Text Analysis for Business

Final Essay

Ana Rodrigo de Pablo

CID: 02490419

PART A

Question 1

I start the assignment by splitting the data into a training set and test set. The test set must contain all calls during fiscal year 2012. In consequence, the training set contains all calls executed prior to 2012. The reason for splitting the data this way is to train the model on historical data (before 2012) and evaluate its performance on unseen data (during 2012). By training the model on historical data, it can learn patterns and relationships in the data, which can then be used to make predictions on the test data from 2012. This allows us to assess how well the model generalizes to new, unseen data and evaluate its predictive performance.

Question 2

I trained a LASSO model using bigrams and trigrams from the opening speeches as features to predict the reported earnings per share. The image below illustrates the trade-off between model complexity and prediction accuracy. At lower values of log(lambda), the model exhibits overfitting, capturing noise in the training data but performing poorly on new data. As log(lambda) increases, the model's complexity decreases (less features are added into the model), leading to a gradual reduction in MSE as the model generalizes better. However, beyond a certain threshold (from 1400 to 200 features in the model), increasing log(lambda) leads to underfitting, causing MSE to rise again as the model becomes too simplistic. The plot aids in identifying the optimal regularization strength that balances bias and variance, enabling accurate predictions of earnings per share based on opening speeches.

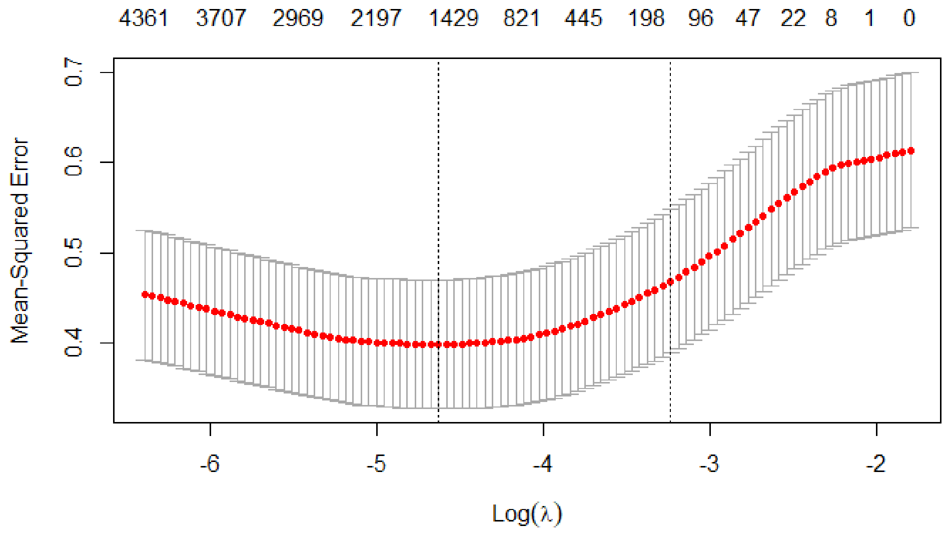

Figure 1: LASSO model using bigrams and trigrams

Figure 2 contains the coefficients of the ngram features that predict high and low earnings. We observe that bigrams or trigrams that contain the words “loss” are on the left side of the plot, hence associated with a negative coefficient and leading to a prediction of lower earnings per share. On the other hand, the right side of the plot contains words like “earn”, “billion”, or “growth”, which predict high earnings per share. The dollar sign “$”, although spread across all the plot is more present on the right side, hence associated with a positive coefficient. Finally, “insur_market” is the most significant ngram indicating higher earnings per share.

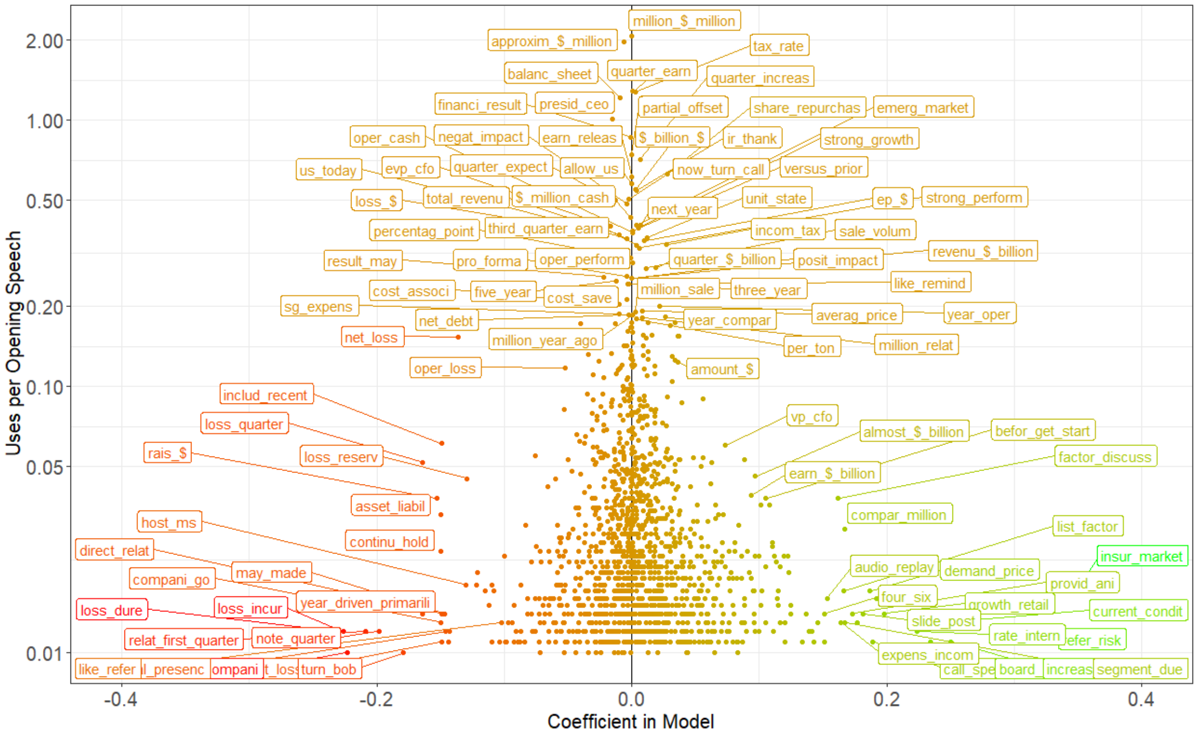

Figure 2: Coefficient plot for the LASSO model using bigrams and trigrams

Question 3

For this question I trained a second LASSO model using word2vec embeddings from the opening speeches as features, and a third model that combines these two feature sets. The table below contains the accuracy estimates of the three models.

Model	Accuracy	Lower Bound	Upper Bound
Ngrams	70.05	68.49	71.62
Vector Embeddings	66.86	65.25	68.47
Ngrams + Vectors	70.39	68.83	71.95

The accuracy results show the different mechanisms of models trained using ngrams vs vector embeddings. Ngrams overall perform better by themselves with the ability of capturing local context and specific word patterns in the text from the opening speeches, while vectors capture the semantic similarity between words. By combining both types of features, the third model can leverage the strengths of each feature type, improving the predictive performance to 70.39.

Furthermore, to avoid overfitting I used the "lambda.min" parameter when making predictions or extracting coefficients from the glmnet models. This parameter selection helps in controlling the model complexity and preventing overfitting by shrinking the less important coefficients towards zero.

