HAAS SCHOOL OF BUSINESS

Leveraging EDGAR Filings in Portfolio Construction

GROUP 2

Ayushi Garg
Anjali Jha
Abhilasha Dey
Anuj Choudhury
Santhosh Suresh Kumar

March 1, 2024

Supervisor: Dr. Ali Kakhbod

Abstract

In this study, we explore the application of Natural Language Processing (NLP) techniques to analyze EDGAR filings, with the aim of enhancing portfolio construction strategies. Drawing inspiration from the insights of the "Lazy Prices" [4] study, we extend our investigation to cover the years 2015 to 2018, enabling us to develop and test various long-short investment strategies. We examine the impact of different rebalancing frequencies and holding durations to fine-tune our approach for optimal performance. The seminal study identified a significant annual alpha of 22% between 1995 and 2014 by focusing on the distinction between companies that changed their reporting language ("changers") and those that did not ("non-changers"), with a strategy involving monthly rebalancing and a three-month lookback period. Extending this analysis to include data through 2018, we observe a 15% alpha and a Sharpe ratio of 3.81 for 6 month lookback period . Additionally, we delve into the enhancement of our model through the integration of advanced transformer-based NLP architectures, such as BERT, which further amplifies the efficacy of our investment strategy. Moreover, we assess the practicality and tradability of our approach by specifically examining its performance within the S&P 500 index.Implementation details are available in the github repository

Contents

1 Introduction

This paper delves into the strategic utilization of corporate disclosures in financial markets, leveraging advanced Natural Language Processing (NLP) techniques, such as sentiment analysis, similarity assessment, and sophisticated models like BERT. These tools can empower investors to craft effective long-short strategies. Our research focuses on the period from 1995 to 2018, analyzing an extensive dataset of 10-K and 10-Q filings from U.S. corporations, sourced from the EDGAR database.

The foundational study differentiated between "changers" and "non-changers," uncovering an impressive 22% annual alpha from 1995 to 2014. Our approach segments the dataset into a training period (1995-2014) and a testing period (2015-2020) to pinpoint the optimal lookback period and rebalancing frequency. We evaluate our portfolio's performance not just on returns but also on critical metrics for investors, such as Sharpe ratio and drawdown, achieving a notable Sharpe ratio of 3.38 during the testing period, indicating robust out-of-sample performance.

By incorporating context-aware models like BERT, we enhance our ability to distinguish between "changers" and "non-changers" with greater precision, offering deeper insights into the nuances of corporate disclosures and their market implications. This refined approach yields promising returns and outperforms traditional Bag of Words models over a focused two-year study period (2012-2013), suggesting a fruitful avenue for further research.

We also examine the strategy's robustness by focusing on the S&P 500 universe, ensuring that the alpha generated is not solely driven by illiquid or small-cap stocks, which could render the strategy non-viable. This restriction still results in a satisfactory Sharpe ratio of 1.02 across the entire period, with the strategy demonstrating resilience even during recessionary phases.

Further refinement involves analyzing specific sections of the filings, with the Management's Discussion and Analysis (MD&A) section proving the most fruitful, delivering the highest Sharpe ratio of 1.53, indicative of its rich informational content.

Additionally, we explore the potential of sentiment analysis applied to financial reports, utilizing a comprehensive Loughran-McDonald (LM) sentiment dictionary to calculate sentiment scores. This analysis aims to deepen our understanding of the market implications of the tones set in corporate communications. However, portfolios based on sentiment analysis show less robust out-of-sample performance.

Overall, this paper offers a comprehensive exploration of the use of corporate disclosures in financial market strategies, extending traditional methods to provide more nuanced insights and improved strategies. By integrating cutting-edge NLP techniques and innovative portfolio categorization methods, we furnish investors with refined tools for making informed decisions in the rapidly evolving financial landscape.

2 Literature Review

The literature review contextualizes the current study within the broader research landscape, encompassing three main areas: underreaction in stock prices, textual analysis in finance and accounting, and the information content of firms' disclosure choices.

2.1 Underreaction in Stock Prices and Investor Inattention

In his review, Tetlock [17] delves into the subtleties of information transmission in financial markets, emphasizing how investor reactions to news—ranging from underreactions to novel information to overreactions to outdated news—pose significant implications for asset pricing and corporate finance. His analysis spans a variety of informative events and transmission mechanisms, highlighting the critical role of both numeric and nonnumeric information in shaping market outcomes and investor behavior.

Da, Engelberg, and Gao [5] and Ben-Raphael, Da, and Israelson [2] employ measures of investor attention from Google and Bloomberg search activities, respectively. Engelberg, Sasseville, and Williams[8] show links between TV ratings spikes during the "Mad Money" show and overreaction in stock prices. The current paper adds to this literature by documenting a distinct form of investor inattention affecting a broad cross-section of firms, particularly focusing on the crucial corporate disclosure of annual reports.

2.2 Textual Analysis in Finance and Accounting

The field of textual analysis within finance and accounting has witnessed a rapid expansion, leveraging sophisticated computational techniques to extract meaningful insights from the extensive textual data produced by firms. This innovative approach not only deepens our understanding of market dynamics but also provides predictive insights into financial outcomes based on the qualitative aspects of corporate disclosures.

In their research, Loughran and McDonald [14] focus on the development and application of textual analysis methodologies within the field of finance and accounting. They critically evaluate the use of traditional and novel textual analysis tools in interpreting financial documents, aiming to enhance the understanding of how narrative disclosures and managerial sentiment in earnings calls and SEC filings can influence market outcomes and investment decisions.

Furthering this research, Li [12] investigates the tendency of managers to obscure financial information in public disclosures, particularly when their firm's performance is poor, supporting the management obfuscation hypothesis. This hypothesis suggests that managers disclose information more transparently when their firms are doing well, a pattern previously identified in research. Li aims to empirically validate this behavior, examining the strategic communication choices made by managers in relation to their firm's performance.

Li [13] explores how the Forward-Looking Statements (FLS) within the Management Discussion and Analysis (MD&A) sections of 10-Q and 10-K filings predict future performance. Utilizing a Naïve Bayesian machine learning algorithm, Li discovers that the tone of FLS correlates with several factors like current performance and firm characteristics, and significantly, it is positively linked with future earnings.

Nelson and Pritchard [15] delved into the use of cautionary language, establishing its connection to a firm's litigation risk, while Feldman et al. (2008)[9] explored how the sentiment in the Management's Discussion and Analysis (MD&A) section correlates with both contemporaneous and future stock G.

The field has also benefitted from the application of machine learning and natural language processing (NLP) techniques, which have refined the capability to quantify and analyze sentiment, tone, and complexity in financial texts. In their research, Fisher, Garnsey, and Hughes[10] conduct a comprehensive synthesis of the existing literature on Natural Language Processing (NLP) within the fields of accounting, auditing, and finance. They explore how NLP technologies are applied to analyze textual documents in these domains to extract insights, make predictions, and develop new methods and tools to enhance understanding and knowledge. Their work identifies the current state of NLP application in these fields and outlines potential directions for future research, emphasizing the bridge between human-computer communication and the precision required for financial analysis and reporting.

