Text Analytics on Earnings Calls

Van Anh Nguyenova

June 16, 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.

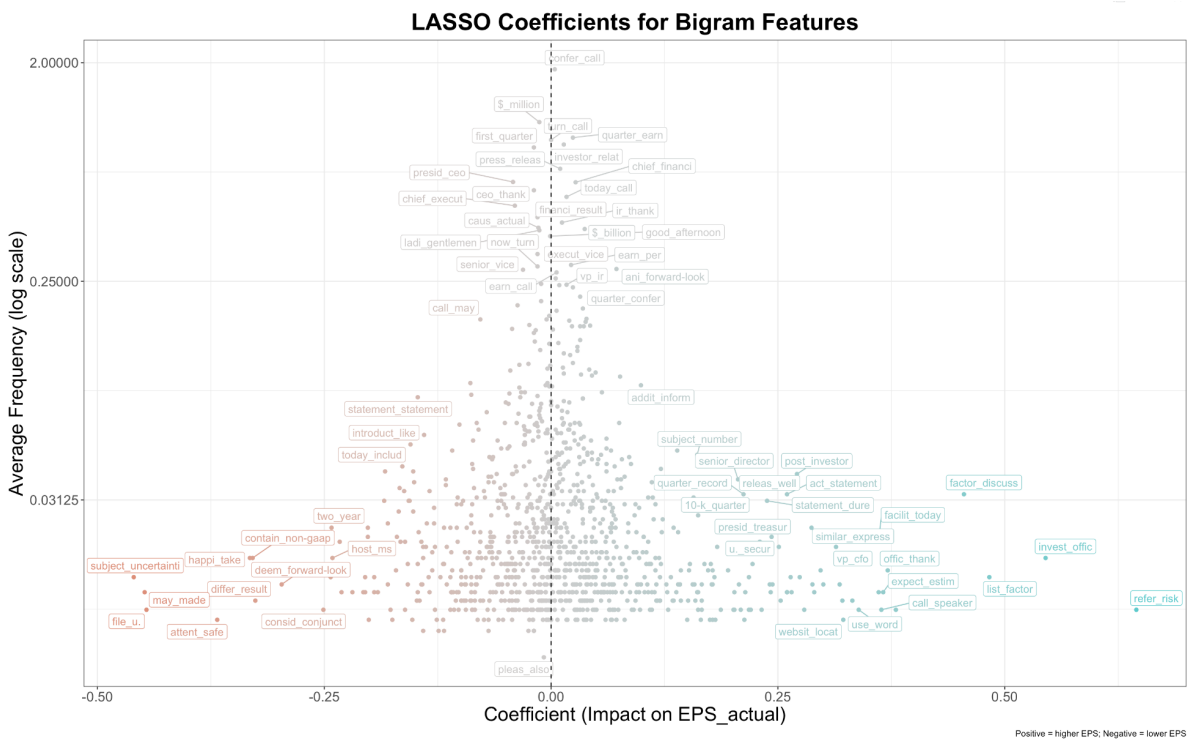

Figure 1: LASSO Coefficients for Bigram Features

The LASSO regression model has identified key bigrams from the opening speeches that correlate with higher and lower earnings per share (EPS_actual). The plot shows these bigrams along two dimensions: their impact on EPS (x-axis) and their average frequency in speeches (y-axis, log scale). Bigrams with positive coefficients are associated with higher EPS, while those with negative coefficients indicate lower EPS.

On the right side of the plot, positive bigrams, such as "invest_offic", "refer_risk," and "post_investor" suggest discussions around growth opportunities, strategic investments, or managing risks in a way that aligns with business success. These terms likely indicate confidence, future-oriented planning, or addressing important financial concerns, all of which can signal strong performance and higher earnings.

On the other hand, negative bigrams like "attent_safe", "subject_uncertainit" and "differ_result" appear to be linked with uncertainty or caution in the speech. 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 like "forward-look_statement," "first_quarter," and "investor_relat" are common across 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 extend our analysis beyond bigram features, we utilized word embeddings to represent each earnings call opening speech as a vector. This word embedding approach offers a semantically richer representation of the speeches and serves as a complementary model to the bigram-based LASSO model developed in earlier.