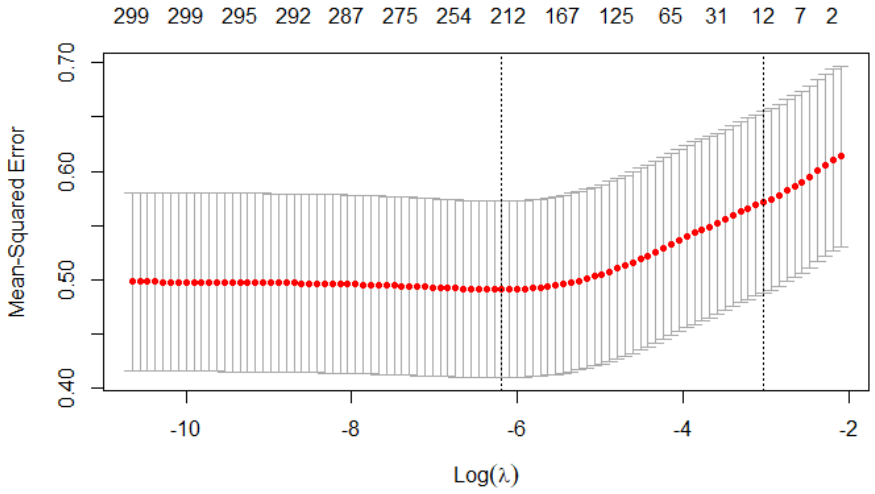

Figure 3: LASSO model using word2vec embeddings

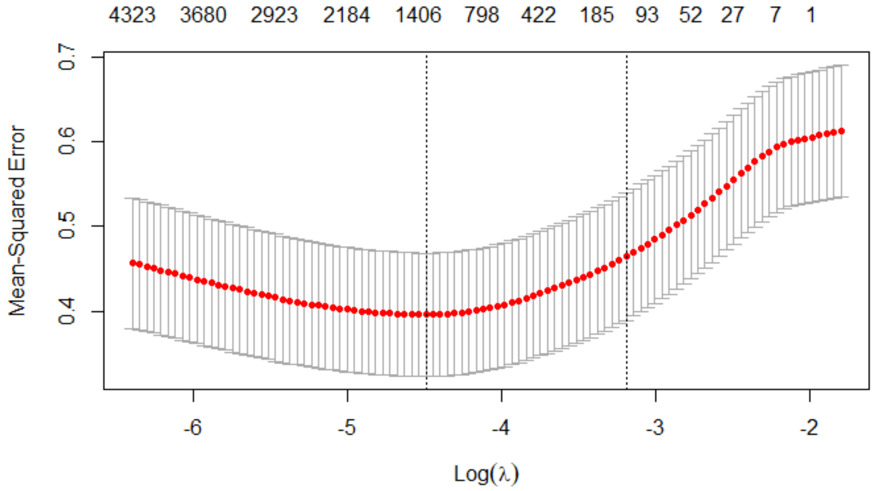

Figure 4: LASSO model using a combination of ngrams and vector embeddings

Question 4

For this question I created two benchmark models to compare against my trained models: A linear regression of wordcount on EPS, and a random guess benchmark. Figure 5 shows the average accuracy scores of the three initial models and the two benchmarks.

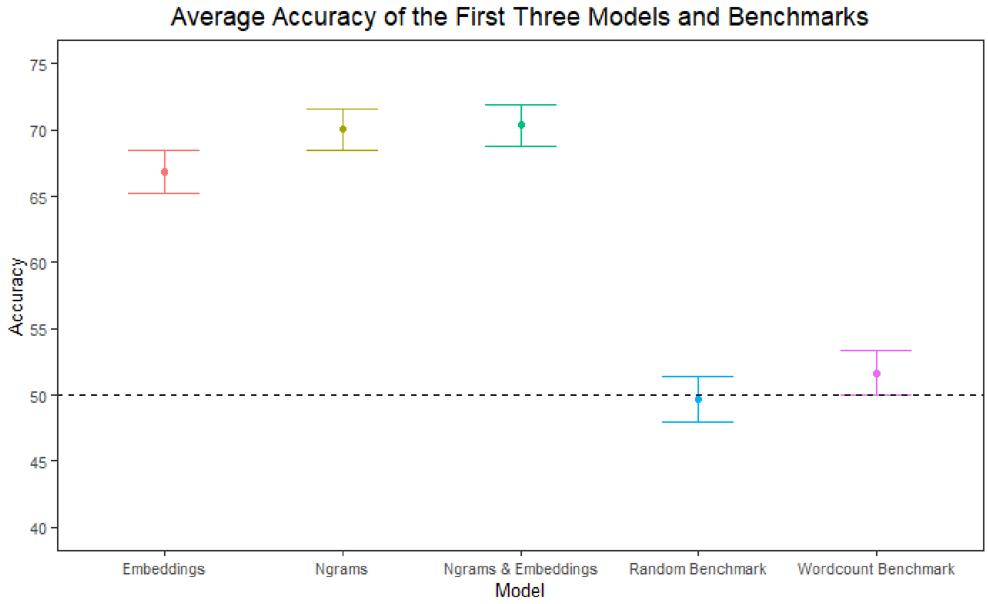

Figure 5: Comparison of accuracy scores of the five models

From the figure above we can see that the most accurate model in predicting earnings per share based on the narrative of the opening speeches, is the model trained using both ngrams and vector embeddings. The model captures the contextual information and relationships between words in the narrative through the ngrams feature, as well as the semantic meaning of the text through vector embeddings. By leveraging both types of features, the model can better understand the nuances and patterns within the narrative, leading to the best predictive performance.

In addition, the model trained on ngrams, although it lacks semantic meaning, also performs quite well and it is just slightly less accurate than the previous; whereas the model trained only on vector embeddings is less accurate, although still performs much better than both the benchmark models. This is because vector embeddings can capture the semantic similarity between words but fail in capturing local context.

In terms of the benchmarks, they are both quite simple models that do not consider the relationships between words or context of the narrative when making predictions, hence they have a very low accuracy. This demonstrates that studying and understanding semantic relationships and context of the words improves prediction power significantly compared to just guessing (random model) or counting the words per speech (wordcount model).

Question 5

The first example consists of a case where earnings per share (EPS) was high (higher than 4.5), while the second example consists of a low EPS (negative value). In both cases, vectors model offsets ngrams model in terms of predictive accuracy.

Opening Speech	**Values**
operator: greetings and welcome to the newmarket corporation third quarter 2012 financial results conference call. at this time, all participants are in a listen-only mode. a brief question-and-answer session will follow the formal presentation. as a reminder, this conference is being recorded.it is now my pleasure to introduce your host david fiorenza. thank you, sir. you may begin.david a. fiorenza - vp, treasurer and principal financial officer: good morning. thank you. thanks for joining us to discuss our third quarter performance. with me today is teddy gottwald. i have a few planned comments after which teddy has a few about the special dividend and then after that we'll be happy to take your questions.as a reminder, some of the comments we will make today are forward-looking statements within the meaning of the private securities litigation reform act of 1995. we believe we base our statements on reasonable expectations and assumptions within the bounds of what we know about our business and operations.however, we offer no assurance that actual results will not differ materially from our expectations due to uncertainties [...]	EPS Actual: 4.91 Ngrams Prediction: 0.60 Ngrams Error: 4.31 Vectors Prediction: 1.34 Vectors Error: 3.57
operator: good morning and welcome to the fourth quarter 2012 axis capital earnings conference call. all participants will be in listen-only mode. please note this event is being recorded.i would now like to turn the conference over to linda ventresca. please go ahead.linda ventresca - evp and corporate development officer: thank you, amy, and good morning ladies and gentlemen. i am happy to welcome you to our conference call to discuss the financial results for axis capital for the fourth quarter and the year ended december 31, 2012. our earnings press release and financial supplements were issued yesterday evening after the market closed. if you would like copies please visit the investor information section of our website at www.axiscapital.com.we've set aside an hour for today's call which is also available as an audio webcast through the investor information section of our website. with me on today's call are albert benchimol, our president and ceo, and joseph henry, our cfo [...]	EPS Actual: -0.23 Ngrams Prediction: 2.55 Ngrams Error: 2.78 Vectors Prediction: 0.56 Vectors Error: 0.79

After reading the examples in detail and going through few more cases, it seems that the vector model is likely to be more accurate in outliers (very high or low values of earning per share). The reason for this may be due to its ability to capture semantic similarity and underlying meaning of the text, which may be even more evident and insightful in outliers' speeches. For instance, the first example contains the following sentence: *"The company is optimistic about its future, with expectations to grow future volumes somewhat faster than the industry rate of 1% to 2% per year"*. If we consider the meaning of the text (as vectors model does) it is quite obvious that EPS will be high, whereas if we just focus just on bigrams and trigrams (as ngram model does) there are not that strong indicators that EPS will be high.

To improve this and close this gap in outliers, I suggest setting some benchmarks within the ngrams model such as a minimum document frequency, maximum features to optimize the performance, or experimenting with different parameters such as ngram range. All this could help us improve ngrams prediction in outlier cases.

Question 6

For this question I used the Distributed Dictionary Representation (DDR) technique to compute similarity of all the speeches to the Loughran & McDonald positive emotion dictionary. Moreover, I computed the traditional dictionary approach, again using the L&M positive emotion dictionary. The accuracy scores are shown in the table below.