Large language models have revolutionized the field of Natural Language Processing (NLP) by delivering unparalleled accuracy and understanding in tasks ranging from text classification to sentiment analysis. Building upon this foundation, Yang, UY, and Huang [18] have innovatively extended the capabilities of these models into the financial realm. They developed FinBERT, a specialized version of the BERT model, by pretraining it with a vast collection of financial communications. This approach not only fills a crucial gap in financial NLP applications but also sets a benchmark for domain-specific model performance, as evidenced by their model's success in financial sentiment classification tasks.

Diaz, Njoroge, and Shane (2020) [6] explored the repercussions of 10K report complexity on capital market price discovery. They reveal that complexity hinders analysts' ability to mitigate information asymmetry and negatively affects their response to 10K information, thereby exacerbating market inefficiency. Their work underscores the critical influence of financial document complexity on the accuracy of analyst forecasts and the broader implications for market dynamics.

Bannouh, Genga, and Peeters [1] examine the influence of Reporting Lag—the duration required for a company to submit its annual or quarterly reports—on subsequent stock performance. Their study demonstrates that companies that report promptly enjoy a notable premium over those with delayed filings. Delving into the causes behind Reporting Lag, they identify firm and document characteristics as significant factors. Shorter lags correlate with positive earnings surprises, improved operational efficiency, greater consistency between reports, and more favorable sentiment relative to preceding reports. Additionally, their findings suggest that extended Reporting Lags may indicate management's postponement in releasing adverse news, highlighting the lag as a potential indicator of underlying firm dynamics.

The integration of textual analysis into financial modeling and forecasting represents a significant paradigm shift towards a more nuanced understanding of market dynamics. It acknowledges that numerical data in financial statements tell only part of the story, with the remainder hidden within the accompanying narratives. As computational power and analytical techniques continue to evolve, the potential of textual analysis in finance and accounting is vast, offering profound implications for investors, regulators, and academics alike. The current study contributes to this growing body of research by utilizing a document similarity measure to predict future returns, even after controlling for disclosure sentiment, thus underscoring the predictive power of qualitative aspects of corporate disclosures.

Padyšák's[16] study delves into the profitability of conducting textual analyses on 10-K and 10-Q filings, focusing on the impact of language similarity. His findings indicate that the congruence of positive language across filings emerges as the most lucrative strategy. An intriguing discovery from his research is that stocks with lower levels of positive language similarity significantly surpass those with higher levels in terms of performance. This phenomenon, known as the positive similarity effect, cannot be accounted for by traditional asset pricing models or shifts in report sentiment, marking it as a unique anomaly within financial markets. Padyšák highlights that employing this strategy in a long-only format proves highly profitable, while its application in a long-short framework delivers remarkable consistency and risk-adjusted returns, with a notable figure of 0.84.

2.3 Information Content of Firms' Disclosure Choices

The information content of firms' disclosure choices is explored by various studies. Lee[11] finds that less earnings-related information is incorporated into stock prices for firms with longer or less readable 10-Qs. Dyer, Lang, and Stice-Lawrence[7] observe the lengthening and increased complexity of 10-Ks. Brown and Tucker[3] focus on changes in the MD&A section, noting a decline in its usefulness. In contrast, the current paper reveals that changes in 10-Ks are remarkably useful, predicting significant negative returns in the future, challenging the notion that these documents have become less informative.

3 Data and Methodology

In this section, we outline the data sources and the methodology employed in constructing the sample and conducting the analysis.

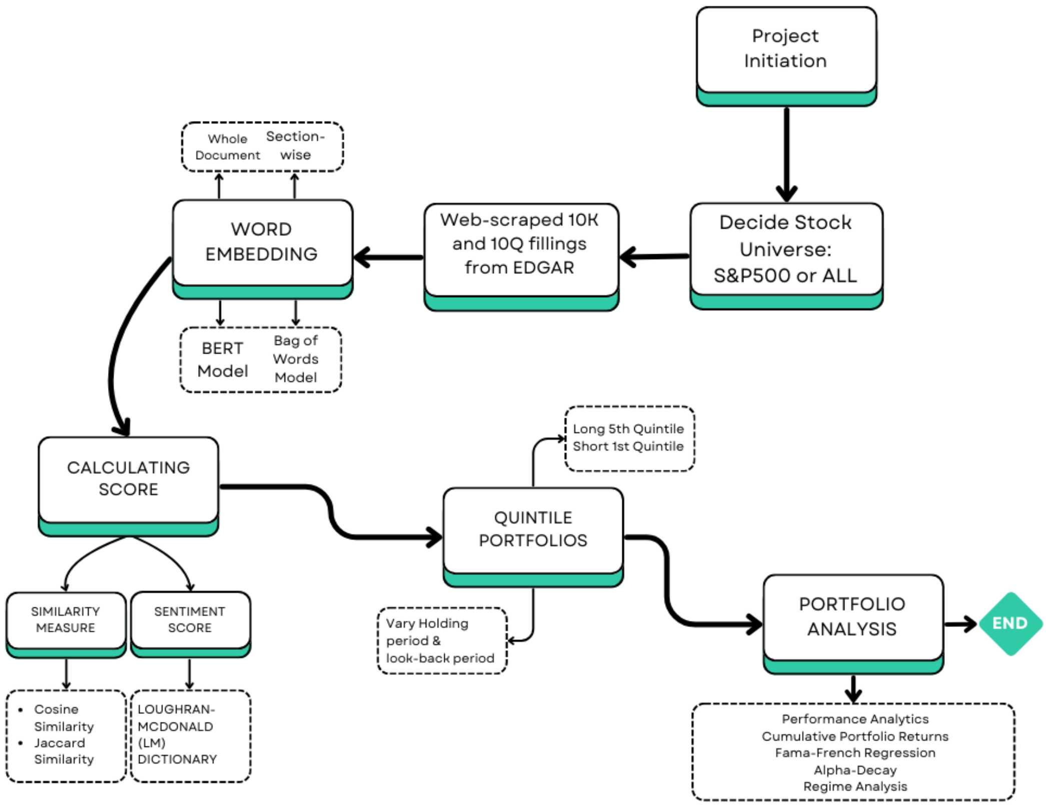


Figure 3.1: Workflow

3.1 Data Sources

Our dataset is constructed from various sources, primarily drawing from the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) website. We download complete 10-K, 10-K405, 10-KSB, and 10-Q filings spanning the years 1995 to 2018. These filings are obtained in HTML text format and encompass all information submitted with each firm's file, including exhibits, graphics, XBRL files, PDF files, and Excel files.

Similar to Lazy Prices (2011), our focus centers on the textual content of the documents. We extract the main text from each 10-K and 10-Q, eliminating tables (if their numeric character content exceeds 15%), HTML tags, XBRL tables, exhibits, ASCII-encoded PDFs, graphics, XLS, and other binary files.

We use monthly stock returns from the Center for Research in Security Prices (CRSP) and firms' book value of equity and earnings per share from Compustat. We also obtain sentiment category identifiers from Loughran and McDonald (2011)'s Master Dictionary

3.2 Similarity Measures

We assess the quarter-on-quarter resemblances between 10-Q and 10-K filings employing two distinguished similarity metrics derived from the domains of linguistics, textual similarity, and natural-language processing (NLP): (i) Cosine similarity and (ii) Jaccard similarity.

