Whispers of Wall Street: Predicting Stock Reactions with BoW and LLMs

NLP For Finance

Written by:

Emil William Hansen (102058)

Supervised by:

Hamid Boustanifar[1] & Sasan Mansouri[2]

Comments:

A complete replication code and additional information can be found here.

[1] EDHEC Business School
[2] University of Groningen

1 Introduction

In today's financial markets, the story is not told by numbers alone; Textual sentiment plays an equally critical role. This report investigates whether sentiment in earnings calls, extracted via BoW and LLM methods, predicts stock market reactions, hypothesizing that LLMs capture nuanced sentiment better than BoW due to their contextual understanding. Specifically, I examine whether positive or negative sentiment in earnings calls influences investor behavior. Using Bag-of-Words (BoW) and Large-Language-models (LLM) and S&P 500 earnings call transcripts from 2015 to 2021, including Q&A interactions and CompStat data, my objective is to assess the market impact of managerial tone and language.

2 Methodology

To investigate the relationship in question, two primary methods are employed. Initially, I used a Bag-of-Words (BoW) approach. In this method, I gather the words marked as positive and negative, determine their respective difference, and then normalize this by dividing by the presentation's length. I employ three different dictionaries: the Harvard General Inquirer (GI) dictionary, the Loughran McDonald 2011 (LM) dictionary, and a custom one developed by Grok 3 (CU)[3]. I also consider negations, applying a threshold of three words. This means that if a negation occurs within three words before the target word, it reverses its polarity from positive to negative or vice versa.

Next, I assess two types of models within the realm of Large Language Models (LLMs), namely BERT and GPT. For BERT, I employ three submodels: financialbert, distilroberta, and financial-roberta, each pre-trained for financial sentiment analysis[4]. In the case of BERT models, I process sections of each presentation to capture a ranking, subsequently averaging these across each presentation to compute a sentiment score. Additionally, I use OpenAI's gpt-3.5-turbo, instructing it to act as a financial sentiment analysis expert and assign rankings to each presentation. The initial aim was to utilize this for presentations, questions, and answers and to incorporate the earning surprise into the model, as this is known beforehand. However, due to limited (empty) API credits, it could not be highlighted in the report, despite being added within the code. Analyzing a smaller data set, I observed notable sentiment discrepancies regarding presentations and questions from analysts, a dynamic may

[3]For comprehensive information on the dictionaries, visit: GI, LM, CU

[4]Further information on these models can be found at: financialbert, distilroberta, financial-roberta

worth exploring further (with more GPU and API credits).

I proceed by performing regressions using these sentiment scores individually to determine if they possess any explanatory power. For control variables within the regressions, I include the scaled earnings surprise, SurpDec, and SurpDec squared, along with the net income for the specified quarter (appropriately scaled), and also the count of analysts who have changed their stock recommendations, both positively and negatively. In addition, I divide each sentiment score into two, depending on whether the monetary surprise is positive or negative. Moreover, I integrate time-fixed effects using both years and quarters, in addition to company-fixed effects. All regression variations employing different sentiment values are presented in Section 5 of my code, but due to space constraints, they are not detailed here. Instead, I only incorporate regressions without fixed effects, as well as those with fixed effects by year and company[5]. In every regression conducted, the dependent variable is the cumulative abnormal return from one day before up to one day after the earnings call, adjusted using the Fama-French three-factor model plus a momentum factor.

3 Results

Initially, Appendix A.1 displays the correlations among my sentiment scores and other variables. The observed correlation for the BoW models is lower than expected, highlighting the output's sensitivity to the input. Introducing negations into the BoW models results in negligible changes in correlation. Concerning my BERT models, they show notable correlations with the BoW models, some reaching up to 0.72, suggesting that these models can capture similar text relationships. Moreover, a high correlation is observed within the BERT models themselves, indicating a similar operational manner. Interestingly, the GPT model shows almost no correlation with the other models. Worth mentioning is that the GPT model processes the entire presentation, whereas the BERT models handle chunks, averaging scores. One might argue that fragment averaging is superior, as presentations tend to be generally positive, making it difficult to detect negative aspects when examining entire content[6]. Furthermore, the second correlation matrix reveals that my scores and variables have minimal correlation with each other.

Consulting Appendix A.2, it's apparent that BoW sentiment scores generally fail to yield a notable correlation. However, dividing sentiments based on whether the earnings surprise is positive or negative

[5]Employing either year or quarter for time-fixed effects yields the same results, since most of the explanatory power is derived from the company-fixed effects. This can be seen in the code, where I test both.