Model	**Accuracy**	**Lower Bound**	**Upper Bound**
DDR	53.58	52.21	54.95
Traditional	52.44	51.07	53.81

We can see that both have a similar accuracy, with the DDR technique being slightly better. The main reason why the traditional dictionary may have a lower performance, is due to its lack of ability to understand context. In the traditional dictionary approach, I used the dfm_lookup function to match the words in the text against a predefined dictionary (in this case, the Loughran & McDonald dictionary), causing us to treat words as independent units. With this technique we end up with a model that relies solely on exact word matches within the dictionary and it does not account for the context in which the words appear or the semantic relationships between words. On the other hand, if we use a distributed dictionary representation approach, we can capture the semantic meaning of text by using pretrained vector word embeddings.

This is the main reason why overall the model using Distributed Dictionaries ends up having better results than the traditional one, as it is better at understanding slight nuances in the text.

Figure 6 compares these two accuracy scores with the other two benchmark models built previously.

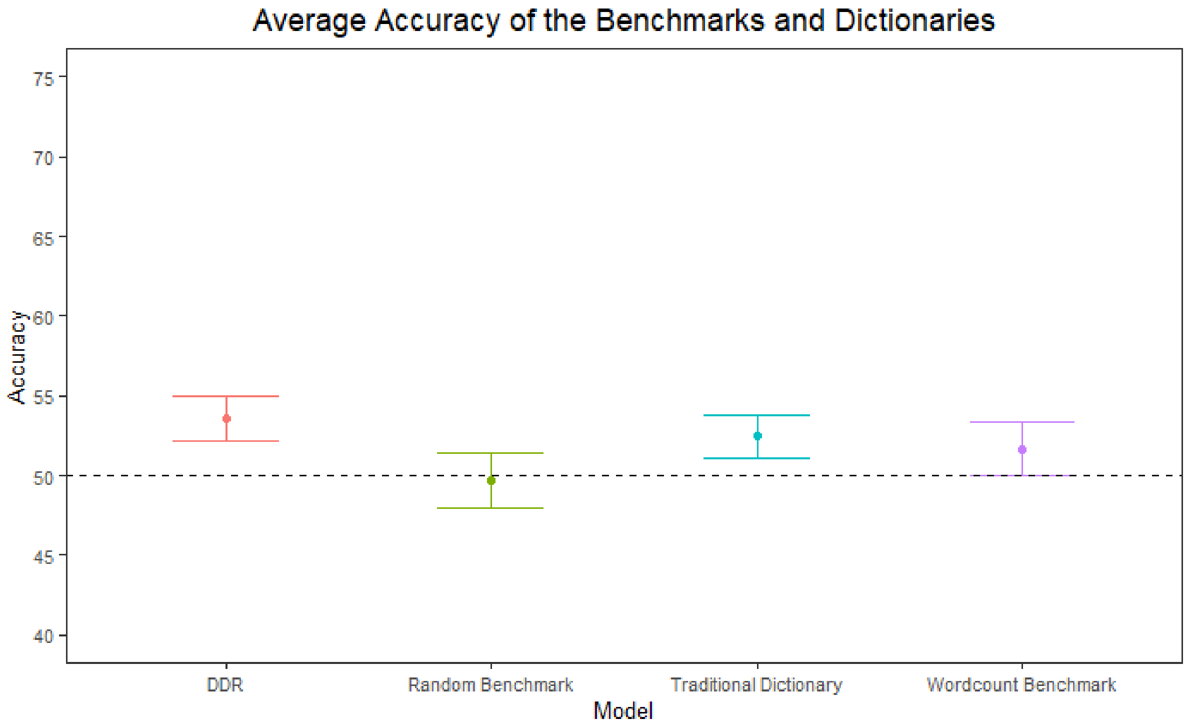

Figure 6: Comparison of accuracy scores of the dictionary and benchmark models

The dictionary models have an accuracy similar to the benchmark models. This is surprising as my expectation was that dictionary models accuracy would be closer to ngrams or vector embeddings rather than wordcount or random guess. The reason for this drop in accuracy could be associated to the dataset used for training and testing the models. The dataset is quite big (~11,000 opening speeches); however, the context and theme are similar among all opening speeches, hence the dataset lacks variability and diversity in terms of language use, topic coverage, or sentence structure. This may not fully challenge the capabilities of more advanced models like DDR and as a result, simpler models perform comparably well.

Question 7

After filtering for the 448 companies that have entries for all four quarters of both FY 2011 and FY 2012 (eight speeches total), I calculated the similarity of each speech to its matching first speech using word2vec. Then, I calculated the average similarity for each of the other seven quarters. Figure 7 highlights these results.

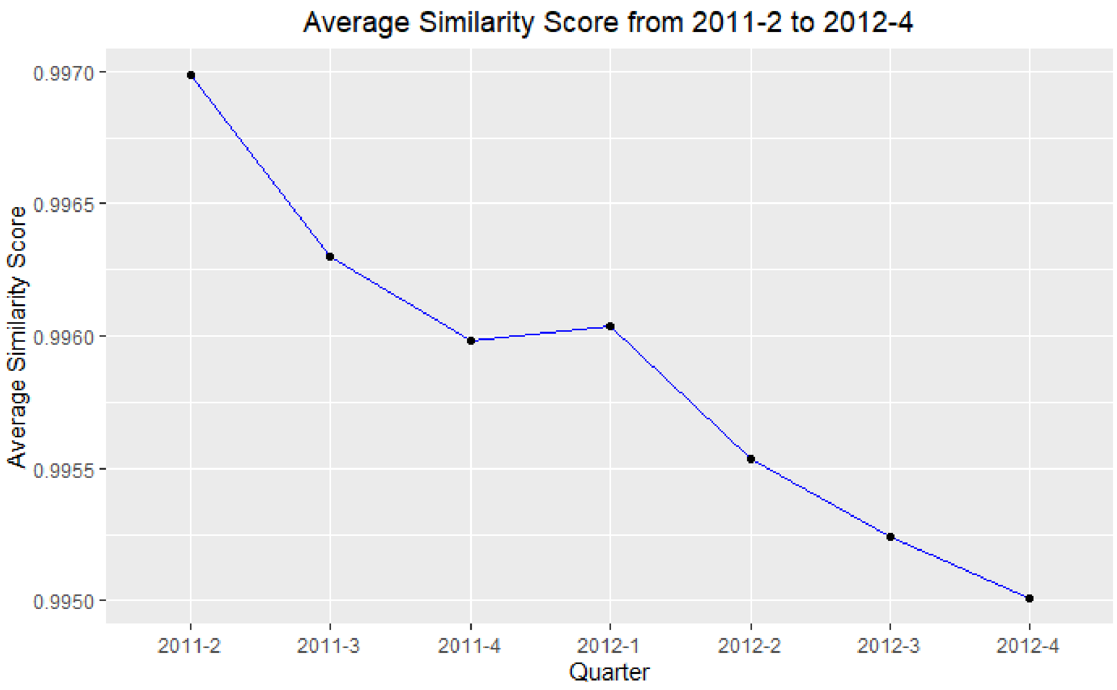

Figure 7: Average similarity score for the 448 companies of interest

We can see how similarity decreases over time. Nonetheless, there is a small increase in the first quarter of 2012, which may indicate that despite the decreasing trend, speeches during the first quarter of the year may have more things in common. We can interpret the decreasing trend in similarity over time as a reflection of companies adapting their messaging or focusing on different topics as time progresses. This could be due to changing market conditions, shifts in corporate strategy, or evolving investor concerns. The slight increase in similarity observed in the first quarter of 2012 may suggest that companies tend to revisit or emphasize certain key messages at the beginning of each year, possibly related to annual reports, strategic plans, or market outlooks.

Question 8

I will now start using the question-and-answer data. Figure 8 shows the relationship between the asker order (for the first 20 askers in each call) and the number of questions they ask. The blue line depicts a decreasing trend in terms of number of questions asked, however it is quite steady. We can see that from asker order 1 to 15, everyone asks on average the same amount of questions, whereas from asker order 15 to 20 the decrease seems to be a bit more evident.

Overall, there is not strong evidence to conclude that askers early in the call ask more questions than askers late in the call. This low variability or slow decreasing trend suggests that factors other than asker order may influence the number of questions asked, such as the complexity of the topics discussed (leading to everyone having lots of questions) or the level of engagement of the audience (if the audience is very engaged all 20 askers may have a lot of questions).

