ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

SENTIMENT ANALYSIS OF EARNING CALLS TRANSCRIPTS

FIN-407

SPRING SEMESTER 2024

GROUP 2

LOÏC NICOLLERAT

FRANCESCO PIO NUMERO

TEO DE XUAN JUSTIN

DAVID VENCATO

Submission Date : 7 June 2024

Chapter 1

Introduction

In this project, we review literature to understand the use of sentiment in finance and trading, as well as current methods to analyse sentiment of financial reports. We explore the use of K-means clustering, lexicon-based analysis (LBA) and topic modeling to impute sentiment labels to earnings call data and discuss their strengths and weaknesses. Finally, we constructed various trading strategies utilising sentiment scores and tested their performance, but none were found to outperform the market. We analyse the reasons behind their underperformance and limitations to our trading strategies, and posit various improvement measures.

1.1 Review of literature

Earnings call transcripts provide a wealth of qualitative information that can significantly impact investor sentiment and market behavior. Over the years, sentiment analysis has emerged as a powerful tool to systematically extract and quantify insights coming from the language and tone used by executives, facilitating better investment decisions and market predictions.

In the seminal work of Tetlock [6], he demonstrated that negative sentiment in media articles could predict stock price, highlighting the potential of sentiment analysis in financial contexts . Next year, in [3] was published an application of sentiment analysis to earnings call transcripts, providing evidence that the sentiment expressed in these calls could predict future stock returns. This research emphasized the importance of using domain-specific lexicons to capture the nuances of financial language accurately. Building on this foundation, Loughran and McDonald [5] developed a sentiment lexicon specifically tailored to financial documents, addressing the limitations of general-purpose lexicons and significantly improving the accuracy of sentiment analysis in earnings call transcripts and other financial texts.

Further advancements were made in [2], where authors proposed combining domain-specific lexicons with machine learning techniques to enhance sentiment detection in financial texts.

Given the proven impact of sentiment analysis on financial forecasting, the next step involves the rigorous preprocessing of earnings calls.

1.2 Data Preprocessing

With sentiment analysis, the raw text derived from earnings call transcripts necessitates meticulous preprocessing to ensure the reliability and accuracy of subsequent machine learning models. Given the nuances and idiosyncrasies inherent in financial discourse, a multi-faceted approach was adopted.

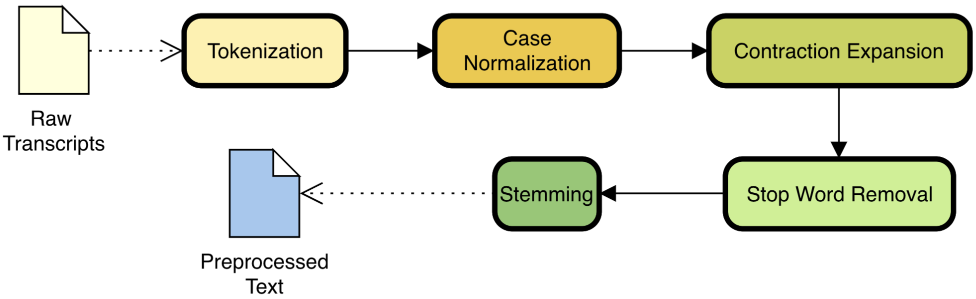


FIGURE 1.1
Text preprocessing

Initially, we extracted from our dataset only the transcripts of the companies present in the S&P500 index for each year so that the stocks now have higher trading volumes and computation time to work on them is quicker without loss of relevant information. Then, the transcripts underwent tokenization, a process that segmented the continuous text into discrete units known as tokens. This involved delineating punctuation from words and utilizing whitespace as the primary delimiter. To further standardize the text, case normalization was employed, rendering all characters lowercase. This mitigated the potential for the model to interpret capitalized and lowercase variations of the same word as distinct entities. Contractions, such as "don't" and "isn't," were expanded to their full forms, enhancing the granularity of subsequent analysis.

Moreover, stop words—common words like "the," "a," and "an"—were systematically removed. While these words contribute to grammatical structure, their prevalence renders them less informative for discerning financial sentiment. Stemming, a rudimentary form of lemmatization, was then implemented to reduce words to their base or root form. This consolidated semantically related terms, such as "investing," "investment," and "investor," under a unified stem. The resultant vocabulary, characterized by greater parsimony and conceptual cohesion, served as the foundation for subsequent sentiment analysis.

In sum, the comprehensive text preprocessing pipeline implemented in this study served to transform the raw, unstructured text of earnings call transcripts into a format amenable to machine learning analysis. This meticulous preparation ensured the integrity and reliability of the subsequent sentiment analysis, thereby augmenting its utility in financial decision-making processes.

1.3 SUMMARY STATISTICS ON THE DATA

The dataset comprises earnings call transcripts, with each entry containing the date, exchange, ticker, quarter, and the full transcript (prepared remarks, Q&A session, and attendees). In total, there are 17,221 transcripts spanning from 2019 to 2023, covering 2,876 unique tickers.

Figure 1.2 shows the distribution of the 30 largest tickers in the dataset. This distribution is relatively homogenous, which is good because no ticker will have too strong an influence on the dataset.

However, Figure 1.3 clearly shows a disparity in the distribution of transcripts from year to year, with a strong dominance of the year 2021. No measures have been taken to balance this non-homogeneity, as the publication date should not strongly influence the result of a sentiment analysis. However, this non-homogeneous distribution should be kept in mind.

