Enhancing Stock Market Return Predictions from 10-K Filings: Leveraging the Loughran–McDonald Dictionary and Sentiment Analysis with BERT

Radu Mistreanu

Vrije Universiteit Amsterdam, 1081HV Amsterdam, Netherlands
r.mistreanu@student.vu.nl

Abstract. Stock market prediction using natural language processing (NLP) is a promising new area of research in financial analytics. Accurately assessing a firm's financial performance from annual reports is challenging, even for state-of-the-art NLP pipelines. To address this, we create a novel tool for predicting stock market returns immediately following publication of the 10-K financial report. First, we introduce a data preparation pipeline based on 10-K filings of S&P500 companies over a period of ten years. Second, we combine the Loughran–McDonald dictionary and pre-trained BERT models to extract relevant information from 10-K reports and calculate sentiment scores. Correlation between sentiment score and stock returns is statistically significant for 3-day and 5-day returns. Third, we benchmark and analyse the predictive capabilities of different transformer models on the 10-K dataset. Best performance is achieved using FinBERT, highlighting the importance of domain-specific fine-tuning for financial sentiment analysis. Even if effect is not causal, financial tone is an important proxy for understanding the impact of information on stock returns. Our experiments highlight promising pathways towards more accurate automatic textual analysis tools in finance.

Keywords: Stock Market Prediction · Sentiment Analysis · Transformer Models.

1 Introduction

The financial industry is among the most data-intensive sectors of the global economy. Each moment, the global finance ecosystem is producing petabytes of structured and unstructured data from retail and corporate banking, capital markets, insurance services, financial services security, and many others. Furthermore, financial institutions, banks, and financial technology (Fintech) firms leverage large amounts of data to analyse financial markets and instruments, track and manage financial assets, predict and anticipate market behaviour and economic forces, and deploy various analytic and predictive tools [17]. This wealth of data often needs to be documented, analysed, and communicated through financial reports, ranging from daily internal expense reports, all the way to comprehensive annual reports and audits.

Historically, such analysis of financial data revolved around structured, numerical data involving statistical and time-series analyses. However, natural language processing (NLP) developments in the last few decades have enabled, for the first time, the use of unstructured, textual data for generating new insights in the financial domain. With an increase in the volume of textual data [7], and an increase in the ratio of data being in textual form [12], NLP is becoming an indispensable tool for processing financial data.

One of the most important sources of textual information and of notable significance for financial reporting is the 10-K report. It is an annual financial document mandated by the U.S. Securities and Exchange Commission (SEC), ensuring that publicly traded companies disclose a thorough and standardised set of information. Given its importance as a comprehensive review of a company's financial condition and business operations over an entire fiscal year, 10-K reports have been used extensively in NLP tasks for predicting stock volatility [16], IPO valuation [18], brand equity [10], or period returns [13].

Pertaining to the challenges mentioned above, 10-K reports are getting longer without any increase in the informativeness of the content [6]. The disclosures are getting more redundant and verbose, while being less specific and readable. The task of extracting important information from the reports for decision-support is mostly done manually and is very labour-intensive. This points to an increased need for approaches that automatically analyse annual reports, increasing objectivity and efficiency.

One of the most influential papers tackling automatic textual analysis of 10-K reports is Loughran & McDonald (2011) [13]. The authors show that word lists developed for other disciplines misclassify common words in financial texts, and create a comprehensive dictionary of terms to that can be used in many financial sentiment analysis tasks. The bag-of-words (BoW) representations of 10-K reports introduced in this paper are found to be predictors of market reactions around the filing date of the 10-K report. Word counts in the Loughran-McDonald dictionary are linked to filing returns, trading volume, stock volatility, and unexpected earnings. For example, increased negative word counts are correlated with lower filing returns immediately following publication of the 10-K report, and the effect is monotonic and statistically significant.

However, BoW representations of text treat words independently and ignore the order in which they appear. They cannot capture context derived from the sequence of words, and they lack understanding of the semantic relationships between related words. All subsequent advances in NLP aim to improve on these deficiencies by encoding word dependencies, context-specific meanings and underlying stylistic nuances like figures of speech and idioms.

At the root of ongoing breakthroughs in NLP capabilities are the current state-of-the-art transformer models [19]. Through contextual word embeddings and complex self-attention mechanisms, transformers have significantly advanced tasks such as language modelling, text generation, machine translation, and sentiment analysis. These models excel in capturing long-range dependencies and contextual nuances in language, leading to state-of-the-art performance across

various NLP benchmarks. Moreover, transformers have facilitated transfer learning, allowing pre-trained models to be fine-tuned on specific tasks with minimal data, accelerating research and development in the field.