[6]I refrained from this approach due to credit limitations, fearing running out.

enhances the relationship, and the addition of negations slightly strengthens the outcomes. Introducing controls for fixed effects leads to more noteworthy results, although none reach the 5% significance threshold; some are nearly significant with negative earnings surprises. The inclusion of fixed effects captures much of the data variability, tackling omitted variable bias, and achieves the highest R^2 at around 13.3%. When both fixed effects and sentiment division are applied, the GI and CU dictionaries produce the greatest R^2, beating the LM dictionary by 0.1%. Regarding the influence of sentiment based on BoW models, the coefficient signs exhibit variability, showing both positive and negative values, which does not provide sufficient evidence for any conclusive interpretation.

In further examination of LLMs, Appendix A.3 reveals encouraging findings even before accounting for fixed effects. I find that two sentiment scores, FinancialBERT and Financial-RoBERTa, are significant at the level 10%, while DistilRoBERTa reaches significance at the level 1%. In particular, DistilRoBERTa records the highest R^2 among models that do not include fixed effects. Interestingly, while aggregated sentiment shows significant outcomes, dividing it according to the direction of earnings surprises removes significance. This implies that the market reacts to the overall tone of earnings calls rather than to the sentiment linked specifically to positive or negative news. The ability of LLMs to discern this comprehensive effect illustrates their advantage in identifying nuanced sentiment patterns overlooked by simpler models. Incorporating fixed effects without splitting sentiment still results in a highly significant outcome at the 1% level in my primary model, indicating that the observed relationship is not solely due to omitted company or time factors. However, segmenting sentiment while controlling for fixed effects does not result in a significant link, as expected. Generally, across LLMs, sentiment coefficients are negative, suggesting that the market frowns upon excessively positive earnings presentations, preferring honest over inflated statements. The persistent negative sentiment coefficients in LLM regressions imply that investors may be skeptical of overly optimistic managerial tones, possibly seeing them as efforts to obscure weaknesses rather than convey real confidence. The outstanding performance of DistilRoBERTa is probably due to its distilled architecture, which balances computational efficiency and contextual awareness, allowing it to detect subtle changes in sentiment that BoW and even other LLMs overlook. This underscores the importance of pre-trained, domain-specific models in financial NLP, where subtle language influences market behavior.

In conclusion, Bag of Words (BoW) sentiment analysis does not yield significant results due to its inability to capture intricate data relationships. The BoW models are highly sensitive to input and determining which words elicit positive or negative reactions. Generally, BoW performs slightly better

when earnings fall short of expectations, although this relationship lacks significance. Conversely, Large Language Models (LLMs) outperform BoW approaches, with the pre-trained BERT model proving to be superior in terms of significance. An important consideration is that chunking the input for GPT can enhance significance, suggesting a potential future strategy. Additionally, examining the sentiment of analysts' questions may also prove beneficial. Furthermore, observations indicate that the market tends to penalize overly optimistic presentations, favoring an honest delivery instead. Ultimately, a well-prepared pre-trained BERT model emerges as the most effective tool for sentiment analysis of earnings call presentations.

A Appendix

A.1 Correlation Matrices

A.1.1 Correlation Matrix of Sentiment Scores

Table 1: Correlation Matrix of Sentiment Scores (BoW vs. LLM Models)

	BoW-LM	**BoW-GI**	**BoW-CU**	**BoW-LM-Neg**	**BoW-GI-Neg**	**BoW-CU-Neg**	**FinBERT**	**DistilRoBERTa**	**Fin-RoBERTa**	**GPT-3.5**
BoW-LM	**1.00**	0.39	0.56	0.99	0.39	0.55	0.29	0.32	0.41	0.11
BoW-GI	0.39	**1.00**	0.35	0.39	**1.00**	0.35	0.18	0.19	0.25	0.06
BoW-CU	0.56	0.35	**1.00**	0.56	0.35	0.99	0.28	0.29	0.40	0.11
BoW-LM-Neg	0.99	0.39	0.56	**1.00**	0.39	0.56	0.29	0.31	0.41	0.11
BoW-GI-Neg	0.39	**1.00**	0.35	0.39	**1.00**	0.35	0.18	0.19	0.25	0.06
BoW-CU-Neg	0.55	0.35	0.99	0.56	0.35	**1.00**	0.27	0.28	0.39	0.11
FinBERT	0.29	0.18	0.28	0.29	0.18	0.27	**1.00**	0.61	0.65	0.10
DistilRoBERTa	0.32	0.19	0.29	0.31	0.19	0.28	0.61	**1.00**	0.72	0.13
Fin-RoBERTa	0.41	0.25	0.40	0.41	0.25	0.39	0.65	0.72	**1.00**	0.14
GPT-3.5	0.11	0.06	0.11	0.11	0.06	0.11	0.10	0.13	0.14	**1.00**