Cosine similarity

Cosine similarity measures the cosine of the angle between two non-zero vectors of an inner product space. This measure is widely used to compare the similarity between two documents in text analysis and information retrieval. In the context of document similarity, it quantifies the cosine of the angle between two document vectors, which represent the documents in terms of their term frequencies.

Given two documents, D_1 and D_2, with their term frequency vectors represented as V_1 and V_2 respectively, the cosine similarity (S_{cos}) between these documents is calculated using the dot product of V_1 and V_2 divided by the product of their magnitudes. Mathematically, it can be expressed as:

$$S_{\text{cos}}(D_1, D_2) = \frac{V_1 \cdot V_2}{\|V_1\| \|V_2\|} = \frac{\sum_{i=1}^{n}(tf_{i,D_1} \times tf_{i,D_2})}{\sqrt{\sum_{i=1}^{n}(tf_{i,D_1})^2} \times \sqrt{\sum_{i=1}^{n}(tf_{i,D_2})^2}}$$

Where:

- tf_{i,D_1} and tf_{i,D_2} are the term frequencies of term i in documents D_1 and D_2, respectively.
- n is the number of unique terms in the union of the sets of terms occurring in both documents.

Jaccard Similarity

Jaccard similarity, on the other hand, is a measure used for comparing the similarity and diversity of sample sets. It evaluates similarity between finite sets, and is defined as the size of the intersection divided by the size of the union of the sample sets.

For documents D_1 and D_2, let S_1 and S_2 be the sets of unique terms in each document, respectively. The Jaccard similarity (S_{jac}) is calculated as:

$$S_{\text{jac}}(D_1, D_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

This formula represents the ratio of the number of unique terms present in both documents to the total number of unique terms in both documents. Unlike cosine similarity, Jaccard similarity does not take into account the frequency of terms, focusing instead on the presence or absence of terms in the document sets.

These similarity measures can be effectively used to compare the textual content of 10-Q and 10-K filings, providing insights into the changes in corporate disclosures, risk factors, and management discussions across different reporting periods.

3.3 NLP Models

Sentiment and similarity analysis in financial texts, such as 10-K and 10-Q reports, demands methods that accurately capture nuanced expressions of outlook and performance, given the specialized language and complex expressions inherent to financial documents. The following are the most common models used for this task:

3.3.1 Bag of Words-based Models

The Bag of Words (BoW) model is a simplistic yet powerful approach to text analysis in natural language processing (NLP). It represents text data by counting the occurrence of words within a document, disregarding the order and context in which the words appear. This model transforms text documents into numerical feature vectors where each unique word in the text corpus corresponds to a feature.

The main advantage of the BoW model is its ease of implementation and interpretation. However, its simplicity also leads to limitations, such as the inability to capture semantic relationships between words and the context in which they appear. Despite these drawbacks, BoW-based models have been extensively used in various applications, including document classification, spam detection, and sentiment analysis. Amongst the Bag of Words based models, we use the LM dictionary model since it allows for a more nuanced understanding of text by incorporating linguistic and semantic richness into the analysis. Unlike simpler Bag of Words approaches that merely count word frequencies, the LM (Linguistic Model) dictionary model leverages predefined linguistic rules and dictionaries that can capture the context, sentiment, and thematic structure of the text.

The Bag of Words(BoW) model offers a straightforward mechanism for sentiment scoring by evaluating the balance between positive and negative terms within a document. Given the formal nature of 10-K and 10-Q reports, this method is adapted as follows:

$$\text{Sentiment Score} = \frac{\text{\# Positive Terms} - \text{\# Negative Terms} - \text{\# Litigious Terms}}{\text{Total \# Terms}} \tag{3.1}$$

Loughran and McDonald (2011)'s Master Dictionary has been used to extract the lists of negative, positive or litigious words.

Loughran-McDonald (LM) Dictionary This is a specialized lexicon developed for the analysis of financial texts, such as corporate filings, earnings reports and analyst reports. The genesis of the LM Dictionary was motivated by the observation that standard sentiment dictionaries frequently misclassify financial terminology, leading to inaccurate sentiment assessments.

The LM Dictionary was constructed through a rigorous process that involved analyzing a vast corpus of financial documents, including SEC filings, earnings reports, and analyst notes.

Key financial terms and phrases were meticulously extracted and classified into sentiment categories such as positive, negative, uncertainty, and litigious, among others. This classification was refined through an iterative process of expert review, ensuring the dictionary reflects the nuanced language of finance. The construction of the LM Dictionary blends automated data analysis with deep financial expertise, resulting in a specialized tool tailored for accurate sentiment analysis in financial contexts.

3.3.2 Transformer-based Models

Transformer-based models represent a significant advancement in NLP, introduced by Vaswani et al in their landmark paper, "Attention is All You Need" These models leverage the transformer architecture, which is built upon the mechanism of self-attention, enabling the model to weigh the significance of different words within a sentence.

Unlike previous models that process text sequentially, transformers process entire texts in parallel, drastically reducing training times and enabling the capture of complex contextual relationships between words. This architecture has paved the way for the development of highly sophisticated NLP models such as BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pretrained Transformer), and their variants, which have set new benchmarks in tasks like language understanding, text generation, translation, and question-answering.

Bidirectional Encoder Representations from Transformers BERT models offer a sophisticated method for generating document embeddings, advancing beyond the capabilities of traditional dictionary-based approaches. Rather than simply identifying the presence or absence of words, BERT's embeddings capture the semantic proximity between terms within a document. This results in a rich vector representation that reflects the document's content with greater depth, considering the complex relationships and contextual meanings of words. By employing these embeddings in combination with established similarity measures, such as cosine similarity, researchers can accurately assess the similarity between two documents. This approach not only acknowledges the nuanced use of language but also significantly improves the accuracy of thematic and sentiment comparisons across texts, offering a refined methodology for document similarity assessment in scientific studies.

One thing to note here, is that to obtain the embeddings, BERT processes text by splitting it into fixed-length sequences, meaning it can only consider a limited context around each word. To ensure maximum context is being captured we obtain section specific embedding, by using max pooling across of batch of sentences corresponding to a section. The resulting vector obtained from max pooling serves as a compact representation of the input sequence, capturing salient features or characteristics that are deemed most important based on the maximum values across the sequence of embeddings.

3.4 Portfolio Construction & Backtesting

Our portfolio construction method involves analyzing monthly company filings, specifically 10-K and 10-Q reports. We employ similarity metrics, such as cosine and Jaccard similarity, to compute similarity scores between current and previous filings, enabling us to categorize companies into quintiles. Companies in the fifth quintile, which show high similarity, are selected for long positions, whereas those in the first quintile, exhibiting low similarity, are targeted for short positions.

Portfolios are maintained for periods of 1, 3, or 6 months and are rebalanced accordingly. To identify the optimal lookback and holding periods, we split the data into a training period (1996-2014) and a testing period (2015-2018), applying the best-performing models from the training period to the testing period for validation. Additionally, we incorporate sentiment analysis by constructing portfolios based on sentiment, favoring companies with positive sentiment scores for long positions and those with negative scores for shorting.