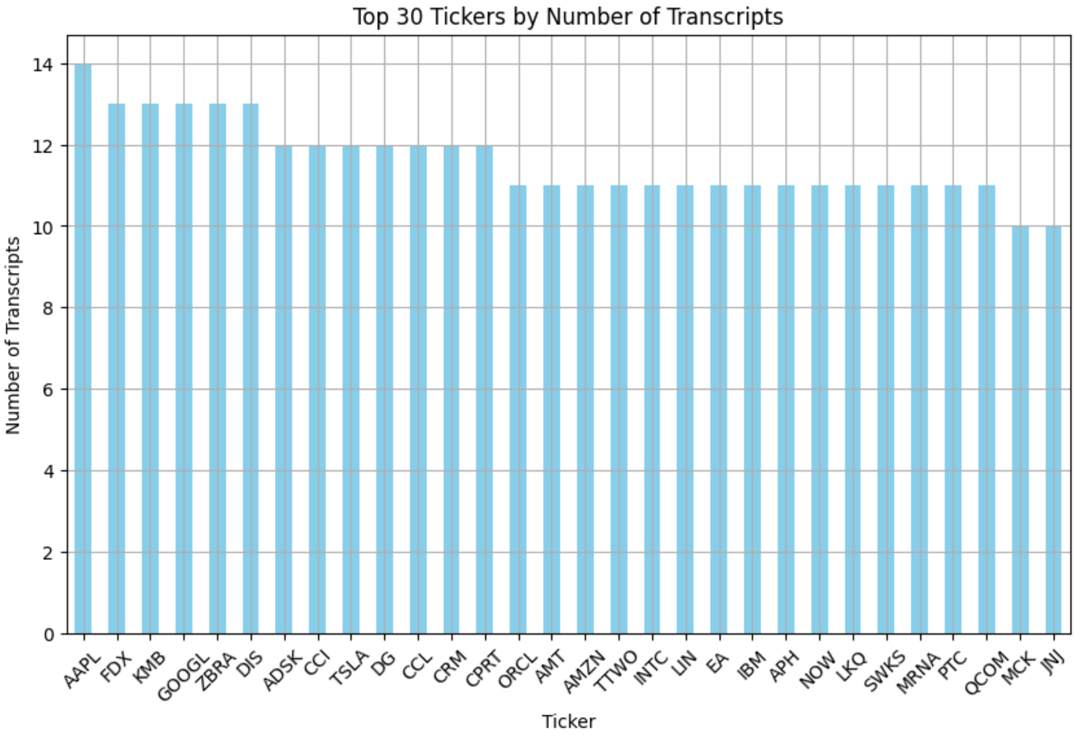


FIGURE 1.2
Top 30 tickers

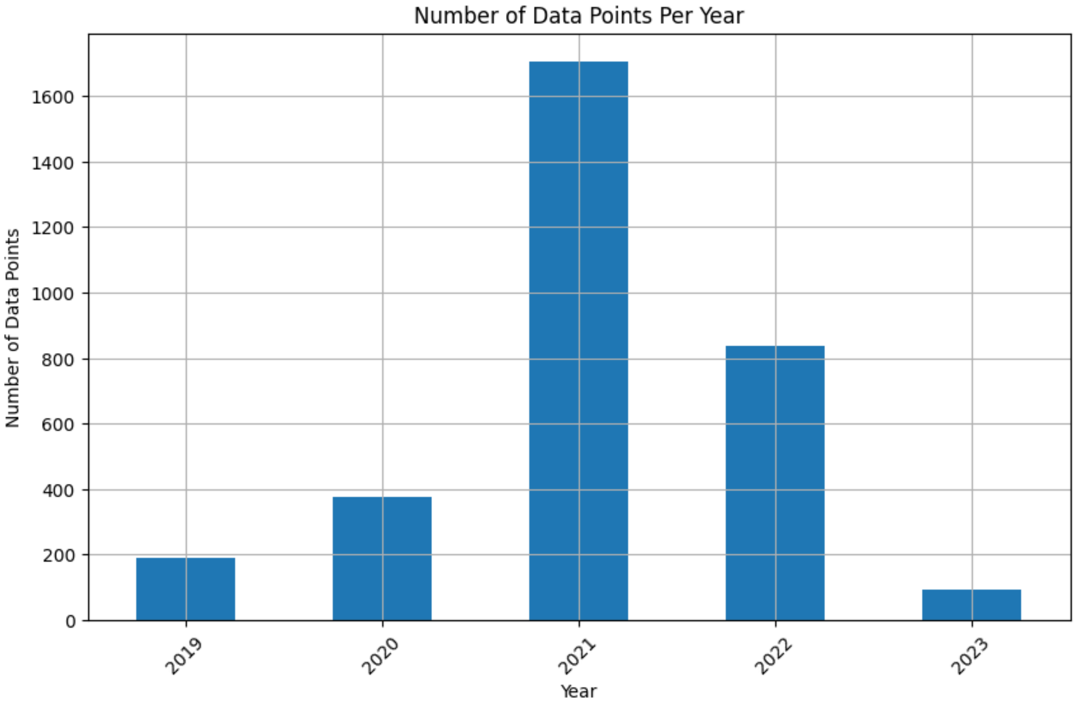


FIGURE 1.3
Years transcipts repartition

1.4 GOALS

In the previous sections, we explored relevant literature, prepared and processed the earnings call transcript data, and summarized key statistical findings. Building on this foundation, we now outline our goals for the analysis.
First, we are going to label the preprocessed dataset using three different unsupervised methods: K-mean clustering, Lexicon Based Analysis, and Topic Modeling.
Then, through the sentiment labels obtained, we aim to construct a portfolio and derive different trading strategies. This step will involve translating the sentiment analysis into actionable investment decisions.
Finally, we will interpret the results by evaluating the performance of the portfolio and understanding the influence of sentiment on stock returns. This evaluation will help us assess the effectiveness of our approach and its potential impact on financial forecasting and decision-making.

CHAPTER 2

METHODS

2.1 K-MEAN CLUSTERING

K-Means is a popular clustering algorithm that partitions a dataset into K distinct, non-overlapping clusters [1]. It aims to minimize the variance within each cluster, making the intra-cluster data points as similar as possible while maximizing the distance between different clusters. The algorithm starts with a random selection of K centroids and iteratively assigns each data point to the nearest cluster, recalculating the centroids based on the means of the data points in each cluster until convergence is achieved.

2.1.1 TF-IDF METHOD

The Term Frequency-Inverse Document Frequency (TF-IDF) method is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. It is primarily used to convert text data into a format that can be easily analyzed and processed by machine learning algorithms, a process known as vectorization. [4]

Why Vectorization?
Vectorization is essential in natural language processing because it transforms text into a numerical format, allowing algorithms to perform calculations. Most machine learning algorithms, including K-Means clustering, cannot directly process raw text; they require numerical input data. Here's why vectorizing text data is crucial:

- **Standardization:** Vectorization standardizes text data into a consistent format (numerical vectors), enabling the application of statistical and machine learning methods.
- **Importance Weighting:** TF-IDF helps highlight the words that are more relevant for analysis by reducing the weight of common words that appear frequently across documents but are less informative.
- **Dimensionality Reduction:** It helps in reducing the dimensionality of the text data by focusing only on relevant terms rather than the entire set of possible words.

The TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents that contain the word, which helps to adjust for the fact that some words appear more frequently in general.

$$\text{TF-IDF} = (\text{TF}) \times (\text{IDF}), \tag{2.1}$$

where $\text{TF} = \frac{\text{Number of times term appears in a document}}{\text{Total number of terms in the document}}$, and $\text{IDF} = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing the term}}\right)$.

This formula ensures that terms are properly weighted according to their importance and rarity across the document set, making TF-IDF a powerful tool for text analysis and sentiment analysis tasks.

2.1.2 RESULTS

The performance of the K-Means clustering algorithm was evaluated using both the Elbow Method and the Silhouette Score. These two metrics were plotted together in a single figure to provide a comprehensive view of the clustering performance.