Notes: This table reports pairwise Pearson correlations between sentiment scores extracted from earnings call transcripts using Bag-of-Words (BoW) and Large Language Models (LLMs). BoW models include Loughran-McDonald (LM), Harvard GI (GI), and Custom (CU) dictionaries. Suffix "-Neg" denotes negation handling within a three-word threshold. LLMs include FinBERT, DistilRoBERTa, Fin-RoBERTa, and GPT-3.5. Correlations are rounded to two decimal places.

Table 1 illustrates the correlation among my sentiment scores. Notably, in the Bag-of-Words (BoW) methods, the correlation varies significantly based on the choice of words, highlighting the models' sensitivity to minor input variations. Introducing negations does not substantially affect the correlation, which remains above 0.99. The BERT models exhibit strong intercorrelation, yet exhibit weaker correlation with the BoW approach. This suggests that while BERT shares some similarities with BoW, it likely captures broader concepts by analyzing entire sentences/chunks instead of merely tallying individual words. However, the GPT model shows very low correlations across all these methods, which is intriguing. Higher correlations might be expected, but this is not the case. This could be due to the model processing entire presentations, which tend to yield an overall positive sentiment score. Such results might question the current methodology and suggest that segmenting presentations could be a more effective research strategy.

A.1.2 Correlation Matrix of Sentiment Scores including Control Variables

Table 2: Correlation Matrix of Sentiment Scores (BoW vs. LLM Models) with Additional Variables

	BoW-LM	BoW-GI	BoW-CU	BoW-LM-Neg	BoW-GI-Neg	BoW-CU-Neg	FinBERT	DistilRoBERTa	Fin-RoBERTa	GPT-3.5	Earnings Surprise	Earnings Surprise^2	Net Income	NUMUP	NUMDOWN	CAR-11-Carhart
BoW-LM	**1.00**	0.39	0.56	0.99	0.39	0.55	0.29	0.32	0.41	0.11	-0.05	-0.05	-0.01	-0.02	0.02	-0.00
BoW-GI	0.39	**1.00**	0.35	0.39	**1.00**	0.35	0.18	0.19	0.25	0.06	0.03	0.03	-0.02	0.03	0.01	-0.02
BoW-CU	0.56	0.35	**1.00**	0.56	0.35	0.99	0.28	0.29	0.40	0.11	-0.02	-0.02	-0.02	0.03	-0.01	-0.02
BoW-LM-Neg	0.99	0.39	0.56	**1.00**	0.39	0.56	0.29	0.31	0.41	0.11	-0.05	-0.05	-0.01	-0.02	0.02	-0.00
BoW-GI-Neg	0.39	**1.00**	0.35	0.39	**1.00**	0.35	0.18	0.19	0.25	0.06	0.03	0.03	-0.02	0.03	0.01	-0.02
BoW-CU-Neg	0.55	0.35	0.99	0.56	0.35	**1.00**	0.27	0.28	0.39	0.11	-0.02	-0.03	-0.02	0.03	-0.01	-0.02
FinBERT	0.29	0.18	0.28	0.29	0.18	0.27	**1.00**	0.61	0.65	0.10	0.00	0.01	-0.00	0.04	0.02	-0.04
DistilRoBERTa	0.32	0.19	0.29	0.31	0.19	0.28	0.61	**1.00**	0.72	0.13	-0.02	-0.02	-0.03	0.02	0.02	-0.05
Fin-RoBERTa	0.41	0.25	0.40	0.41	0.25	0.39	0.65	0.72	**1.00**	0.14	-0.01	-0.01	-0.02	0.02	0.03	-0.03
GPT-3.5	0.11	0.06	0.11	0.11	0.06	0.11	0.10	0.13	0.14	**1.00**	0.00	0.01	-0.00	-0.01	-0.01	-0.02
Earnings Surprise	-0.05	0.03	-0.02	-0.05	0.03	-0.02	0.00	-0.02	-0.01	0.00	**1.00**	0.94	0.04	0.12	-0.16	0.26
Earnings Surprise^2	-0.05	0.03	-0.02	-0.05	0.03	-0.03	0.01	-0.02	-0.01	0.01	0.94	**1.00**	0.02	0.14	-0.11	0.23
Net Income	-0.01	-0.02	-0.02	-0.01	-0.02	-0.02	-0.00	-0.03	-0.02	-0.00	0.04	0.02	**1.00**	0.05	0.02	0.04
NUMUP	-0.02	0.03	0.03	-0.02	0.03	0.03	0.04	0.02	0.02	-0.01	0.12	0.14	0.05	**1.00**	0.08	-0.02
NUMDOWN	0.02	0.01	-0.01	0.02	0.01	-0.01	0.02	0.02	0.03	-0.01	-0.16	-0.11	0.02	0.08	**1.00**	-0.00
CAR-11-Carhart	-0.00	-0.02	-0.02	-0.00	-0.02	-0.02	-0.04	-0.05	-0.03	-0.02	0.26	0.23	0.04	-0.02	-0.00	**1.00**

