Can News Headlines be Traded?

Maria Manuel Rodrigues de Freitas

Dissertation written under the supervision of professor Dan Tran

Dissertation submitted in partial fulfilment of requirements for the MSc in Finance, at the Universidade Católica Portuguesa, Tuesday 13th September, 2022.

Acknowledgements

I am very thankful to my supervisor for his guidance and patience in helping me wrap up this stage of my academic path and to professor José Fais for his support throughout my Master's program.

I also want to thank the most important people in my life, my family, for their unconditional support and for pushing me to want to be a better version of myself.

Abstract

Papers have documented a relation between news and financial market movements. We explore this observation on the firm level from a practical investment perspective with the aid of ML methods. Various implementations of NLP models are trained to produce numerical sentiment value from the news, from which the best model is the state-of-the-art 'Finbert' plus a SVM with rbf kernel. The model is used on the news of S&P500 constituents retrieved from 'Reuters Newswire' between the 1st of December 2020 and the 31st of March 2022. Finally, sentiment is aggregated daily to create Long, Short and Long-Short portfolios with 100 and 200 companies. We find that the relationship between sentiment and return is stronger on the same day, with some value being retained the following day. Namely, the Long-Short portfolio achieves the best performance, displaying a significant positive alpha on Fama-French factors and a low r-square. However, the profitability of the strategies does not hold when considering transaction costs of 10bp. A further analysis using the EWCT technique to limit turnover shows that some profitability can still be achieved but only for the Long portfolio, which beats the benchmark by a small margin. This fact is solidified by the different sentiment proxies, which also demonstrate the potential for some profitability in the Long-Short portfolio, highlighting news volume as an essential component of the sentiment. We also find that lower turnover limits in the EWCT strategy provide better returns, meaning that sentiment momentum has value.

Tittle: Can News Headlines be Traded?

Author: Maria Manuel Rodrigues de Freitas

Keywords: Sentiment, News, Natural Language Processing, FinBert, Transaction Costs, Investment Startegy, Turnover Optimization, News Speed Assimilation

Resumo

Jornais académicos documentam uma relação entre notícias e movimentos do mercado financeiro. Exploramos esta observação a nível da empresa segundo uma perspectiva de investimento, com a ajuda de métodos de ML. Diversas implementações de modelos de NLP foram treinadas para produzir uma classificação para as notícias, das quais o melhor modelo é o 'Finbert' mais um SVM com kernel rbf. O modelo seleccionado é aplicado às notícias publicadas pela 'Reuters Newswire' entre 1 de Dezembro de 2020 e 31 de Março de 2022 relativamente aos constituintes do S&P500. Finalmente, o sentimento é agregado em sinais diários utilizados para criar estratégias de investimento com posições Longas, Curtas e Longas-Curtas, testadas com 100 e 200 empresas. Verificamos que a relação entre sentimento e retorno é mais forte no mesmo dia, com algum valor a ser retido no dia seguinte. Nomeadamente, a estratégia Longa-Curta atinge o melhor desempenho, apresentando um alfa positivo significativo em factores Fama-French e um r-squared baixo. No entanto, a rentabilidade das estratégias não se mantém quando se consideram custos de transacção de 10bp. Uma análise adicional utilizando a técnica EWCT para limitar o volume de transacções mostra que a estratégia Longa consegue alguma rentabilidade, ultrapassando o benchmark por uma pequena margem. Este facto é solidificado pelos sentimentos alternativos, que também demonstram potencial na estratégia Longa-Curta, evidenciando o volume de notícias como um componente essencial do sentimento. Verificamos também que limites baixos de transacções na EWCT, proporcionam melhores retornos, o que significa que a tendência do sentimento tem valor.

Tittle: As notícias podem ser usadas para investir?

Author: Maria Manuel Rodrigues de Freitas

Keywords: Sentimento, Notícias, Processamento de Linguagem Natural, FinBert, Custos de Transação, Estratégia de Investimento, Otimização de Volume de Transações, Velociade da Assimilação de Notícias

Table of Contents

List of Figures

List of Tables

Acronyms

BERT Bidirectional Encoder Representations from Transformers

BOW Bag of words

CUSIP Committee on Uniform Security Identification Procedures

DT Decision Tree

DOW Dow Jones Industrial Average

DSS Decision Support System

EMH Efficient market hypothesis

EW Equal Weighted

EWCT Exponentially-weighted calendar time

GI General Inquirer

GICS Global Industry Classification Standard

GNvn GoogleNews-vectors-negative300

H-IV4 Harvard psychosocial dictionary

LM Loughran and McDonald word list

LR Logistic Regression

ML Machine Learning

NB Naive Bayes

NER Named Entity Recognition

NLP Natural language processing

NLTK Natural Language Tool Kit

NN Neural Network

NYSE New York Stock Exchange

RF Random Forest

POS Part-of-speech

RIC Reuters Instrument Code

SEC U.S. Securities and Exchange Commission

SVM Support Vector Machine

TF-IDF Term frequency - Inverse document frequency

TR-EK Thomson Reuters Eikon Data API for Python

UGC User Generated Content

VW Value Weighted

WRDS Wharton Research Data Services

WSJ The Wall Street Journal

Chapter 1

Introduction

Relevance

The EMH dictates that rational investors incorporate all available information into pricing Fama (1970). Since its inception, this theorem has been the basis of numerous papers exploring the flow of quantitative information in capital markets and their impact on stocks. Contrarily, behavioural finance presents that investors can be "emotional" and have biases originating from market sentiment, which triggers fluctuations in the value of companies Nofsinger (2005).

Until recently, the analysis of qualitative market information associated with sentiment was lacking. However, advances in NLP allowed for a more reliable analysis of different financial corpora Jegadeesh and Wu (2013). Within the available economic sources, news headlines contain relevant information about events in a less ambiguous way than the corpus Radinsky et al. (2012). Moreover, the news is a crucial source of information for many investors Koratamaddi et al. (2021) as shown by a glitch in the Bloomberg news service in 2008, which incorrectly reported the death of Steve Jobs, the CEO of Apple, leading to a 5.4 % intra-day drop in the company's stock.

Despite the increased literature around textual sentiment extraction, a significant portion still makes simplistic assumptions in their analysis. Frequently, words are seen as independent from one another, allowing for scenarios where the negation of a positive word (e.g., "good") is not understood Tetlock (2007). Also, the magnitude of the sentiment embedded in words is often forgone to enable more accessible classifications built on the news count according to a binary scale (positive or negative). Finally, very few papers comprehensively analyse sentiment from an economic point of view. Commonly, trading strategies built around sentiment lack execution

considerations and do not account for trading costs. Thereby, most papers report unrealistically good performances, leaving open the question of what is their actual profitability.

Project Scope

This thesis focuses on producing a reliable sentiment score from news headlines and analyzing the economic viability of using it for portfolio creation. For that purpose, we outline two main tasks. First, apply state-of-the-art and commonly used NLP models to classify news headlines and assess the performance improvement of such models in contrast to each other. Second, provide a comprehensive comparison of different investment strategies and the impact on their performance when considering realistic assumptions such as trading costs, while answering questions like:

- Is sentiment assimilated by returns only on the publication date?
- Is there value in sentiment after adjusting for well-known risk factors?
- Can costs be minimized in the portfolio creation enough to allow for profitability?

To answer these questions and provide a complete pipeline from news to strategy, the remainder of this dissertation is organized as follows. Chapter 2 provides an overview of the literature on the subject and highlights important concepts. Chapter 3 explores the data sources used, including one for training the ML model and another for the economic analysis, consisting of news headlines provided by the Reuters Newswires service between the 1st of December 2020 and the 31st of March 2022. Chapter 4 focuses on the methodology applied for news classification and the creation of investment strategies. Chapter 5 contains an analysis and discussion of the results, and finally, in Chapter 6, we present the main conclusions, limitations, and possible future avenues for research.

Chapter 2

Literature Review

2.1 Overview

The prediction of stock market movements and how they incorporate information has been a subject of extensive research in the academic field. Notably, two models widely contrasted are the EMH and behavioural finance.

Within this scope, (Akerlof, 1978) and (Milgrom and Stokey, 1982) argue that markets only function if there are noise traders to provide the liquidity for rational ones. This relation is also identified by (Shiller, 1980) and (Milgrom and Stokey, 1982), which suggest it allows for a market overreaction to new information. On the other hand, (Cutler et al., 1988) gives proof of market underreaction to the news.

The question is no longer whether investor sentiment affects stock prices but how to measure and quantify it (Baker and Wurgler, 2006). In that regard, research has mainly focused on two interpretations of sentiment: (I) Textual sentiment classified according to tone, polarity and subjectivity; (II) Investor sentiment, described by (Baker and Wurgler, 2006) as "a belief about future cash flows and investment risks that is not justified by the facts at hand" (Kearney and Liu, 2014).

Focusing on the first definition, data mining and machine learning advancements have allowed for better textual sentiment analysis, enabling ample research on its relationship with returns. Thereby four main hypotheses have emerged:

i) Media sentiment predicts investor sentiment, impacting short-term returns.

ii) Investor sentiment predicts media sentiment, which reflects past stock movements.

iii) Media sentiment contains new information not incorporated into prices.

iv) Media sentiment contains stale information.

Most notably, (Antweiler and Frank, 1998) found that news predicts abnormal returns, which reversed (hypothesis 'i'). Similarly, (Tetlock, 2007) finds that media pessimism predicts a downward pressure on prices, which reverses to fundamentals. Moreover, they discover that changes in market returns are also reflected in the following news sentiment (hypotheses 'i' and 'ii'). On the other (Tumarkin and Whitelaw, 2001) find that bulletin board postings cannot predict returns and that strong positive returns preceded days with unusual activity and positive opinions.

Other textual analyses, such as 10K fillings, report a positive relationship between tone and future earnings after controlling for known determinants of performance (Li, 2010). Also, in line with hypothesis 'iii' (Li et al., 2014) present that news can contain some fundamental information. Finally, (Dewally, 2003) discovered that two newsgroups provided no value to the readers between April 1999 and February 2001 (hypothesis iv).

2.2 Textual Data Types

Textual analysis can be employed to capture opinions, emotions or relevant information. Different texts contain distinct types of information with specific advantages and disadvantages. In finance, three main types are commonly used: corporate documents (e.g., conference calls and filings); news media (e.g., news and analyst reports); and UGC such as social media (e.g., Twitter) (Marty et al., 2020).

Corporate documents are official releases written by insiders who know the company better, and as of 2005, SEC guidelines ensure some consistency in reports (Loughran and McDonald, 2016). They are firm-specific and contain forward-looking information. Nevertheless, it is unlikely that they include the whole truth and manipulations or biases are possible. They have a low frequency, usually quarterly and annually, and limited explorable events (e.g., annual reports) (Kearney and Liu, 2014). These characteristics make corporate documents more appropriate for cross-sectional analysis and event studies on the future outcome of individual firms.

News media can emerge from various sources with different qualities. They reflect aspects of the past or schedules instead of providing foresight and consist mainly of views from company outsiders. At the same time, multiple press channels are acknowledged as credible sources

relevant to fundamentals (Schumaker et al., 2012). They are flexible in the context (e.g. market-level or firm-level), range of entities and events, and produced continuously in large volume and frequency (monthly, weekly, daily, and even intra-day levels) (Marty et al., 2020). As such, news media are suitable for studying multiple events over varying time and entities.

UGC is primarily written by uninformed investors in an open and unregulated environment, resulting in a less accurate and noisier source. Their low information-to-noise ratio makes the data cleaning and processing more challenging (Kearney and Liu, 2014). Similarly to news media, they have flexibility in frequency, events, and market and, since 2016, have rapidly become the dominant source of media (Gan et al., 2020). They are more relevant for testing behavioural finance than market efficiency and are a good way of capturing smaller investors' sentiment. Moreover, despite the noise (Black, 1986) contends that a vast number of small events can originate more causal sentiment than the one provided by a smaller number of large events.

2.3 Data Treatment and Classification

Textual data is high dimensional unstructured data that follows specific linguistic rules making its meaning hard to capture. A critical step in text mining is feature processing – usually consisting of feature extraction, feature selection and feature representation – that depending on the goal, might require the implementation of all three stages of the process or not (Yazdani et al., 2017).

2.3.1 Feature Extraction

Feature extraction is a parsing procedure that breaks down documents into meaningful units (or tokens), like words or sentences (Yazdani et al., 2017).

Commonly, this process is used to identify words through word segmentation (or tokenization) (Schumaker and Chen, 2009). A simple approach is a BOW which uses the frequency of discrete words to represent text in a term-document matrix. Chosen by its ease of implementation and transparency (Gidofalvi and Elkan, 2001), it has the downside of not capturing semantics and assuming word independence. Extensions of this model slightly overcome this aspect by using a combination of words, as is the case of N-grams (skip-grams), which use (non)continuous sequences of n-words. However, they suffer from data sparsity and high dimensionality (Liu,

2020).

Higher-level methods use sentence tokenization, allowing the capture of semantics. For these cases, the assumption of independence between tokens, which are the sentences, is more solid as they usually have structural and grammatical independence (Grefenstette and Tapanainen, 1994). For instance, POS tagging labels words according to classes and syntactic categories, indentifying negations and intensifiers, which can affect the meaning of other words. Another seldom used method is the NER, which allows recognizing entities, such as firms, within textual media.

2.3.2 Feature Selection

Used to retain the most relevant information, it allows for reducing the dimensionality and noise of the inputs. Generally, it includes processes like removing redundant information like stop words (e.g. articles and prepositions), eliminating special characters or punctuation or "simplifying" expressions. For the latter, the conventional techniques used are Stemming, which removes (pre)suffixes (e.g. -ing, -ed), reducing the word to its steam and Lemmatization, which preserves only the root of words (Marty et al., 2020).