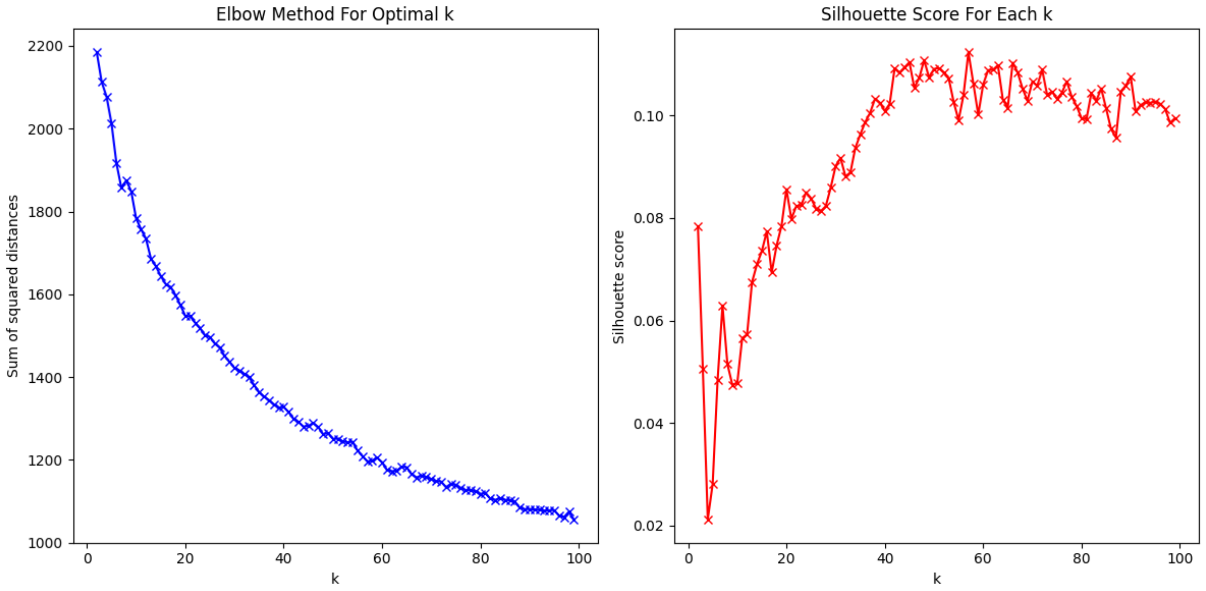


FIGURE 2.1
Combined plots showing the Elbow Method and Silhouette Scores. The absence of a clear elbow and low silhouette scores indicate poor clustering performance.

The figure (Figure 2.1) illustrates that there is no distinct elbow, which complicates the determination of an optimal number of clusters. Additionally, the Silhouette Scores are uniformly low, suggesting that the clusters formed do not have well-defined, cohesive structures. These findings indicate that the earnings call transcripts do not naturally segregate into distinct sentiment groups using the K-Means clustering approach.

This outcome might be due to the inherent complexities and subtleties of natural language used in financial documents, as well as the limitations of K-Means in capturing the nuanced differences in sentiment expressions without contextual cues. Further analysis with advanced techniques that consider semantic understanding may provide better results.

2.1.3 STRENGTHS AND WEAKNESSES

Strengths:

- **Simplicity and Scalability:** K-Means is straightforward to implement and scales well to large datasets.
- **Efficiency:** It is computationally efficient in medium-sized datasets, especially when the number of dimensions and clusters is moderate.

Weaknesses:

- **Assumption of Spherical Clusters:** K-Means assumes that clusters are spherical and of similar size, which might not be the case in text data.
- **Sensitivity to Initialization and Noise:** The algorithm is sensitive to the initial choice of centroids and is easily influenced by noise and outlier data points.
- **Difficulty in Handling Non-Numerical Data:** While TF-IDF helps in transforming text to numerical data, K-Means still struggles with the nuanced and contextual nature of text data.
- **Lack of Contextual Understanding:** K-Means and TF-IDF are not designed to understand the context within which words are used. This is particularly problematic in financial texts, where context heavily influences meaning. Phrases that might indicate financial distress or success often require understanding of the broader financial situation, industry trends, and economic conditions, which these methods cannot comprehend.

These characteristics suggest that while K-Means can provide preliminary insights into sentiment categorization, it falls short in applications requiring deep understanding of text semantics and context, such as detailed sentiment analysis in financial documents. Alternative techniques that can incorporate contextual understanding, like NLP models with embeddings that capture semantic meaning, may be more effective.

2.2 LEXICON BASED ANALYSIS

Lexicon-based analysis is a popular method for extracting meaningful information from texts using dictionaries of words (lexicons) associated with specific sentiments. We use this method to analyze and label our dataset of earning call transcripts. Our algorithm is based on Loughran-McDonald dictionary. In this view, we will examine the construction of this dictionary to better understand our algorithm.

The Loughran-McDonald dictionary

It is a specialized lexicon designed to capture sentiment in financial texts more accurately than general-purpose sentiment dictionaries. It was developed by Tim Loughran and Bill McDonald and first introduced in their 2011 paper [5].

They identified terms that frequently appear in financial contexts trying to understand the typical connotations that they have in financial documents and reviewed a subset of the most frequent words to classify them into sentiment categories.

The specific sentiment categories are:

- Positive [1]: words that typically convey a positive sentiment in financial contexts, such as "growth" and "profit."
- Negative [2]: words that convey negative sentiment, such as "loss" and "risk."
- Uncertainty [3]: words that indicate uncertainty or ambiguity, like "approximately" and "uncertain."
- Litigious [4]: words related to legal and regulatory matters, for example "litigation" and "regulatory."
- Constraining [5]: words indicating limitations, restrictions, or factors that may hinder performance or growth, often associated with negative sentiment and may include terms such as "restricted," "limited".
- Strong Modal [6]: words that indicate strong conviction or necessity, such as "must" and "will."
- Weak Modal [7]: words that indicate weak conviction or possibility, for example "may" and "could."

Given the absence of words categorized as ’strong modal’ or ’weak modal’ in our dataset, these two sentiments were not considered in the algorithm implementation.

2.2.1 ALGORITHM

The steps of our implemented algorithm are:

- Assign a sentiment vector v to each datapoint D. Each entries v_i is obtained by

$$v_i = \sum_{\substack{\omega \text{ word} \\ \omega \in \text{ D}}} \mathbb{1}_{\{\text{L.M. = i}\}}(\omega)$$

  where $i \in \{1, ..., 5\}$ and $L.M.$ is seen as a function that takes a word and returns its sentiment;

- we compute the weighted average of the entries to get the label of D:

$$label = \frac{w^T v}{|D|}$$

  where w is the weights vector defined as:

$$w^T = (1, -1, -0.5, -0.7, -0.8).$$

Remark 2.2.1 *The values of the weights are chosen in order to reflect the degree of which sentiments are considered unfovarable.*

In the result we will see that the number of positive words is dominant in the earning transcript calls, but the final weighted score is not close to 1. This happens because the weights work well with data and allow us to treat them in a more realistic way. The heterogeneity of the labels is one of the strengths of this method. We analyze them in detailed, together with the weaknesses, in the following section.

	Pos.	Uncert.	Neg.	Litig.	Constr.
Doc1	110	55	48	12	6

TABLE 2.1
Sentiment vector of transcript 1.