Notes: This table reports pairwise Pearson correlations between sentiment scores extracted from earnings call transcripts using Bag-of-Words (BoW) and Large Language Models (LLMs), along with additional financial variables. BoW models include Loughran-McDonald (LM), Harvard GI (GI), and Custom (CU) dictionaries. Suffix "-Neg" denotes negation handling within a three-word threshold. LLMs include FinBERT, DistilRoBERTa, Fin-RoBERTa, and GPT-3.5. Additional variables include Earnings Surprise, Earnings Surprise^2 (squared), Net Income, NUMUP (number of upward revisions), NUMDOWN (number of downward revisions), and CAR-11-Carhart (cumulative abnormal returns using the Carhart model). Correlations are rounded to two decimal places.

Table 2 presents the same results as Table 1, but includes the control variables from the regressions. This was added merely to demonstrate that none of the variables exhibit correlation, except for earnings surprise and earnings surprise squared, which is expected due to how they are constructed. Additionally, these variables do not correlate with our sentiment values, an intriguing finding considering that one might anticipate a more positive presentation with favorable results. However, this could be attributed to the generally positive tone maintained by management in response to market reactions.

A.2 BoW Regressions

A.2.1 Regression Results for BoW Sentiment Analysis (No Fixed Effects)

Table 3: Regression Results for BoW Sentiment Analysis (No Fixed Effects)

	Without Negation			With Negation		
Variable	**LM**	**GI**	**CU**	**LM**	**GI**	**CU**
Alpha	-0.011***	-0.009***	-0.012***	-0.011***	-0.009***	-0.012***
	(-4.930)	(-3.406)	(-6.595)	(-4.938)	(-3.416)	(-6.568)
Earnings Surprise	0.012***	0.012***	0.012***	0.012***	0.012***	0.012***
	(7.011)	(7.027)	(7.009)	(7.011)	(7.024)	(7.011)
Earnings Surprise^2	-0.001*	-0.001*	-0.001*	-0.001*	-0.001*	-0.001*
	(-2.183)	(-2.193)	(-2.196)	(-2.181)	(-2.191)	(-2.197)
Net Income	0.001	0.001	0.001	0.001	0.001	0.001
	(1.824)	(1.786)	(1.802)	(1.823)	(1.786)	(1.804)
NUMUP	-0.001**	-0.001**	-0.001**	-0.001**	-0.001**	-0.001**
	(-2.938)	(-2.910)	(-2.916)	(-2.939)	(-2.905)	(-2.922)
NUMDOWN	0.001**	0.001**	0.001**	0.001**	0.001**	0.001**
	(2.755)	(2.779)	(2.750)	(2.756)	(2.781)	(2.753)
Sentiment	0.030	-0.039	-0.070	0.033	-0.039	-0.052
	(0.528)	(-1.234)	(-0.820)	(0.566)	(-1.237)	(-0.604)
Fixed Effects	No	No	No	No	No	No
R^2	0.075	0.075	0.075	0.075	0.075	0.075

Notes: This table reports regression results examining the impact of sentiment scores from earnings call transcripts on market reactions, using the Bag-of-Words (BoW) approach with Loughran-McDonald (LM), Harvard GI (GI), and Custom (CU) dictionaries. Columns under "With Negation" account for negations (threshold of three words), while those under "Without Negation" do not. No company- or year-fixed effects are included, and sentiment is not split by earnings surprise. T-statistics are in parentheses below coefficients. Significance levels: *** p<0.001, ** p<0.01, * p<0.05.