A common method used in finance consists in leaning on existing dictionaries and lexicon created by experts ((Tetlock, 2007),(Engelberg, 2008),(Loughran and McDonald, 2011)). Notably, in early research, the GI H-IV4 word lists (Stone et al., 1966) were the standard. However, as they use general linguistics premised on sociology and psychology, they can misclassify words like homonyms (e.g. "crude" can refer to the oil industry). Nowadays, instead, the use of word lists specific to the finance and accounting domain is predominant. Examples include (Henry, 2008) and (Loughran and McDonald, 2011) (LM) word lists.

However, using a dictionary or lexicon-based approach limits words to the context and time of the dictionaries, which are not updated frequently and might not exist specific to the problem at hand. Thus, authors like (Li, 2010) advocate for applying statistical methods, such as TF-IDF and information gain, which eliminate irrelevant words according to thresholds.

A less common technique uses the returns of companies, to which the textual media refers, as an "external market feedback" to determine word importance. One such case is (Jegadeesh and Wu, 2013) which regresses the returns against the number of words to derive weights for each word. More recently, (Ke et al., 2019) use returns to create a list of relevant words via predictive screening and then assign prediction weights to each word through topic modelling.

2.3.3 Feature Representation

Feature representation is an optional step that transforms selected features into a numerical format. Often the representation utilizes the dictionary or statistical probabilistic scores used in feature selection. However, it can also result from more advanced techniques such as hierarchical clustering and word embedding (Marty et al., 2020).

Word embedding models take a corpus of text and learn to represent the words within a vector space, where close vectors mean similar context (Vajjala et al., 2020). The downside is that they are costly to train, both in time and computing power. Some papers overcome these obstacles by relying on pre-trained models publicly available, consisting in key-value pairs where the key is a word, and the value is a vector. (Souma et al., 2019) uses GloVe trained on Wikipedia documents, while (Kim and Lee, 2018) uses Googles´s word2vec model trained on the news.

2.3.4 Data Classification

With the data processed, the next step is scoring it, which can vary in (i) the type of score/task and (ii) the model used for scoring.

Regarding the first, the most common is a classification-based analysis that produces a discrete sentiment variable. For instance, binary classifications can be used to capture the directionality of sentiment and, potentially, the market. More complex versions use multiclass models, as is the case of (Bouazizi and Ohtsuki, 2016) that use seven different classes (e.g. Love, Hate) when studying twitter messages.

A more challenging and less used scoring is based on regressions, which attempt to classify across a continuous scale, with larger absolute values meaning higher magnitude signals. For instance, (Jiang et al., 2017) uses ensemble regression algorithms to attribute a score between -1 and 1 to news headlines and microblog messages.

In terms of the model used for scoring, a considerable body of research has used analytical methods in conjunction with a dictionary or lexicon-based approaches (Marty et al., 2020). They perform arithmetic calculations such as sums and averages of words along their dictionary categories (Loughran and McDonald, 2011).

Models of increasing popularity are ML methods which can identify (non)linear patterns between words and their sentiment (Li, 2010). Within these, the most used are supervised models, which require a labelled training set. Some examples include: (i) NB (Antweiler and Frank, 2004), which uses the posterior probability of the document belonging to a given class (e.g.

positive or negative) based on the words present in the corpus; (ii) SVM (Kalyani et al., 2016) which separate data in hyperplanes and has shown good performed in stock market prediction from news headlines (Prosky et al., 2017); and (iii) DT that divide data hierarchically based on different text features (Aggarwal and Zhai, 2012).

Lastly, some sentiment analyses use a hybrid approach that combines multiple techniques (Sankar and Subramaniyaswamy, 2018). For simplicity, we only presented, in-depth, in the methodology, the best performing models.

2.4 Theoretical Findings

(Shiller, 2000) argued that news media are essential in setting the stage for market movements and inciting them. He suggests that news content drives market sentiment. Formally, (Chan, 2003) find a strong drift after bad news, seen mainly in smaller and more illiquid stocks, which supports some theories of investor over and underreaction. Likewise, (Antweiler and Frank, 2006) verifies an overreaction of investors by employing a NB algorithm to WSJ news between 1973 and 2000. However, this reaction is followed by a short-term reversal, which takes more time to occur during recessions. Using the WSJ but focusing on the column "Abreast of the Market" (Tetlock, 2007) shows that pessimism predicts daily stock returns between 1984 and 1999, especially for smaller-cap companies. However, negative sentiment pressure on returns is also found to be followed by a reversion to fundamentals within a trading week. This is interpreted by (DellaVigna and Pollet, 2009) as traders catching up to the market, on Monday, after re-evaluating their investments at the end of the week. Extending this work, (Tetlock et al., 2008) analyzes the impact of negative words within the WSJ and DOW News Service stories from 1980 to 2004 for S&P 500 companies. They report three main findings: (i) the fraction of negative words can forecast low company earnings; (ii) negative words predict negative abnormal returns on the following day; (iii) the predictability is higher for stories focusing on fundamentals.

Using another well-known journal, the New York Times (Garcia, 2013) also supports that stocks react to the news of the DOW between 1905 to 2005. Moreover, they find that news impact on returns is particularly evident in recessions and on Monday, which is affected by news written on Saturdays and Sundays. Nonetheless, this impact reverses over the following trading days. (Hillert et al., 2014) in a study of 2.2 million news between 1989 and 2010 find that news

exacerbates the momentum effect. In other words, firms with excessive media coverage experience stronger momentum, with the impact also depending on aspects like the level of uncertainty of the stock and investor individualism. On news coverage, (Ferguson et al., 2015) also find the tone and volume of news to predict returns significantly, especially in lower visibility firms.

More recent work like (Heston and Sinha, 2017) confirms previous research by finding that daily news predicts stock returns for around one or two days. (Uhl et al., 2015) find longer-lasting effects over the news when aggregated over a week, but their research lacks transparency into the construction of tone and sentiment.

2.5 Sentiment Trading Strategies

Following theoretical discoveries, papers have attempted to analyse the economic utility of sentiment through different trading strategies. Some use real-time market data and NLP to develop algorithmic trading systems. Others regard it as a DSS with use in financial markets and banking (Gagnon, 2012) as evidenced by the development of services such as Thomson Reuters News Analytics which provides customers with scores of polarity and news novelty.

Methodologies for economic anylzis vary significantly across papers, namely in terms of four components: (i) the benchmark and metrics used for comparison; (ii) the degree of trading execution considerations; (iii) the complexity of the trading strategy applied and inputs considered in the same.

2.5.1 Benchmark

Suitable benchmarks must be employed to provide an unbiased and realistic evaluation. According to (Cremers et al., 2012), to determine the value added of a startegy the benchmark should be a passive portfolio with similar risk.

Within the equity markets, the prevalent approaches include using stock indexes, such as the S&P500, DOW or Frank Russel indexes. They are standard benchmarks because of their liquidity, diversification, and ability to capture broad market aspects, making them attractive for an average investor. The S&P500 is particularly common as it focuses on large-cap U.S companies and is considered a good representation of the U.S economy. However, it can be criticised for its disproportionate weight towards larger companies. Similarly, Russel also captures the U.S economy, including a slightly higher number of stocks – 12% more than the S&P500 – and

mid-cap stocks that are a bit more volatile. Lastly, the DOW encompasses a much smaller set of stocks, focusing on blue chips and specific sectors.

2.5.2 Trading Considerations

To analyze a strategy's performance as comprehensively and realistically as possible, aspects like trading costs and media timing need to be considered.

Trading costs include direct costs, like brokerage and taxes, and implicit costs, such as market impact cost, price movement, and opportunity cost (Domowitz et al., 2001). These costs impact strategies' profitability, especially sentiment ones, due to their high turnover (Graf, 2011). However, prior research usually neglects them, generating unrealistic studies, particularly for strategies at higher frequencies (e.g. intraday or daily).

Trading costs depend on the underlying trading security or market (Yang et al., 2017). (Fong and Yong, 2005) use 50 bps in their analysis of global equity markets, while (Harris and Yilmaz, 2009) apply a lower rate of 10bps for foreign market exchange. For large-cap stocks on NYSE (Chan and Lakonishok, 1997) show that the average round-trip cost is 20bp. More recently, (Frazzini et al., 2018) reported that asset managers incur approximate daily costs of 10bp for large stocks and 20bps for smaller ones.

Regarding the media timing issue, the main problem rests in the time it takes to retrieve and process textual clues to then be able to act on them. Typically, analyzes are built on the premise that the sentiment signal is retrievable in a concise amount of time, which is, in most cases, unrealistic. As a solution, some authors like (Ke et al., 2019) are careful in forming portfolios only at the market open of each day, giving leeway to process information.

2.5.3 Types of Trading Strategies

Trading strategies can differ significantly depending on how they are constructed. We highlight three main types of distinctions in terms: (i) the degree of complexity, as in the mechanism that triggers a given transaction; (ii) the frequency at which positions are adjusted; (iii) and how the signal is interpreted.

Complexity

Strategies can range from simpler human-defined rule-based models to ML optimization focused on determining the optimal combination of trading indicators (Hochreiter, 2016).

The simpler approach can be a condition like "buy stock X when sentiment is above threshold Y". However, these thresholds can be arbitrary and small changes can generate very distinct performances (Feuerriegel and Prendinger, 2016). Instead, an alternative is to compare the relative value of sentiment either between stocks or past performances. For instance, (Zhang and Skiena, 2010) ranks stocks according to their daily sentiment and then goes long and short on an equal number of positive and negative stocks, respectively.

Similarly, (Wang et al., 2021) apply a majority vote strategy by having the past three days' sentiment determine the ranking and then buy/sell the top/bottom 20%. (de Oliveira Carosia et al., 2021) also uses past information, which he compares against the current one to determine which transaction to make.

More complex strategies include leaning-to-rank algorithms like ListNet, which employ ML to rank stocks based on a group of inputs by optimizing their weights. (Song et al., 2017) use six metrics with information on investor sentiment and previous market performance to fine-tune their models. Another ML optimization approach is genetic programming, which, inspired by biological breeding and Darwinism, starts by generating random programs and then selects the fittest for reproduction (Yang et al., 2017).

Frequency

Most papers focuses on relations at smaller frequencies. For instance, (Yang et al., 2017),(Makrehchi et al., 2013) and (Katayama and Tsuda, 2020) create daily sentiment measures based on averages or counts news sentiment and then use them as signal to adjust their portfolios. Recently, however, high-frequency trading has been gaining more popularity, having gone from 10% of all equity trades in the 2000s to 50% by 2018 ((Gao and Liu, 2020),(Zaharudin et al., 2022)), with papers finding economic value in it (Sun et al., 2016). Fewer attempts have been made to leverage sentiment in lower-frequency trading. (Uhl, 2014) employs a strategy that is only adjusted once a month, while (de Oliveira Carosia et al., 2021) makes weekly adjustments.

Signal Interpretation

There are two main types of strategy, the follow the loser approach or follow the winner. For many years, the most implemented strategy has been to go long(short) on firms with positive(negative) news ((Tetlock, 2007), (Engelberg, 2008)). However, the opposite approach has also been recognized as profitable, attributed to a mean reversion in financial markets (Li et al., 2016)

2.6 Economic Findings

Findings on the profitability of strategies built on just sentiment tend to vary significantly. However, most of them have in common a lack of practical and cost considerations. For instance, (Zhang and Skiena, 2010) using insufficient metrics and ignoring costs, find profitability between 2005 and 2009 for a long-short strategy built on NYSE stocks. Similarly, ignoring costs (Schumaker et al., 2012) report returns between 3.30% and 2.41% for portfolios constructed on S&P500 stocks over 5 weeks. (Song et al., 2017) run experiments on long-only and long-short portfolios built around 128 highly liquid stocks. Their strategies are adjusted at a weekly frequency spanning 12 years and provide Sharpe ratios of 0.55 and 1.29 for the long-only and long-short portfolios. Additionally, they find lower volatility periods to be beneficial for the performance of the strategies. (Palmer and Schäfer, 2020) achieve average monthly factor-adjusted returns of 0.7% in a long-short portfolio of DOW stocks from 2010 to 2018.

Contrarily, (Kelly and Ahmad, 2018) employ a more realistic analysis that explores daily adjusted strategies built on the DOW from 2005 to 2015, achieving a Sharpe ratio of 0.39, beating the 0.16 achieved by just holding all the stocks. (Ke et al., 2019) achieve net Sharpe ratios as high as 2.3 on long-short portfolios constructed on the DOW Newswires database between 1989 and 2017.

Chapter 3

Data

We utilize three primary data sources, two of textual character, namely news headlines for NLP training and for evaluating investment strategy and numerical information (stock returns).

3.1 News Headline for NLP Training

For the NLP training, a larger dataset is compiled by aggregating the following smaller publicly accessible datasets with annotated sentiment.

The **'SemEval-2017 Task 5'** is a series of international NLP research workshops to advance the current state-of-the-art in semantic analysis (Cortis et al., 2017). As part of the 11th workshop, a dataset on financial texts was provided, including two tracks, Microblog messages and news headlines. We only consider the news headline corpus, which consists of 1,780 news statements collected from 20,000 RSS feeds (e.g. Reuters and Bloomberg). Each headline was manually annotated by three independent financial experts and further consolidated by a fourth domain expert into labels ranging from -1 (negative) to 1 (positive). Finally, the quality of the annotations was determined following a similar methodology to (Takala et al., 2014), resulting in a gold standard containing 1,647 labelled data points. Then the non-published labelled data used for participant evaluation is excluded, leaving 1,156 headlines, of which 658 are positive (sentiment score > 0), 460 negative (sentiment score < 0) and 38 neutral Table A.1.