Monthly returns are calculated using CRSP data, with returns capped at the 1st and 99th percentiles to reduce the influence of outliers. The performance of the portfolios is assessed based on annualized returns and the Sharpe ratio.

To ensure the model's robustness across different market regimes, we backtested it during both recessionary and expansion phases. The recession phase is defined as the period from December 2007 to June 2009, while the expansion phase is considered to be the period following June 2009.

4 Exploratory Data Analysis

Consolidating key points from the EDA:

- **Number of stocks filings per Year**: Figure 4.1 The filings have been consistent over last few years
- **Changes over time:** As seen in Figure 4.2, we observe that the number of unique words used increases over time on an aggregate level, with key changes observed for Telecommunication sector, which was identified by Fama-Fench 48 industry classification.
- **Data distribution across industries :**Figure 4.3 shows the distribution of different CIKs in our dataset across different industries. The dataset has more stocks pertaining to manufacturing, financial services, real estate and telecommunication which is inline with the industry distribution in broader market such as S&P500
- **Distribution across industries:** We do no observe any industry (industry based on Fama Fench 48 factor classification) to have under usage of words, but certain industries do seem to be more verbose in their filings, on further investigation looks like these are pertaining to industries such as Pharmaceutical, utilities, tobacco and coal as shown in 4.4
- **Avg Sentiment Score Over Years**:4.5 Consitent with the higher proportion of negative words in the dictionary, average sentiment score seems to be negative
- **Sentiment Proportion in LM dictionary**:4.6 LM dictionary contains higher proportion of negative sentiment words than positive sentiment.

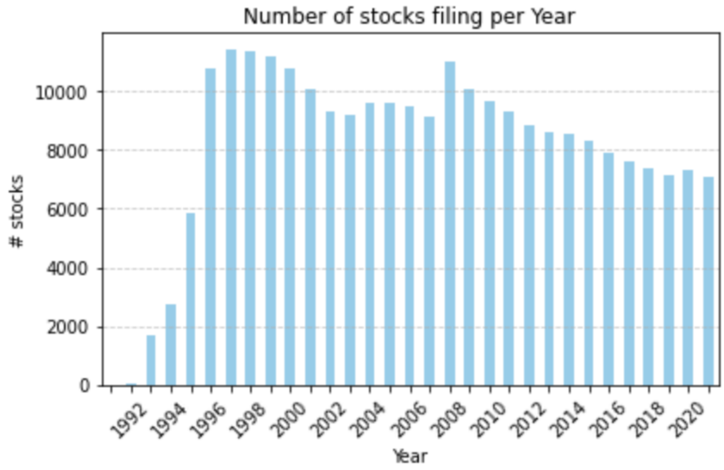
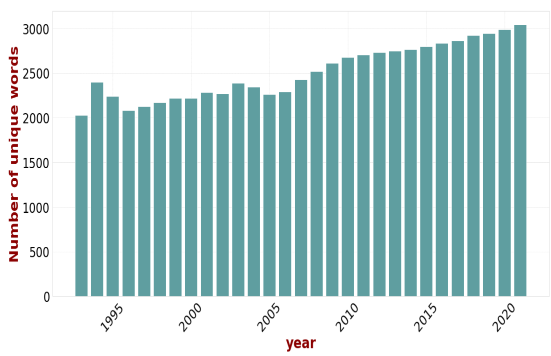
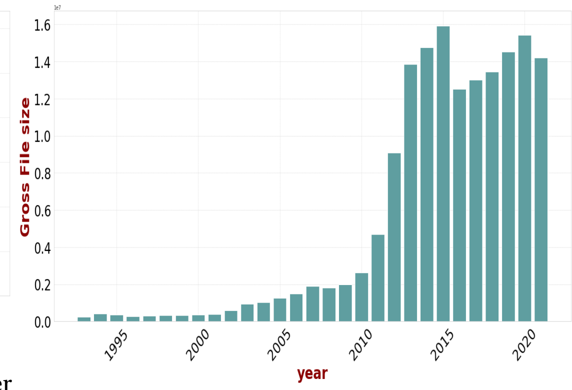