The LASSO model trained on word2vec embeddings achieved an accuracy of 62.17%, with a 95% confidence interval ranging from 60.51% to 63.83%. This indicates a moderately strong association between the semantic content of the opening speeches and the actual earnings per share (EPS), although it performed slightly worse than the bigram-based LASSO model. The relatively narrow confidence interval suggests stable performance across the test set. These results highlight that word embeddings capture meaningful latent information that correlates with financial outcomes, supporting their utility as predictive features in earnings call analysis.

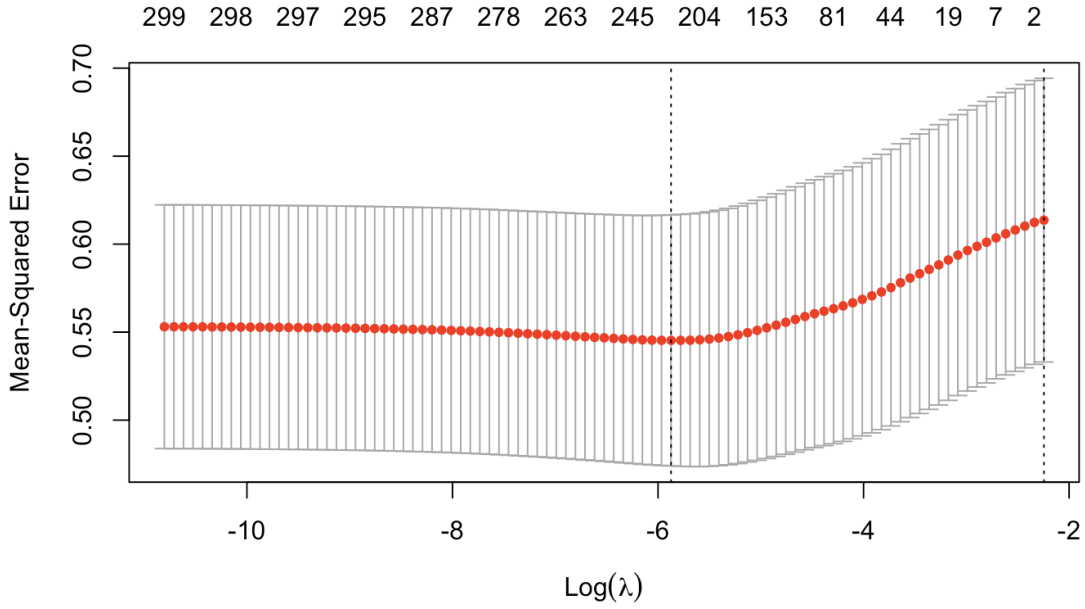

Figure 2: Cross-validation plot for LASSO with word2vec.

The cross-validation plot above displays the mean-squared error (MSE) across a range of regularization strengths (`log(λ)`) for the LASSO model using word2vec embeddings. The red dots represent the average MSE at each λ, while the vertical grey lines indicate ± 1 standard error.

The minimum MSE achieved is approximately 0.55, which occurs around $\log(\lambda) \approx -6$, the value chosen as `lambda.min`. At this point, the model retains nearly 245 non-zero coefficients, indicating the number of features actively contributing to prediction. This point is marked by the leftmost vertical dashed line. The second vertical line on the right corresponds to `lambda.1se`, a more conservative choice that selects the largest λ within one standard error of the minimum — this would favor a sparser model.

