ERASMUS UNIVERSITY ROTTERDAM

ERASMUS SCHOOL OF ECONOMICS

BSc Thesis: Economics & Business Economics

Major Financial Economics

Predicting the flip of a "fair" coin

A study on sentiment analysis and stock returns using machine learning classifiers

Author: Andrei Dumitrescu

Student number: 598088

Thesis supervisor: Dr. Amy Li

Second assessor: Dr. Clint Howard

Finish date: 27.06.2024

NON-PLAGIARISM STATEMENT

By submitting this thesis, the author declares to have written this thesis completely by himself/herself, and not to have used sources or resources other than the ones mentioned. All sources used, quotes and citations that were literally taken from publications, or that were in close accordance with the meaning of those publications, are indicated as such.

ABSTRACT

This research studies the ability of sentiment at predicting firms' short term stock returns through modern means of analysis. 1,2 million news articles from 2014-2024 are analyzed through the lenses of a natural-language-processing model (NLP), distil-ROBERTA, which are used to construct numerical sentiments aggregated at the firm, publisher and daily dimensions. Decision Tree and Random Forest classifiers are employed to find the non-linear relationships and patterns between the vector of sentiments and tomorrow's stock movement across 29 large U.S firms, achieving an average accuracy of 52% on out-of-sample data. Using this slight advantage over a random coinflip chance, a trading strategy employing the classifiers' predictions yields significant abnormal returns after altering models' confidence parameter. This research is the first to use the publisher as an aggregation element of sentiment and highlights the use of machine learning models to uncover the non-linear relationship between sentiment and stock market movements.

Keywords: Natural Language Processing Model, Sentiment analysis, Machine Learning Classifier, Inefficient Market, Trading Strategy.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1. Introduction

One of the masterminds behind the historic short squeeze of the GameStop stock, Keith Grill, was called for a committee hearing for realizing a potential market inefficiency that leading investment banks did not see themselves (U.S. House Committee on Financial Services Democrats, 2021). The lawsuits for releasing pertinent information with value implications from public information were ultimately dismissed (Business Insider, 2024), but his story raises an important issue about social media attention and its effects on stock prices. Media comments are defined and classified by Ahmad et al. (2016), into two categories: *sentiment*, which does not provide any correlations with stock returns, and *news*, which contain value-relevant information that may impact stock returns. For example, an article announcing the opening of a new office may be seen as *sentiment* when predicting stock returns, whereas an article mentioning a new lawsuit a firm is facing can be classified as *news*. Negative *news* tends to have a negative effect on stock prices and a self-financed portfolio can be built to have excess returns of 7.3% per year (Tetlock, 2007). Exploring the dynamics of media-expressed tone on firms' stock returns with the premise that prices should fully reflect a firm's value in an efficient capital market, could provide insights into potential inefficiencies and behavioral biases in the market's assessment of a firm's value.

Despite current literature being quite diverse in data used and types of methodologies they find similar findings of sentiment on stock returns. In evaluating sentiment, Hu et al. (2021) calculate the sentiment score as the average positive emojis on a given reddit post and use linear regression to find a significant effect of *submission tone* on tomorrow's stock price. Ahmad et al. (2016) use Loughran & McDonald's (2011) dictionary of 2300 negative finance words to construct a daily *negative tone* when parsing through 5.5 million news articles. Using a fixed effects model, they find significantly that the *negative tone*, up to the fifth lag, have predictive capabilities of tomorrow's returns which are like the findings of Khan et al. (2020), that used different machine learning models to evaluate the effect twitter sentiment may have on stock returns. These plausible findings and strong predictability of stock returns using social media attention, are summarized in a model by Pedersen (2022), with the explanation that in the short term, naïve investors are likely to act upon what they see in the news and ride the bubble. This bubble however has a destined reversal effect (Pedersen, 2022) which was present in the findings of Ahmad et al. (2016).

The methodology of the papers above can be split into two sections, calculating the sentiment and the method of evaluating its effect on stock returns. In terms of calculating the sentiment using a larger dictionary will increase the range of sentiment from negative to positive, opposed to negative to not-very-negative sentiment scale used by Ahmad et al. (2016); this also explains why their self-financing portfolio, compromised of shorting negative-talked about firms and longing the least-negative talked about firms, was mainly driven by the short leg of this strategy. In terms of evaluating the effectiveness of using sentiment as a predictor for stock returns, machine learning classifier models better capture non-linear relationships (Khan et al., 2020) than traditional linear or logistic regressions. Current literature lacks in the complexity of determining the sentiment as well as the models evaluating its effect on stock returns. By adding to the work of Ahmad et al. (2016) by employing state-of-the-art rule-based model for opinion mining and using supervised machine learning models (Khan et al., 2020), the research question is formulated as follows:

How and to what extent does sentiment tone derived from news articles impact firms' short term stock returns?

Inspired by the method of sample firms used by Ahmad et al. (2016), 30 non-financial companies including Apple, Microsoft, Verizon and Walmart will be evaluated from January 2014 to May 2024. The reason behind the time frame is to have the most current data and ensure a trading strategy that is applicable today; using older data to infer causal inferences may not be reliable. For the sample companies, articles from the ProQuest database will be downloaded in the specified time frame and the title will be used as the primary source of data for creating the sentiment. In answering the research question, the media-expressed tone will be evaluated by using a natural language processing model for financial sentiment analysis, distil-ROBERTA[1], where the range of sentiment lies between -1 and 1, where -1 is seen as very negative and +1 very positive sentiment.

Contrary to existing research, this sentiment score will be compiled at the publisher-daily level. Since the ProQuest article database features multiple publishers, such as Business Insider, Bloomberg Wire Service, Wall Street Journal, and Dow Jones Institutional News, aggregating sentiment scores daily could obscure valuable insights; some publishers may not significantly impact the predictability of stock returns and may skew results when aggregating at the daily dimension. This research paper is the first to assess the publisher dimension in the context of stock market prediction. The number of independent variables will be equal to the number of different publishers multiplied by 10 lags which will allow the Decision Tree Classifier and Random Forest Classifier (RFC), the best performing models at predicting stock movement (Khan et al., 2020), to tailor the weights of the independent variables, publisher's sentiment score, to each individual company analyzed. This is motivated by the findings of Khan et al. (2020) as certain stock markets may be more influenced (Microsoft) by news than others (Nokia, MSI), and Ahmad et al. (2016), that mention the need of machine learning models to analyze such effect. Additionally, the dependent

[1] The distil-ROBERTA model is tuned on financial news data set and has 4,445,280 downloads monthly. It uses 82.1 million parameters and is twice as fast as the ROBERTA base model due to its computational efficiencies and can be found on hugging-face, a free platform for lexicon sharing: https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis. It was developed by an entity named mrm8488, last updated in January 2024.

variable will be the same one used by (Khan et al., 2020) which is a binary classification of tomorrow's stock movement based on direction: a 0 would classify as *negative movement* and 1 would classify as *positive movement*.

