Sentiment Analysis and Finance: Constructing a Zero Investment Portfolio using Transformers and Lexical Analysis.

Student Number: 11061990

Word Count: 9537

This dissertation is submitted as part of MSc Quantitative Finance at The University of Manchester - Manchester Business School

Acknowledgements

I would like to thank the staff at The University of Manchester for their guidance, support and, encouragement throughout the process of writing this dissertation. I would especially like to thank my supervisor Eghbal Rahimikia, whose extensive knowledge, feedback and advice throughout this process has guided this dissertation immensely.

Abstract

News articles have a great impact on stock markets. This paper aims to find methods to predict the market direction of the upcoming weeks. The research is focused on the top 20 companies in the Healthcare sector. The news articles serve as qualitative data, while the weekly returns are used as quantitative data. This study utilizes techniques like Term Frequency-Inverse Document Frequency coupled with the Loughran McDonald Dictionary, to assign a sentiment to text data. The deep learning models employed for predicting future market sentiment are BERT, FinBERT and, RoBERTa. The performance of these models is evaluated using accuracy score, precision, recall and F1 score. RoBERTa displays best accuracy of 53.67% while FinBERT displays 38% which is the worst. Contrastingly, it is observed that FinBERT yields the best portfolio results based on a Sharpe Ratio of 1.29, Sortino Ratio of 2.23, and highest average returns of 22.76%. It was discovered that initial months of 2020 witnessed negative returns followed by a short-lived upturn, before the returns reversed back to a lesser volatile range.

1. Introduction

Data has grown into a pivotal force influencing decisions in the age of information. This information plays a major role in one of the most significant sectors in today's day and age, namely the financial sector. The financial sector isn't merely steered by quantitative data; rather, it's the combination of qualitative and quantitative data that has inspired this dissertation (Tatsat et al., 2020).

In an ever-shifting environment of financial markets, market return predictability is a domain that attracts the attention of not only large businesses but also academics (Bruno Miranda Henrique et al., 2019). The prospect of making new strides in prediction models for future returns is a very intriguing domain. The Efficient Market Hypothesis states that efficient markets reflect information instantaneously making it impossible for investors to consistently generate abnormal returns using historical or public information (Fama, 1970). The ability to make predictions about the stock price movement and achieve returns higher than the market not only disproves this theory that forms the base of finance, but also gives certain investors the capacity to benefit from anomalous returns based purely on market data.

This dissertation explores the range and scope of market predictability within the Healthcare sector, often regarded as the home of innovation and humans' pursuit of good health (Miotto et al., 2017). Seated at the junction of innovation, ethics, and good health, the healthcare industry is a multifaceted territory. Its spectrum ranges from pharmaceutical giants pioneering fresh therapies to biotechnology firms harnessing genetic discoveries for personalized medicine. Healthcare companies, given their dependence on research and cutting-edge innovations, are extremely sensitive to the outcomes of clinical trials, regulatory approvals, newer advancements in the field, etc. In 2019 Biogen a biotech company, ended their clinical trials for an Alzheimer's drug called aducanumab (Knopman et al., 2020). The company's stock price tanked by more than 29%, that week (YunLi626, 2019). This example emphasizes the financial interdependencies of these corporations with the outcomes of their operational achievements as well as the public sentiment created by these endeavors.

Investor decision-making is a crucial driver of market movements, but this decision-making process is influenced by several investor emotions. Risk aversion is a popular concept in finance that talks about how the fear of a potential loss leads investors to make more

conservative choices. In times of ambiguity and downturns, investors may choose to liquidate their positions in the market (Holt & Laury, 2002). Individuals feel the pain of losses more than they enjoy the pleasure of gains, this emotional bias is a driver of investment decisions (Wang et al., 2016).

Sentiment analysis is a process of leveraging computational tools to identify and classify sentiments expressed in a corpus, mainly to discover whether the writer's emotions towards an issue or event are negative, positive, or neutral (Mishev et al., 2020). Sentiment Analysis is a valuable tool to combat the erratic nature of financial markets in healthcare. Market returns are very volatile during times of distress and situations where the public sentiment is easily swayed in either positive or negative direction. Sentiment analysis aims to bridge the gap between human emotions and the financial markets (Sohangir et al., 2018). Deciphering sentiments from news articles can provide insightful information about how emotions and psychological factors cause shifts in the markets. Early detection of shifts in the market based on sentiments before they are fully reflected in the price movements, gives investors the advantage to invest or divest in securities before their peers and monetize on short-term disparities also known as arbitrage opportunities. Interpreting market sentiments can help mitigate risks and develop risk management strategies. Pessimistic sentiments can highlight potential risks associated with the company or security, while positive sentiments stress on the capabilities of the company (Smailović et al., 1970). In 2016, Vertex Pharmaceuticals' cystic fibrosis, Orkambi was approved by the FDA (Vertex Pharmaceuticals Incorporated, 2018). There was an increase of positive sentiments about the company which in turn led to an increase in the stock price.

Human beings are more sensitive to negative information than they are to positive information, due to this reason sentiment analysis in finance aims to focus on deciphering the contextual meanings of more negative words than positive words (Mishev et al., 2020). In recent years there have been significant advancements in the development of tools to extract the sentiment from news surrounding companies. Sentiment analysis is divided into 3 main categories: the sentiment dictionaries-based approach, machine learning models-based approach, and finally deep learning models-based approach (Zhao et al., 2020).

The dictionaries-based approach relies on identifying words that carry information about the article while removing words that add no significance to the sentence; for example: articles in

the English language do not add any significance to the sentiment of a sentence. The most commonly used dictionaries in finance are Harvard IV-4 (HIV4) (Stone et al., 1963) and Loughran and McDonald's (LM) (Loughran & McDonald, 2011). Wordlists developed for other research areas misclassified words in a financial text, and thus Loughran and McDonald assembled a dictionary that identifies the category of financial texts more efficiently (Mishev et al., 2020). Another approach to assign a value to the article is using TF-IDF. It is a statistical method to represent text, it identifies the relevance of a word within a document, within a collection of documents; the document here refers to news articles. This method assigns a numerical value to the text which can be classified as positive negative or neutral based on its polarity (Aizawa, 2003).

There have been significant developments in the use of transformers for Natural Language Processing (NLP), in 2018, Devlin et al. used the transformer architecture to develop a Bidirectional model called BERT (Bidirectional Encoder Representations from Transformers). This model paved the way for further improvements in NLP. FinBERT is a language model based on BERT specifically trained for NLP tasks. This model is pre-trained on 1.8 million financial news articles from the Reuters TRC2 dataset, and it is said to achieve a 15% higher accuracy than the BERT model (Yang et al., 2020). RoBERTa is another model derived from the BERT architecture, it is trained on a dataset ten times bigger with different hyperparameters than BERT. (Liu et al., 2019)

This study aims to utilize these state-of-the-art models to derive new results and add to the existing research. The models are trained on an in-sample period of 2005-2015 and tested on an out-of-sample period of 2016-2022. The results from these models are used to develop a zero-investment portfolio based on different trading strategies and portfolio weightings. The portfolios created based on each model are compared based on their Sharpe ratio, Sortino ratio, mean, standard deviation, the Fama-French 3-factor model, and 5-factor model (Womack & Zhang, 2003) (Foye, 2018).

The concept of creating a zero-investment portfolio based on textual data is innovative and expands the scope of this dissertation. The ability to create portfolios based on sentiments, without the need for additional financial leverage, can revolutionize investment strategies as well as risk management approaches. Through meticulous analysis and implementation of sentiment-driven strategies, this dissertation aims to display that these portfolios can yield

competitive returns. This approach offers a pathway to optimize investment practices. Empirically showcasing the effectiveness of sentiment-based portfolios this study underlines the relevance of integrating finance and deep learning.

This research explores the intricacies of finance, textual news, and deep learning models. The study goes beyond theoretical explorations as it dives into the real-world applications of sentiment analysis, it links speculative findings to practical investment approaches, and assesses the quality of the results and the effectiveness of both the models and the strategies. The findings and conclusions are not definitive but provide a steppingstone for further research within this domain.

2. Literature Review

The following literature review offers a comprehensive investigation of existing academic literature, research studies, and theories pertinent to this research topic. It includes a meticulous analysis of relevant published material to render an in-depth review of the methodologies adopted within the scope of this study.

2.1 Sentiment Analysis in Finance

The prediction of future stock returns sheds light on many theories over the years, one of them being the efficient market hypothesis (EMH) (Fama, 1970). EMH states that the price of a security reflects all the available market information and that all investors have access to this information (Fama, 1970). It believes that markets are perfect and any change in investor sentiment is instantaneously reflected in the stock prices. Malkiel (2003) summarises and evaluates the EMH and critiques it. The article draws comparisons between a random walk and the EMH, where future price changes illustrate random movements from previous prices as prices are said to follow an unpredictable pattern. The research discusses the evidence of short-term momentums and how they exist in the market, deviating from the idea of a random walk. Malkiel (2003) examines the 1987 market crash and the 1990 tech bubble and reports that predictable arbitrage opportunities were non-existent during these events. The conclusion drawn from this research is that markets may not be perfectly random but short-term momentums in the market do not produce profitable arbitrage opportunities. These theories were considered idealistic, and several scholars made breakthroughs from this ideology. Two psychology professors Daniel and Amos in 1979 theorized the concept that psychology and behavior impact stock returns. Tetlock (2007) investigates the effects of media content on the stock market. Tetlock constructs a pessimism factor from Wall Street Journal content using textual analysis. The pessimism factor is found to be predictive of the downward pressure on market prices, which is followed by reversion indicating a sentiment effect. Trading volumes increased after negative sentiments appeared in the Wall Street Journal. Another interesting finding is that the pessimism factor also forecasts high market volume, further confirming the effect of sentiment proxy. “A tale of company fundamentals vs sentiment driven pricing: The case of GameStop” highlights the strength of online communities like Reddit’s WallStreetBets in influencing stock prices based on discussions and sentiments shared online (Umar et al., 2021).

Bollen et al. (2011) explore the correlation between moods derived from daily Twitter feeds and the Dow Jones Industrial Average (DJIA). The authors use two mood-tracking tools namely OpinionFinder and Google Profile of Mood States. The former categorizes moods into negative and positive while the latter measures mood in 6 dimensions. A Self-organizing Fuzzy neural network and Granger causality are used to examine the predictive capacity of moods. The conclusion shows a parallel between the results of the mood-tracking tools and the neural network implying that specific moods can increase the accuracy of DJIA predictions by 87.6%. “Stock price reaction to news and no-news” is a publication that uses a database of news headlines and surveys the stock returns after public news. These stocks are compared to stocks that do not have any public news. The research concludes that stocks associated with public news experience momentum while the stocks without news experience no momentum (Chan, 2003). These examples suggest that markets and news are interdependent entities and thus there is a need to evaluate the public sentiments.

2.2 Sentiment Analysis Using Different Methodologies

“Content analysis stands or falls by its categories” (p. 92) (Berelson, 1952). This statement describes the need for sophisticated methods to segregate different emotions embedded within articles. Words in the English language and in financial context have different connotations, this often leads to misclassification of words in a financial text. The Loughran McDonald (LM) Dictionary (2011) was created specifically to tackle this issue of analyzing financial text and deriving its true contextual meaning. The Harvard-IV-4 TagNeg (H4N) file is used for generic sentiment analysis, but it is not as useful for financial text. It categorizes words such as tax, capital, cost, etc., as negative words while in a financial context, they would mean otherwise.

Words in a sentence derive meaning from the word preceding them. This paved the way for a transition from dictionary-based methods to more sophisticated deep learning models for sentiment analysis.

Early deep learning models introduced for this task were Recurrent Neural Network (RNN) models such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). These are sequence-to-sequence models; this essentially means that the model takes a sequence of inputs and yields another sequence (Young et al., 2018). “Recent Trends in Deep Learning Based Natural Language Processing” discusses the network architecture of these models.

RNNs can store information about previous words in a sentence. They have a hidden state that captures information from previous time steps, this means that they retain contextual information about the previous words which aids in predictions. RNNs suffer from a vanishing gradient problem. Gradients used to update the network begin to vanish after several backpropagations (Young et al., 2018). This problem was addressed by RNN variants – LSTMs and GRUs. LSTMs have specialized memory cells that tackle long-range dependencies and mitigate the vanishing gradient problem. GRUs use the RNN architecture while reducing the long-range dependency issue. Recently, there has been a shift in the deep learning models used for lexical analysis.

"Attention is all you need" introduces the transformer architecture that is based on multi-headed self-attention mechanisms (Vaswani et al., 2017). The authors describe the self-attention as the ability of the model to focus on different parts of the input sentence. The self-attention layer replaces the recurrent layer in RNNs, this improves computational complexity and introduces long-range dependencies between the words in a sentence. A transformer can focus on the context of the sentence while encoding it (Vaswani et al., 2017).