2.3 LASSO Regression Combined Features

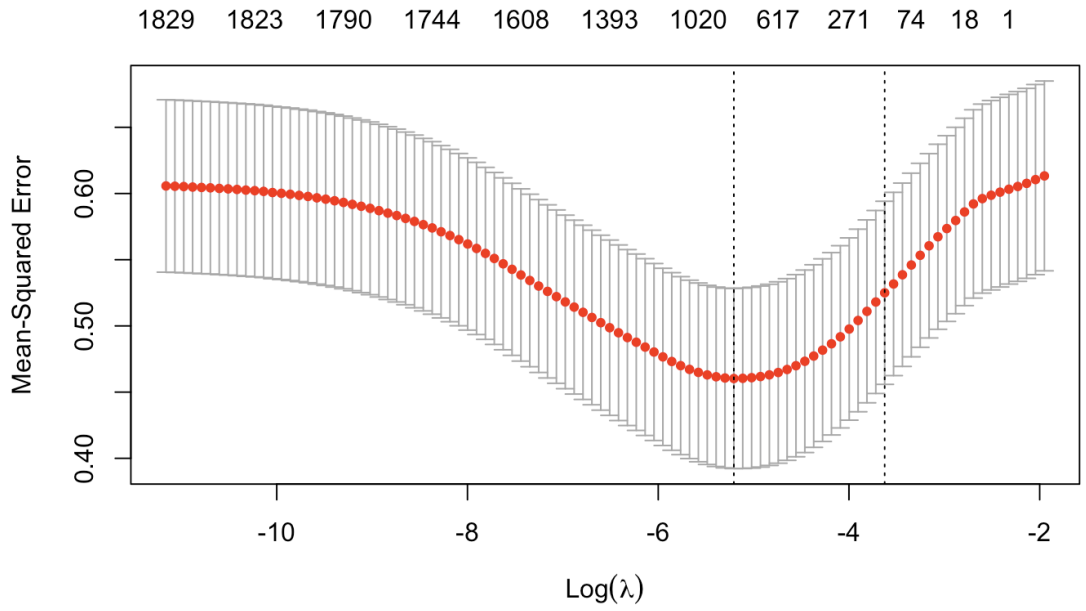

Figure 3: Cross-validation plot for LASSO with combined features.

Based on the plot, MSE varies as `log(λ)` changes. For `log(λ)` values around -5, the MSE is at its lowest, indicating the optimal value of λ for the model. As λ decreases further (i.e., `log(λ)` becomes more negative), the MSE starts to rise, suggesting that too little regularization leads to overfitting. On the other hand, as λ increases (`log(λ)` becomes more positive), the MSE also increases, suggesting that too much regularization leads to underfitting.

The vertical dotted lines, positioned around `log(λ)` ≈ -5, indicate the range where the MSE is minimized. The range between `log(λ)` ≈ -5 and `log(λ)` ≈ -4 corresponds to the optimal regularization region, where the model achieves the best balance between bias and variance.

To explore whether integrating lexical and semantic features improves performance, we trained a third LASSO model using a combined feature set that merges bigram frequencies and word2vec embeddings. The resulting model achieved an accuracy of 66.98%, with a 95% confidence interval ranging from 65.37% to 68.59% based on Kendall's tau.

Compared to the bigram-only model (68.77%) and the word2vec-only model (62.17%), the combined model performed better than the semantic model alone but slightly worse than the bigram model. This suggests that while word embeddings provide additional semantic depth, the bigram features carried most of the predictive power in this case. Nonetheless, the combined model still maintained a strong and stable performance, with nearly 1,000 non-zero coefficients at the optimal lambda (`lambda.min`), as shown in the cross-validation plot.

2.4 Benchmarks

To contextualize the performance of our trained models, we introduced two simple benchmark features: word count and sentiment score of the truncated opening speeches. Word

count was calculated as the number of alphabetic tokens in each speech, while sentiment was computed using the `sentiment_by()` function to obtain the average sentiment score per document. We then evaluated their predictive accuracy with respect to `EPS_actual` using the `kendall_acc()` function.

