MSc Data Science for Decision Making
Academic year 2022 - 2023

Decoding Abnormal Returns:
Unraveling Insights from Pharmaceutical Sector Earnings Calls through Graph-Enhanced Text Analysis

Vicente Lisboa, Lucas Santos and Davis Thomas

Abstract

Decoding Abnormal Returns: Unraveling Insights from Pharmaceutical Sector Earnings Calls through Graph-Enhanced Text Analysis

Vicente Lisboa[1], Lucas Santos[2] and Davis Thomas[3]

1] vicente.lisboa@bse.eu
2] lucas.santos@bse.eu
3] davis.thomas@bse.eu

This study[1] explores the potential correlation between textual data obtained through Natural language processing (NLP) techniques from earnings calls and its impact on enhancing the accuracy of stock price forecasting for the top ten pharmaceutical companies by adopting an event study methodology framework to estimate the cumulative abnormal return in the post-earnings-call 30-day window. This research lends support to a graph representation of earnings call transcripts, as the behavior of the management speakers relative to the analysts during the earnings call is a good predictor of the return of the stocks in a short time frame.

Keywords: Textual data; Stock price forecasting; Earnings calls, Natural Language Processing (NLP); Graph features

[1]We express our gratitude to the supervising and advising researchers, Christian Brownlees and Hannes Mueller. We extend our appreciation to Eric Matamoros from Novartis for presenting us with our thesis challenge and offering valuable feedback. We also acknowledge the support provided by PhD student Elliot Motte.

Contents

1 Introduction

The investigation of factors that contribute to enhancing performance in stock market prediction has garnered significant attention from both scholars and market participants. This interest stems from the potential exploitation of even minor advantages, which can yield substantial long-term impacts. For instance, the Medallion fund, an algorithmic hedge fund developed by Jim Simons that holds one of the most impressive track records in the industry, achieved an impressive annual gross return of 66% between 1988 and 2018, despite a win ratio of only 51%. This highlights the impact over the long-term that small advantages have in trading strategies.

In the era of big data, abundant information is available for analysis, and researchers have explored various data sources to improve stock price forecasting models. One such source is the textual data obtained from earnings calls of publicly traded companies. Earnings calls are scheduled conference calls held by companies shortly after they release their quarterly financial statements. During these calls, company executives, including the CEO, CFO, and other key executives, present a summary of the financial results for the quarter, discussing relevant financial metrics such as revenue, earnings, expenses, and profit margins. They also highlight significant events, developments, challenges, and future prospects that have impacted or could impact the company's performance. As an example, a distinctly important topic in the earnings calls for pharmaceutical companies besides financial highlights is the R&D (Research and Development) expenses, since these companies are characterized by significantly higher median R&D expense as a fraction of revenue than other large companies in other sectors. (LEDLEY et al., 2020)

In recent years, there has been a growing recognition of the potential correlation between textual data from earnings calls and the accuracy of stock price forecasting, as the textual data encompasses information-rich transcripts of these calls, capturing the insights, perspectives, and market sentiments expressed by company executives. Extracting valuable features from these transcripts and incorporating them into forecasting models can potentially improve their predictive power.

This research aims to explore the potential correlation between textual data obtained from earnings calls and its ability to enhance the accuracy of stock price forecasting for the top ten pharmaceutical companies. The study employs natural language processing (NLP) techniques to extract meaningful features from the transcripts and analyze their impact on forecasting models. The practical implications of this investigation are significant for market participants, including investors, traders, and financial analysts. By improving the accuracy of stock price forecasting, these individuals can make more informed decisions in their respective roles.

In this study, nine different models were developed and analyzed to predict Cumulative Abnormal Return (CAR) based on various features. The models were formulated by combining different sets of features, and their performance was evaluated using both ordinary least squares (OLS) and XGBoost algorithms. The XGBoost models consistently outperformed the OLS model in terms of Mean Squared Error (MSE), indicating that they provided better overall fit to the data. Among the XGBoost models, Model incorporating the graph features and earnings surprise emerged as best performer. The inclusion of graph features, such as the average Euclidean distance between management and analysts, positively influenced the predicted

CAR, suggesting the importance of management-analyst communication during earnings calls.

The structure of this work is organized as follows: Section 2 offers a comprehensive literature review that encompasses existing research on stock price prediction and the utilization of textual data. Sections 3 and 4 outline the methodology employed in this study, along with an overview of the data collection process. The results and analysis of the experiments conducted are presented in Section 5, followed by a discussion of the findings. Finally, Section 6 concludes the thesis, highlighting its limitations, and explores potential avenues for future research.

2 Literature Review

The Efficient Market Theory (EMH) is grounded in the notion that market prices adequately incorporate all available information in fully efficient markets. The work of (FAMA, 1970) established the three forms of efficiency (weak, semi-strong, and strong) and subsequent studies have further investigated and refined this hypothesis. Researchers have examined diverse aspects, including anomalies that could be exploited in trading strategies, behavioral biases, and the influence of new information on market efficiency.

Despite the adepts of the EMH being skeptical about strategies that "beat" the market consistently, finding strategies and how to improve the performance of the portfolios in the stock market with successful trading strategies were the goal of many researchers during the years. The research done by (BALL; BROWN, 1968) stands as a seminal effort that sought to explore inefficiencies in the market. It was the first study to examine the correlation between earnings surprises and stock market returns. The authors tried to assess the usefulness of existing accounting income numbers by comparing the performance of portfolios built with stocks with high earnings surprise and low surprise and built the idea that developed the concept of Post-Earnings Announcement Drift (PEAD). According to (FINK, 2021) PEAD refers to the phenomenon that stock prices tend to continue to drift upward (downward) following earnings announcements when the quarterly earnings were above (below) expectations.

Several studies have utilized the Price Earnings Announcement Drift (PEAD) anomaly proposed by Ball anf Bown as a starting point for their research. For instance, the work of (KAESTNER 1, 2006) introduces an event study methodology to measure the price behavior, represented by the cumulative abnormal return, following earnings releases to forecast post-event returns. The PEAD and the Event Methodology approach combined was used to measure the impact of the release of text documents on the stock prices. The study done by (LEE et al., 2014) focused on forecasting stock price changes in response to events reported in the 8-K document. The results indicated that the incorporation of text features slightly enhanced prediction performance. It is worth highlighting that this work approached the prediction task as a classification, forecasting the direction of the return. In the same way, the findings of (AKITA et al., 2016) show that the use of text features improves the results of the prediction for Recurrent Neural Networks and Long-Short Term Memory networks and that distributed representations of textual information perform better than Bag-of-Words (BoW) methods.

Most of the works that extract text features to predict stock movement rely on traditional and simple Natural Language Processing methods. Compared to traditional manual content

analysis, the dictionary-based approach significantly improves the efficiency of text classification tasks. Researchers create sets of keywords that correspond to specific groupings, such as topics, attributes, or stakeholders, that they want to detect in the text. Each document is then analyzed to identify the presence of these keywords. A comprehensive review of these methods is done by (RIFFE et al., 2019). These manually curated lists of words represent the constructs they aim to identify and can be used for several simple automated content analysis tasks: word counts, keyword-in-context, and co-occurrence of words to improve contextual information extraction.

In that sense, techniques like Latent Dirichlet Allocation (LDA), a probabilistic topic modeling technique used to uncover latent topics within a collection of documents, represent a improvement. It is an unsupervised learning algorithm that represents each document as a mixture of topics and each topic as a distribution of words. Such latent variable models have the potential to add value to large document collections by discovering interpretable, low-dimensional subspaces. Coherence-based topic model evaluation proposed by (MIMNO et al., 2011) improves the quality of the low-dimensional subspaces (topics) that are obviously flawed to human domain experts. This leads to the construction of topic allocation vectors that are richer in information retrieval and can provide more meaningful insights into the document collection.

In order to extract different features from the text, graph-based representations emerge to exploit relations in documents offering a more expressive document encoding than a bag of words approach, and this approach is validated and evaluated on classification tasks in (JIANG et al., 2010). (SALAMAT et al., 2022) encodes corpus-wide features with graph representation, and extends this to a transductive hybrid approach composed of an unsupervised node representation learning model followed by a node classification/edge prediction model, which is then developed to classify stock market technical analysis reports.

(GHOSAL et al., 2019b) proposed an unconventional approach to emotion recognition in conversation transcripts, by modeling the graph nature of a conversation to leverage inter-speaker dependency to improve the classification of the text. Several authors have attempted to extend the hierarchical representation of data that graphs allow to improve performance on prediction tasks, among them (CHENG et al., 2022), where a multi-model graph neural network is used to construct a heterogeneous knowledge graph which is then used for financial time-series data prediction. This work aims to build upon these techniques to leverage the contextual information in earnings calls, especially by capturing the speaker interactions to better predict the post-announcement reaction of investors.

Based on the literature review conducted to guide the feature creation process, the next step is choose the most suitable algorithm for the task. In this context, previous studies, such as the work by (FAUZAN; MURFI, 2018), have reported favorable outcomes with the XGBoost model. For instance, in the domain of insurance claim prediction, the XGBoost model exhibited superior accuracy when compared to other tree-based methods (e.g., AdaBoost, Random Forest) and Neural Networks

3 Methodology

The literature review covers the methodologies used on this work, in this section the methodological steps taken are described and detailed. Starting from the methodology selected to estimate the performance of the models and after detailing the methodological steps to generate the text futures.

3.1 Event Study Methodology

The Event Study Methodology (ESM) is a widely used approach in empirical research that focuses on analyzing the impact of specific events on financial markets. In the case of an earnings call, this methodology can be applied to assess the market's reaction to the information disclosed during such calls. This work follows an adapted version of the one applied in (KAESTNER 1, 2006).

The event window is defined as the date of the call and the two days before and after the event. The selection of this window is influenced by various factors, including the duration necessary for the market to fully incorporate and process the information provided during the event. The pre-event window of 720 days is used to estimate the βs used to estimate the abnormal returns (a smaller window would result in βs with high variance, the distribution of the values is available in Figure 14) and the post-event period is selected as the thirty days after the event. After selecting the windows of analysis, pursuant to ESM, the following steps were adopted:

1. Estimation of Normal Returns: To understand the normal behavior of stock prices during the event window, the expected returns of the stocks in the absence of the event were computed. This estimation is done using a benchmark index such as the SP500.

2. Calculation of Abnormal Returns: Abnormal returns represent the difference between the actual returns observed during the event window and the estimated normal returns. These abnormal returns are considered to be the impact of the event on stock prices.