The augmented methodology of using the NLP model to translate articles into sentiment scores and using machine learning models to answer the research question will provide insights about price discovery as well as the means of employing new and sophisticated models to create an arbitrage opportunity. This research will distinguish between predicting stock direction (Khan et al., 2020), and predicting magnitude of returns (Ahmad et al., 2016) based on a potential trading strategy, regressed on the Fama & French (1993) risk factors, using models' predictions; answering the research question will depend on the answers of these sections' results.

Under the Efficient Market Hypothesis (EMT) the stock market movement follows a Random Walk, unexplainable and as random as a coin toss (50% accuracy) in the short term. Across a two-year time frame, on average this research finds that the RFC yields 52% accuracy, in line with the accuracies of Khan et al. (2020). Relating to the coinflip analogy, on average the models trained offer an edge to investors that wish at predicting the stock direction the next day. Multiple portfolios were created including a long, short and long-short portfolios, where the 29 company-specific predictions regressed on the Fama & French (1993) risk factors, yield, under certain circumstances, significant abnormal returns over the time frame May 2022-May 2024.

The remainder of the paper is structured following these sections. Section 2 will discuss prior literature and their relevant findings while sections 3 and 4 will talk about the data and methodology employed in this research. Section 5 will present the results in the context of the two hypotheses which will be stated in section 2 in answering the research question. Section 6 will discuss the main findings and compare results with the literature presented prior along with the limitations of this paper. Section 7 will provide a short conclusion and summary of this research.

2. Theoretical Framework

The relationship between sentiment analysis and stock prices has been extensively studied over the past 20 years through different methods and techniques of both the predictor and the predicted. The theoretical framework will be split into three sections: sentiment analysis, stock returns and the relationship between the two. The main research across the fields will be chronologically explained to showcase how each of the three elements has developed over time but also why and through what means it has developed.

2.1 Sentiment analysis

The traditional definition of sentiment analysis is about the creation of an opinion spectrum of an object within loose but defined boundaries such as negative, neutral, and positive or emotional boundaries such as happiness and anger (Mäntylä et al., 2018). The first paper to study public opinion and thus begin the roots of sentiment analysis was by Stagner (1940), that assessed the population's opinion on World War 2. Sentiment analysis is rooted in the sociology field with survey- based methods given to the public or experts to understand the population's opinion on political matters. It is important to recognize that access to public opinion was difficult and only with the access of the world-wide-web did the topic of sentiment analysis exponentially grew in research (Mäntylä et al., 2018).

With growing access to data, different and better methods are also employed. The methods in sentiment analysis drastically change with computer technology advancements and opinion mining is developed. Opinion mining is synonymous with modern sentiment analysis through machine learning and NLP means in creating the sentiment spectrum (Pang & Lee, 2008). Tokenization (Guo, 1997) is the fundamental technique in developing NLP models, where lexical structures are translated into tokens or words, the unit of analysis of NLP models. Singular tokens do not have much meaning by themselves in sentences, the same way one cannot understand the context of a sentence given one word. This analogy can be exemplified by the title of Loughran & McDonald's (2011) paper "When is a liability not a liability?". Given the token "liability", this may have a negative connotation in normal colloquial speech but given financial vocabulary, the word "liability" is a word that can have different meanings depending on the context is in. The tokens given to an NLP model, or training data, to understand out-of-sample observations are subject to the context they are being analyzed in, and the size of the lexical structures, which is the reason why different lexicons, dictionaries and models exist in the opinion mining realm.

Endless possibilities of opinion mining arise in both methods employed and types of contexts. Yu & Hatzivassiloglou (2003) use a supervised machine learning program to detect polarity of sentences by using Naïve Bayes Classifier (Leung, 2007) where an NLP model is trained on a dictionary of sentences in the context of product reviews. Extensive literature exists on classifying movie reviews (Basari et al., 2013) into positive or negative movie reviews and on microblogging analysis (Pak & Paroubek, 2010) that develop an NLP to classify news documents into positive, neutral, and negative sentiment. Researchers can also choose to use pre-trained lexicons available to everyone, allowing researchers to use appropriate NLPs according to their context and lexical structures including the Stanford CoreNLP (Manning et al., 2014) and VADER (Hutto & Gilbert, 2014) which is generally seen as the best for the microblogging-sentence contextual data as it contains tokens for emojis and slang terms.

With original roots embedded in the sociology domain of extracting public opinion on political matters, computer science and machine learning have expanded the domain of sentiment analysis across vast domains. Sentiment is an extremely valuable factor, and this research paper will use modern sentiment analysis tools in answering the research question.

2.2 Stock pricing

From a theoretical perspective, there is the long-lasting efficient market hypothesis (EMT) developed by Fama (1970) that simply states that prices of stocks reflect all information available about them. This version of the EMT is classified as the semi-strong form of efficiency. With investors being informed about prices, trading securities under the EMT should be impossible to have higher returns than the market; it would be impossible to have excess returns without taking more risk. The EMT, and the fact that prices follow a Random Walk, was proven empirically multiple times (Alexander, 1961; Kendall & Hill, 1953; Roberts, 1959). Given this rather pessimistic view, that the stock market movements are as predictable as a coin toss, researchers have attempted to develop models and found factors to explain how excess returns can be explained.

When taking on systematic risk an investor will be compensated with higher expected returns, in a linear fashion (Sharpe, 1964). Further anomaly findings continued (Elbannan, 2014): smaller size firms show higher returns (Banz, 1981), highly leveraged firms show higher returns (Bhandari, 1988) and firms with high ratios of book to equity Stattman (1980) experience higher returns than their relative counterparts. The mentioned anomalies are all to be inputted in addition to the systematic risk factor (Sharpe, 1964) under a new model by Fama & French (1993).

The three-factor (Fama & French, 1993) model has been the model to beat for nearly two decades now. Carhart, 1997 use a momentum factor (Jegadeesh & Titman, 1993) in addition to the three, defined as the speed of price changes in the stock, where high momentum is associated with higher returns, thus building upon the weak form of EMT (Fama, 1970) where current stock price is affected by past stock price changes.

There now exists a plethora of research exploring how each model explains excess returns thus highlighting different forms of the EMT, but also new anomalies, highlighting that there is still some unexplained variation of excess returns (Woo et al., 2020) which the leading models (Fama & French, 1993 and Carhart, 1997) cannot explain, resulting into significant abnormal returns. Studies that challenge the EMT though anomalies include seasonal (Nippani & Arize, 2008), where U.S corporate bonds have higher returns in January compared to other months, reversal (De Bondt et al., 1985) where a past losing portfolio outperforms a past winner portfolio with a reversal happening every three to five years, and behavioral (Patel et al., 1991) where there is a herd effect where irrational investors follow others' actions resulting in short-term abnormal returns.

2.3 Sentiment analysis and stock pricing

Given the background on sentiment analysis and asset pricing models it is imperative to further showcase how the two variables are related from an econometric standpoint where the sentiment will be the predictor of the predicted stock returns. From a chronological perspective, sentiment analysis found its way in the domain of finance with the progress of the NLPs in the 2000s; this section will delve into the newest findings without explaining the oldest findings.