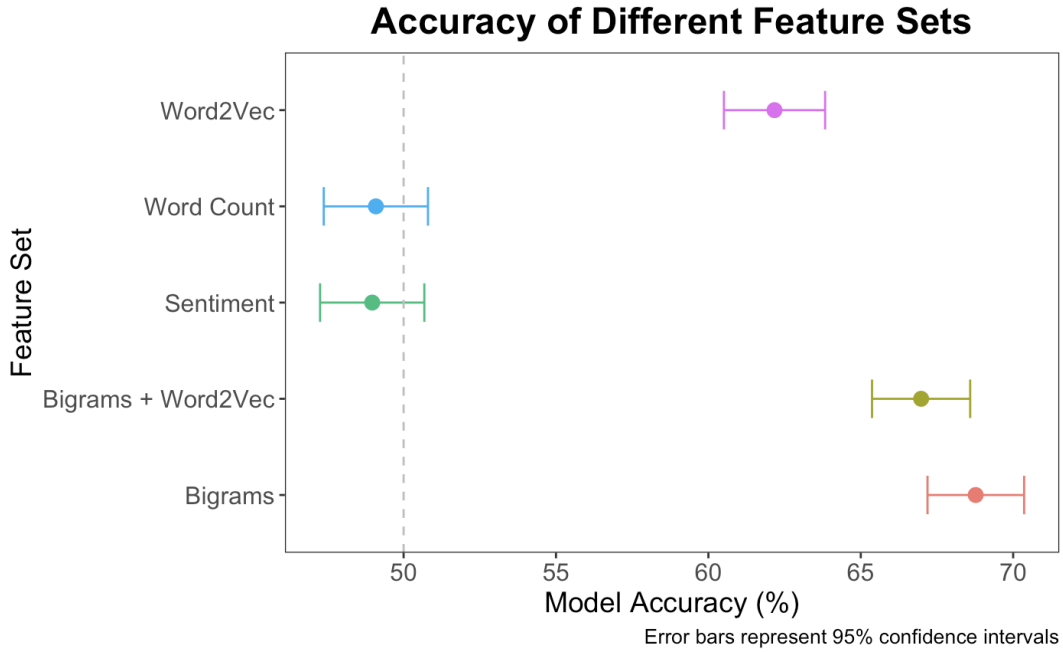

Figure 4: Accuracy of Different Feature Sets with 95% Confidence Intervals.

Figure 4 displays the model accuracy of different feature sets in predicting actual earnings per share (EPS), along with 95% confidence intervals. Among the five evaluated approaches, the bigram-based LASSO model achieved the highest accuracy at approximately 68.77%, followed closely by the combined model (bigrams + word2vec) at 66.98%. The word2vec-only model reached 62.17%, suggesting that semantic features contribute predictive value, though not as strongly as lexical n-grams.

In contrast, the two benchmark models, based on word count and sentiment score, performed slightly below the 50% baseline, highlighting their limited standalone predictive power. The horizontal dashed line at 50% represents chance-level performance, against which all three LASSO models clearly outperform.

Overall, the plot demonstrates that while simple heuristics provide minimal insight, more sophisticated text representations, particularly bigrams, are effective for modeling financial outcomes from corporate communication.

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).

All the 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

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

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.

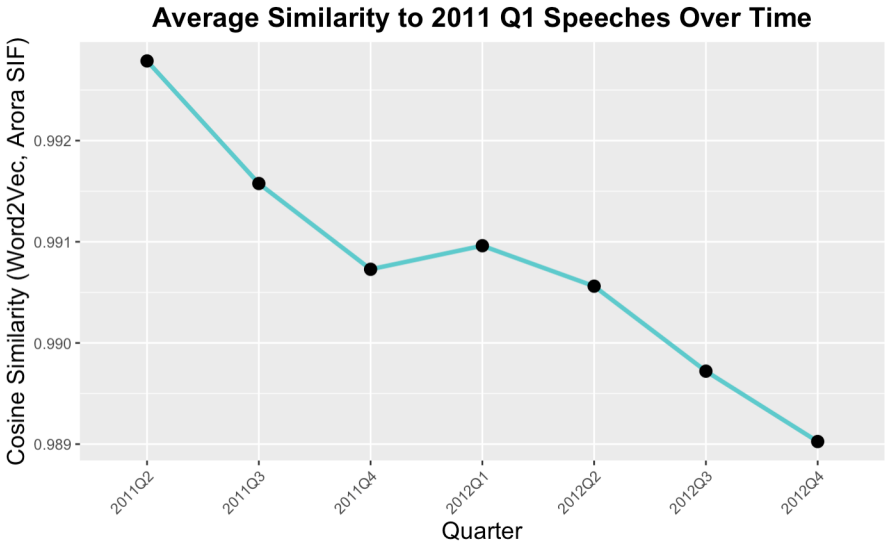

Figure 5: Average Similarity to 2011 Q1 Speeches Over Time.

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 the speeches evolved slightly over time, possibly reflecting changing business conditions or shifts 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 (EPS). Two separate models are trained: one using the text of the first ten questions from each call, and another using the corresponding answers.

