Text Analytics on Earnings Calls

Van Anh Nguyenova

June 17, 2025

Contents

1 Introduction

Earnings calls are an essential component of financial communication, where executives provide updates on a company's performance and answer questions from analysts. These transcripts contain valuable linguistic signals that can reflect corporate sentiment, transparency, and confidence. This project leverages text analytics techniques to explore how language used during earnings calls relates to business outcomes.

The analysis focuses on three key objectives: (1) predicting whether companies meet their earnings-per-share (EPS) expectations using textual features, (2) comparing the tone and politeness between analyst questions and executive responses, and (3) classifying transcripts into fiscal quarters based on early responses. By combining ngram-based features, word embeddings, politeness markers, and classification models, the project demonstrates how natural language processing can uncover patterns in financial discourse that support investor decision-making and business analysis.

2 Predicting EPS Using Text

2.1 LASSO Regression with Bigrams

We begin our analysis by training a LASSO model, using only bigrams from the opening speeches as features to predict the reported earnings per share.

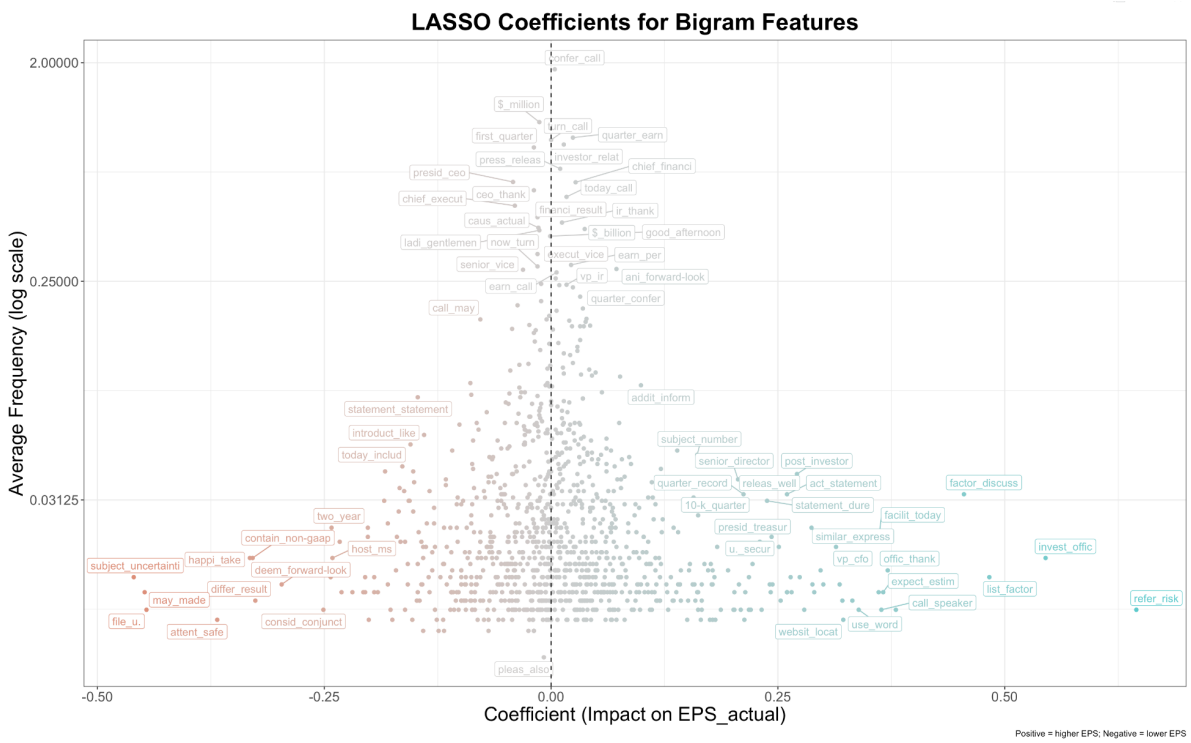

Figure 1: LASSO Coefficients for Bigram Features

The LASSO model identified specific phrases in the opening speeches of earnings calls that are linked to actual earnings outcomes. The accompanying chart illustrates how frequently these phrases appear and whether they tend to be used by companies with higher or lower earnings.

Phrases associated with higher earnings, such as "invest office," "refer risk," and "post investor", often signal confidence, proactive risk management, and a focus on growth opportunities. These terms likely indicate confidence, future-oriented planning, or addressing important financial concerns, all of which can signal strong performance and higher earnings.

