Numerical Claim Detection in Finance: A New Financial Dataset, Weak-Supervision Model, and Market Analysis

Agam Shah♠ , **Pratvi Shah**♣ , **Arkaprabha Banerjee**♣ , **Anushka Singh**♢
Arnav Hiray♠ , **Dheeraj Eidnani**♠ , **Bhaskar Chaudhury**♣ , **Sudheer Chava**♠
♠ Georgia Institute of Technology
♣ DA-IICT
♢ IIT-Kharagpur

Abstract

In this paper, we investigate the influence of claims in analyst reports and earnings calls on financial market returns, considering them as significant quarterly events for publicly traded companies. To facilitate a comprehensive analysis, we construct a new financial dataset for the claim detection task in the financial domain. We benchmark various language models on this dataset and propose a novel weak-supervision model that incorporates the knowledge of subject matter experts (SMEs) in the aggregation function, outperforming existing approaches. Furthermore, we demonstrate the practical utility of our proposed model by constructing a novel measure “optimism". Furthermore, we observed the dependence of earnings surprise and return on our optimism measure. Our dataset, models, and code will be made publicly (under CC BY 4.0 license) available on GitHub and Hugging Face.

1 Introduction

The release of ChatGPT on November 30th, 2022 began a race to build better-performing large language models (LLMs). While many companies benefited from this surge, the clear winner was the GPU manufacturer Nvidia. On May 24th, 2023, Nvidia surprised the market with earnings that beat analyst expectations by 44.26%[1]. Nvidia’s stock price was up 24.37% the next day. Such an event is largely driven by valuable information within analyst reports and earnings calls, highlighting their importance.

Earnings conference calls are a quarterly event where the company’s top executives provide performance reports of the company over the last quarter (3 months). Between the two earnings calls analyst from various financial institutions analyze and provide earnings estimates and recommendations. For example, Jegadeesh and Kim (2010) has documented that there is a significant stock market reaction to analysts’ recommendations (ratings). However, analyst ratings can be biased (Michaely and Womack, 1999; Corwin et al., 2017; Coleman et al., 2021). Therefore it is important to understand whether the ratings are backed by strong numerical financial claims in the analyst’s report. To evaluate the ratings’ reliability, the extraction of numerical financial claims is a necessary task. Further, the sentences with a claim have a higher density of forward-looking information. Related, extraction of numerical ESG claims from earnings call transcripts, can help better understand whether companies do walk the talk on their environment and social responsibility claims (Chava et al., 2021). The importance of mentioned examples necessitates the numerical claim detection task in the Finance domain.

A key component of this paper is the identification of Numeric Financial Sentences. Specifically, Numeric Financial Sentences include a financial term, a numeric value, and either a currency or percentage symbol. Chen et al. (2020) first introduced the categorization of sentences into ‘in-claim’ and ‘out-of-claim’ specifically in the Mandarin language. Expanding on their foundation, we define an ‘in-claim’ sentence as one presenting a speculative financial forecast. Conversely, an ‘out-of-claim’ sentence presents a numerical statement about a past event, transitioning from a mere claim to a confirmed fact. For clarity, ‘in-claim’ sentences can also be termed "financial forecasts" whereas ‘out-of-claim’ can be labeled as "established financials." Every Numeric Financial Sentence that is not a speculative financial forecast (in-claim) is an ‘out-of-claim’ sentence. Figure 1 illustrates the identification of Numeric Financial Sentences as well as distinguishing between “in-claim" and “out-of-claim" sentences.

Correspondence to Agam Shah {ashah482@gatech.edu}

[1] Estimated earnings per share (EPS) was $0.61 and actual EPS was $0.88. Source: `https://www.alphaquery.com/stock/NVDA/earnings-history`

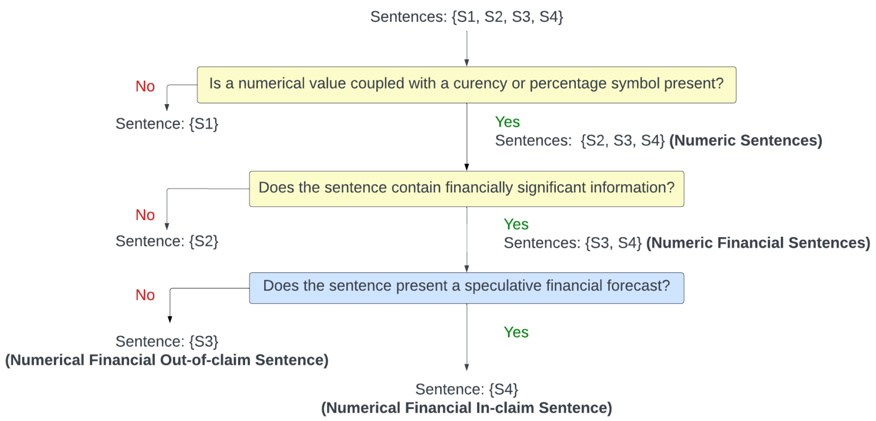
Sentences: {S1, S2, S3, S4}
Is a numerical value coupled with a curency or percentage symbol present?
No
Sentence: {S1}
Yes
Sentences: {S2, S3, S4} (Numeric Sentences)
Does the sentence contain financially significant information?
No
Sentence: {S2}
Yes
Sentences: {S3, S4} (Numeric Financial Sentences)
Does the sentence present a speculative financial forecast?
No
Sentence: {S3}
(Numerical Financial Out-of-claim Sentence)
Yes
Sentence: {S4}
(Numerical Financial In-claim Sentence)

Figure 1: Example of In-claim and Out-of-claim sentences.
S1: "We also continued to grow our total active installed base by adding new customers."
S2: "Adjusted operating margins of over 41% were above the midpoint of guidance, as we balanced our strategic investments with prudent discretionary spend."
S3: "In q2, we achieved a record $4.39 billion in revenue, representing 15% year-over-year growth."
S4: "Operating income is expected between $2.1 billion and $3.6 billion."