This research paper will answer the research questions through two hypotheses: one related to stock movement and the other to stock returns. In the domain of sentiment analysis, research papers have used opinion mining to either predict stock movement (direction) or stock returns (changes). This section will thus be split into two: one that analyzes findings of literature with the dependent variable as stock movement and one that analyzes through the lenses of researchers that used stock returns as their dependent variable.

2.3.1. Sentiment analysis and stock movement

Boudoukh et al. (2013) demonstrated through textual analysis that specific types of news articles significantly impact stock movement. Using an article database from the Dow Jones News over a 25-year time frame (1979-2004) they analyze both the firm specific tone as well as at the market level through the lens of The Stock Sonar, an augmented dictionary of Loughran & McDonald (2011) using positive and negative words developed by Feldman et al. (2011). Their aggregation method was the category of news (Financial, Legal, Product etc.) and daily dimension. They find that news with negative sentiment drives stock prices down, while positive news has the opposite effect. When using a 10 year-sample period, when regressing on the Fama-French 3 risk factors (Fama & French, 1993) on sample companies making a 1/N weighted portfolio, they find a significant constant, alpha, in certain portfolios. In conclusion, their paper uses opinion mining to find the correlation between sentiment and stock movement and use that as a potential portfolio strategy which has abnormal returns.

Ho & Huang (2021) further advanced this research by integrating sentiment analysis with candlestick chart representations. Their findings support the hypothesis that sentiment indicators extracted from media can provide early signals of stock price movements. Using Twitter data, Ho & Huang (2021) analyze it through the VADER (Hutto & Gilbert, 2014) model to get a sentiment spectrum to predict the stock movement of five sample companies: Apples, Tesla, IBM, Amazon and Alphabet. Their paper uses machine learning classifiers among the Random Forest Classifier also used by Khan et al. (2020) to classify stock market movements across similar but larger sample companies but were more concentrated on which models best predict stock movement through sentiment; the Random Forest Classifier has the highest accuracy out of all the models and the most consistent in predicting stock direction tomorrow. It is important to note that due to Khan et al. (2020) larger sample size he finds that despite using the best machine learning models there are certain stock may be more influenced (Microsoft) by news than others (Nokia). A larger sample size will also be employed in this research to highlight stocks that may be influenced more than others by news. A hypothesis relating media-tone to stock movement can be formulated as follows:

Hypothesis 1: Media-expressed sentiment tone has explanatory power in predicting stock movement.

Under the null hypothesis, media-expressed sentiment does not have any explanatory power in predicting stock movement and can be regarded as noise; the null is what this research attempts at rejecting. The sign of the effect is clear in current literature, positive sentiment is identified with positive movement; this research will mainly delve into finding inner complexities of sentiment. Inspired by Boudoukh et al. (2013) by using two aggregation dimensions, this research will explore sentiment tone at the publisher-daily dimensions.

2.3.2 Sentiment analysis and stock returns

Hu et al. (2021) use reddit emojis posted on forums specialized in predicting stock returns, opposed to word dictionaries and use 3 lags of their sentiment spectrum (rational, naïve and fanatic) to find that there is a significant effect of media-aggregated-tone in predicting stock returns. Ahmad et al. (2016) found that media-expressed negative tone correlates with significant declines in firm-level stock returns, through a very complete analysis of 5.5 million news articles. Both Ferguson et al. (2014), that used 300,000 articles, and Ahmad et al. (2016) use Loughran's & McDonald's (2011) dictionary to analyze news articles both highlighting the predictive power of sentiment on stock movement. They also find that making a portfolio when one longs positively talked about stocks and shorts the negative talked about stocks there are significant abnormal returns when regressing the returns on the Fama & French (1993) three factor model.

All the mentioned literature used daily aggregated sentiment to predicted tomorrow's stock market returns; the reason behind this is that there is a slight herding effect behavior which creates a temporary bubble that is to be burst in the very near future (Pedersen, 2022). Under the EMT, information should be directly priced in, but information takes time to be inferred by the market price. In the first hypothesis, sentiment will be used to predict the stock movement and those predictions will act as tools for portfolio creation and the second hypothesis can be formulated as follows:

Hypothesis 2: A trading strategy driven by sentiment results in significant abnormal returns.

The null hypothesis is that a trading strategy driven by sentiment results in portfolios with no significant abnormal returns; this research attempts to reject the null.

The relationship between stock prices and stock movement is complex, influenced by factors including sentiment, news content, and media coverage. Advances in textual analysis and NLP have significantly enhanced the ability to analyze these factors, providing deeper insights into the mechanisms driving stock price movements. This research will be using Ahmad et al. (2016) as the base methodology, using a larger data sample but similar method and answering their further research request to analyze sentiment and stock returns through machine learning means, thus employing Khan et al. (2020) method of using a binary machine learning classifier model. Inspired by using an additional aggregation dimension (Boudoukh et al., 2013), this paper will use the publisher as an aggregation tool; the first to do so in the context of opinion mining and stock returns.

3. Data

The data collection can be split into four different parts: sentiment, financial data, risk factors and descriptive statistics. The section below will provide information on the data sources used and manipulation of such data in creating the dependent variable and independent variable along with descriptive statistics where relevant.

3.1 Sentiment – the independent variable

This research will use data from the ProQuest database, which offers downloadable access to 15 different databases of news, blogs, company reports, trade journals and market reports. With the assumption that more articles will be published mentioning sizeable companies, the firms used in this research will be the top 30 non-financial companies[2] listed on the Fortune US 500 in 2013. As titles contain the most relevant information about an article, the main approach to gather the relevant articles for this study will be through a filter where only English news and blogs mentioning either the name or ticker, of the respective company, in the title will be downloaded. For simplicity, the word *articles* will refer to the titles of news and blogs, the primary source of data used in this research. Using this technique, the *articles* will be accurately assigned to the respective companies they reference. The timeframe of this dataset is from January 1st, 2014, to May 1st, 2024.

Articles contain words and non-numeric elements; an NLP model will be used to transform the titles into numerical, sentiment values. The model used to get a sentiment tone will be the *distil-ROBERTA-finetuned-financial-news-sentiment-analysis* (distil-ROBERTA), a pre-trained NLP model on financial news which will ensure sentiment values are based on a financial vocabulary. The model was chosen by both the context of data, financial news, and size of lexical structures as the model is limited to understanding only the first 512 tokens of the *article*. The distil-ROBERTA model will be parsed through all *articles* giving three outputs, a negative, neutral and positive probability of sentiment all, which will be used to create a compounded sentiment score to have a continuous variable. Using -1, 0, and 1 as numeric weights of the respective sentiment probability, the sentiment score for each observation will be calculated as follows: a sentiment closer to 1 represents a positive sentiment of the *article*, closer to 0 represents a neutral sentiment and closer to -1 represents a negative sentiment.

$$sentiment_i = -1 * \left(negative_{probability\ i}\right) + 0 * \left(neutral_{probability\ i}\right) + 1 * (positive_{probability\ i})$$

[2] Financial companies include companies involved in banking. Fannie Mae, J.P. Morgan Chase & Co., Bank of America Corp., Wells Fargo, Prudential Financial and Freddie Mac were removed due to this criterion. Berkshire Hathaway was kept as it was deemed a conglomerate.