2.2.2 STRENGTHS AND WEAKNESSES

Interpretability of labels VS Struggles with understanding contex

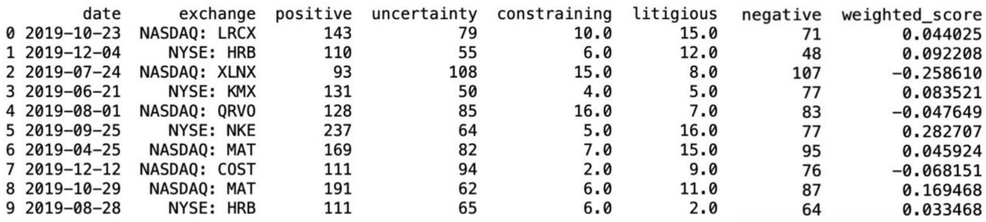
- The Lexicon Based Analysis offers simplicity and interpretability, in fact its computational cost is not expensive and the labels are derived directly from the presence of specific words in the lexicon, so that everyone can understand and interpret how the sentiment score is calculated, increasing trust and confidence in the results.
- On the other hand, it presents limitations in understanding the context of documents because of the assignment of a value to each word not to take into account where the word appears in the corpus. For instance, the word "growth" might be positive in one context but negative in another if discussing unsustainable growth. This context insensitivity can lead to inaccuracies in sentiment scoring.

Easy comparison of labels VS Difficulty in sentiment nuances

- The method provides a weighted sentiment score between -1 and 1, which offers a clear and quantifiable measure of the overall sentiment in a transcript. This numerical output facilitates comparison across different documents and time periods.

- Earning transcript calls, by their nature, are generally positive and their weighted score will always tend to be positive, despite the penalization given by negative sentiments. In conclusion, it is not easy to distinguish finer sentiment nuances between documents and a trading strategy based on this method may turn out to be bad.

In our case, despite the high number of positive words (see Table 2.2), the contribution of the weights associated with negative feelings worked well, allowing the labels to concentrate more in a range between -0.4 and 0.4 (Fig. 2.2) with a Gaussian distribution (for more details see in the Appendix 4)



```
         date      exchange  positive  uncertainty  constraining  litigious  negative  weighted_score
0 2019-10-23  NASDAQ: LRCX       143           79          10.0       15.0        71        0.044025
1 2019-12-04     NYSE: HRB       110           55           6.0       12.0        48        0.092208
2 2019-07-24  NASDAQ: XLNX        93          108          15.0        8.0       107       -0.258610
3 2019-06-21     NYSE: KMX       131           50           4.0        5.0        77        0.083521
4 2019-08-01  NASDAQ: QRVO       128           85          16.0        7.0        83       -0.047649
5 2019-09-25     NYSE: NKE       237           64           5.0       16.0        77        0.282707
6 2019-04-25   NASDAQ: MAT       169           82           7.0       15.0        95        0.045924
7 2019-12-12  NASDAQ: COST       111           94           2.0        9.0        76       -0.068151
8 2019-10-29   NASDAQ: MAT       191           62           6.0       11.0        87        0.169468
9 2019-08-28     NYSE: HRB       111           65           6.0        2.0        64        0.033468
```

TABLE 2.2
The first ten sentiment vectors.

Different weights to different sentiments VS Equal weights to all words

- We are not interested just in "positive" and "negative" sentiment but also in other categories, such as "litigious", "constraining", "uncertainty". These sentiments provide a more granularity and grade of details in the data.

- Unfortunately, even if we assign different weights to different sentiment, the method can not accurately reflect the strength of the sentiment, treating all words equally. For instance, words like "excellent" and "good" both indicate positive sentiment, but "excellent" suggests a stronger positive feeling. So we don't have a differentiate between levels of positivity, negativity and other used sentiments. This can lead to misleading results, especially if certain words are more impactful than others.

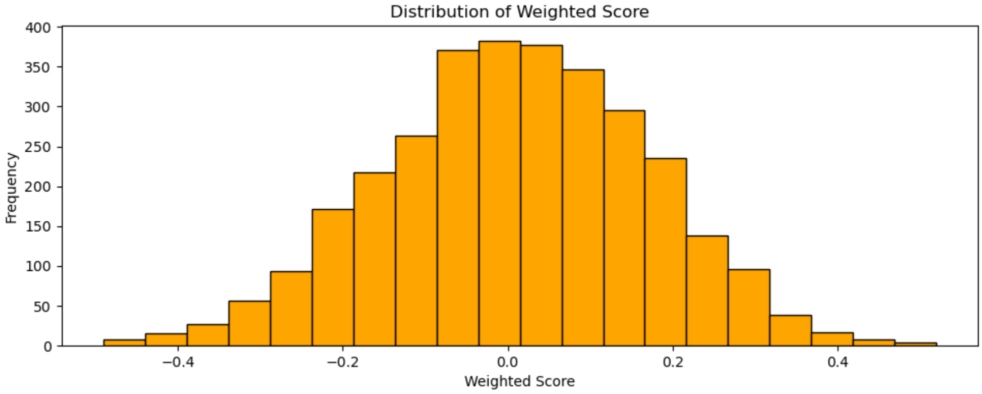


FIGURE 2.2
Distribution of "Weighted Score" label.

2.3 TOPIC MODELING

Topic modeling is a type of statistical model used to discover abstract topics within a collection of documents. It aims to identify patterns in the use of words and phrases, grouping them into one or more topics that capture the underlying themes in the text. These topics help in understanding large volumes of textual data by revealing the structure that pervades a corpus. Since our algorithm for generating labels relies on LDA and VADER, it's important to recall what these instruments entail.

Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model that is used for uncovering the hidden thematic structure in a collection of documents. LDA assumes that documents are a mixture of a limited number of topics and that each topic is characterized by a distribution of words. By analyzing the word distributions across documents, LDA identifies sets of words that frequently co-occur, thereby defining topics.

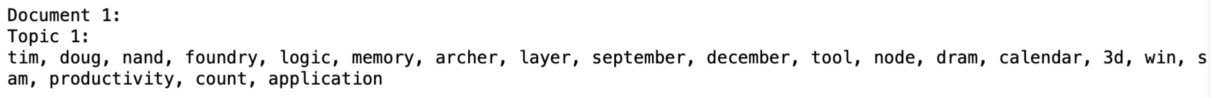
Valence Aware Dictionary and Sentiment Reasoner (VADER)

VADER is a lexicon and rule-based sentiment analysis tool designed to perform well on text from social media, but it is also applicable to other domains. It assigns sentiment scores based on the valence of words, taking into account various linguistic features such as intensifiers or polarity of words. It provides positive, negative, neutral, and compound scores to capture the overall sentiment of the text. VADER's ability to accurately analyze the sentiment of short texts, makes it particularly suitable for our algorithm as explained in the next paragraph.

ALGORITHM

In our approach to topic modeling, we employ a two-step algorithm that leverages LDA and VADER to effectively extract and analyze themes from textual data. This method ensures that we capture the essential topic within each document and assess the sentiment associated with these topics, providing a comprehensive understanding of the content.