Table 3 presents the results of our regressions, highlighting that all control variables, in addition to sentiment, show significance, with net income being significant at the 10% level. In particular, the GI dictionary comes closest to significance. Interestingly, the score signs differ: LM yields a positive coefficient, while the other two produce negative ones, raising questions about whether the market reacts more to negative or positive scores. This table encapsulates the uncertainty inherent in the BoW approach, as the performance of the model varies greatly with different dictionaries. Moreover, the presence of substantial significance in the alphas signals potential omitted variable(s).

A.2.2 Regression Results for BoW Sentiment Analysis with Split Sentiment (No Fixed Effects)

Table 4: Regression Results for BoW Sentiment Analysis with Split Sentiment (No Fixed Effects)

	Without Negation			With Negation		
Variable	**LM**	**GI**	**CU**	**LM**	**GI**	**CU**
Alpha	-0.010***	-0.011***	-0.011***	-0.010***	-0.011***	-0.011***
	(-4.416)	(-4.247)	(-6.417)	(-4.416)	(-4.247)	(-6.417)
Earnings Surprise	0.010***	0.014***	0.011***	0.010***	0.014***	0.011***
	(4.818)	(5.742)	(6.608)	(4.818)	(5.742)	(6.608)
Earnings Surprise^2	-0.001	-0.001**	-0.001*	-0.001	-0.001**	-0.001*
	(-1.424)	(-2.587)	(-1.960)	(-1.424)	(-2.587)	(-1.960)
Net Income	0.001	0.001	0.001	0.001	0.001	0.001
	(1.822)	(1.831)	(1.837)	(1.822)	(1.831)	(1.837)
NUMUP	-0.001**	-0.001**	-0.001**	-0.001**	-0.001**	-0.001**
	(-2.908)	(-2.940)	(-2.915)	(-2.908)	(-2.940)	(-2.915)
NUMDOWN	0.001**	0.001**	0.001**	0.001**	0.001**	0.001**
	(2.741)	(2.758)	(2.755)	(2.741)	(2.758)	(2.755)
Sentiment (E[EPS] > 0)	0.020	-0.057	-0.140	0.166	0.035	0.272
	(0.305)	(-1.630)	(-1.411)	(1.361)	(0.591)	(1.368)
Sentiment (E[EPS] < 0)	0.166	0.035	0.272	0.020	-0.057	-0.140
	(1.361)	(0.591)	(1.368)	(0.305)	(-1.630)	(-1.411)
Fixed Effects	No	No	No	No	No	No
R^2	0.076	0.076	0.076	0.076	0.076	0.076

Notes: This table reports regression results examining the impact of sentiment scores from earnings call transcripts on market reactions, using the Bag-of-Words (BoW) approach with Loughran-McDonald (LM), Harvard GI (GI), and Custom (CU) dictionaries. Sentiment is split based on earnings surprise: "E[EPS] > 0" for positive surprises and "E[EPS] < 0" for negative surprises. Columns under "With Negation" account for negations (threshold of three words), while those under "Without Negation" do not. No company- or year-fixed effects are included. T-statistics are in parentheses below coefficients. Significance levels: *** p<0.001, ** p<0.01, * p<0.05.

Continuing from Table 3, I now categorize the sentiment based on positive or negative earnings surprises, as illustrated in Table 4. However, while some results become more pronounced, they remain statistically insignificant. It is evident that certain models perform better at predicting sentiment when the earnings surprise is less than anticipated and vice versa. Although there is a slight increase in R^2, the difference is not substantial.

A.2.3 Regression Results for BoW Sentiment Analysis with Company and Year Fixed Effects

Table 5: Regression Results for BoW Sentiment Analysis with Company and Year Fixed Effects

	Without Negation			With Negation		
Variable	**LM**	**GI**	**CU**	**LM**	**GI**	**CU**
Alpha	0.001	0.004	-0.001	0.001	0.004	-0.000
	(0.056)	(0.283)	(-0.039)	(0.059)	(0.283)	(-0.027)
Earnings Surprise	0.010***	0.010***	0.010***	0.010***	0.010***	0.010***
	(5.664)	(5.680)	(5.673)	(5.664)	(5.678)	(5.673)
Earnings Surprise^2	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	(-0.929)	(-0.935)	(-0.942)	(-0.928)	(-0.933)	(-0.944)
Net Income	0.002	0.002	0.002	0.002	0.002	0.002
	(1.322)	(1.286)	(1.290)	(1.323)	(1.283)	(1.288)
NUMUP	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	(-1.000)	(-1.008)	(-1.017)	(-1.001)	(-1.002)	(-1.013)
NUMDOWN	0.001**	0.001**	0.001**	0.001**	0.001**	0.001**
	(2.594)	(2.583)	(2.554)	(2.595)	(2.588)	(2.562)
Sentiment	0.035	-0.045	-0.118	0.038	-0.045	-0.098
	(0.523)	(-1.216)	(-1.228)	(0.559)	(-1.224)	(-1.007)
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.132	0.132	0.132	0.132	0.132	0.132