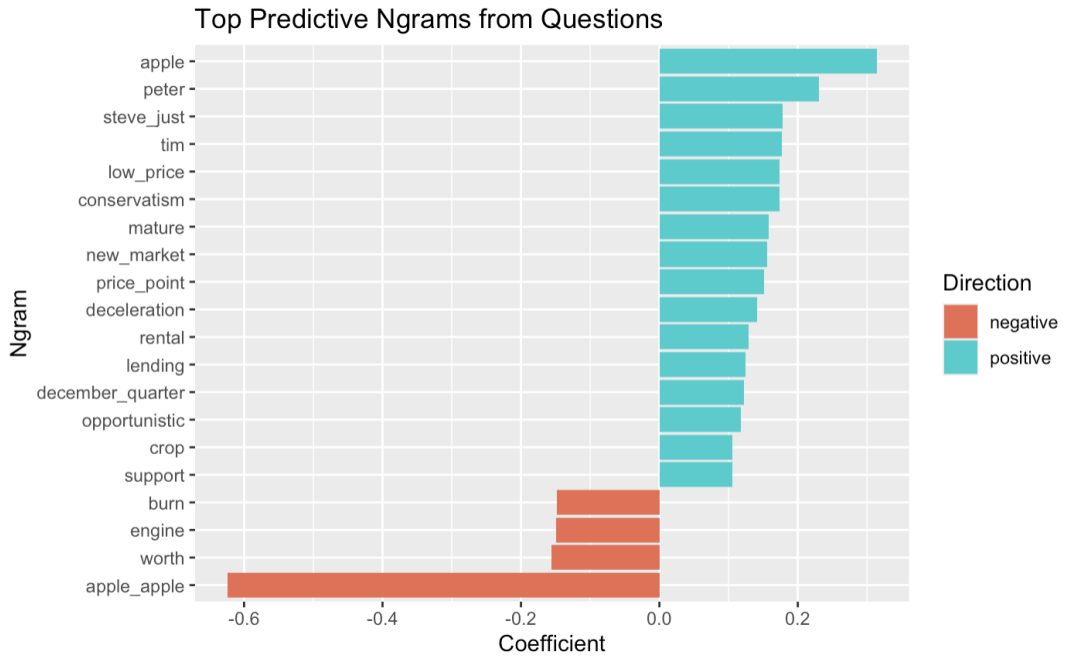

Figure 6: Top Predictive Ngrams from Questions

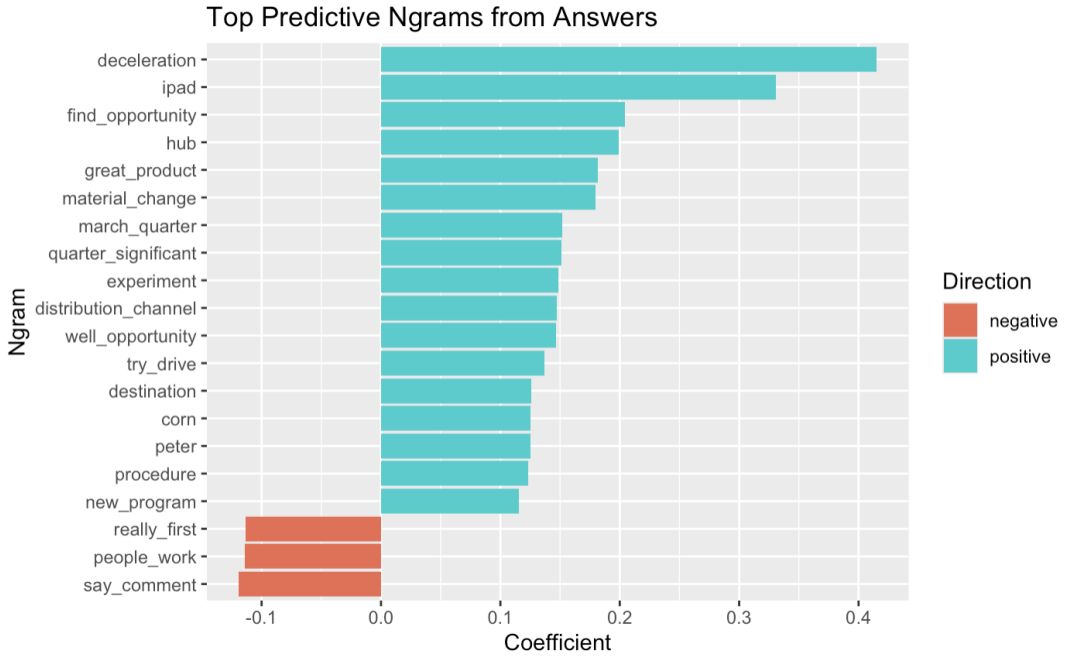

Figure 7: Top Predictive Ngrams from Answers

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 ngrams suggest that analyst questions referencing specific people or market-related terms (e.g. `apple`, `peter`, `steve_just`, `new_market`) are indicative of higher-than-expected EPS. The inclusion of `apple` and `steve_just` likely refers to the company Apple and its leadership,

which may reflect positive sentiment or optimism surrounding the company, indicating strong performance. These terms may signal more optimistic or forward-looking inquiries. However, negatively associated ngrams like `burn`, `worth`, or `engine` may signal concern or uncertainty, and possibly also relate to engine as an industry term, potentially aligning with lower EPS outcomes. Interestingly, the term `apple_apple` also appears among the negative coefficients, with the strongest effect on EPS. This contradicts the earlier positive association with Apple, suggesting that there may be ambiguity in how this term is interpreted in different contexts.