A major challenge for building or training predictive models is the scarcity of labeled data (Zhang et al., 2021; Ratner et al., 2017). Supervised learning often involves a significant amount of manual labeling of data which is often infeasible for large datasets. In such scenarios, one can leverage weak-supervision-based learning methods (Varma and Ré, 2018) or fine-tune the pre-trained language model. Weak-supervision is a process that leverages slightly noisy or imprecise labeling functions (lfs) to label vast amounts of unlabeled data (Ratner et al., 2020; Lison et al., 2021). The strength of the weak-supervision model lies in these imperfect labels, when combined, producing reliable predictive models (Lison et al., 2021; Zhang et al., 2021). However, a crucial component involves the development of effective lfs for a given raw dataset systematically rather than manual annotation (Lison et al., 2021).

The aim of our work is to derive financially significant information from the quarterly analyst reports and earnings calls by categorizing each numerical sentence as in-claim or out-of-claim. Our major contributions through this paper are the following:

- We introduce a new task of claim detection (in English) with a labeled dataset.
- We build clean, tokenized, and annotated open-source datasets based on earnings calls.
- We introduce a weak-supervision model with a novel aggregation function.
- We benchmark a wide range of language models for the claim detection task.
- We develop a novel measure of optimism and validate its usefulness in predicting various financial indicators.

2 Related Work

NLP in Finance Finance is one of the most attractive domains for the application of NLP. Araci (2019) and Liu et al. (2020) presented pre-trained language models for the Finance domain. There are multiple datasets specifically catered for applications of NLP in finance including question answering dataset created by Chen et al. (2021) and Maia et al. (2018), and also a NER dataset constructed by Shah et al. (2023b) for the financial domain. There is a vast body of literature on undertaking sentiment analysis tasks on financial data(Maia et al., 2018; Malo et al., 2014; Day and Lee, 2016; Akhtar et al., 2017).

Works of Li et al. (2020) and Sawhney et al. (2020) were centered around predicting volatility using earnings call transcripts in the domain of risk management. Chava et al. (2022) measure the firm level inflation exposure by fine-tuning RoBERTa (Liu et al., 2019), while Li et al. (2021) leveraged word-embeddings to measure the corporate culture. Moreover, Nguyen et al. (2021) and Hu and Ma (2021) used multimodal machine learning for credit rating prediction and measurement of persuasiveness respectively. Shah et al. (2023a) investigated the impact of monetary policy communication on financial markets. Cao et al. (2020) critically examined the evolution of corporate disclosure in recent years, influenced by the rising application of NLP in Finance. Our research focuses on identifying numerical financial claims from a vast set of English analyst reports and earnings calls using a weak-supervision model. This differs from Chen et al. (2020), which targets numeric claim detection in a smaller Chinese language dataset.

Weak-Supervision In order to reduce the complexities associated with manual labeling, several standard techniques such as semi-supervised learning (Chapelle et al., 2009), transfer learning (Pan and Yang, 2010), and active learning (Settles, 2009) have been employed. However, many researchers (Meng et al., 2018; Kartchner et al., 2020) and practitioners also employ weak-supervision-based models to further reduce the computational costs while retaining the accuracy of the labeled data. Weak-supervision models were primarily developed in a bid to replace standard labeling techniques with models which can leverage slightly noisy or imprecise sources to label vast amounts of data (Ratner et al., 2020). Techniques such as distant supervision (Mintz et al., 2009) and crowd-sourced labels (Yuen et al., 2011) are often associated with weak-supervision-based models, however, they tend to have limited coverage and accuracy (Lison et al., 2021). In the case where we have noisy labels from multiple sources available, there have been efforts made to use majority vote, weighted majority vote (Ratner et al., 2020), and other label-models (Yu et al., 2022; Zhang et al., 2022).

3 Dataset

We collect two categories of text and financial market datasets. Analyst reports are procured from a proprietary source while earnings call transcripts are collected in a manner that allows us to make the resulting dataset open-source.

3.1 Analyst Reports

The raw dataset consists of quarterly analyst reports (in English) for a large number of public firms in the U.S. These analyst reports were collected from Zacks Equity Research and were available to us through the Nexis Uni license[2]. Before the data is passed on to models it is standardized in order to maintain consistency for subsequent steps.

The text documents are split into sentences using multiple regex-based rules. We employ regex-based rules as they typically are significantly faster with similar accuracy compared to standard libraries in sentence tokenization. Afterward, numeric sentences containing statistical information (i.e. sentences consisting of a numeric value coupled with a currency or percentage symbol) are extracted, in order to ensure their numerical relevance (Chen et al., 2019). This numerical condition filter reduced the number of sentences by 66.7%.

The next step in the pipeline is to use a whitelisting technique in order to only retain sentences that contain financially significant information. This is done by cross-referencing each sentence with a financial dictionary which includes a comprehensive list of technical terms related to the financial market and corresponding literature. The financial dictionary used in this study was developed by Shah et al. (2022) and contains more than 8,200 financially significant terms. To verify whether a sentence contains any financially significant information, each word in the sentence is cross-referenced with the dictionary. If none of the words in the sentence exist in the dictionary, then the sentence is marked as irrelevant. Filtering using the financial dictionary reduced the dataset by an additional 17.2%. From Table 1, we can clearly observe that this two-tier filtering method enriched the data by retaining only 27.5% of the sentences from the original data.