In "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" Devlin et al., (2019) introduce BERT as a state-of-the-art bidirectional model. Unlike other sequential models that interpret context left-to-right or right-to-left, BERT can capture information in both directions simultaneously (Devlin et al., 2019). "BERT for Stock Market Sentiment Analysis" compares BERT to a Naïve Bayes classifier (NB) and a support vector machine (SVM). It demonstrated the outperformance of the bidirectional model over the machine learning (Sousa et al., 2019). Singh et al., (2021) use BERT to predict sentiments of tweets during COVID-19 in "Sentiment analysis on the impact of coronavirus in social life using the BERT model". The paper is used as a baseline to understand how well the model performed during a time of distress and a time that was crucial for the healthcare sector. The model delivered an accuracy of 94% on the validation set suggesting its suitability for this research (Singh et al., 2021).

In "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models", Araci (2019) introduces a new transformer model FinBERT curated specifically for the financial domain. The author implements FinBERT by further pre-training the BERT model on a financial corpus and fine-tuning the model for sentiment classification. TRC2-financial corpus of 29 million words from Reuters is used to pre-train the BERT model. The model is

fine-tuned on labeled financial sentiment datasets; this improves its capacity to capture nuances of financial context. Araci (2019) concludes his research by displaying the outperformance of the FinBERT model over previous machine learning models. “Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers” performs sentiment analysis using dictionary-based approaches, statistical methods, machine learning models, and transformer-based models (Mishev et al., 2020). The authors compare the model performances based on the Mathews Correlation Coefficient (MCC). The MCC score for transformer models is higher indicating their superiority over lexicons and statistical methods (Mishev et al., 2020).

The RoBERTa model introduced in “RoBERTa: A Robustly Optimized BERT Pretraining Approach” is an improved version of the BERT model (Liu et al., 2019). The authors propose key design modifications to the BERT model to address its undertraining issues. The article concludes that by optimizing hyperparameters and the training data in a more efficient way, state-of-the-art models can be achieved using the original BERT architecture. Bozanta et al., (2021) use this model in “Sentiment Analysis of StockTwits Using Transformer Models” to classify tweets from a financial microblog. RoBERTa outperforms the other deep learning models, despite this RoBERTa still misclassifies some tweets. The authors suggest fine-tuning and training the model on more batches to solve this problem (Bozanta et al., 2021).

The final area of research is based on the concept of constructing a portfolio, as it is of crucial importance to this study. The concept of constructing a portfolio based on trading strategies and sentiment analysis has gained traction due to its possible monetary benefits. Chen et al., (2021), use a hybrid machine learning model - eXtreme Gradient Boosting with a firefly algorithm to predict the stock return. The paper uses these stock returns to create a portfolio based on a mean-variance model. The conclusions drawn showed that the portfolio created based on the machine learning model performed better than traditional methods in terms of risks and returns but not in terms of stock price prediction.

Chen et al., (2020) employ machine learning models along with sentiment analysis to create investment strategies in a paper titled “A quantitative investment model based on random forest and sentiment analysis”. The paper incorporates a random forest (RF) model with a sentiment analysis (SA) method to perform quantitative and qualitative analysis. The empirical results show that the portfolio obtained using the RF-SA model generated a higher

return than the Shanghai Composite Index (Chen et al., 2020). “Predicting Returns with Text Data” introduces a novel methodology to predict asset returns using textual data. This paper discusses various trading strategies and portfolio creation methodology that benefits this study (Ke et al., 2021). They use a probabilistic model to link returns via a latent score. The model displays a Sharpe ratio of 4.21 for an equal-weighted portfolio which is much higher than its alternatives. (Ke et al., 2021).

3. Data and Research Design

The data used in this paper surrounds companies in the Healthcare sector. This study aims to employ machine learning and natural language processing to predict market direction. This paper employs a mixed-method research design (George, 2022). The qualitative data comprises headlines and bodies of news articles that are used to predict the direction of the market return. While the quantitative data is the daily returns data transformed to weekly returns. This section discusses the chronological procedure of cleaning the data for the transformer models, to using the model outcomes to create portfolios. The data pre-processing techniques for the in-sample and out-of-sample periods are kept consistent to maintain homogeneity within the data.

3.1 Data Collection

The data is collected from Wharton Research Data Service (WRDS) which is a research platform that provides analysts with financial data obtained from various resources such as companies and financial markets. The Centre for Research in Security Prices (CRSP) is a provider of historical stock data on WRDS. The daily stock returns are extracted from CRSP based on the CUSIP numbers and Tickers. The news data is collected from Key Developments under Compustat – Capital IQ which is another data provider on WRDS. The data is split into 2 periods, namely in the in-sample period and the out-of-sample period. The in-sample period ranges from 1st January 2005 to 31st December 2015, it consists of 10 years of data. The in-sample data is used to train the deep learning models. The transformer models are evaluated and tested on the out-of-sample period which is collected for 6 years, from 1st January 2016 to 31st December 2022. The daily returns and news articles are collected only for companies in the Healthcare sector. The market capitalisations for these companies at the start of the in-sample period and out-sample period are collected. The Loughran McDonald dictionary is used to assign a polarity to each word, to bifurcate positive and negative words. Lastly, the Fama-French daily factors for the 3-factor and 5-factor models were collected from WRDS. The quantitative data contains daily stock returns. The textual data on the other hand consists of key developments data such as the headline of the articles, the body, and the date the article was announced.

3.2 Returns Data Pre-processing

The daily returns are extracted from CRSP. The in-sample period consisted of 289 companies, out of which 171 companies did not trade for the entirety of the in-sample period. These companies were removed from the dataset. Of the remaining 118 companies only 20 are selected. These 20 companies were selected based on their market capitalization in 2005, which is the beginning of the in-sample period. The stock tickers for these companies are PFE, JNJ, AMGN, ABT, MRK, LLY, MDT, UNH, BMY, TGT, BSX, CAH, BAX, SYK, CVS, DHR, GSK, GILD, BDX, TEVA. The companies are kept uniform through the in-sample and out-sample periods. The aggregation of three weeks' worth of returns for each company is used to maintain uniformity with the textual data. This is called the rolling window methodology (Perera, 2016). This approach is employed to address missing values within textual data; for example, a scenario where certain weeks exhibit no discussions or news relating to the company. The absence of information is considered as no change in sentiment throughout the week.

3.3 Textual Data Pre-processing

The textual data being the core part of this analysis is transformed using several meticulous steps.

Cleaning the data

Text-based information is more complex than numerical data due to high dimensionality, contextual ambiguity, repetitive words but most importantly due to its unstructured format. The data may consist of numbers, symbols, punctuations, links, URLs, and brackets, all these characters are not required for analysing the nature of a sentence but they increase the dimensionality of the data and cause a problem while correctly providing a sentiment label as they are not interpretable by deep learning models. The first step is to eliminate the above-mentioned additional characters. This is done using Python's libraries such as NumPy, Pandas, re, nltk, etc. The next step is to eliminate stop words, these are common words used in the English language that hold little to no significance when estimating the sentiment of a statement. Articles and prepositions in the English language are common stop words. The nltk library in Python contains English stop words that can be used to further pre-process the text.

Tokenising Data

Tokenising the data means breaking a larger piece of text into smaller words called tokens, in this way, each word can be analysed. This step is essential as the words need to be broken down to make them readable by the models and it reduces the dimensionality of the input data.

Stemming the data

A simple word can be written in different forms like its adverb or noun form. The word 'sad' is an adjective, its noun form is 'sadness' and its adverb form is 'sadly', they all point to the same meaning, the aim of stemming words is to reduce the dimensionality of the data and to match all these words to the same root, this also reduces redundant words. Thus to analyse every word separately to find its respective root tokenised data is used. This process is done using Python's Porter Stemmer.

Lemmatizing the data

A word can have different inflected forms, for example, 'I have called' and 'I am calling'. Lemmatization aims to group these words so they can be analysed together. This process also uses tokenised data. Lemmatization in Python is done using the WordNetLemmatizer.

3.4 Term Frequency-Inverse Document Frequency (TF-IDF)

Numerical data can easily be interpreted by machine learning models and thus it is easier to work with, but the same cannot be said about textual data. There is a need to transform qualitative data into quantitative values. Term Frequency-Inverse Document Frequency is a natural language processing technique to combat this task. TF-IDF gauges the importance of a word in a document. It comprises two components - term frequency and inverse document frequency (Haddi et al., 2013). Term frequency, as the name suggests measures the importance of a word in a document based on its frequency in the document. While the second component; the inverse document frequency measures the number of documents the particular word appeared in (Haddi et al., 2013). One of the earlier research projects done in TF-IDF was on "Term-weighting approaches in automatic text retrieval" (Salton & Buckley, 1988). "A study of information retrieval weighting schemes for sentiment analysis" uses variants of TF-IDF for sentiment analysis (Paltoglou & Thelwall, 2010). The basic principle

behind TF-IDF is that a word is important if it occurs often in a single document and appears less in other documents. (Chiny et al., 2021)

For this research, each article is considered as an independent document.

$$tf_i = \frac{n_i}{\sum_k n_k}$$

$$idf_i = log(\frac{|D|}{d_j : t_j \epsilon d_j})$$

$$tfidf_i = tf_i * idf_i$$

The first part is the TF term and the second is the IDF term. The TF-IDF weights are calculated by multiplying the TF and IDF values. The words are given a polarity based on the LM dictionary.

3.5 Loughran McDonald Dictionary

The Loughran McDonald (LM) Dictionary contains over 80000 words, it was created by Tim Loughran and Bill McDonald. The words are segregated into negative, positive, uncertainty, litigious, etc. LM dictionary is superior in terms of financial text classifications when compared to the H4N file. It was found that the Harvard file classified 73.8% of the words as negative but in finance, they would not fall into that category (Loughran & McDonald, 2011). The Harvard IV-4 sentiment dictionary (HVD), LM dictionary, and other approaches like Bag of Words (BoW), are used to analyze the sentiments of financial text. Hong Kong Stock Exchange prices and news were used for this research. The sentiment analysis models based on the dictionaries outperform the BoW approach. The LM Dictionary performed better than HVD in the validation and testing sets making it the most efficient approach for this research. For this study each word is assigned a score using TF-IDF, which is multiplied by a polarity score of 1 or -1 based on the LM dictionary, 1 is for positive words and -1 is for negative words. The final statement sentiment is derived from a summation of the scores of the individual words.

3.6 Assigning a Sentiment

The TF-IDF score for each word is combined to get the overall sentiment of the article. A negative score indicates the direction for the week will be negative. Similarly, positive

sentiment shows optimistic conditions for the company, while neutral sentiments display indifference. The portfolio creation only utilises companies with neutral sentiments when the particular week has no negative or positive direction for the remaining companies. The reason behind this methodology is that no stocks will be held in the zero investment portfolio, they will either be short-sold or bought for a long position in the portfolio.

3.7 Transformer Models

The following section introduces the models used in this research.

3.7.1 BERT Model

Bidirectional Encoder Representations from Transformers (BERT) is a machine learning model developed by Google. The Bert model uses a transformer architecture to capture contextual nuances that traditional machine learning models fail to do. The transformer architecture is represented in Figure 3.1. The transformer consists of an encoder and a decoder. The encoder consists of 6 identical layers with 2 sub-layers within each layer. The first sublayer is a multi-head self-attention mechanism that can infer different parts of the sentence. The next is a position-wise fully connected feed-forward neural network. The decoder further consists of the same structure and sub-layers but also has a third multi-head attention layer for the output generated from the encoder. Each layer in the encoder and decoder is followed by layer normalisation (Vaswani et al., 2017). The BERT transformer is said to be bidirectional or non-directional as unlike directional models, the transformer encoder reads the sequence and searches for important parts, then it makes an embedding for each word based on their relevance in the sentence. This enables it to grasp complicated nuances and word dependencies within sentences. The decoder on the other hand takes the output from the encoder and turns it back into a translated version of the text output. Devlin et al., (2019) discuss the pre-training process of the model, it involves 2 unsupervised learning tasks, namely the masked language model (MLM) and next sentence prediction (NSP). Figure 3.2 shows the pre-training and fine-tuning tasks to build the original BERT model.

This study makes use of the pre-trained ‘Bert-base-uncased’ model for sentiment classification. This model contains 12 transformer layers. The uncased aspect in the model indicates that the model vocabulary only consists of lowercase characters, this reduces the

size of the vocabulary. Firstly, the input text is tokenized using the BERT Tokenizer available in the Hugging Face transformer library in Python. These tokens are further assigned an embedding vector that captures the meanings of the sub-tokens and their relationships in the sentence. The model provides positional embedding to highlight the position of each token and extract contextual information. The first token in every sequence is a special token [CLS], it stands for classification. A classification layer is added on top of the [CLS] token output, it includes weights and biases. A pooling mechanism is applied to the aggregate representation of the special token [CLS], it is either a max pooling or mean pooling. It means that the maximum value or the mean value from each dimension is selected. The pooling mechanism is used to highlight the most salient features across the [CLS] output. The aggregated result is sent through a SoftMax function that produces probabilities for each class. The class with the highest probability is chosen as the sentiment for the sentence.