The ProQuest database offers access to two important variables for each *article*, the *publisher* and *publication date*. A *publisher* is defined in this context as an entity that writes at least 1 *article* in the specified time frame, and the *publication date* as the date of publication of the article in YY-MM-DD format. Aggregating the sentiment data to the daily dimension assumes that every *article* posted on a specific day is equally weighted in the calculation of the daily aggregated sentiment. The assumption will be avoided by aggregating to the publisher level, *P*, and publication date, *T*, which means that there will be a sentiment score for each *publisher* every day. Under this technique, the assumption made is that all *articles* published by the same *publisher* on a particular *publication date*, are equally weighted in the calculation of the sentiment.

Moreover, there may be *publishers* that offer different kind of explanatory power for different companies, or they may write about different companies at different frequencies, each company will be treated as a separate *case*, *C*, showcasing the complexity of this research. For each *case*, only the *publishers* that have posted at least 500 times during the frame, *F*, will be considered as *publishers*, and the aggregation technique above will be applied to return a daily-*publisher*-*case* specific *sentiment* value ($s_{P,T,C}$). This will increase computational efficiency while also limiting the amount of noise with too many *publishers.* Finally, the *publishers* that have no articles for a given day that fulfill the criteria, will receive a sentiment score of 0 which ensures no missing values in the vector of sentiments which is a fair assumption. Having no *articles* being written means that nothing happened worth writing, hence, the neutral sentiment.

$$s_{P,T,C} = \begin{cases} \sum_{i=1}^{n} sentiment_i & if\ F_{P,C} \geq 500 \\ 0 & otherwise \end{cases}$$

A final additional step will be classifying the *sentiment* aggregated on Saturday and Sunday as additional sentiment values for the day of Friday. Essentially, Friday's sentiment values are the weighted average of the sentiment vectors calculated on Friday, Saturday and Sunday. This does not violate any internal validity concerns.

Furthermore, different lags of different *publisher* sentiment values may provide different explanatory power in each *case*; providing correlation and autocorrelations for each *case* and for each *publisher* in providing a motivation behind the number of lags would not be efficient; as each *case* will be treated individually, this research will simply give the model the flexibility to use any specific lags of any *publisher*. The explanatory power, or weights assigned to each *publisher* and lagged *publisher*-specific sentiment values up to the 10^{th} lag will be decided by the model; the *independent variable*, $S_{C,T}$, is thus a vector of daily *case*-specific sentiments with dimension, D_C is shown below. It is important to note that giving more lags to the models will result in noisy predictions and giving less lags defeats the purpose of using machine learning classifiers which are innately good at assigning weights to relevant independent variables.

$$L = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}_{10*1}$$

$$S_{C,T} = \begin{bmatrix} L \cdot s_{p1,T,C} \\ L \cdot s_{p2,T,C} \\ \dots \\ L \cdot s_{pP,T,C} \end{bmatrix} \; with\ dimension\ D_c$$

3.2 Stock Movement and Returns – the dependent variable

Using the Yahoo Finance API daily stock financial data was downloaded. This research will attempt to predict the movement or direction of the stock price tomorrow, *T+1*. An issue that arises is the look-ahead bias of using the adjusted close prices in determining the stock movement the next day; there may be *articles* being published after market close meaning that the *independent variable* will contain future information in predicting tomorrow's stock movement; given the aggregation technique for Friday's sentiment vector, using adjusted close prices will introduce great internal validity concerns. This can be solved by simply taking the movement tomorrow of the stock from open, *Op,* to close, *Cl*, which solves the overlap issue. The *independent variable* will be taking all sentiment information from the current and past days to predict the *stock movement* of the next day. A binary classification will be used, where a classification of 1 means *positive movement* tomorrow and 0 a *negative movement* tomorrow of the respective case for each day. An *observation* under the above conditions counts if there has been at least 1 *article* published for a tradable day.

$$stock\ movement_{C,T} = \begin{cases} 1\ if\ P_{T+1,\ Cl} - P_{T+1,\ Op} > 0 \\ 0\ if\ P_{T+1,\ Cl} - P_{T+1,\ Op} < 0 \end{cases}$$ [3]

It is important to note that the *dependent variable* is irrespective of magnitude of returns. The magnitude of returns will also be analyzed in the form of a portfolio creation thus splitting the results section into two separate sections; one that analyzes the explanatory power of the *independent variable* in

[3] The Close prices and Open prices were also used to calculate the stock returns for that day. This is being done to maximize the internal validity of this research.

predicting the *dependent variable*, answering thus the first hypothesis, and one that analyzes the magnitude of returns given the predicted movement in form of a portfolio, answering the second hypothesis.

3.3 Risk Factors for portfolio creation

To showcase how the predictions do in terms of magnitude and not just the direction of the stock returns for the given day portfolios will be created. With the literature reviewed (Ahmad et al., 2016 and Ferguson et al., 2014) using the classic 3 factor model (Fama & French, 1993) with market premium (MKT-RF), size (SMB) and book to market equity (HML) risk factors will be used to regress on the portfolios' returns. The risk factors will be taken from the Kenneth R. French data library. These will be used as control variables, to find whether sentiment-predicted movement yields significant abnormal returns.

3.4 Descriptive Statistics

The company Express Scripts was removed from the data sample as it stopped being publicly traded on December 20th, 2018 (Sweeney, 2018), it had considerably less data than the other 29 sample companies and it was not worth keeping in the data sample. After removing *articles* that have invalid *publishers* given the criteria mentioned in section 3.1.3, descriptive statistics for the primary source of data are in Table 1. After the selection criteria was applied, only tradable days were included in the descriptive statistics since there is simply no data on weekends under the current selection method.

A total of 1,274,924 *articles* were analyzed given the constraints mentioned and the sample companies show great variation in terms of primary data. Apple has the highest article count with 132,236 along with IBM at 110,547. Target has the lowest article count (13,149) and tied having the lowest articles per day (5) with AmerisourceBergen and Cardinal Health. Target has 540 tradable days with no articles. The number of articles per day varies also, in the range of 5 (AmerisourceBergen) and 35 (Apple). Apple and Boeing have the most complete data with only 1 tradable day without an article. Further, the number of valid *publishers* also varies and can be directly correlated to the number of *articles* for that *case*. AmerisourceBergen, Cardinal Health, Target, Valero and Walgreen have the lowest number of valid *publishers* with 5 and Apple has the highest number with 34. Ultimately, Apple for example will have 340 independent variables.

Table 1: Descriptive statistics of companies' articles.