3. Computing the target variables of interest: The cumulative sum of the abnormal returns (Cumulative Abnormal Return) over the event window (CAR1) and the post-event window (CAR2) gives a better insight into the impact of the earnings call, as comparing these two variables allows to measure the impact of the Post-Earnings Announcement Drift (PEAD).

The event study methodology provides a structured framework for assessing the effects of the earnings call on stock prices. It allows to quantify and analyze market reactions, providing valuable insights into the relationship between earnings release and financial markets.

3.1.1 Cumulative Abnormal Returns (CAR)

The process to estimate the Cumulative Abnormal Return (CAR) for the post-event window for each stock is based on the Capital Asset Pricing Model (CAPM) introduced by (SHARPE, 1964) and (LINTNER, 1965). The equation introduced in (KAESTNER 1, 2006) was adapted by adding the β to represent the sensibility of the stock to market returns.

$$AR_{i,t} = R_{i,t} - \beta R_{m,t} \tag{1}$$

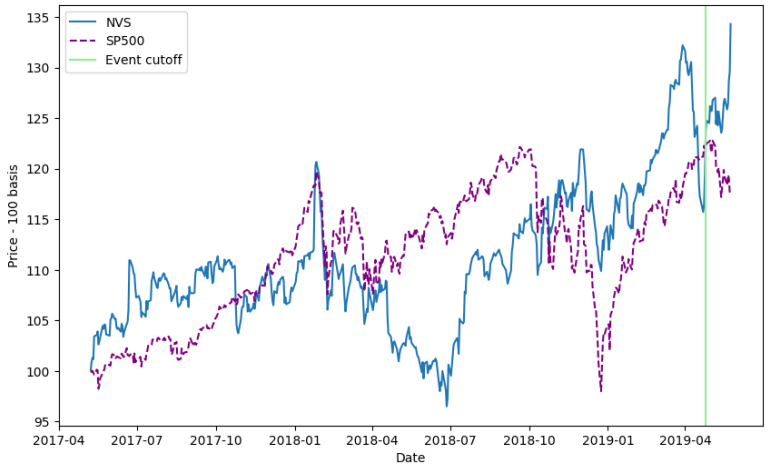

Figure 1: Evolution of NVS vs S&P500 100 basis
$\beta = 0.63$, CAR $= 0.1010$

Equation 1 shows the procedure to calculate the abnormal returns, where $R_{i,t}$ is stock i's daily return at time t and $R_{m,t}$ is representative of the return of the market portfolio, which in this case is the return of the S&P500. The Cumulative Abnormal Return is obtained by summing daily abnormal returns of stock i for various event windows following (and excluding) the announcement day as shown in equation 2:

$$CAR_i(q) = \sum_{t=1}^{q} AR_{i,t} \tag{2}$$

Figure 1 provides a graphical representation of the calculation process for the CAR. The period leading up to the event cutoff (a 720-day window) is utilized to calculate the β coefficient. Subsequently, the CAR is computed by comparing the expected return based on the performance of the overall market (represented by the SP index) with the actual return of Novartis.

3.2 Topic Modelling with Dictionary based methods

Unlike Latent Dirichlet Allocation (LDA) or other topic modeling techniques which learn topics from the data itself, dictionary-based topic modeling relies on pre-existing knowledge or domain-specific dictionaries to assign documents to predefined topics, thus increasing the interpretability of the model. Treating the corpus as a bag of words, the topic distributions were computed based on predefined dictionaries created by research groups with financial expertise. This lends interpretability to the topic allocations extracted from the text data, and is invaluable in gaining insights into the industry's financial landscape, thereby aiding decision-making processes.

Basing topic allocation extraction on dictionary methods proposed by (RIFFE et al., 2019) and using the dictionaries created by (S&P GLOBAL MARKET INTELLIGENCE, 2020), dictionary-based methods of topic detection in transcripts were implemented. The market-moving trigram dictionaries broadly cover six categories of topic tags: revenue, operating income, earnings, cash flow, profitability and shareholder return, and this is shown in

Table 1. Unigrams that represent the directionality of the text are also constructed as a dictionary, and these can be used in conjunction with the topics themselves as directional descriptors that capture the direction of the market-moving topics.

Topic	Keywords
Revenue	sales, revenue, top line, top bottom line, net revenue, organic revenue growth, organic sales growth, operational sales
Earnings	eps, earnings, earnings per share, net income, bottom line, top bottom line
Profitability	margin, gross margin, operating margin, return invested capital, return capital
Operating Income	ebit, operating income, operating profit, operating earning
Cashflow	cash flow, operating cash flow, cash flow operations, free cash flow
Shareholder Return	buyback, dividends, dividend per share, share repurchase, repurchased million shares
Positive Words	increase, increased, increases, increasing, increasingly, expand, expanded, expanding, expands, expansion, expansions, grow, grows, grew, growth, growths, improve, improved, improves, improvement, improvements, strong, stronger, strongest, strongly
Negative Words	decline, declined, declines, declining, deteriorate, deteriorates, deteriorated, deteriorating, compress, compressed, compresses, compressing, compressible, compression, reduce, reduces, reduced, reducing, reduction, reductions, weak, weaker, weakest, weaken, weakens, weakened, weakening, weakness, weaknesses

Table 1: Topic Dictionaries

3.3 Cosine similarity

Cosine similarity is a metric used to measure the similarity between two vectors projected in a multi-dimensional plane. It assesses the orientation of vectors rather than their magnitude. Cosine similarity finds wide-ranging applications in text documents and machine learning algorithms. Specifically, in the domain of text documents, cosine similarity is employed to gauge the similarity between two documents based on their content. By representing documents as vectors, with each dimension corresponding to a word or term, cosine similarity facilitates efficient text comparison and retrieval. This metric proves valuable in tasks such as document

clustering and analyzing semantic similarity.

3.3.1 Cosine similarity based on LDA topic allocation vectors

To perform cosine similarity analysis using earnings calls each transcript needs to be represented as a vector of values. In this study, Latent Dirichlet Allocation (LDA) is used exclusively for this purpose, to capture an embedding space for the earnings calls, by creating a vector of topic distributions for each transcript. The best model is chosen by optimizing the coherence score to obtain better semantic descriptors of the corpus. Employing this methodology enables the algorithm to extract and organize conversations centered around coherent topics. This approach has been extensively evaluated against human perceptions of similarity to provide lower-dimensional document representations that can be used in visualizations and in computing similarity between documents, in the work of (TOWNE; ROSÉ; HERBSLEB, 2016). Using this method, the topic vector of an earnings call is compared to the past five, the past ten, and the past twenty earnings calls of the same company, and the cosine similarity is computed for each transcript in the dataset. For the oldest earnings calls the value is zero since there is no observation before, and this modification does not distort the feature value.

3.3.2 Cosine Similarity using TF-IDF Vectors

The TF-IDF measure is commonly used in natural language processing and information retrieval to evaluate the importance of words in a document within a collection of documents. It takes into account both the frequency of a term within a document (term frequency) and its rarity across the entire document collection (inverse document frequency).

In the context of this study, Sam et al. utilized a similar approach in their work on measuring patent novelty, as referenced in (ARTS; HOU; GOMEZ, 2021). They employed L2 normalization of the TF-IDF vectors for each document and subsequently calculated the cosine similarity between a focal patent and all patents filed within the five years preceding the focal patent. This similarity measure allowed them to assess the distinctiveness of the focal patent in comparison to prior patents.

By leveraging TF-IDF vectors, researchers can capture the relative importance of words in a document and enable comparisons between different documents based on their content. This technique plays a crucial role in various applications, including document clustering, information retrieval, and text classification, as it helps identify relevant and distinctive features within a corpus of text.

This approach is adopted to create a feature that calculates the novelty of the focal earnings call compared to all earnings calls across companies in the past. The corpus is cleaned of stopwords and punctuation, and lemmatization is applied to convert the words to their base form considering the context. The transcripts are processed to remove any digits or alphanumeric combinations, and the resulting text is tokenized, and the transcripts are then vectorized into term frequency (TF-IDF) representation using a count vectorizer. This conversion resulted in a sparse matrix containing the weights assigned to each word within each document, which can be seen in Table 2. The parameters of the vectorizer include setting the minimum document frequency to 10, and the maximum document frequency to 300. Bi-grams and tri-grams

are also included to capture word sequences.

To ensure consistency in the term frequencies, the TF matrix is then normalized using the L2 norm. This step is crucial for accurately calculating cosine similarity between documents. Following the normalization the cosine similarities between the current transcript and all previous transcripts are computed. This computation is performed for each transcript in the dataset, resulting in a comprehensive analysis of similarities across all available transcripts, compared to all previous earnings calls.

Ticker	carcinoma	diabetes	gaap	diagnosis	eps	profit
PFE	0.029776	0.000000	0.014888	0.007444	0.044664	0.000000
LLY	0.000000	0.034806	0.110218	0.000000	0.034806	0.005801
NVO	0.000000	0.152848	0.000000	0.000000	0.000000	0.062271
JNJ	0.000000	0.011825	0.017737	0.000000	0.041387	0.005912

Table 2: Document-Term Matrix for selected features

3.4 Graph Representation

Treating the earnings calls as a text document, or an article, and using methods that leverage understanding the text as a bag of words fails to sufficiently capture the information present in the transcript. The earnings call has multiple elements at play, vying for the attention of the listening investor. The participants in the conversation, ie, the CEO, the speakers from the company and the analysts have different motives. The representatives of the company will try to present the news as positively as possible, and the analysts aim to shine light on either the details of the presentation or on events surrounding the company that has not been discussed in the call.

To capture this interplay, the conversation should be modeled as a dynamic interaction between speakers, and this can be done by creating a graph representation of the conversation using nodes for the speakers and their corresponding utterances. Utterances are the entire contiguous segment of the conversation spoken by an individual speaker. SOTA Emotion Recognition in Conversation(ERC) methodologies, such as (GHOSAL et al., 2019a), (SHEN et al., 2021) advocate for this approach to represent a conversation.

In this approach, every transcript undergoes preprocessing to eliminate punctuation and other characters. The text corpus is then segmented into speakers and utterances, with careful tracking of the positions of each utterance and speaker. The utterance nodes are linked to their respective speaker nodes and sequentially connected to one another based on their positions in the conversation. Additionally, speaker nodes are connected to one another if those speakers engaged in a dialogue. Finally, all utterance nodes belonging to a single speaker are sequentially connected to each other.