Type	# Sentences
Total sentences	8,583,093
Total numeric sentences	2,857,567
Total numeric-financial sentences	2,364,977

Table 1: Change in the size of the dataset based on filtration.

3.2 Earnings Call Transcripts

To make our work more impactful, we also collect earnings call transcripts for NASDAQ 100 companies from their investor relation page. We were successfully able to write individual scripts for 78 out of 100 NASDAQ companies. As all the companies in this list are public companies, their data can be accessed and shared publicly which allows us to open-source the resulting dataset. Collecting data till March of 2023 results in a total of 1,085 earnings call transcripts. The biggest advantage of writing separate scripts for each company is that

---

[2] Nexis Uni license doesn't authorize republication of full or partial text. To solve this problem, we also collect and construct a dataset from earnings calls which can be made public under CC BY 4.0 license.

it allows us to keep adding more transcripts every quarter increasing the size of the dataset shared over time. We apply text processing (tokenization, numerical filter, financial dictionary filter) on earnings call transcripts similar to what is used for analyst reports.

3.3 Comparison with Related Dataset

In this section we compare our proposed datasets with NumClaim (Chen et al., 2020), an expert-annotated dataset in the Chinese language. Our dataset of raw analyst reports in the English Language from 1,530 major companies over the period of 2017-20 is significantly larger than NumClaim or other associated datasets. Our open-sourced dataset from collected earnings call transcripts is also larger than the NumClaim dataset. The detailed comparison of our datasets with NumClaim is provided in Table 2.

3.4 Financial Market Data

Stock Price and Earnings Surprise Data We collect stock price data from Polygon.io[3] starting January 1st, 2017. We collect the actual earnings per share (EPS) and forecasted median EPS from the I/B/E/S dataset.

Sector Data For each firm in our dataset, we collect sector information by collecting GSECTOR classification from the annual fundamental COMPUSTAT database. GSECTOR maps each company to one of the twelve sectors.

3.5 Sampling and Manual Annotation

From the complete raw dataset of 87,536 analyst reports and 1,084 earnings call transcripts, we sample data and annotate sentences. The sampled dataset consisted of 96 analyst reports consisting of two files per sector per year, accounting for about 2,681 unique financial-numeric sentences. We also sample 12 earnings call transcripts randomly consisting of two files per year, consisting of 498 financial-numeric sentences. This set was manually annotated and assigned 'in-claim' or 'out-claim' labels by two of the authors with a basic background in finance and domain-specific knowledge gained from examples supplied by a financial expert co-author. The annotator agreement was 99.21% and 95.78% for analyst reports and earnings call transcripts respectively. Any disagreement between the two annotators was resolved with the help of a third expert who has a deeper understanding of finance. Train, validation, and test split for both categories of the dataset are provided in Table 3.

4 Models

In this section, we provide details of the four categories of models we have used. Initially, we provide detail on the proposed weak-supervision model with the customized aggregation function. In order to provide a comprehensive benchmark for the claim detection task and comparison with proposed weak-supervision model, we add Bi-LSTM, six BERT architecture-based PLMs, and two generative LLMs.

4.1 Weak-Supervision Model

For implementing a weak-supervision model we use the Snorkel library (Ratner et al., 2017), leveraging its inherent pipeline structure for generating labels for each data segment and then passing the outputs through the customized aggregation function.

Labeling functions used in our model include rule-based pattern matching combined with part-of-speech (POS) tag constraints for some phrases. We create seventeen labeling functions for the categorization of results and also make use of multiple other labeling functions in order to divide sentences representing assertions or written in the past tense. These labeling functions are listed in Table 4. More details on the construction of the labeling function can be found in Appendix B.

The output of the labeling functions needs to be aggregated to decide the final label of the sentence. Unlike other models, we use independent and weighted labeling functions with weights based on the level of confidence assigned by SMEs. We have considered two levels of in-claim sentences resulting in a total of four types of return values (-1: Out-of-claim sentence, 0: Abstain, 1: Low confidence while making a claim, and 2: High confidence while making a claim). In the final results, both levels have been considered for in-claim sentences. This fine-grained categorization helps us understand the results better and opens room for future fine-tuning of the models. For our model, each labeling function classifies a sentence independently, and hence we consider the 'max' as our aggregating function as shown below:

---

[3] `https://polygon.io/stocks`

Dataset	Analyst Reports	Earnings Calls	NumClaim (Chen et al., 2020)
Language	English	English	Chinese
Year	2017-20	2017-23	NA
Sector Information	Yes	Yes	No
# Stocks	1,530	78	NA
# Files	87,536	1,085	NA
# Words	167,301,873	11,641,673	42,594
# Numeric Sentences	2,857,567	48,686	5,144
# Numeric Financial Sentences	2,364,977	41,013	NA
# Numeric Financial Claim Sentences	336,252	5362	1,233

Table 2: Comparison of our datasets with NumClaim (Chen et al., 2020) dataset.

Dataset	Train	Validation	Test
Analyst Reports	1,715	429	537
Earnings Calls	318	80	100

Table 3: Size of train, validation, and test split for two categories of the data

$$label\ (x_i) = \begin{cases} 1, & \max(lf_1(x_i), ...lf_n(x_i)) > 0; \\ & \forall lf_j(x_i) \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

where x_i is i^{th} sentence, $lf_j(x)$ is j^{th} labeling function, and $label(x_i)$ is label of i^{th} sentence.