Company	**Articles**	**Mean per day**	**Days with no articles**	**Publishers**
ADM	26,122	8	189	10
AmerisourceBergen	14,211	5	323	5
Apple	132,236	35	1	34
AT&T	59,108	16	6	22
Berkshire Hathway	34,139	10	105	19
Boeing	80,092	22	1	30
Cardinal Health	15,520	5	295	5
Chevron	42,369	12	113	12
Costco	33,199	10	153	10
CVS Health	32,044	9	67	10
ExxonMobil	41,051	12	102	11
Ford	44,615	12	15	17
General Electric	41,061	12	58	13
General Motors	41,171	12	10	17
Home Depot	50,328	14	26	14
HP	25,571	9	421	7
IBM	110,547	31	14	17
Kroger	24,909	7	114	10
Marathon Petroleum	18,008	6	310	6
McKesson	17,609	6	523	6
Microsoft	117,761	32	12	26
Phillips 66	24,620	7	213	7
Procter & Gamble	41,824	12	141	8
Target	13,149	5	540	5
UnitedHealth Group	31,042	9	215	6
Valero	19,842	6	369	5
Verizon	63,873	18	8	23
Walgreen	20,075	7	146	5
Walmart	58,828	16	14	18

Notes: This table presents descriptive statistics of the articles downloaded from the ProQuest database in the time frame January 2014 – May 2024 with only tradable days being considered. The values were calculated after the selection criteria and once the weekend articles were classified as Friday articles.

4. Methodology

The section below explains the method and technique used for this research. The methodology can be split into two sections, the machine learning methods of analyzing and predicting the *stock movement*, the magnitude evaluation of predicted stock movements through different portfolios.

4.1 Machine Learning classifiers

As this research will assume that the relationship between the *dependent variable* and *independent variable* is non-linear, machine learning models that can capture non-linear patterns will be used. This paper will use the Decision Tree Classifier, due to its interpretability and will be used as a base model, and the Random Forest Classifier, the best performing model in Khan et al. (2020) research. The reason behind using a classifier and supervised machine learning methods is that the *dependent variable* is observable, hence the term *supervised*, through two different *classes*, hence the term *classifier*, that of *positive movement* (1) and *negative movement* (0). Both models will predict the stock movement at *T+1* based on the *observation's sentiment* at *T* in the specified vector form.

For machine learning methods and this type of research two sub-samples will be created from the original dataset of all *observations*. The *case* specific data set will be ordered by date and the data will be split into two. The training set (80%) will be used to train the models mentioned above, which allows the models to learn the non-linear relationships between *stock movement* and *sentiment*; on this data set, the *hyperparameters*, defined as the model's unique tuning parameters for controlling overfitting and underfitting, will be chosen based on *accuracy* over the k-fold validation within the training set. In this case a 7-fold validation splits the training set into 7 equal sub-data sets and uses them to train and validate its own predictions by maximizing training accuracy. A*ccuracy* is defined as the proportion of the summed true positives (TP) and true negatives (TN) to the sum of TP, TN, false positives (FP) and false negatives (FN).

$$Accuracy_C = \frac{TP + TN}{TP + TN + FP + FN}$$

The validation data set (20%) will be used to validate the best model on out-of-sample data, where predictions will be used in constructing a portfolio. The method above ensures a tradable strategy starting in May 2022 if the best model hyperparameters are used.

Further, a model that is overfitted describes the trained data well (high training accuracy), but when subject to out-of-sample it performs poorly (low validation accuracy) while an underfitted model is when a model is not trained enough to capture the complex patters in the dataset.

4.1.1 Decision Tree Classifier

The main goal of a decision tree classifier (DTC) is to split the whole data sample (*A*) into smaller sub-samples until each observation belongs to one of the *classes*, *k*, or a stopping criterion is met. The splitting starts at an initial node, called the root node, which attempts to separate the proportion, *p*, of instances of each class in the node by minimizing the Gini Impurity (GI) thus using the Classification and Regression Trees (CART) algorithm. A split occurs when the GI loss function can be lowered through an element in the vector of *sentiment* that best splits the data into two subsequent nodes.

$$Gini\ Impurity\ (A) = 1 - \sum_{i=1}^{k} p_i^2$$

When splitting is possible, this creates depth, defined as the maximum number of allowed splits from the root node. Depth along with the minimum number of samples in a leaf are the hyperparameters that will be tuned based on the accuracy of the predictions on a 7-fold validation technique. For each *case*, decision tree classifier models with depths, *DE,* and minimum sample at a leaf, *MLS*, from range of 1 to 35 and 1 to 25 over 7 folds of the training set, respectively, will be trained and the model with the highest cross-validated accuracy will be deemed the best model.

The validation data will be fitted on to the best model. A prediction, is one in which the observation's $S_{C,T}$ falls into multiple nodes, R_m, where the c_m is the predicted class determined by the majority *class* at R_m. It is important to notice that every observation in the validation data set will fall into only one leaf node R_l, hence the indicator function, *I*, taking a value of 1 when the observation's input falls into R_l. The output is a vector of two probabilities, the distribution of *classes* of all training instances at R_l. Since there are only two classes, a confidence classification interval, θ of 0.5 can be employed to have predictions for all observations. For hypothesis 1, the majority class ($\theta = 0.5$) will be used and for hypothesis 2, this cutoff point will be altered. The relationship between the output and input of sentiment are captured by Molnar (2024) and simplified below:

$$movement\ pro\widehat{babil}ities\ DTC_{C,T} = \widehat{DTC}(S_{C,T}) = \sum_{m=1}^{M} c_m I\{S_{C,T} \in R_m\} \quad where \quad c_m = \begin{bmatrix} p(0) \\ p(1) \end{bmatrix}$$

$$stock\ mov\widehat{emen}t\ DTC_{C,T} = \begin{cases} 0, & p(0)_{\widehat{DTC}(S_{C,T})} > \theta \\ 1, & p(1)_{\widehat{DTC}(S_{C,T})} > \theta \end{cases}$$

4.1.2 Random Forest classifier

The random forest classifier (RFC) works in a similar way to the decision tree; essentially multiple trees are being modeled, hence the name forest. The model uses the square root of the D_C as the number, *r*, of randomly chosen predictors, in each tree, and uses a different random subset of the training data sampled, also known as bootstrapping, hence the name random. Each tree will have a random sample of predictors and subset of the training data. This will control overfitting and underfitting a lot better than the DTC.

A split in an individual tree can only happen when the Gini Impurity can be lowered by an element in the randomized allocation of parameters for that tree. In other words, the RFC works the exact same way a DTC works but it trains multiple trees with each tree looking at a random allocation of parameters. The hyperparameters modelled on to the training set will be the number of trees, *TR*, ranging from 1 to 80, the depth, *DE*, ranging from 1 to 30, and the minimum leaf sample, *MLS*, ranging from 1 to 25.

The classification of predictions works in a similar way to the DTC, where the observation's $S_{C,T}$s are used to navigate to a leaf node, but this time, in each of the trees (TR) of the RFC. The distribution of classes for all observations is thus a vector of probabilities of size *TR*, where the average class distribution is taken.

$$movement\ \widehat{probabilities}\ RFC_{C,T} = \widehat{RFC}\left(S_{C,T}\right) = \frac{1}{TR}\sum_{TR=1}^{TR}\sum_{m=1}^{M} c_m I\left\{S_{C,T} \in R_m\right\}\ where\ c_m = \begin{bmatrix} p(0) \\ p(1) \end{bmatrix}$$

$$stock\ \widehat{movement}\ RFC_{C,T} = \begin{cases} 0, & p(0)_{\widehat{RFC}(S_{C,T})} > \theta \\ 1, & p(1)_{\widehat{RFC}(S_{C,T})} > \theta \end{cases}$$

4.2 Portfolio Creation

This research will evaluate how a trading strategy would perform, and three different portfolios will be created. Given the two *classes*, a long portfolio will be constructed given than the predictions are of *positive movement* tomorrow, a short portfolio, given that the prediction are of *negative movement* tomorrow. Both portfolios will then be aggregated to the daily dimension to have one observation per day. A long-short strategy will be evaluated, respectively. Portfolios returns will be regressed on the Fama & French (1993) daily risk factors and the constant, will be evaluated in terms of sign, and significance to uncover the potential abnormal returns this strategy would yield.