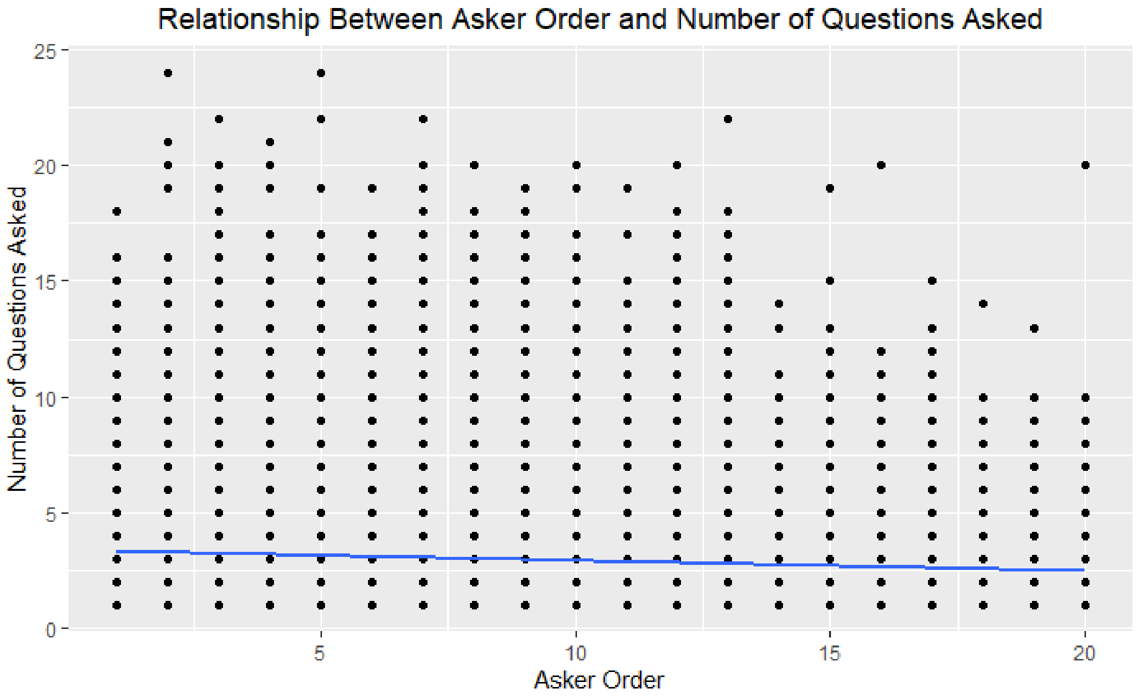

Figure 8: Relationship between the asker order and number of questions asked

Question 9

For this question I trained two new separate models: A model using the text of the first ten questions from each call, and another one with the first ten answers from each call. Figures 9 and 10 show the ngram coefficient plots (unigrams and bigrams) showing the features of the companies' Q&A that predict EPS_actual in each model.

In figure 9 we observe that the appearance of words like "procedure", "flood", or "prior_quarter" in the questions is associated with higher earnings per share, whereas "worth" or "first_want" are associated with lower earnings per share. On the other hand, figure 10 indicates that the presence of "worker" in the answers in associated with lower EPS, whereas "went_back", "thrill", or "grow_faster" indicate higher EPS.

Under both figures the information seems to be quite clustered in the middle, with no strong indicators of lower of higher earnings per share based on the words mentioned during the questions and answers (i.e. there are barely left or right outliers).

Something remarkable is that there seems to be a tendency of unigrams being most insightful than bigrams under both scenarios, with the former appearing more often on the extremes of the plot.

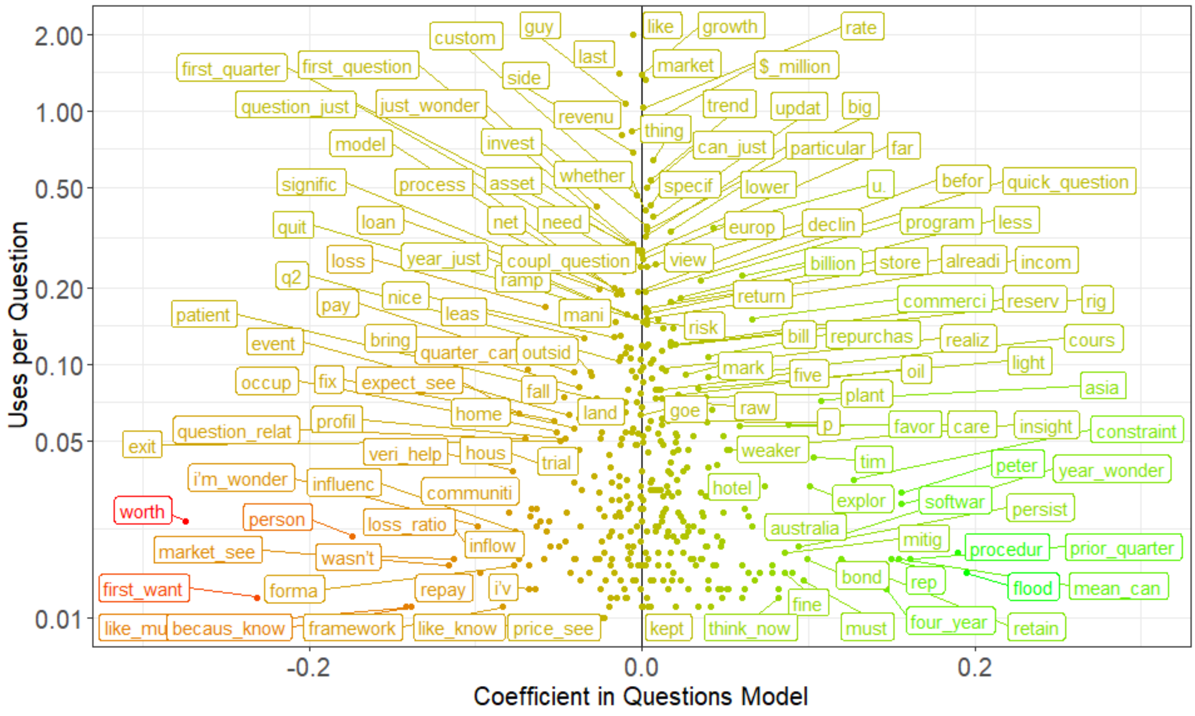

Figure 9: Coefficient plot for the first 10 questions of each call

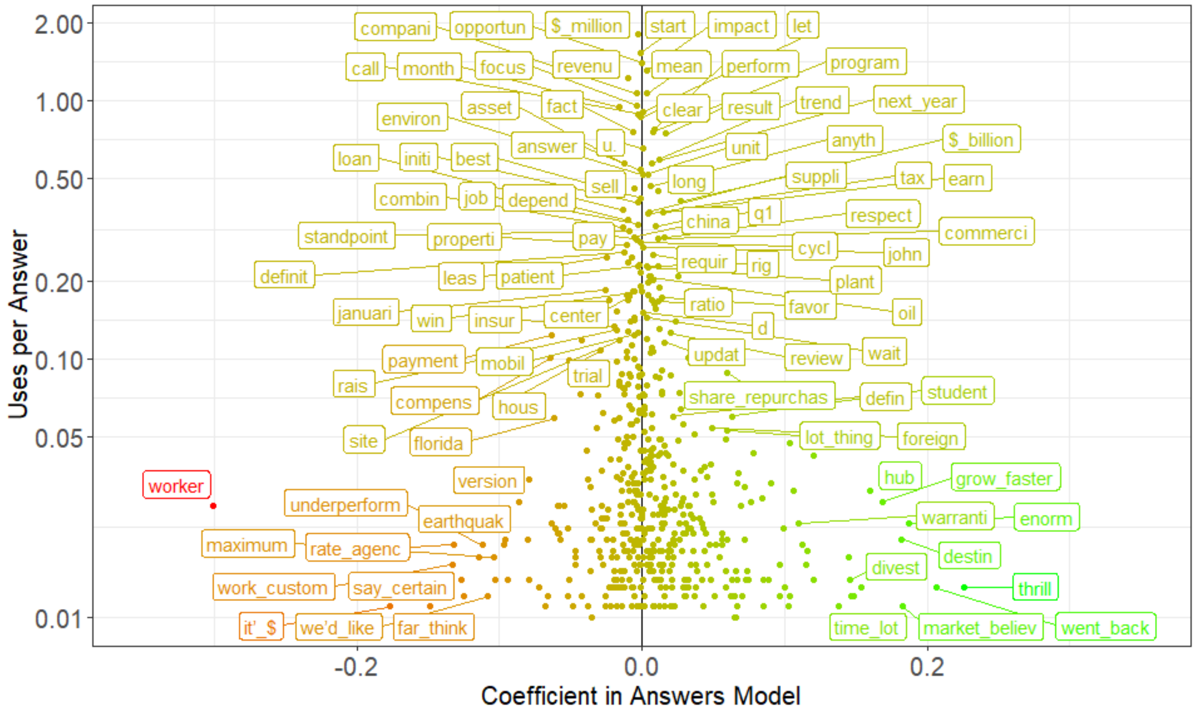

Figure 10: Coefficient plot for the first 10 answers of each call

Question 10

For this question I trained a LASSO model using bigrams and trigrams for the answers set. The objective was to understand if the accuracy improves when combining bigrams and trigrams (rather than unigrams and trigrams) as this causes our LASSO model to gather more contextual information from the narrative.