Of particular importance for the present paper, BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model extensively used in textual analysis and NLP pipelines [5]. BERT is an encoder model that takes a sequence of text as an input, and produces a vector that encapsulates complex word dependencies, meaning, and context. This vector can then be used text classification, sentiment analysis, information retrieval, question answering, and other tasks.

While BERT is a generalised model trained on large and diverse corpora of texts, there exist specialised BERT models trained on text specific to a single discipline or field of study. FinBERT [1] is a specialized variant of BERT, tailored specifically for financial domain applications. Developed by further pre-training the original BERT model on a large corpus of financial texts, FinBERT outperforms state-of-the-art machine learning methods for financial tasks. This specialisation enables FinBERT to better handle the jargon, context, and unique language patterns prevalent in financial texts, making it an invaluable tool for tasks like market sentiment analysis, risk assessment, and automated financial advising.

Moreover, transformer models exhibit remarkable generalisability. Their architectures and training paradigm ensure that learned dependencies carry over to similar tasks. Models that have been trained for one specific task or on one specific dataset retain most of their predictive power for related datasets. For example, a transformer model that has been trained to predict stock volatilities from financial reports should maintain some of its inferencing capabilities for predicting volatility from news articles instead. This is a valuable characteristic of transformers that is sought-after in financial applications of NLP, given the relatively small quantity of publicly available training data when comparing finance to other disciplines.

These factors ensure that transformers are being used extensively in financial applications. In a review of NLP applications in FinTech [3], over half of the papers surveyed employ transformer architectures in their pipelines. For example, BERT models are utilised to encode textual data related to the fear index and on S&P 500 index movement prediction [20], and to develop models for foreign currency exchange market predictions [4].

Problem Definition and Contribution. Assessing a firm's financial performance through annual reports, particularly the 10-K filings, remains a challenging task even for advanced NLP pipelines. Current methodologies may be limited by the complexity, verbosity, and often redundant parts of these documents, making it difficult to extract meaningful insights accurately. The goal of the present paper is to address these limitations by introducing a novel NLP pipeline for the task of predicting stock market returns immediately following the publication of 10-K financial reports. The proposed tool integrates sentiment

analysis using the Loughran-McDonald dictionary and pre-trained transformer models to evaluate the relationship between sentiment analysis of the report and the market reaction. Our research aims to highlight promising avenues towards developing better automatic financial text analysis tools.

The main contributions of this study are as follows:

- We introduce a data preparation pipeline for extracting sentiment scores based on 10-K filings from S&P 500 companies over a ten-year period.
- We develop a sentiment analysis tool by applying the Loughran-McDonald dictionary and pre-trained BERT models to calculate sentiment scores for 10-K reports, demonstrating a statistically significant correlation with stock returns[1].
- We benchmark several model configurations and different BERT algorithms to identify the most effective model for predicting stock market returns from 10-K reports.

2 Related Work

As presented before, Loughran & McDonald (2011) [13] provides the basis for significant parts of the present study. The authors observe that word lists developed for disciplines other than finance misclassify as negative three quarters of words that are not considered negative in financial contexts. They introduce a comprehensive dictionary of terms (from now on referred to by its abbreviation LMD) which includes six word lists (positive, negative, uncertain, litigious, strong modal, and weak modal). These lists were created from a large corpus of 10-K filings and are strong predictors for financial texts. They show that their custom word lists provide a more precise measure of tone in financial reports, significantly correlating with market reactions like 10-K filing returns, trading volume, return volatility, fraud, and unexpected earnings. The present study utilises the positive and negative words lists and the same measurement of filing returns, while addressing the limitations of BoW representations using BERT models.

Textual analysis of 10-K reports using word count models and term weighing is a compelling and proven technique. Support vector regression models and unigram and bigram representations of 10-K reports have been successfully used to predict stock price volatility [11]. In the authors' words, "a very simple representation of the text" can rival models based on past volatility for predicting the target variable. Moreover, LMD word lists have been incorporated alongside information-retrieval term-weighing models for stock volatility predictions, and the models outperform state-of-the-art regression models based on historic prices [16].