- The first step in our algorithm involves using LDA to identify the primary subject within each document. In particular, we extract one topic per document, represented by a list of 20 words (example in Fig. 2.3).



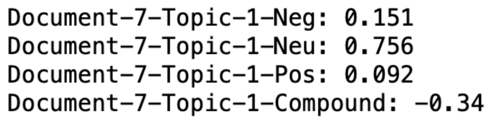
```
Document 1:
Topic 1:
tim, doug, nand, foundry, logic, memory, archer, layer, september, december, tool, node, dram, calendar, 3d, win, s
am, productivity, count, application
```

FIGURE 2.3
LDA output for "Document 1".

  This distilled representation succinctly captures the core theme of the document, facilitating further analysis.
  An inquiry may arise as to why we didn't utilize a greater number of topics for topic modeling. The constraint lay in computational resources; executing the code proved excessively time-consuming. Consequently, we opted for the most salient topic, striking a balance between computational efficiency and comprehending the overarching context of the transcript calls.

- Once the topics are extracted, we proceed to the second step, which involves sentiment analysis using VADER. We specifically apply VADER to the words extracted from the topic generated by LDA. So, in this method, the label of a document is the compound score given to its associated topic. By analyzing the sentiment of these key words, we gain insights into the emotional tone associated with the primary theme of each document.
  As we said before, one of the significant advantages of VADER is its ability to work effectively with a small number of words, making it a good choice for our use case where each topic consists of only 20 words. Initially, one might consider employing the Loughran-Mcdonald dictionary and weighted

score from the lexicon-based analysis, as used in a previous section. However, when applied to such a small number of words, we encountered a practical challenge. The resulting sentiment labels were nearly identical and vanishing. This limitation arises due to the algorithm's structure based on Loughran-Mcdonald. To be completely honest, as can be seen in Fig. 2.5, even with this method, we have a high frequency of null labels; however, we still have greater heterogeneity compared to the LBA algorithm, and this is the focal point for setting up a portfolio strategy, as we will see in the following section.



```
Document-7-Topic-1-Neg: 0.151
Document-7-Topic-1-Neu: 0.756
Document-7-Topic-1-Pos: 0.092
Document-7-Topic-1-Compound: -0.34
```

FIGURE 2.4
VADER output on the topic's words of "Document 7".

2.3.1 STRENGTHS AND WEAKNESSES

Captures underlying topics VS Computational cost

- The LDA and VADER approach excels in capturing the underlying topics present in the document. By using LDA, the algorithm identifies the primary topic within each document, providing a concise representation of its core theme. This enables a deeper understanding of the content and facilitates the extraction of valuable insights from the data.
- However, one of the significant drawbacks of this approach is its complexity and computational expense. Running the LDA can require substantial computational resources and time. The computational burden increases significantly with larger datasets or more complex models, making it impractical for some applications and limiting scalability.

Heterogeneity of results VS VADER is not optimized for financial texts

- The LDA and VADER approach provides heterogeneous results in sentiment analysis. Having a diverse range of sentiment scores is advantageous for labeling, as it allows for a more nuanced understanding of the sentiment expressed in the document.
- However, a limitation of using VADER is that it is not specifically optimized for financial texts. While it performs well on social media content, it may not capture the specific nuances and terminology present in financial documents. This lack of optimization can lead to potential inaccuracies or biases in sentiment analysis when applied to financial texts.

Easy comparison of labels VS Hard interpretation of the topics

- As in lexicon based analysis, this approach provides a straightforward method for comparing sentiment labels. The compound score from VADER ranges from -1 to 1, allowing for easy comparison and interpretation of sentiment across documents. This simplicity facilitates the labeling process, enabling quick and efficient analysis of sentiment.
- However, a challenge of this approach is the interpretation of the results for the topics extracted by LDA. While the compound score provides a clear sentiment label, understanding the underlying themes represented by the topics may be more complex. The topics extracted by LDA may require further analysis and interpretation to discern their significance and relevance to the document content.

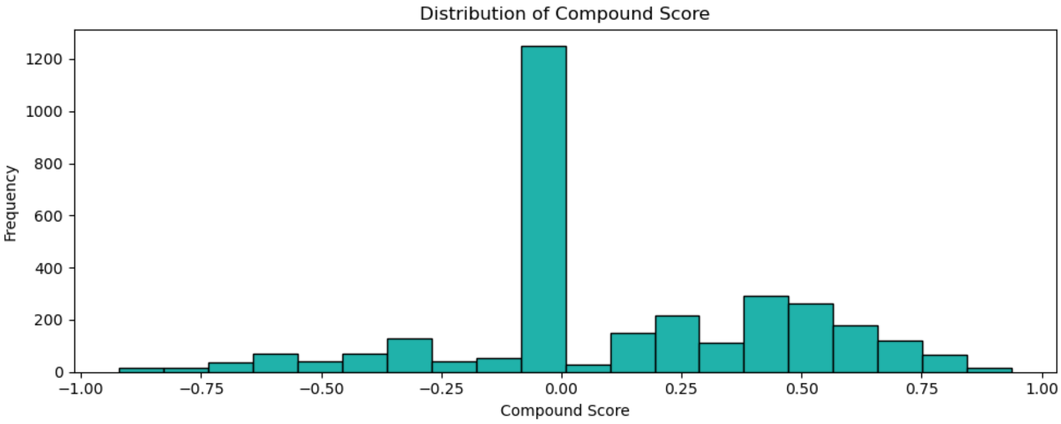


FIGURE 2.5
Distribution of "Compound Score" label.

CHAPTER 3

STRATEGY

In this chapter, we present our trading strategy, utilising sentiment labels generated from lexicon-based analysis and topic modeling with weighted and compound scores respectively. Due to the characteristics of earnings calls transcripts (having similar vocabulary, nuanced language etc.), expected limitations in the use of K-means clustering meant that we ultimately only utilised sentiment labels from the two earlier-described methods.

To derive the final dataset used to test strategies, we merge into a single, consolidated dataframe, five earlier dataframes split by year, which contain the price data of stocks in the S&P500 from 2019 to 2023, and the sentiment of their respective earnings calls transcripts. We then explore making buying decisions using sentiment data and selling decisions carried out with different rules, and evaluate their overall performance over the time spanned by our dataset with the price labels in our dataset. Finally, we compare our strategy with the "Boglehead" strategy (named after John Bogle, Vanguard's Founder). In this study, we specifically define this strategy as simply just buying and holding SPY, an ETF reflective of the performance of the S&P 500. Ultimately, even if our strategies seem profitable in the long-run, their risk-adjusted return should be evaluated to assess whether they outperform the market.