Notes: This table reports regression results examining the impact of sentiment scores from earnings call transcripts on market reactions, using the Bag-of-Words (BoW) approach with Loughran-McDonald (LM), Harvard GI (GI), and Custom (CU) dictionaries. Columns under "With Negation" account for negations (threshold of three words), while those under "Without Negation" do not. All regressions include company- and year-fixed effects. T-statistics are in parentheses below coefficients. Significance levels: *** p<0.001, ** p<0.01, * p<0.05.

Table 5 has been updated to incorporate fixed year and fixed company effects, which significantly enhance the R^2. An intriguing observation is that with the inclusion of these fixed effects, nearly all our variables lose significance, suggesting that many omitted variables—previously captured to some extent by these variables—are now accounted for by the fixed effects. Regarding sentiment scores, their significance diminishes across the board.

A.2.4 Regression Results for BoW Sentiment Analysis with Split Sentiment and Fixed Effects

Table 6: Regression Results for BoW Sentiment Analysis with Split Sentiment and Fixed Effects

	Without Negation			**With Negation**		
Variable	**LM**	**GI**	**CU**	**LM**	**GI**	**CU**
Alpha	0.002	0.002	-0.001	0.002	0.002	-0.001
	(0.130)	(0.168)	(-0.051)	(0.130)	(0.168)	(-0.051)
Earnings Surprise	0.009***	0.014***	0.010***	0.009***	0.014***	0.010***
	(3.909)	(5.194)	(5.296)	(3.909)	(5.194)	(5.296)
Earnings Surprise^2	-0.000	-0.001	-0.000	-0.000	-0.001	-0.000
	(-0.384)	(-1.705)	(-0.724)	(-0.384)	(-1.705)	(-0.724)
Net Income	0.002	0.002	0.002	0.002	0.002	0.002
	(1.324)	(1.351)	(1.314)	(1.324)	(1.351)	(1.314)
NUMUP	-0.000	-0.001	-0.001	-0.000	-0.001	-0.001
	(-0.961)	(-1.048)	(-0.994)	(-0.961)	(-1.048)	(-0.994)
NUMDOWN	0.001**	0.001**	0.001**	0.001**	0.001**	0.001**
	(2.599)	(2.576)	(2.581)	(2.599)	(2.576)	(2.581)
Sentiment (E[EPS] > 0)	0.032	-0.072	-0.190	0.183	0.047	0.232
	(0.441)	(-1.845)	(-1.735)	(1.391)	(0.751)	(1.101)
Sentiment (E[EPS] < 0)	0.183	0.047	0.232	0.032	-0.072	-0.190
	(1.391)	(0.751)	(1.101)	(0.441)	(-1.845)	(-1.735)
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.132	0.133	0.133	0.132	0.133	0.133

Notes: This table reports regression results examining the impact of sentiment scores from earnings call transcripts on market reactions, using the Bag-of-Words (BoW) approach with Loughran-McDonald (LM), Harvard GI (GI), and Custom (CU) dictionaries. Sentiment is split based on earnings surprise: "E[EPS] > 0" for positive surprises and "E[EPS] < 0" for negative surprises. Columns under "With Negation" account for negations (threshold of three words), while those under "Without Negation" do not. All regressions include company- and year-fixed effects. T-statistics are in parentheses below coefficients. Significance levels: *** p<0.001, ** p<0.01, * p<0.05.

Table 6 presents the outputs of the segmented regressions accounting for fixed variables. The results indicate that the same variables remain significant, but the significance of our sentiment values is increased. This suggests that even after accounting for fixed effects, segmenting them can better elucidate the variation in the returns.