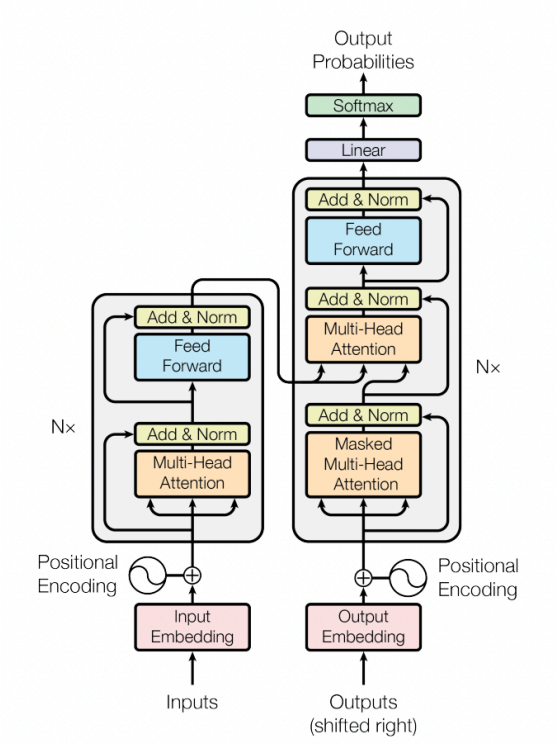

Figure 3.1 Transformer Architecture (Vaswani et al., 2017)

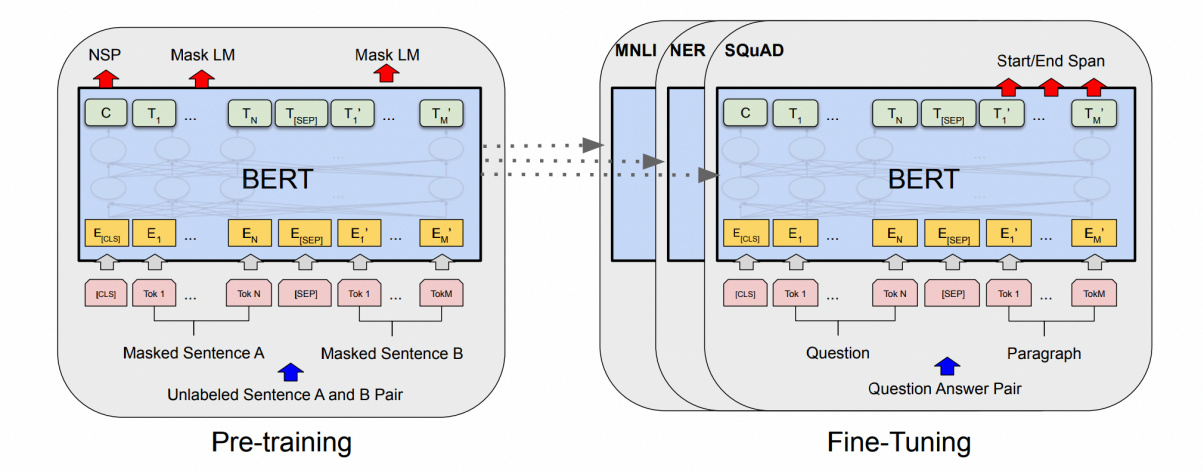

Figure 3.2 The image shows the pre-training and fine-tuning tasks for the BERT model (Devlin et al., 2018).

3.7.2 FinBERT Model

FinBERT is a specialised model developed to specifically analyse financial text. It is adapted from the BERT model and follows a similar architecture. FinBERT is fine-tuned on financial news articles and data, which helps it understand the unique terminologies and nuances of the financial domain (Araci, 2019). It is a more domain-specific model unlike, BERT which is a general-purpose model for sentiment analysis.

3.7.3 RoBERTa Model

RoBERTa is also adapted from the BERT model, it follows a similar architecture but some features have been optimized to create a more efficient model. “RoBERTa: A Robustly Optimized BERT Pretraining Approach” discusses the framework of the RoBERTa model (Lui et al., 2019). The MLM from the original BERT model is modified by duplicating the training data mask the sentence in 10 different ways over 40 epochs. The BERT model was further trained to predict whether the document segments came from the same document or not using NSP loss (Lui et al., 2019). This study uses the “roberta-base” model.

All three models have a base architecture with a few differences that can greatly impact the model accuracies. The models return probabilities of the statement being positive or negative or neutral. Trading is done based on these probabilities.

3.8 Model Accuracy Measures

The accuracies of the transformer models are evaluated using 4 metrics, namely accuracy score, precision, recall, and F1 score. These metrics are implemented using the scikit-learn package in Python. The following formulas are observed in the article "Accuracy, Precision, Recall or F1?" by Shung (2018).

Accuracy score

It computes the sum of the true positive and true negative classifications; true positive and negative classifications refer to the accurate outcomes. This sum is divided by the total number of articles.

$$accuracy = \frac{TP + TN}{number\ of\ articles}$$

$TP = True\ positives$

$TN = True\ negatives$

Precision

Precision is the ratio of true positives to the sum of true positives and false positives. This metric is calculated for each label, in this case, the labels are negative positive, and neutral. The average score from the three labels is taken to give the precision score for a multi-label model.

$$precision = \frac{TP}{TP + FP}$$

$TP = true\ positives$

$FP = false\ Positives$

Recall

Recall measures the ratio of true positives accurately predicted out of the total number of positives. Similar to precision, the average score is derived for a multi-label classification.

$$recall = \frac{TP}{TP + FN}$$

$TP = true\ positives$

$FN = false\ negatives$

F1 score

The F1 score is a harmonic mean of precision and recall. It provides a single score that represents the values of both these metrics. Similar to precision and recall an average of the classes is taken in case of multi-label classification.

$$F1\ score = 2 * \frac{precision * recall}{precision + recall}$$

3.9 Zero Investment Portfolio

A zero-investment portfolio is a collection of assets that require the investor to have no investment in the portfolio. The probabilities generated from the model are ranked in decreasing order. In case of short-selling, the company's with the least ranks are used, while the stocks ranked the highest are utilized for the positive side.

The portfolio is created based on 3 different trading strategies, and two different weighting methods (Ke et al., 2019).

3.9.1 Trading strategies

1. Only buying stocks of companies that show a positive sentiment for the coming week (Long Position)
2. Short-selling stocks of firms that display negative sentiment for week.
3. Finally, short-selling stocks with higher negative scores and using the returns to buy stocks with higher positive scores.

3.9.2 Weighting methods

1. Equal Weighted –

   Each company is assigned the same weight, here weight refers to the importance of the asset in the portfolio. In an equal-weighted portfolio, each asset contributes equally to the portfolio's overall performance. The weight (w_i) of each asset is :

$$w_i = \frac{1}{number\ of\ securities\ in\ the\ portfolio}$$

2. Value Weighted –

It is a type of portfolio where the weights of individual assets are determined based on their market capitalization. Market capitalization is calculated by multiplying the current price of an asset by the total number of outstanding shares. In a value-weighted portfolio, assets with larger market capitalizations have higher weights in the portfolio, while small-cap companies have lesser weights. This method reflects the company's true size as compared to its peers. The weight (w_i) of each asset given is by:

$$w_i = \frac{market\ capitalization\ of\ company\ i}{total\ market\ capitalization}$$

The investment portfolio is created using the top 5 stocks trading each week; either the 5 stocks are bought or/and sold each week.

3.9.3 Portfolio Accuracy Measures

The portfolio performance under each model and trading strategy is compared based on the following metrics.

Average portfolio returns

The mean of the returns of a portfolio. This metric is calculated for weekly returns which are further transformed into annualised returns.

$$annual\ return\ = ((1 + weekly\ return)^{1/n} - 1) * n * 100$$

$n = the\ time\ period, in\ this\ case\ 52$

Portfolio standard deviation

This metric measures how far the returns deviate from the mean return of the portfolio. The standard deviations are stated in annual terms as well. The method to convert them is the following:

$$annual\ deviation = weekly\ standard\ deviation * \sqrt{52}$$

Sharpe Ratio

Measures the risk-adjusted returns of a portfolio. It uses the standard deviation of the portfolio's return. The initial ratios generated need to be annualised using the following:

$$Sharpe\ ratio = \frac{r_p - r_f}{\sigma_p} \text{ (Sharpe, 1998)}$$

$$annual\ Sharpe\ ratio\ = weekly\ Sharpe\ ratio * \sqrt{52}$$

$r_p = portfolio\ return$

$r_f = risk\ free\ rate$

$\sigma_p = portfolio\ standard\ deviation$

Sortino ratio

This metric similar to Sharpe ratio evaluates the risk-adjusted returns of a portfolio but instead of using total risk it uses downside deviation. Like the Sharpe ratio, the Sortino ratio must be annualised using the following:

$$Sortino\ ratio = \frac{r_p - r_f}{\sigma_D} \text{ (Rollinger \& Hoffman, 2013)}$$

$$annual\ Sortino\ ratio\ = weekly\ sortino\ ratio * \sqrt{52}$$

$r_p = portfolio\ return$

$r_f = risk\ free\ rate$

$\sigma_p = the\ standard\ deviation\ of\ downside\ returns$

Fama-French 3-factor model

This model aims to explain stock returns using 3 additional factors. The factors for the model are :

Market risk – the difference between the market return and risk-free rate. (Borchert et al., 2003)

Small minus Big (SMB) – it captures the performance difference between small-cap and large-cap stocks.

High minus Low (HML) – it evaluates the performance difference between value stocks and growth stocks.

$$r_A = r_f + \beta_A(r_M - r_f) + s_A SMB + h_A HML$$

$r_A = portfolio\ return$

$r_f = risk\ free\ rate$

$r_M = market\ return$

$B_A, s_A, h_A = factor\ coefficients$

Fama-French 5-factor model

As the name suggests it is a 5-factor model and an extension of the 3-factor model. It includes two additional factors apart from the ones mentioned above. (Chiah et al., 2016)

Robust minus weak (RMW) – it represents the performance difference between high profitability and low profitability stocks

Conservative minus Aggressive (CMA) – it portrays the difference between conservative and aggressive stocks.

$$r_A = r_f + \beta_A(r_M - r_f) + s_A SMB + h_A HML + w_A SMB + c_A HML$$

$B_A, s_A, h_A, w_A, c_A = factor\ coefficients$

4. Analysis and Discussion

This section presents the key findings derived from an extensive analysis and examines the quality of these discoveries. These results provide valuable insights and implications within the context of this research.

4.1 Analyzing the pre-processed data

The previous section highlights the methodologies employed to preprocess qualitative and quantitative data. As mentioned earlier in-sample dataset consisted of 289 companies, was narrowed down to 20 companies. These were selected based their market capitalizations which is a measure of the company's size.

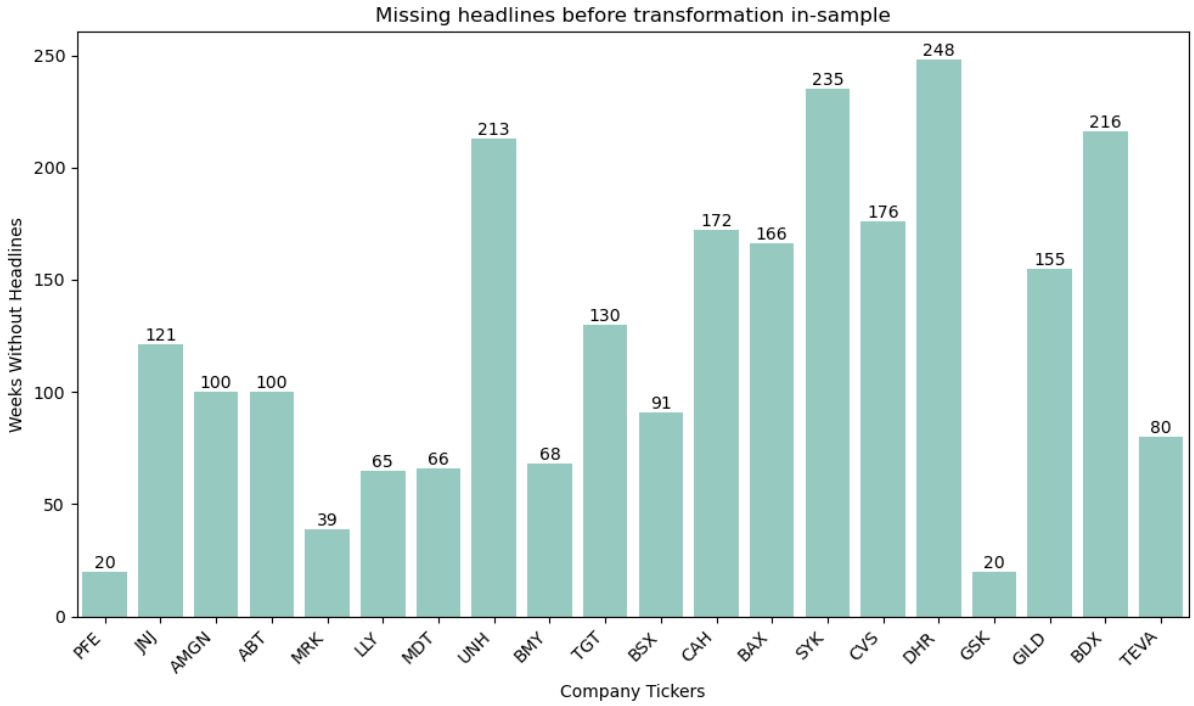

Figure 4.1.1 Missing headlines before transformation across companies for in-sample data