4.2 Generative LLMs

To understand the capabilities of current state-of-the-art (SOTA) generative LLMs' in a zero-shot manner, we add a zero-shot ChatGPT[4] performance benchmark in our study. We use the "gpt-3.5-turbo" model with 1,000 max tokens for output, and a 0.0 temperature value. The ChatGPT API was accessed on June 12th, 2023. In a recent article, Rogers et al. (2023) made a case for why closed models like ChatGPT make bad baselines. In order to understand where SOTA open-source LLMs stand in comparison to ChatGPT and fine-tuned models, we also test the Falcon-7B-Instruct model with zero-shot. We use the following zero-shot prompt for both models:

"Discard all the previous instructions. Behave like you are an expert sentence sentiment classifier. Classify the following sentence into 'IN-CLAIM', or 'OUTOFCLAIM' class. Label 'IN-CLAIM' if consist of a claim and not just factual past or present information, or 'OUTOFCLAIM' if it has just factual past or present information. Provide the label in the first line and provide a short explanation in the second line. The sentence: {sentence}"

4.3 Bi-LSTM

In the realm of text classification problems, Long Short-Term Memory (LSTM) was a popular recurrent neural network architecture (Hochreiter and Schmidhuber, 1997). An enhanced approach to LSTM is the Bidirectional LSTM (Bi-LSTM), which processes input in both directions (Schuster and Paliwal, 1997). In order to assess the efficacy of Recurrent Neural Networks (RNNs) in claim detection, we employ the Bi-LSTM model on the datasets we have developed. Instead of training it from scratch, we initialize the embedding layer of the Bi-LSTM using 300-dimensional GloVe embeddings trained using Common Crawl (Pennington et al., 2014). We employ a grid search approach to identify the optimal hyperparameters for each model, considering four different learning rates (1e-4, 1e-5, 1e-6, 1e-7) and four different batch sizes (32, 16, 8, 4). In our training process, we employ a maximum of 100 epochs, incorporating early stopping criteria. In cases where the validation F1 score does not exhibit an improvement of greater than or equal to 1e-2 over the subsequent 7 epochs, we designate the previously saved best model as the final fine-tuned model.

4.4 PLMs

In order to establish a performance benchmark, our study encompasses a range of transformer-based (Vaswani et al., 2017) models of varying sizes. For the small models, we employ BERT (Devlin et al., 2018), FinBERT (Yang et al., 2020), FLANG-BERT (Shah et al., 2022), and RoBERTa (Liu et al., 2019). Within the category of large models, we incorporate BERT-large (Devlin et al., 2018) and RoBERTa-large (Liu et al., 2019). To avoid over-fitting on financial text, we refrain from conducting any pre-training on these models prior to fine-tuning. For PLMs, we employ grid-search, fine-tuning, and early stopping similar to what we used for Bi-LSTM. The experiments are conducted

---

[4] `https://chat.openai.com/`

Used to detect	Output	Type	Keyword or phrase
High Confidence out-of-claim (Past Tense or Assertions)	-1/0	Phrase Matching	reasons to buy:, reasons to sell:, was, were, declares quarterly dividend, last earnings report, recorded
Low Confidence in-claim	1/0	Phrase Matching	earnings guidance to, touted to, entitle to
High Confidence in-claim	2/0	Lemmatized Word matching	expect, anticipate, predict, forecast, envision, contemplate
High Confidence in-claim	2/0	POS Tag for word “project"	VBN, VB, VBD, VBG, VBP, VBZ
High Confidence in-claim	2/0	Phrase Matching	to be, likely to, on track to, intends to, aims to, to incur, pegged at

Table 4: Labeling Functions used in weak-supervision model. SpaCy Lemmatizer has been used for labeling functions involving lemmatized word matching.

using PyTorch (Paszke et al., 2019) on an NVIDIA RTX A6000 GPU. Each model is initialized with the pre-trained version from the Transformers library provided by Huggingface (Wolf et al., 2020).

5 Results

In this section, we present the results obtained using the above models and provide a detailed analysis of the outcomes.

5.1 Weak-Supervision Model

The performance in Table 5, highlights how well our Weak-Supervision based model performs when compared with Manually Annotated Data. In order to make sure that there is no contamination issue between the labeling functions and annotated data, we perform a robustness check in Appendix A.

Ablation #1: Number of Labeling Functions Figure 2, shows how the accuracy of the model changes depending on the number of labeling functions. For this plot, we initially computed the contribution of each labeling function (Table 4, High confidence and Low Confidence in-claim) towards the detection of in-claim sentences and then considered the addition of new labeling function at each step to ensure the steepest ascent to saturation. At each step, in addition to one new labeling function, all labeling functions present in Table 4 for Past Tense and Assertions, were also used. They either abstain or classify sentences as out-of-claim and help improve the classification of out-of-claim sentences. From the plot, we can notice that after around thirteen labeling functions, the addition of new labeling functions does not produce any change in the accuracy. In fact, increasing labeling functions thereafter leads to a minor decrease in accuracy. This suggests that we can effectively capture the required trends for classification in this setting with thirteen labeling functions.

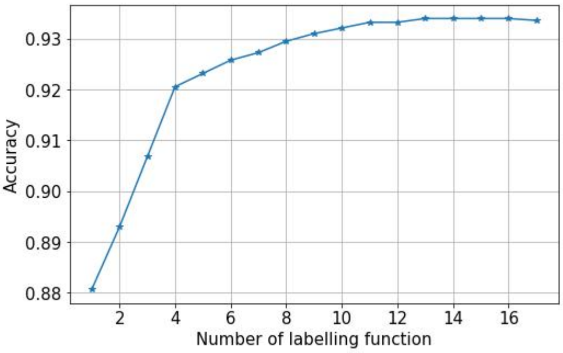

Figure 2: Accuracy v/s Number of labeling functions. Note: This is accuracy, not F1 score.