The **'Financial Phrase Bank'** is a collection of 4,840 sentences sampled from financial news texts and company press releases (Malo et al., 2014). The data is categorized into positive, negative or neutral, following annotation by 16 people with adequate background knowledge of financial markets. In this manner, each sentence was attributed between 5 to 8 annotations

aggregated based on majority agreement. Four alternative datasets are provided centred on the strength of the majority agreement, from which, in this dissertation, the version of sentences with more than 50% agreement is chosen. Out of the 4,846 headlines, there are 2,879 neutral (60%) 1,363 positive (28%) and 604 negative (12%).

After combining the datasets mentioned above, we eliminated all duplicated headlines. This avoids inducing ML models to an error by disregarding cases of headlines with opposing annotations. This is important due to the lack of a NER mechanisms in the models applied. Finally, neutral headlines (2,890) are eliminated to focus on binary classification. The final clean dataset contains 2,736 unique headlines, of which 1,778 positive (65%) and 958 negative (35%), constituting unbalanced data. Headlines, both positive ad negative, have, on average, a length of 17 words with a standard deviation of 9.

3.2 News Headline for Investment Strategy

To evaluate investment strategies and analyze the economic viability of sentiment, we focus on the S&P500 companies. This choice is made because (i) they are representative of the U.S market by accounting for approximately 75% of the total market capitalization (Tetlock et al., 2008), (ii) there is ample and frequent news on the companies included in the index and (iii) the stocks are mostly liquid allowing for more realistic trading strategies.

A corpus of news headlines for each of the S&P500 constituents, between 1st September 2020 and 31st March 2022, is retrieved using the TR-EK API for Python. The timeline is restricted by the limited historical access of TR-EK.

First, a list of all the companies in the index and their respective RIC was compiled using the 'Get Data' method to obtain the current list of constituents and the chronological list of joiners and leavers. Secondly, news headlines were retrieved using the 'Get News Headline' method to request all English news for each RIC. Resultingly, we obtained a set of headlines complete with the timestamp, RIC, and source identification. Thirdly, we used the 'Convert Symbols' method to obtain CUSIP and Ticker to aggregate the data with returns later. Also, the 'Get Data' method was used again to obtain the GICS Sectors for statistics.

A corpus of 1,330,190 news headlines is obtained, filtered to exclude document fillings announcements from the sources 'Edgar Fillings' and 'Global Filings' and earning calls transcripts from 'Event Transcripts News'. Similarly, all news for the company 'Moody's' is removed, as

their content concentrates on rating changes of other companies. Then, to avoid including stale news, only the first occurrence of a repetitive headline is kept by removing any duplicates considering headlines and RIC code.

The final dataset consists of 763,169 headlines characterized by a common occurrence of the words 'shares', 'earnings', 'buy', 'sell' and 'covid19'. These are expected to be reliable and valuable sources that are used by professional investors(Hillert et al., 2014).

3.2.1 Data Characteristics

Concerning the level of media coverage per day, a given company has three articles written about them on average, Table 3.1.

Decile [Min-Max]	***Average Headlines per Day***			***Companies***
	Mean	**Median**	**% days without articles**	**n**
First [0.0 - 0.6]	0.50	0.52	79.35	53
Second [0.6 - 0.8]	0.73	0.73	69.83	54
Third [0.8 - 1.0]	0.93	0.93	61.31	53
Fourth [1.0 - 1.2]	1.11	1.13	57.28	52
Fifth [1.2 - 1.4]	1.30	1.29	50.26	53
Sixth [1.4 - 1.6]	1.49	1.50	46.58	52
Seventh [1.6 - 2.0]	1.79	1.79	44.47	53
Eight [2.0 - 2.8]	2.33	2.29	35.76	53
Nineth [2.8 - 4.9]	3.60	3.37	27.49	53
Tenth [4.9 - 5.6]	12.43	8.90	10.87	53
Total	2.62	1.40	48.35	529

Table 3.1: Average Headlines per day by Decile

The companies' deciles highlight that the average media coverage level can differ substantially between firms. Namely, the ones in the 10th decile have three and a half times as many articles written compared to the ones in the 9th decile. On the other hand, the difference between other subsequent deciles is much smaller, with the highest increase in the average daily headlines being 47% when going from the first to the second decile. Overall, just a small number of companies within the S&P500 have multiple different news published daily. For these, the sentiment signal is expected to be more encompassing but possibly include some noise.

It is also noteworthy that up until the 9th decile, the mean and median coverage are close to each other, meaning that the distribution of the average daily headlines is practically symmetrical. Contrarily, in the 10th decile, the mean is 40% higher than the group's median. This right-skewed distribution indicates that a small group of companies have much broader coverage than the other companies within this decile.

There is a decreasing tendency in days not covered as deciles increase. Nonetheless, until the 5th decile, most companies are not covered for at least half of the time, and at most in the 10th decile, companies have, on average, 10.9% days without articles. In other words, every week, there is approximately only one day without news coverage for companies in the 10th decile. Comparatively, for the 1st decile, there are only two days in which the companies have news published about them.

From the deciles analyzed, the 10th has the largest range of values, with a minimum and maximum average daily headlines of 4.9 and 45.6, respectively. For that purpose, it is valuable to look at the statistics of the top companies based on average daily headlines, Table 3.2.

Company	***Average Headlines per Day***				***Companies***
	Mean	**Median**	**Max**	**% days without articles**	**Total**
Pfizer	45.57	33.00	379	26292	10.75
Amazon	37.99	37.00	204	21921	0.69
Facebook	34.09	28.00	276	19668	0.69
Microsoft	29.66	28.00	157	17111	0.17
Apple	28.12	24.00	228	16227	0.52
Google	23.29	21.00	96	13437	1.56

Table 3.2: Top Companies by News Coverage

Pfizer is by far the most mentioned company with 26,292 news despite having the most number of days non-covered from the timeframe analyzed (0.25%) when considering the top companies. This was expected given that despite their role in vaccine distribution, Covid19 started to be less talked about throughout 2021.

Looking at the top companies in terms of coverage, it is also notable that the sectors in which they operate are relatively similar. More in-depth, Table 3.3 shows that 'Communication Services' have a much higher coverage in daily quantity and consistency over time. The only other sector with a similar percentage of days covered is the 'Energy' sector, which might be attributed to the increasing focus on sustainable energy. For all the other sectors, on average, there is coverage for half of the relevant period.

Regarding data sources, almost 50% of the news comes from Reuters, despite the available 173 sources, Table A.2.

GICS Sector	*Average Headlines per Day*			*Companies*
	Mean	**Median**	**% days without articles**	**n**
Communication Services	6.58	2.70	32.42	27
Consumer Discretionary	2.96	1.42	48.44	65
Consumer Staples	2.16	1.60	46.00	32
Energy	2.82	2.13	35.96	25
Financials	2.86	1.55	44.97	67
Health Care	3.00	1.31	52.74	68
Industrials	2.00	1.15	51.94	73
Information Technology	2.70	1.28	49.88	82
Materials	1.34	1.15	53.19	29
Real Estate	1.01	0.96	56.35	32
Utilities	1.52	1.33	46.70	29

Table 3.3: Average Headlines per day by GICS Sector

3.3 Returns

Stock values are retrieved from CRSP daily returns and made available through WRDS. The variable Holding Period Returns is used, consisting of returns adjusted for dividends and stock splits.

Chapter 4

Methodology

The methodology is subdivided into two main part. The first part focuses on extracting sentiment from the news through NLP techniques, while the second explores different trading strategies and analyses their viability. In this section, we present an overview of the methodology utilized. The steps applied were implemented primarily in python and are available on GitHub[1].

4.1 NLP Sentiment Analysis

The sentiment analysis aims to process data and use it to train a supervised ML method capable of classifying headlines according to two classes (positive and negative) with the highest precision possible. We use the data described in section 3.1 to train and test different processing methodologies and ML models. Moreover, to have a sample of new data reserved for evaluating the models, we choose an 80/20 split for training and testing, respectively.

For the training, two main NLP techniques were used – word embedding and BERT – to produce numerical outputs that could be further refined through standard ML techniques to output a binary classification. Despite relying on existing pre-trained models, both NLP techniques have their intricacies. For one, each model requires specific pre-processing to replicate the steps undertaken by the underlying pre-trained model. Secondly, they produce vector representations of the textual corpus with different refinement levels.

For clarity, the pre-processing and classifications of each approach are presented in Figure 4.1, and further explained hereafter. Note that both approaches use the same ML models to refine the vector representations into the sentiment classification.

[1] https://github.com/Maria-Freitas/Dissertation2022

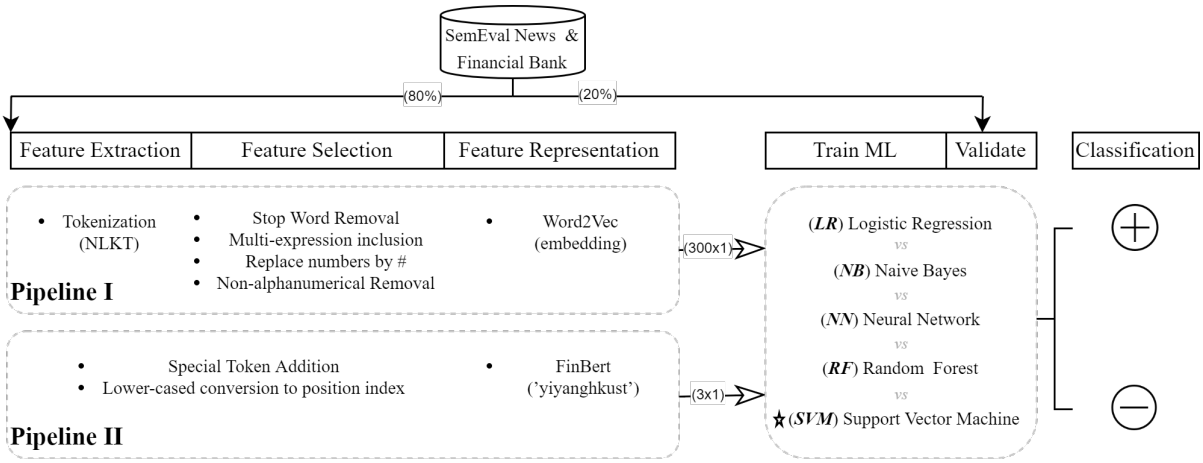

Figure 4.1: Methodology Overview

4.1.1 Data Processing

Concerning '***pipeline I***', each news headline was first tokenized using the Python library NLTK. Then a custom pre-processor was applied, performing the following transformations: (i) removal of author-defined stop words "a", "and", "to, and "of"; (ii) incorporate multi-expressions (e.g., "New_York instead of "New, York"); (iii) replace numbers by hashtags; (iv) and remove non-alphanumerical characters, such as punctuation and special characters.

The resulting tokens are then searched for in the vocabulary of Google's word2vec model GNvn, and the corresponding semantic vector is retrieved. Then, a headlines' vector is computed as the average of the corresponding word vectors and used as input to ML techniques trained to produce a sentiment score.

Introduced by (Mikolov et al., 2013) as part of an open-source project by Google, the GNvn is a word embedding model that outputs numerical representations of word features. These numerical vectors are derived from training a NN on a 6 billion words data set of Google News, with mini-batch asynchronous gradient descent and an adaptive learning rate (Adagrad). Precisely, when trained, each vector's weights are adjusted to minimize the loss function so that similar words are close in a 300-dimension vector space. In turn, this enables the detection of mathematical similarities that take into consideration a word's meaning based on past appearances (e.g., "man" is to "boy"; what "woman" is to "girl") (Rong, 2014).

For '***pipeline II***', the transformations performed mimicked the ones employed in the original pre-trained model used, which is the fine-tuned version 'yiyanghkust/finbert-tone' of a FinBert. Specifically, using the module 'transformers' available in Python, a tokenizer is initiated that: (i) splits each headline into tokens; (ii) adds two unique tokens to denote the beginning and ending

of a sentence ([CLS] and [SEP] tokens, respectively); (iii) converts tokens into indexes of the tokenizer vocabulary and if not present attributes the unknown token ('[UNK]'); (iv) and pads or truncates the sentence to max length. The output of these transformations is used as input for the pre-trained FinBert model, which outputs a 3-feature vector.

4.1.2 Document Scoring with ML

The resulting vectors from each pipeline, 300-feature for '***pipeline I***' and 3-feature for '***pipeline II***', were used as inputs for multiple ML models, trained separately, to solve for a binary solution. Models tested include SVM with different kernels, RF, LR, NB and NN.

The following subsection presents the primary considerations in the training and evaluation of each ML model. No distinction is made between pipelines, as in both cases, the models were trained with the same objective.

Train, Evaluation & Decision Criteria

To evaluate the models' performance, both in terms of optimal parameter calibration and model selection, the decision criteria chosen was the mean F1-Score of each class defined as the harmonic mean of recall and precision (Sokolova and Lapalme, 2009):

$$F1_c = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} \tag{4.1}$$

with

$$recall_c = \frac{C_c}{G_c} \qquad precision_c = \frac{C_c}{S_c} \tag{4.2}$$

where C_c are the correctly predicted instances of class *c*, G_c are all the instances of class *c* and S_c are all the predicted instances of class c.

For the optimizations of the parameter within each given model, 80% of the whole dataset was used. Each ML was trained based on different combinations of hyper-parameters, using the "Gridsearch" functionality, part of the "Scikit" Python's library. Notably, the "Gridsearch" was programmed to provide the hyper-parameter combination that allowed for the best test F1-Score, while using a 5-fold cross-validation.

This process provided quantifiable metrics that served as decision criteria. The selection of the best parameters within each ML model was based on the best F1-Score of validation and on

the gap between the training and validation scores, indicative of the possibility of overfitting. As for the choice between ML models, we used the same metrics but calculated them utilizing the entire training and test data (80% and 20% of the whole dataset, respectively), plus some considerations on the models. Note that the model never sees the previously mentioned test data when training.