The graph representations were created using the StellarGraph library (DATA61, 2018), which has support for heterogeneous directional graphs for this use case. Taking the example of an earnings call released by Pfizer in the second quarter of 2009. The major company speakers in this call, Chuck Triano, Jeff Kindler, Frank D'Amelio, Ian Read, Martin Mackay, and Amy

Schulman can be observed in a speaker-filtered subgraph in Figure 2.

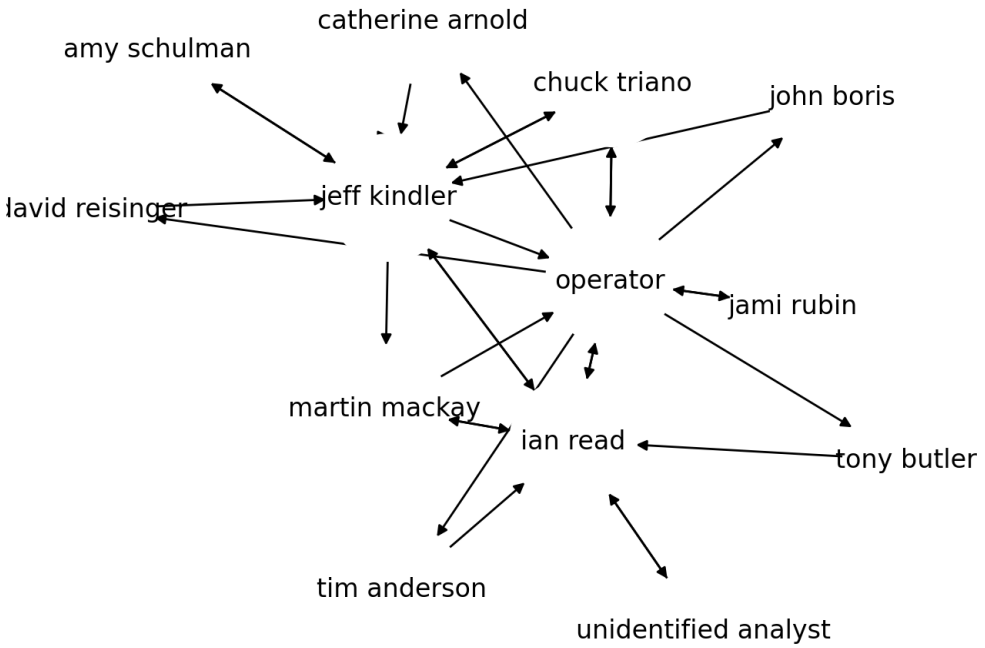

Figure 2: Subgraph of speakers of an earnings call: PFE, second quarter, 2009

It's also possible to see from the complete graph visualization of the transcript in Figure 3 that some speakers play a more central role in the conversation than others, and hence any embedding extracted from this transcript should account for the role that the speaker plays in the conversation, further validating the modeling approach.

Text embeddings for the utterance nodes can be obtained using sentence transformers developed by (REIMERS; GUREVYCH, 2019). The "all-mpnet-base-v2" model is chosen as the embedder for its impressive performance on sentence embedding benchmarks. As the context size of the model is limited to 384 words, the embeddings for each viable segment of the utterance were averaged. The embedder returns a vector of length 768 to represent each segment. Experiments with different embedding lengths revealed that better separation and richer embedding spaces were obtained using this length.

In contrast to learning a transcript/graph level embedding, which would be analogous to a text embedding of the whole text, learning a node-level feature representation for the speaker nodes in the graph from the nodes in its neighborhood offers more expressivity and potential for interpretability. To this end, the embeddings of the text nodes are aggregated using a general, inductive framework that leverages node feature information proposed by (HAMILTON; YING; LESKOVEC, 2017), a graph neural network, GraphSAGE. The loss function of this model forces close nodes to have similar representations while distant nodes will have different representations. This algorithm demonstrates robust generalization capabilities, particularly when applied to previously unseen nodes. By leveraging the graph's structural properties, the algorithm effectively samples and combines features from the neighborhood of a given node,

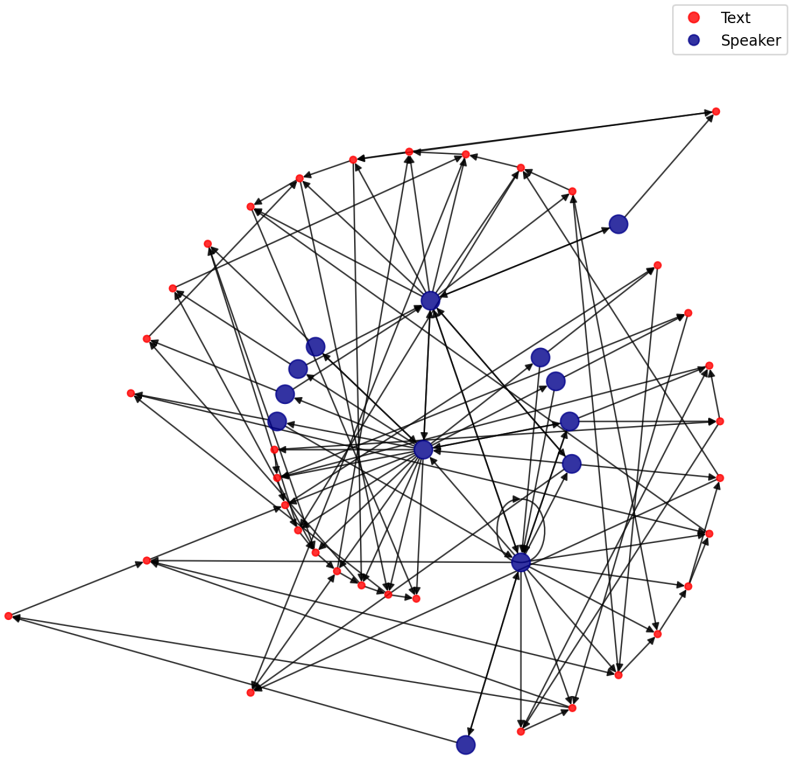

Figure 3: Graph Representation of an earnings call transcript: PFE, second quarter, 2009

resulting in the generation of a comprehensive embedding specific to that node.

Given the heterogeneous nature of the transcript graph, it becomes essential for the model to consider the node types and incorporate a mechanism for appropriately weighting the aggregated information from diverse node types. To address this requirement, an adaptation of GraphSAGE called HinSAGE is employed. HinSAGE offers the necessary flexibility and expressivity to aggregate embeddings from nodes of varying types, thus enabling a more comprehensive analysis of the transcript graph.

Instead of training the embeddings using the abnormal returns as the target to optimize the embedding space for the regression task, the generation of general-purpose embeddings is prioritized, especially considering the dataset constraints. To achieve this unsupervised feature learning using Deep Graph Infomax (DGI) (VELIČKOVIĆ et al., 2018) is performed. DGI trains using data in the format (target,context), where the target are the speaker nodes. The context nodes are generated from sampling from the neighborhood of the target node, at a depth of 2, to limit sampling from nodes too far away from the speaker, as this would result in embeddings that were identical for all speakers in a transcript. DGI works by learning to distinguish between the original graph and a graph sample corrupted by a corruption procedure, which randomly shuffles node features. Using this procedure to train the weights of the encoder model(HinSAGE), embeddings are generated for the nodes of interest.

3.5 Prediction and evaluation method

Following the feature generation, for the regression problem, 18 models using 9 distinct combinations of features were estimated, being one linear model (Ordinary Least Squares - OLS) and one machine learning algorithm (XGBoost Regressor).

To evaluate this, two error-based metrics (MAPE and MSE - the two most popular error-based metrics to evaluate Machine Learning Algorithms predicting returns (DESSAIN, 2022)) were used to evaluate the regression models.

3.5.1 Error Based metrics

The first one is the MSE, it is a popular metric for evaluating predictive models in finance for simplicity and interpretability, which make it valuable for comparing models and assessing forecasting accuracy. However, its sensitivity to outliers and lack of directional distinction in financial contexts, where overestimation and underestimation can have different impact, the metric could not fully capture the desired evaluation.

To have broader view about the results the models were evaluate using the Mean Absolute Percentage Error (MAPE) and Mean Squared Error (MSE), since MSE is sensitive to outliers and large errors due to the squared term, while MAPE focuses on the relative errors, making it more robust to the scale of the data. By considering both metrics, it's possible to gain a balanced perspective on the model's performance, considering both absolute and relative errors.

3.5.2 Model Explainability

Model explainability assess the importance of features to determine which ones contribute the most to the model's learning process. Firstly, SHAP analysis was performed. SHAP (SHapley Additive exPlanations) value is a model-agnostic technique for interpreting machine learning models by assigning importance scores to each feature in a prediction. It is based on the concept of Shapley values from cooperative game theory. SHAP values provide a unified measure of feature importance that takes into account the interactions and dependencies between features. They quantify the contribution of each feature towards the prediction compared to the average prediction.

Another way to analyze the model is through local interpretability. In this case, LIME (Local Interpretable Model-Agnostic Explanations) was used to examine how different features can be interpreted. LIME is a model-agnostic technique that provides explanations for individual predictions by approximating the behavior of the underlying model in the local vicinity of the instance being explained. It helps to shed light on the factors that contribute to specific predictions by creating a simpler, interpretable model that approximates the original model's behavior around the instance of interest. By applying LIME to a Novartis Earning Call, the aim was to understand how the various features can be interpreted in the context of the prediction. This approach allows for a detailed examination of the importance and impact of different features on the model's predictions, providing valuable insights into the decision-making process of the model.

4 Data

The following section introduces the different sources of date used and the feature creation.

4.1 Source

From the Financial Modeling Prep API, an extensive dataset was obtained comprising the top 10 pharmaceutical industries based on their revenue. The dataset includes the following companies in alphabetical order: AbbVie (AABV), AstraZeneca (AZN), Bristol-Myers Squibb Co (BMY), Johnson Johnson (JNJ), Eli Lilly And Co (LLY), Merck Co Inc (MRK), Novo Nordisk A/S (NVO), Novartis AG (NVS), Pfizer Inc. (PFE), and Rogers Corp (ROG). These companies represent prominent players in the pharmaceutical sector, and analyzing their financial data can provide valuable insights into the industry's dynamics. The dataset encompasses a comprehensive range of daily stock prices for each company from the years 2005 to 2022 and serves as a fundamental component in this analysis, enabling the exploration of correlations, identification of patterns, and evaluation of the companies' growth trajectories over the years.