Figure 4.1: Number of stocks filing per Year



(a) Change in number of unique words used over time



(b) Gross File size across years

Figure 4.2: Change over time

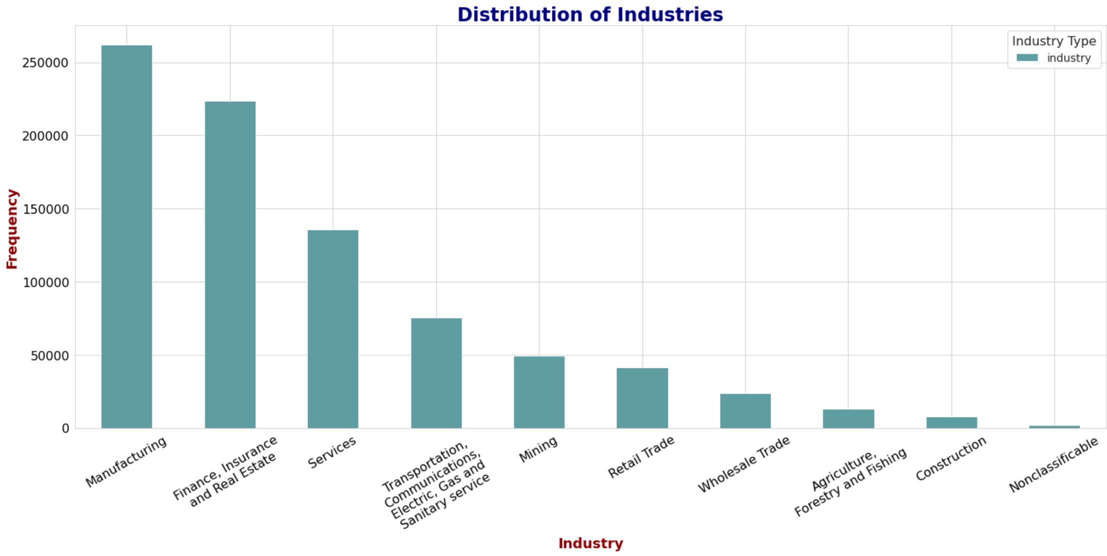


Figure 4.3: Data distribution across industries

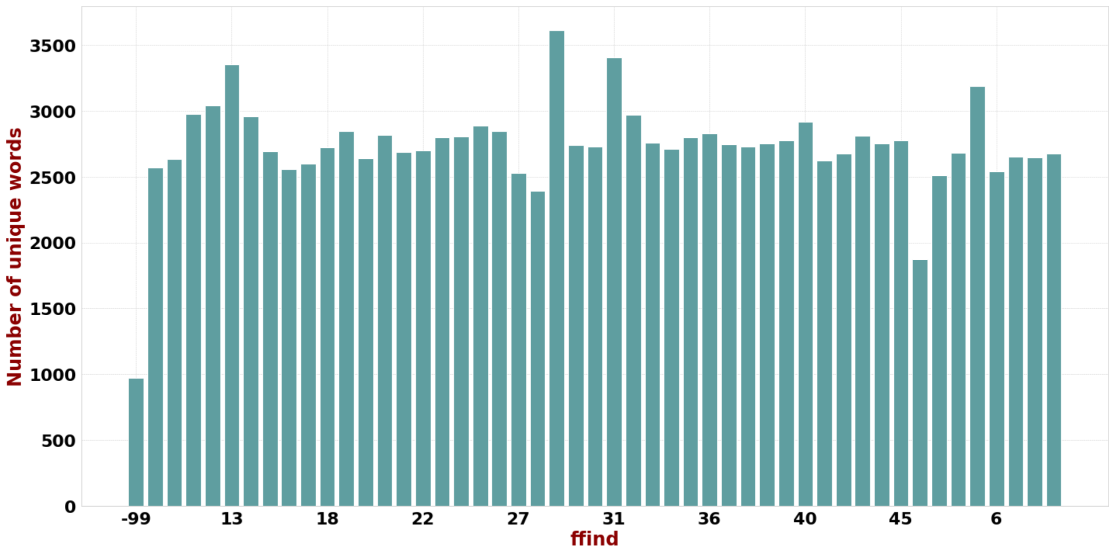


Figure 4.4: Word usage distribution across FF-industries

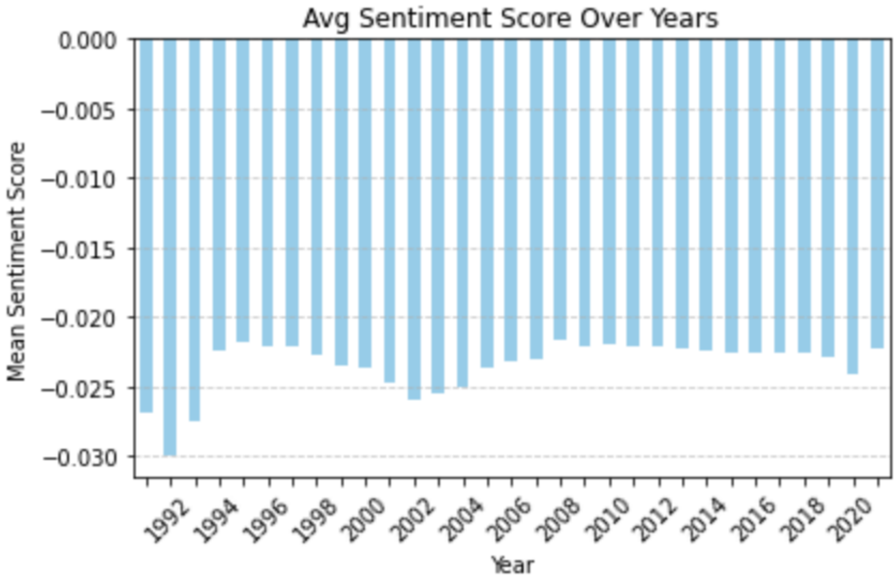


Figure 4.5: Avg Sentiment Score Over Years

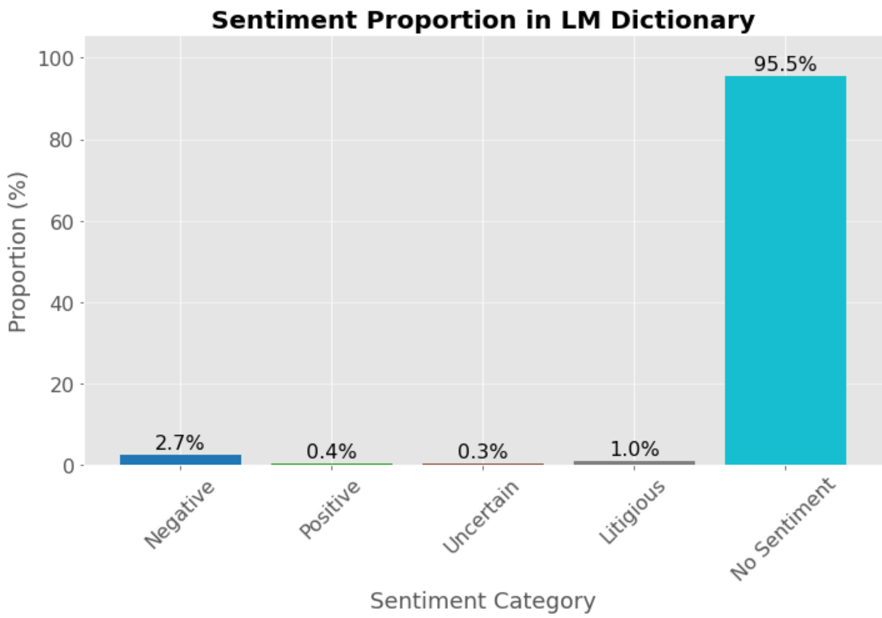


Figure 4.6: Sentiment Proportion in LM dictionary

5 RESULTS

5.1 SENTIMENT BASED STRATEGY

Sentiment analysis has been widely applied to various textual sources, including news articles, analyst transcripts, customer reviews etc to develop investment strategies. Similarly we can apply it to financial documents such as 10Ks and 10Qs, utilizing sentiment words extracted from dictionaries like the Loughran and McDonald Financial Sentiment dictionary. Upon examining the results of such strategy as shown in Fig 5.1, we observe that it performs well within the period spanning from 1995 to 2008. This timeframe coincides with the period used for training the sentiment dictionary, and the strategy's success during this period is consistent with the dictionary's effectiveness within its training data. However, when assessing the out-of-sample performance, the strategy's effectiveness is not as robust. Over time, these financial reports have become more complex, incorporating a broader range of sentiment-laden language that may not be adequately covered in existing sentiment dictionaries such as the LM (Loughran and McDonald) dictionary. Additionally, as markets evolve, they may become more efficient in directly incorporating the sentiment expressed in these reports into asset prices.

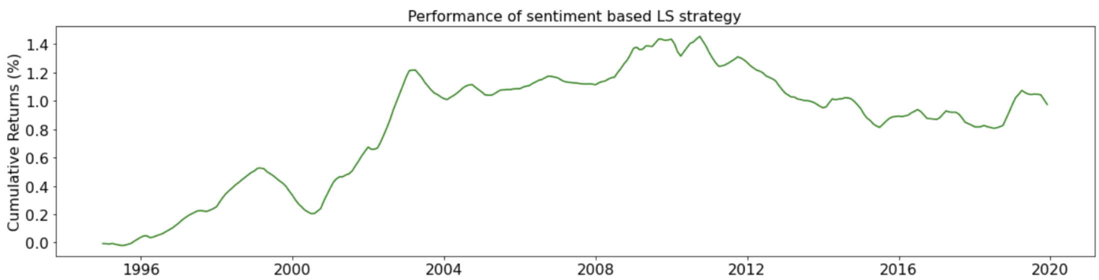


Figure 5.1: Performance of Long Short Strategy based on LM Dictionary Sentiment Score

5.2 CHANGERS VS NON CHANGERS STRATEGY : BAG OF WORDS MODEL

Next, we analyze the performance of a long-short strategy based on quarter-over-quarter changes in financial reports. To delineate the optimal holding period and lookback period for constructing such a long-short portfolio, we partition the time frame into two distinct periods: an in-sample period spanning from 1995 to 2014, and an out-of-sample period from 2015 to 2018. The results for similarities computed using similarities such as Jaccard and Cosine can be seen in the Table 5.1 and Table 5.2 respectively.