Ablation #2: Aggregation Functions We consider majority voting and Snorkel’s aggregation function (Ratner et al., 2017) as baseline aggregation functions for comparative analysis. The accuracy of baseline aggregation functions along with our aggregation function is reported in Table 6. For all three models, the same set of labeling functions is used and they only differ in the aggregation part.[5] The result highlights the importance of the construction of a customized aggregation function for a weak-supervision model where a small set of labeling functions are complete and less noisy.

5.2 Generative LLMs

Zero-shot ChatGPT fails to outperform both weak-supervision and fine-tuned Bi-LSTM and PLMs based on BERT architecture. It still achieves impressive performance without having access to any labeled data. The finding here is in line with the survey done by Pikuliak (2023), which finds that zero-shot ChatGPT fails to outperform fine-tuned models on more than 77% of NLP tasks. Falcon-7B-Instruct with zero-shot performs worst and the

---

[5] We do not perform any post-processing on the output to convert abstain label to one of the labels.

Zero-Shot Models	**No Train/Analyst Reports (AR)**		**No Train/Earnings Calls (EC)**	
Weak-Supervision	0.9272 (0.0116)		0.9382 (0.0213)	
ChatGPT-3.5-Turbo	0.8189 (0.0135)		0.7371 (0.0334)	
Falcon-7B-Instruct	0.2008 (0.0134)		0.1780 (0.0062)	
Fine-Tuned Models	**AR/AR**	**EC/AR**	**AR/EC**	**EC/EC**
Bi-LSTM	0.9309 (0.0235)	0.8244 (0.0332)	0.8961 (0.0236)	0.8892 (0.0375)
BERT-base-uncased	0.9532 (0.0192)	0.9269 (0.0150)	**0.9251** (0.0113)	0.9376 (0.0205)
FinBERT-base	0.9617 (0.0076)	**0.9381** (0.0112)	0.9209 (0.0257)	0.9279 (0.0135)
FLANG-BERT-base	0.9611 (0.0137)	0.9270 (0.0109)	0.9119 (0.0257)	0.9363 (0.0089)
RoBERTa-base	0.9615 (0.0091)	0.9319 (0.0131)	0.8906 (0.0301)	**0.9563** (0.0036)
BERT-large-uncased	0.9539 (0.0111)	0.9183 (0.0063)	0.9197 (0.0349)	0.9416 (0.0349)
RoBERTa-large	**0.9642** (0.0069)	0.9381 (0.0138)	0.8975 (0.0244)	0.9427 (0.0153)

Table 5: In the table, A/B indicates that the model is fine-tuned on dataset A and tested on dataset B. All values are F1 scores. An average of 3 seeds was used for all models. The standard deviation of F1 scores is reported in parentheses. Generative LLMs and weak-supervision models are tested as zero-shot while all other models are fine-tuned with training data.

Aggr. Funtion	**AR**	**EC**
Majority Vote	0.4274 (0.0208)	0.5313 (0.0427)
Snorkel's WMV	0.4269 (0.0204)	0.5309 (0.0372)
Ours	0.9272 (0.0116)	0.9382 (0.0213)

Table 6: Performance comparison of our aggregation function with baseline aggregation functions. All values are F1 scores. An average of 3 seeds was used for all models. The standard deviation of F1 scores is reported in parentheses.

F1 score is much lower than ChatGPT highlighting the gap between open and closed LLMs.

5.3 Bi-LSTM

The Bi-LSTM model outperforms the weak-supervision model on analyst reports data but doesn't outperform on earnings call data. The potential reason can be the larger fine-tuning dataset available for analyst reports. It doesn't outperform the model based on BERT on any of the four configurations.

5.4 PLMs

The fine-tuned models utilizing the BERT architecture demonstrate superior performance compared to other model classes, emphasizing the significant value gained from annotated data. Intriguingly, the model that achieves the highest performance within a particular train-test dataset category does not necessarily exhibit the best performance on transfer learning datasets. This finding underscores the importance of separate data annotation. Notably, the RoBERTa model emerges as the top performer within the same train-test data category.

6 Market Analysis

6.1 Experiment Setup

Construction of the Optimism Measure We use our weak-supervision model to label all the financial numeric sentences in the analyst reports and earnings calls as in-claim or out-of-claim. We then filter the sentences and only keep in-claim sentences to evaluate predictions.

We further label each in-claim sentence as 'positive', 'negative', or 'neutral' using the fine-tuned sentiment analysis model specifically for the financial domain. The model is fine-tuned for financial sentiment analysis using the pre-trained FinBERT (Araci, 2019). We then use labeled sentences in each document to generate a document-level measure of analyst optimism for document i using the following formula:

$$\text{Optimism}_i = 100 \times \frac{\text{Pos. In-claim}_i - \text{Neg. In-claim}_i}{\text{Total Sentences}_i} \quad (1)$$

where Pos. In-claim$_i$ and Neg. In-claim$_i$ are the number of positive and negative in-claim sentences respectively in document i after the filter, and Total Sentences$_i$ is the total number of sentences in the document.

