Quora Insincere Questions Classification

Jialing Li
jl9716@nyu.edu

Erqian Wang
ew1708@nyu.edu

1 Introduction

The increasing volume of user-generated content on platforms like Quora poses significant challenges and opportunities for content moderation. Questions, the fundamental units of these platforms, vary widely in quality and intent. Among them, distinguishing between sincere and insincere questions is crucial for maintaining the integrity of discussions and information exchange. Insincere questions, which may be misleading, based on false premises, or intended to provoke rather than seek genuine answers, can undermine the quality of discourse. This project leverages natural language processing (NLP) techniques to identify linguistic and stylistic features that characterize sincere and insincere questions, aiming to enhance automatic moderation tools and contribute to more effective management of online content.

2 Related Works

Recent research in the domain of question classification has explored the interplay between linguistic features and underlying intentions, particularly concerning the detection of insincere questions. Rie et al. (2012)[1], and Elena et al. (2020)[2] have investigated the significance of lexical diversity, as measured by Type-Token Ratio and Guiraud's Index, in text classification. Complementarily, studies by Julian, Manar and Herenia (2023)[3] have highlighted the importance of readability metrics in identifying linguistic clarity as a distinguishing factor.

Methodologically, researchers have adopted diverse approaches, from statistical analyses to machine learning algorithms like BERT and distilBERT. Data collection typically involves scraping user-generated questions from online platforms, with manual or automated annotation for sincerity classification.

Our research extends previous studies by employing a multifaceted analytical framework that combines Representative N-grams Selection, Latent Dirichlet Allocation (LDA) and readability metrics to explore the thematic and lexical structure of user-generated questions. After establishing an

Preprint. Under review.

understanding of the textual composition, we classify questions as sincere or insincere using logistic regression models for their transparency and computational efficiency. To enhance interpretability, SHAP (SHapley Additive exPlanations) is integrated to elucidate how specific linguistic features influence these classifications. We believe a combination of these methods can enhance the understanding of linguistic features and latent topics influencing classification decisions.

3 Theory and Hypotheses

3.1 Framework

This research operates on the premise that linguistic features, such as lexical diversity and readability, are indicative of the underlying intentions in user-generated questions. The distinct nature of insincere questions is hypothesized to manifest through specific patterns in their textual composition.

3.2 Hypotheses

Based on this premise, we test the following hypotheses.

3.2.1 Lexical Diversity Hypotheses

Insincere questions will demonstrate different levels of lexical diversity (measured by Type-Token Ratio and Guiraud's Index) compared to sincere questions, reflecting variation in their linguistic complexity.

3.2.2 Readability

There will be significant differences in readability scores (measured by Flesch Reading Ease and Dale-Chall Score) between sincere and insincere questions, potentially due to the differing complexity and clarity of language used.

4 Data and Methods

4.1 Data

This project utilizes a dataset from Quora comprising over 1 million user-submitted questions categorized into two types: sincere and insincere. An insincere question is one that is intended to make a statement rather than seek helpful answers. It can be identified by a non-neutral or exaggerated tone, disparaging or inflammatory content, lack of grounding in reality, or use of shocking sexual content.

4.2 Representative N-grams Selection

To identify the most representative unigrams and bigrams for sincere and insincere questions, the Term Frequency-Inverse Document Frequency (TF-IDF) method is employed. This statistical measure evaluates how relevant a word is to a document in a collection of documents, which in this case is the set of questions. For each type of question, we compute the TF-IDF score for every term. We then rank the terms based on their TF-IDF scores. The top 5 unigrams and bigrams are selected as most representative of sincere and insincere questions.

4.3 Topic Modeling

Latent Dirichlet Allocation (LDA) is used for topic modeling to discover abstract topics within the questions. The model is applied separately to the corpora of sincere and insincere questions to uncover the underlying themes that pervade each category. The process involves:

- Dictionary and Corpus Creation: A dictionary of all unique tokens is created, and each document (question) is converted into a bag-of-words format.
- Model Training: The LDA model is trained on the bag-of-words corpus, specifying the number of topics (set to 5). The model iteratively learns to assign topics to each document and words to each topic, based on their distribution across the entire corpus.
- Topic and Word Distribution: For each topic, the model provides a distribution over the vocabulary, highlighting words that are most likely to appear in that topic. The top 5 words from each topic are extracted to represent the theme of the topic.