Table 5.1: Portfolio Performance (1995-2014): Jaccard Similarity

Holding Period	Lookback Period	Sharpe Ratio	Max Drawdown	Hit Rate	Monthly Returns	Monthly Vol
1	1	1.73	-12%	69%	124.0bps	249.0bps
1	3	2.79	-7%	79%	117.0bps	145.0bps
1	6	3.8	-4%	84%	108.0bps	99.0bps
3	3	2.34	-4%	91%	88.33bps	131.06bps
3	6	2.53	-4%	89%	81.33bps	111.43bps
6	6	2.04	-3%	93%	63.67bps	107.78bps

Table 5.2: Portfolio Performance(1995-2014): Cosine Similarity

Holding Period	Lookback Period	Sharpe Ratio	Max Drawdown	Hit Rate	Monthly Returns	Monthly Vol
1	1	1.27	-19%	67%	89.0bps	244.0bps
1	3	1.94	-10%	74%	75.0bps	135.0bps
1	6	2.61	-9%	76%	73.0bps	97.0bps
3	3	1.86	-6%	82%	62.0bps	115.47bps
3	6	2.07	-5%	84%	60.33bps	100.46bps
6	6	1.58	-5%	89%	46.33bps	101.25bps

The portfolios exhibiting a one-month holding period alongside a six-month lookback period demonstrate the highest Sharpe ratio in-sample. It gives robust performance in out-of-sample data as well, as can be seen in Table 5.3

Table 5.3: Portfolio Performance (2015-2018)
Look Back Period = 6m, Holding Period = 1m

	Cosine Similarity	Jaccard Similarity
Sharpe Ratio	3.21	3.38
Max Drawdown	-4%	-5%
Hit Rate	87%	85%
CAGR	10.94%	13.14%
Monthly Returns	87 bps	103 bps
Monthly Volatility	94 bps	106 bps

As illustrated in Table 5.4, the long-short portfolio demonstrates significant alpha, even after adjusting for market risk factors, according to both the CAPM and the Fama-French three-factor and five-factor models.

	Model	LS	Q1	Q2	Q3	Q4	Q5
0	CAPM	0.42	0.53	0.8	0.6	0.76	0.96
	(T-stat)	(3.44)	(2.64)	(3.76)	(2.96)	(3.71)	(4.72)
1	3 Factor Model	0.43	0.49	0.76	0.55	0.72	0.91
	(T-stat)	(3.53)	(2.45)	(3.6)	(2.78)	(3.54)	(4.57)
2	5 Factor Model	0.4	0.49	0.76	0.52	0.74	0.9
	(T-stat)	(3.17)	(2.36)	(3.45)	(2.55)	(3.5)	(4.35)

Table 5.4: α and t-stat values for different models.

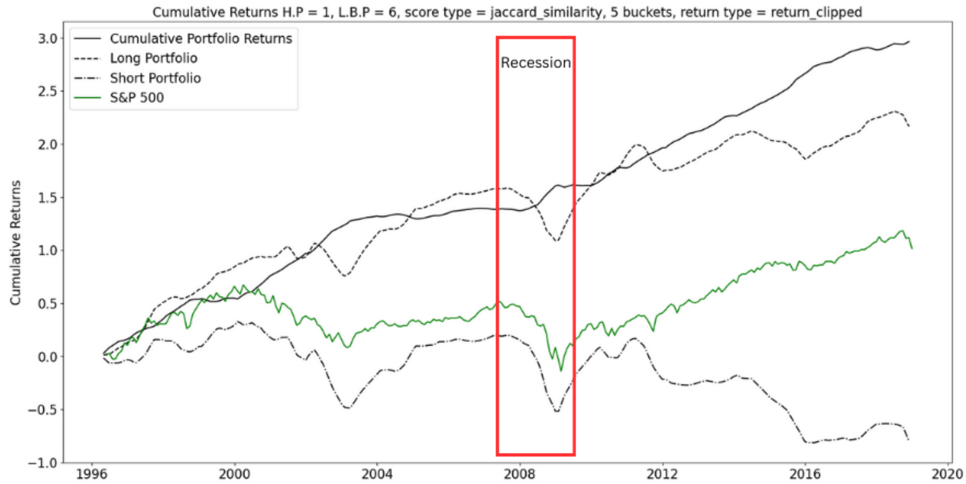


Figure 5.2: Jaccard Similarity Holding Period: 1 month, Lookback Period: 6 months

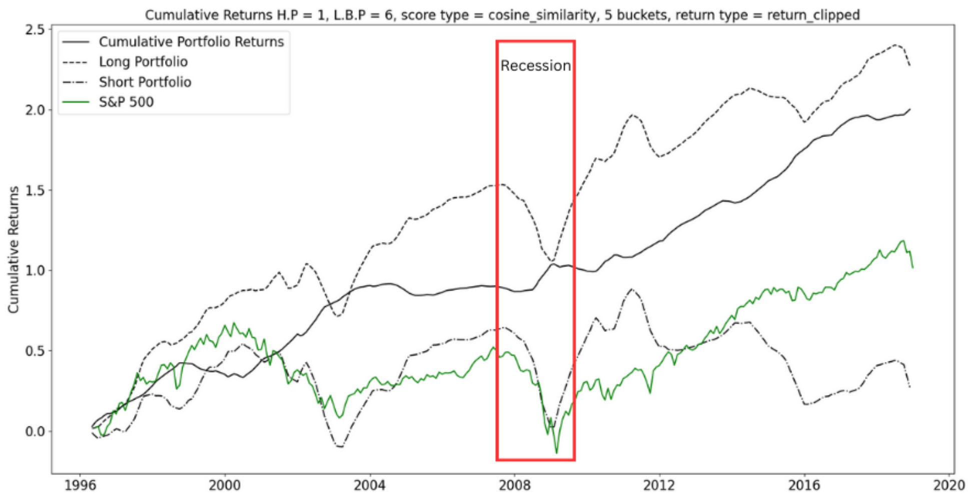


Figure 5.3: Cosine Similarity Holding Period: 1 month, Lookback Period: 6 months

From Fig 5.2 and 5.3, it is evident that the long-short strategy outperforms the S&P 500. However, it's notable that the breadth of this portfolio (all US stocks filing 10Ks and 10Qs) surpasses that of the S&P 500, indicating a broader selection of stocks. We're aware that achieving higher Sharpe ratios is often easier with small-cap stocks, but this advantage may diminish when considering the trading costs associated with small-cap stocks. To assess the robustness of this strategy, we evaluate its performance by confining the stock universe exclusively to the S&P 500.