In contrast, phrases such as "attent safe," "subject uncertainty," and "differ result" are more commonly used by companies with lower earnings. These terms might suggest challenges, risk-averse strategies, or a lack of clarity about the company's financial health, which could indicate lower earnings. The presence of such language could signal issues that investors or stakeholders might interpret as negative signals, potentially leading to lower confidence and lower earnings.

A significant number of bigrams have coefficients close to zero, indicating that many features do not strongly influence the prediction of EPS. For example, terms such as 'forward-looking statement', 'first quarter', 'investor relations' are common on many earnings calls, indicating general financial discussions rather than strong performance signals.

Thus, the positive words tend to convey optimism, growth, and stability, while the negative words are more associated with uncertainty, caution, or potential risks, all of which can significantly influence how investors perceive a company's future performance.

2.2 LASSO Regression with Word2Vec

To build on our earlier analysis, we used word embeddings to capture the underlying meaning of words in each company's opening speech. This method creates a more nuanced representation of the text and serves as a complementary approach to our previous model based on simple word pairs.

The resulting model achieved an accuracy of 62.2%, with a narrow confidence interval, indicating consistent performance. While this was slightly lower than the accuracy of the bigram-based model, it confirms that the tone and language used in earnings calls carry valuable signals about a company's financial results. These findings underscore the potential of advanced text analysis tools to support earnings prediction and investor insight.

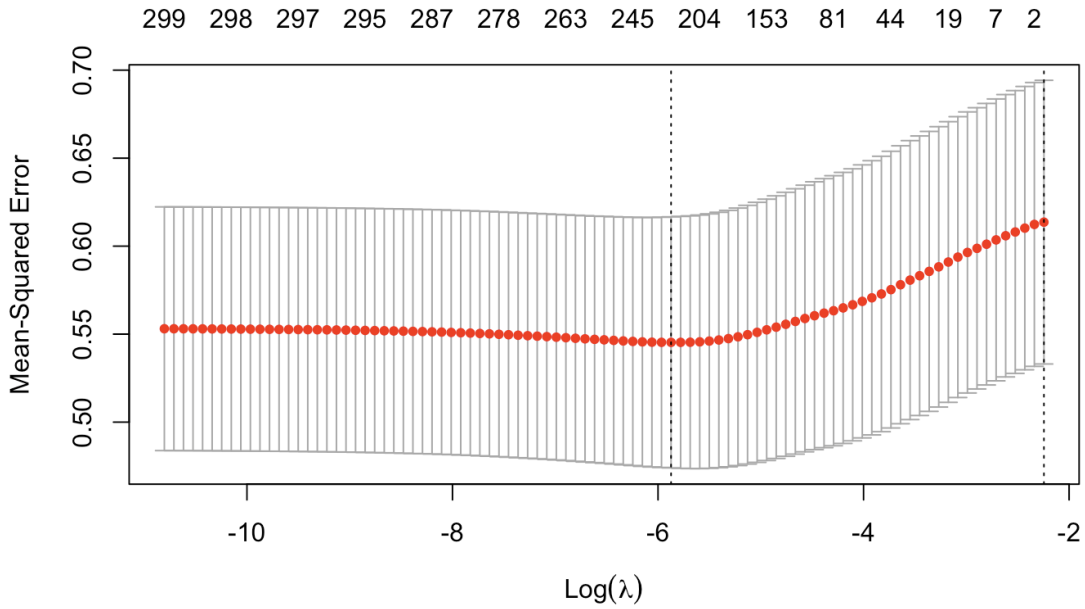

Figure 2: Cross-validation plot for LASSO with word2vec.

The chart above shows how the model's error rate changes with different levels of complexity. The lowest error was observed when the model included around 245 key features, striking a balance between accuracy and simplicity. This point, marked by the first dashed line, was selected as the model's optimal setting. A second, more conservative option, represented by the second dashed line, would simplify the model further by including fewer features, which may improve generalizability at the cost of slightly reduced accuracy.

2.3 LASSO Regression Combined Features

The chart illustrates how the model's prediction error changes as we adjust its complexity. The optimal performance occurs when the model balances simplicity and detail,