4.4 Logistic Regression and SHAP

Logistic regression, in conjunction with k-fold cross-validation and TF-IDF vectorization, served as the backbone of the model training and hyperparameter tuning.

- Data Preprocessing: Stop words are removed and all letters are converted into lowercase. After being tokenized, the corpus is transformed into numerical representations through TF-IDF vectorization and Count Vectorization. The entire corpus is then separated into training and testing sets.
- Model Training: Logistic regression, integrated with k-fold cross-validation (k=5), is trained on the TF-IDF-transformed training data. During training, the model performance is assessed using out-of-fold predictions iteratively to refine the predictions.
- Model Prediction: Model predictions are generated by applying the trained model to the test set, determining whether each question is classified as sincere or insincere.

Insights into the contributions of individual features to model predictions are gained by employing SHAP. Visualization techniques, such as beeswarm plots and force plots, are employed to elucidate feature importance and provide detailed explanations of individual predictions.

5 Results and Discussion

5.1 Hypothesis Testing

The Mann-Whitney U tests conducted to assess the differences between sincere and insincere questions in terms of Type-Token Ratio (TTR), Guiraud's Index, Flesch Reading Ease (FRE), and Dale-Chall (DC) scores yielded statistically significant results (p-value ≈ 0) for all metrics tested. These findings support our hypotheses regarding linguistic distinctions between sincere and insincere questions:

- **Lexical Diversity:** Insincere questions demonstrated a lower TTR (0.936) compared to sincere questions (0.961), contradicting the initial expectation of higher lexical diversity. However, the Guiraud Index was higher for insincere questions (3.93) than for sincere ones (3.47), suggesting a nuanced aspect of lexical diversity where insincere questions employ a wider variety of words, but perhaps in a more repetitive manner.

- **Readability:** The results also indicated a significant difference in readability scores, with insincere questions having a lower FRE score (70.3) compared to sincere questions (74.9), suggesting that insincere content may be deliberately complex or obfuscated. In contrast, the DC score was higher for insincere questions (9.21) than for sincere questions (8.99), aligning with the hypothesis that insincere questions might use language that is less common or more difficult, aligning with their potentially deceptive nature.

5.2 Representative Unigrams and Bigrams

The top 5 representative unigrams and bigrams of each type of questions are listed below.

Unigram (Insincere)	**Unigram (Sincere)**	**Bigram (Insincere)**	**Bigram (Sincere)**
people	sexmates	donald trump	severity bracket
women	sheer	black people	sew sons
trump	witted	white people	segregation black
like	thrill	united states	schools burned
men	proterrorist	trump supporters	say better

Table 1: Top 5 Unigrams and Bigrams for Sincere and Insincere Questions

- Top unigrams in insincere questions included terms like "women", "trump", and "men", reflecting potentially controversial or provocative topics that may be employed to stir discussions or debates.

- Top unigrams in sincere questions revealed a surprising set of terms such as "sexmates", "sheer", "witted", and "thrill", which did not seem to align clearly with typical informational or earnest inquiries, suggesting possible misclassification or nuances not captured by the labels.

- Top bigrams in insincere questions such as "donald trump", "black people", and "white people" indicate a focus on socially and politically charged subjects, likely to incite strong reactions or engage specific audience sentiments.

- Top bigrams in sincere questions like "severity bracket", "sew sons", and "segregation black" showed a mixture of technical, familial, and socially relevant phrases, which may indicate deeper, less sensational content compared to insincere queries.

5.3 Topic Modeling

	Sincere Questions	**Insincere Questions**
Topic 1	would, like, people, life, someone	women, men, white, people, black
Topic 2	get, best, good, job, year	people, jews, indians, hate, years
Topic 3	best, world, make, learn, book	would, hillary, clinton, earth, obama
Topic 4	best, way, whats, work, quora	people, quora, trump, think, many
Topic 5	india, would, us, people, country	trump, muslims, us, india, people

Table 2: Top Topics from Sincere and Insincere Questions

5.3.1 Topic from Sincere Questions

The topics derived from sincere questions suggest a focus on personal improvement, career advice, and general inquiries about life and learning.

1. **General Life Advice:** The first topic includes terms suggesting discussions centered around hypothetical scenarios or personal life choices.

2. **Career and Skills:** The second topic implies a focus on career advice, job opportunities, and yearly evaluations or achievements.

3. **Educational Content:** The third topic indicates an interest in learning, reading, and understanding global perspectives or contexts.