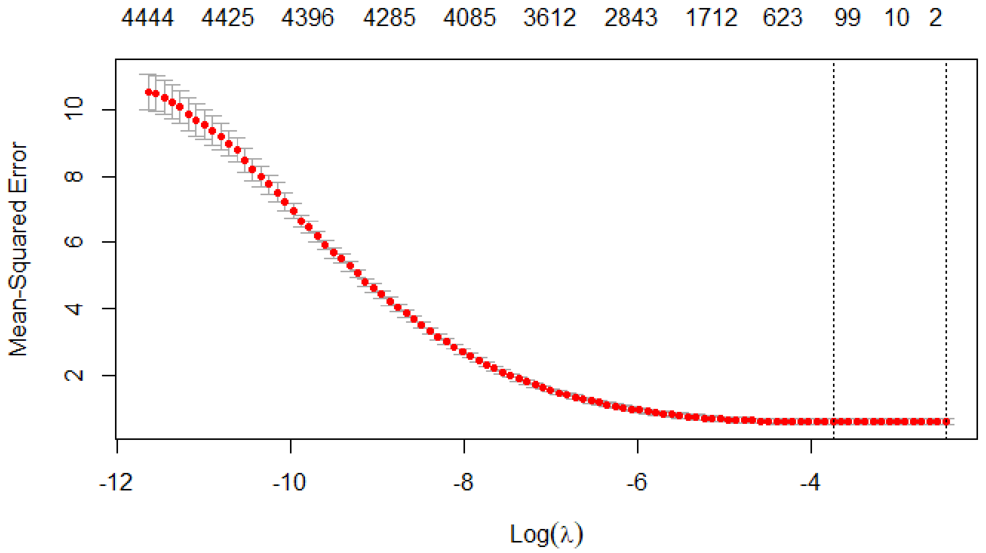

Figure 11: LASSO model using bigrams and trigrams from the answers set

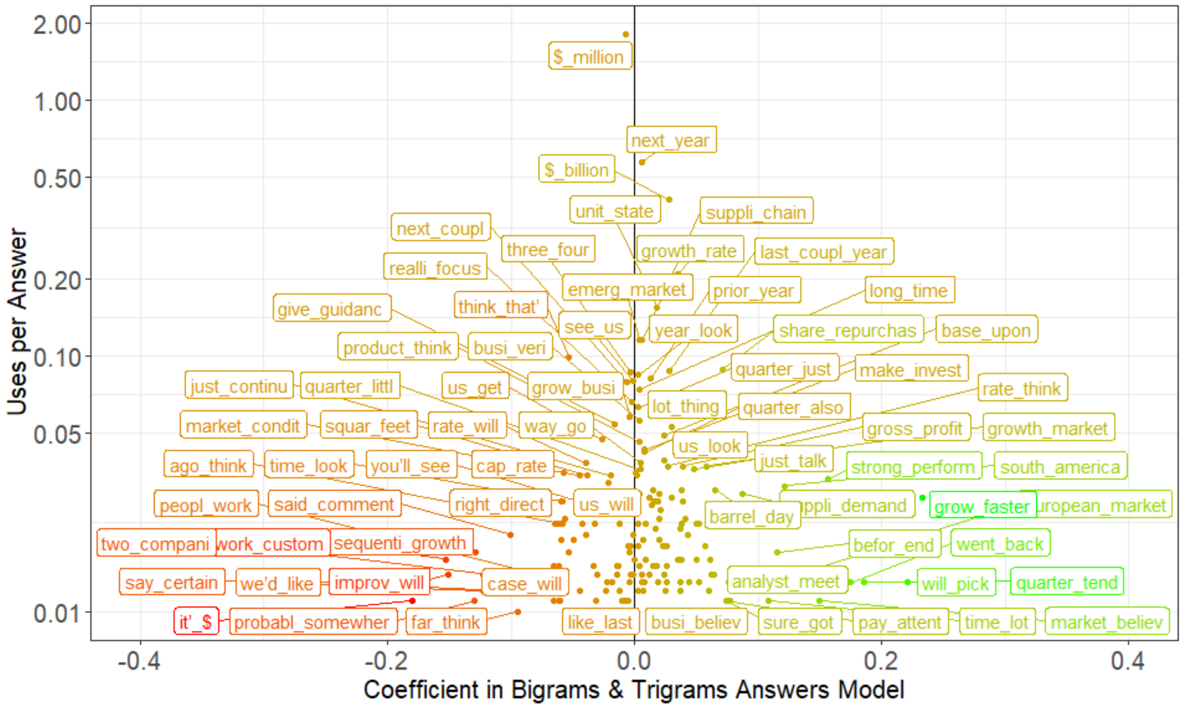

Figure 12: Coefficient plot for the first 10 answers using bigrams and trigrams

From figure 12 we can see that the plot is now less clustered and contains more ngrams that have stronger impact on EPS (left and right extremes). At a high level we could say that ngrams related with “work” tend to indicate lower earnings per share, whereas ngrams related with “grow” or “market” tend to indicate higher EPS.

However, when comparing the accuracy of the two models built in question 9 against the one just built, we observe a huge drop in predictive accuracy, meaning that the inclusion of trigrams in the model does not offset the removal of unigrams.

Model	Accuracy	Lower Bound	Upper Bound
Unigrams + Bigrams Questions	58.58	56.89	60.26
Unigrams + Bigrams Answers	59.51	57.83	61.18
Bigrams + Trigrams Answers	55.20	53.50	56.91

Unigrams and bigrams model for answers achieved a slightly higher accuracy than the same one for questions, with a score of 59.51% vs. 58.58%. This indicates that the combination of unigrams and bigrams in answers contributes to better predictive performance.

In addition, despite the inclusion of trigrams in the answers model, which theoretically should capture more nuanced language patterns, the accuracy decreased. The decrease suggests that trigrams alone may not be as informative or relevant for predicting earnings per share compared to a combination of unigrams and bigrams.

Overall, the results demonstrate that while including more complex language features such as trigrams may seem beneficial, it's crucial to strike a balance between complexity and predictive power. In this case, the models including only unigrams and bigrams outperformed the model with bigrams and trigrams, suggesting that simpler models can sometimes be more effective for this task.

My next action was to build three benchmark models I could compare the accuracy of the three above models with. The benchmark models built were: A linear regression of the word count of the first ten questions on EPS, a linear regression of the word count of the first ten answers on EPS, and a random guess model.

Figure 13 shows a comparison of the average accuracy across all six models. We can see that the ngram models are the most accurate, wordcount models have a drop in accuracy, and finally random guess has an accuracy below 50%. The reason for these disparities in accuracy is because wordcount models capture only the crude relationship between the length of the text and EPS, without considering the specific language used. The drop in accuracy compared to the ngram models suggests that the linguistic content of the text provides more predictive power than just its length. Moreover, the random guess model serves as a baseline for comparison, where predictions are made randomly without considering any information from the text. As expected, the random guess model performs the worst among all models, as it does not consider any text or contextual information.