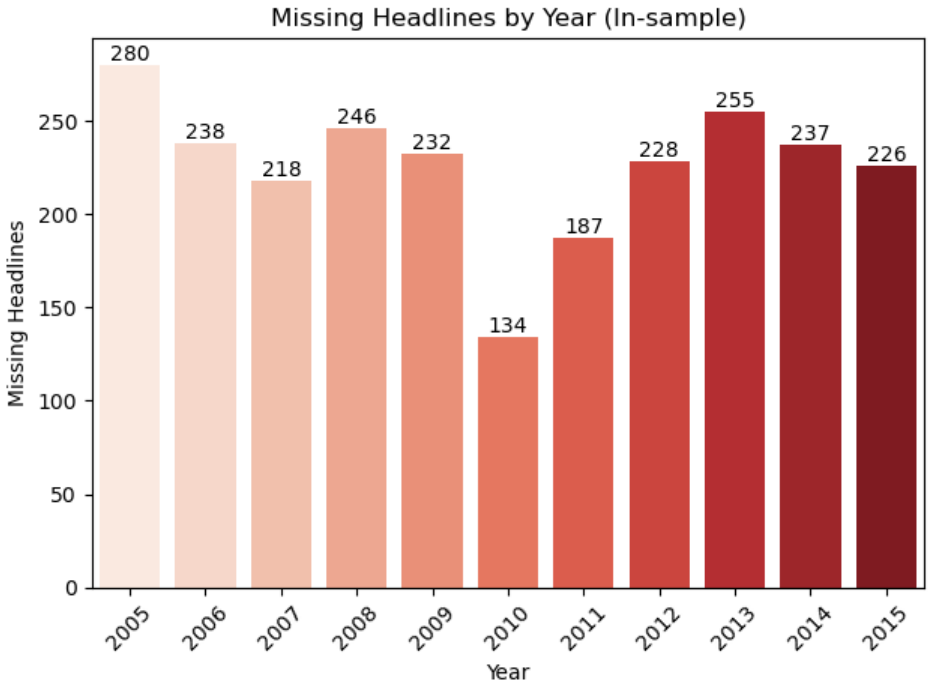

Figure 4.1.2 Missing headlines across the in-sample period over the years.

The above figure shows the spread of missing values across the 20 companies after preprocessing. In the analysis of the in-sample text data a notable observation came to light regarding the absence of headlines for many companies over several weeks. Figure 4.1.1 effectively illustrates the distribution of these instances of missing headlines. It was observed that Danaher Corp, Stryker Corp Becton Dickinson & Co, and UnitedHealth Group Inc. had the greatest number of missing data.

A graphical representation of the distribution of this incomplete data across the 10-year period is elucidated in Figure 4.1.2. The first year in the in-sample data has 280 missing values which is highest for the in-sample period. While the year 2010 has only 134 missing articles. Approximately 21.6% of the dataset had no news articles prompting the need for a different strategy. Missing data can affect model performances; the model may be biased due to misrepresentation of labels. This can also cause underfitting issues as the model does not have enough information.

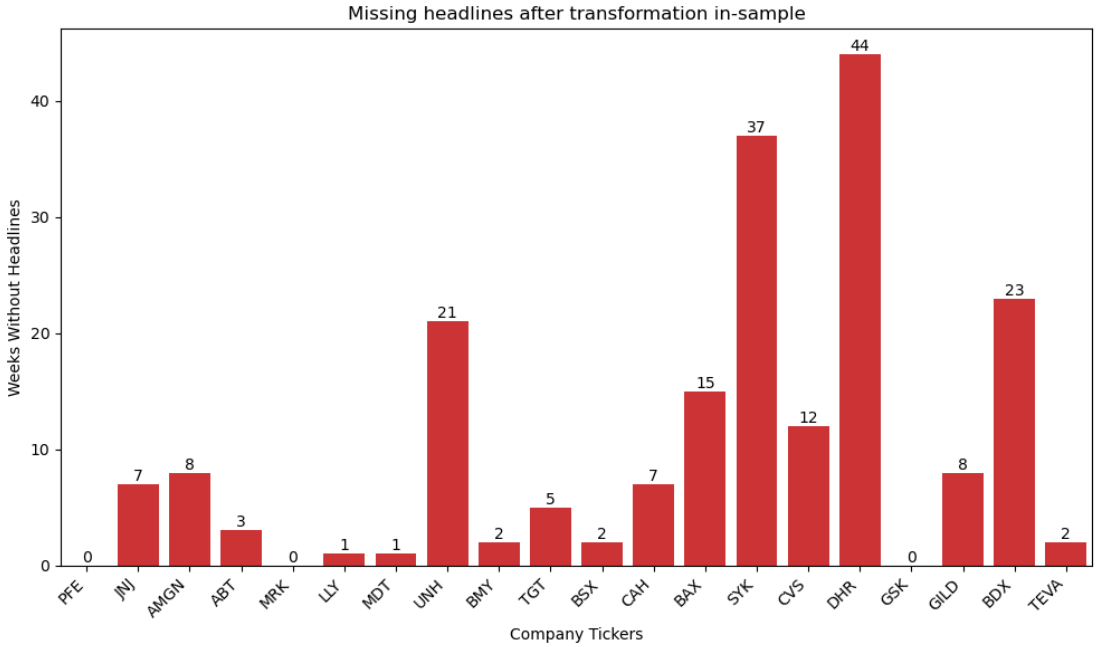

Figure 4.1.3 Missing headlines after transformation across companies for in-sample data.

Consequently, to tackle this issue, a rolling window strategy is applied to transform the values of the from an estimation period of one week to 3 weeks. Figure 4.1.3 illustrates the company-wide distribution of missing data after reshaping the observation period by increasing the number of weeks. The missing textual information reduced to 1.27%. The number of weeks under consideration for the quantitative data are increased as well to maintain uniformity in the dataset.

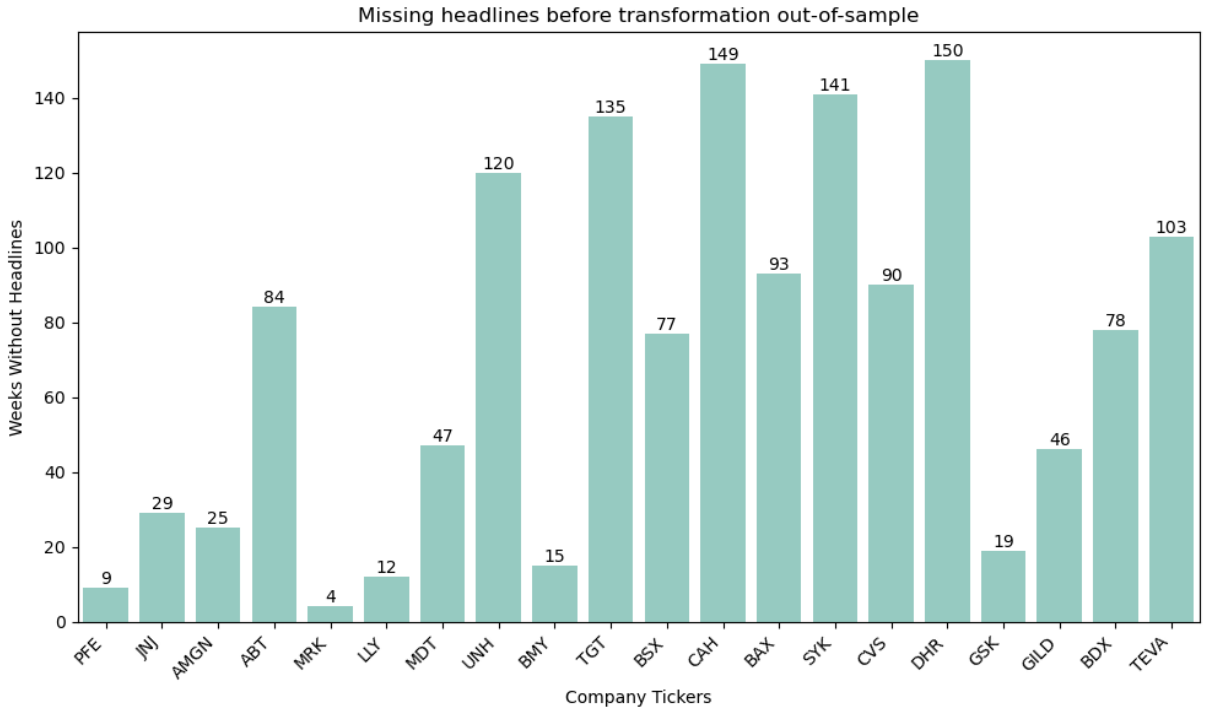

Figure 4.1.4 Missing headlines before transformation for out-of-sample data.

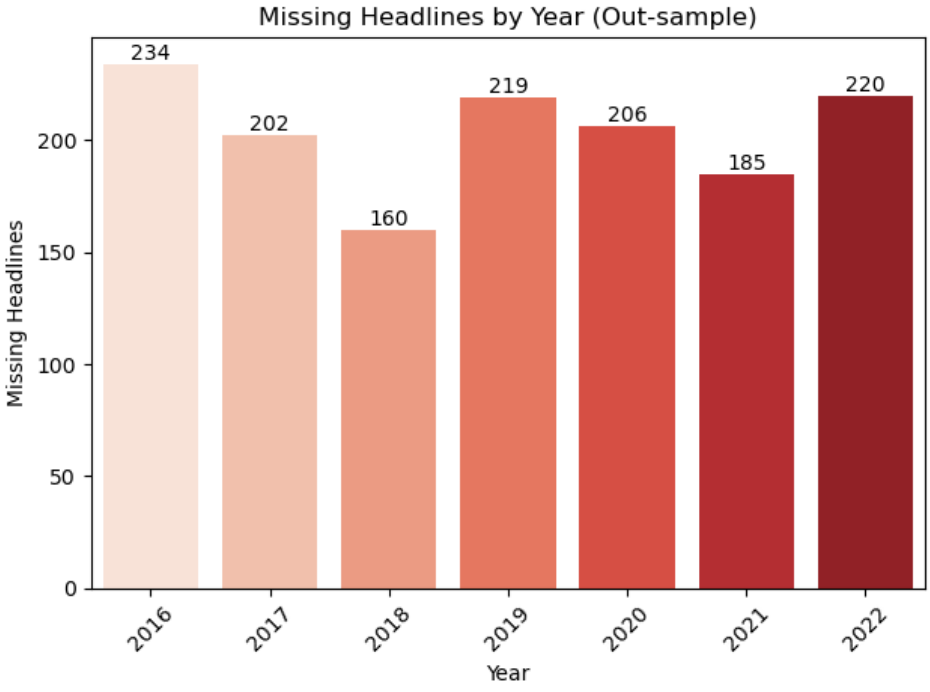

Figure 4.1.5 Missing headlines across the out-of-sample period over the years.

Figure 4.1.4 represents the spread of articles across the out-of-sample period for the 20 companies. Upon examining the out-of-sample period, it is visible that approximately 19.5% of the headlines were missing across the 20 companies in a period of 6 years.
When comparing the two data sets, it can clearly be seen that Danaher Corp has highest missing news across both periods. On the other hand, Pfizer, GSK plc, Merk & Co Inc had relatively low missing values across the same periods.
Furthermore figure 4.1.5 outlines the distribution of missing values across the years. The year 2018 had the lowest missing information, while 2016 had the most. 2019 saw the beginning of the covid-19 pandemic that devastated the world and put severe strain in the healthcare sectors across the globe. In 2019, missing values grew at a very high rate. Following the beginning of the pandemic, missing headlines began to decline year on year as lockdowns were mandated across the world and people relied on news articles as an important source of receiving updates. Once these restrictions were lifted across the globe, in 2022, missing headlines in the industry rose to its highest point since 2016.