4. **Practical Guidance:** The fourth topic reflects questions seeking the best methods or practices in various aspects of life, including using platforms like Quora.

5. **Cultural and National Discussions:** The fifth topic highlights discussions that are more nationally or culturally specific, often reflecting broader societal issues.

The representative words and bigrams further underline the thematic distinctions, with insincere questions gravitating towards hot-button issues and potentially manipulative language, whereas sincere questions, though varied, tend to engage with a broader and possibly more constructive range of topics.

5.3.2 Topics from Insincere Questions

In contrast, the topics from insincere questions largely revolve around controversial, polarizing, or provocative themes.

1. **Gender and Race:** The first topic emphasizes discussions on gender and racial dynamics, often in a controversial or confrontational context.
2. **Cultural and Religious Bias:** The second topic suggests a focus on cultural, religious, or racial prejudices, possibly reflecting contentious or biased viewpoints.
3. **Political Figures and Theories:** The third topic includes discussions about political figures and conspiracy theories, indicating a politically charged dialogue.
4. **Social Media Discourse:** The fourth topic reflects meta-discussions about the platform itself and political opinions, indicating a use of social media for spreading specific narratives or opinions.
5. **International and Political Issues:** The fifth topic points to international politics and religious issues, often involving heated debates or polarized opinions.

The topic modeling results clearly delineate the nature of sincere and insincere questions on Quora. Sincere questions tend to be more oriented towards self-improvement, educational content, and practical advice. These are typically constructive and aimed at personal or communal growth. In contrast, insincere questions are characterized by their focus on controversial topics, often touching on sensitive issues like race, religion, and politics. Such questions may be crafted to provoke, mislead, or stir controversy rather than to seek or spread knowledge.

These distinctions not only provide insights into the motivations behind different types of questions but also help in developing targeted strategies for content moderation and community engagement on social platforms. The findings underscore the importance of context and thematic content in distinguishing between sincere and insincere discourse, offering valuable cues for automated systems designed to maintain the quality and integrity of user interactions online.

5.4 Logistic Regression and SHAP Analysis

The logistic regression model demonstrated strong performance on the test dataset, achieving an F1 score of 0.76 and an accuracy of 0.84. This suggests that the model effectively balances precision and recall, while it correctly predicts the outcomes for a substantial majority of the cases.

5.4.1 Model Global Interpretability

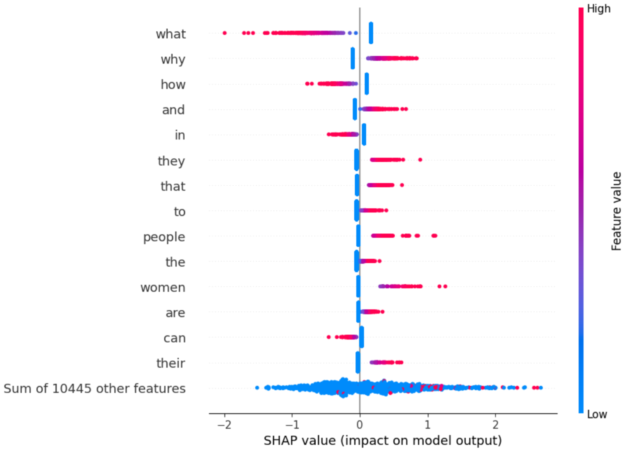

Figure 1: Feature Importance

In this analysis, we employ feature importance plot from SHAP to to quantify the impact each feature has on the model's predictions.

In analyzing the classification of questions as sincere or insincere based on SHAP values, distinct patterns emerge regarding the use of specific words. Words such as "what", "how", "in", and "can" typically contribute to classifications of insincerity. These words often frame questions that are direct, perhaps demanding specific information or challenging the respondent in a way that might be perceived as confrontational or aggressive. For example, questions like "What makes you think that is true?" or "How can you justify your actions?" use these words to demand justification or probe in a manner that may be seen as provocative.

Conversely, words like "why", "and", "they", "that", "to", "people", "women", and "their" tend to indicate sincerity. These words are more frequently used in questions that aim to explore deeper understanding or involve relational and societal contexts. Questions such as "Why do people believe this?" or "What can we learn from their experiences?" show a genuine curiosity and a desire to engage with broader issues or group dynamics. This suggests that sincere questions are characterized by their explorative nature and inclusivity, reflecting an intent to connect and understand rather than confront.

