Assignment 2

Amazon Dataset

December 3rd 2024

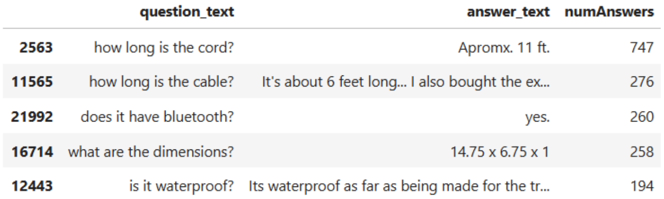
Dataset Description

For our analysis, we utilized the **Amazon Question/Answer Data**, which we downloaded from `https://cseweb.ucsd.edu/~jmcauley/datasets/amazon/qa/`. The dataset contains detailed information about questions and answers related to Amazon products. The original dataset includes the following columns:

- **asin**: Product ID
- **questions**: JSON list of questions related to the product, with the following attributes:
  - **askerID**: ID of the asker
  - **questionTime**: Timestamp of when the question was posted
  - **questionText**: The text of the question
  - **answers**: JSON list of answers associated with the question, with the following details:
    - **answererID**: ID of the answerer
    - **answerTime**: Timestamp of when the answer was posted
    - **helpful**: A list showing the number of helpful votes and total votes
    - **answerType**: Indicates whether the answer is "Yes/No" or open-ended
    - **answerScore**: A helpfulness score ranging from 0 to 1

Exploratory Analysis

Top 5 Questions with High Engagement



	question_text	answer_text	numAnswers
2563	how long is the cord?	Apromx. 11 ft.	747
11565	how long is the cable?	It's about 6 feet long... I also bought the ex...	276
21992	does it have bluetooth?	yes.	260
16714	what are the dimensions?	14.75 x 6.75 x 1	258
12443	is it waterproof?	Its waterproof as far as being made for the tr...	194

Figure 1: Top 5 Questions with Example Answers

After preprocessing the data by removing duplicate questions and standardizing them to lowercase, the dataset comprises 213,639 unique entries. Among these, the top 5 questions received the highest number of answers.

Our initial analysis suggests that shorter questions tend to attract more engagement, as they are clearer and quicker to understand. Users are more inclined to respond to concise questions since they require less time and effort to process compared to longer, more complex ones. This observation guided our decision to consider text length as a feature in our baseline model.

Distribution of Number of Answers

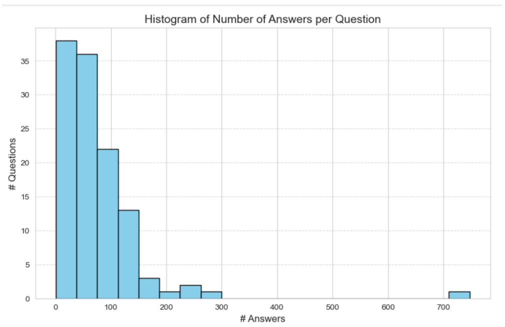


Figure 2: Distribution of Number of Answers per Question

The distribution of answers per question was analyzed to understand engagement patterns across the dataset. The plot reveals a highly skewed distribution, with most questions receiving fewer than 100 answers. Notably, an extreme outlier was identified, where one question garnered approximately 700 responses.

This right-skewed distribution highlights a significant disparity in user engagement, where only a small subset of questions receives a disproportionately high number of answers.

Predictive Task

Our predictive task focuses on predicting the number of responses a question will receive, framed as a regression problem.

Feature Selection and Justification - based on our exploratory analysis, we observed that shorter and concise questions tend to attract higher engagement. This insight justifies the inclusion of text length as a key feature. Additionally, we employ text analysis techniques such as TF-IDF to capture the semantic content of the questions and sentiment analysis to examine emotional tone, as these may influence user engagement.