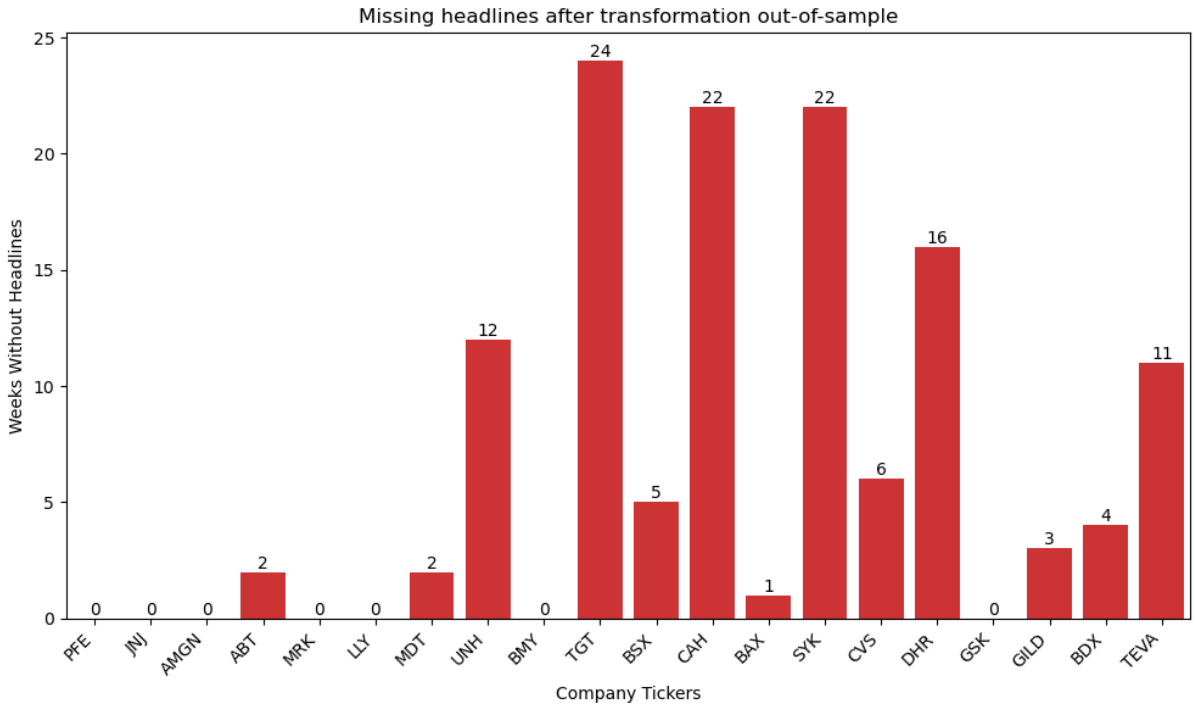

Figure 4.1.6 Missing headlines after transformation across companies for out-of-sample data

The same approach of tackling missing information was applied to the out-of-sample data. The number of weeks for used forecasting is increased to 3 weeks. Figure 4.1.6 shows the reduction in missing articles from 19.5% to 1.78% showing a significant change.

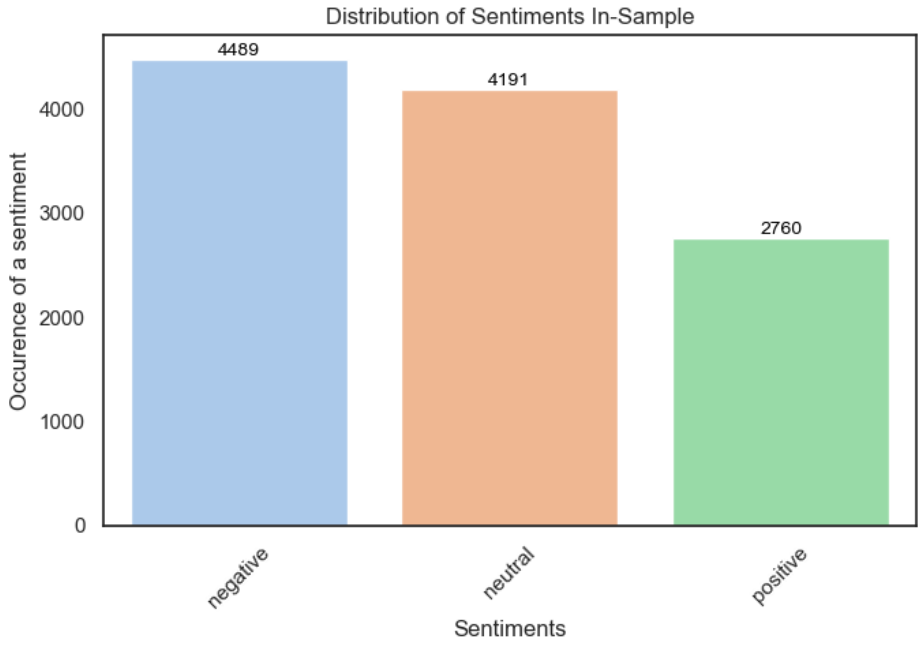

Figure 4.1.7 Sentiment distribution across the in-sample data

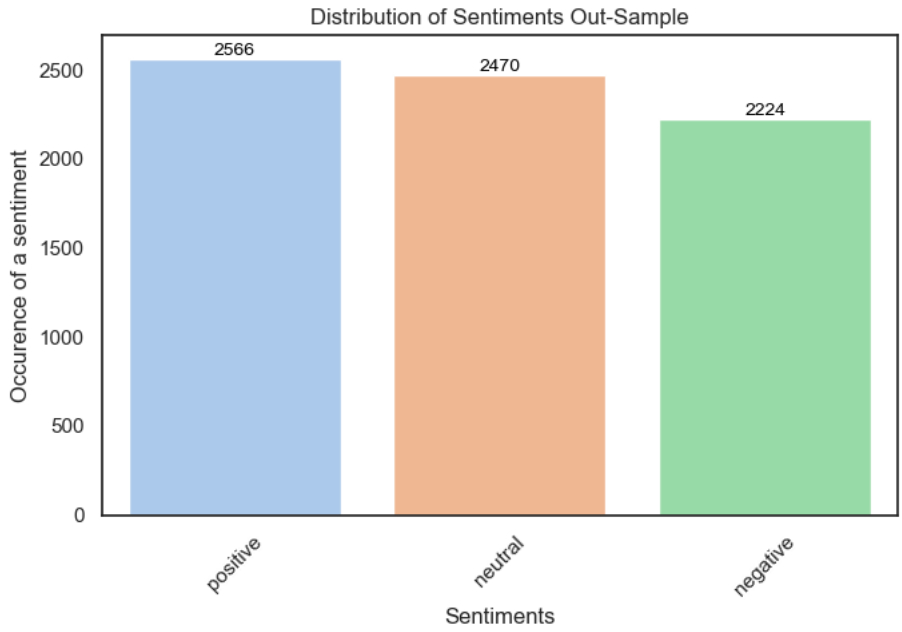

Figure 4.1.8 Sentiment distribution across out-of-sample data

Figure 4.1.7 captures the sentiment distribution in articles for the in-sample period. With negative instances accounting for 39%, neutral occurrences constituting 37%, and positives making up 24% of the dataset.

Conversely, figure 4.1.8, portrays the subsequent out-of-sample period. The two periods provide an intriguing contrast in the sentiment distribution during their respective time frames. There is a more balanced allocation between the sentiments with for this period. Positive sentiments account for 35%, neutral at 34% and negative at 31%. Negative sentiments are more prevalent during the in-sample whereas the out-of-sample period exhibits more positive sentiments.

	companyname	words	counts
0	PFE	[inc, present, confer, announc, annual]	[2345, 533, 500, 463, 304]
1	JNJ	[confer, present, earn, annual, oct]	[295, 292, 207, 175, 134]
2	AMGN	[inc, announc, confer, present, result]	[1101, 319, 291, 270, 256]
3	ABT	[laboratori, earn, announc, confer, result]	[1043, 290, 258, 206, 203]
4	MRK	[inc, co, announc, confer, present]	[1734, 1694, 410, 404, 398]
5	LLY	[eli, co, compani, announc, confer]	[1610, 1038, 472, 395, 338]
6	MDT	[medtron, inc, announc, confer, present]	[1704, 1327, 439, 276, 229]
7	UNH	[group, inc, earn, incorpor, announc]	[746, 497, 276, 213, 174]
8	BMY	[compani, co, announc, confer, present]	[891, 476, 399, 303, 275]
9	TGT	[corp, result, report, earn, sale]	[1060, 342, 308, 270, 247]
10	BSX	[scientif, corpor, announc, earn, quarter]	[1259, 745, 420, 303, 265]
11	CAH	[cardin, inc, earn, confer, announc]	[1000, 909, 228, 212, 159]
12	BAX	[intern, inc, earn, quarter, confer]	[804, 775, 288, 201, 188]
13	SYK	[corp, corpor, earn, confer, quarter]	[416, 258, 238, 200, 150]
14	CVS	[cv, caremark, corpor, earn, corp]	[907, 609, 591, 283, 220]
15	DHR	[corp, earn, quarter, confer, guidanc]	[537, 264, 162, 161, 148]
16	GSK	[plc, buyback, share, announc, equiti]	[2221, 580, 578, 565, 564]
17	GILD	[scienc, inc, confer, present, announc]	[858, 847, 321, 227, 224]
18	BDX	[compani, earn, confer, present, result]	[579, 227, 193, 142, 141]
19	TEVA	[pharmaceut, industri, limit, ltd, announc]	[1505, 1308, 792, 502, 455]

Table 4.1.1 Top recurring words for each company in the in-sample period

Words	Counts
announce	16746
earn	14846
confer	14173
present	12190
result	11674

Table 4.1.2 Top recurring words for the In-sample

Table 4.1.1 presents the primary five recurring words appearing in the articles for each company. These words have been subjected to stemming. The top 5 recurring words from all articles in the in-sample period are displayed in Table 4.1.2 along with the frequency counts for each word. It is noteworthy that, all these words are classified as neutral in the financial context.

	companyname	word	counts
0	PFE	[present, announc, confer, annual, oct]	[587, 484, 426, 376, 257]
1	JNJ	[present, confer, annual, oct, announc]	[407, 393, 231, 186, 172]
2	AMGN	[present, confer, announc, annual, sep]	[386, 316, 316, 233, 170]
3	ABT	[laboratori, announc, earn, confer, result]	[435, 160, 109, 90, 90]
4	MRK	[present, announc, confer, annual, keytruda]	[512, 459, 372, 287, 259]
5	LLY	[eli, compani, announc, present, confer]	[1255, 1197, 444, 399, 299]
6	MDT	[medtron, plc, announc, confer, present]	[988, 754, 279, 212, 197]
7	UNH	[group, incorpor, offer, fixedincom, announc]	[677, 572, 219, 215, 127]
8	BMY	[compani, announc, present, confer, annual]	[1147, 488, 442, 326, 276]
9	TGT	[corpor, announc, earn, quarter, present]	[411, 140, 127, 92, 69]
10	BSX	[scientif, corpor, confer, earn, present]	[660, 583, 208, 180, 152]
11	CAH	[cardin, earn, confer, announc, present]	[515, 113, 109, 98, 89]
12	BAX	[intern, confer, earn, car, announc]	[525, 159, 128, 123, 123]
13	SYK	[corpor, earn, announc, confer, result]	[427, 122, 110, 94, 80]
14	CVS	[cv, corpor, earn, announc, end]	[762, 550, 164, 156, 126]
15	DHR	[corpor, earn, confer, end, quarter]	[385, 128, 103, 95, 93]
16	GSK	[plc, present, confer, gsk, announc]	[924, 429, 360, 290, 256]
17	GILD	[scienc, announc, present, confer, annual]	[923, 303, 243, 192, 150]
18	BDX	[compani, announc, confer, bd, present]	[659, 187, 166, 136, 115]
19	TEVA	[pharmaceut, industri, limit, announc, ltd]	[643, 573, 460, 228, 145]

Table 4.1.3 Top recurring words for each company in the out-of-sample period

Words	Counts
announce	4544
present	4162
confer	3938
company	3139
annual	2851

Table 4.1.4 Top recurring words for the Out-of-sample

A company-wise breakdown for the most preeminent words in the out-of-sample period are showcased in Table 4.1.3. Conversely, table 4.1.4 tabulates the top 5 recurring words in the entire out-of-sample period.

'Announce', 'Present', and 'Confer' were the three words that appeared the most frequently during both the periods. A similarity between the two periods is the presence of neutral words as the most prominent in both periods.