A.3 LLMs Regressions

A.3.1 Regression Results for LLM Sentiment Analysis (No Fixed Effects)

Table 7: Regression Results for LLM Sentiment Analysis (No Fixed Effects)

Variable	FinancialBERT	DistilRoBERTa	Financial-RoBERTa	GPT-3.5
Alpha	0.010	0.018	0.005	0.009
	(0.866)	(1.525)	(0.505)	(0.511)
Earnings Surprise	0.011***	0.012***	0.011***	0.011***
	(6.994)	(7.016)	(6.988)	(6.993)
Earnings Surprise^2	-0.001*	-0.001*	-0.001*	-0.001*
	(-2.170)	(-2.209)	(-2.171)	(-2.170)
Net Income	0.001	0.001	0.001	0.001
	(1.809)	(1.749)	(1.777)	(1.815)
NUMUP	-0.001**	-0.001**	-0.001**	-0.001**
	(-2.879)	(-2.898)	(-2.913)	(-2.954)
NUMDOWN	0.001**	0.001**	0.001**	0.001**
	(2.794)	(2.812)	(2.803)	(2.747)
Sentiment	-0.014	-0.017**	-0.010	-0.003
	(-1.863)	(-2.521)	(-1.675)	(-1.170)
Fixed Effects	No	No	No	No
R^2	0.076	0.077	0.076	0.075

Notes: This table reports regression results examining the impact of sentiment scores from earnings call transcripts on market reactions, using Large Language Models (LLMs): FinancialBERT, DistilRoBERTa, and Financial-RoBERTa (BERT-based), and GPT-3.5-turbo (OpenAI). No negations, company-fixed effects, or year-fixed effects are included, and sentiment is not split by earnings surprise. T-statistics are in parentheses below coefficients. Significance levels: *** p<0.001, ** p<0.01, * p<0.05.

Moving on to LLMs, Table 7 illustrates our sentiment score, derived using LLMs instead of BoW. For all other variables, the relationships observed are consistent with those found when using BoW. Notably, in terms of significance, the LLM models surpass all BoW methods, with one model (DistilRoBERTa) achieving significance at the 1% level. It is important to note that this is prior to incorporating fixed effects. An additional observation from the table is that the intercept has lost its significance, indicating that LLM sentiment values account for what was previously captured by the intercept in the BoW methods.

A.3.2 Regression Results for LLM Sentiment Analysis with Split Sentiment (No Fixed Effects)

Table 8: Regression Results for LLM Sentiment Analysis with Split Sentiment (No Fixed Effects)

Variable	FinancialBERT	DistilRoBERTa	Financial-RoBERTa	GPT-3.5
Alpha	-0.009**	-0.008*	-0.009**	-0.010**
	(-2.294)	(-2.042)	(-2.277)	(-2.436)
Earnings Surprise	0.011**	0.011**	0.011**	0.009*
	(2.664)	(2.452)	(2.717)	(1.978)
Earnings Surprise^2	-0.001	-0.001	-0.001	-0.000
	(-1.154)	(-0.960)	(-1.162)	(-0.557)
Net Income	0.001	0.001	0.001	0.001
	(1.809)	(1.789)	(1.802)	(1.783)
NUMUP	-0.001**	-0.001**	-0.001**	-0.001**
	(-2.891)	(-2.867)	(-2.889)	(-2.882)
NUMDOWN	0.001**	0.001**	0.001**	0.001**
	(2.791)	(2.812)	(2.797)	(2.800)
Sentiment (E[EPS] > 0)	-0.002	-0.001	-0.002	0.000
	(-0.429)	(-0.434)	(-0.468)	(0.240)
Sentiment (E[EPS] < 0)	-0.002	-0.003	-0.002	-0.001
	(-0.460)	(-0.815)	(-0.516)	(-0.815)
Fixed Effects	No	No	No	No
R^2	0.075	0.075	0.075	0.075

Notes: This table reports regression results examining the impact of sentiment scores from earnings call transcripts on market reactions, using Large Language Models (LLMs): FinancialBERT, DistilRoBERTa, and Financial-RoBERTa (BERT-based), and GPT-3.5-turbo (OpenAI). Sentiment is split based on earnings surprise: "E[EPS] > 0" for positive surprises and "E[EPS] < 0" for negative surprises. No negations or company- and year-fixed effects are included. T-statistics are in parentheses below coefficients. Significance levels: *** p<0.001, ** p<0.01, * p<0.05.

Proceeding to Table 8, I observe that our significance level decreases when the measure is divided. The intercept becomes significant once more, suggesting it accounts for part of the explanatory power initially attributed to sentiment. This highlights that while the overall model generally explains abnormal returns well, it may not effectively do so when isolating cases of higher or lower than anticipated earnings, indicating it captures more complex relationships within the data.