In the answer-based model (see Figure 7), positively weighted terms such as `deceleration`, or `find_opportunity`, or `great_product` suggest that when executives speak directly about opportunities, innovation, or product strength, these are often associated with better earnings results. Notably, `deceleration` has the most impactful positive coefficient here (it also appears positively in the question-based model), implying that when companies proactively acknowledge slowing trends, it can build investor trust. We also observe the recurrence of certain terms that appeared as top predictive ngrams in the question-based model, such as `peter` and `ipad`, which likely relate to Apple. Negative terms like `say_comment` or `really_first` might indicate vague or deflective language, potentially raising red flags about the company's performance.

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.

The corresponding plot (see Figure 8) presents these findings visually, with each point representing the Kendall accuracy and its 95% confidence interval. Cross-context predictions show a slight drop in accuracy but do not exhibit noticeably wider intervals.

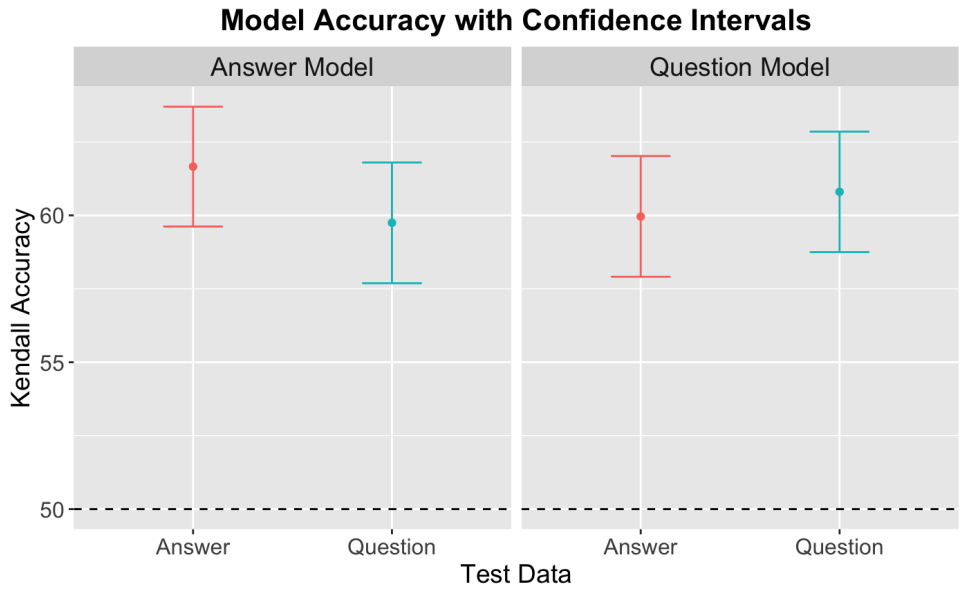

Figure 8: Model Accuracy with Confidence Intervals for Answer and Question Models.