Empirical Specification We use the following empirical specification for market analysis.

$$Y_{i,t} = \alpha + \beta \times \text{Optimism}_{i,t} + \epsilon_{i,t} \quad (2)$$

Here $Y_{i,t}$ is the outcome variable of interest for firm i at time t, α is a constant term, and $\epsilon_{i,t}$ is an error term. The coefficient (β) will help us understand the influence of Optimism$_{i,t}$ on the outcome variable ($Y_{i,t}$).

Outcome (Y)	Constant (α)	Beta (β)
Earn. Surp.	0.1665 ***	-1.8643 ***
CAR [+2, +30]	1.0777 ***	-36.4158 ***
CAR [+2, +60]	1.0924 ***	-58.3560 ***

Table 7: Market analysis result based on the empirical regression. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

6.2 Post Earnings Prediction

We examine the relation between optimism in analyst reports for a company in a specific quarter and its effect on earnings. Using earnings-based metrics, we perform a regression as per Eq 2 using earnings call transcripts and analyst report data. For quarters with multiple reports on one stock, we aggregate sentences and claims to compute Optimism_i. We focus on optimism measures that are not zero, indicating either optimism or pessimism.

Earnings Surprise (%) The Earning Surprise (%) is calculated by subtracting the median EPS (in the last 90 days) from the actual EPS. The difference is scaled by the stock price at the end of the quarter and multiplied by 100. This method aligns with Chava et al. (2022).

The Earnings Surprise (%) is set as the outcome variable ($Y_{i,t}$). The results in Table 7 show a significant link between optimism and the Earnings Surprise (%). A negative β coefficient indicates that with every unit rise in optimism in analyst reports, the Earnings Surprise (%) drops. This implies that heightened optimism in reports often leads to the actual EPS underperforming expectations. This "false optimism" aligns with previous studies like (Coleman et al., 2021), highlighting analysts' tendency to overestimate firm performance.

Cumulative Abnormal Returns We further aim to explore the influence of optimism in analyst reports on the magnitude of cumulative abnormal return (CAR) post-earnings. CAR for a firm represents the total daily abnormal stock return in the period after a specific event, in our context, the firm's earnings conference call.

We analyze two CAR time frames. CAR[+2, +30] is the cumulative abnormal for the [+2,+30] trading day window post-earnings call, as determined by Chava et al. (2022). The same methodology is used to calculate CAR[+2, +60] as well.

Table 7 shows that greater optimism in analyst reports corresponds with a larger decline in cumulative abnormal return. This emphasizes the 'false optimism' trend in reports, where increased optimism leads to greater discrepancies from actual outcomes, leading to a larger negative cumulative abnormal return.

The prevailing notion in finance literature is that analysts often exhibit an over-optimistic bias. While Francis and Philbrick (1993) and Barber et al. (2007) believe this bias helps maintain good ties with corporate insiders, Michaely and Womack (1999) sees it as a means for personal financial gains. Recently, Brown et al. (2022) found that analysts favor firms with attributes like high debt or fluctuating earnings. This suggests such firms might exaggerate earnings, potentially through manipulation. Our market analysis aligning with these theories reinforces our method's accuracy and the financial relevance of our study.

7 Conclusion

Our work presents claim based labeled dataset in the English language alongside presenting a weak-supervision model with a standalone accuracy of 93%. Developed customized aggregation function outperforms baseline aggregation functions. We also benchmark various language models and compare the performance with the weak-supervision model. We show the application of claim detection by generating a measure of optimism from the weak-supervision model. We also validate the measure by studying its applicability in predicting earnings surprise, abnormal returns, and earnings call optimism. We release our models, code, and benchmark data (for earnings call transcripts only) on Hugging Face and GitHub. We also note that the trained model for claim detection can be used on other financial texts.

Limitations

By acknowledging the following limitations, we pave the way for future research to address these areas and further enhance the understanding and applicability of our approach.

- *Limited Scope of Text Data*: Our analysis is restricted to analyst reports and earnings calls, excluding other potentially valuable text datasets such as related news articles and investor presentations. Incorporating these additional sources of information could provide a more comprehensive understanding of pre-earnings drifts.

- *Exclusion of Audio and Video Features*: Our measure construction does not utilize audio or video features from earnings calls, which may contain supplementary information.

- *Narrow Range of Language Models*: Although we benchmark various models, we do not include Large Language Models (LLMs) such as LLaMA and MPT. Exploring other generative LLMs, including zero-shot and few-shot learning approaches, could further enrich our analysis.

- *Omission of Alternative Weak-Supervision Models*: We do not explore multiple end models, such as the confidence-based sampling with contrastive loss proposed in the COSINE framework by Yu et al. (2020). Incorporating such alternative weak-supervision models could offer additional insights and improve the robustness of our approach.

- *Absence of Trading Strategy Construction*: Although we provide a market analysis based on the proposed measure of optimism, we do not construct and backtest trading strategies. Future work could explore the development and evaluation of trading strategies using our measure as a basis.

Ethics Statement

Our work adheres to ethical considerations, although we acknowledge certain biases and limitations in our study. We do not identify any potential risks stemming from our research; however, we recognize the presence of geographic and gender biases in our analysis.

- *Geographic Bias*: Our study focuses solely on publicly listed companies in the United States of America, which introduces a geographic bias. The findings may not be fully representative of global firms and markets.

- *Gender Bias*: We acknowledge the gender bias present in our study due to the predominant representation of male analysts, CEOs, and CFOs.

- *Data Ethics*: The data used in our study, derived from publicly available sources, does not raise ethical concerns. All raw data is obtained from public companies that are obligated to disclose information under the guidance of the SEC and are subject to public scrutiny.