Given the substantial length of 10-K reports, most textual analyses employ some form of textual extraction. While some previous analyses utilise specific

[1] Model and data are available on https://github.com/radu-mistreanu/10K-LMD-FinBERT

sections of the report, [6, 8], others extract forward-looking statements (FLS), defined as "short sentences that contain information likely to have (...) a direct effect in the foreseeable future," for their predictive models [15]. The current paper adopts both paradigms, by examining sections 1A, 3, 7, and 7A of the 10-K report, and employing sentence extraction using the LMD word lists and forward-looking statements.

Transformer architectures have also been applied to 10-K textual analyses. For example, BERT models have been used to extract FLS from 10-K filings and to determine financial sentiment scores from FLS for predicting stock prices [8]. Although the authors' paper focuses on FLS extraction pipelines rather than the predictive power of the model, their sentiment analysis pipeline is robust and sound. In this study, we have chosen to enhance and extend their pipeline by incorporating sentence extraction based on the LMD word lists.

To our knowledge, no study has examined the inferencing capabilities of an NLP pipeline that combines the Loughran-McDonald dictionary with sentiment analysis using transformer models for predicting stock returns after the publication of 10-K reports. This study aims to leverage the statistically significant correlations identified by Loughran & McDonald (2011) along with the context-rich processing capabilities of state-of-the-art transformer models. Unlike previous studies, this paper measures stock returns following the release of 10-K reports using a novel NLP pipeline that integrates LMD with transformer-based sentiment analysis.

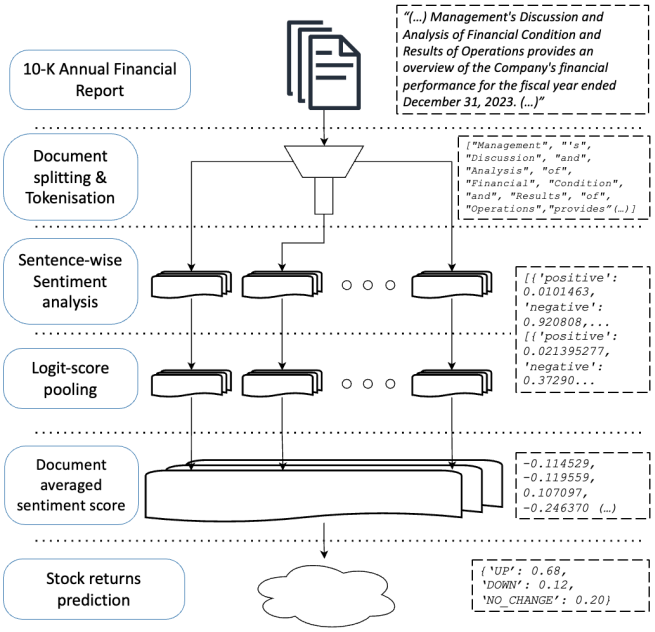

Fig. 1. High-level overview of the model architecture.

3 Methodology

This section outlines the methodology behind the three main contributions of this paper. It is divided into sub-sections that cover data collection and processing, model architectures, design decisions, and assumptions relevant to each part of the study. The sub-sections, in chronological order, describe: the novel data processing pipeline for extracting relevant sentences based on LMD from a large dataset of publicly available 10-K reports, the sentiment analysis pipeline for anticipating short-term market movements, and benchmarking of various model configurations.

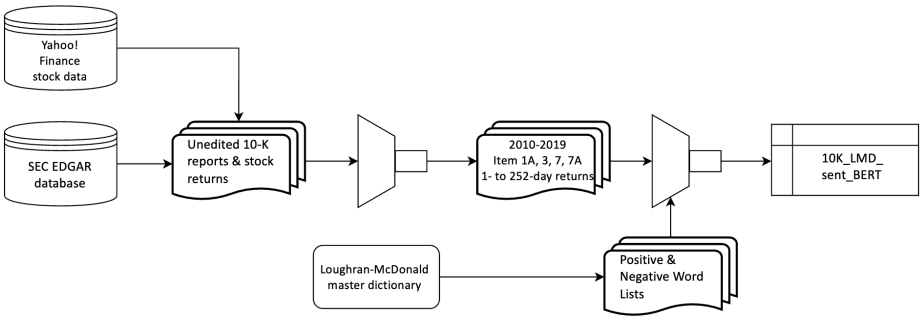

Fig. 2. Detailed diagram of data processing pipeline.