The **earnings calls transcripts** of the pharmaceutical companies were obtained from the same source. These transcripts were published on a quarterly basis within the specified timeframe. Earnings calls serve as valuable resources, documenting the discussions and presentations held among company executives, analysts, and investors. Through the analysis of these transcripts, insights can be gained regarding various aspects including financial performance, strategic initiatives, RD updates, regulatory challenges, and market outlooks. Examining the language employed by company representatives during these calls enables a deeper understanding of their perspectives, priorities, and communication strategies.

Furthermore, the analysis includes reported and expected earnings per share (EPS) figures for each company on a quarterly basis collected from Refinitiv Eikon (LISBOA; SANTOS; THOMAS, 2023). EPS is a crucial financial metric that measures the profitability of a company and is widely used by investors to assess its financial performance. By comparing the reported EPS figures with the expected EPS, it's possible to analyze the extent to which companies meet or exceed market expectations. This analysis helps us evaluate the companies' ability to generate profits, manage costs, and deliver value to shareholders.

4.2 Features

4.2.1 Target - Cumulative Abnormal Return

The target variable is the Cumulative Abnormal Return (CAR). As detailed in the methodology section, it is the difference between the expected return and the observed return given the historic of the stock. Figure 4 illustrates the assessment of class imbalance in the target variable based on ticker. The bars in the figure represent the sign of the target variable. There is a tendency for the target variable to be more positive than negative. Is possible to infer from the figure that the CAR has slightly more positive values than negative values in general, for some stocks (NVO and ABBV) the imbalance is bigger, with one class representing at least $\frac{3}{4}$ of the observations.

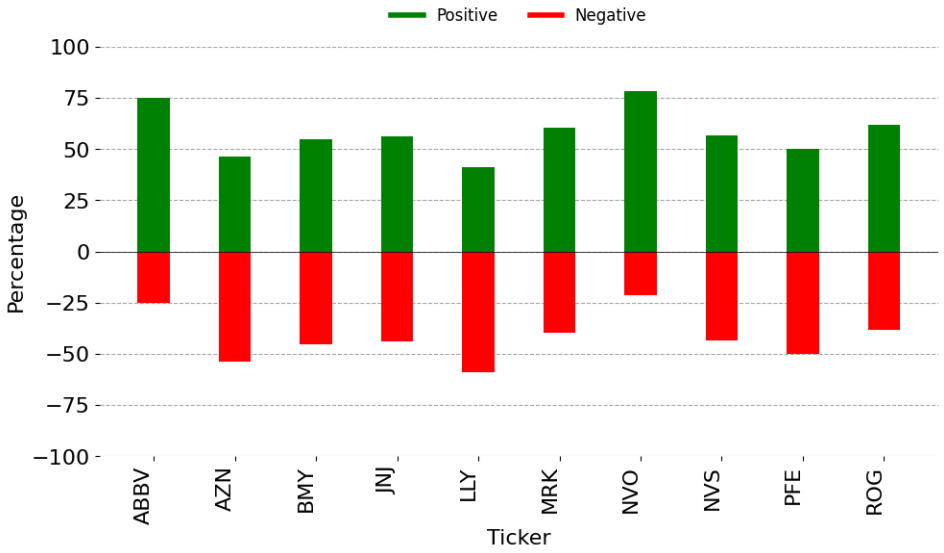

Figure 4: Class imbalance of target variable by ticker

4.2.2 Dummies

Generation of binary variables that take a value of 0 or 1 to indicate the presence or absence of a particular category. These variables serve the purpose of categorizing and distinguishing the companies within the model, allowing for the incorporation of company-specific effects or characteristics. By including dummy variables, the model can capture any potential differences or variations in the data across different pharmaceutical companies, enabling a more comprehensive analysis of their individual impacts on the dependent variable or outcome of interest.

4.2.3 Earnings Per Share

The estimation of Earnings Per Share (EPS) by companies holds significant importance as it influences decision-making processes within the organization. Periodically, companies provide EPS estimates, and any errors in these estimations can have substantial implications. To incorporate this element of surprise into the model, a percentage calculation is performed to determine the difference between the expected EPS and the reported EPS, this was called the surprise effect in the model.

The EPS surprise serves as a valuable metric by allowing for a comparison between a company's actual earnings performance and the expectations set by the market. A positive EPS surprise indicates that the reported EPS exceeds the expected EPS, signifying that the company has surpassed market expectations. Conversely, a negative EPS surprise indicates that the reported EPS falls short of the anticipated value, indicating that the company's earnings have failed to meet the market's expectations.

4.2.4 Technical Features

The inclusion of technical features in the study of stock prices for a company enhances the analysis by providing insights into price patterns, trends, market sentiment, and risk character-

istics. These features, derived from historical price and volume data, go beyond fundamental factors and offer a quantitative representation of market dynamics and investor behavior. By incorporating technical indicators and measures, such as moving averages, support and resistance levels, volatility indicators, and oscillators, analysts and investors gain a comprehensive understanding of the stock's behavior and can make more informed investment decisions. The combination of fundamental and technical analysis improves the accuracy of stock price estimations and forecasting by considering both intrinsic value and market-driven factors. During the feature definition process, particular attention was paid to avoid any overlap between different quarters in order to mitigate the potential risk of data leakage. This precautionary step was taken to ensure that each feature is constructed in a manner that maintains the integrity of the dataset. By preventing overlap, the independence and reliability of the information within each quarter are preserved

- **Relative Strength Index (RSI)** :

  Is a momentum oscillator that measures the speed and change of price movements. It compares the magnitude of recent gains and losses over a specific period and provides a relative strength value between 0 and 100. For this project was used the index for 5, 14, and 50 days. The RSI helps identify overbought and oversold conditions in the stock, indicating potential trend reversals or continuation. It's an important indicator used in technical analysis to assess the stock's price momentum and potential buying or selling opportunities.

- **Percentage changes of the stock prices** :

  The analysis percentage changes can reveal crucial information about the stock's volatility and price movements. Larger percentage changes may indicate periods of significant price fluctuations, while smaller changes may suggest more stable price behavior. Monitoring these changes over time can help identify patterns, trends, and potential turning points in the stock's performance. As features were used the percentage change of the stock price within the following periods: 1 week, 1 month and 1 quarter.

4.2.5 Topic Features

- **Topic Allocations** :

  Topic modeling consists of the extraction of meaningful themes or topics from a collection of documents. It allows uncovering latent patterns, recurring themes, and hidden structures within textual data without prior knowledge of the specific topics present, by analyzing the co-occurrence and distribution of words across documents.
  For the topic features, sentences with at least one instance of a topic word are flagged, and the count of the sentences for each topic is normalized by the total number of sentences in the transcript.

- **Topic Allocations - Directional** :

  The same methodology can be extended to directional descriptors. For the directional topic features, the total count of sentences with both a topic tag and a directional descriptor are normalized by the total number of sentences, thus constructing signals for directional market moving topics.

The calculation of directional features in the context of topic analysis aims to determine the sentiment associated with specific topics present in earning calls. This method examines the content of each document and evaluates sentences within the earning calls to assess the sentiment direction of various predetermined topics. The method follows a systematic approach by considering the presence of relevant words related to each topic, positive and negative sentiment indicators.

This process yields the average directional sentiment for each topic, providing insights into the prevailing sentiment associated with these specific financial indicators in the pharmaceutical industry's earning calls. By quantifying the sentiment direction of these market-moving topics the aim is to gain a deeper understanding of market sentiment, financial performance, and potential trends within the pharmaceutical industry.

4.2.6 Cosine Similarity

- Cosine similarities based on the LDA topic vectors are constructed by comparing the topic vectors of transcripts within the same company. For each transcript, the average similarity is computed over the past 5, 10, and 20 transcripts. Analysis of the feature importance of these features revealed that average cosine similarity over the past 10 calls has better predictive power.As a result, this particular feature is retained in the model.

- Cosine similarities based on the TF-IDF vectors are constructed by comparing the vectors of the focal transcript with all previous transcripts. The higher dimensionality of these vectors captures a richer embedding space, however the semantic similarity of these vectors is more difficult to interpret.

4.2.7 Graph Features

The embeddings for the speakers from the company are averaged to create a "company embedding" for each transcript. For the analysts in the transcript, the mean, the maximum and the minimum of the cosine similarities of each analyst with the company embedding were computed. The same procedure is repeated for the Euclidean distance between the company and analyst embeddings. The embedding space of the nodes reveals that the speaker embeddings cluster together, away from the nodes of the analysts, as in figure 5.

Filtering for fewer companies, from figure 6 the company embeddings for NVS tend to be more dispersed compared to those of PFE, and it's possible can also observe that several companies seem to cluster together in the embedding space, namely, BMY, MRK and PFE.

From this cursory analysis of the embedding space, it's clear that these spatial distances between the embeddings as a difference in contextual information contributed to the overall conversation by the speaker. The embeddings which are spaced closer are more likely to address similar topics, and therefore are closer together in the embedding space. This implies that if answers to questions posed by analysts are not detailed and extensive, or if the manager tries to avoid the question or otherwise obfuscate the truth, the embeddings will show more significant divergence.