4.1.3 Top Models

Overall, the models in '***pipeline II***' displayed the best performance according to the previously defined characteristics and other considerations analyzed in section 5.1.

'***Pipeline II***' uses the intuition of transfer learning by taking an existing pre-trained model and complementing it with an extra layer to convert a 3-class model to a 2-class one. The pre-trained model was FinBERT available by initializing a BERT classifier using the same Python module as the tokenizer. Consisting in the state of the art of sentiment analysis in finance, FinBERT (fine-tuned version) is a finance-specific BERT model published by (Yang et al., 2020). Trained on 4.9 billion tokens – including corporate reports, earning conference call transcripts and analyst reports – and further fine-tuned on the Financial Phrase Bank dataset, the model provides a 3-layer output corresponding to the likelihood of the corpus being neutral, positive, or negative.

First introduced in 2018 by researchers at Google AI Language, the BERT model consists of a multi-layer bidirectional transformer encoder. This model receives as input a single (or pair of) sentence(s) and represents each of its tokens by summing the corresponding token, segment, and position embedding. The resulting vectors are used to pre-train the BERT through a technique called Masked-Language Modeling that randomly masks words in a sentence and then tries to predict them by looking at the sentence bidirectionally at the same time. It also uses a "next sentence prediction" that jointly pre-trains text-pair representations. The results from BERT can then be used in downstream tasks by fine-tuning the model. For a deeper understanding of transformers' architecture, we refer to (Vaswani et al., 2017).

Following the representation of the sentences by a 3-feature vector produced by FinBert, from the ML models tested to receive that input and solve for a positive or negative classification, the one with the best overall performance is the SVM using the 'Radial basis function' kernel.

The SVM is a supervised ML technique, first introduced by (Boser et al., 1992) with polynomial kernels and by (Cortes and Vapnik, 1995) with general kernels, having been motivated

by classification algorithms for separating data such as (Fisher, 1936) and (Rosenblatt, 1961). In simpler terms, the model aims at separating the data into multiple categories with the help of a boundary to differentiate them while transforming the dimension of the data if necessary.

Formally, given a set of data points ($\overrightarrow{x}$,y) where $\overrightarrow{x}$ is a vector with D attributes ($\overrightarrow{x} \epsilon R^D$) and y its label, a hyperplane is defined as to separate the data points by their respective labels. This hyperplane, called maximum margin separator, constitutes a decision boundary defined to maximize the margin between points of both classes (Figure 4.2). It follows that points with unknown labels can be inferred by their position in the hyperplane.

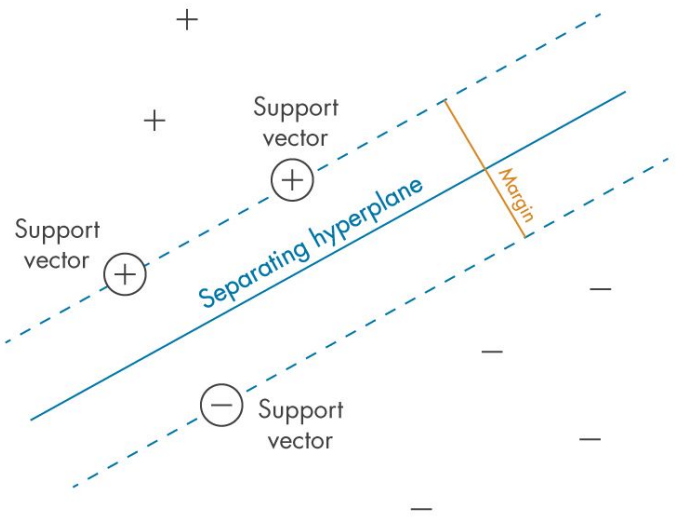

Figure 4.2: SVM Mechanics

However, not all datasets can be strictly separated by their labels using a linear hyperplane. Thus, a "kernel" function can be applied to transform the data and map it into higher dimensional spaces where the classes can be separated. We refer to (Christmann and Steinwart, 2008) for an overview of different kernel functions and their mathematical underlings and to (Powell, 1987) for a more specific explanation of the rbf kernel. Overall, SVM is one of the most used classification methods due to its generalization ability, discriminatory power and optimal solutions (Cervantes et al., 2020). It allows for the transformation of problems into quadratic optimizations, ensuring that the solution found is a global optimum, contrary to other ML techniques that can provide a local maximum, as is the case of neural networks (Russell, 2010).

4.2 Investment Strategies

4.2.1 Sentiment Indicator

The steps defined in 4.1 for the best performing pipeline are conducted on the dataset reserved for investment analysis to obtain a tone classification of positive or negative.

The confidence with which news headlines are classified within each signal is not all the same. News located further away from the separating hyperplane is classified more confidently. However, this confidence score can be arbitrary, vary significantly in magnitudes and have little interpretability. In contrast, probabilities have a specific scale and meaning to them. As SVM do not output probabilities natively, contrary to models like the LR, a probability calibration method needs to be used. Within the methods available, the Platt scaling Platt et al. (1999) is chosen as it is directly available by the python library used for the SVM and is a commonly used and accepted method. It applies a LR on the SVM scores fit by additional cross-validation. Resultingly, probabilities are obtained, which are scaled between -0.5 to 0.5, with values below 0 being negative and those above positive. Thereby, despite having solved for a binary model, using probabilities allows for preserving the tone's polarity and introducing a component of how certain the classification is.

Different methods were investigated to aggregate these into company-specific sentiment daily signals. Ultimately, we chose to aggregate news following (Sinha, 2016) and (Souza et al., 2015) as:

$$Sent_{i,t} = \frac{HL_{Sent}pos_{i,t} + HL_{Sent}neg_{i,t}}{HL_{vol_{i,t}}} \tag{4.3}$$

where $HL_{Sent}pos_{i,t}$ and $HL_{Sent}neg_{i,t}$ represent the total positive and negative sentiment for firm *i* on day *t*, respectively. $HL_{vol_{i,t}}$ is the total number of news for company *i* on day *t*.

For the period considered to make up a day, we follow (Ke et al., 2019) and recognize sentiment on day *t* to use all news available between 9:30 am of the day until 9:00 am of the next day. This signal is used to trade on *t+1*. News between 9:00 am and 9:30am of any given day are excluded, Figure 4.3. This methodology was chosen because it recognizes the earliest time most traders can access the market and that overnight news can be challenging to act on before opening.

Regarding the frequency of the indicator, as hinted previously, daily is selected because it is the frequency with the most literary support for the existence of a relationship between

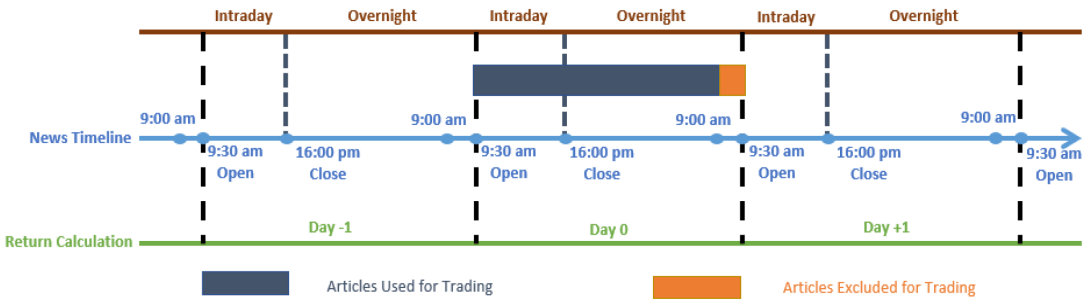

Figure 4.3: News Timeline
Source: (Ke et al., 2019)

returns and sentiment. Additionally, from a trading perspective, it can be executed by a trader without requiring complex algorithms and fast processing power. Thereby it is more manageable than an intraday signal, which is usually only leveraged by funds specialized in high-frequency trading, given their particular investment style and process. On the downside, daily strategies can originate a significant number of transactions which negatively affect profitability, therefore requiring optimization in terms of turnover.

4.2.2 Portfolios

Three basic strategies are explored, consisting of a simplistic short (”S”) and long (”L”) only allocations, as well as a long-short (”L-S”) strategy. All are built with a maximum position of 100 stocks and adjusted daily based on the previous day’s sentiment signal. Trading only 100 stocks per day is deemed acceptable given the limited number of stocks with frequent news in our sample and reduced negative classifications. Nevertheless, results are also provided for portfolios with up to 200 stocks for robustness check.

The amount allocated to each stock every day is also contrasted by performing equal-weighted (”EW”) and value-weighted(”VW”) versions of each portfolio. EW provide a robust means of evaluating the predictive power of sentiment independent of firm size, being closer to the way hedge funds use news text for portfolio creation. In contrast, VW can potentially provide practical trade benefits (lower transaction costs) and be justifiable for economic reasons by attributing more value to larger more productive firms (Ke et al., 2019).

Finally, a version of the portfolios is also built based on the follow the loser approach, which anticipating a contrarian movement from the market dictates that a position is taken in opposition to that of the signal (e.g., if positive, go short).

These portfolios are used to test the profitability of investment strategies as is but also highlight the economic magnitude of the sentiment measure as a predictive signal. To evaluate their risk-reward we present the maximum drawdown, represented as the peak-to-trough decline relative to the value at peak, as well as daily turnover as defined in Champagne et al. (2018):

$$\frac{1}{2T}\sum_{t=1}^{T}\left(\sum_{i} \mid w_{i,t+1} - \frac{w_{i,t}\,(1+y_{i,t+1})}{1+\sum_{j} w_{j,t}y_{j,t+1}} \mid\right) \tag{4.4}$$

where $w_{i,t}$ is the weight of stock *i* in the portfolio at time *t*.

We also report the annualized Sharpe ratio and average returns as percentages. Moreover, we compute alphas and r-squares from regressions on Fama-French factors to understand their exposure to standard aggregate risk factors.

To evaluate the real-world profitability of the strategies we present a version of the top portfolios, including a trading cost of 10bp for each transaction, chosen due to the nature of the stocks studied, mostly large-cap ones.

As the strategies rely on daily adjustments based on sentiment, which can be very volatile, an attempt to optimize the allocation and reduce costs is also tested by following the exponentially-weighted calendar time (EWCT) strategy presented by (Ke et al., 2019). The strategy is characterized by turning over only a fixed portion of the existing portfolio daily, as defined by a user-chosen γ parameter. Specifically, starting from an EW portfolio with *N* stocks (at *t*) each subsequent day (at *t+1*) only a portion γ of all existing positions is liquidated and reallocated to an EW portfolio base on previous day sentiment. This mechanism rewards stocks with consistent performances while penalizing those with very volatile signals.

In the strategy, γ, can take values between 0 and 1, with a value of 1 meaning full liquidation of positions to be replaced based on new sentiment signals. In contrast, a 0 γ means that the portfolio would be the one created on the first day and never change. By varying this turnover parameter, we can influence the importance of assimilating new information and discarding old by choosing a high γ versus rewarding trends and limiting the turnover through a lower γ. Moreover, the EWCT guarantees a daily turnover smaller or equal to γ.

Chapter 5

Discussion

5.1 Sentiment Analysis

5.1.1 ML models Evaluation

The F1-Scores of the training (size of 2188) and test (size of 548) data for each experimented model are shown in Table 5.1

Processsing	**ML techinque**	**Model Selection**	
		Train	**Validation**
None	Baseline Model	39.91%	37.23%
Word2Vec	Linear SVM	88.24%	79.54%
	Logistic Regression	87.14%	78.35%
	Naive Bayes	63.45%	63.45%
	Neural Networks	99.38%	80.86%
	Poly SVM	97.32%	80.38%
	Random Forest	98.81%	66.63%
	Rbf SVM	96.55%	80.70%
	Sigmoid SVM	72.23%	67.35%
FinBert	Basic	81.56%	81.29%
	SVM	86.15%	84.20%
	Logistic Regression	87.16%	84.71%
	Naive Bayes	86.32%	84.16%
	Neural Networks	87.36%	85.15%
	Poly SVM	86.78%	84.68%
	Rbf SVM	87.68%	85.46%
	RFC	87.92%	85.10%
	Sigmoid SVM	84.47%	83.39%

Table 5.1: ML Models F1-Score Comparison

The worst performing model, on both training and test set, is by far the majority class baseline, which predicts all headlines as positive (F1-Score around 38.57%). Comparatively, all other models show a significantly better performance than the baseline. This was expected be-

cause of the comparison metric chosen, which differs from an accuracy measure that would have shown a 65% score.

Regarding the F1-Scores of the training set used for model selection, the only models with scores above 90% are four models, all part of pipeline I. Specifically, they are the RF (98.81%), NN (99.38%), SVM with Rbf (96.55%) and Poly (97.32%) kernel. However, their test scores are considerably lower, with the RF experiencing the highest decrease of 32.18% in score from train to test, while the other models decrease around 17%. With a lower test score, this gap between train and test score is evidence of overfitting, meaning the model is fitted too close to the training set, picking up its noise. Thus, despite the SVM with rbf kernel and NN having the best test scores from pipeline I, which are only around 4% lower than the ones from pipelines II, they are not adequate models given the degree of overfitting.

The test scores are lower than the train scores for all models, except for the NB, for which both F1-Scores are 63.45%. To some extent, these results can be explained by the dataset size and model inputs. Nevertheless, there is a clear difference between the models applied to data processing according to pipeline I (300-feature-long vectors) and the data processed by pipeline II (3-feature-long vectors). The first displays a more significant gap overall between training and test scores, as seen in the previous examples, while the latter appears to be better at generalizing.