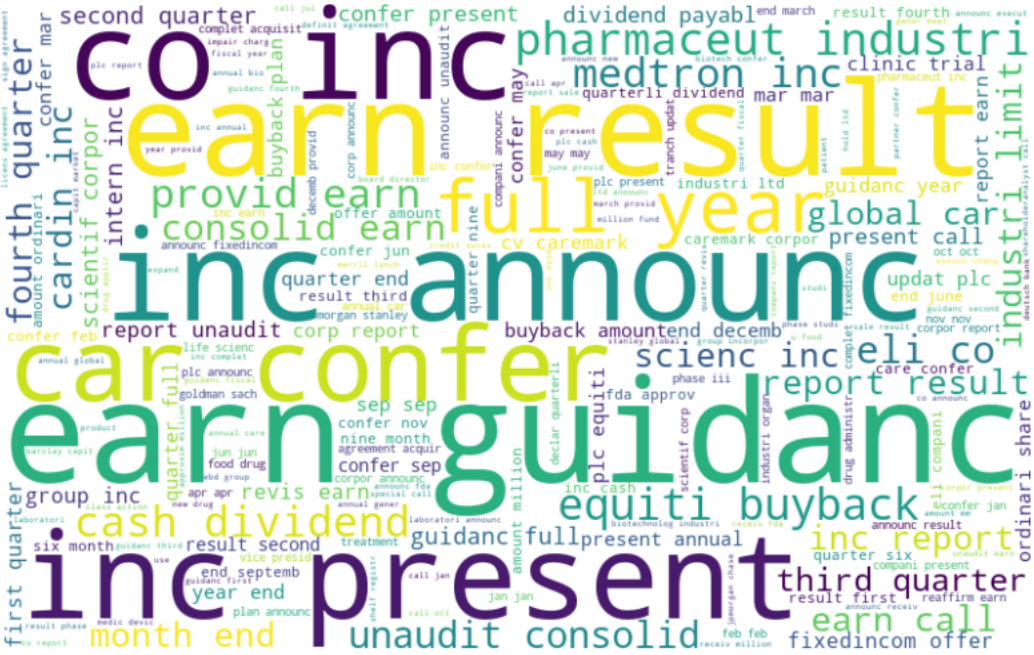

Figure 4.1.9 Word-cloud for the in-sample data

Figure 4.1.10 Word-cloud for Out-of-sample data

A word cloud visually represents the frequency of a word in a piece text by adjusting the size of each word accordingly. In this context, 4.1.9 is a world cloud generated for the in-sample data. Words like ‘equity buy back’, ‘cash dividends’ appear often and can signify either positive or negative company outlooks. Conversely, the presence of the term ‘unaudited’ carries potential negative connotations. The word cloud results for the out-sample period are shown in figure 4.1.10. The words “fixed income” and “offer” appear with a high frequency. Both these words also are contextually neutral. The in-sample and out-of-sample word clouds have a few overlapping words such as confer, guidance, result, present, etc. Apart from these the rest of the cloud looks dissimilar indicating the difference in words utilised during the two periods.

4.2 Model Results

The following section delves into the performance of the models and portfolios. The models are compared based on statistical metrics. For each model six different portfolios are constructed. They are compared based on risk-adjusted measures, returns and the amount of variance in them.

Model	Testing Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
BERT	47	66	47	30.8
FinBERT	38	56.3	33.6	17.3
RoBERTa	53.67	28.8	53.67	37.49

Table 4.2.1 Model Evaluation against accuracy, precision, recall, F1 score

The results of the assessment of the three transformer models namely BERT, FinBERT and RoBERTa are reported in table 4.2.1. The model performances were evaluated through training on the in-sample data and subsequent testing on the out-of-sample data. The evaluation of models was undertaken using four pivotal metrics namely accuracy score, precision, recall, and F1 score.

Accuracy score quantifies the ratio of accurate predictions to the total number of predictions made. It only predicts the number of correctly predicted outcomes by a model, but it does not break the categories down. Among these models, RoBERTa emerged as the frontrunner, with an accuracy score of 53%. BERT demonstrated an accuracy of 47%. On the other hand, the performance of FinBERT was at a mere 38%. A key factor contributing to FinBERT's diminished accuracy was the model's tendency to mislabel news articles as 'neutral' or 'positive'. The RoBERTa model is essentially a Robustly Optimized BERT model, in essence this means that RoBERTa must outperform BERT, which as per the results it did.

Precision measures the percentage of results which are relevant. In contrast Recall measures the percent of total relevant results correctly classified by the model. The final metric F1 score, offers a value that captures the precision and recall of a model.

The evaluation metrics presented in table 4.2.1 show a distinct hierarchy in for the model performances. From the table it is evident that RoBERTa demonstrated the highest accuracy, recall and F1 score, followed by BERT and FinBERT. FinBERT showed a relatively high precision score of 56.3%; however, this is offset by low recall score of 33.6%, indicating that it misclassifies a significant number of articles. All 3 models had relatively low F1 scores, but FinBERT had the worst.

High precision scores indicate that when the model classifies an instance as positive it is usually correct. Precision is a key metric when investors try to exclude models that generate high false positive values. Conversely, high recall scores indicate the model is impressive at finding most of the positive outcomes. This plays an important role in scenarios where missing out on positive instances would be very costly. F1 score is a better indicator is because it handles imbalances between classes better than accuracy score and it combines precision and recall. A risk averse would prefer to have a higher precision over recall as it would mean they avoid false positives and that the model predicts the positives accurately.

	BERT	FinBERT	RoBERTa	LM Dictionary
l_ew	79.04	81.97	76.89	76.87
s_ew	64.59	92.08	92.16	73.68
ls_ew	14.45	-10.13	-15.71	3.19
l_vw	86.38	92.57	83.70	91.84
s_vw	73.01	90.41	92.05	85.28
ls_vw	13.37	2.15	-8.34	6.56

Table 4.2.2 Cumulative returns for all the models.

The outcomes produced by the models serve as base for constructing different types of portfolios. Table 4.2.2 mentions the cumulative returns for each of these models across all the trading strategies. Cumulative returns are the total change in an investment over a specific period of time. A similar trend was observed for the LM dictionary and BERT, FinBERT models under cumulative returns. The three models display the highest cumulative returns under the long value weighted (LVW) portfolio. Contrastingly, for RoBERTa the best returns were found under the short value weighted (SVW) portfolio. The table indicates that the LVW portfolio under the FinBERT model produces the highest cumulative returns. While the long-short equal weighted (L-SEW) portfolio under the RoBERTa model generates the least returns. This metric can be used to track patterns over a long investment horizon.