Predictions are made by assigning the observation a *class* given the cut-off point, θ. The term θ, was used to describe the confidence constraint which can be seen as a cut-off point; only predictions that fall above the cut-off point will be valid. Multiple portfolios can be constructed altering this parameter.

5. Results

The results will be split into two sections as there are results of two models, for two different hypotheses. The first part will cover the stock direction predictions of the models and the second part, portfolio returns of the models will be analyzed, which will help answer the second hypothesis.

5.1 Hypothesis 1 Results

Interpretability of coefficients in machine learning is difficult and extensive. Conceptually, patterns and correlations are found in each node's data subset and there could be drastically different patterns across the different *cases* and through the different branches of the trees. The method of interpretability will be through the hyperparameters calculated by the models on the training set of each *case*, and the training (*ACC_TRAIN)* and validation (*ACC_VALID*). The predictions made are only shown given $\theta = 0.5$, meaning that every observation in the dataset will have a prediction.

5.1.1 DTC Results

The DTC can find patterns within the 7-split training set, and it manages to get a training of 0.52 on average, shown in Table 2. There is not a lot of variation in the training accuracy, the range is between 0.52 and 0.54 which can be attributed to the idea that the model simply has more data to work with (Apple, Ford and IBM) all which have substantially higher number of *publishers* and *articles.*

In terms of hyperparameters the DTC (Table 2) can clearly show which *cases* are affected by sentiment and which are not by looking at the depth of each *case*. The range is from 1 (ExxonMobil and Ford) to 25 (IBM). With 1 depth as the best model, the DTC simply can find just one sentiment value in the vector that minimizes the Gini Coefficient. There are *cases* that are not reactive to sentiment. The data aggregation technique is being used by DTC to its maximum potential as there are *cases* with large amounts of depth thus finding intricate patterns between multiple relevant *publishers* or lags. On average, there were 11 depths being used across the *cases* and a minimum leaf sample of 13.

The minimum sample leaf is a measure of overfit and it can be considered as a hyperparameter tuned on the data rather than being *case* specific. Overfitting is not fully taken care of under this hyperparameter as there are *cases* with a very low minimum sample leaf such as Phillips 66, HP, Target, Target and Walgreen with 1, 2, 2, 2 and 2 respectively.

In terms of validation accuracy, it is expected that it will not deviate a lot from the training accuracy due to the way the 7-fold cross validation works but there seems to be a variety of accuracies in the range of 0.46 (General Motors) and 0.55 (Apple). This is one downside of the DTC that the variation in validation accuracies is high while the training accuracy yields great results.

Table 2. Best RFC hyperparameters and accuracies by *case*

Company	DE	MLS	ACC_TRAIN	ACC_VALID
ADM	3	13	0.53	0.50
AmerisourceBergen	23	6	0.54	0.50
Apple	4	8	0.52	0.54
AT&T	7	18	0.55	0.50
Berkshire Hathway	18	18	0.53	0.49
Boeing	19	11	0.53	0.53
Cardinal Health	4	13	0.52	0.51
Chevron	12	23	0.53	0.50
Costco	3	19	0.52	0.55
CVS Health	7	20	0.52	0.50
ExxonMobil	1	15	0.52	0.48
Ford	1	12	0.55	0.52
General Electric	2	16	0.53	0.53
General Motors	19	5	0.52	0.51
Home Depot	2	20	0.56	0.53
HP	19	2	0.52	0.49
IBM	25	16	0.51	0.51
Kroger	5	6	0.52	0.52
Marathon Petroleum	13	13	0.52	0.47
McKesson	12	18	0.52	0.53
Microsoft	6	21	0.53	0.46
Phillips 66	17	1	0.52	0.51
Procter & Gamble	5	24	0.54	0.52
Target	22	2	0.51	0.51
UnitedHealth Group	6	13	0.52	0.54
Valero	13	4	0.53	0.52
Verizon	23	13	0.52	0.50
Walgreen	5	2	0.52	0.50
Walmart	9	13	0.52	0.52
Average	***11***	***13***	***0.53***	***0.51***

Notes: This table presents the hyperparameters that performed best on the DTC 7-fold cross validated training set compromised of 80% of the data. The same best model was then applied to the validation data set (20%) of the remaining data where accuracy was measured again.

5.1.2 RFC Results

The RFC achieved an average training accuracy of 0.53 (Table 3). The training accuracies range from 0.52 to 0.56, suggesting the model's consistency in finding patterns across all *cases*. This variability can be attributed to differences in the amount of data available for each *case* and the complexity of the underlying patterns, like the DTC. One important *case* to highlight is Microsoft as in the DTC despite having many *articles* and *publishers* it performs the worst on the validation set (0.46) but when the RFC is used 0.52 is achieved. This highlights one of the advantages of the *random* and *forest* elements of the RFC as it is more consistent and reliable.

Analyzing the hyperparameters in Table 3, the depth varies significantly across cases, from as low as 1 (IBM) to as high as 23 (Walgreen). The number of trees (TR) used in the forest also varies from as few as 4 (Berkshire Hathaway, Costco, ExxonMobil, Target, Walmart) to as many as 68 (Kroger). These variations indicate that while some *cases*' stock movements can be predicted with simpler models, others require more complex models to capture the underlying patterns. Very similar to the DTC, the variation of the best hyperparameters stems from the aggregation technique used. From a large selection of sentiment scores, the RFC picks and chooses which element in the sample of randomly picked predictors best predicts stock movement.

The hyperparameter tuning suggests that overfitting is managed to some extent but not completely avoided. For instance, companies with low minimum leaf samples, such as ExxonMobil and Phillips 66, may still suffer from overfitting despite the overall effort to mitigate it. On average, the RFC uses a tree depth of 10, a minimum leaf sample of 14, and 23 trees, indicating a balance between model complexity and generalizability.

The range of validation accuracies remains wide, from 0.46 (General Electric) to 0.55 (AmerisourceBergen, Apple, CVS Health, and Walmart), but only 4 *cases* have a validation accuracy under 0.50 highlighting the advantage of the RFC, lower variation in results pointing towards better predictions. The average validation accuracy is 0.52, slightly lower than the training accuracy but still above a coin flip, supporting the hypothesis that the model has predictive capabilities given $\theta = 0.5$.