Focusing on the models based on FinBert, they have all superior test F1-Scores compared to pipeline I models, ranging from 81.29% to 85.46%. This is anticipated as the central part of the tone recognition is made by the FinBert model – not influenced by the training data – with the additional ML layer trained on the dataset serving only to fine-tune the model into a binary classification. For the same reason, models trained on fewer parameters are expected to perform better, as more complex models are easier to overfit. The top 2 performing models following FinBert are the SVM with rbf kernel (85.46% F1-score) and the NN (85.15% F1-score), with the others following very closely. When comparing the test scores of these two models against their train scores, both the Rbf SVM (87.68%/ 85.46%) and NN (87.32%/ 86.39%) can be said to overfit by 1.03 times, a value which we deemed acceptable.

As all the other Finbert Models, excluding the original, perform very similarly in terms of overfitting and overall test score, selecting either for the classification of headlines would be acceptable. Nonetheless, we chose to use the SVM with rbf kernel has SVMs have been reliably used on many headline classification attempts, with authors such as (Prosky et al., 2017) finding the use of the model with the rbf kernel to have an outstanding performance.

5.1.2 Top Model Assessment

Table 5.2 displays the F1-score and other performance metrics, by class and the overall macro average, for the test data unseen by the model.

	Negative	Positive	Macro Average
Precision	85.51%	86.51%	86.00%
Recall	79.37%	90.77%	85.07%
F1-Score	82.33%	88.59%	85.46%
Instances	223	325	533

Table 5.2: Test Data Metrics

Comparing the in-class performance, all metrics are superior for positive classification, which was expected due to the imbalance of classes, as there is approximately one negative news for every one and a half positive. In particular, in terms of recall, out of all the positive news, the models accurately predict 90.77% of the time, while for negative news, that value is around 11% points lower (neg. recall 79.37%). This means that from all positive news, only 9.2% are predicted as negatives in opposition to the 20.6% of negative news predicted as positive. This discrepancy in recall higher than the precision means that the model will tend to predict more positive news. More similarly across classes is the precision, where in both, for every ten predictions of the respective class, on average, nine are corrected, positive precision 86.51% and negative precision 85.51%. To better illustrate these points, the confusion matrix is present in Figure 5.1.

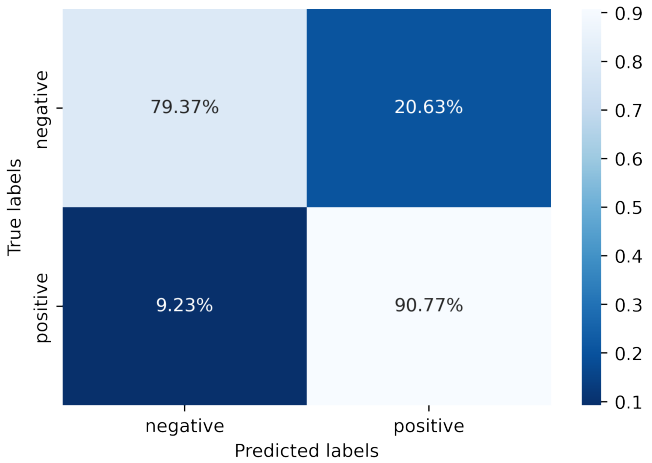

Figure 5.1: Top Model Confusion Matrix

The macro average precision and recall are very close, which is desirable for the problem at hand as there is no preference between capturing all positive news with the chance of wrongly

classifying negative (higher recall) or making sure that we get positive predictions right by classifying less as positive (higher precision). Moreover, the model seems to provide reliable classification, as shown by its accuracy of 86.13%, which consists of the percentage of correct predictions.

To understand how the selected model performs against other papers' models, we choose examples with similar underlying datasets and classification problems to ensure a fair comparison (Hagenau et al., 2013).

(Mishev et al., 2020) analyze the same data set combination while keeping three classes (positive, negative, and neutral). For simpler models such as a lexicon-based one using H-IV4 they achieve an F1-Score of 62.5%, while for statistical methods such as the H-IV4 81.3%. Word encoders like word2vec achieved a score of 79.2%, sentence encoders 74.4% and transformers like FinBERT 89.3%. Overall, among all models tested, F1-Scores vary between 62.5% and 94.7%.

The second most relevant paper results to compare would be those focused on at least one of the datasets used in the combination employed in this dissertation. For instance, (Malo et al., 2014) analyzes the financial phrase bank data verifying an in-class F1-Score between 58.5% and 74.43% in the sentences with >50% annotator agreement.

We deem the model selected as sufficiently good, as the performance falls within the range achieved in previous literature.

5.1.3 Descriptive Statistics Investment News

In this subsection, the news classified, based on the top model, are analyzed through descriptive statistics and visualizations.

Figure 5.2 shows the distribution of sentiment. As previously outlined, the values can range from -0.5 to 0.5, with those above 0 signalling positive news and those below negative. Accordingly, we can see a distribution with a minimum and maximum of approximately -0.50 and 0.49, respectively.

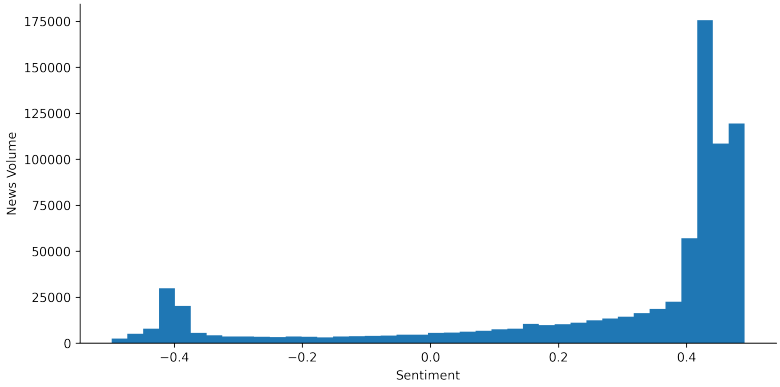

Figure 5.2: Sentiment Distribution

More noteworthy is the sentiment distribution, with an average of 0.27 and a median of 0.42 showing a positive bias. This means that it includes more positive than negative news, a fact also highlighted by the negative skewness of -1.54. This was expected to some extent, given the discrepancy between class scores. Nonetheless, the period analyzed is characterized by an abnormal amount of companies' valuation growth motivated by a risk-free rate of 0%. Such conditions can partly explain such a considerable concentration of positive news.

To better understand the results and possible problems with the classifications, we explore some of the top and bottom classified headlines for Amazon, Table A.3. The following observations are drawn:

1) Some news articles are not directly related to the company and are still identified by Reuters as being indirectly linked. This is the case for industry and economy-level headlines (e.g., "U.S. FTC meeting will scrutinize Big Tech's small deals"). As Reuters will presumably associate these headlines with the other companies in the sector or part of the economy, we do not expect their presence to inflate the sentiments of each company. Instead, they ensure that some systematic information is included beyond the intrinsic one.

2) There appears to be some inexactness in the Reuters entity recognition mechanism, as shown by the headline "Refinitiv Newscasts - Ecuador oil pipeline ruptures in the Amazon", which highlights a case where the Amazon forest was recognized as the company.

Having looked at the news distribution, it is also important to explore statistics on the daily aggregate signals. Tables 5.3 and 5.4 present the daily sentiment by company grouped by deciles of media coverage and GICS, respectively.

Decile	Mean	Median	Skewness	Kurtosis
First	0.31	0.32	-0.60	0.17
Second	0.31	0.31	-0.54	0.23
Third	0.31	0.31	-0.51	0.19
Fourth	0.30	0.30	-0.62	0.19
Fifth	0.31	0.32	-0.37	0.21
Sixth	0.31	0.31	-0.05	0.23
Seventh	0.29	0.29	0.09	0.21
Eight	0.28	0.28	-0.25	0.19
Night	0.29	0.29	0.63	0.19
Tenth	0.27	0.27	0.11	0.16

Table 5.3: Average Daily Sentiment by Decile of Daily Coverage

Sentiment is very similar across all deciles, with less-covered companies showing, on average, a slightly higher daily sentiment than those more covered. Another insight is that companies in the extreme deciles have a lower (Fisher) kurtosis and fewer outliers. However, the difference is minimal, with all deciles having moderate tails.

GICS Sector	Mean	Median	Skewness	Kurtosis
Information Technology	0.32	0.33	-0.86	0.17
Health Care	0.32	0.32	-0.20	0.24
Industrials	0.30	0.31	-1.11	0.16
Utilities	0.30	0.31	-0.75	0.26
Real Estate	0.31	0.31	-0.17	0.25
Financials	0.29	0.30	-0.41	0.21
Materials	0.30	0.30	0.12	0.23
Communication Services	0.27	0.28	-0.83	0.16
Consumer Discretionary	0.28	0.28	0.03	0.19
Consumer Staples	0.27	0.28	-0.27	0.21
Energy	0.26	0.26	1.87	0.19

Table 5.4: Average Daily Sentiment by GICS

Sentiment across sector shows no significant differences, with the highest sentiment sector, "Information Technology" (0.32), differing from the smallest one, "Energy" (0.26), by 0.06.

Figure 5.3 shows the time series of average weekly sentiment, S&P500 returns and news volume over the sample period. Weekly data is chosen for better readability of the graph.

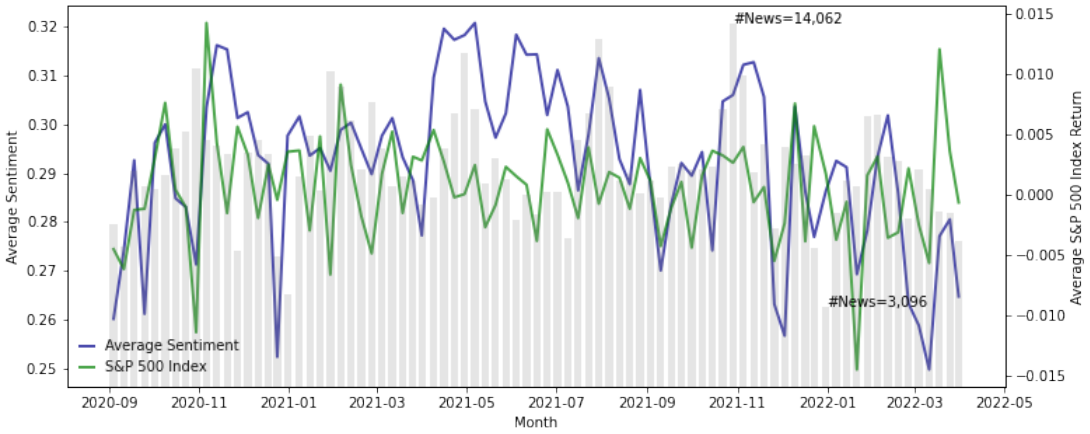

Figure 5.3: Weekly average Sentiment, S&P 500 Returns and News Volume

Average weekly sentiment for all companies seems to be positively correlated with the S&P500 returns and the news volume. Nonetheless, sentiment movements are much more pronounced, as seen right before January of 2021, when sentiment fell to some of its lowest levels, while the S&P saw just a tiny shock. One can speculate from this that there is a tendency for news to sensationalize occurrences, which rational investors are aware of and take precautions. Moreover, the sentiment seems to be translated into returns rather quickly, which might not leave room for taking advantage of sentiment signals outside intra-day trading. In terms of news volume, there seems to be a notably higher coverage around September, presumably due to fiscal quarter announcements, while at the beginning of the year following New Year's Eve, there is less activity.

5.2 Economic and Investment Analysis

5.2.1 Speed of Information Assimilation

To investigate the speed of information assimilation, we analyze the association between news sentiment on day *t* with returns on day *t-1*, day *t* and day *t+1*. This is achieved by creating EWportfolios with (up to) 100 stocks built on the *t* day's sentiment.

While the relation between *t* sentiment and *t+1* returns represents an implementable strategy, further analyzed in the following subsections, the other two associations have a different interpretation. Due to the timing of the news signal, a trader cannot take positions and exploit the returns, unless for the case of returns at *t* through high-frequency trading, which would, however, carry considerations not accounted for in our analysis. Their inclusion is simply a way to explore the correlation between returns and sentiment from the perspective of economic return units. Table 5.5 and Figure 5.4 highlight visually and through summary statistics this relation for the portfolios created.

Formation	Sharpe Ratio	Turnover	Average Return	Max Drawdown	FF3		FF5		FF5+MOM	
					α	R2	α	R2	α	R2
				Day -1						
L	0.39**	74.46	1.39	-6.43	0.17	90.46	0.19	90.90	0.16	91.77
S	0.12	91.29	0.56	-29.08	1.99	77.91	1.90***	78.67	1.96***	80.98
L-S	0.87***	86.84	1.18	-2.34	0.90***	41.82	0.87***	44.27	0.87***	44.28
				Day 0						
L	0.60***	74.45	2.16	-6.56	1.01***	90.08	1.05***	90.81	1.01***	92.07
S	0.57***	91.54	2.53	-8.60	3.92***	80.60	3.85***	81.05	3.89***	82.13
L-S	1.79***	86.94	2.33	-1.04	2.13***	34.57	2.10***	36.70	2.10***	36.77
				Day +1						
L	0.30*	74.52	1.1	-7.0	-0.08	89.9	-0.04	90.8	-0.08	92.2
S	-0.24	91.70	-1.1	-47.3	0.30	79.5	0.21	80.2	0.28	83.4
L-S	0.22	87.1	0.3	-4.1	0.06***	40.3	0.04	42.3	0.04	42.4

***, ** and * denote p-values <0.01, <0.05 and <0.10, respectively

Table 5.5: Price Responses On Days -1, 0 and +1

The Day-1 strategy (dashed line) displays the relation between news sentiment and previous day returns. It quantifies the extent to which sentiment picks up on stale news. Given the moderately good Sharpe ratios, there seems to be a significant portion of daily news containing repeated or known information by market participants.