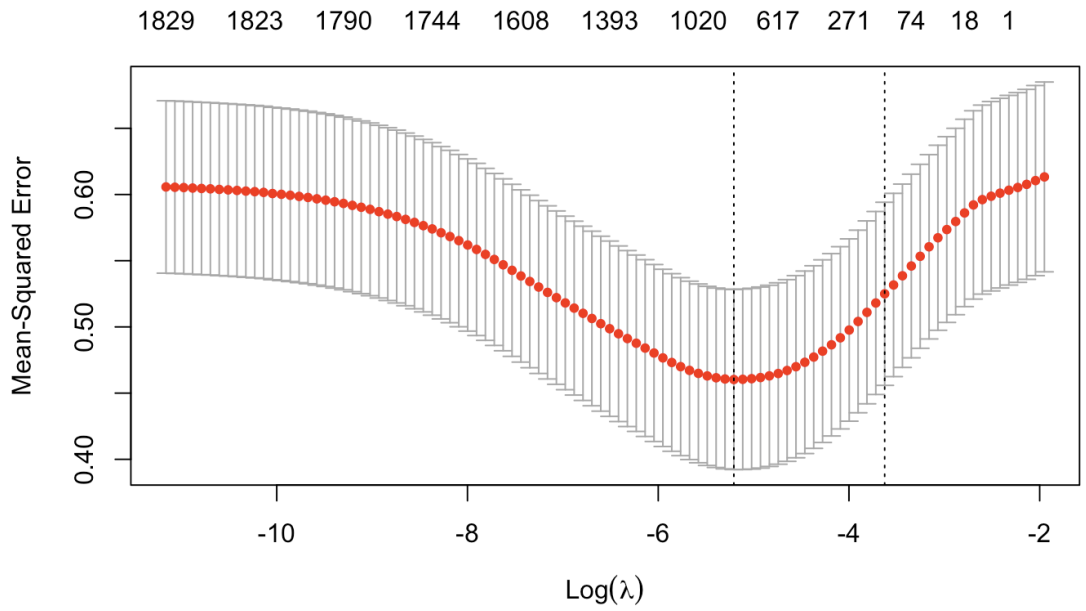

Figure 3: Cross-validation plot for LASSO with combined features.

around a tuning value that includes approximately 1,000 impactful terms. This sweet spot minimizes error without overfitting or underfitting the data.

To test whether combining two types of information, word frequency patterns (bigrams) and semantic context (word embeddings), would enhance results, we trained a third model using both feature types. This combined approach achieved an accuracy of nearly 67%, which outperformed the word embedding model alone but was slightly below the accuracy of the bigram-only model.

These findings suggest that while semantic context adds value, the most predictive power came from frequent word pairings in earnings call language. The combined model remains a strong performer and provides a more holistic representation of how executives communicate earnings outcomes.

2.4 Benchmarks

To help evaluate the strength of our predictive models, we compared them against two simple benchmark features: the length of each speech (measured by word count) and the overall tone (measured by sentiment score). These basic indicators served as reference points to assess whether more advanced text analysis methods added meaningful predictive value. As expected, the simple benchmarks provided limited insight on their own, reinforcing the importance of using more sophisticated language features to understand financial communication.

Figure 4 illustrates how well different types of language features predict a company's actual EPS, including statistical confidence intervals. The best-performing model used frequent two-word phrases (bigrams), achieving nearly 69% precision. A combined model using both bigrams and semantic word embeddings followed closely at 67%, while the model using word embeddings alone reached about 62%.

In comparison, simple benchmarks such as word count and sentiment fell below the 50% threshold, showing that they offer limited value on their own. The horizontal line