Table 3. Best RFC hyperparameters and accuracies by *case*

Company	**DE**	**MLS**	**TR**	**ACC_TRAIN**	**ACC_VALIDATE**
ADM	17	21	40	0.53	0.52
AmerisourceBergen	9	21	12	0.54	0.55
Apple	5	23	60	0.55	0.54
AT&T	19	3	16	0.54	0.53
Berkshire Hathway	15	19	4	0.52	0.49
Boeing	4	17	13	0.53	0.50
Cardinal Health	17	23	24	0.52	0.49
Chevron	13	13	12	0.54	0.54
Costco	11	25	4	0.53	0.53
CVS Health	5	19	32	0.54	0.54
ExxonMobil	13	1	4	0.53	0.48
Ford	5	5	24	0.56	0.51
General Electric	5	3	55	0.54	0.46
General Motors	19	9	16	0.53	0.52
Home Depot	5	15	24	0.54	0.52
HP	3	11	52	0.52	0.50
IBM	1	23	8	0.53	0.54
Kroger	3	23	68	0.53	0.53
Marathon Petroleum	9	13	8	0.53	0.52
McKesson	7	17	24	0.53	0.54
Microsoft	3	7	8	0.55	0.52
Phillips 66	11	1	12	0.52	0.52
Procter & Gamble	23	25	56	0.54	0.54
Target	3	11	4	0.52	0.51
UnitedHealth Group	3	25	16	0.53	0.52
Valero	19	25	20	0.55	0.53
Verizon	7	7	44	0.54	0.51
Walgreen	23	1	4	0.53	0.53
Walmart	13	13	4	0.53	0.55
Average	***10***	***14***	***23***	***0.53***	***0.52***

Notes: This table presents the hyperparameters that performed best on the RFC 7-fold cross validated training set compromised of 80% of the data. The same best model was then applied to the validation data set (20%) of the remaining data where accuracy was measured again.

5.2. Hypothesis 2 Results

Before making the portfolios, it is important to set appropriate cut-off points, θ, for the different portfolios given the output of the models. Figure 1 shows descriptive statistics the distribution of the positive class probability[4] of each observation in the validation data set inputted into each of the two models. Both distributions are symmetric which is expected when only using two predictive classes. The whiskers of the boxplots show a larger spread in favor of the DTC which shows one of the disadvantages of the DTC, the larger variation in output when compared to the RFC. In terms of outliers, due to certain *cases* having a very low minimum sample leaf there are observations that are overfitted, where the leaf node has a high 0 or 1 class distribution.

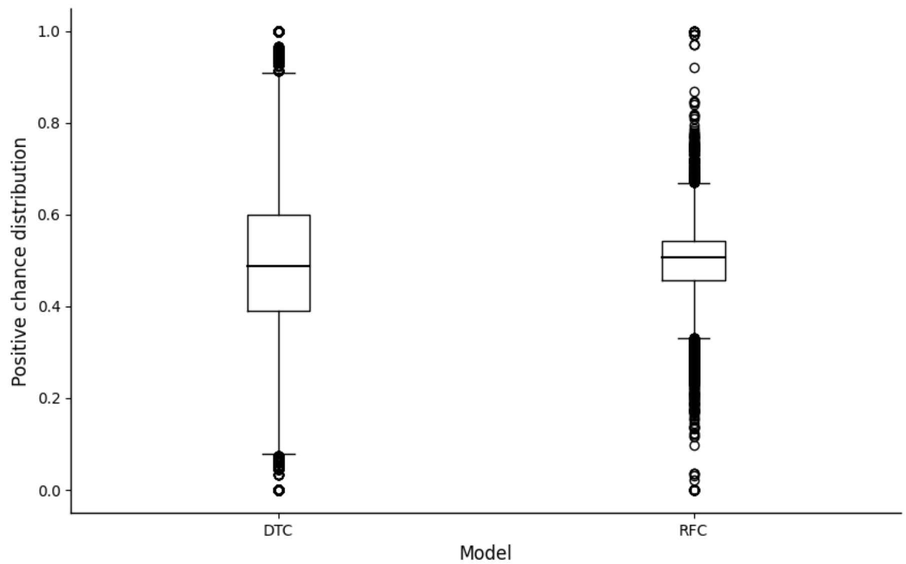

Figure 1: Boxplot of the positive chance distribution class of the DTC and RFC models' observations

The minimum sample leaf was the only hyperparameter used in controlling for overfitting and while it does control for overfitting well, to avoid further spurious relations for the portfolio analysis observations that have a *p(1)* higher than 0.95 or lower than 0.05 for both the DTC and RFC predictions.

Given Figure 1, appropriate cut-off points can be used for each model to uncover how portfolio's abnormal returns alter by changing this additional parameter. Cut-off points in steps of 0.05 will be analyzed from the range of 0.5 to 0.80. The scope of the analysis will be directed at evaluating the intercept in terms of significance and sign after regressing the portfolios return on the Fama & French (1993) factors. For simplicity, the short portfolio returns were multiplied by -1[5], which switches the sign of all coefficients including that of the intercept. Long portfolios will consist of buying positive movement signals, short

[4] The negative class probability distribution offers the distribution of 1-p(1) following the same pattern as Figure 1.
[5] A negative sign of the intercept in any regression across the long, short and long-short strategies will mean that strategy yields negative returns if followed.

portfolios will consist of shorting negative movement signals, and long-short portfolios will be allowed to long and short in line with the model predictions.

5.2.1. DTC Results

Despite the DTC's simplicity compared to more complex models, it provides insightful results across different trading strategies. As θ increases, the intercept for the long portfolio (Table 4) is not affected showing that the DTC confidence constraint is not correlated with increased accuracy. This relates again to the fundamentals of the DTC; with only one tree being built, patterns can be found within the dataset, but they do not translate to causal effects. For the short portfolios (Table 5), increasing θ does increase the magnitude of the constant but insignificantly. The best long portfolio ($\theta = 0.75$) and short portfolio ($\theta = 0.75$) are shown in Figure 1. While both strategies yield positive returns, they can all be explained by the systematic risk, size factor and value factors as they are all significant for the portfolios mentioned.

The DTC long-short portfolios (Table 6), yield insightful results as θ is increased the intercept increases faster than the standard error and at cut off points 0.7 and 0.75 the strategy yields significant abnormal returns to the 0.1 significance level. The best long-short portfolio is with $\theta = 0.75$, which coincides with the best long and short θ and its graphical performance is shown in Figure 1 also. The best performing short strategy performs quite poorly when compared to the long strategy highlighting the poorer performance of the DTC, returns which can be explained due to the natural market movement and risk factors.