5 Politeness Analysis

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

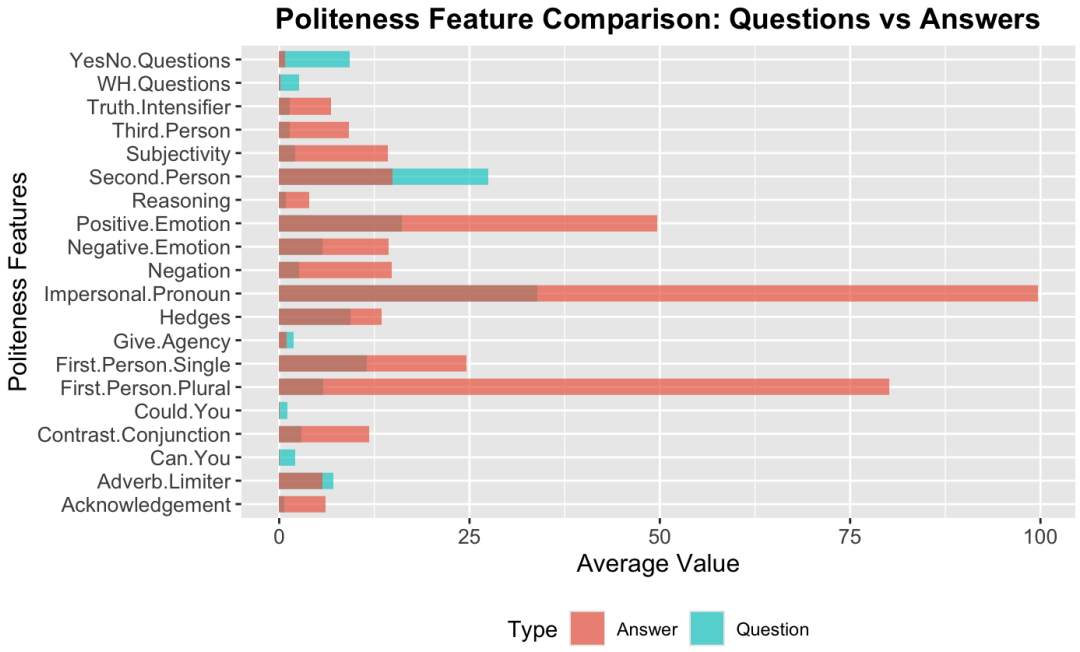

Figure 9: Politeness Feature Comparison: Questions vs Answers

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 or inquiry.

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.

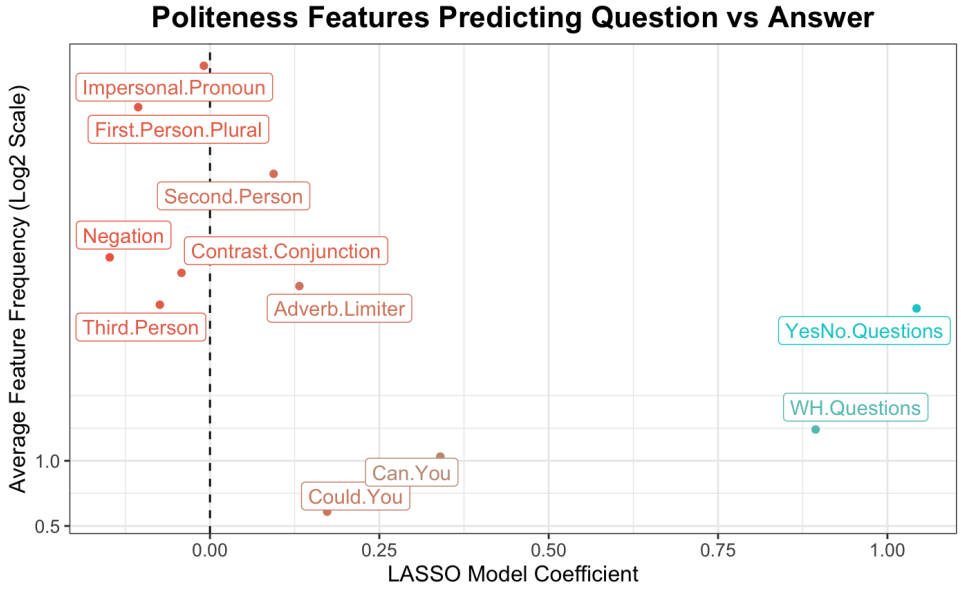

Figure 10: Politeness Features Predicting Question vs Answer

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 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 more frequent in 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

The model's accuracy for each quarter (1 through 4) has been calculated and displayed in a table. For Quarter 1, the accuracy is approximately 28.42%, for Quarter 2 it is 21.44%, for Quarter 3 it is 71.89%, and for Quarter 4 it is 19.04%.

The results show a high accuracy for Quarter 3 compared to the other quarters, with a significant drop in performance for the first, second, and fourth quarters. This suggests that the model is better at predicting Quarter 3, but struggles with the other quarters, particularly Quarter 4, where accuracy is the lowest.

The model's binary accuracy for distinguishing between Quarter 1 and other quarters is approximately 73.5%. This indicates that the model is correctly classifying whether a call is from the first quarter or not in about 73.5% of the cases.

This performance suggests that the model is somewhat effective at identifying calls from Quarter 1, but there is still room for improvement. While a binary accuracy of 73.5% is relatively strong, further model optimization, additional features, or a more complex model may help improve the accuracy and reduce errors, especially for the less well-predicted quarters.

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.