3.1 Data Collection and Processing

The SEC has established the EDGAR (Electronic Data Gathering, Analysis, and Retrieval) system, which provides public online access to 10-K reports and other filings, enhancing transparency and accessibility. This facilitates the collection of reports for automatic textual analysis and the dissemination of high-quality, open-source datasets containing extensive 10-K reports. As a result, researchers, investors, and policymakers can more effectively analyse financial data and trends, fostering a more informed and efficient market.

The dataset used in this study originates from such a source[2]. It includes annual reports of all S&P 500 historical constituents from 2010-2022, sourced from SEC EDGAR Form 10-K filings, along with n-day future returns of each firm's stock price from the filing date. Data collection and processing were done using a scraper algorithm [14]. We chose to retain only the 10-K filings from 2010-2019 to exclude financial data affected by the COVID-19 pandemic and its associated market uncertainty. Predicting stock returns during such instability is beyond the scope of this research. By focusing on a period of relative market

[2] https://huggingface.co/datasets/jlohding/sp500-edgar-10k

stability, we aim to achieve more accurate predictions and avoid the confounding variables related to the pandemic years.

Each 10-K report entry in the dataset is divided into sections. This study focuses on sections 1A (Risk Factors), 3 (Legal Proceedings), 7 (Management's Discussion and Analysis of Financial Condition and Results of Operations), and 7A (Quantitative and Qualitative Disclosures About Market Risk), as these have been identified as the most informative regarding forward-looking statements [8]. Additionally, for each 10-K entry, the dataset includes corresponding n-day stock returns following the 10-K filing, ranging from 1-day to 252-day returns. Previous studies [9] have identified that investor response to 10-K filings persists for 5 days after the filing date. Therefore, it is expected that sentiment scores identified by the model will show a statistically significant correlation with 1- to 5-day returns.

Out of the six words lists composing the Loughran-McDonald dictionary, we use the positive and negative words list for extracting relevant sentences from the 10-K dataset, as these two lists are most closely correlated with stock returns [13]. Since publication in 2011, the Loughran-McDonald master dictionary has been updated yearly to include the words that are the strongest predictor of financial mood, and latest version at the time of writing contains a total of 2692 words in the positive and negative categories that have been used in the present study.

Across the whole 10-K dataset, sections 1A, 3, 7, and 7A of each entry contain an average of 900 sentences. Extracting only the sentences that contain words from the LMD, each financial report retains an average of 300 sentences deemed relevant for stock return prediction. It is hypothesised that most of the sentences not captured by the dictionary will not carry information that is relevant to the financial mood of the overall report, and will have no meaningful effect on model performance. Any instance in the dataset that retained less than 10% of the report or was deemed to contain too few sentences for achieving a good prediction, and was eliminated from the final dataset.

3.2 Model Architecture

Following from the data preparation step presented above, the current section will describe the model architecture and analysis pipeline for the 10-K dataset. Figure 1 describes a high-level overview of the textual analysis pipeline implemented in this project. It shows the progression from raw textual data from the reports, through to sentence splitting and tokenisation, leading to sentence-wise sentiment analysis, score pooling and prediction generation.

BERT models trained for sentiment analysis process individual tokenised sentences and generate probability distributions over three sentiment classes (positive, neutral, and negative). By default, they output the class with the highest probability. To obtain a sentiment score for entire documents, we use the output from the softmax activation layer, resulting in a one-dimensional vector with values for positive, neutral, and negative scores for each sentence. We then average the positive and negative scores across all relevant sentences in the report. The

final document-wide score is calculated by subtracting the average negative score from the average positive score. This approach [1] produces an intuitive score: -1 represents a completely negative document, 0 a perfectly neutral document, and 1 a perfectly positive document.

Due to the sentence selection process using the LMD positive and negative word lists, the raw text for each instance in the dataset carries minimal long-term dependencies and context beyond each sentence. Individual sentences serve as the ideal minimum unit of meaning for sentiment analysis, since dependencies between sentences have a negligible effect on the overall sentiment [8]. Sentence-wise sentiment analysis also addresses one of the main limitations of BERT models, namely, the small maximum token size. Nearly every sentence in the 10-K dataset can be represented in its entirety without truncation or loss of meaning.

N-day stock returns are calculated as the ratio of the stock price n days after filing to the price at the time of filing. For example, a 3-day return value of 1.02 indicates a 2% increase in stock value three days after the report filing. In the first part of the analysis, this ratio is used to calculate the correlation between document sentiment scores and stock returns. In the second part, the numerical values are converted to categorical variables to indicate whether the stock has shifted in value beyond a certain significant threshold. The models' performance is evaluated based on a classification task, predicting whether the stock is likely to experience a significant shift in value based on the company's financial disclosure in the 10-K filing.