Data Preprocessing - To prepare the dataset for modeling and ensure useful feature extraction, we performed the following preprocessing steps:

- **Added Column:** Created a new feature, *numAnswers*, which counts the number of answers associated with each question to serve as the target variable for the regression task.
- **Filter:** Removed repeating questions, reducing the dataset size from 867,921 questions to 215,868 unique entries.
- **Text Prepossessing:** Cleaned the text data by removing stop words, stripping punctuation, and converting all text to lowercase for consistency in relevant models.

These steps ensure the data is clean, structured, and optimized for feature extraction, reducing noise and enhancing model performance.

Model Evaluation and Validity - to assess the validity of our predictions and ensure their significance, we will split the dataset into training, validation, and test sets to evaluate model performance on unseen data. Use metrics Mean Squared Error (MSE) to measure prediction accuracy. Where MSE is given by

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

To confirm the model's significance, we will compare its performance to a baseline model.

Model

In this section, we outline the models designed to predict the number of responses each question will receive. The models vary in complexity and the features they use, enabling a comparative analysis of different feature representations and their effectiveness.

Baseline Model

The baseline model uses text length as the sole feature. Using the text length for the answers in the model we are trying to understand the correlation with question's users would have. This simple approach serves as a baseline for evaluating the performance of more sophisticated models. With an MSE of 17.495, it provides a clear point of comparison for subsequent models.

First Model: Sentiment Compound Score

Building on the baseline model, the first model incorporates sentiment analysis as an additional feature. Using the compound sentiment score extracted via the *SentimentIntensityAnalyzer* from *nltk*, this model evaluates the tone of a question on a scale from -1 (very negative) to 1 (very positive). The hypothesis is that questions with more positive or neutral sentiments attract higher engagement. This model achieved the lowest MSE of 8.663, demonstrating that sentiment, combined with text length, is a significant predictor of engagement.

The Mean Squared Error (MSE) for the SentimentIntensityAnalyzer predictions is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where:

- y_i: True *numAnswers* count for the i-th instance.
- $\hat{y}_i$: Predicted *numAnswers* count from the SentimentIntensityAnalyzer for the i-th instance.
- n: Total number of instances.

Second Model: TF-IDF without Preprocessing

The second model employs a more advanced feature extraction technique: TF-IDF (Term Frequency-Inverse Document Frequency) using the raw question text. This approach captures the importance of specific terms in a question relative to the entire dataset. While it provides additional insight into word importance, this model yielded a higher MSE of 24.033, possibly due to noise from unprocessed text data.

The Mean Squared Error (MSE) for a model trained on TF-IDF without preprocessing is given by:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where:

- y_i: True target value for the i-th instance.

- $\hat{y}_i$: Predicted value for the i-th instance, obtained using TF-IDF features.
- n: Total number of instances.

Third Model: TF-IDF with Lowercasing and Punctuation Removal

To address the limitations of raw text, the third model preprocesses the question text by lowercasing and removing punctuation before applying TF-IDF. These preprocessing steps standardize the text and reduce variability caused by capitalization and unnecessary characters. This model showed marginal improvement over the second model, achieving an MSE of 23.476, suggesting that preprocessing helps, but additional refinement is needed.

The Mean Squared Error (MSE) for a model trained on TF-IDF features (with lowercasing and punctuation removal) is given by:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where:

- y_i: True target value for the i-th instance.
- $\hat{y}_i$: Predicted value for the i-th instance, obtained using TF-IDF features.
- n: Total number of instances.

The TF-IDF features are computed after preprocessing the text:

- All text is converted to lowercase.
- Punctuation is removed before generating the TF-IDF matrix.

Fourth Model: TF-IDF with Lowercasing, Punctuation Removal, and Stopword Removal

The fourth model extends preprocessing by also removing stopwords (common words like “the” and “and”). By reducing noise and focusing on more meaningful terms, this approach further refines the features. This model achieved an MSE of 18.113, demonstrating that removing stopwords improves performance compared to the previous TF-IDF models. However, it still did not outperform the Sentiment Model.