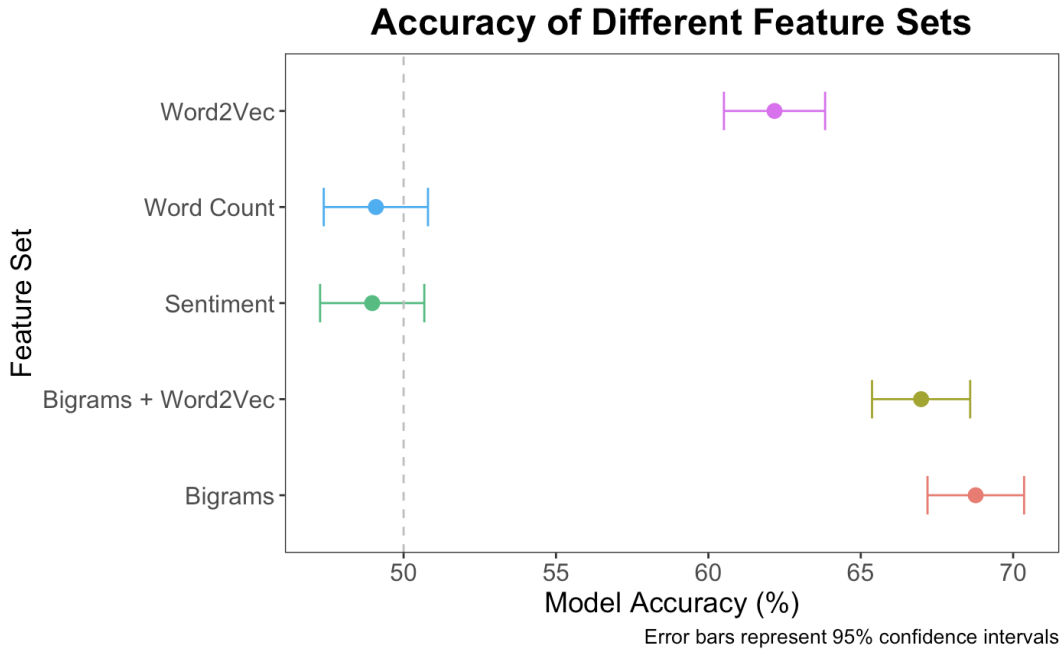

Figure 4: Accuracy of Different Feature Sets with 95% Confidence Intervals.

in the figure marks this baseline, emphasizing how much better the advanced models perform.

These results underscore that nuanced language patterns, especially short word combinations, are powerful tools for understanding financial outcomes from earnings calls, far surpassing basic text metrics.

3 Temporal Similarity in Speeches

We identified 448 companies that have complete earnings call data for all four quarters in both FY 2011 and FY 2012, totaling eight speeches per company. In this section, the goal is to compute the average similarity between each company's Q1 2011 speech and its seven subsequent speeches (Q2–Q4 of 2011 and Q1–Q4 of 2012).

FY	FQ	avg_similarity	quarter
2011	2	0.9927877	2011Q2
2011	3	0.9915761	2011Q3
2011	4	0.9907283	2011Q4
2012	1	0.9909612	2012Q1
2012	2	0.9905611	2012Q2
2012	3	0.9897214	2012Q3
2012	4	0.9890257	2012Q4

Table 1: Average Similarity Scores for Earnings Calls Across Quarters

All similarity values are very high, indicating that the opening speeches throughout the quarters are very similar to the first one in 2011. The similarities range from 0.993 to 0.989, showing only gradual slight decreases. Notably, in Q1 2012, the similarity even

increased slightly. This increase could be due to Q1 being the first quarter of the new fiscal year, potentially leading to a more standardized or familiar opening speech.

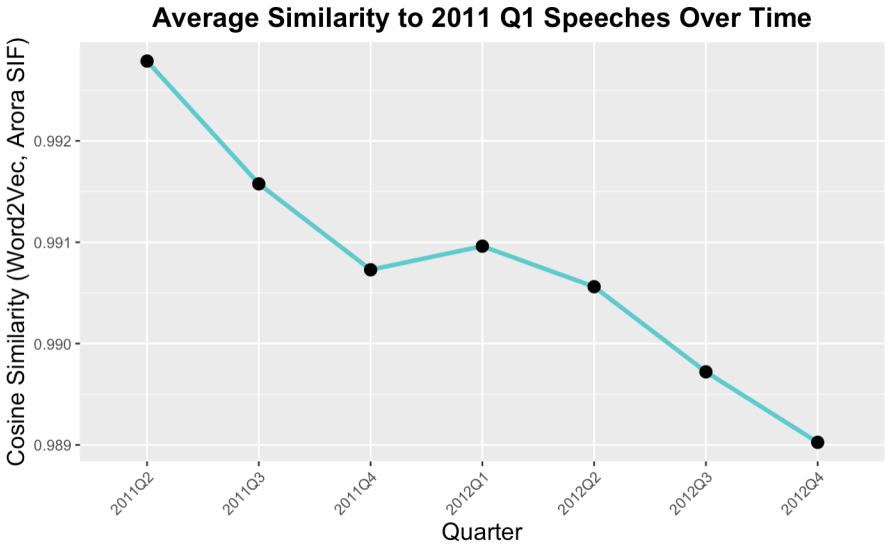

Figure 5: Average Similarity to 2011 Q1 Speeches Over Time.

Figure 5 additionally highlights the gradual decrease in similarity. The most noticeable decline occurs from Q2 2011 to Q1 2012, with Q1 2012 showing a slight increase compared to Q4 2011. Overall, the trend demonstrates a gradual decrease in similarity, suggesting that speeches evolved slightly over time, possibly reflecting changing business conditions or changes in corporate communication strategies.

4 Q&A Text Analysis

We now turn to the question-and-answer portion of the earnings calls to examine which features of the Q&A content help predict a company's actual earnings per share. Two separate models are trained: one using the text of the first ten questions from each call, and another using the corresponding answers.

4.1 Predicting EPS from Questions vs. Answers

The results reveal distinct patterns in the linguistic features that are predictive of a company's actual earnings per share (EPS).