The strategy demonstrates robustness across various stock selections, exhibiting promising results with Sharpe ratios of 1.02 and 0.72 for cosine and Jaccard similarity respectively, within the S&P 500 stock universe (Table 5.5). The decrease in Sharpe ratio from above 2 to approximately 1 underscores a key theory explaining the strategy's effectiveness: the tendency for investors to overlook critical details found in company filings such as 10Ks/10Qs, despite containing valuable insights into stocks. These insights often remain unincorporated into prices until they become public news. Given that S&P 500 stocks represent some of the largest and most widely followed companies in the US, they attract significant investor attention, leading to greater market efficiency. Analyzing return plots over time reveals that the portfolio consistently outperforms the S&P 500, particularly during recessions, suggesting its potential as an effective hedge. Furthermore, since the universe of stocks is limited to the S&P 500, it demonstrates effectiveness for large-cap stocks as well.

Based on the regime analysis results presented in Table 5.6, it is evident that this strategy exhibits a performance that surpasses that of the S&P500 index while demonstrating a regime-agnostic behavior.

Table 5.5: Performance Metrics on S&P500 stocks

Metric	**Cosine Similarity**	**Jaccard Similarity**
Sharpe Ratio	1.02	0.72
Max Drawdown	-12%	-13%
Hit Rate	60%	57%
Monthly Returns	49 bps	40 bps
Monthly Volatility	168 bps	197 bps

Table 5.6: Regime analysis for all filings

Metric	**Cosine**			**Jaccard**		
	Recession	**Expansion**	**Overall**	**Recession**	**Expansion**	**Overall**
Sharpe Ratio	2.22	3.49	2.61	3.16	5.02	3.80
Max Drawdown (in %)	-2	-4	-9	-2	-2	-4
Hit Rate	77%	81%	76%	83%	87%	84%
CAGR	10.69%	10.69%	9.13%	16.49%	15.20%	13.82%
Monthly Returns	85.0bps	85.0bps	73.0bps	128.0bps	118.0bps	108.0bps
Monthly Volatility	133.0bps	84.0bps	97.0bps	141.0bps	82.0bps	99.0bps

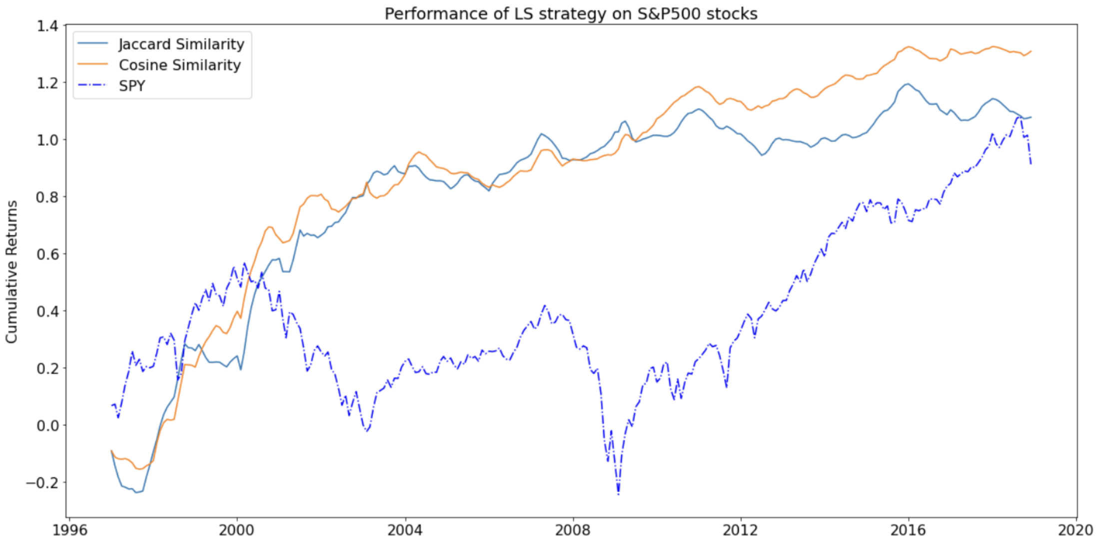


Figure 5.4: Performance of Long-Short strategy on SP500 stocks

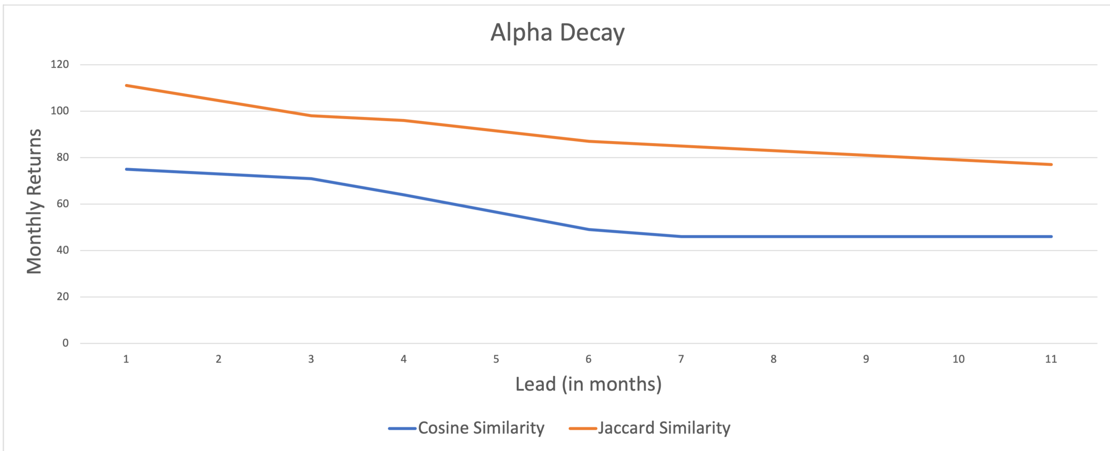


Figure 5.5: Returns decay

We endeavored to examine how altering the initial period of holding the long-short portfolio based on the time elapsed since the filing date affects its impact. Our findings reveal that as the months progress since the filing, there is a diminishing alpha (Fig 5.5), indicating investors' tendency to underreact to these filings.

5.3 Changers vs Non Changers Strategy : BERT Model

Given BERT models' ability to consider contextual nuances, we anticipate their superior ability in distinguishing between changers and non-changers. This prompts our exploration of this approach using FinBERT. However, employing such models demands substantial computational resources and time. Consequently, we opt to test this strategy on the 2012-2013 timeframe for S&P 500 stocks, a domain traditionally challenging for alpha generation. The selection of this timeframe was arbitrary, aiming for robustness in our findings. Encouragingly, we get promising results (Table 5.7) for the BERT model within this constrained period. This suggests the potential utility of deploying a robust long-short strategy using BERT model post extensive backtesting across extended timeframes and further validation checks

Given BERT model has certain limitations, i.e., BERT processes text by splitting it into fixed-length sequences, meaning it can only consider a limited context around each word. This can be a limitation when dealing with long-range dependencies or understanding the context of very long documents. Hence, we compute a section-wise embedding to ensure maximum context capture as opposed to applying it over the entire text and then computing the embedding of the document.