Trading Framework
Let us start with stating the hypotheses we consider and the initial setup of each of our strategies:

- we set an initial balance at 25,000 USD - could be an arbitrary number, but we set it such that a casual investor would begin with this amount since it is a defined minimum amount to hold for a margin account under pattern day trading rules;
- utilise only a long-only portfolio, trading equities. (i.e. no derivatives or other financial instruments);
- invest at most 10% of portfolio value in a long position by default;
- use yfinance to fetch prices if data is not available - we use our existing price data (e.g. the price_day_after column) as far as possible to reduce computational resources needed to get additional price data.
- we consider only self-financing portfolios;
- assume that there is no trading cost.

Test Various Strategies
Run I. Varying Buying Conditions

- **Strategy 1**: Buy if weighted score is above 0.3, an arbitrary threshold to retrieve stocks with positive sentiment implied by weighted_score, **with sentiment scores about 2 s.d. above mean**.

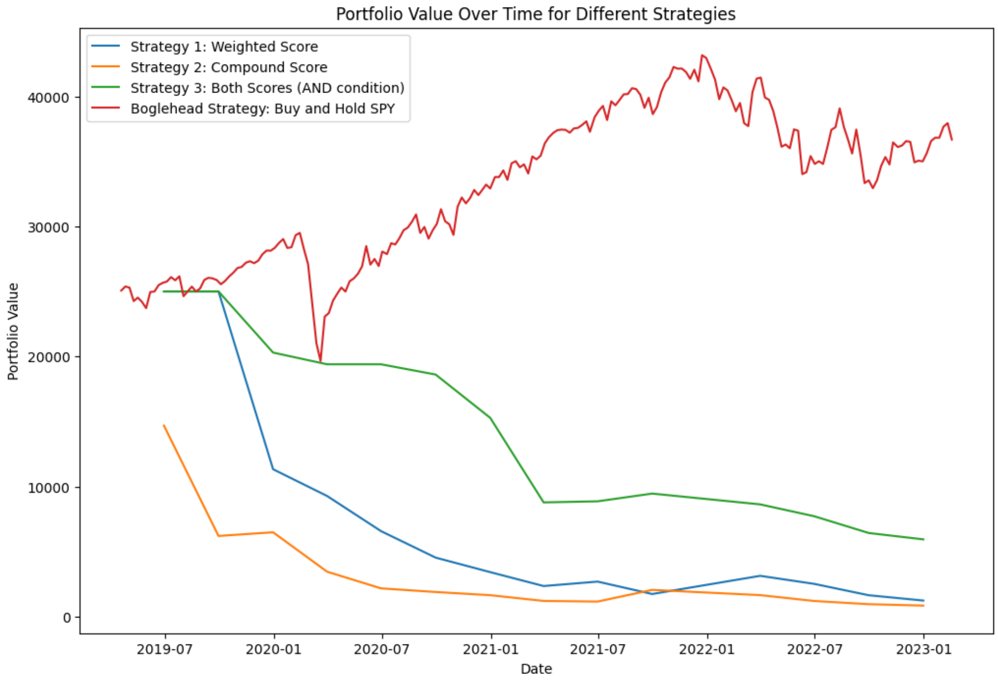
- **Strategy 2**: Buy if compound score is above 0.5, an arbitrary threshold to retrieve stocks with positive sentiment implied from compound score, **with sentiment scores about 2 s.d. above mean**.
- **Strategy 3**: Buy if weighted score is above 0.3 **and** compound score is above 0.5, to test the performance of a strategy with relatively stricter buying conditions.
- Benchmark **Boglehead Strategy**: Buy and hold SPY (S&P500) - an indicator of market performance.

Rationale & Additional Trading Conditions

- **Dealing with Sentiment Turnaround (becoming more negative)**:
  - Sell the stock at close the day after a future quarter's earnings, if the *weighted score* is below -0.2 or *compound score* is below 0.
  - Otherwise, just continue holding the stock.
- **Buying Stocks with more Positive Sentiment**:

  If a quarter has passed since owning the stock **and** another stock has an earnings report with more than +0.2 difference in *weighted score* or *compound score*:
  - Sell the earliest-held position's stock at close the day after the other stock's earnings call.
  - Buy the other stock at the same time.

Result:



The results show that all strategies underperformed compared to the benchmark Boglehead strategy, even if Strategy 3 demonstrated stronger performance than Strategies 1 and 2. Indeed, the incorporation in Strategy 3 of a dual criterion for making buy decisions may have filtered out potential false positive signals that could occur with either criterion alone.

Moreover, the additional requirement in Strategy 3 for both the weighted and compound scores to meet specific thresholds could act as a risk management measure. By demanding a higher level of confidence in the positive sentiment signals, Strategy 3 may have avoided investments in stocks with weaker sentiment indications, which could have mitigated downside risk compared to Strategies 1 and 2.

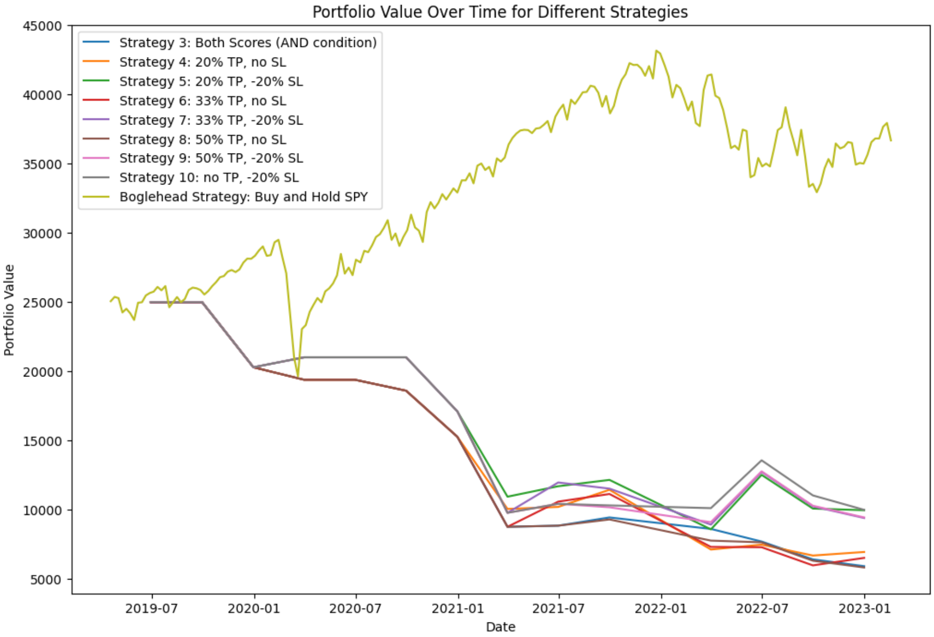
In conclusion, our strategies 1 and 2 are bankrupt. Strategy 3 is based on the constraints in 1 and 2, so it buys much less stock but is bankrupt itself in time.