In the question-based model (see Figure 6), several positively associated phrases suggest that analyst questions referencing specific individuals or market-related topics, such as "apple", "peter", "steve_just", and "new_market", are linked with higher-than-expected earnings per share (EPS). References to "apple" and "steve_just" likely pertain to Apple Inc. and its leadership, signaling positive sentiment or optimism, which may be indicative of strong company performance. These terms may reflect more forward-looking or confident lines of inquiry. Conversely, negatively associated terms such as "burn", "worth", and "engine" may point to investor concerns or areas of uncertainty. The term "apple_apple" also appeared among the most negatively weighted, despite the earlier positive references to Apple, suggesting that repeated mentions or specific usage may influence interpretation in different ways.

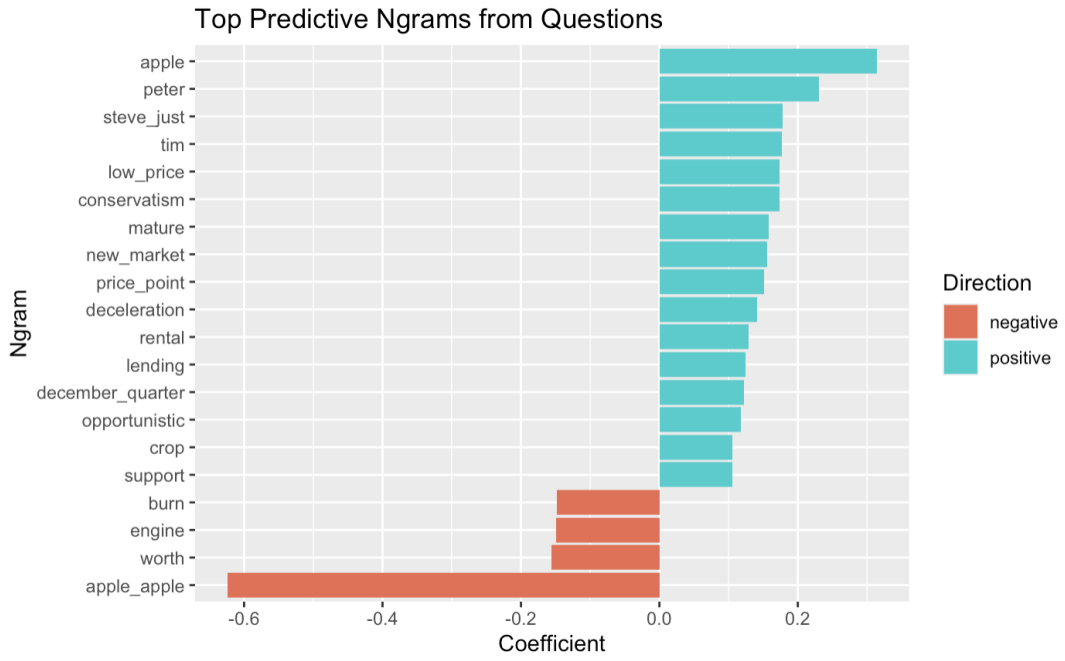

Figure 6: Top Predictive Ngrams from Questions

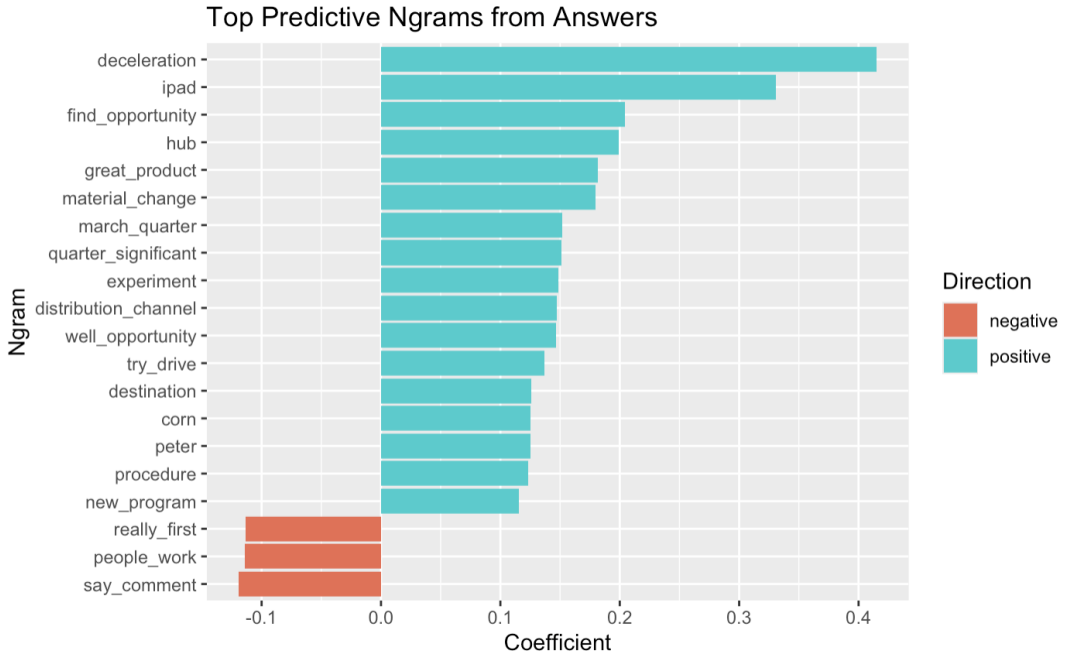

Figure 7: Top Predictive Ngrams from Answers

In the answer-based model (see Figure 7), positively weighted terms like "deceleration", "find_opportunity", and "great_product" suggest that when executives address business opportunities, innovation, or product quality explicitly, these responses are typically linked to stronger financial results. Notably, "deceleration" carries the most positive weight in this model, and also appears in the question-based model, implying that transparent acknowledgment of slowing growth can build investor trust. Some terms, such as "peter" and "ipad", also appear in both models, reinforcing their relevance across different