To generate the final prediction on the classification task described above, the document sentiment scores are processed by a logistic regression algorithm. The threshold for what constitutes a sudden shift in value for this study is set to capture the 25% top and bottom price events in the data. Logistic regression was chosen for its easily interpretable output (a probability distribution over the three classes) and computational efficiency. Given that most information is already encoded in the document sentiment score, the low-dimensionality of the data lends itself well to logistic regression, and does not benefit from models such as support vector machines, which excel on high-dimensional data and complex relationships in the data.

The stock value thresholds for what constitutes a sudden shift in value will lead to a notable class imbalance in the dataset. Given the emphasis on identifying and flagging financial reports which may see considerable spikes in value, preliminary testing has revealed that oversampling the minority classes leads to better performance metrics. Therefore, the training set for all benchmarked models will contain balanced class distributions obtained using Synthetic Minority Over-sampling Technique (SMOTE) which generates synthetic samples for the minority class by interpolating between existing samples.

The final design consideration relates to studying the effect of the sentence selection using the LMD and sentence-wise sentiment analysis on the predictive power of the model. To test this, an additional transformer model that does not use LMD selection and sentence-wise sentiment analysis is introduced. This

model takes as input entire 10-K sequences and is trained on direct sequence classification without the proxy of sentiment scores. To get around the limited token size of traditional BERT models which is insufficient for entire 10-K report sections, the model chosen for this analysis is Longformer [2]. Due to a different attention mechanism, the Longformer model has a maximum token size 8 times larger than BERT, while retaining similar performance. The performance of this model is included in the results section below.

3.3 Baseline Methods

For contrastive experiments, we consider whether financial text-specific fine-tuning contributes to better inferencing capabilities, and quantify the impact of sentence selection using LMD on extracting relevant information for stock price prediction performance. As such, the current paper investigates BERT-base, FinBERT, and Longformer-base. For BERT-base and FinBERT, their outputs will be passed through the logistic regression algorithm mentioned above, and all three will be compared on the classification task described in the previous section.

Related to the first goal, one of the main aims of the current experimental setup is to establish whether domain-specific fine-tuning leads to a corresponding increase in predictive power. As presented in previous sections [13], words that may be considered negative in normal speech represent normal terms in financial applications (loss, liability, cost, etc.). It is hypothesised that general-purpose sentiment analysis models like base BERT will exhibit the same behaviour, leading to a skewed outlook of the overall sentiment of a 10-K report.

The design requirement of providing a practical tool to support investors' decision-making and complement traditional financial analytics impacts how model performance should be measured. Since the model's goal is to flag reports indicating a sudden shift in stock value based on financial tone, the model's sensitivity (the proportion of actual positives correctly identified) is the key metric for determining its usefulness. Addressing the dataset's class imbalance with SMOTE increases sensitivity but can negatively affect other performance metrics, such as accuracy and precision, resulting in a trade-off. As the final judgment lies with investors and financial analysts, prioritizing sensitivity (recall) over precision is desirable.

4 Results

After calculating 10-K document-averaged sentiment scores, the numerical n-day stock returns ratios were used to study correlation with sentiment scores. Table 1 shows each of the n-day stock returns timeframes and their corresponding Pearson correlation coefficients and p-values. As expected, FinBERT produces sentiment scores that have a statistically significant positive correlation with 3-day and 5-day stock returns. Correlation of 1-day returns is also notable, but fails short of achieving statistical significance (p-value 0.064). Correlation decreases

with time frames longer than 5-days, becoming negligible past 10-day returns, and the results are not statistically significant. Sentiment scores calculated using BERT show negligible correlation with stock returns, and the results are not statistically significant.

These findings are consistent with literature and past analyses of 10-K documents [9]. Moreover, BoW models using the same words lists as the present study were also found to be correlated with stock returns [13]. What this experiment shows is that state-of-the-art financial sentiment analysis using LMD word lists and fine-tuned transformer models is also a predictor of stock returns immediately following 10-K publication.

Table 1. Correlation between sentiment scores and stock returns.