This can be easily verified by taking the case of a single earnings call and observing the value of the metrics produced. For Novartis, from the fourth earnings call transcript in 2013, the presentation by the management fails to anticipate the questions brought up by the analysts. In the presentation, a general results presentation is given by the CEO who expresses

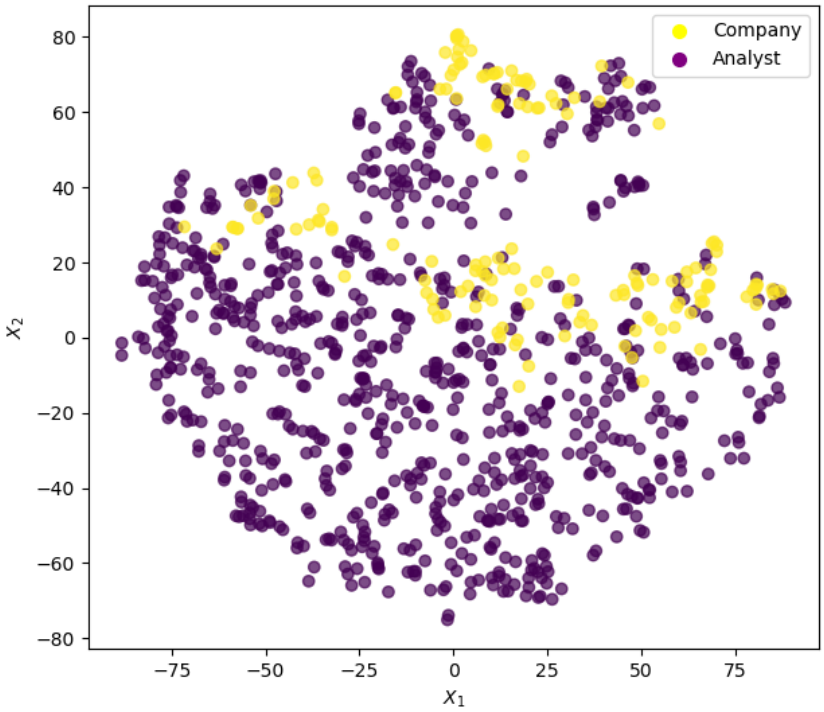

Figure 5: TSNE visualization of GraphSAGE embeddings for speaker nodes of both companies and analysts

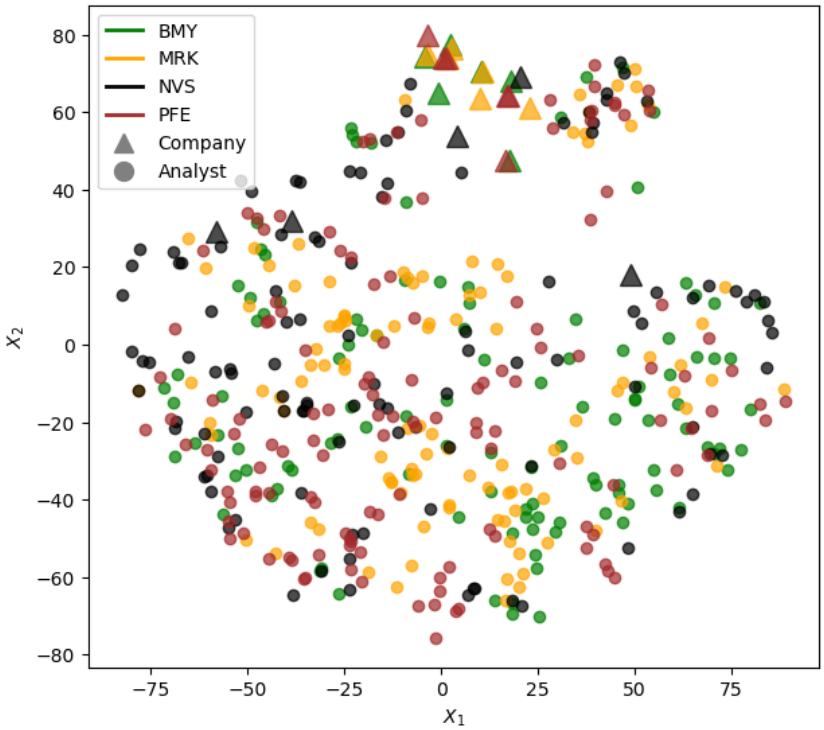

Figure 6: TSNE embeddings of speaker nodes for BMY, MRK, NVS and PFE

optimism for innovation and promises efficiency and integrity. *"...drive economic growth, a lot of people are saying look, a healthy population is key. What do we need to do is to improve the efficiency of our healthcare systems. So I think based on all of this, if you look at Novartis and the way that we are positioned, I think we're positioned well for the future. Looking at our results ..."*. However, the analysts were more focused on details including but not limited to specific products of Novartis: *"..Airflusal, the Danish approved the medium and the high dose equivalent of Advair, but it seems like the subsequent approvals have only really endorsed the higher dose. Can you explain to us why?.."*. This can easily explain the higher value of the Euclidean distance between the company and analyst embeddings.

4.3 Data Exploration

4.3.1 Relation between Earnings Calls and Stock Prices

This study aims to explore the potential relationship between earnings calls and stock price movements. As mentioned earlier, the target variable of interest is CAR (Cumulative Abnormal Return). To illustrate the analysis, Figure 7 presents a 42-day timeframe centered around the release of an earning call associated with the highest CAR value. The earning call is depicted as a green band in the figure.

The day prior to the publication of the earning call, there is a noticeable upward shift in the price, indicating a 5% increase. This observation suggests a potential anticipation or positive market sentiment surrounding the upcoming earning call. However, it is also possible to observe similar price behavior in the days following the release of the earning call.

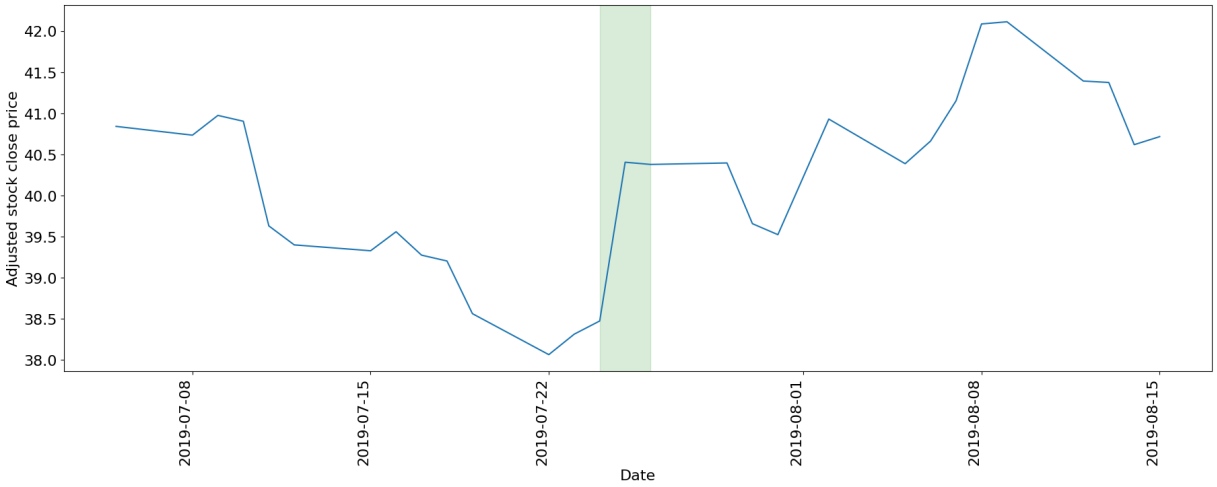

Figure 7: Earnings call release and stock price movement - BMY

4.3.2 Earnings calls - Trigrams

A trigram refers to a sequence of three consecutive words within a text or document.Trigrams capture the relationship and context between three adjacent words and can provide valuable insights into the language patterns, syntactic structures, and semantic meaning within a given text.

Trigrams are useful for text analysis because they offer a more nuanced understanding of the textual content compared to individual words or bigrams (sequences of two consecutive words). By considering the relationship between three words, trigrams capture richer context and can reveal hidden patterns, dependencies, and co-occurrences in the text.

To understand the potential correlations between patterns in earnings calls and company performance, an analysis framework was developed. Initially, a specific period was carefully chosen during which access to earnings calls from multiple companies was available. From this period, two companies were selected based on their stock price growth. Eli Lilly (LLY) was chosen to represent the company with higher growth, exhibiting a growth rate of 246%, while Pfizer (PFE) represented the company with lower growth, showing a growth rate of 77%. The performance for each company can is available in the table 5 in the appendix.

Figure 8 presents the trigrams extracted from the transcripts of all earnings calls conducted by LLY and PFE within the selected period. The trigrams captured in the earnings calls of LLY provide valuable insights into the specific areas of discussion that potentially contributed to their higher growth. Similarly, Figure 9 displays the trigrams extracted from the earnings calls of PFE, shedding light on the topics associated with their comparatively lower grow.

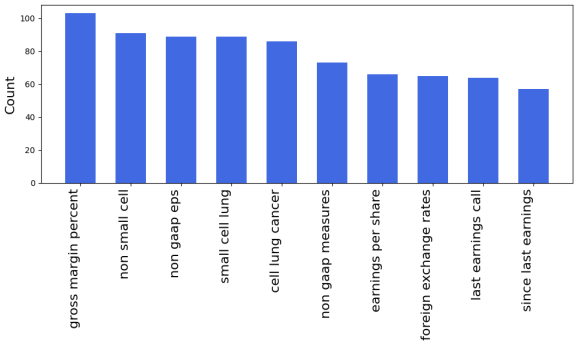

Figure 8: Top Trigrams - LLY - (2013-2018)

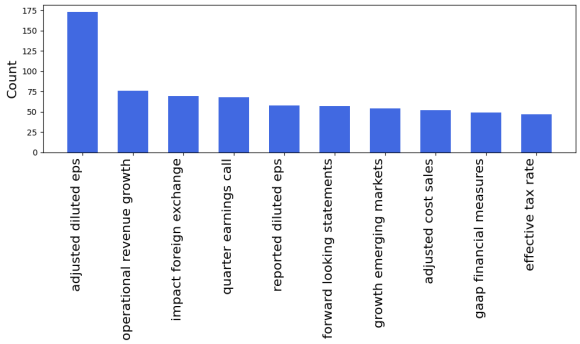

Figure 9: Top Trigrams - PFE - (2013-2018)

Through the analysis of trigrams associated with a company's performance over a span of seven years, valuable insights can be gained regarding its financial and operational aspects. A stark contrast can be observed between the two groups. In the case of LLY, the identified trigrams highlight topics related to cancer treatment, such as "non small cell lung cancer," alongside economic analysis indicators like "gross margin percent." On the other hand, PFE's trigrams predominantly revolve around economic subjects, including terms like "adjusted diluted EPS," "operational revenue growth," and "impact foreign exchange"

With the objective of analyzing the conversations surrounding the identified trigrams in greater detail, a selection of sentences was extracted. These statements offer valuable insights into the discussions and key topics of the company. Some notable statements include:

- *"Excluding this FX effect, our **gross margins percent** increased by nearly 220 basis points, going from 74.9% in last year's quarter to 77.1% this quarter, driven by manufacturing efficiencies.* - LLY(2017-04-26)
- *"If the lung cancer trial necitumumab and ramucirumab are positive, we will be positioned off a comprehensive treatment strategy for **non-small cell** lung cancer patient"* -

LLY(2013-01-29),

- *"Positive results from a Phase 3 study of Cyramza for first line EGFR* ***non-small cell lung cancer****"* - LLY(2019-04-30)
- *"As a result, foreign exchange negatively impacted second quarter* ***adjusted diluted EPS*** *by approximately $0.06 compared with the year-ago quarter."* - PFE(2015-07-28)
- *"In the second quarter, we delivered solid financial results with 2%* ***operational revenue growth****, 4% adjusted diluted EPS growth and 6% operational revenue growth in our Biopharmaceuticals Group compared with the year-ago quarter."* - PFE(2019-07-29)
- *"As I previously mentioned,* ***foreign exchange negatively impacted*** *first quarter 2017 revenues by approximately $116 million or 1% and positively impacted adjusted cost of sales, adjusted SI&A expenses, and adjusted R&D expenses in the aggregate by $61 million or 1%."* - PFE(2017-05-02)

For LLY, the statements highlight positive aspects of the company's performance, such as an increase in gross margin percent driven by manufacturing efficiencies and the potential for comprehensive treatment strategies in non-small cell lung cancer, similar as reported in (HUBERMAN; REGEV, 2001) where EntreMed's stock price pumped after a *New York Times* article. LLY is actively engaged in developing innovative therapies and medications to address various types of cancer and it´s research and development efforts focus on discovering novel treatment options, improving existing therapies, and advancing the understanding of cancer biology. These indications suggest a favorable outlook in their growth prospects. On the other hand, PFE's sentences are more focus in the performance of the company, mention the negative impact of foreign exchange on adjusted diluted EPS and revenue, albeit with solid operational revenue growth in certain segments. These findings imply that PFE may face challenges related to currency fluctuations but still demonstrates operational growth.