types of discourse. On the other hand, phrases such as "say_comment" and "really_first" are negatively associated, potentially reflecting vague or non-committal language that may raise concerns about performance clarity.

4.2 Testing Transfer Learning Between Question and Answer Models

Next, we evaluate the accuracy of the two models trained earlier. This includes both in-context evaluation—testing the question-based model on question data and the answer-based model on answer data—and cross-context transfer learning, where we apply the question-based model to answer data and the answer-based model to question data. This comparison helps assess how well the models generalize across different types of conversational input.

Predicted On	Trained On Questions	Trained On Answers
Questions	60.80	59.75
Answers	59.96	61.66

Table 2: Model Accuracy for Different Predicted and Trained Data

In-context evaluations show that the question-based model performs best when tested on question data (60.80%), and the answer-based model performs best on answer data (61.66%), as expected. These represent the diagonal elements in the matrix, where each model is evaluated on the same type of text it was trained on. Cross-context evaluations yield slightly lower accuracy scores (59.96% and 59.75%, respectively), though the drop is not as substantial as one might expect. Still, this suggests that while there is some predictive overlap between question and answer texts, the models still perform best within their original training context.

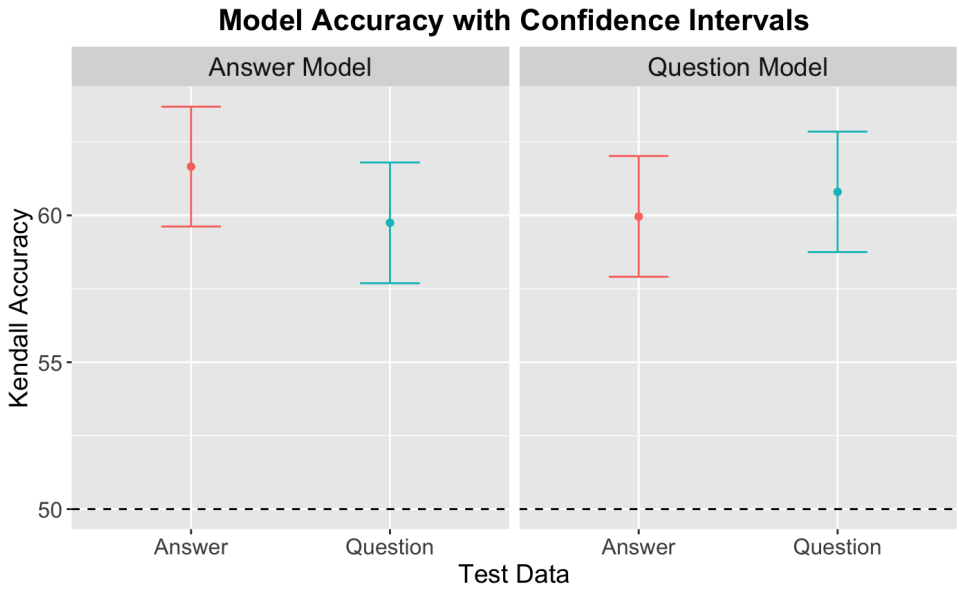

Figure 8: Model Accuracy with Confidence Intervals for Answer and Question Models.