Model		n-day returns										
		1	3	5	10	20	40	60	80	100	150	252
BERT	corr	-0.009	-0.013	-0.086	-0.012	-0.054	0.016	0.056	0.009	-0.042	-0.094	-0.094
	p-value	0.830	0.767	0.063	0.786	0.236	0.720	0.220	0.831	0.361	0.041	0.088
FinBERT	corr	0.085	**0.115**	**0.101**	0.051	-0.056	-0.053	-0.029	-0.033	-0.048	-0.015	-0.008
	p-value	0.064	**0.012**	**0.027**	0.269	0.221	0.251	0.529	0.471	0.296	0.737	0.852

The findings demonstrate that financial tone (expressed through document sentiment scores) is an important proxy for the markets' reaction to financial disclosures, and can used to anticipate stock returns up to five days after 10-K filing. While the effect is meaningful, financial sentiment alone cannot explain most of the variation in market movements. Given this, the best use case for such a model is as a support tool for investors that flags reports where financial tone may indicate a significant shift in stock value. As such, focusing on 3-day returns and turning the pipeline to the classification task described in the methodology section, Table 2 shows the results of the benchmarking step.

Table 2. Performance metrics of the tested models.

Model	Weighted accuracy	Recall [UP]	Recall [DOWN]
BERT	0.06	0.56	0.36
FinBERT	**0.74**	**0.64**	**0.67**
Longformer	0.30	0.00	0.00

The performance results highlighted in Table 2 confirm the assumptions mentioned in the methodology section. First, FinBERT outperforms the other configurations, pointing to the importance of fine-tuning in the context of financial sentiment analysis. As theorised, general-purpose sentiment analysis models struggle to accurately represent specialised financial vocabulary, leading to a skewed perception of the tone of the document, and therefore, lower predictive power. Second, LMD significantly enhances the inferencing capabilities of

the model. By using LMD word lists for sentence selection, the models encode essential information necessary for accurate predictions. Without extracting sentences pertinent to financial sentiment, the Longformer model cannot distinguish reports that are associated with sudden shifts in value. Following its training phase, the Longformer model predicts only the majority class (stock value will not change).

5 Discussion

This exploratory NLP analysis and proof-of-concept pipeline for stock market predictions reveals compelling insights into automatic textual analysis of financial reports and prediction models in finance. Firstly, the findings of this study extend the correlation identified in BoW models [13] to transformer models, indicating that the inherent market reaction to financial disclosures can be gauged and predicted using sentiment analysis. Secondly, this paper provides an early step towards more accurate automatic textual analysis tools in finance. Thirdly, even if effect is not causal, financial tone is an important proxy for understanding the impact of information on stock returns.

The real value of the system presented in this paper is in generating new insights using NLP, enhancing conventional analysis methods that rely on statistics, historical data, and time series. Instead of replacing traditional approaches, this system aims to complement the decision-making processes of investors and analysts. By harnessing the capabilities of state-of-the-art NLP tools, this study offers insights that are beyond the reach of conventional methods. For instance, the model captures how market expectations correlate with financial disclosures via financial tone, a discovery unattainable through statistical analyses and historical price data.

However, the current model has significant limitations with regards to its predictive power. Analysis of the results indicates limited accuracy, suggesting that the model should not be trusted blindly and can only be recommended as a support tool. Stock market prediction remains an incredibly complex process that cannot be fully captured by any single model. The current paper has identified a statistically significant correlation between financial tone and stock returns, but it is unable to explain and fully predict price movements. Given the low dimensionality of the data, more powerful classifiers would not increase performance. Integrating additional features like historical data streams could enhance the model's predictive capabilities and applicability in broader financial markets.

Moreover, this study purposefully avoids periods of market uncertainty such as the global financial crisis or the COVID-19 pandemic to maintain the focus on regular market conditions and to establish a baseline understanding of how transformer-based sentiment analysis performs under stable economic environments. This deliberate choice ensures that the findings reflect the typical operational conditions of the market rather than exceptional circumstances that

could skew results or misrepresent the model's effectiveness. Future research would be needed to explore its robustness in more turbulent economic climates.

Generalisability is a key advantage of transformer models that can extend the utility of the described model to other sources of financial data beyond 10-K reports. For instance, these models can be applied to quarterly earnings reports, news articles, analyst notes, or even social media posts related to stocks and companies for predicting market reactions and trends. Future research could explore the adaptation and optimisation of the proposed pipeline for these different data sources, leading towards broader applications in financial analytics.

Future work may also involve a more comprehensive strategy of evaluating the performance of the model through multi-agent simulations of the stock market environment. By simulating interactions among various agents representing different investor behaviours and market conditions, the model's predictive capabilities can be systematically tested under diverse scenarios. This approach would provide valuable insights into how the proposed pipeline might react to dynamic market dynamics, including periods of heightened volatility.