4.3.3 Relation of target variable with graph features

Figure 10 shows the relation between the target variable and the different graph features. The concentration of points in the middle of the scatter plots suggests that there is a tendency for the two variables to have values that are closer to their average or median. This could indicate a moderate level of correlation between the variables, and that the relation between those variables is a non-linear relation, making models that can capture this behavior more suitable for prediction.

5 Results

This section presents the main findings of this study. Nine different models were performed, which were formulated by combining the features outlined earlier. The models were defined considering the set of features. These models underwent testing using both ordinary least squares (OLS) and XGBoost algorithms. The Table 3 presented below showcases the details and outcomes of these models. The "X" represents the features used in each model.

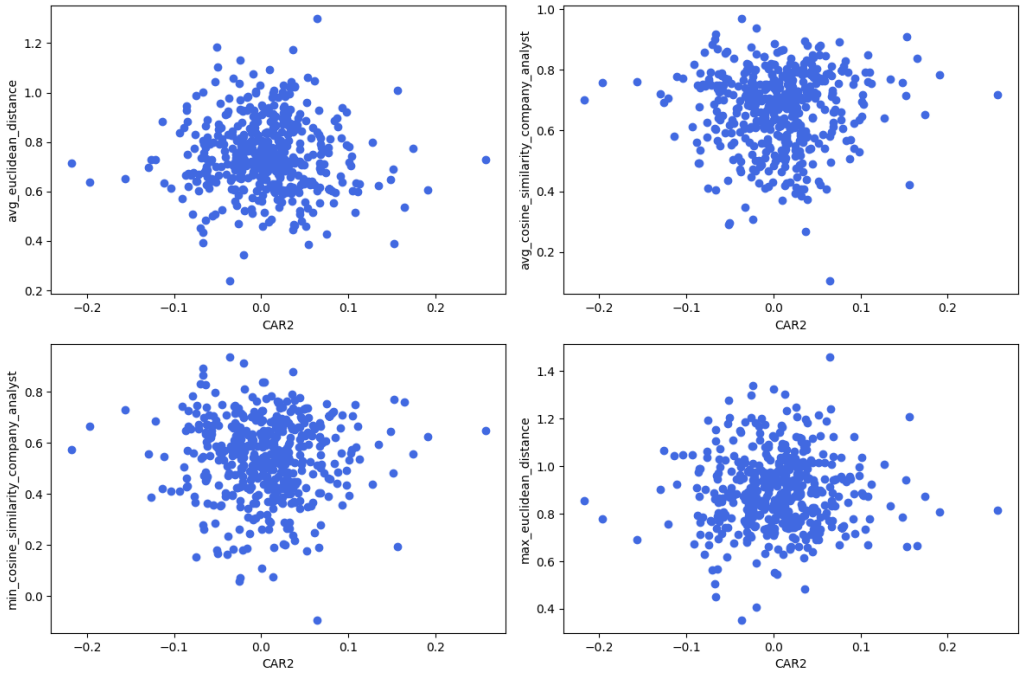

Figure 10: Scatter plot CAR - Graph features

The selection of features was conducted with the aim of evaluating the performance of each text feature category, namely Topic Features Directional, Cosine Features, and Graph Features, in relation to the Earnings Surprise parameter. Additionally, an incremental approach was adopted to incorporate the remaining text features into the best model. Furthermore, a separate model was built solely based on text features. The baseline model is selected as the model with solely earnings surprise as a feature, as previous works have determined that it is one of the best indicators to predict stock movement after an earnings call.

Features	Model								
	1	2	3	4	5	6	7	8	9
Earnings Surprise	X	X	X	X	X	X	X	X	
Technical Features + Dummies		X						X	
Topic Features Directional			X				X	X	X
Cosine Features				X		X	X	X	X
Graph Features					X	X	X	X	X

Table 3: Model description

5.1 Reported metrics

The results of all the models are presented in Table 4.

Overall, XGBoost models outperform the OLS models in both MSE across almost all models. The lower MSE values of the XGBoost model suggest that the model's predicted val-

	MSE		MAPE	
Model	XG_Boost	OLS	XG_Boost	OLS
Model_1	0.00260	0.00274	185.42077	192.21726
Model_2	0.00296	0.00291	210.24134	207.69373
Model_3	0.00274	0.00276	352.69783	180.67165
Model_4	0.00272	0.00270	155.35934	170.26032
Model_5	**0.00251**	0.00276	**172.35694**	192.48048
Model_6	0.00261	0.00270	157.82400	177.43963
Model_7	0.00271	0.00271	225.94707	155.86358
Model_8	0.00283	0.00287	194.06934	235.15643
Model_9	0.00261	0.00271	211.56809	150.54951

Table 4: Mean Squared Error performance and Mean Absolute Percentage Error.

ues deviate less from the actual values, indicating a stronger overall fit to the data. For the other hand, The lower MAPE values of the model indicate a smaller average percentage error between the predicted values and the actual values, highlighting its superior accuracy in capturing the true patterns in the data. Overall, these results indicate that the XGBoost model is superior to the OLS model in terms of predictive performance, which intuitively supports the proposition that the relation between abnormal return and model features is non-linear. Adding text features using cosine similarity measures and topic modeling were not able to beat the simpler baseline model based on the earnings surprise.

Among the XGBoost models, Model 5 exhibited the highest performance, with the lowest MSE and one of the lowest MAPE values. This suggests that using the surprise effect and the graph features is most effective in capturing the underlying patterns and trends in the data. Further model explainability and interpretation will be focused on this model. Considering MSE as the primary metric for comparison, only model 5 reported a better result than the baseline which takes into consideration only the earnings surprise element.

Model 6 and Model 9 reported a good performance, with a very similar MSE compared to the baseline. It's interesting to see that the information carried by the earnings surprise, the best indicator according to the literature to forecast stock abnormal return in the post-event window, can be achieved using only text features like in model 9. All of the XGBoost models that performed better or similar to the baseline contains the graph features, indicating this is the most relevant set of features for forecasting abnormal returns according to our model.

5.2 Model Explainability

Figure 11 provides a visual representation of the feature importance of the model. The F scores take into account both the frequency of the feature's appearance in the decision trees and the improvement in the model's performance when the feature is used for splitting. It shows that the "eps surprise" feature has the highest impact on the model and is considerably more important than the other features of the model.

An interesting observation from the Figure 12 with SHAP values, is the *avg_ euclidean_* -

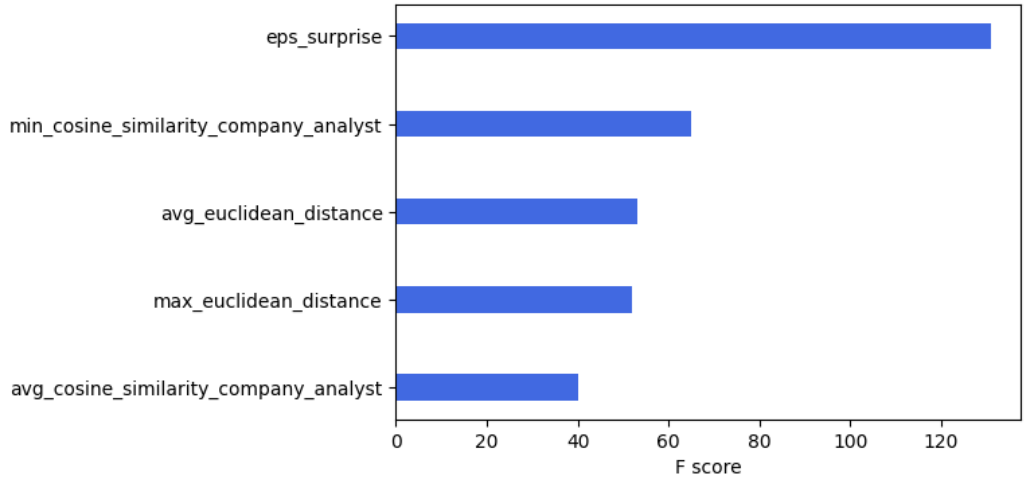

Figure 11: Feature importance - Best model

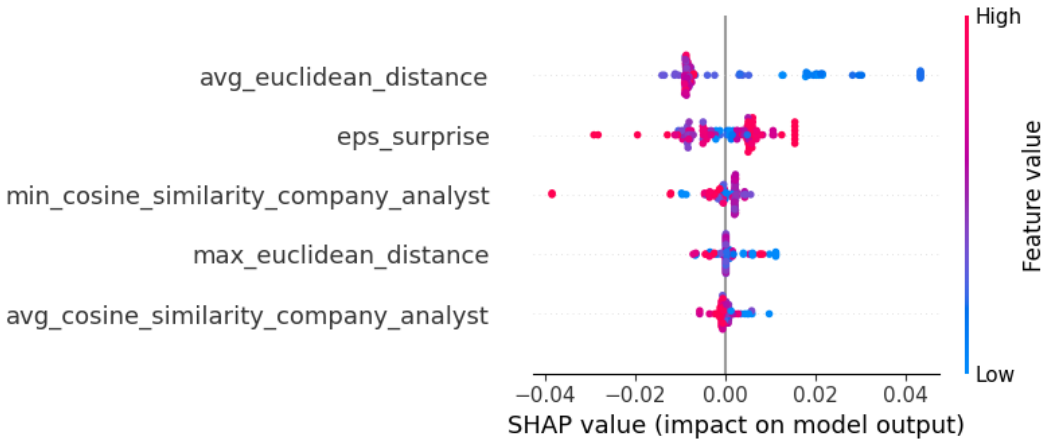

Figure 12: Model interpretability - SHAP values

distance feature, which measures the Euclidean distance between the analyst and management node embeddings. The position of the embeddings is influenced by the utterances of the speakers. From the SHAP values, if the distance is low (the management and the analysts talk about related topics or the managers have answered the questions asked by the speakers in sufficient detail so as to influence their respective speaker embeddings), the impact on the predicted Cumulative Abnormal Return (CAR), that is, the abnormal return in the next 30 days after the earnings call, is positive.