The Mean Squared Error (MSE) for the fourth model, trained on TF-IDF features with lowercasing, punctuation removal, and stopword removal, is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where:

- y_i: True target value for the i-th instance.
- $\hat{y}_i$: Predicted value for the i-th instance, obtained using TF-IDF features.
- n: Total number of instances.

The TF-IDF features are generated after applying the following preprocessing steps to the text:

- All text is converted to lowercase.
- Punctuation is removed.
- Stopwords (common words like ”the,” ”and,” ”is,” etc.) are removed.

Model	**Description**	**MSE**
Baseline Model	Text length as the only feature	17.495
Sentiment Model	Text length + Sentiment compound	8.663
TF-IDF Model	TF-IDF without preprocessing	24.033
Preprocessed TF-IDF Model	TF-IDF + lowercase/punctuation removal	23.476
Stopword Removal Model	TF-IDF + lowercase/punctuation removal/stopword removal	18.113

Table 1: Overview of the different models built for predicting the number of responses.

Transformation	**Description**	**MSE**
No Transformation	Original features (length + sentiment)	8.663
Log Transformation	Apply `log(x)` to features	9.194
Log1p Transformation	Apply `log1p(x)` to features	8.738
Squared Transformation	Square the feature values	8.719

Table 2: MSE for different transformations on the first model (length + sentiment).

Summary of Models

The Sentiment Model outperforms the other models tested, leveraging a combination of text length and sentiment compound to predict the number of responses effectively. When comparing different feature transformations, the model using original features (text length + sentiment) achieved the lowest Mean Squared Error (MSE) of 8.663. In contrast, applying log transformations resulted in a higher MSE of 9.194, log1p transformation yielded 8.738, and squared transformation gave 8.719. This suggests that the original features provide the most accurate and relevant predictors for this task.

Related Literature Analysis

Text-based predictive tasks are a well-researched area in Natural Language Processing (NLP), with numerous studies leveraging machine learning to extract meaningful insights and improve engagement. A particularly relevant study by Abid Ali Awan analyzed the Internet News and Consumer Engagement dataset to predict article popularity. The objective of Awan's work—to provide insights into crafting more engaging article titles—parallels our goal of identifying "good" questions that foster active engagement within a Question-and-Answer platform.

Dataset Origins and Usage - Our study utilizes the Amazon Question-and-Answer dataset, which has been previously used to study consumer behavior, including question clarity and answer helpfulness. Unlike Awan's focus on news articles, we analyze question text to predict engagement by modeling the number of answers received. The shared focus on engagement guided our adoption of similar text preprocessing techniques.

Similar Datasets and Methods - Other studies in text-based predictive tasks have utilized datasets such as Quora Question Pairs Dataset:

Used for identifying duplicate questions and improving question clarity; Yahoo! Answers Comprehensive Questions and Answers Dataset: Applied in predicting best answers and engagement levels based on question features. These datasets focus on similar themes of engagement and clarity, providing valuable insights into feature extraction and modeling strategies. State-of-the-art methods in these studies include Gradient Boosting Models (e.g., LightGBM, XGBoost) for their effectiveness in handling structured and text-based data, and Neural Networks with Embeddings (e.g., Word2Vec, BERT) for capturing semantic relationships in text.

While we did not employ neural network models due to computational constraints, we borrowed ideas such as TF-IDF vectorization, sentiment analysis, and baseline regressors from these works.

Features and Insights - the features used in Awan's study included sentiment Intensity Scores which are derived from the text to evaluate tone and mood, TF-IDF Representations for capturing term importance relative to the corpus, and engagement Metrics: As the target variable, measuring article popularity.

We incorporated similar features, such as text length as a baseline predictor, sentiment compound scores to assess the tone of the questions, and TF-IDF features with varying levels of preprocessing, including stopword removal and punctuation standardization. By adapting these features, we could align our models with effective techniques in prior work while tailoring them to the specific context of Question-and-Answer platforms.