6 Conclusion

This paper set out to introduce a novel NLP pipeline integrating sentiment analysis with pre-trained transformer models to predict stock market returns from 10-K financial reports. By leveraging state-of-the-art NLP techniques and comprehensive financial dictionaries, the pipeline aims to provide a robust tool for financial sentiment analysis and subsequent stock return prediction. Capturing market response to financial disclosures through sentiment analysis is a compelling new method of generating insights in the broader field of financial analytics.

The study demonstrates a practical and effective approach for anticipating short-term market movements, making it a valuable tool for investors and analysts. The use of transformer models, such as FinBERT, coupled with the LMD word lists, enables the extraction of sentiment scores that correlate with stock price changes, providing actionable insights from the 10-K reports. This method not only enhances traditional financial analysis, but also offers a scalable solution for processing large volumes of financial documents.

While this research does not imply causality, the observed correlations between financial tone and stock returns are still meaningful. Financial tone serves as a proxy for market sentiment, reflecting how investors might react to a company's financial disclosures. This underscores the importance of sentiment analysis in understanding market dynamics and investor behaviour.

The report set out to address several key research questions: Can transformer-based sentiment analysis of 10-K reports predict stock returns? How effective is the proposed NLP pipeline in extracting relevant financial sentiment? What is the relationship between sentiment scores and stock price movements? These questions have been answered by demonstrating that the proposed pipeline can

effectively extract meaningful sentiment scores from 10-K reports and that these scores exhibit a statistically significant correlation with short-term stock returns.

The model will not put financial analysts out of business anytime soon. Nevertheless, it provides meaningful conclusions that can augment traditional financial analysis. While human expertise remains irreplaceable, the integration of advanced NLP techniques offers a complementary tool that enhances the analytical capabilities of financial professionals. By automating sentiment extraction and highlighting potential stock movements, the model aids analysts in making more informed decisions.

In conclusion, the introduced NLP pipeline represents a significant advancement in financial sentiment analysis, providing a reliable method for predicting stock market returns based on 10-K financial reports. However, the model's predictive power is limited, and it should be used as a support tool rather than a standalone predictor. Future research should explore its robustness in turbulent economic climates and its applicability to other financial data sources like earnings reports and news articles. Additionally, multi-agent simulations could systematically test the model's predictive capabilities under diverse market scenarios, providing further insights into its performance. This study lays the groundwork for more sophisticated applications of NLP in financial analytics, ultimately contributing to a more efficient and informed market.

Acknowledgments. I would like to express my gratitude to my thesis supervisor, dr Jieying Chen, for her guidance and support throughout the course of this research. Her expertise, insightful feedback, and patience have been invaluable in shaping this thesis. The amount of time and attention dr Chen dedicated to helping me went far beyond what is expected of a supervisor. Without her guidance and assistance, this work would not have been possible. I am profoundly grateful for her influence on my academic and professional journey. Thank you for being an exceptional supervisor.

References

1. Araci, D.: FinBERT: Financial Sentiment Analysis with Pre-trained Language Models (2019). https://doi.org/10.48550/ARXIV.1908.10063, https://arxiv.org/abs/1908.10063
2. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: the long-document transformer (Dec 2020). https://doi.org/10.48550/arXiv.2004.05150, http://arxiv.org/abs/2004.05150, arXiv:2004.05150 [cs]
3. Chen, C.C., Huang, H.H., Chen, H.H.: NLP in FinTech Applications: Past, Present and Future (2020). https://doi.org/10.48550/ARXIV.2005.01320, https://arxiv.org/abs/2005.01320
4. Chen, D., Ma, S., Harimoto, K., Bao, R., Su, Q., Sun, X.: Group, Extract and Aggregate: Summarizing a Large Amount of Finance News for Forex Movement Prediction. In: Hahn, U., Hoste, V., Zhang, Z. (eds.) Proceedings of the Second Workshop on Economics and Natural Language Processing. pp. 41–50. Association for Computational Linguistics, Hong Kong (Nov 2019). https://doi.org/10.18653/v1/D19-5106, https://aclanthology.org/D19-5106