- *Language Model Ethics*: The language models employed (with proper citation) in our research are publicly available and fall under license categories that permit their use for our intended purposes. While most models employed are publicly available, it is important to note that ChatGPT's prompt answers will not be made public due to licensing conditions. We acknowledge the environmental impact of large pre-training of language models and mitigate this by limiting our work to fine-tuning existing models.

- *Annotation Ethics*: All annotations were performed by the authors, ensuring that no additional ethical concerns arise from the annotation process.

- *Hyperparameter Reporting*: In the interest of clarity and readability, we refrain from reporting the best hyperparameters found through grid search in the main paper. Instead, we will make all grid search results, including hyperparameter information, publicly available on GitHub. This transparency allows interested readers to access detailed information on our experimental setup.

- *Publicly Available Data*: We specify the datasets that will be made publicly available and indicate the applicable licenses under which they will be shared.

By acknowledging these ethical considerations and limitations, we strive to maintain transparency and promote responsible research practices.

References

Md Shad Akhtar, Abhishek Kumar, Deepanway Ghosal, Asif Ekbal, and Pushpak Bhattacharyya. 2017. A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 540–546.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Brad M Barber, Reuven Lehavy, and Brett Trueman. 2007. Comparing the stock recommendation performance of investment banks and independent research firms. *Journal of financial economics*, 85(2):490–517.

Anna Bergman Brown, Guoyu Lin, and Aner Zhou. 2022. Analysts' forecast optimism: The effects of managers' incentives on analysts' forecasts. *Journal of Behavioral and Experimental Finance*, 35:100708.

Sean Cao, Wei Jiang, Baozhong Yang, and Alan L Zhang. 2020. How to talk when a machine is listening: Corporate disclosure in the age of ai. Technical report, National Bureau of Economic Research.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542.

Sudheer Chava, Wendi Du, and Baridhi Malakar. 2021. Do managers walk the talk on environmental and social issues? *Available at SSRN 3900814*.

Sudheer Chava, Wendi Du, Agam Shah, and Linghang Zeng. 2022. Measuring firm-level inflation exposure: A deep learning approach. *Available at SSRN 4228332*.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Numeral attachment with auxiliary tasks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1161–1164.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Numclaim: Investor's fine-grained claim detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1973–1976.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Braiden Coleman, Kenneth J Merkley, and Joseph Pacelli. 2021. Human versus machine: A comparison of robo-analyst and traditional research analyst investment recommendations. *The Accounting Review, Forthcoming*.

Shane A Corwin, Stephannie A Larocque, and Mike A Stegemoller. 2017. Investment banking relationships and analyst affiliation bias: The impact of the global settlement on sanctioned and non-sanctioned banks. *Journal of Financial Economics*, 124(3):614–631.

Min-Yuh Day and Chia-Chou Lee. 2016. Deep learning for financial sentiment analysis on finance news providers. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1127–1134. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jennifer Francis and Donna Philbrick. 1993. Analysts' decisions as products of a multi-task environment. *Journal of Accounting Research*, 31(2):216–230.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Allen Hu and Song Ma. 2021. Persuading investors: A video-based study. Technical report, National Bureau of Economic Research.

Narasimhan Jegadeesh and Woojin Kim. 2010. Do analysts herd? an analysis of recommendations and market reactions. *The Review of Financial Studies*, 23(2):901–937.

David Kartchner, Wendi Ren, David Nakajima An, Chao Zhang, and Cassie S Mitchell. 2020. Regal: Rule-generative active learning for model-in-the-loop weak supervision. *Advances in neural information processing systems*.

Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. 2020. Maec: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3063–3070.

Kai Li, Feng Mai, Rui Shen, and Xinyan Yan. 2021. Measuring corporate culture using machine learning. *The Review of Financial Studies*, 34(7):3265–3315.

Pierre Lison, Jeremy Barnes, and Aliaksandr Hubin. 2021. skweak: Weak supervision made easy for nlp. *arXiv preprint arXiv:2104.09683*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *IJCAI*, pages 4513–4519.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2018. Weakly-supervised neural text classification. In *proceedings of the 27th ACM International Conference on information and knowledge management*, pages 983–992.

Roni Michaely and Kent L Womack. 1999. Conflict of interest and the credibility of underwriter analyst recommendations. *The Review of Financial Studies*, 12(4):653–686.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Cuong V Nguyen, Sanjiv R Das, John He, Shenghua Yue, Vinay Hanumaiah, Xavier Ragot, and Li Zhang. 2021. Multimodal machine learning for credit modeling. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1754–1759. IEEE.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Matúš Pikuliak. 2023. Chatgpt survey: Performance on nlp datasets. https://www.opensamizdat.com/posts/chatgpt_survey.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11 (3), page 269. NIH Public Access.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2020. Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, 29(2):709–730.

Anna Rogers, Niranjan Balasubramanian, Leon Derczynski, Jesse Dodge, Alexander Koller, Sasha Luccioni, Maarten Sap, Roy Schwartz, Noah A. Smith, and Emma Strubell. 2023. Closed ai models make bad baselines.

Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Shah. 2020. Voltage: volatility forecasting via text-audio fusion with graph convolution networks for earnings calls. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8001–8013.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Burr Settles. 2009. Active learning literature survey. *University of Wisconsin-Madison Department of Computer Sciences*.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023a. Trillion dollar words: A new financial dataset, task & market analysis. *arXiv preprint arXiv:2305.07972*.

Agam Shah, Ruchit Vithani, Abhinav Gullapalli, and Sudheer Chava. 2023b. Finer: Financial named entity recognition dataset and weak-supervision model. *arXiv preprint arXiv:2302.11157*.

Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When FLUE meets FLANG: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2335, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paroma Varma and Christopher Ré. 2018. Snuba: Automating weak supervision to label training data. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 12 (3), page 223. NIH Public Access.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *CoRR*, abs/2006.08097.

Yue Yu, Lingkai Kong, Jieyu Zhang, Rongzhi Zhang, and Chao Zhang. 2022. Actune: Uncertainty-based active self-training for active fine-tuning of pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1422–1436.

Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. 2020. Fine-tuning pretrained language model with weak supervision: A contrastive-regularized self-training approach. *arXiv preprint arXiv:2010.07835*.

Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. 2011. A survey of crowdsourcing systems. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 766–773. IEEE.

Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. 2021. Wrench: A comprehensive benchmark for weak supervision. *arXiv preprint arXiv:2109.11377*.

Rongzhi Zhang, Yue Yu, Pranav Shetty, Le Song, and Chao Zhang. 2022. Prboost: Prompt-based rule discovery and boosting for interactive weakly-supervised learning. *arXiv preprint arXiv:2203.09735*.

A Robustness Check

From a data engineering perspective, there can be concern about the model design and gold data construction as the authors who designed the weak-supervision model have annotated the data. This can lead to exaggerated performance on the data, which may taint the test set. To ensure that there is no contamination issue in the weak-supervision model and it is generalizable, we get the same test dataset annotated separately by four annotators with master's degrees in Quantitative Finance. These annotators were hired by the department as Graduate Assistants based on merit and were paid a $20 per hour salary for their work which is more than double the federal minimum wage and higher than the highest minimum wage ($15.74 in Washington, D.C) in the USA. The rates are standard and in compliance with ethical standards. These annotators had no information about the rules/patterns used in our weak-supervision model. Each sample in the test dataset is annotated by two annotators, and we drop the observations where there is a disagreement among annotators. [6] The F1 score of the weak-supervision model on a dataset annotated by non-authors is 0.9281 which is close to a score of 0.9272 on the author-annotated dataset. We also recalculate the F1 score of the model based on the author-annotated labels after dropping observations dropped in a non-author annotated dataset. The model gives a higher mean F1 score of 0.9360 which is expected as ambiguous sentences are dropped. Overall these results show the robustness of our model on the dataset annotated by annotators who don't have knowledge of the rules used in the weak-supervision model. From here onwards, the performance is always calculated on a gold dataset created by authors.

B Labeling Functions Methodology

The following illustrates the methodology adopted by us while choosing the rules to define the weak-supervision mode. All rules were acknowledged post detailed analysis of sample documents distributed over sector and time :

1. Phrases often provided definitive information about a given sentence in a document and in most cases they had a fairly consistent linguistic composition. It was exploited to both identify out-of-claim and in-claim sentences.
2. Certain phrases such as "reasons to buy", "reasons to sell" or the presence of words which are indicative of past tense such as "was", "were" are characteristic of out-of-claim sentences, since they indicated either facts or events which happened in the past.
3. The alternate adoption of phrase matching was to identify in-claim sentences. This mostly consisted of a verb form indicative of a probabilistic event (eg: likely, intends) coupled with a preposition (usually "to" or "at"). Based on the ambiguity of the resulting phrase they were either categorised as a high-confidence claim or a low-confidence one.
4. In a bid to capture the effect of a few other verb forms indicative of a probabilistic event, we also chose to look at its lemmatized form to reduce inflectional usage and use the base token for a more holistic evaluation over multiple usage formats.

---

[6] There is 98.59% agreement between two annotators.

		ES (%)	CAR [+2,+30]	CAR [+2, +60]
Sentence Type/Subset	**Average Sentences**	**Adj.** β	**Adj.** β	**Adj.** β
Unfiltered	98	-.0005***	-0.0175***	-.0282***
Numeric	26	-.0028***	-.062***	-.0995***
Numeric Financial	21.6	-.00228***	-.0716***	-.1142***
Numeric Financial In-claim	3.7	-.01427***	-.2800***	-.4481***

Table 8: Ablation on market analysis, highlighting the importance and information density of "in-claim" sentences. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

5. POS tags were also derived for "project" as a word wherever present. This was done to segregate its usage as a verb. Its usage as a verb was usually observed to be adopted while making claims or predictions.

C Ablation Study: Market Analysis

To understand the influence of "in-claim" sentences on market sentiment, we introduce the optimism measure in section 6, outlining its implications. In this section, we carry out an ablation study to better understand the impact of "in-claim" sentences. As such, we compute the optimism score for four sentence subsets: Unfiltered, Numerical, Numerical Financial, and Numerical Financial "In-claim" sentences for each file. For example, the optimism score for a subset of Numerical sentences for document i is given by:

$$\text{Optimism (Numerical)}_i = 100 \times \frac{\text{Pos. Numerical}_i - \text{Neg. Numerical}_i}{\text{Total Sentences}_i}$$

We standard normalize these scores for uniform comparison by deducting their mean and dividing by the standard deviation. As the beta coefficient lacks full context, to factor in the size of the sentence subset, we adjusted each coefficient by the average sentence count, terming it as the adjusted beta. This illustrates the information density in each filtered sentence set. When examining the Earnings Surprise (%) columns of Table 8 the Adjusted Beta for Earnings Surprise increases, implying that a mere average of 3.7 "in-claim" sentences holds crucial information. This highlights the high information density of our filtered sentences. While we aren't dismissing the importance of other sentences, our analysis reveals that the ones we've extracted are the most informative on a per-sentence basis.