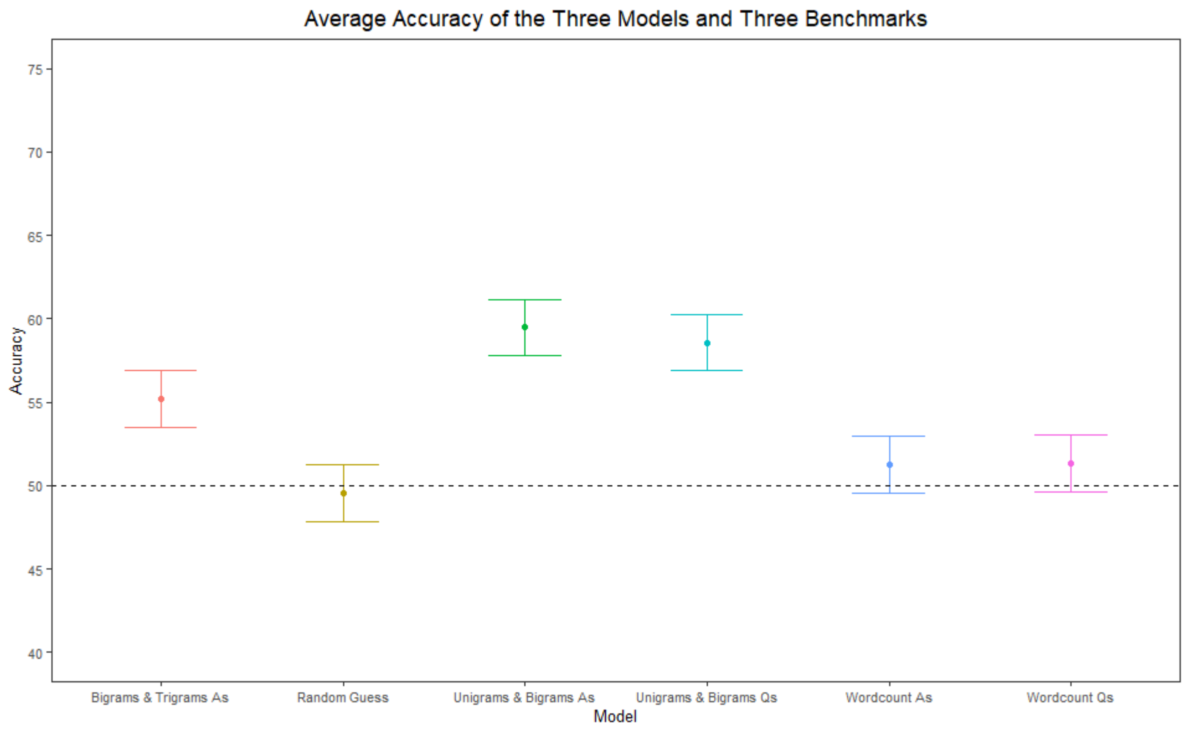

Figure 13: Comparison of average accuracy scores across the six models

Question 11

I ran the politeness package from Spacy on our dataset and gained further insights into the linguistic patterns in the data. The politeness plot (figure 14) shows the feature difference in the questions and answers text. My conclusions are the following:

The prevalence of first-person plural and impersonal pronouns in the questions suggests that questions are more likely to be formulated from a general or group perspective. Typically, when a question is formulated this way, it may not ask or provide detailed information about the questioner themselves, but rather aim to gain insights from the broader market outlook. The use of "you" or "we" highlights the impersonal tone of the question, indicating that the questioner may be addressing a broader audience or seeking information that applies universally.

On the other hand, the answers are characterized by the use of impersonal pronouns and second-person language. This indicates that the company answering the questions tends to address the whole audience rather than focusing on individual questioners. This approach creates a sense of inclusivity and engages the broader audience in the conversation.

Furthermore, questions tend to exhibit more positive emotions and subjectivity. This may indicate that questioners are acknowledging the company's efforts to address their concerns or are expressing optimism about the company's performance and future prospects. It is in the nature of questions to display subjectivity, as questioners often seek clarification or express opinions rather than providing objective information.

I also noted the usage of yes/no questions, wh-questions (who, what, where), and hedges (indicators of uncertainty) in the answers' narratives. Answerers may use these to ask questions to the audience or follow up on the askers' questions to show interest and engage with the topic further. This can help the company understand the audience sentiments and dive deeper into the topic, ensuring that they address any concerns or uncertainties effectively.

Finally, I observed the presence of adverb limiters such as "just," "only," and "simply", as well as negations. These linguistic features are often used to minimize statements or emphasize limitations, potentially indicating caution or modesty in the company's responses. By acknowledging limitations or constraints, the company may seek to manage expectations or convey transparency about their actions and decisions.

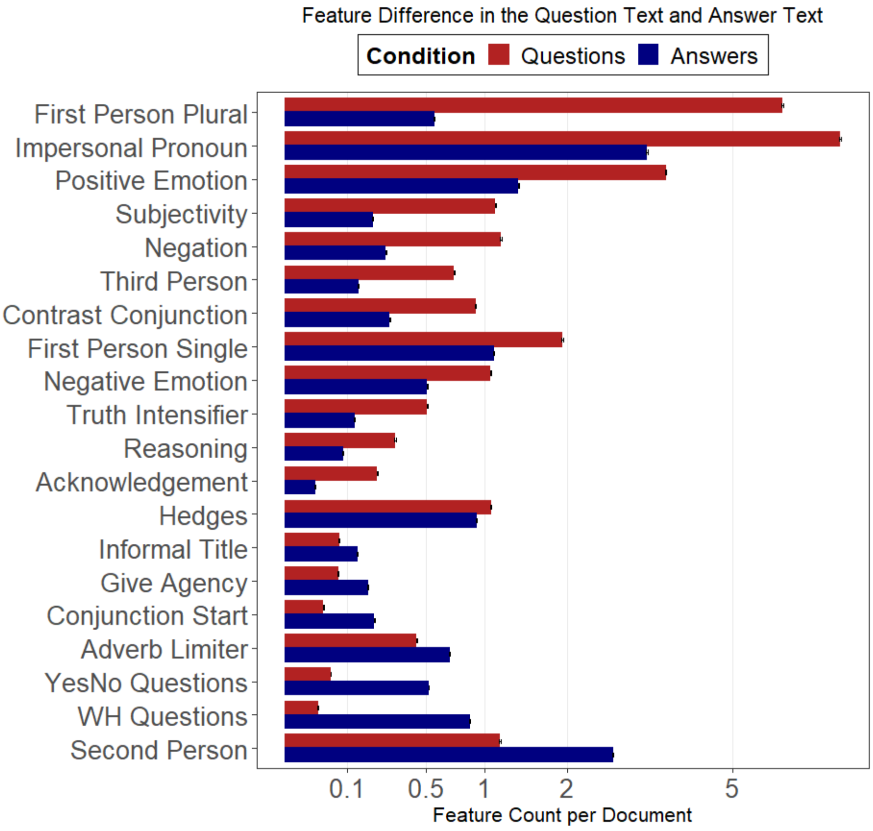

Figure 14: Politeness plot

Overall, these linguistic patterns reflect the dynamics of quarterly earnings calls, where questioners seek insights from a broader market perspective while companies aim to engage with and address the concerns of their entire audience. This can influence the tone and perception of the communication, potentially impacting investors' interpretations of the company's performance and strategy.

Following this, I trained a LASSO model to predict whether a turn is a question or an answer using the politeness features. Its corresponding coefficient plot is shown in figure 15. The results are aligned with the politeness plot, with the bigger indicators of the type of turn being “WH questions”, “Yes/No questions” and “Please”. The rest of the features remain close to the 0.0 coefficient, indicating that they do not shed much light in understanding whether a turn is a question or an answer.