A.3.3 Regression Results for LLM Sentiment Analysis with Company and Year Fixed Effects

Table 9: Regression Results for LLM Sentiment Analysis with Company and Year Fixed Effects

Variable	FinancialBERT	DistilRoBERTa	Financial-RoBERTa	GPT-3.5
Alpha	0.019	0.030	0.019	0.027
	(0.989)	(1.635)	(1.059)	(1.166)
Earnings Surprise	0.010***	0.011***	0.010***	0.010***
	(5.689)	(5.714)	(5.696)	(5.620)
Earnings Surprise^2	-0.000	-0.000	-0.000	-0.000
	(-0.957)	(-0.998)	(-0.959)	(-0.891)
Net Income	0.002	0.002	0.002	0.002
	(1.304)	(1.275)	(1.289)	(1.302)
NUMUP	-0.001	-0.001	-0.001	-0.001
	(-1.023)	(-1.087)	(-1.083)	(-1.052)
NUMDOWN	0.001**	0.001**	0.001**	0.001**
	(2.574)	(2.522)	(2.540)	(2.541)
Sentiment	-0.012	-0.019**	-0.012	-0.003
	(-1.458)	(-2.475)	(-1.770)	(-1.451)
Fixed Effects	Yes	Yes	Yes	Yes
R^2	0.132	0.134	0.133	0.132

Notes: This table reports regression results examining the impact of sentiment scores from earnings call transcripts on market reactions, using Large Language Models (LLMs): FinancialBERT, DistilRoBERTa, and Financial-RoBERTa (BERT-based), and GPT-3.5-turbo (OpenAI). All regressions include company- and year-fixed effects; no negations are applied, and sentiment is not split by earnings surprise. T-statistics are in parentheses below coefficients. Significance levels: *** p<0.001, ** p<0.01, * p<0.05.

In Table 9, it is evident that even after accounting for fixed effects, the sentiment score for DistilRoBERTa remains significant at the level 1%. Furthermore, I observe that the value R^2 reaches the highest among all regressions, indicating that this regression provides the most explanatory power. Regarding the sentiment scores of other LLMs, I find that one is significant at the 10% level, while the rest are not significant. All exhibit a negative sign, suggesting that as management becomes more positive to some extent, it leads to negative abnormal returns.

A.3.4 Regression Results for LLM Sentiment Analysis with Split Sentiment and Fixed Effects

Table 10: Regression Results for LLM Sentiment Analysis with Split Sentiment and Fixed Effects

Variable	FinancialBERT	DistilRoBERTa	Financial-RoBERTa	GPT-3.5
Alpha	0.004	0.005	0.004	0.004
	(0.252)	(0.357)	(0.302)	(0.258)
Earnings Surprise	0.011**	0.010*	0.011**	0.009
	(2.454)	(2.227)	(2.496)	(1.892)
Earnings Surprise^2	-0.001	-0.000	-0.001	-0.000
	(-0.718)	(-0.515)	(-0.710)	(-0.239)
Net Income	0.002	0.002	0.002	0.002
	(1.310)	(1.302)	(1.306)	(1.302)
NUMUP	-0.001	-0.001	-0.001	-0.001
	(-1.010)	(-1.009)	(-1.021)	(-1.004)
NUMDOWN	0.001**	0.001**	0.001**	0.001**
	2.592)	(2.594)	(2.588)	(2.596)
Sentiment (E[EPS] > 0)	-0.003	-0.003	-0.003	-0.000
	(-0.773)	(-0.862)	(-0.937)	(-0.343)
Sentiment (E[EPS] < 0)	-0.002	-0.004	-0.003	-0.001
	(-0.436)	(-0.902)	(-0.619)	(-0.865)
Fixed Effects	Yes	Yes	Yes	Yes
R^2	0.132	0.132	0.132	0.132

Notes: This table reports regression results examining the impact of sentiment scores from earnings call transcripts on market reactions, using Large Language Models (LLMs): FinancialBERT, DistilRoBERTa, and Financial-RoBERTa (BERT-based), and GPT-3.5-turbo (OpenAI). Sentiment is split based on earnings surprise: "E[EPS] > 0" for positive surprises and "E[EPS] < 0" for negative surprises. All regressions include company- and year-fixed effects; no negations are applied. T-statistics are in parentheses below coefficients. Significance levels: *** p<0.001, ** p<0.01, * p<0.05.

Finally, Table 10 presents the segmented sentiment score analysis while accounting for fixed effects. The intercepts lose their significance due to these controls, and, consistent with Table 8, all sentiment scores are not significant.