The corresponding plot (see Figure 8) presents these findings visually, with each point

representing the precision scores and its 95% confidence interval. Cross-context predictions show a slight drop in accuracy but do not exhibit noticeably wider intervals.

5 Politeness Analysis

Now, we focus on extracting and analyzing politeness features from the question and answer texts.

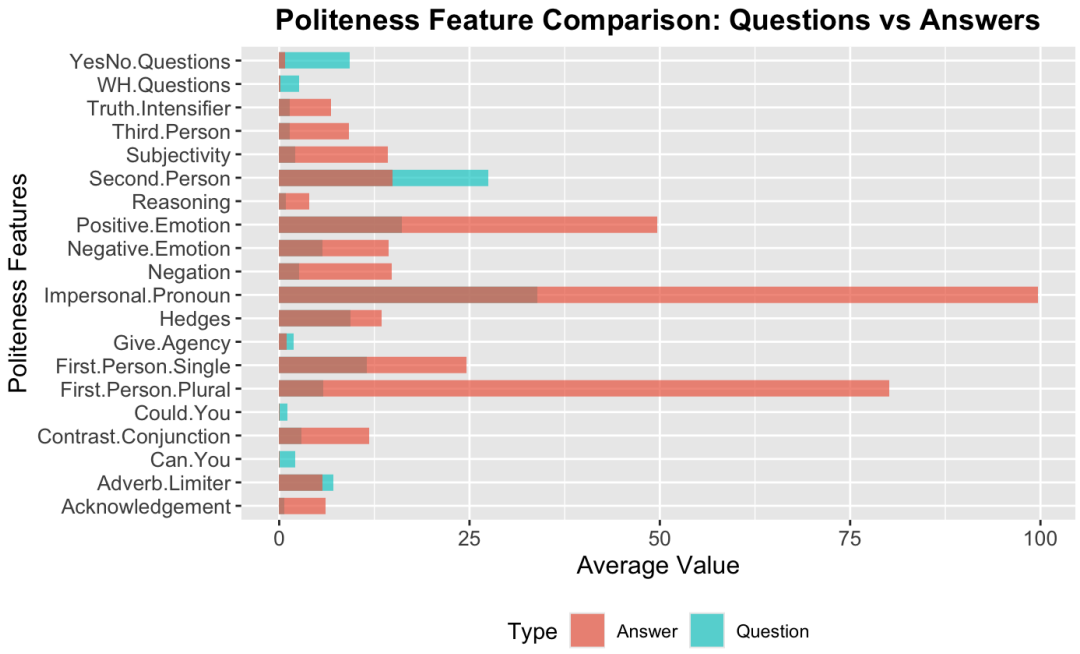

Figure 9: Politeness Feature Comparison: Questions vs Answers

We observe some notable trends from the analysis of the politeness feature differences between questions and answers (see Figure 9). The features with the largest differences are those reflecting more formal and detached language, such as first-person plural and impersonal pronouns, which are predominantly more common in answers than in questions. This indicates that answers tend to be more formal and structured, often involving more general statements or shared experiences, likely reflective of the way answers often address a broader audience or offer explanations.

On the other hand, features that suggest direct engagement, like second-person pronouns and yes/no questions, are more prominent in questions. This reflects the more personal and direct nature of questions, where engagement with the recipient is often more pronounced.

Other notable differences include the use of hedges, reasoning, and negation, which are more frequent in answers. Hedges and reasoning reflect the cautious and explanatory tone often used when providing answers, while negation highlights the clarifications or corrections that are commonly necessary in responses.

Overall, the data shows that answers are generally more formal and structured, using features that involve clarification and explanation, while questions tend to focus on directness, engagement, and the expression of curiosity.

5.1 Predicting Text Type: Question vs. Answer

In this section, a LASSO model is trained to predict whether a turn is a question or an answer based on politeness features.