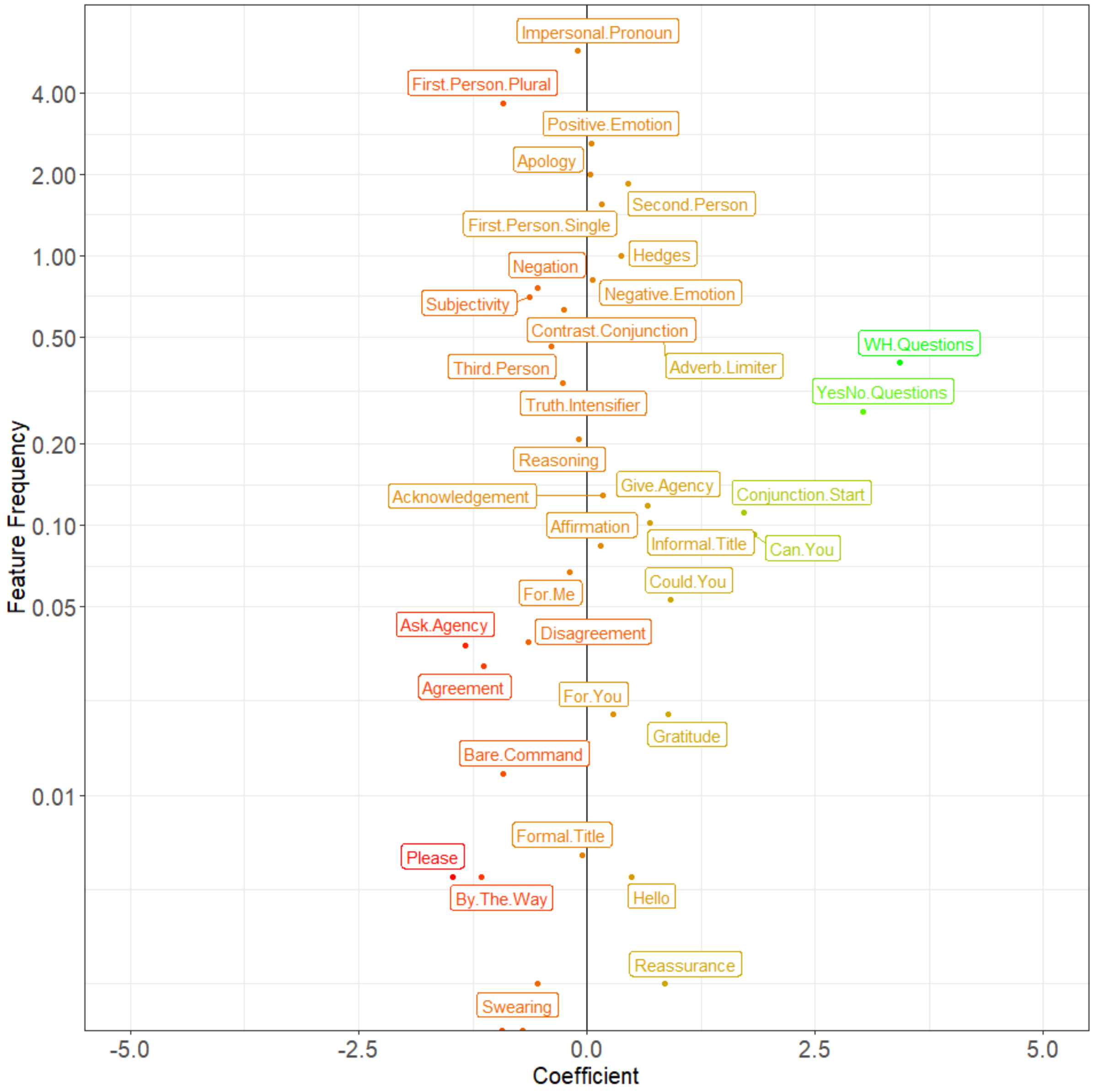

Figure 15: Coefficient plot indicating whether a turn is a question or an answer using the politeness features

Question 12

To understand the difference between the answers that companies give during their first quarter calls, compared to calls during other quarters, I first pre-processed the data and then trained a LASSO model using unigram and bigram features. The resulting coefficient plot is shown in figure 16.

The right extreme contains ngrams such as “first_quarter”, “q1”, or “march” which indicate that they are taking about the current quarter. The left extreme contains ngrams such as “next_year”, “third_quarter”, or “q3”. These findings indicate that during first quarter calls, companies tend to focus more on the current quarter, as reflected by the presence of the terms previously mentioned.

The above suggests a temporal pattern in the language used during earnings calls, with companies adjusting their focus and discourse based on the current or upcoming financial periods. Such insights can be valuable for understanding how companies communicate their performance and outlook to investors and stakeholders over time.

In addition, all other terms such as “impact”, “demand”, “strateg”, etc. which used to be indicators of earnings per share before, now lay on the middle of the plot, not providing major significance in identifying calls executed during the first quarter.

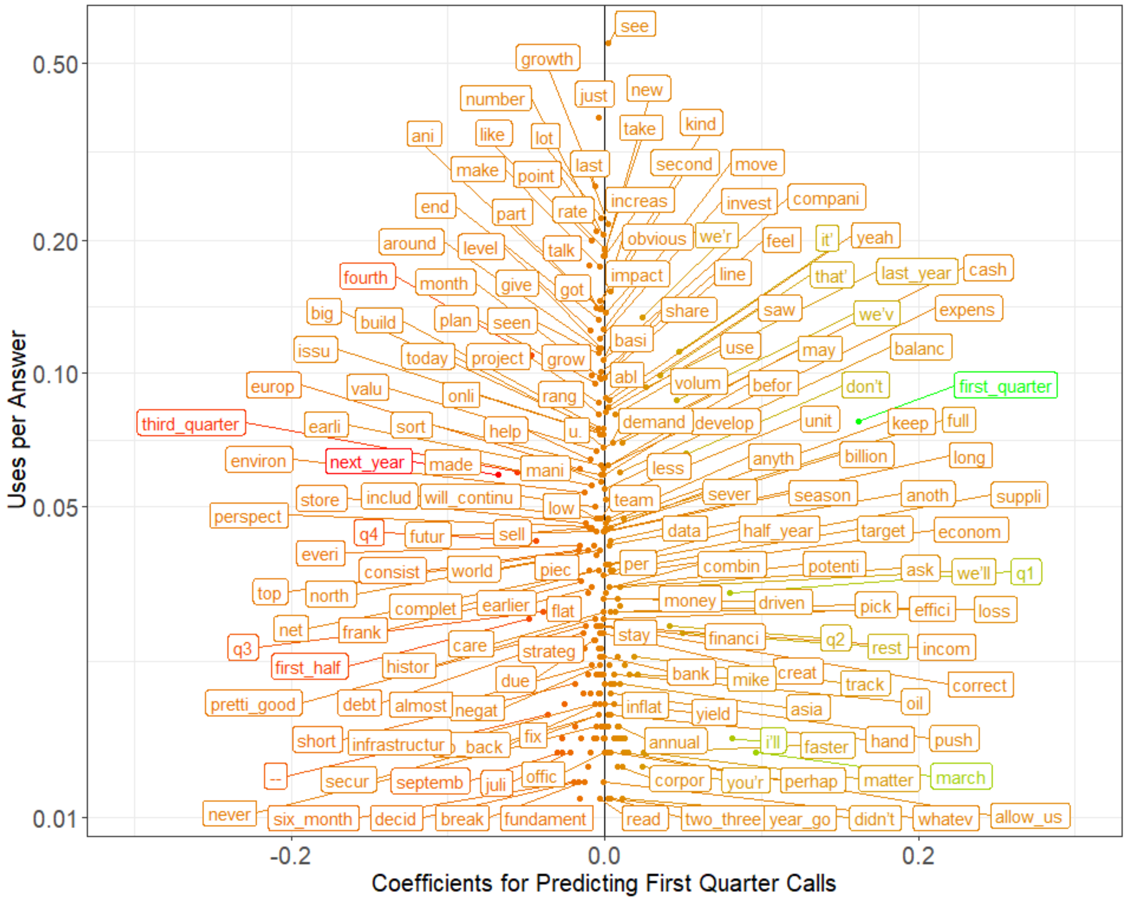

Figure 16: Coefficient plot for the first five questions of first quarter calls

Question 13

I first trained a multinomial model to predict which of the four quarters a call comes from. The confusion matrix to show the accuracy across all four quarters is shown below.

		Predicted			
		Quarter 1	Quarter 2	Quarter 3	Quarter 4
Actuals	Quarter 1	681	348	867	3264
	Quarter 2	448	678	1246	2852
	Quarter 3	211	227	1698	2918
	Quarter 4	329	173	921	3858

The diagonal elements represent the number of correct predictions for each quarter. For example, in Quarter 1, the model predicted correctly 681 calls, while in Quarter 2, it predicted 678 calls correctly.

The off-diagonal elements represent the errors made by the model. For instance, it incorrectly classified 448 calls from Quarter 2 as Quarter 1, and 348 calls from Quarter 1 as Quarter 2.

The most common errors appear to be misclassifications of Quarter 4, as the model overpredicts Quarter 4 when in reality the calls belong to Quarter 1, 2, or 3. This could be because the narrative contains words like "last" or "year" that suggest it is the last quarter or end of the year and hence are predicted as Quarter 4, but in reality they may be referring about something else.

I then used the multinomial output to generate a binary result that classified a call from the first quarter, or any other quarter. To achieve this, I considered the probabilities predicted by the multinomial model for each quarter. If the highest probability corresponds to the first quarter, I classified the call as from the first quarter. Otherwise, I classified it as from any other quarter.

Finally, I calculated the accuracy of the models in question 12 & 13 in predicting the binary outcome (i.e. first vs. all other quarters). The accuracy results are plotted in figure 17. The multinomial model has an accuracy of 55.34% while the LASSO model has a slightly higher accuracy of 55.89%. The similarity in accuracies suggests that both models are performing similarly well in distinguishing calls from the first quarter from those from other quarters. This indicates that the features used for training the models are effective in capturing the differences between calls from different quarters.

Nonetheless, the accuracy for both models is around 55%, which is not that high considering it would be close to a random guess benchmark. We could increase the accuracy of both models it by analysing and including additional features that may better capture the differences between calls from the first quarter and those from other quarters. This could include sentiment analysis, topic modelling or topic correlation.

