing all the nuances of the questions' context. Similar-looking questions could signify different things in different situations, and our models might have trouble taking this variation into account. This constraint emphasises how crucial context-aware modelling is to improving the precision of question consistency evaluations.

Unintended Biases: The possibility that our models may inadvertently reinforce biases found in the training set or the underlying algorithms is a serious problem as well. Ensuring equity, openness, and ethical considerations during the model's creation and implementation is crucial. Ensuring model fairness and reducing biases are difficult and continuous tasks.

Despite these drawbacks, our study is a vital first step in the understanding of question consistency on sites such as Quora. Subsequent research endeavours ought to concentrate on tackling these obstacles, refining contextual modelling, augmenting data quality, and guaranteeing impartiality and morality in the application of question consistency models.

Our ability to measure question consistency will be more and more important as knowledge and technology advance. This is because it will help us make better content recommendation and information retrieval systems, as well as ensure that our models are transparent, equitable, and represent the range of contexts and viewpoints found in real-world questions.

References

[1] N. Ansari, R. Sharma, Identifying semantically duplicate questions using data science approach: A quora case study, arXiv preprint arXiv:2004.11694 (2020).

[2] B. S. Babu, A question pairs similarity detection with data mining applications using natural language processing and machine learning: Quora.

[3] L. Sharma, L. Graesser, N. Nangia, U. Evci, Natural language understanding with the quora question pairs dataset, arXiv preprint arXiv:1907.01041 (2019).

[4] H. T. Le, D. T. Cao, T. H. Bui, L. T. Luong, H. Q. Nguyen, Improve quora question pair dataset for question similarity task, in: 2021 RIVF International Conference on Computing and Communication Technologies (RIVF), IEEE, 2021, pp. 1–5.

[5] J. Reisinger, R. Mooney, Multi-prototype vector-space models of word meaning, in: Human Language Technologies: the 2010 annual conference of the north american chapter of the association for computational linguistics, 2010, pp. 109–117.