State-of-the-Art Methods - State-of-the-art methods for engagement prediction often rely on advanced machine learning models such as Light Gradient Boosting Machines (LightGBM): Awan's study demonstrated the effectiveness of LightGBM for predicting article popularity, outperforming Random Forest and logistic regression models.Neural Networks with Contextual Embeddings (e.g., BERT) that used in more recent studies for nuanced understanding of textual data.

Although our study employs simpler regression models, the use of linear regression with engineered features aligns with the goals of interpretability and computational feasibility. Borrowing from prior works, we emphasized feature preprocessing and sentiment analysis as critical components of our predictive task.

Awan's research concluded that text features such as sentiment and word importance are critical for predicting engagement but highlighted that popularity does not always equate to quality content. Similarly, we found that sentiment and text length are key predictors, while raw TF-IDF features underperform without preprocessing.

Unlike Awan's use of complex models like LightGBM, our simpler linear regression approach achieved competitive results, emphasizing interpretability and computational feasibility. In both cases, the exploration of text features and sentiment analysis highlights their versatility in engagement prediction across different domains. Our findings reinforce the importance of leveraging text preprocessing to enhance model performance, particularly in contexts with structured text data like QA platforms.

Conclusion/Result

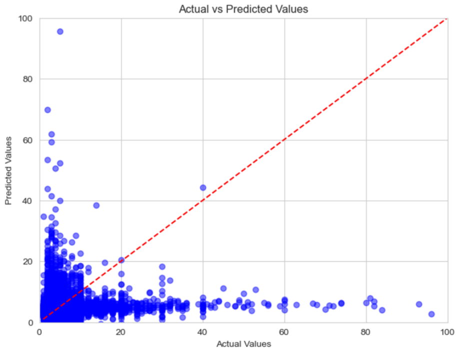


Figure 3: Many low values are misinterpreted as high and vice versa. Most values and predictions are under 20.

Our analysis aimed to predict the number of answers a question might receive using text features from Amazon's Questions and Answer dataset. The goal was to identify key text characteristics that can drive higher user engagement. We started with a baseline model using question length, then incorporated various text preprocessing techniques such as sentiment analysis, TF-IDF with lowercasing and punctuation removal, and sentiment compound scores to explore which feature representations yielded the best performance.

The results revealed that the Sentiment Model, which combined text length and sentiment score, achieved the lowest mean squared error (MSE) of 8.663. This outperformed other models, including those using just TF-IDF or simple length-based features. Sentiment analysis added valuable context, improving prediction accuracy by capturing the emotional tone behind the questions.

Our model's parameters, including the sentiment compound score, showed the strongest predictive power. The text length feature alone provided useful but less nuanced insights, while more complex features like TF-IDF and sentiment analysis better-captured patterns of user engagement. This suggests that while length remains an important factor, sentiment plays a critical role in how users engage with questions.

The proposed model succeeded because it leveraged multiple layers of text analysis to provide a more comprehensive prediction. The success of sentiment analysis shows that the tone of a question can significantly influence user responses. Conversely, simpler models that only considered text length or raw TF-IDF scores did not capture this nuance, leading to higher prediction errors. Therefore, sentiment analysis, combined with text length, proved to be the most effective for this task, while other models failed to fully capture the richness of the data.

Reference

Awan, Abid Ali. “Analyzing and Predicting Consumer Engagement”. Oct 27, 2021. `https://towardsdatascience.com/analyzing-and-predicting-consumer-engagement-8b3229f`

Kaggle. ”Yahoo! Answers dataset”. `https://www.kaggle.com/datasets/jarupula/yahoo-answers-dataset/data?select=test.csv`

Sharma, Lakshay. ”Natural Language Understanding with the Quora Question Pairs Dataset”. Jul 1, 2019. `https://arxiv.org/abs/1907.01041`