Some practical interpretations can be derived from these observations:

- The managers should prepare their presentations by anticipating the questions of the analysts better. Designing their speech around the topics that the analysts are interested in will ensure that their speaker embeddings are closer together.
- However it is important to recognize that this does not imply causation, as the managers could be trying to put a positive spin on a poor earnings quarter, and any embedding distance could just be indicative of managers trying to sideline questions that would show the situation in a poor light.

Another inference from Figure 12 is that a high positive surprise in the reported value of the EPS against the expected EPS positively impacts the predicted CAR, which is intuitive. Sparse high EPS values which are counter to this trend can be observed in the SHAP values.

To examine the local interpretability of the model on a specific earnings call, we apply LIME which can be seen in Figure 13. In the graph, each feature is represented by a bar, and the length of the bar indicates the magnitude and direction of its contribution. Positive values indicate that the feature positively contributes to the prediction, while negative values indicate a negative contribution. In this specific case, the value of the target variable was 0.01869 and the predictive value 0.01026.

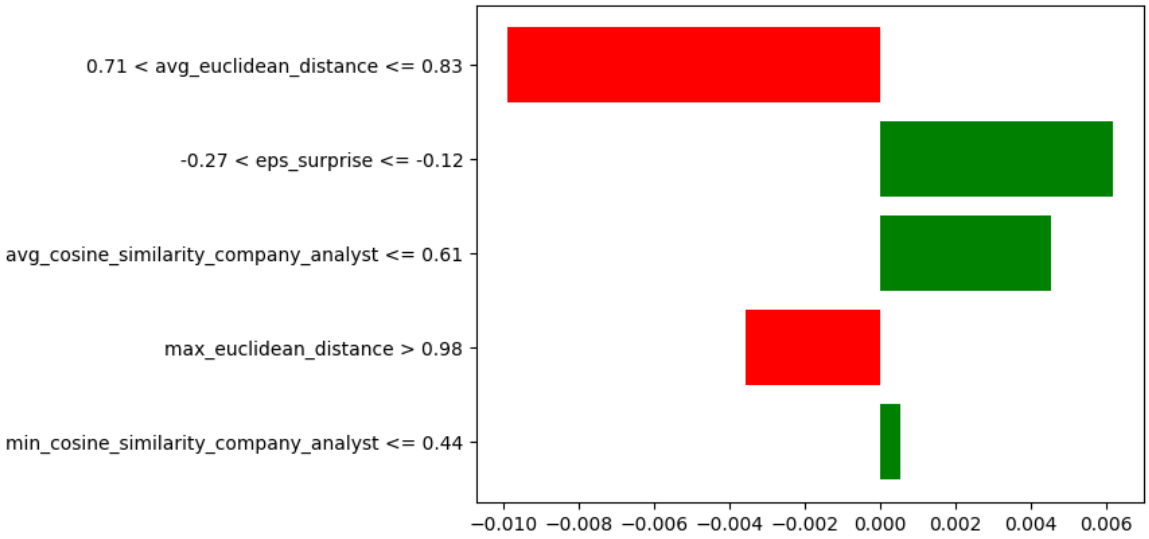

Figure 13: Model interpretability - Lime - NVS(2014-01-29)

This model, provides an intuitive and a counter-intuitive conclusion simultaneously. According to LIME, a high Euclidean distance impacts the predicted CAR negatively, while a negative surprise (EPS reported below market expectation) impacts the target positively. As noted previously, in the SHAP figures it is possible to find values that follow this counter-intuitive pattern for the EPS surprise as well, which apply to this specific observation. It is important to remember that compared to SHAP the LIME is a local interpretability technique, so the inferences from the LIME diagram can only be applied to the specific earnings call.

6 Conclusion

This study contributes to the literature by following a new approach to generate textual features from the earnings calls to forecast stock returns. Nine different models were developed and analyzed to predict Cumulative Abnormal Return (CAR) based on various features. The models were formulated by combining different sets of features, and their performance was evaluated using both ordinary least squares (OLS) and XGBoost algorithms. The results presented in Table 4 demonstrate that the XGBoost models consistently outperform the OLS model in terms of Mean Squared Error (MSE) while the OLS models report better results for

Mean Absolute Percentage Error (MAPE).

Among the XGBoost models, Model 5 emerged as the best performer, exhibiting the lowest MSE and the third-lowest MAPE. This indicates that incorporating the graph features to the earnings surprise yields the most accurate predictions of Cumulative Abnormal Returns. Furthermore, interpretability with SHAP values reveals that a low average Euclidean distance between the management and analysts, which indicates more significant overlap of topics discussed, positively influences the predicted CAR. This finding has practical implications for improving management-analyst communication during earnings calls.

In conclusion, the results highlight the importance of incorporating earnings surprise effects and graph features in capturing underlying patterns. The findings suggest that attempts to measure the alignment of manager and analyst discussions in an earnings call can be a good predictor of the abnormal return of the stock after an earnings call. Understanding the textual data surrounding an earnings call provides investors and market participants with valuable insights for making informed decisions.

6.1 Limitations and extensions

One of the limitations of our model lies in the target variable. Instead of the one-factor model which was used to calculate the Cumulative Abnormal Return, the use of a different approach like the 4-factor model of (JEGADEESH; TITMAN, 1993) could generate more robust results.
Sentence transformers used in the model currently rely on a general-purpose embedding model that has a competitive benchmark performance. An interesting comparison could be made of the value in using models fine-tuned for the financial domain, like FinBERT. Fine-tuned speaker level embeddings for the prediction level task with a supervised graph neural network using more data could improve the predictive power of the embedding space. Graph labeling of speakers currently relies on the likely positions of the management speakers in the earnings call transcript. Scraping official documents to obtain the proper names of the board members, thereby enhancing the labeling method in the graph could ensure that the graph construction is more robust. Furthermore, an analysis can be conducted under a causality framework to explore whether the proximity of nodes in the earnings call data influences abnormal returns, or if the relationship works in the opposite direction.

References

AKITA, Ryo et al. Deep learning for stock prediction using numerical and textual information. In: IEEE. 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS). [S.l.: s.n.], 2016. P. 1–6.

ARTS, Sam; HOU, Jianan; GOMEZ, Juan Carlos. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. **Research Policy**, v. 50, n. 2, p. 104144, 2021. ISSN 0048-7333. DOI: https://doi.org/10.1016/j.respol.2020.104144. Available from: <https://www.sciencedirect.com/science/article/pii/S0048733320302195>.

BALL, Ray; BROWN, Philip. An empirical evaluation of accounting income numbers. **Journal of accounting research**, JSTOR, p. 159–178, 1968.

CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. In: PROCEEDINGS of the 22nd acm sigkdd international conference on knowledge discovery and data mining. [S.l.: s.n.], 2016. P. 785–794.

CHENG, Dawei et al. Financial time series forecasting with multi-modality graph neural network. **Pattern Recognition**, Elsevier, v. 121, p. 108218, 2022.

DATA61, CSIRO's. **StellarGraph Machine Learning Library**. [S.l.]: GitHub, 2018. https://github.com/stellargraph/stellargraph.

DESSAIN, Jean. Machine learning models predicting returns: Why most popular performance metrics are misleading and proposal for an efficient metric. **Expert Systems with Applications**, Elsevier, v. 199, p. 116970, 2022.

FAMA, Eugene F. Efficient capital markets: A review of theory and empirical work. **The journal of Finance**, JSTOR, v. 25, n. 2, p. 383–417, 1970.

FAUZAN, Muhammad Arief; MURFI, Hendri. The accuracy of XGBoost for insurance claim prediction. **Int. J. Adv. Soft Comput. Appl**, v. 10, n. 2, p. 159–171, 2018.

FINK, Josef. A review of the post-earnings-announcement drift. **Journal of Behavioral and Experimental Finance**, Elsevier, v. 29, p. 100446, 2021.

GHOSAL, Deepanway et al. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In: PROCEEDINGS of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, Nov. 2019. P. 154–164. DOI: 10.18653/v1/D19-1015. Available from: <https://aclanthology.org/D19-1015>.

GHOSAL, Deepanway et al. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. **CoRR**, abs/1908.11540, 2019. arXiv: 1908.11540. Available from: <http://arxiv.org/abs/1908.11540>.

HAMILTON, William L.; YING, Rex; LESKOVEC, Jure. Inductive Representation Learning on Large Graphs. **CoRR**, abs/1706.02216, 2017. arXiv: 1706.02216. Available from: <http://arxiv.org/abs/1706.02216>.

HUBERMAN, Gur; REGEV, Tomer. Contagious speculation and a cure for cancer: A nonevent that made stock prices soar. **The Journal of Finance**, Wiley Online Library, v. 56, n. 1, p. 387–396, 2001.

JEGADEESH, Narasimhan; TITMAN, Sheridan. Returns to buying winners and selling losers: Implications for stock market efficiency. **The Journal of finance**, Wiley Online Library, v. 48, n. 1, p. 65–91, 1993.

JIANG, Chuntao et al. Text classification using graph mining-based feature extraction. **Knowledge-Based Systems**, v. 23, n. 4, p. 302–308, 2010. Artificial Intelligence 2009. ISSN 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2009.11.010. Available from: <https://www.sciencedirect.com/science/article/pii/S095070510900152X>.

KAESTNER 1, Michael. Anomalous Price Behavior Following Earnings Surprises: Does Representativeness Cause Overreaction? **Finance**, Cairn/Softwin, v. 27, n. 2, p. 5–31, 2006.