5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018). https://doi.org/10.48550/ARXIV.1810.04805, https://arxiv.org/abs/1810.04805
6. Dyer, T., Lang, M.H., Stice-Lawrence, L.: The Ever-Expanding 10-K: Why are 10-Ks Getting so Much Longer (and Does it Matter)? SSRN Electronic Journal (2016). https://doi.org/10.2139/ssrn.2741682, http://www.ssrn.com/abstract=2741682
7. Fisher, I.E., Garnsey, M.R., Hughes, M.E.: 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 (Jul 2016). https://doi.org/10.1002/isaf.1386, https://onlinelibrary.wiley.com/doi/10.1002/isaf.1386
8. Glodd, A., Hristova, D.: Extraction of Forward-looking Financial Information for Stock Price Prediction from Annual Reports Using NLP Techniques (Jan 2023), https://hdl.handle.net/10125/103313
9. Griffin, P.A.: Got Information? Investor Response to Form 10-K and Form 10-Q EDGAR Filings. Review of Accounting Studies **8**(4), 433–460 (Dec 2003). https://doi.org/10.1023/A:1027351630866, https://doi.org/10.1023/A:1027351630866
10. Huang, C.Y., Liu, P.Y., Xie, S.M.: Predicting brand equity by text-analyzing annual reports. International Journal of Market Research **62**(3), 300–313 (May 2020). https://doi.org/10.1177/1470785319883201, http://journals.sagepub.com/doi/10.1177/1470785319883201
11. Kogan, S., Levin, D., Routledge, B.R., Sagi, J.S., Smith, N.A.: Predicting risk from financial reports with regression. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09. p. 272. Association for Computational Linguistics, Boulder, Colorado (2009). https://doi.org/10.3115/1620754.1620794, http://portal.acm.org/citation.cfm?doid=1620754.1620794
12. Lewis, C., Young, S.: Fad or future? Automated analysis of financial text and its implications for corporate reporting. Accounting and Business Research **49**(5), 587–615 (Jul 2019). https://doi.org/10.1080/00014788.2019.1611730, https://www.tandfonline.com/doi/full/10.1080/00014788.2019.1611730
13. Loughran, T., Mcdonald, B.: When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. The Journal of Finance **66**(1), 35–65 (Feb 2011). https://doi.org/10.1111/j.1540-6261.2010.01625.x, https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2010.01625.x
14. Loukas, L., Fergadiotis, M., Androutsopoulos, I., Malakasiotis, P.: EDGAR-CORPUS: Billions of Tokens Make The World Go Round. In: Proceedings of the Third Workshop on Economics and Natural Language Processing. pp. 13–18. Association for Computational Linguistics, Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.econlp-1.2, https://aclanthology.org/2021.econlp-1.2
15. Noce, L., Zamberletti, A., Gallo, I., Piccoli, G., Rodriguez, J.A.: Automatic Prediction of Future Business Conditions. In: Przepiórkowski, A., Ogrodniczuk, M. (eds.) Advances in Natural Language Processing. pp. 371–383. Springer International Publishing, Cham (2014). https://doi.org/10.1007/978-3-319-10888-9_37
16. Rekabsaz, N., Lupu, M., Baklanov, A., Dür, A., Andersson, L., Hanbury, A.: Volatility Prediction using Financial Disclosures Sentiments with Word Embedding-based IR Models. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp.

1712–1721. Association for Computational Linguistics, Vancouver, Canada (2017). https://doi.org/10.18653/v1/P17-1157, http://aclweb.org/anthology/P17-1157
17. Soldatos, J., Kyriazis, D. (eds.): Big Data and Artificial Intelligence in Digital Finance: Increasing Personalization and Trust in Digital Finance using Big Data and AI. Springer Nature (2022). https://doi.org/10.1007/978-3-030-94590-9, https://library.oapen.org/handle/20.500.12657/54429
18. Tao, J., Deokar, A.V., Deshmukh, A.: Analysing forward-looking statements in initial public offering prospectuses: a text analytics approach. Journal of Business Analytics **1**(1), 54–70 (Jan 2018). https://doi.org/10.1080/2573234X.2018.1507604, https://www.tandfonline.com/doi/full/10.1080/2573234X.2018.1507604
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need (2017). https://doi.org/10.48550/ARXIV.1706.03762, https://arxiv.org/abs/1706.03762
20. Yang, L., Dong, R., Ng, T.L.J., Xu, Y.: Leveraging BERT to Improve the FEARS Index for Stock Forecasting. In: Chen, C.C., Huang, H.H., Takamura, H., Chen, H.H. (eds.) Proceedings of the First Workshop on Financial Technology and Natural Language Processing. pp. 54–60. Macao, China (Aug 2019), https://aclanthology.org/W19-5509