The Day 0 strategy (solid line) displays the relation between news sentiment and returns on the same day. It shows how well the sentiment score captures fresh, value-relevant news not previously incorporated into prices. As shown by the highest Sharpe ratios compared to the other day lags, on average, price responses are most concentrated on the same day that news arrives.

The Day+1 strategy (dotted line) displays the relation between news sentiment and subsequent day returns. It captures the extent to which information in the sentiment score is incorporated with a delay, consisting of the implementable trading strategy further analyzed in Section 5.2.2. Given the metrics shown, there seems to be some information in the news not fully incorporated into market prices instantly. However, this is relatively small and exists mainly for positive signals.

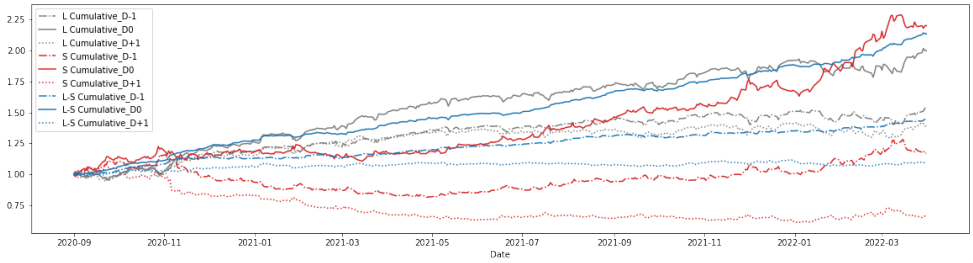

Figure 5.4: Cumulative returns in Days -1, 0 and +1

Regarding the type of strategy (L, S or L-S), the S portfolio has a lower performance across all three relationships analyzed, having, however, a similar Sharpe ratio to the L portfolio and the highest returns on Day 0 strategies. This suggests that negative news is more quickly incorporated into market prices and that stale news does not carry as much value. A possible interpretation of this might be that in a period characterized by "free money" and investment resulting from government incentives and lower interest rates, the overall market expectation was for positive news and better performance of companies. As such, negative news that contradicted the trend prompted quicker adjustments. In other words, good news has little impact following a period of superior market performance, while bad ones cause a more prominent reaction. David (1997) and Veronesi (1999) also observed this phenomenon for the aggregate market and justified it based on two reasons: (i) bad news cause investors to question the good state of the market; (ii) increased uncertainty leads risk-averse investors to require higher returns, increasing discount rates.

It is also to note that the performance of portfolio S deteriorates sharply in Day +1, being the only occurrence from the table with negative returns. This can be understood as the existence of an overreaction to news as soon as they are published, which is corrected by the next day. The presence of what appears to be a mean reversion could potentially be exploited by taking long positions following negative signals.

Focusing on the L-S and L portfolios, at Day -1 and Day 0 strategies, they attain very sim-

ilar average returns, differing in their volatility which puts the L-S portfolio at a Sharpe ratio of approximately 2.5 times higher than the L portfolio. This lower volatility can possibly be explained by the close-to-neutral nature of the L-S portfolio. This trend is only broken at Day +1 strategy, with the returns of the L-S portfolio being drastically smaller, conceivably driven by the short positions in the portfolio.

Finally, in terms of alpha returns, only on Day 0 are there significantly positive values for all portfolios after adjusting for well-known risk factors, which implies that investment on the same day can yield economic profits beyond systematic risk-adjusted returns. However, much of this value seems to be forgone in Day +1, where only the L-S portfolio achieves a significant positive alpha when controlling for the Fama-French factors and relatively low r-squared. Notwithstanding, the L portfolio holds for Day +1 the highest Sharpe ratio, which is significantly different from zero and above the benchmark.

5.2.2 Tradable Portfolio Strategies

This subsection focuses on the main strategies created based on the *t-1* sentiment score, as defined in the methodology. These are the L, S and L-S portfolios presented as EW and VW and constructed to be implementable strategies. The relevant results are presented in Table 5.6, from which four main insights can be drawn out.

Formation	Sharpe Ratio	Turnover	Average Return	Max Drawdown	FF3		FF5		FF5+MOM	
					α	R2	α	R2	α	R2
EW L	0.30*	74.52	1.1	-7.0	-0.08	89.9	-0.04	90.8	-0.08	92.2
EW S	-0.24	91.70	-1.1	-47.3	0.30	79.5	0.21	80.2	0.28	83.4
EW L-S	0.22	87.1	0.3	-4.1	0.06***	40.3	0.04	42.3	0.04	42.4
VW L	0.29*	76.06	0.99	-7.92	0.05	87.97	0.06	88.90	0.04	89.33
VW S	-0.07	93.01	-0.27	-28.04	0.76**	76.72	0.72**	77.07	0.73**	77.14
VW L-S	0.45**	87.64	0.75	-4.24	0.55**	24.17	0.54**	25.17	0.54**	25.24
Benchmark	0.24	-	0.84	-10.21	-0.03	99.10	-0.06	99.47	-0.06*	99.50

***, ** and * denote p-values <0.01, <0.05 and <0.10, respectively

Table 5.6: Basic Investment Strategies Performance

First, the performance, as seen by the Sharpe ratio, shows no consistent difference between the VW and EW. One could argue that the VW are slightly higher on average, but the difference is not enough to infer if the predictive power of sentiment is more substantial for small or large stocks.

Second, the L portfolio has a significant positive Sharpe ratio (0.34 EW and 0.29 VW), while the S does not (-0.24 EW and -0.07 VW). This is partly due to the market equity risk premium that the long side naturally earns while the short side pays. However, as highlighted in

the previous subsection analysis, the negative performance of the short side seems to stem from the quick integration of sentiment information in the given day, followed by an adjustment in the next.

Third, only the L-S portfolios have a low exposure to common aggregate risk factors, which have at most 43% daily r-squared when regressed on Fama-French factors, making it so that a large portion of the average returns is alpha. Specifically, the alphas are more consistently positive and significant for the VW portfolio.

Finally, despite the bad performance of the S portfolios, all others present a Sharpe ratio close to or higher than the S&P500 benchmark. Namely, the L portfolios attain significantly higher returns than the benchmark but are penalized by higher volatility. In contrast, the L-S portfolios have less volatility, given their closeness to a zero-net investment, but have fewer returns. Besides the volatility values as defined by the standard deviation, the max drawdown for L-S and L portfolios is low, which suggests that despite having more frequent fluctuations, very few are extreme ones in a negative direction.

Overall some of the portfolios achieve a better performance than the benchmark. However, the difference in performance is relatively small. Moreover, the created portfolios have very high levels of turnover, meaning high transaction costs, in opposition to the benchmark, which is passively held. For a fair performance comparison, costs need to be included in determining if some strategies can remain profitable and do so enough to compensate for the need for active management of the portfolio.

5.2.3 Alternative Portfolio Strategies

To investigate the sensitivity of our results to varying some of the portfolio implementations, we explore some alternative portfolios, shown in Table 5.7.

	Trading 200 stocks				**Follow the Loser**			
Formation	Turnover	Sharpe Ratio	Average Return	Max Drawdown	Turnover	Sharpe Ratio	Average Return	Max Drawdown
EW L	53.31	0.38**	1.34	-7.47	74.74	-0.30*	-1.11	-40.32
EW S	91.69	-0.24	-1.07	-47.50	91.54	0.24	1.07	-13.13
EW L-S	78.16	0.31*	0.66	-4.54	87.09	-0.23	-0.31	-11.87
VW L	53.29	0.34**	1.17	-9.30	76.15	-0.29*	-0.99	-39.71
VW S	93.01	-0.07	-0.27	-28.04	92.89	0.07	0.27	-17.23
VW L-S	79.23	0.37**	0.77	-5.10	87.64	-0.45**	-0.75	-23.29

***, ** and * denote p-values <0.01, <0.05 and <0.10, respectively

Table 5.7: Alternative Investment Strategies Performance

The basic portfolios are adjusted to allow (up to) 200 stocks, which impacts mainly long positions as the number of stocks with negative news is limited. As for the L-S and L portfolios, the

increase in stocks causes an increase in returns without significantly affecting volatility in terms of standard deviation and maximum drawdown, likely due to the higher degree of diversification it provides.

Next, motivated by the discoveries on the speed of information assimilation for the different types of signals, results are presented for portfolios built on the follow the looser approach. Specifically, we short positions with positive signals and vice-versa. As expected, the S portfolio achieved good returns, especially for the EW one, which can be interpreted as smaller stocks having larger overreactions followed by corrections when faced with negative news. However, despite the sound returns, the Sharpe ratio is penalized by more fluctuations in values creating more risk.

5.2.4 Top Strategies With Trading Costs

Thus far, the analysis ignored the existence of transaction costs despite their indispensability for assessing a strategy´s profitability, particularly those with high turnover. To understand the practical viability of the strategies explored, we present in Table 5.8 the gross and net returns of some of the best strategies. Other strategies mentioned in the previous sections were also evaluated with costs, showing no noteworthy difference from the results presented.

	Gross		**Net**	
Formation	Return	Sharpe Ratio	Return	Sharpe Ratio
EW L	1.34	0.38**	0.07	0.02
EW S	-1.07	-0.24	-3.26	-0.74***
EW L-S	0.66	0.31*	-1.22	-0.57***
VW L	1.17	0.34**	-0.11	-0.03
VW S	-0.27	-0.07	-2.49	-0.64***
VW L-S	0.77	0.37**	-1.13	-0.54***

***, ** and * denote p-values <0.01, <0.05 and <0.10, respectively

Table 5.8: Gross and Net Performance of Portfolios with 200 companies

The cost considerations lead to a drop in profitability for the L-S portfolios, which are slightly more pronounced in the EWportfolios. Nonetheless, regardless of the strategy employed, introducing costs makes it so that there is practically no viable strategy. Only the EWL-S portfolio, when constructed on 200 companies, manages to retain positive returns, which are negligible and non-significant, making the strategy unattractive.

Trading costs need to be minimized by reducing turnover for the strategies to be feasible. To achieve this, we apply the EWCT optimized method for adjusting the portfolios each day,

defined in the methodology, for which the results for the L portfolio are shown in Table 5.9 for varying turnover limits (γ) between 0.1 and 0.9.

	L	**Gross**		**Net**	
λ	Turnover	Return	Sharpe Ratio	Return	Sharpe Ratio
0.1	5.6	0.64	0.36**	0.50	0.28***
0.2	11.0	0.73	0.37**	0.47	0.23
0.3	16.3	0.82	0.37**	0.42	0.19
0.4	21.6	0.90	0.38**	0.38	0.16
0.5	26.9	0.97	0.38**	0.33	0.13
0.6	32.2	1.05	0.38**	0.27	0.10
0.7	37.4	1.12	0.38**	0.22	0.07
0.8	42.7	1.19	0.38**	0.17	0.05
0.9	47.9	1.27	0.38**	0.12	0.04

***, ** and * denote p-values <0.01, <0.05 and <0.10, respectively

Table 5.9: Gross and Net Performance of Optimized Portfolio L with 200 companies

As the limit to turnover increases, the gross return and Sharpe ratio of the strategies increase. However, this change is much smaller in the Sharpe ratio, suggesting that an increase in risk also accompanies it. In contrast, the net values experience a decrease in reward as the limit to turnover increases. This behaviour is contrary to expectations that returns would decrease with lower (γ) due to a loss of predictive information as the signal is smoothed out. Instead, it seems that there is value to gain from the momentum of sentiment. Thereby, the restriction imposed on turnover not only benefits the returns but also reduces the costs, allowing for a positive Sharpe ratio.

Despite the improvements achieved with the EWCT optimized method, only the L portfolios benefit enough to achieve positive net results, which when the turnover limit is lower peaks to values closer to our benchmark. This conclusion remains true when the number of companies varies and the initial portfolio changes. Specifically, tests were made with 100 and 200 companies as well as with a starting EWportfolio defined by the first set of sentiment signals and one with the entire universe of stocks. The latter was specifically tested to determine if the initial signal, the only one fully incorporated, had a significant impact on the further adjustments made.

5.3 Daily Sentiment Robustness

To ensure that the economic findings remain valid for different interpretations of sentiment, robustness tests are led by following other proposed methods of sentiment aggregation.

The explored aggregation until now consists of an average, which is easily affected by outliers and ignores the total amount of news. The first alternative sentiment aggregation, method explored is the relative measure proposed by (Antweiler and Frank, 2004) expressed as:

$$Sent^{*}_{i,t} = \ln\left(\frac{1 + HL_{Sent}pos_{i,t}}{1 + HL_{Sent}neg_{i,t}}\right) \tag{5.1}$$

where $HL_{Sent}pos_{i,t}$ and $HL_{Sent}neg_{i,t}$ represent the sentiment of positive and negative news headlines for firm *i* on day *t*.

Note that using the natural logarithm in this method also reduces the impact of excessively large headline volumes.

The second method is an absolute sentiment which also captures the news volume but is more sensitive to large concentrations of news on a given side of the polarity. Following (Antweiler and Frank, 2004) and (Souza et al., 2015) it can be presented as:

$$Sent^{**}_{i,t} = HL_{Sent}pos_{i,t} - HL_{Sent}neg_{i,t} \tag{5.2}$$

where the variables can be read the same as in the other equation.

The main results from running statistics with these alternative sentiment measures can be found in the Appendix. The values obtained for both proxies support previous findings, highlighting an even higher chance for profitability, hinting that the news volume is essential information. More comprehensively, the following differences applicable to both proxy measures can be highlighted:

I) The superior performance on Day 0 is reinforced, but it is also shown a higher chance for profitability for the L and L-S portfolios on the two other day shifts analyzed.