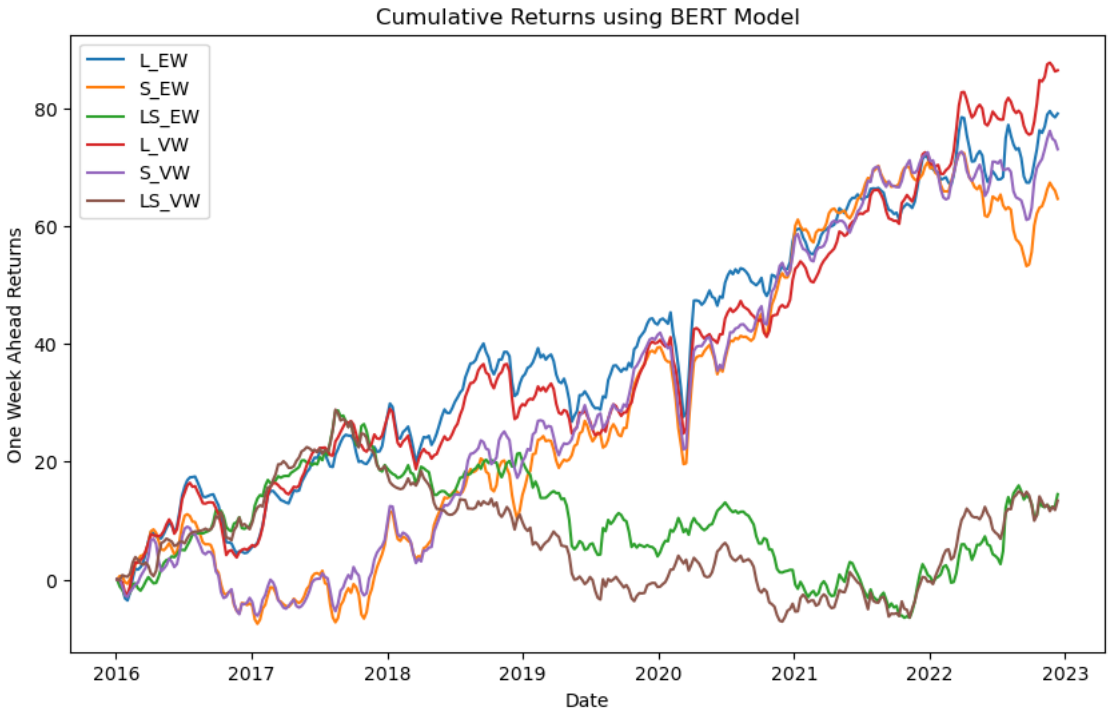

Fig 4.2.1 One week ahead cumulative returns using BERT

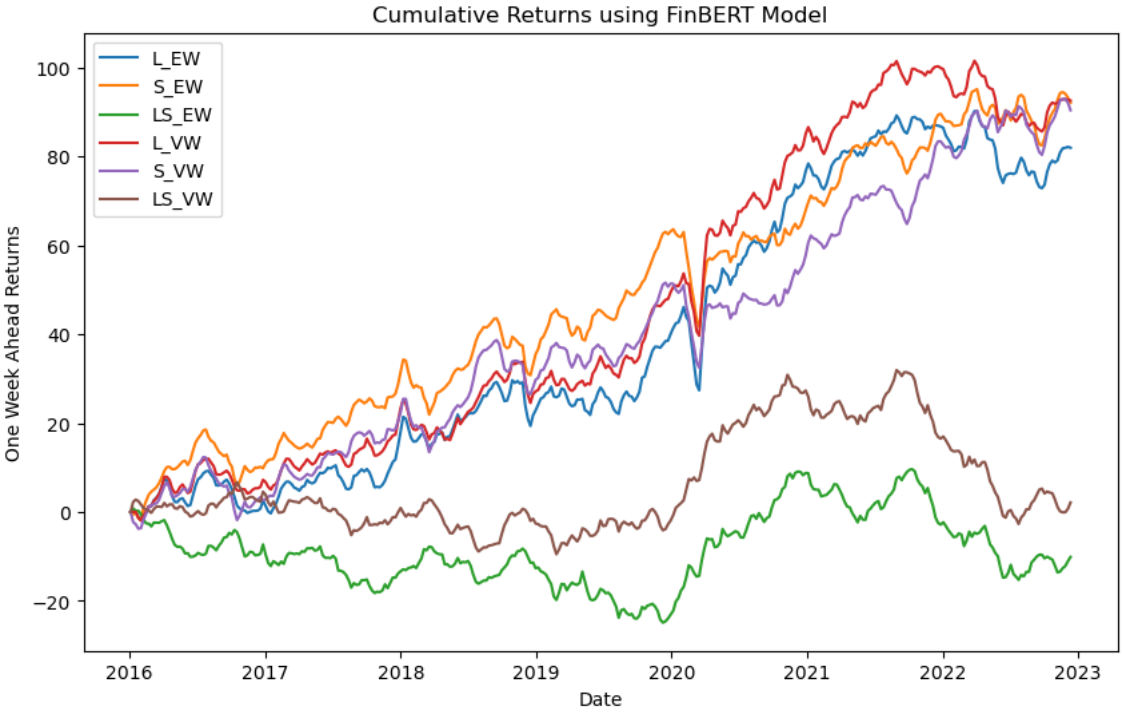

Fig 4.2.2 One week ahead cumulative returns using FinBERT

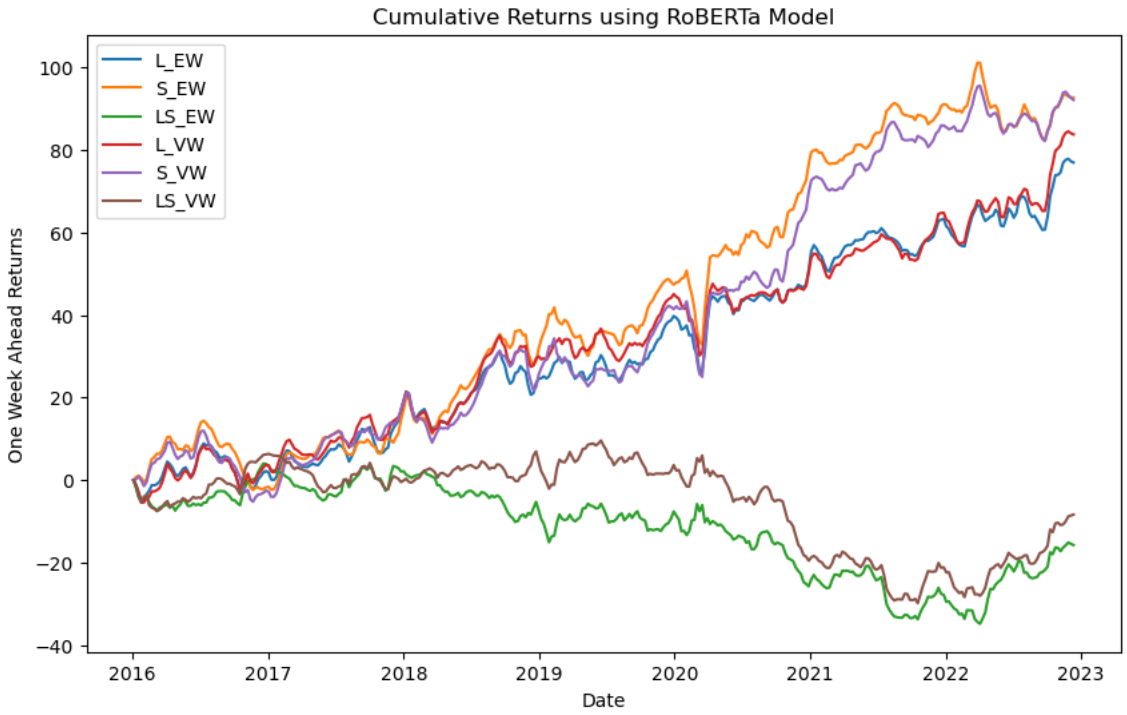

Fig 4.2.3 One week ahead cumulative returns using RoBERTa

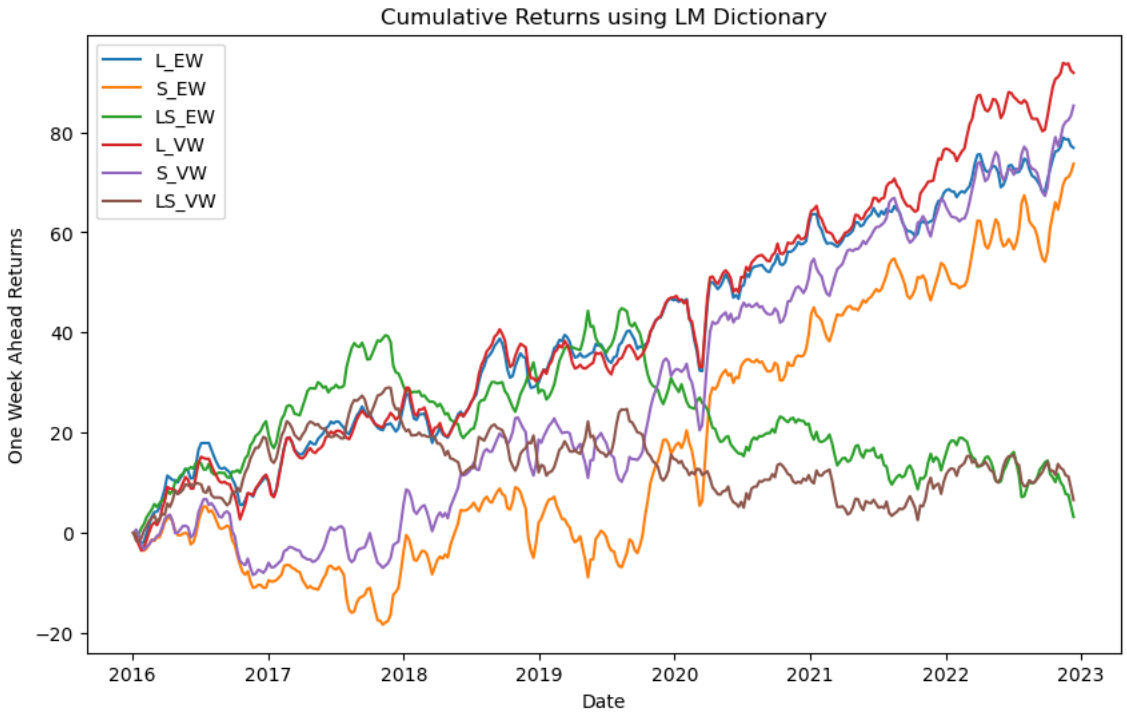

Fig 4.2.4 One week ahead cumulative returns using LM Dictionary

Figures 4.2.1, 4.2.2, 4.2.3 and 4.2.4 visually illustrate the cumulative returns for each portfolio under the BERT, FinBERT, RoBERTa and, LM Dictionary respectively.

The portfolio performances for each model are tabulated below. For the BERT model the returns generated before 2019 are negative, but the onset of covid-19 leads to sharp increases in the returns of short portfolios. A similar trend was observed in case of the LM model as well. This is an interesting observation as it highlights abnormal short-selling activities during the initial months of the pandemic (Luu et al, 2023). Investors began short-selling overpriced stocks during the begin of covid-19. Indicating that there were more overpriced stocks in these two models than FinBERT.

	Average(%)	Std	Sharpe ratio	Sortino ratio	ff3 alpha	ff3 r2 (%)	ff5 alpha	ff5 r2 (%)
l_ew	19.74	10.56	1.03	1.62	0.17	13.67	0.17	13.71
s_ew	16.40	10.38	0.85	1.32	0.13	14.78	0.13	15.06
ls_ew	3.90	8.20	0.20	0.35	0.04	0.30	0.04	0.35
l_vw	21.39	9.66	1.24	1.91	0.20	11.14	0.20	11.18
s_vw	18.36	10.43	0.96	1.49	0.16	12.35	0.16	12.44
ls_vw	3.62	7.67	0.20	0.33	0.04	0.40	0.04	0.40

Table 4.2.3 BERT portfolio results. Expected returns, Sharpe ratio, standard deviations and Sortino ratio are annualized.

	Average(%)	Std	Sharpe ratio	Sortino ratio	ff3 alpha	ff3 r2 (%)	ff5 alpha	ff5 r2 (%)
l_ew	20.40	10.92	1.04	1.63	0.03	56.53	0.03	56.37
s_ew	22.66	9.88	1.29	1.82	0.07	56.14	0.07	55.99
ls_ew	-2.83	7.89	-0.24	-0.37	-0.04	1.34	-0.04	1.34
l_vw	22.76	10.24	1.25	2.20	0.08	48.02	0.08	47.85
s_vw	22.29	9.85	1.27	1.91	0.08	46.80	0.08	46.68
ls_vw	0.59	7.85	-0.01	-0.02	0.00	1.00	0.00	1.07

Table 4.2.4 FinBERT portfolio results. Expected returns, standard deviations, Sharpe ratio and Sortino ratio are annualized.

	Average(%)	Std	Sharpe ratio	Sortino ratio	ff3 alpha	ff3 r2 (%)	ff5 alpha	ff5 r2 (%)
l_ew	19.25	10.02	1.06	1.76	0.17	11.01	0.17	11.05
s_ew	22.77	10.72	1.20	1.91	0.20	17.41	0.20	17.39
ls_ew	-4.43	7.91	-0.34	-0.54	-0.03	2.44	-0.03	2.40
l_vw	20.79	9.79	1.18	1.92	0.20	9.31	0.20	9.34
s_vw	22.65	10.51	1.25	2.14	0.20	15.86	0.20	15.81
ls_vw	-2.32	7.89	-0.20	-0.32	-0.01	3.55	-0.01	3.51

Table 4.2.5 RoBERTa portfolio results. Expected returns, standard deviations, Sharpe ratio and Sortino ratio are annualized.

	Average(%)	Std	Sharpe ratio	Sortino ratio	ff3 alpha	ff3 r2 (%)	ff5 alpha	ff5 r2 (%)
l_ew	19.24	9.50	1.12	1.85	0.21	2.65	0.17	13.41
s_ew	18.51	12.27	0.83	1.45	0.20	1.39	0.15	15.57
ls_ew	0.87	9.25	0.00	0.01	0.01	0.19	0.02	3.26
l_vw	22.60	9.86	1.26	2.12	0.25	2.34	0.21	10.26
s_vw	21.14	11.33	1.04	1.80	0.23	1.30	0.19	13.43
ls_vw	1.79	8.94	0.06	0.09	0.02	0.62	0.03	2.59

Table 4.2.6 LM portfolio results. Expected returns, standard deviations, Sharpe ratio and Sortino ratio are annualized.

Tables 4.2.3, 4.2.4, 4.2.5, and 4.2.6 provides a cross sectional view of the portfolio performances of the BERT, FinBERT, RoBERTa, and LM dictionary respectively. The average returns, standard deviation, Sharpe ratio and, Sortino ratio are presented in annualized terms.

The expected value of a portfolio provides the investor with an estimate of what the overall profit or loss can resemble. The expected returns exhibit a trend similar to the cumulative returns. The RoBERTa model shows the highest returns for a SVW portfolio, while the remaining models depict higher returns for a LVW portfolio.
The FinBERT model shows the highest average return of 22.76% for a LVW portfolio. RoBERTa's SVW portfolio comes after, followed by LM and then the BERT model.

Standard deviation of returns is a measure of how far variables move compared to the mean of the portfolio. Analysts use this measure to inspect the consistency of returns. The L-SVW portfolios under all four models have the least standard deviation. This is observed because long-short strategies aim to minimize market risk, by hedging returns. This strategy exploits downside and upside movements in the market.

Sharpe ratios measure the portfolio's risk-adjusted performance (Sharpe, 1998). Higher Sharpe ratios indicate better risk-adjusted performance. The short equal weighted (SEW) portfolio under the FinBERT model has the best Sharpe ratio of 1.29. LM's highest ratio of 1.26 was found under the LVW portfolio. RoBERTa's SVW portfolio shows a slightly lower value of 1.25, followed by BERT's LVW portfolio with a value of 1.24. Sharpe ratios above 1 indicate good portfolio performance as they generate excess returns relative to volatility.

The differentiating factor between the Sharpe and Sortino ratio is their volatilities (Rollinger & Hoffman, 2013). FinBERT, BERT and LM have highest scores of 2.2, 1.91, 2.12 respectively for the LVW portfolio. While RoBERTa deviates from this trend with a score of 2.14 for the SVW portfolio. Sortino ratios above 2 are considered good investments, while a score below 2 is regarded as poor. The Sortino ratio only considers downside volatility and is more relevant from a risk-averse investor's perspective as they are more concerned with losses (Holt et al., 2002).

The Fama-French 3 factor and 5 factor r^2 values represent the amount of variance in the portfolio that can be explained by the respective factors of the model. Higher r^2 ratios indicate a better explanation of the returns. The SEW portfolio for the FinBERT model shows the greatest r^2 values of 56.53% for the 3-factor model and 56.37 % for the 5-factor model.

	Average(%)	Std	Sharpe ratio	Sortino ratio	ff3 alpha	ff3 r2 (%)	ff5 alpha	ff5 r2 (%)
FinBERT l_vw	22.76	10.24	1.25	2.20	0.08	48.02	0.08	47.85
RoBERTa s_vw	22.65	10.51	1.25	2.14	0.20	15.86	0.20	15.81
LM l_vw	22.60	9.86	1.26	2.12	0.25	2.34	0.21	10.26
BERT l_vw	21.39	9.66	1.24	1.91	0.20	11.14	0.20	11.18

Table 4.2.7 Comparing the top portfolios from each model.

Table 4.2.7 enumerates the top portfolios from each model. The FinBERT model shows high average returns, along with Sharpe and Sortino ratios. Finally, the Fama-French factors indicate highest explained variance suggesting that its overall performance was the best. An interesting fact to highlight here is that, despite the performance of the FinBERT model being the least impressive it shows the best results in terms of portfolio performance.

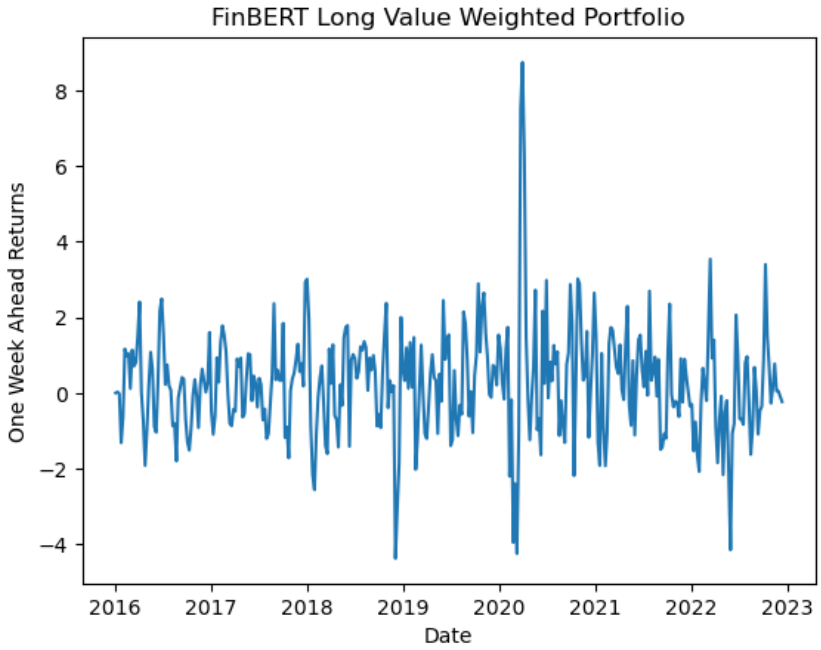

Figure 4.2.5 Weekly returns of the FinBERT long Value weighted portfolio

Figure 4.2.5 shows the weekly returns for the FinBERT model. The initial months of 2020, the beginning of the covid-19 pandemic saw abnormally high and low returns. This pattern

was reversed during mid 2020. The low returns were a cause of high short-selling activities. While the high returns can be linked to decreased positions in short-selling activities after the first few months of covid-19.