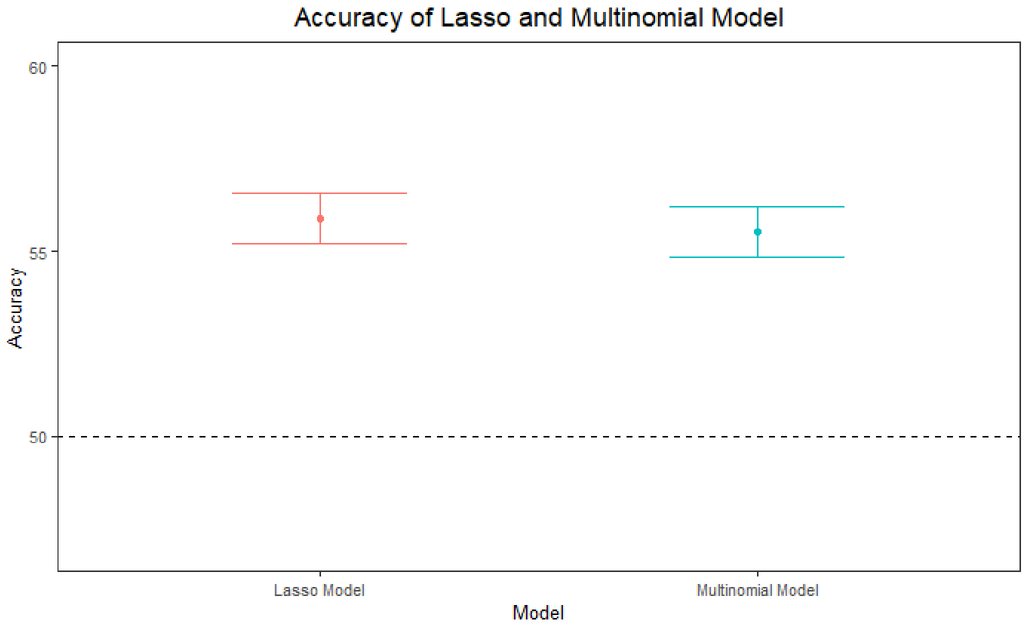

Figure 17: Accuracy of LASSO and multinomial models

PART B

I began by calculating the EPS surprise, which measures the percentage change in earnings from before to after the announcement. To ensure accuracy, I divided the difference between EPS_actual and EPS_consens by the absolute value of EPS_actual, considering that EPS can be negative.

After this, I split the data into positive and negative EPS surprise groups, I further divided each group into training (FY prior to 2012) and test sets (FY 2012), and trained LASSO models using unigrams. I used unigrams only as we saw during part A that this is usually the most accurate and insightful ngram for this dataset. My aim is to compare the speeches of companies with decreasing EPS against those with increasing EPS, so that companies can adapt their speech to the economic outlook and top industries.

Using labelTopics, I identified the top topics and their proportions in each group. Then, I explored what each topic was about and came up with labels to describe the 7 topics. The labeled topics in figures 1B and 2B provide insight into the main themes discussed.

For companies with a positive EPS surprise, the dominant topic was Automotive, indicating that these companies managed to increase their EPS compared to the initial forecast. Conversely, for those with a negative EPS surprise, the primary topic was Financial, suggesting that in this industry their actual EPS was lower than predicted.

Some topics, such as Banking, Consulting, Oil & Gas, and Healthcare, appeared in both figures, indicating that their prominence was independent of EPS performance. My assumption is that these topics were top topics, and depending on the company, the EPS could have been lower or higher than forecasted. However, it's difficult to identify a general trend across all companies for these four industries.

Let’s now bring our attention to the “Real Estate” topic in figure 2B. It is not surprising to find that topic there as we are working with speeches from 2010 to 2012, just 2 to 4 years after real estate bubble of 2008. The economic effects were likely still being felt, and earnings for real estate companies may have been lower than expected due to ongoing economic challenges and market uncertainties.

To explore my hypothesis I created a word cloud of the words in the Real Estate topic (figure 3B). The cloud confirms our assumptions, showing not only words directly related to the real estate sector (“portfolio”, “square”, “feet”, “acquisition”, ...), but also words reflecting negative sentiment and instability (“cost”, “expense”, “interest”, or “end”).

In conclusion, the analysis of EPS surprise reveals insights into the dynamics of company expectations, real earnings per share, and their corresponding speeches. While Automotive or Technology industries dominate positive EPS surprise; Financial, Healthcare or Banking topics prevail in negative surprise cases. Finally, we studied the use case of Real Estate companies and its effects from the 2008 bubble.

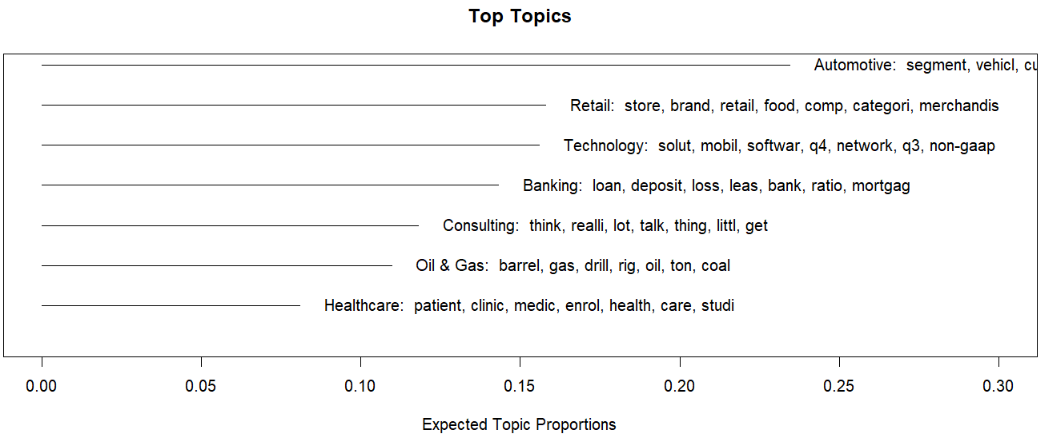

Figure 1B: Seven-topic model for positive EPS_surprise

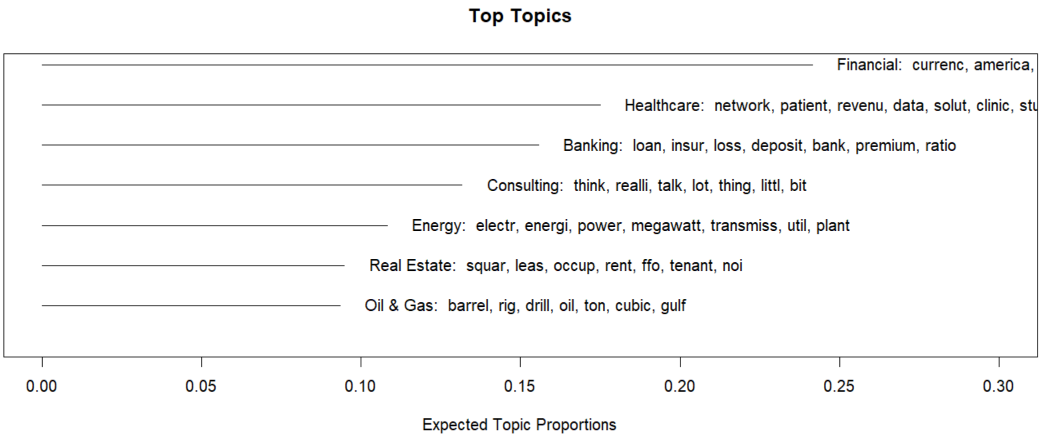

Figure 2B: Seven-topic model for negative EPS_surprise

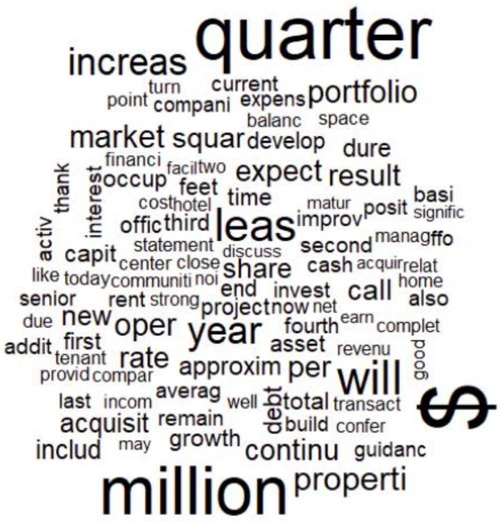

Figure 3B: Word cloud for “Real Estate” topic within the negative EPS_surprise model