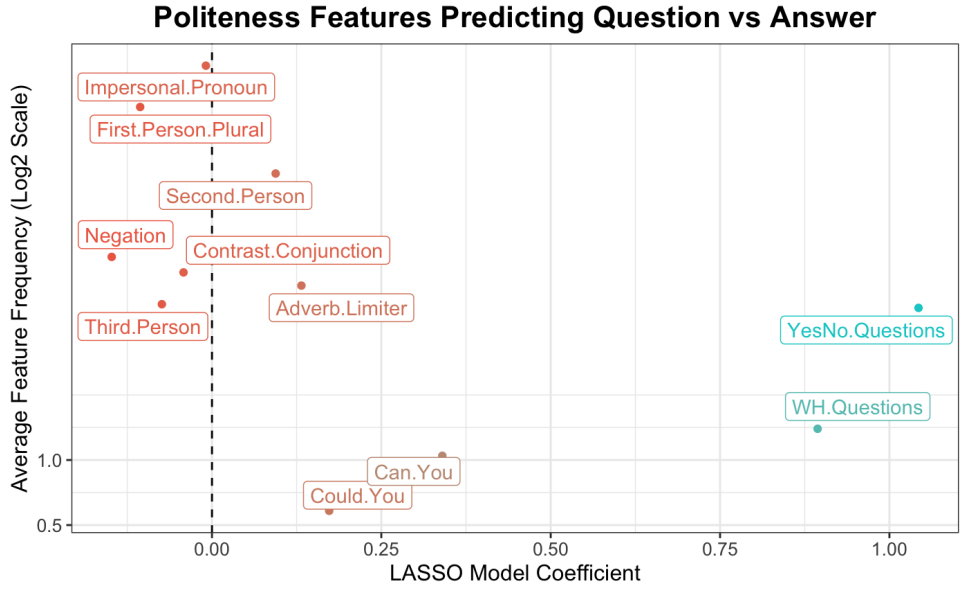

Figure 10: Politeness Features Predicting Question vs Answer

From Figure 10, we can observe that features such as "Impersonal Pronoun," "First Person Plural," and "Second Person" have negative coefficients and higher average frequencies, suggesting they are more prominent in prediciting answers. This likely reflects the use of more formal and impersonal language in answers.

On the other hand, features like "YesNo Questions" and "WH Questions" have higher coefficients for questions, as expected. These features, representing yes/no and WH-type questions (e.g., who, what, where), play a significant role in identifying questions. Additionally, features like "Can.You" and "Could.You" have lower coefficients, indicating they have a smaller impact on distinguishing between questions and answers but still lean towards being more frequent in questions.

Overall, the plot reveals that formal language markers such as first-person plural and impersonal pronouns are prominent in predicting answers, while more direct and engaging features like question types (yes/no and WH) are more common in questions.

6 Classifying Quarter Based on Answers

Finally, we predict the quarter of the earnings call.

Predicted 1	Predicted 2	Predicted 3	Predicted 4	Actual
767	393	262	310	1
290	615	263	193	2
1403	1645	2091	1470	3
239	216	293	464	4

Table 3: Confusion Matrix

The most common errors occur when Quarter 3 is misclassified as Quarter 4, Quarter 2, and Quarter 1, with 1470, 1645, and 1403 incorrect predictions, respectively. These errors suggest that the model struggles to classify Quarter 1, Quarter 2, and Quarter 4 correctly, likely due to overlapping features in the text data.

The multinomial logistic regression model achieved an accuracy of approximately 36.07% in predicting the fiscal quarter (FQ) across all four categories. This relatively low accuracy indicates that the model is not performing optimally and struggles to correctly classify the quarter in most cases. While this accuracy provides some insight into the model's performance, there is considerable room for improvement.

Quarter	**Accuracy**
1	0.2841793
2	0.2143604
3	0.7188037
4	0.1903980

Table 4: Accuracy per Quarter

Table 4 presents the model's accuracy in predicting each fiscal quarter. The model performed best in identifying Quarter 3, achieving an accuracy of approximately 71.9%. In contrast, performance for Quarters 1, 2, and 4 was considerably lower—28.4%, 21.4%, and 19.0%, respectively. These results suggest that the model is most effective at distinguishing Quarter 3 but has limited predictive accuracy for the remaining quarters, particularly Quarter 4.

7 Conclusion

This analysis provides key insights into how language used in earnings calls correlates with financial performance and stakeholder communication strategies. Predictive models leveraging bigrams and word embeddings proved effective in estimating earnings-per-share (EPS), with bigram-based models offering the most accurate forecasts. These findings suggest that the phrasing and tone in executive communication can serve as reliable indicators of company performance.

The analysis of Q&A interactions further revealed systematic linguistic differences between analyst questions and executive responses. Questions tended to be more direct and engagement-focused, while answers were more formal and impersonal. This distinction was confirmed through politeness feature modeling, highlighting consistent communication styles across earnings calls.

Although the model showed limited accuracy in classifying fiscal quarters from early responses, its strong binary classification performance for Q1 suggests opportunities for refining models to capture temporal cues more effectively.

Overall, the results emphasize the strategic importance of communication in financial reporting. Firms can leverage these insights to enhance messaging consistency, anticipate investor sentiment, and improve transparency in earnings calls.