5. Limitations

The transformer models utilized requires refinements as these did not meet the optimal standard of prediction. This being said, model performance is impacted by numerous factors one of the key reasons for lower model performance can be the disparities between the in-sample data and the out-of-sample data. The in-sample data contained more negative sentiments while the out-of-sample data contained more positive sentiments. The distribution of sentiments in the out-of-sample period was more balanced than the in-sample period.

Further improvements can be made using fine-tuning techniques. Optimizing hyperparameters can increase the classification capacity of the models. Although a limitation to this is the computationally intensive nature of the process. Notably fine-tuning models may not always generate favorable results as they can introduce overfitting biases.

It is crucial to observe that strategies such as the long-short positions did not present impressive results, suggesting the need for different approaches to improve their performances. Different trading strategies can be employed. Employing a contrarian strategy, where the investor selects stocks with sentiment scores that are contrary to market trends. In situations where the stock shows promising returns but the sentiment is negative the investor must buy the stock. An alternative to this can be building a portfolio based on specific events like clinical trial results, FDA approvals or earnings releases. The investor must gauge the sentiment changes before and after these events and determine the weight of each security in the portfolio. Furthermore, several asset allocation methods can be used. Mean Variance Optimization technique aims to maximise the mean while ensuring the variance is minimized. In addition, Monte Carlo simulations are another efficient tool to optimize portfolio weights. This algorithm generates various asset allocations and picks the one that meets the investor's criteria. Lastly, the number of stocks purchased and sold can be varied to track the changes in the portfolio performance.

Investors can only make informed investment choices if they have metrics that provide them with insights about a portfolios' financial goals. Thus, it is important to pick the correct metrics. A few of the metrics mentioned do not always serve as optimal comparison measures. The average of portfolio returns is heavily impacted by highly positive or negative returns. Mean as an indicator is not reliable as it can be skewed to either the positive or negative side based on the outliers. This can blindside the investor by estimating incorrect

future returns. The standard deviation of returns focuses on the total dispersion from the mean, but investors are mainly concerned with the negative deviations from their expected returns. This is why standard deviation is a generic measure and does not always provide insightful information. The r^2 values only display the variance that can be explained by the factors used. It is not an indicator of whether the portfolio actually performs well and thus this metric should be used combination with other metrics.

The Sharpe and Sortino ratios give different risk adjusted measures of the portfolio. The Sharpe ratio uses the portfolios standard deviation in its denominator, assuming that returns are normally distributed, but this is not always true. In case of extreme events like covid-19 add volatility to the market and are non-normal events. For non-normal distributions standard deviation may not capture tail risks. This can drive poor investment decisions as investors may view a stock as less volatile than it truly is.

The Sortino ratio only considers downside volatility. Investors concerned with capital growth or upside volatilities cannot use this indicator. It also does not account for diversification within in the portfolio. Thus, none of these metrics are perfect and performance decisions need to be based on all of these metrics combined.

6. Conclusion

In conclusion, this study presents the intricate relationship between public emotions, financial markets, stock directions, and the predictive capacity of transformer models. Although, many researchers have established the interdependence between public sentiments and market movements this research utilizes a mixed-method research design that incorporates dictionary-based and statistical like TF-IDF while leveraging transformer models like FinBERT, BERT, and RoBERTa.

The outcomes of this investigation based on the predictive power of transformer models, while noteworthy, do not demonstrate high statistical significance. The RoBERTa model emerges as the top performer with an accuracy of 53.67% and an F1 score of 37.49%. In contrast, the FinBERT model trained specifically on financial corpus misclassifies a significant proportion of data as "neutral". While the BERT displayed an accuracy of 47% with high a precision score of 66%. Precision scores are more essential to risk-averse investors as higher scores suggest lower false trading signals.

An innovative strategy presented in this study is the concept of building a zero-investment portfolio using transformer models. The results are demonstrated using a cross-sectional analysis by segregating them based on the model and trading strategy. Despite its shortcomings in the overall model performance, the FinBERT model distinguishes itself with an impressive performance for the LVW portfolio. With a Sharpe ratio of 1.29 and a Sortino Ratio of 2.23.

The analysis covers an intriguing discovery observed in the weekly returns for FinBERT's LVW during the COVID-19 pandemic in 2020. During this timeframe, the volatility of the stock market is displayed. With the initial months of 2020 yielding negative returns, followed by a sharp reversal to positive returns. These observations underscore the link between widespread emotions of fear and distress at the onset of the pandemic and their influence on stock returns in the healthcare sector. An initial surge in short-selling activities by investors was noticed; they sold securities they believed were overpriced. This trend was observed to be short-lived. Subsequent reductions in short positions lead to a surge in prices driving the returns back up for a limited period. These findings align with the results appearing in "Short-selling activities in the time of COVID-19" by Luu et al., (2023).

The paradoxical findings about the FinBERT model highlight the complex interplay between sentiment prediction models and portfolio performances, suggesting the need for additional research and investigation in this dynamic domain. This paper contributes to the existing literature by demonstrating the potential and the limitations of using transformer models for sentiment classification. Nevertheless, this paper discusses portfolio creation methodologies and provides evidence that a zero-investment portfolio can be created to generate higher returns than the market. Further improvements to this procedure may contribute to refining textual analysis and portfolio creation.

Citations

Aizawa, A., 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, *39*(1), pp.45-65.

Araci, D. (2019) Finbert: Financial sentiment analysis with pre-trained language models, arXiv.org. Available at: https://arxiv.org/abs/1908.10063 (Accessed: 30 August 2023).

Berelson, Bernard R., 1952,Content Analysis in Communication Research(The Free Press, Glen-coe, IL)

Bollen, J., Mao, H. and Zeng, X., 2011. Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), pp.1-8.

Bozanta, A., Angco, S., Cevik, M. and Basar, A., 2021, December. Sentiment analysis of stocktwits using transformer models. In *2021 20th IEEE international conference on machine learning and applications (ICMLA)* (pp. 1253-1258). IEEE.

Chan, W.S., 2003. Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of financial economics*, *70*(2), pp.223-260.

Chen, M., Zhang, Z., Shen, J., Deng, Z., He, J. and Huang, S., 2020, June. A quantitative investment model based on random forest and sentiment analysis. In Journal of Physics: Conference Series (Vol. 1575, No. 1, p. 012083). IOP Publishing.

Chen, W., Zhang, H., Mehlawat, M.K. and Jia, L., 2021. Mean–variance portfolio optimization using machine learning-based stock price prediction. *Applied Soft Computing*, *100*, p.106943.

Chiah, M., Chai, D., Zhong, A. and Li, S., 2016. A Better Model? An empirical investigation of the Fama–French five-factor model in Australia. *International Review of Finance*, *16*(4), pp.595-638.

Chiny, M., Chihab, M., Bencharef, O. and Chihab, Y., 2021. LSTM, VADER and TF-IDF based hybrid sentiment analysis model. *International Journal of Advanced Computer Science and Applications*, *12*(7).

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fama, E.F., 1970. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, *25*(2), pp.383-417.

FDA approves ORKAMBI® (lumacaftor/ivacaftor) as first medicine to treat the underlying cause of cystic fibrosis for children ages 2-5 years with most common form of the disease (2018) *Vertex Pharmaceuticals*. Available at: https://investors.vrtx.com/news-releases/news-release-details/fda-approves-orkambir-lumacaftorivacaftor-first-medicine-treat

Foye, J., 2018. A comprehensive test of the Fama-French five-factor model in emerging markets. *Emerging Markets Review*, *37*, pp.199-222.

George, T. (2022) Mixed methods research: Definition, guide, & examples, Scribbr. Available at: https://www.scribbr.co.uk/research-methods/mixed-methods/

Haddi, E., Liu, X. and Shi, Y., 2013. The role of text pre-processing in sentiment analysis. *Procedia computer science*, *17*, pp.26-32.

Henrique, B.M., Sobreiro, V.A. and Kimura, H., 2019. Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, *124*, pp.226-251.

Holt, C.A. and Laury, S.K., 2002. Risk aversion and incentive effects. *American economic review*, *92*(5), pp.1644-1655.

Kahneman, D. and Tversky, A. (1979) Prospect theory: An analysis of decision under risk.

Ke, Z.T., Kelly, B.T. and Xiu, D., 2019. *Predicting returns with text data* (No. w26186). National Bureau of Economic Research.

Knopman, D.S., Jones, D.T. and Greicius, M.D., 2021. Failure to demonstrate efficacy of aducanumab: An analysis of the EMERGE and ENGAGE trials as reported by Biogen, December 2019. *Alzheimer's & Dementia*, *17*(4), pp.696-701.

Li, X., Xie, H., Chen, L., Wang, J. and Deng, X., 2014. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, *69*, pp.14-23.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loughran, T. and McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance*, *66*(1), pp.35-65.

Luu, E., Xu, F. and Zheng, L., 2023. Short-selling activities in the time of COVID-19. *The British Accounting Review*, p.101216.

Malkiel, B.G. (2003) *A random walk down wall street*. New York: W.W. Norton & Company.

Mao, H., Counts, S. and Bollen, J., 2011. Predicting financial markets: Comparing survey, news, twitter and search engine data. *arXiv preprint arXiv:1112.1051*.

Miotto, R., Wang, F., Wang, S., Jiang, X. and Dudley, J.T., 2018. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, *19*(6), pp.1236-1246.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L.T. and Trajanov, D., 2020. Evaluation of sentiment analysis in finance: from lexicons to transformers. *IEEE access*, *8*, pp.131662-131682.

Paltoglou, G. and Thelwall, M., 2010, July. A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1386-1395).

Perera, S. (2016) *Rolling window regression: A simple approach for time series next value predictions*, *Medium*. Available at: https://medium.com/making-sense-of-data/time-series-next-value-prediction-using-regression-over-a-rolling-window-228f0acae363.

Rochester, NY: University of Rochester, Graduate School of Management, Managerial Economics Research Center.

Rollinger, T.N. and Hoffman, S.T., 2013. Sortino: a 'sharper'ratio. *Chicago, Illinois: Red Rock Capital*.

Salton, G. and Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), pp.513-523.

Sharpe, W.F., 1998. The sharpe ratio. *Streetwise–the Best of the Journal of Portfolio Management*, *3*, pp.169-85.

Shung, K.P. (2018) Accuracy, precision, recall or F1?, Medium. Available at: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9.

Singh, M., Jakhar, A.K. and Pandey, S., 2021. Sentiment analysis on the impact of coronavirus in social life using the BERT model. *Social Network Analysis and Mining*, *11*(1), p.33.

Smailović, J., Grčar, M., Lavrač, N. and Žnidaršič, M., 2013. Predictive sentiment analysis of tweets: A stock market application. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data: Third International Workshop, HCI-KDD 2013, Held at SouthCHI 2013, Maribor, Slovenia, July 1-3, 2013. Proceedings* (pp. 77-88). Springer Berlin Heidelberg.

Sohangir, S., Wang, D., Pomeranets, A. and Khoshgoftaar, T.M., 2018. Big Data: Deep Learning for financial sentiment analysis. *Journal of Big Data*, *5*(1), pp.1-25.

Sousa, M.G., Sakiyama, K., de Souza Rodrigues, L., Moraes, P.H., Fernandes, E.R. and Matsubara, E.T., 2019, November. BERT for stock market sentiment analysis. In *2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI)* (pp. 1597-1601). IEEE.

Stone, P.J. and Hunt, E.B., 1963, May. A computer approach to content analysis: studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, spring joint computer conference* (pp. 241-256).

Tatsat, H., Puri, S. and Lookabaugh, B., 2020. *Machine Learning and Data Science Blueprints for Finance*. O'Reilly Media.

Tetlock, P.C., 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, *62*(3), pp.1139-1168.

Umar, Z., Gubareva, M., Yousaf, I. and Ali, S., 2021. A tale of company fundamentals vs sentiment driven pricing: The case of GameStop. *Journal of Behavioral and Experimental Finance*, *30*, p.100501.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, M., Rieger, M.O. and Hens, T. (2016) 'The impact of culture on loss aversion', *Journal of Behavioral Decision Making*, 30(2), pp. 270–281. doi:10.1002/bdm.1941.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. and Davison, J., 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Womack, K.L. and Zhang, Y., 2003. Understanding risk and return, the CAPM, and the Fama-French three-factor model. *Available at SSRN 481881*.

Yang, Y., Uy, M.C.S. and Huang, A., 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Young, T., Hazarika, D., Poria, S. and Cambria, E., 2018. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, *13*(3), pp.55-75.

YunLi626 (2019) *Biogen posts its the worst day in 14 years after ending trial for Blockbuster Alzheimer's drug*, *CNBC*. Available at: https://www.cnbc.com/2019/03/21/biogen-shares-plunge-more-than-25percent-after-ending-trial-for-alzheimers-drug-aducanumab.html

Zhao, L., Li, L., Zheng, X. and Zhang, J., 2021, May. A BERT based sentiment analysis and key entity detection approach for online financial texts. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (pp. 1233-1238). IEEE.