II) In the basic strategies, the Sharpe ratio of the L and L-S portfolios saw an improvement regarding the EWallocation, rising to statistically significant values of at least 1.5 times higher than the benchmark. Similarly, we report higher significantly positive alphas for these portfolios, except for the L portfolio built on the absolute proxy for which the positive alpha is not statistically significant. On the other hand, the VW portfolios remain

mostly the same, with the L-S suffering a reduction in the Sharpe ratio but retaining positive alpha returns. In contrast to our ambiguous results when comparing EWand VW built on the original sentiment, these proxies reinforce the idea that there is a superior predictive power for smaller stocks, supporting discoveries by (Tetlock, 2007).

III) In the strategies when accounting for costs, there is also a slight improvement in the L portfolios, which display positive Sharpe ratios for all the weight allocation types and number of companies tested. Moreover, when the method for turnover optimization is applied there are also L-S portfolios that, for low (γ) values, can achieve profitability close to that of the benchmark.

Chapter 6

Conclusion

This dissertation aims to adapt a state-of-the-art text classification algorithm to produce actionable sentiment signals from news and analyze the economic viability of strategies built on said signals. For that purpose, we retrieve a sample of S&P500 constituents' returns from WRDS and headlines through the TR-EK service between 1st September 2020 and 31st March 2022.

Concerning the classification model, it is found that a plethora of models can achieve good classification results with an acceptable degree of accuracy when trained for binary classification. This seems to be partly due to the use of pre-trained models, which are reinforced as versatile models capable of being fine-tuned on small data while preserving a good performance. Moreover, beyond the binary classification, using probabilities associated with the classification offers some value by retaining the nature of the signal (positive or negative) while capturing its degree of certainty. These can be leveraged in creating portfolios by allowing thresholds and/or evaluating texts relative to one another. Resultingly, instead of reacting to all binary signals equally, there can be more selectivity and control over the number used.

We find that sentiment is assimilated into returns relatively quickly, with some value still holding into the day after the publication of the headline. This is particularly true for the L-S portfolio, especially when built on the alternative sentiment measures, where we see significantly positive alpha and reduced r-squared. The latter highlights that the frequency of news is also an essential component.

Additionally, there is evidence that negative news is reacted to more intensively, requiring adjustments on the following day, which does not occur for the other portfolios. Nonetheless, even when implementing a follow the loser approach for the S portfolio, it still has a lower average performance than the others, suggesting that the next day's price correction is minor.

In terms of profitability, we found daily adjusted portfolios built on our defined measure of sentiment to achieve a performance close to that of our benchmark and, in some cases, beat it when ignoring transaction costs. However, the inclusion of costs, as is, drives practically all tested strategies to negative return levels due to their high turnover. To solve this problem, methods such as the EWCT , which limits the daily rate of turnover, are found to significantly control costs while creating a reward system for good performers and penalizing volatile ones. This double benefit from the strategy contrasts with the original authors' findings, which observe a trade-off between costs and returns.

In the end, applying the EWCT allows for Sharpe ratios above the benchmark for the L portfolio. However, this may stem from the market economic condition, which was bullish for the period analyzed. In contrast, the L-S portfolio with a low turnover retains positive risk-adjusted returns and a significant amount of returns from the sentiment but falls slightly below the S&P500.

Limitations

In terms of the data used, the data employed to train the ML model consists of a limited amount of examples with a slight imbalance, increasing the likelihood of overfitting or bias. Given the challenge in finding good quality large samples of public labelled data, one possible avenue to surpass this obstacle is by exploring models that use frameworks where the labels are the returns instead of a human-given score based on the text tone.

On the other hand, the OS (non-labelled) data is a larger pool of news data but has constraints regarding the market on which it focuses and the timeline used. The timeline limitations are particularly harmful as our data capture mainly an expansion period, not allowing us to understand how the strategies hold following market changes. Enhanced research could explore and compare different markets in terms of country or asset or extend the sentiment source to other mediums, such as social media or corporate documents. This latter path would potentially offer the advantage of fewer limitations to the availability of data in terms of size and period range.

Second, to convert headlines to an actionable numerical signal, different textual processing methodologies, part of the NPL field, are tested with a focus on pre-trained models. Within NLP, many more models and tools exist which are worth exploring, one case being techniques

around NER and grammatical relation detection. The latter can be used in news cases mentioning multiple companies with opposing polarity, extending research beyond the company or macro level.

Third, different ways to leverage sentiment signals from an economic standpoint can be explored. Two possible paths are: (a) different frequency of calibration of the strategies, which can help surpass the high costs of daily adjustments; (b) utilization of more advanced models, such as ranking ML models, which can be used to get a better stock selection based on different sentiment proxies and complementary variables capturing aspects like economic period (expansion or contraction).

Headmost, one of the focuses of this dissertation is on exploring the validity of using sentiment extracted from firm news in creating profitable investment strategies. However, the use of only sentiment as a signal is limited. Instead, it might provide more value when used with other relevant variables (e.g. analyst recommendations, returns, trading volume).

Bibliography

Aggarwal, C. C. and Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer.

Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in economics*, pages 235–251. Elsevier.

Antweiler, W. and Frank, M. (1998). Do us stock markets typically overreact to corporate news stories? ssrn.

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of finance*, 59(3):1259–1294.

Antweiler, W. and Frank, M. Z. (2006). Do us stock markets typically overreact to corporate news stories? *Available at SSRN 878091*.

Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The journal of Finance*, 61(4):1645–1680.

Black, F. (1986). Noise. *The journal of finance*, 41(3):528–543.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

Bouazizi, M. and Ohtsuki, T. (2016). Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in twitter. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., and Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215.

Champagne, C., Karoui, A., and Patel, S. (2018). Portfolio turnover activity and mutual fund performance. *Managerial Finance*.

Chan, L. K. and Lakonishok, J. (1997). Institutional equity trading costs: Nyse versus nasdaq. *The Journal of Finance*, 52(2):713–735.

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

Christmann, A. and Steinwart, I. (2008). Support vector machines.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., and Davis, B. (2017). Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. Association for Computational Linguistics (ACL).

Cremers, M., Petajisto, A., and Zitzewitz, E. (2012). Should benchmark indices have alpha? revisiting performance evaluation. Technical report, National Bureau of Economic Research.

Cutler, D. M., Poterba, J. M., and Summers, L. H. (1988). What moves stock prices?

David, A. (1997). Fluctuating confidence in stock markets: Implications for returns and volatility. *Journal of Financial and Quantitative Analysis*, 32(4):427–462.

de Oliveira Carosia, A. E., Coelho, G. P., and da Silva, A. E. A. (2021). Investment strategies applied to the brazilian stock market: a methodology based on sentiment analysis with deep learning. *Expert Systems with Applications*, 184:115470.

DellaVigna, S. and Pollet, J. M. (2009). Investor inattention and friday earnings announcements. *The Journal of Finance*, 64(2):709–749.

Dewally, M. (2003). Internet investment advice: Investing with a rock of salt. *Financial Analysts Journal*, 59(4):65–77.

Domowitz, I., Glen, J., and Madhavan, A. (2001). Liquidity, volatility and equity trading costs across countries and over time. *International Finance*, 4(2):221–255.

Engelberg, J. (2008). Costly information processing: Evidence from earnings announcements. In *AFA 2009 San Francisco meetings paper*.

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

Ferguson, N. J., Philip, D., Lam, H., and Guo, J. M. (2015). Media content and stock returns: The predictive power of press. *Multinational Finance Journal*, 19(1):1–31.

Feuerriegel, S. and Prendinger, H. (2016). News-based trading strategies. *Decision Support Systems*, 90:65–74.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Fong, W. M. and Yong, L. H. (2005). Chasing trends: recursive moving average trading rules and internet stocks. *Journal of Empirical Finance*, 12(1):43–76.

Frazzini, A., Israel, R., and Moskowitz, T. J. (2018). Trading costs. *Available at SSRN 3229719*.

Gagnon, S. (2012). Rules-based integration of news-trading algorithms. *The Journal of Trading*, 8(1):15–27.

Gan, B., Alexeev, V., Bird, R., and Yeung, D. (2020). Sensitivity to sentiment: News vs social media. *International Review of Financial Analysis*, 67:101390.

Gao, B. and Liu, X. (2020). Intraday sentiment and market returns. *International Review of Economics & Finance*, 69:48–62.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300.

Gidofalvi, G. and Elkan, C. (2001). Using news articles to predict stock price movements. *Department of Computer Science and Engineering, University of California, San Diego*, page 17.

Graf, F. (2011). Mechanically extracted company signals and their impact on stock and credit markets.

Grefenstette, G. and Tapanainen, P. (1994). What is a word, what is a sentence?: problems of tokenisation.

Hagenau, M., Liebmann, M., and Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, 55(3):685–697.

Harris, R. D. and Yilmaz, F. (2009). A momentum trading strategy based on the low frequency component of the exchange rate. *Journal of Banking & Finance*, 33(9):1575–1585.

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *The Journal of Business Communication (1973)*, 45(4):363–407.

Heston, S. L. and Sinha, N. R. (2017). News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts Journal*, 73(3):67–83.

Hillert, A., Jacobs, H., and Müller, S. (2014). Media makes momentum. *The Review of Financial Studies*, 27(12):3467–3501.

Hochreiter, R. (2016). Computing trading strategies based on financial sentiment data using evolutionary optimization. In *International Conference on Soft Computing-MENDEL*, pages 181–191. Springer.

Jegadeesh, N. and Wu, D. (2013). Word power: A new approach for content analysis. *Journal of financial economics*, 110(3):712–729.

Jiang, M., Lan, M., and Wu, Y. (2017). Ecnu at semeval-2017 task 5: An ensemble of regression algorithms with effective features for fine-grained sentiment analysis in financial domain. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 888–893.

Kalyani, J., Bharathi, P., Jyothi, P., et al. (2016). Stock trend prediction using news sentiment analysis. *arXiv preprint arXiv:1607.01958*.

Katayama, D. and Tsuda, K. (2020). A method of using news sentiment for stock investment strategy. *Procedia Computer Science*, 176:1971–1980.

Ke, Z. T., Kelly, B. T., and Xiu, D. (2019). Predicting returns with text data. Technical report, National Bureau of Economic Research.

Kearney, C. and Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185.

Kelly, S. and Ahmad, K. (2018). Estimating the impact of domain-specific news sentiment on financial assets. *Knowledge-Based Systems*, 150:116–126.

Kim, D. and Lee, Y. (2018). News based stock market sentiment lexicon acquisition using word2vec. *The Journal of Bigdata*, 3(1):13–20.

Koratamaddi, P., Wadhwani, K., Gupta, M., and Sanjeevi, S. G. (2021). Market sentiment-aware deep reinforcement learning approach for stock portfolio allocation. *Engineering Science and Technology, an International Journal*, 24(4):848–859.

Li, F. (2010). The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102.

Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., and Chen, Y. (2014). The effect of news and public mood on stock movements. *Information Sciences*, 278:826–840.

Li, Y., Wu, J., and Bu, H. (2016). When quantitative trading meets machine learning: A pilot survey. In *2016 13th International Conference on Service Systems and Service Management (ICSSSM)*, pages 1–6. IEEE.

Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

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):35–65.

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Makrehchi, M., Shah, S., and Liao, W. (2013). Stock prediction using event-based sentiment analysis. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 337–342. IEEE.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., and Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Marty, T., Vanstone, B., and Hahn, T. (2020). News media analytics in finance: a survey. *Accounting & Finance*, 60(2):1385–1434.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Milgrom, P. and Stokey, N. (1982). Information, trade and common knowledge. *Journal of economic theory*, 26(1):17–27.

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*, 8:131662–131682.

Nofsinger, J. R. (2005). Social mood and financial economics. *The Journal of Behavioral Finance*, 6(3):144–160.

Palmer, M. and Schäfer, T. (2020). Leveraging textual analyst sentiment for investment. In *International Workshop on Enterprise Applications, Markets and Services in the Finance Industry*, pages 59–74. Springer.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Powell, M. J. (1987). Radial basis functions for multivariable interpolation: a review. *Algorithms for approximation*.

Prosky, J., Song, X., Tan, A., and Zhao, M. (2017). Sentiment predictability for stocks. *arXiv preprint arXiv:1712.05785*.

Radinsky, K., Davidovich, S., and Markovitch, S. (2012). Learning to predict from textual data. *Journal of Artificial Intelligence Research*, 45:641–684.

Rong, X. (2014). word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*.

Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY.

Russell, S. J. (2010). *Artificial intelligence a modern approach*. Pearson Education, Inc.

Sankar, H. and Subramaniyaswamy, V. (2018). Hybrid sentiment classification of reviews using synonym lexicon and word embedding. *International Journal of Pure and Applied Mathematics*, 119(12):13297–13308.

Schumaker, R. P. and Chen, H. (2009). A quantitative stock prediction system based on financial news. *Information Processing & Management*, 45(5):571–583.

Schumaker, R. P., Zhang, Y., Huang, C.-N., and Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, 53(3):458–464.

Shiller, R. C. (2000). Irrational exuberance. *Philosophy and Public Policy Quarterly*, 20(1):18–23.

Shiller, R. J. (1980). Do stock prices move too much to be justified by subsequent changes in dividends?

Sinha, N. R. (2016). Underreaction to news in the us stock market. *Quarterly Journal of Finance*, 6(02):1650005.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Song, Q., Liu, A., and Yang, S. Y. (2017). Stock portfolio selection using learning-to-rank algorithms with news sentiment. *Neurocomputing*, 264:20–28.

Souma, W., Vodenska, I., and Aoyama, H. (2019). Enhanced news sentiment analysis using deep learning methods. *Journal of Computational Social Science*, 2(1):33–46.

Souza, T. T. P., Kolchyna, O., Treleaven, P. C., and Aste, T. (2015). Twitter sentiment analysis applied to finance: A case study in the retail industry. *arXiv preprint arXiv:1507.00784*.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

Sun, L., Najand, M., and Shen, J. (2016). Stock return predictability and investor sentiment: A high-frequency perspective. *Journal of Banking & Finance*, 73:147–164.