Run II: Inclusion of Take-Profit and Stop-Loss Orders
In this run, we use the and-condition offered by strategy 3 and add additional take-profit and stop-loss mechanisms with **Strategies 4 to 10**. Take-profit is a predefined price level at which a trader closes a position to lock in profits; on the other hand stop-loss is a predetermined price threshold set by a trader to limit losses by automatically closing a position if the market moves against them. For example, a 20% take-profit means that a trader has set a target to close their position and secure profits once the price of the asset they are trading increases by 20% from their entry point. On the contrary, a 20% stop-loss closes the position if the price decreases by -20% from the entry point.
In this way, we want to avert inaction that would lead to excessive losses arising from market meltdowns (e.g. COVID pandemic) or potential heavy stock price declines in between earnings calls. By providing a mechanism to lock in profits where there is already a significant profit in a given position, we make our strategy more reactive and robust to price fluctuations outside of earnings calls. We compare the new strategies to Strategy 3 and Boglehead.

Strategy	4	5	6	7	8	9	10
TP	20%	20%	33%	33%	50%	50%	None
SL	None	-20%	None	-20%	None	-20%	-20%

TABLE 3.1
Summary of Trading Strategies 4-10

Results:



The results show that Strategies 5, 9 and 10 with stop-loss mechanisms outperformed those without stop-loss mechanisms like Strategies 3, 4 and 6, although the strategies overall still do not outperform the market. No discernible difference is observed between the strategies testing the presence and value of take-profit mechanisms.

For now, we can see the importance of minimising downside in a trading strategy like this that utilises sentiment on a quarterly basis. Ordinarily, we would be unable to react until the following quarter or even beyond, but the price action itself (e.g. large declines in a short time) informs our strategy to make crucial decisions intra-quarter.

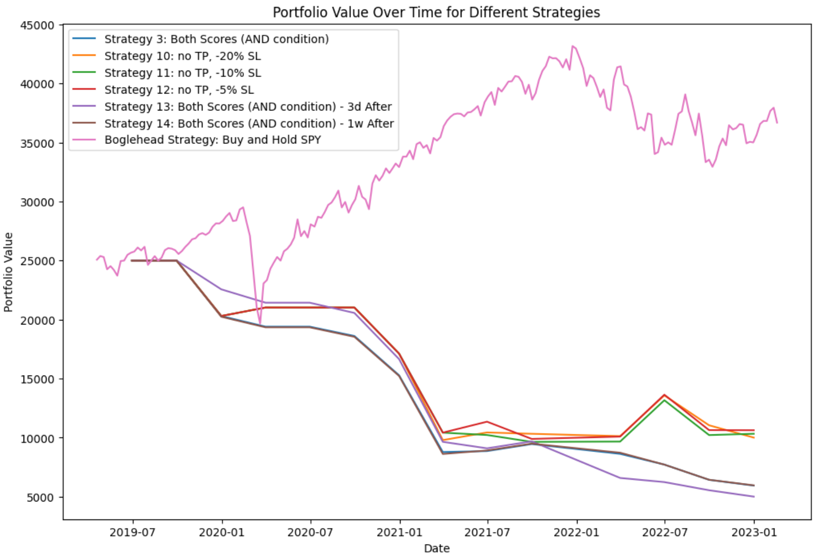
We could also conceptualise more dynamic trading rules on sentiment directly if sentiment data is received on a more regular basis (e.g. weekly/daily sentiment like a fear-greed index), such as from news articles, analysts and from social media like Reddit, but due to the lack of frequent data, there is limited effectiveness in making dynamic trading rules that act on our sentiment labels.
To continue, we explore whether we could optimise a stop-loss value and find other ways to improve the general performance of our trading strategy.

Run III: Varying Stop-Loss Values & Delayed Buying
In this run, we wanted to test how varying the stop-loss threshold percentage-wise, and whether delayed buying of stocks (e.g. buying 3 days later at close instead of day after at close) with positive sentiment could influence the performance of our strategy. The rationale for the former was to explore a balance between limiting downside and having a profitable strategy not carried out due to the stop-loss mechanism, while the rationale for the latter was to explore letting possible short-term trends (possibly negative, such as "selling the news") play out before our strategy capitalised on a long-term uptrend hypothesised by positive sentiment.

No take-profit mechanisms were in place due to inconclusiveness on whether it would positively impact the performance of our strategy.

- **Strategy 11**: No TP, -10% SL
- **Strategy 12**: No TP, -5% SL
- **Strategy 13**: Strategy 3, but buy 3 days after at close
- **Strategy 14**: Strategy 3, but buy 1 week after at close

Results:



The results show no significant differences between the tighter stop-loss values and the earlier Strategy 10 with -20%, and underperformance for the strategies with delayed buying. This is likely because the above difference did not impact the final trading decisions made by the strategy.

We also interpret that a standard stop-loss threshold was found to be sufficient for our strategy, while delayed buying not being viable for our strategy.

Still, we have not found measures or conditions that would ensure that when we trade on sentiment, we are able to outperform the market. We wish to examine the possible reasons.

Examining Trading Decisions Made & Exploring Possible Improvements
There were limitations in our strategies' ability to make trading decisions:

- First, there was limited ability in the strategy to react to a stock's earnings calls. Due to limited price data, our trading strategy could only act earliest at market close the day after earnings are released, where the market would have already digested and reacted to information and sentiment from the earnings call. There are earlier points in time where making trading decisions then would be critical to the overall performance of a trade - such as buying and selling once pre-market trading opens, or at market open the day after an earnings release.

  In our trading strategies, we bought stocks earliest at 1 day after earnings at close, when the expected price action would have largely played out. An improvement to our trading strategies could just be simply buying at the earliest possible time where we have the sentiment value generated, i.e. during pre-market trading, and examine price action over the short-term (even during intra-day) to discern timeframes where price action is correlated with sentiment.

- There was limited intra-day price data for the stocks, and we had limited computing power that could test out both intra-day and inter-day trading spans. It could be possible that sentiment of a given value correlates only with positive or negative price action that lasts only within the week after the earnings call for example, or that positive pre-earnings price action followed by positive earnings sentiment correlates with a post-earnings price decline. If this is such a case, we could have added this into our trading strategy to improve performance.

  Alternatively, we could have also constructed a strategy that remained largely cash-based and swing-traded stocks on a short-term basis using our sentiment labels. They could prove to be more profitable than our trading strategies, especially if we are unable to optimise the trading and portfolio optimisation themselves.

- We did not consider shorting stocks with negative sentiment, due to the inherent risk of unlimited downside and the lack of data on costs associated with shorting (e.g. borrowing costs). But there could be performance driven by shorting given a sufficient amount in capital, and careful modeling of the trading strategy.

- We did not have data that could be used to assign a sentiment label to a whole sector of stocks, or the entire market itself. For example, (assuming that we are trading stocks beyond those in the S&P500) it could be useful to retrieve and utilise the sentiment scores of EV car makers (e.g. TSLA, XPEV, NIO etc.), "meme" stocks, semiconductors, regional banks etc. for regular time intervals (more frequent than quarterly earnings) to conduct trading decisions. An individual stock's performance could have strong correlation with the performance of their respective sector and their sentiment at a given time that is more current than the past quarter, especially if news outside of earnings calls (e.g. tariff announcements etc.) changes existing sentiment reflected in a stock's earnings calls.