LEDLEY, Fred D et al. Profitability of large pharmaceutical companies compared with other large public companies. **Jama**, American Medical Association, v. 323, n. 9, p. 834–843, 2020.

LEE, Heeyoung et al. On the Importance of Text Analysis for Stock Price Prediction. In: LREC. [S.l.: s.n.], 2014. v. 2014, p. 1170–1175.

LINTNER, John. The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. **The Review of Economics and Statistics**, The MIT Press, v. 47, n. 1, p. 13–37, 1965. ISSN 00346535, 15309142. Available from: <http://www.jstor.org/stable/1924119>. Visited on: 1 July 2023.

LISBOA, Vicente; SANTOS, Lucas; THOMAS, Davis. **Historical EPS and expected EPS**. [S.l.: s.n.], 2023. Thomson Reuters Eikon. Retrieved from https://www.eikon.com.

MIMNO, David et al. Optimizing Semantic Coherence in Topic Models. In: PROCEEDINGS of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK.: Association for Computational Linguistics, July 2011. P. 262–272. Available from: <https://aclanthology.org/D11-1024>.

REIMERS, Nils; GUREVYCH, Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. **CoRR**, abs/1908.10084, 2019. arXiv: 1908.10084. Available from: <http://arxiv.org/abs/1908.10084>.

RIFFE, Daniel et al. **Analyzing media messages: Using quantitative content analysis in research**. [S.l.]: Routledge, 2019.

S&P GLOBAL MARKET INTELLIGENCE. **Natural Language Processing III: Application of Natural Language Processing Techniques to Earnings Call Transcripts**. [S.l.], 2020. Available from: <https://www.spglobal.com/marketintelligence/en/documents/nlp-iii-final-013020-10a.pdf>.

SALAMAT, Sara et al. **Text Representation Enrichment Utilizing Graph based Approaches: Stock Market Technical Analysis Case Study**. [S.l.: s.n.], 2022. arXiv: 2211.16103 [cs.LG].

SHARPE, William F. Capital asset prices: A theory of market equilibrium under conditions of risk. **The journal of finance**, Wiley Online Library, v. 19, n. 3, p. 425–442, 1964.

SHEN, Weizhou et al. Directed Acyclic Graph Network for Conversational Emotion Recognition. **CoRR**, abs/2105.12907, 2021. arXiv: 2105.12907. Available from: <https://arxiv.org/abs/2105.12907>.

TOWNE, W. Ben; ROSÉ, Carolyn P.; HERBSLEB, James D. Measuring Similarity Similarly: LDA and Human Perception. **ACM Trans. Intell. Syst. Technol.**, Association for Computing Machinery, New York, NY, USA, v. 8, n. 1, Sept. 2016. ISSN 2157-6904. DOI: 10.1145/2890510. Available from: <https://doi.org/10.1145/2890510>.

VELIČKOVIĆ, Petar et al. **Deep Graph Infomax**. [S.l.: s.n.], 2018. arXiv: 1809.10341 [stat.ML].

A Appendix

A.1 Stock price distribution

The following table shows the stock price distribution of the different companies between 2013 and 2020.

Ticker	Initial Date	Initial price	Final Date	Final price	Growth
LLY	2013-01-29	42.44	2020-07-30	146.73	245.76
AZN	2013-01-31	16.16	2020-07-30	53.82	233.10
JNJ	2013-01-22	54.97	2020-07-16	138.94	152.75
ABBV	2013-04-26	30.52	2020-10-30	76.72	151.37
ROG	2013-02-20	46.35	2020-02-21	110.18	137.71
MRK	2013-02-01	28.86	2020-10-27	68.55	137.51
BMY	2013-01-24	26.67	2020-11-05	58.98	121.14
NVO	2013-01-31	29.93	2020-08-07	61.47	105.34
NVS	2013-01-23	40.50	2020-07-21	76.07	87.85
PFE	2013-01-30	18.24	2020-10-27	32.47	77.99

Table 5: Stock price growth - Pharmaceutical companies

A.2 XGBoost

The XGBoost is a scalable tree boosting system introduced by (CHEN; GUESTRIN, 2016). The objective function at time t is described in 3.

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i) + \Omega(f_t) \tag{3}$$

Since the learners cannot be optimized using traditional optimization methods in Euclidean space it's necessary to train the model in a additive manner.

One of the main differences of XGBoost to other methods is the approach taken to split the nodes, while the other methods use metrics like Gini or Entropy the XGBoost uses the similarity score. Also, XGBoost uses advanced L1 and L2 regularization, which improves the model general performance in comparison with Gradient Boosting.

A.3 Graph Representation

The StellarGraph representation is a directed multigraph with 48,010 nodes and 127,518 edges over all transcripts, in which there are 38,361 text nodes and 9629 speaker nodes. Of these the text nodes have a feature of length 768, which are the text embeddings. A position level encoding of the speakers is also assigned to the speaker nodes.
HinSAGE node generators are capable of sampling and aggregating features of different dimensionality considering the different node types, which account for the handling of these feature vectors. The training loss curve of this model can be seen in Figure 15.

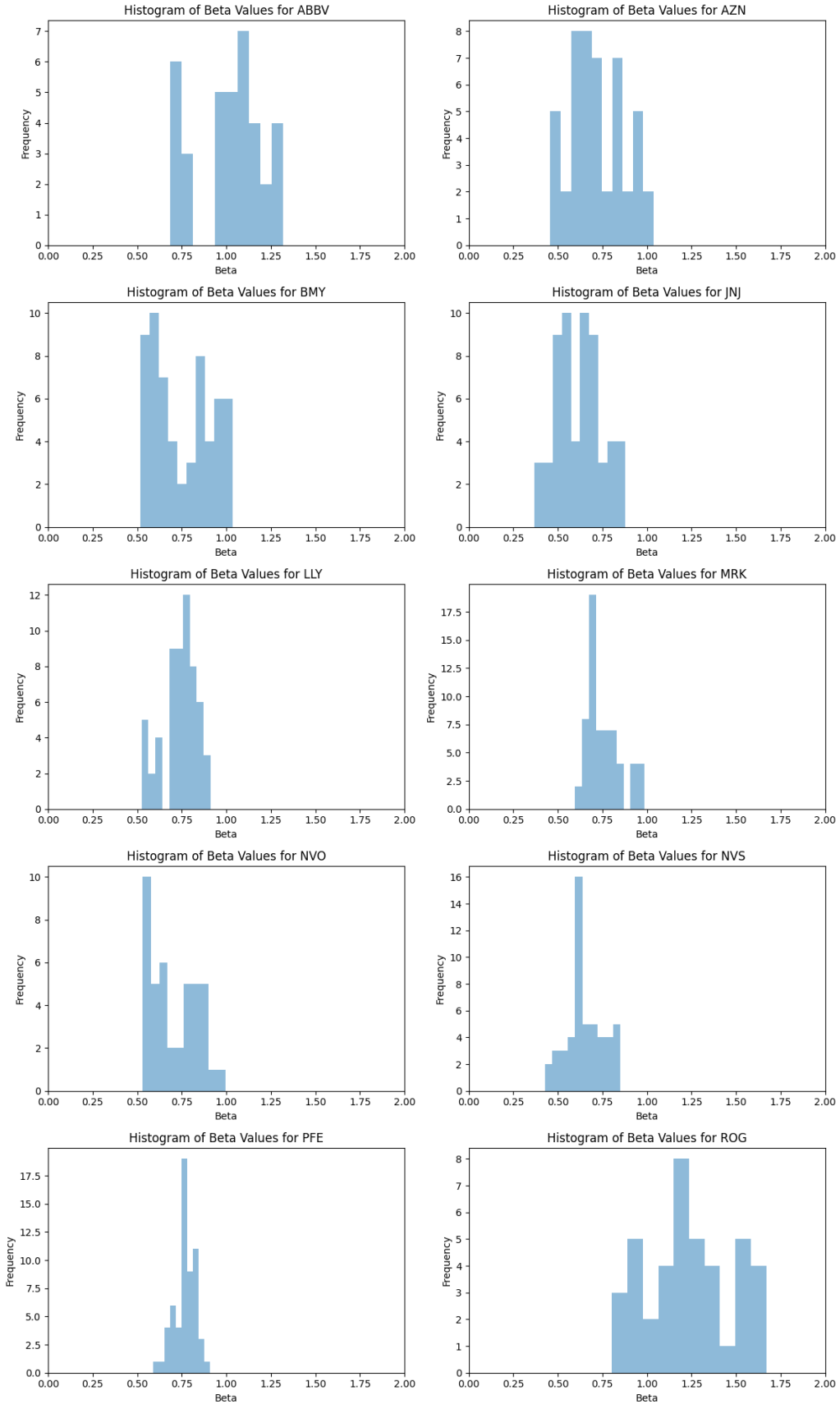

Figure 14: Distribution of the estimated Beta for different stocks considering the 720 days window

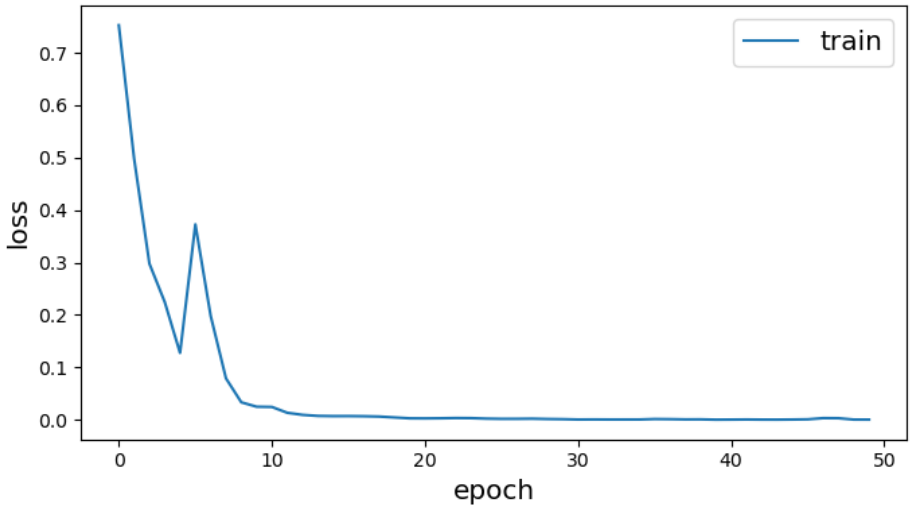

Figure 15: Loss curve of HinSAGE training using DeepGraphInfoMax