Takala, P., Malo, P., Sinha, A., and Ahlgren, O. (2014). Gold-standard for topic-specific sentiment analysis of economic texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2152–2157.

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

Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The journal of finance*, 63(3):1437–1467.

Tumarkin, R. and Whitelaw, R. F. (2001). News or noise? internet postings and stock prices. *Financial Analysts Journal*, 57(3):41–51.

Uhl, M. W. (2014). Reuters sentiment and stock returns. *Journal of Behavioral Finance*, 15(4):287–298.

Uhl, M. W., Pedersen, M., and Malitius, O. (2015). What's in the news? using news sentiment momentum for tactical asset allocation. *The Journal of Portfolio Management*, 41(2):100–112.

Vajjala, S., Majumder, B., Gupta, A., and Surana, H. (2020). *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*. O'Reilly Media.

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.

Veronesi, P. (1999). Stock market overreactions to bad news in good times: a rational expectations equilibrium model. *The Review of Financial Studies*, 12(5):975–1007.

Wang, C., Wang, T., Yuan, C., and Rong, J. Y. (2021). Learning to trade on sentiment. *Journal of Economics and Finance*, pages 1–16.

Yang, S. Y., Mo, S. Y. K., Liu, A., and Kirilenko, A. A. (2017). Genetic programming optimization for a sentiment feedback strength based trading strategy. *Neurocomputing*, 264:29–41.

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

Yazdani, S. F., Murad, M. A. A., Sharef, N. M., Singh, Y. P., and Latiff, A. R. A. (2017). Sentiment classification of financial news using statistical features. *International Journal of Pattern Recognition and Artificial Intelligence*, 31(03):1750006.

Zaharudin, K. Z., Young, M. R., and Hsu, W.-H. (2022). High-frequency trading: Definition, implications, and controversies. *Journal of Economic Surveys*, 36(1):75–107.

Zhang, W. and Skiena, S. (2010). Trading strategies to exploit blog and news sentiment. In *Fourth international aAAI conference on weblogs and social media*.

Appendix

Appendix A

Company	Title	Sentiment	Class
Shell	Shell and BG Shareholders to Vote on Deal at End of January	0.0	Neutral
Halfords	Halfords appoints Jonny Mason as new CFO	0.0	Neutral
Barclays	Barclays share price subdued as bank faces fresh forex probe	-0.373	Negative
HSBC	HSBC Posts Surprise Fourth-Quarter Pretax Loss of $858 Million	-0.756	Negative
Tesco PLC	FTSE ends lower on weaker miners, Tesco outperforms	0.402	Positive
Persimmon	Persimmon share price climbs on 23% rise in full-year revenue	0.975	Positive

Table A.1: SemEval-2017 Classified Headline Examples

News Source	Count	Share
Reuters News	371081	48.62%
Public Technologies	33986	4.45%
CNBC	32478	4.26%
Business Wire	29152	3.82%
Zacks.com	27795	3.64%
Other (168)	268677	35.21%
Total	763169	100%

Table A.2: Top News Sourcees

Date	Headline	Sentiment (as %)	Source
08/11/2021	amazon.com - in coming weeks, will begin process of developing second phase of bellevue 600 project	48.91%	RTRS
05/01/2021	MEDIA-Amazon-backed grocer More retail weighing India IPO at $5 billion value - Bloomberg News	48.88%	RTRS
28/10/2021	amazon.com inc <amzn.o>sees q4 sales $130 blnto 140 bln	48.72%	RTRS
21/01/2022	From online into the real world: Amazon opens its first clothing store. In pics	48.72%	MINTNE
21/09/2021	Amazon sets date for its big fall hardware launch event	48.70%	ASNEWS
03/10/2021	German labour union calls for strikes at seven Amazon locations	-48.69%	RTRS
08/09/2021	U.S. FTC meeting will scrutinize Big Tech's small deals	-48.72%	RTRS
30/01/2022	Refinitiv Newscasts - Ecuador oil pipeline ruptures in the Amazon	-48.79%	RTRS
05/06/2021	Activists block some Amazon warehouses on Black Friday EUROPE	-49.03%	ATLJUR
23/08/2021	MEDIA-Amazon Games Studio head Mike Frazzini steps down - Bloomberg News	-49.08%	RTRS

Table A.3: Amazon Headline Examples

Appendix B: Relative Sentiment Measure Results

Formation	Sharpe Ratio	Turnover	Average Return	Max Drawdown	FF3		FF5		FF5+MOM	
					α	R2	α	R2	α	R2
				Day -1						
L	0.57***	55.49	1.96	-6.62	0.76	93.70	0.77***	93.73	0.74***	94.35
S	0.12	91.29	0.56	-29.08	1.99***	77.91	1.90***	78.67	1.96***	80.98
L-S	1.24***	73.64	1.72	-1.43	1.43***	40.34	1.41***	40.70	1.41***	40.74
				Day 0						
L	0.89***	55.47	3.22	-5.30	2.03***	89.45	2.07***	89.70	2.03***	91.66
S	0.57***	91.54	2.53	-8.60	3.92***	80.60	3.85***	81.05	3.89***	82.13
L-S	2.89***	73.70	3.96	-0.73	3.73***	28.76	3.71***	29.24	3.71***	29.24
				Day +1						
L	0.41**	55.49	1.43	-6.7	0.30	92.8	0.32**	92.9	0.30**	93.6
S	-0.24	91.70	-1.06	-47.3	0.30***	79.5	0.21	80.2	0.28	83.4
L-S	0.51***	73.8	0.72	-4.9	0.48***	34.3	0.46**	34.9	0.47***	36.9

***, ** and * denote p-values <0.01, <0.05 and <0.10, respectively

Table B.4: Price Responses On Days -1, 0 and +1

Formation	Sharpe Ratio	Turnover	Average Return	Max Drawdown	FF3		FF5		FF5+MOM	
					α	R2	α	R2	α	R2
EW L	0.41**	55.49	1.43	-6.7	0.30	92.8	0.32**	92.9	0.30**	93.6
EW S	-0.24	91.70	-1.06	-47.3	0.30***	79.5	0.21	80.2	0.28	83.4
EW L-S	0.51***	73.8	0.72	-4.9	0.48***	34.3	0.46**	34.9	0.47***	36.9
VW L	0.29*	33.63	1.09	-10.25	0.33**	95.03	0.27**	95.63	0.27**	95.63
VW S	-0.07	93.01	-0.27	-28.04	0.76**	76.72	0.72**	77.07	0.73**	77.14
VW L-S	0.29*	50.47	0.91	-8.42	0.43**	82.31	0.36*	83.07	0.37*	83.13

***, ** and * denote p-values <0.01, <0.05 and <0.10, respectively

Table B.5: Basic Investment Strategies Performance

	Gross		Net	
Formation	Return	Sharpe Ratio	Return	Sharpe Ratio
EW L	1.28	0.36**	0.23	0.06
EW S	-1.07	-0.24	-3.26	-0.74***
EW L-S	0.93	0.46***	-0.59	-0.29*
VW L	0.98	0.27	0.42	0.12
VW S	-0.27	-0.07	-2.49	-0.64***
VW L-S	0.98	0.31*	0.08	0.03

***, ** and * denote p-values <0.01, <0.05 and <0.10, respectively

Table B.6: Gross and Net Performance of Portfolios with 200 companies

L		Gross		Net	
λ	Turnover	Return	Sharpe Ratio	Return	Sharpe Ratio
0.1	4.7	0.70	0.34*	0.59	0.28**
0.2	9.1	0.78	0.35**	0.56	0.25
0.3	13.4	0.86	0.35**	0.54	0.22
0.4	17.7	0.93	0.36**	0.51	0.20
0.5	21.9	1.00	0.36**	0.48	0.17
0.6	26.1	1.07	0.37**	0.44	0.15
0.7	30.2	1.13	0.37**	0.40	0.13
0.8	34.4	1.18	0.37**	0.36	0.11
0.9	38.6	1.24	0.37**	0.32	0.09

***, ** and * denote p-values <0.01, <0.05 and <0.10, respectively

Table B.7: Gross and Net Performance of Optimized L Portfolio with 200 companies

L-S		Gross		Net	
λ	Turnover	Return	Sharpe Ratio	Return	Sharpe Ratio
0.1	5.3	0.35	0.28	0.22	0.18***
0.2	10.9	0.41	0.31*	0.15	0.11
0.3	16.7	0.48	0.33*	0.08	0.05
0.4	22.6	0.54	0.35**	0.00	0.00
0.5	28.6	0.60	0.37**	-0.09	-0.06
0.6	34.6	0.65	0.38**	-0.18	-0.11
0.7	40.7	0.69	0.38**	-0.28	-0.16
0.8	46.8	0.73	0.39**	-0.39	-0.21
0.9	52.9	0.77	0.39**	-0.50	-0.25

***, ** and * denote p-values <0.01, <0.05 and <0.10, respectively

Table B.8: Gross and Net Performance of Optimized L-S Portfolio with 200 companies

Appendix C: Absolute Sentiment Measure Results

Formation	Sharpe Ratio	Turnover	Average Return	Max Drawdown	FF3		FF5		FF5+MOM	
					α	R2	α	R2	α	R2
				Day -1						
L	0.52***	50.74	1.80	-7.13	0.61	94.01	0.62***	94.02	0.60***	94.48
S	0.12	91.29	0.56	-29.08	1.99***	77.91	1.90***	78.67	1.96***	80.98
L-S	1.16***	67.49	1.61	-1.43	1.34***	41.15	1.31***	41.68	1.31***	41.68
				Day 0						
L	0.73***	50.72	2.65	-6.03	1.47***	89.70	1.52***	90.01	1.47***	92.03
S	0.57***	91.54	2.53	-8.60	3.92***	80.60	3.85***	81.05	3.89***	82.13
L-S	2.25***	67.55	3.39	-0.79	3.14***	28.93	3.14***	28.98	3.13***	29.63
				Day +1						
L	0.36**	50.72	1.28	-7.2	0.15	93.1	0.18	93.2	0.15	93.8
S	-0.24	91.70	-1.06	-47.3	0.30**	79.5	0.21	80.2	0.28	83.4
L-S	0.42**	67.6	0.60	-5.1	0.37***	36.9	0.35**	37.8	0.37**	40.1

***, ** and * denote p-values <0.01, <0.05 and <0.10, respectively

Table C.9: Price Responses On Days -1, 0 and +1

Formation	Sharpe Ratio	Turnover	Average Return	Max Drawdown	FF3		FF5		FF5+MOM	
					α	R2	α	R2	α	R2
EW L	0.36**	50.72	1.28	-7.2	0.15	93.1	0.18	93.2	0.15	93.8
EW S	-0.24	91.70	-1.06	-47.3	0.30**	79.5	0.21	80.2	0.28	83.4
EW L-S	0.42**	67.6	0.60	-5.1	0.37***	36.9	0.35**	37.8	0.37**	40.1
VW L	0.23	23.80	0.90	-11.54	0.13	95.93	0.07	96.34	0.07	96.34
VW S	-0.07	93.01	-0.27	-28.04	0.76**	76.72	0.72**	77.07	0.73**	77.14
VW L-S	0.24	32.44	0.83	-10.55	0.31*	87.00	0.24	87.72	0.24	87.79

***, ** and * denote p-values <0.01, <0.05 and <0.10, respectively

Table C.10: Basic Investment Strategies Performance

	Gross		Net	
Formation	Return	Sharpe Ratio	Return	Sharpe Ratio
EW L	1.29	0.37**	0.27	0.08
EW S	-1.07	-0.24	-3.26	-0.74***
EW L-S	0.81	0.39**	-0.61	-0.30*
VW L	0.96	0.26	0.47	0.13
VW S	-0.27	-0.07	-2.49	-0.64***
VW L-S	0.81	0.24	0.14	0.04

***, ** and * denote p-values <0.01, <0.05 and <0.10, respectively

Table C.11: Gross and Net Performance of Portfolios with 200 companies

	L	Gross		Net	
λ	Turnover	Return	Sharpe Ratio	Return	Sharpe Ratio
0.1	4.8	0.69	0.34*	0.58	0.28**
0.2	9.3	0.77	0.35**	0.55	0.25***
0.3	13.8	0.85	0.36**	0.52	0.22
0.4	18.2	0.92	0.36**	0.49	0.19
0.5	22.6	0.99	0.36**	0.45	0.17
0.6	26.9	1.05	0.36**	0.41	0.14
0.7	31.2	1.11	0.37**	0.37	0.12
0.8	35.5	1.17	0.37**	0.32	0.10
0.9	39.9	1.23	0.36**	0.27	0.08

***, ** and * denote p-values <0.01, <0.05 and <0.10, respectively

Table C.12: Gross and Net Performance of Optimized L Portfolio with 200 companies

	L-S	Gross		Net	
λ	Turnover	Return	Sharpe Ratio	Return	Sharpe Ratio
0.1	5.3	0.35	0.28	0.22	0.18**
0.2	10.9	0.41	0.31*	0.15	0.11
0.3	16.7	0.48	0.33*	0.08	0.05
0.4	22.6	0.54	0.35**	0.00	0.00
0.5	28.6	0.60	0.37**	-0.09	-0.06
0.6	34.6	0.65	0.38**	-0.18	-0.11
0.7	40.7	0.69	0.38**	-0.28	-0.16
0.8	46.8	0.73	0.39**	-0.39	-0.21
0.9	52.9	0.77	0.39**	-0.50	-0.25

***, ** and * denote p-values <0.01, <0.05 and <0.10, respectively

Table C.13: Gross and Net Performance of Optimized L-S Portfolio with 200 companies