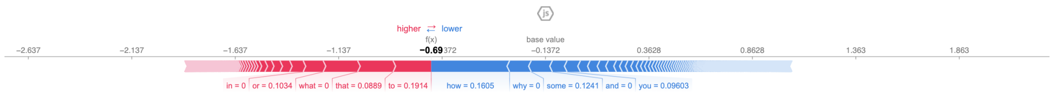

Figure 2: Force Plot for Randomly Selected Instance

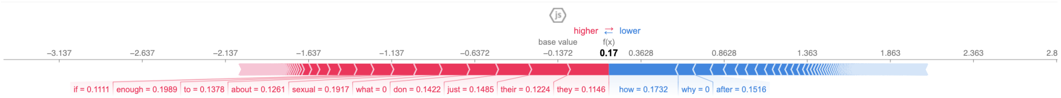

Figure 3: Force Plot for Randomly Selected Instance

5.4.2 Model Local Interpretability

Local interpretability, as shown in some of the randomly selected SHAP force plots for individual questions, provides detailed insights into how these global trends play out in specific instances. For example:

- Insincere Classification: In figure 2, the original question is "Is it bad when explaining how you solved some bug, telling who introduced that bug to the code, not on the purpose of blaming, but to give context to when or how?". Words like "how" and "why" push the prediction towards insincerity. Despite the question's context aiming to clarify rather than blame, the presence of these words aligns with the global trend where such phrasings prompt an insincere label due to their demanding nature.
- Sincere Classification: Conversely, in the question "How can I gather enough courage to talk to a girl after hearing/reading hundreds of stories about how they just press charges of sexual harassment if things don't go their way?"(Figure 3), the words "they", "their", and personal concerns indicated by "talk to a girl" counterbalance the usual insincerity associated with "how", leading to a sincere classification. This reflects the global observation that relational words contribute to sincerity.

6 Conclusion

In conclusion, our analysis of lexical complexity, readability, representative unigrams and bigrams, topic modeling, as well as SHAP analysis on the trained logistic model sheds light on the distinct linguistic features of sincere and insincere questions within online platforms. We demonstrate that insincere questions exhibited a lower Type-Token Ratio (TTR) but a higher Guiraud Index, indicating a nuanced aspect of lexical diversity where insincere questions employ a wider variety of words, albeit potentially in a more repetitive manner. Readability scores reveal that insincere questions tend to be deliberately complex or obfuscated.

Moreover, the topic modeling unveiled intriguing insights. Insincere questions often center around controversial or provocative topics, such as gender and race, aiming to incite strong reactions. Conversely, sincere questions cover a broad spectrum of topics including personal improvement, career advice, and general inquiries about life and learning, reflecting a genuine curiosity and intent to connect.

The logistic regression model and SHAP analysis further elucidate the linguistic patterns distinguishing sincere from insincere questions. By leveraging feature importance plots for global interpretability, we discerned distinct trends in the impact of specific words on question classifications. For instance, words such as "what", "how", "in", and "can" were found to contribute to insincere classifications, indicative of a potentially confrontational or demanding tone. Conversely, words like "why", "and", "they", "that", "to", "people", "women", and "their" were associated with sincerity, reflecting a genuine intent to explore deeper understanding or relational contexts. Moreover, we utilized force plots to explore local interpretability, finding that global and local interpretability align with each other. This comprehensive approach provides valuable insights into the underlying motivations and intentions behind the language used in online communication.

Moving forward, further research could delve deeper into the contextual nuances of language, considering factors such as cultural differences and evolving linguistic trends. Additionally, exploring the application of machine learning techniques in identifying and addressing insincere communication online could enhance trust and credibility in digital interactions.

Our contribution lies in bridging the gap between linguistic analysis and machine learning, offering a comprehensive understanding of language dynamics in online communication.

References

[1] Rie Koizumi and Yo In’nami. Withdrawn: Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, 40:522–532, 12 2012.

[2] Elena Zakharova and Olga Savina. Lexical diversity measures’ review and classification. *Tyumen State University Herald. Humanities Research. Humanitates*, 6:20–34, 01 2020.

[3] Herenia Ponce, Julian Chamizo-Gonzalez, and Manar Al-Mohareb. Annual reports readability from linguistic and communication perspectives: Systematic literature review. *Business and Professional Communication Quarterly*, 86:1–52, 01 2023.