Meanwhile, market sentiment could be tested for use to predict bear markets like in 2020 with the Assassination of Qasem Soleimani and then the COVID pandemic, or in 2022 with inflation. These could have been factored into our trading model to limit downside if a market-wide downturn is forecasted.

- Due to limited data, we were unable to consider the use of financial derivatives for trading or for hedging. Buying options on a stock or a leveraged ETF (e.g. NVDL) instead of the stock itself could be much more profitable than buying the stock itself especially when trading costs are present, but we acknowledge that modeling this successfully to come up with an effective trading strategy would be a much more complex task.

- We did not use discrete sentiment labels for the stocks, due to their lower perceived utility compared to a sentiment score, but it is still possible to make trading strategies with them (in which case we focus more on the trading time horizon rather than the degree of positivity/negativity in sentiment) that could be profitable.

Further Factors for Consideration

- Systematic exploration of discrete levels (e.g. weighted score 0.10 - 0.19) of sentiment scores with different timeframes to find optimal periods to take long or short trading positions.

- Sentiment scoring of a ticker's sector to factor into trading decisions.

- Using the Black-Litterman Model or alternatives as a tool to optimize portfolio allocation, which we have not learnt about until the in-class presentations.

- Exploring the use of modern deep learning methods like FinBERT to impute sentiment labels, and compare it to our existing methods. We have experimented casually with transformers to impute discrete sentiment labels to our data, but this was not included in our final work due to the utility of the continuous sentiment labels already offered by lexicon-based analysis and topic modeling.

Chapter 4

Conclusion

We know that financial markets are inherently complex and unpredictable. There is immense difficulty in crafting trading strategies, including sentiment-based trading strategies, even just to outperform the market. We have largely been successful in adding sentiment labels to our data of stocks, as a demonstration of our understanding of the machine learning techniques used in finance. But the complexity of making trading decisions amidst limited data availability, macroeconomic sentiment changes and the challenge to portfolio optimisation has meant that we were ultimately unable to produce a trading strategy that can outperform the market.

At present, our strategy hinges heavily on the temporal flow of tickers with positive sentiment (i.e. we buy tickers only as and when they release earnings calls which happen to be positive, and not on any other factors). Therefore, our trading strategy of buying positive sentiment and selling negative sentiment is likely a blunt instrument, since our strategy is not yet able to distinguish between stocks of different sectors, different PE values, volumes or price action in order to make trades or to optimise a portfolio.

We would ideally want to buy undervalued stocks with positive sentiment, and possess stocks in our portfolio that are diversified across sectors in order to improve performance. We would ideally trade stocks with strong correlations between their sentiment and definitive price action within a fixed timeframe. But without also modeling concurrent factors that influence the price of a stock, sentiment-based trading strategies would likely be ineffective and bankrupt, particularly when compared to buying and holding ETFs representing a broad market index like the S&P 500. We believe that sentiment analysis is ultimately a useful tool in trading to generate returns, but its use should accompany existing valuation and financial modeling methods for a trade to be successful.

MATHEMATICAL APPENDIX

Let X be the random variable that assigns to each document D its weighted score. Resuming the notation from Section 2.2, let us remember that:

$$X(D) = \frac{w_1 v_1(D) + ... + w_n v_n(D)}{|D|},$$

where each v_i is a random variable for all $i \in \{1, ...n\}$ such that:

$$v_i(D) = \sum_{\substack{\omega_j \text{ word} \\ \omega_j \in \text{ D}}} \mathbb{1}^j_{\{\text{L.M. = i}\}}(\omega_j)$$

and $\mathbb{1}^j_{\{\text{L.M. = i}\}}$ representing the indicator function that has as domain the word of D in position j. In particular, these indicators are indipendent (the sentiment of the word in position j_1 does not influence the sentiment of the word at position j_2 for $j_1 \neq j_2$) and equally distributed given that are Bernoulli with $\mathbb{E}[\mathbb{1}^j_{\{\text{L.M. = i}\}}] = p_i$ and $\sigma^2(\mathbb{1}^j_{\{\text{L.M. = i}\}}) = p_i(1 - p_i)$ (p_i is the probability that a random word has sentiment i). So, from the central limit theorem, it follows that:

$$\frac{v_i - |D|p_i}{\sqrt{|D|p_i(1 - p_i)}} \xrightarrow{|D| \to +\infty} \mathcal{N}(0, 1) \quad \Longrightarrow \quad v_i \xrightarrow{|D| \to +\infty} \mathcal{N}(|D|p_i, |D|p_i(1 - p_i))$$

Now, the v_i's are not indipendent. In fact, knowing that $v_i = n$ tells us that there are n words in the document with sentiment "i", so when we calculate v_j with $j \neq i$, we'll know that there are at most $|D| - n$ words with sentiment "j". However, when $|D|$ becomes large enough, the additional information is not so relevant (it's true that n also grows, but if $|D|$ and n were to increase, for example, both proportionally, $|D| - n$ would still increase). For this reason, except for limited error, we can assume that v_i are independent. Hence, any linear combination of the v_i's is still Gaussian. So, we have:

$$X(D) \xrightarrow{|D| \to +\infty} \mathcal{N}\Big(|D| \sum_1^n p_i, \frac{\sum_1^n \omega_i^2 p_i(1 - p_i)}{|D|}\Big)$$

In particular, when we consider a large number of documents as in our case, the histogram of the frequencies of the weighted scores is expected to be bell-shaped, as shown in Fig.2.2.

BIBLIOGRAPHY

[1] DataScientest. *K-Means Clustering in Machine Learning: A Deep Dive*. Accessed: 2024-04-05. 2023. URL: https://datascientest.com/en/k-means-clustering-in-machine-learning-a-deep-dive.

[2] Shasha Deng, Fei Wang and Lei Liu. 'Knowledge-based temporal reasoning for financial news events'. In: *Proceedings of the 2014 SIAM International Conference on Data Mining*. SIAM. 2014.

[3] E. Henry. 'Are investors influenced by how earnings press releases are written?' In: *The Journal of Business Communication (1973)* 45.4 (2008), pp. 363–407. DOI: 10.1177/0021943608319388.

[4] Cheng-Hui Huang, Jian Yin and Fang Hou. 'A Text Similarity Measurement Combining Word Semantic Information with TF-IDF Method'. In: *Journal of Information Science and Technology* No.5 (2011), pp. 856–864.

[5] Tim Loughran and Bill McDonald. 'When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks'. In: *The Journal of Finance* 66.1 (2011), pp. 35–65. DOI: 10.1111/j.1540-6261.2010.01625.x.

[6] Paul C. Tetlock. 'Giving Content to Investor Sentiment: The Role of Media in the Stock Market'. In: *The Journal of Finance* 62.3 (2007), pp. 1139–1168. DOI: 10.1111/j.1540-6261.2007.01232.x.