We apply the BERT and BOW models across various sections of 10Ks and 10Qs, finding notably improved performance within the MD&A section, aligning with expectations from existing literature. Our observations indicate a concentration of firms' reporting changes within the Management's Discussion and Analysis of Financial Condition and Results of Operations (MD&A) section. This section affords management the highest level of discretion and flexibility regarding content. Furthermore, we find that the wording within the MD&A section reliably predicts significant abnormal returns, surpassing the effects associated with

		Business	Risk	QMR	MDA	S&P 500
Sharpe Ratio	**BERT**	-1.71	-0.4	1	1.53	1.52
	BOW	0.85	0.28	0.09	1.38	
Max Drawdown	**BERT**	-17%	-9%	-3%	-1%	0%
	BOW	-10%	-11%	-8%	-4%	
Hit Rate	**BERT**	37%	0.54	0.58	0.75	0.75
	BOW	50%	0.45	0.58	0.62	
Monthly Returns	**BERT**	-98bps	-24bps	19bps	53bps	1bps
	BOW	45bps	18bps	3bps	34bps	
Monthly Vol	**BERT**	199bps	215bps	67bps	121bps	2bps
	BOW	185bps	229bps	127bps	87bps	

Table 5.7: Performance metrics for different models.

other sections such as "Quantitative and Qualitative Disclosures About Market Risk" (Item 7a) and notably, the "Risk Factors" section (Item 1A).

6 Future Steps for this project

- **Extended Period Study for BERT:** Apply BERT-based analysis over a longer timeframe to evaluate the robustness and consistency of BERT-driven investment strategies under diverse market conditions. This expanded investigatin ensure that this approach remains effective beyond the two-year period chosen in this study, covering a broader array of economic cycles and market fluctuations.
- **Cross-Industry Comparison:** Expand the analysis to encompass a variety of industries, examining the impact of similarity in corporate disclosures on stock performance sector-wise. This involves evaluating the sensitivity of different industries to nuances in financial reports, identifying sector-specific trends, and anomalies. Such an expansion is crucial for understanding how industry-specific factors influence the predictive power of corporate disclosures' qualitative aspects and the robustness of the positive similarity effect in the financial markets.
- **International Markets Analysis:** Extend the study to include international markets, assessing the influence of corporate disclosures on global stock prices. This comparison aims to refine sentiment-based trading strategies by considering various regulatory environments and cultural contexts, enhancing the understanding of global market responsiveness to corporate disclosures. It is particularly relevant for adapting the investment strategies to the diverse transparency and disclosure standards across countries.
- **Analysis of Dependence on Other Factors:** Investigate the interplay between sentiment-based strategies and other market factors such as momentum and value. This analysis is essential to discern whether the market's response to corporate disclosures is influenced

by or coincidental with other financial indicators, thereby offering a more nuanced view of the investment strategy's effectiveness in light of the broader market dynamics.

7 Conclusion

In this report, we investigate the application of Natural Language Processing (NLP) techniques to EDGAR filings for the purpose of refining portfolio construction strategies. Through an in-depth analysis, we developed changers vs non-changers long-short strategies, leveraging the linguistic nuances found in corporate disclosures. Our research extends the insights from the "Lazy Prices" study to include more recent datasets, confirming the presence of alpha, though at a reduced rate, which highlights the enduring significance of textual analysis in financial modeling.

We specifically examined the performance of various NLP methodologies, comparing traditional bag-of-words approaches against more sophisticated transformer-based models like BERT. Our results showcased the superior ability of advanced models to discern the intricate aspects of corporate communications and their consequent effects on market dynamics.

Our analysis primarily focused on the importance of similarity metrics in analyzing corporate financial statements. Utilizing cutting-edge NLP techniques, such as Large Language Models (LLMs), we conducted a quantitative assessment of semantic similarities within an extensive collection of 10-K filings. The creation of a long-short portfolio based on these similarity scores revealed notable alpha, highlighting the effectiveness of employing textual analytics in financial strategy formulation. The consistency of this alpha across different market scenarios attests to the resilience of our approach, further establishing the integration of linguistic analysis as a valuable component in investment strategies. These results not only validate the success of our quantitative approach but also set the stage for future explorations into augmenting portfolio management through advanced textual analyses.

Overall, our findings reinforce the importance of detailed text analysis in deriving meaningful insights for investment strategies, demonstrating its relevance amidst the changing landscape of financial markets and the growing complexity of corporate reporting.

REFERENCES

[1] Karim Bannouh, Derek Geng, and Bas Peeters. Filing, fast and slow: Reporting lag and stock returns. *Available at SSRN 3432446*, 2021.

[2] Azi Ben-Raphael, Zhi Da, and Ryan Israelson. It depends on where you search: Institutional investor attention and underreaction to news. *Review of Financial Studies*, 30:3009–3047, 2017.

[3] Stephen V. Brown and Jennifer Wu Tucker. Large-sample evidence on firms' year-over-year md&a modifications. *Journal of Accounting Research*, 49:309–346, 2011.

[4] Lauren Cohen, Christopher Malloy, and Quoc Nguyen. Lazy prices. *The Journal of Finance*, 75(3):1371–1415, 2020.

[5] Zhi Da, Joey Engelberg, and Pengjie Gao. In search of attention. *Journal of Finance*, 66:1461–1499, 2011.

[6] Jamie Diaz, Kenneth Njoroge, and Philip B Shane. 10k complexity and the market efficiency and information asymmetry implications of analysts' annual earnings forecasts. *Available at SSRN 3337793*, 2020.

[7] Travis Dyer, Mark Lang, and Lorien Stice-Lawrence. The evolution of 10-k textual disclosure: Evidence from latent dirichlet allocation. *Journal of Accounting and Economics*, 64:221–245, 2017.

[8] Joey Engelberg, Caroline Sasseville, and Jared Williams. Market madness? the case of mad money. *Management Science*, 58:351–364, 2012.

[9] Ronen Feldman, Suresh Govindaraj, Joshua Livnat, and Benjamin Segal. The incremental information content of tone and sentiment in management discussion and analysis. *SSRN Electronic Journal*, 2008.

[10] Ingrid E Fisher, Margaret R Garnsey, and Mark E Hughes. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, 23(3):157–214, 2016.

[11] Yen-Jung Lee. The effect of quarterly report readability on information efficiency of stock prices. *Contemporary Accounting Research*, 29:1137–1170, 2012.

[12] Feng Li. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and economics*, 45(2-3):221–247, 2008.

[13] Feng Li. The information content of forward-looking statements in corporate filings - a naïve bayesian machine learning approach. *Journal of Accounting Research*, 48:1049–1102, 2010.

[14] Tim Loughran and Bill McDonald. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54:1187–1230, 2016.

[15] Karen K Nelson and Adam C Pritchard. Litigation risk and voluntary disclosure: The use of meaningful cautionary language. In *2nd annual conference on empirical legal studies paper*, 2007.

[16] Matus Padysak. The positive similarity of company filings and the cross-section of stock returns. *Available at SSRN 3690461*, 2020.

[17] Paul C Tetlock. Information transmission in finance. *Annu. Rev. Financ. Econ.*, 6(1):365–384, 2014.

[18] Yi Yang, Mark Christopher Siy Uy, and Allen Huang. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*, 2020.