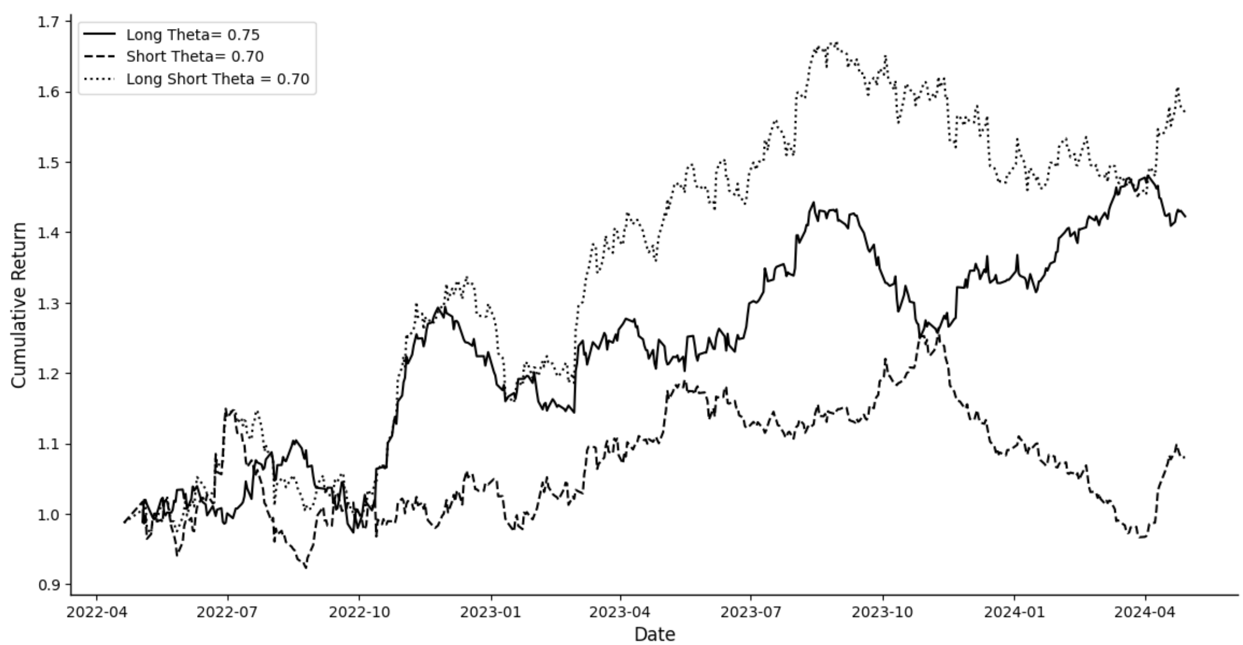

Figure 2: DTC best θ portfolios for long, short and long-short strategies

Table 4. DTC Fama – French portfolio regressions with different θ values – long strategy

	DTC - Long portfolios						
	0.50	0.55	0.60	0.65	0.70	0.75	0.80
Intercept	0.000076	0.000323	0.000361	0.000124	0.000427	0.000597	0.000594
	(0.000270)	(0.000300)	(0.000369)	(0.000406)	(0.000454)	(0.000456)	(0.000522)
Mkt-Rf	0.005279***	0.005167***	0.005416***	0.005100***	0.004602***	0.004800***	0.004843***
	(0.000259)	(0.000287)	(0.000353)	(0.000390)	(0.000438)	(0.000438)	(0.000514)
SMB	-0.001502***	-0.001133**	-0.001642***	-0.001531**	-0.001809***	-0.001654**	-0.002370***
	(0.000415)	(0.000460)	(0.000566)	(0.000619)	(0.000694)	(0.000694)	(0.000802
HML	0.000864***	0.001236***	0.001817***	0.002234***	0.002400***	0.002484***	0.001849***
	(0.000328)	(0.000364)	(0.000448)	(0.000491)	(0.000547)	(0.000547)	(0.000615)
Adj. R2	0.469477	0.406388	0.322681	0.258189	0.191816	0.212129	0.188154
Obs.	502	501	501	494	461	445	371

Notes: Columns represent an OLS regression on the Fama & French (1993) excess return (Mkt-Rf), size (SMB) and value (HML) risk factors and a constant with different θ, cutoff points, of the DTC positive movement model predictions. Returns of the predictions are aggregated to the daily level. This trading strategy buys when the model signals a positive movement signal. The number of observations, representative of the trading days that the models trade on during the period January 2022-May 2024. The standard errors are shown in parentheses under their respective coefficient. The stars present after the coefficients represent the following significance levels: * p<0.1, ** p<0.05 and *** p<0.01.

Table 5. DTC Fama – French portfolio regressions with different θ values – short strategy

	DTC - Short portfolios						
	0.50	0.55	0.60	0.65	0.70	0.75	0.80
Intercept	-0.000102	-0.000106	0.000055	0.000250	0.000471	0.000404	0.000373
	(0.000289)	(0.000298)	(0.000325)	(0.000370)	(0.000417)	(0.000459)	(0.000516)
Mkt-Rf	-0.005547***	-0.005323***	-0.005105***	-0.005623***	-0.005701***	-0.005523***	-0.005680***
	(0.000279)	(0.000287)	(0.000312)	(0.000355)	(0.000399)	(0.000440)	(0.000482)
SMB	-0.000003	0.000017	0.000449	0.000762	0.001258**	0.001311*	0.002960***
	(0.000445)	(0.000459)	(0.000499)	(0.000567)	(0.000637)	(0.000702)	(0.000804)
HML	-0.003695***	-0.003237***	-0.002958***	-0.003302***	-0.003932***	-0.003631***	-0.002501***
	(0.000351)	(0.000362)	(0.000395)	(0.000448)	(0.000502)	(0.000554)	(0.000618)
Adj. R2	0.473831	0.434422	0.366168	0.351660	0.311081	0.261148	0.246662
Obs.	502	502	501	494	484	473	416

Notes: Columns represent an OLS regression on the Fama & French (1993) excess return (Mkt-Rf), size (SMB) and value (HML) risk factors with different θ, cutoff points, of the DTC negative movement model predictions. Returns of the predictions are aggregated to the daily level and multiplied by -1 for simplicity of comparison across the other strategies. This trading strategy is thus limited to only selling when the model signals a negative movement signal. The number of observations, representative of the trading days that the models trade on during the period January 2022-May 2024. The standard errors are shown in parentheses under their respective coefficient. The stars present after the coefficients represent the following significance levels: * p<0.1, ** p<0.05 and *** p<0.01.

Table 6. DTC Fama – French portfolio regressions with different θ values – long-short strategy

	DTC – Long-Short portfolios						
	0.50	0.55	0.60	0.65	0.70	0.75	0.80
Intercept	0.000008	0.000252	0.000447	0.000396	0.000868*	0.000997*	0.000670
	(0.000244)	(0.000284)	(0.000379)	(0.000454)	(0.000515)	(0.000539)	(0.000572)
Mkt-Rf	-0.000232	-0.000124	0.000342	-0.000528	-0.001342***	-0.000983*	-0.001632***
	(0.000234)	(0.000272)	(0.000363)	(0.000436)	(0.000496)	(0.000519)	(0.000553)
SMB	-0.001556***	-0.001164***	-0.001238**	-0.000853	-0.000433	-0.000272	0.001076
	(0.000376)	(0.000436)	(0.000582)	(0.000698)	(0.000791)	(0.000826)	(0.000881)
HML	-0.002831***	-0.001998***	-0.001142**	-0.001026*	-0.001597**	-0.001314**	-0.000904
	(0.000297)	(0.000345)	(0.000460)	(0.000552)	(0.000625)	(0.000654)	(0.000690)
Adj. R2	0.171392	0.068985	0.017247	0.006107	0.017786	0.006791	0.012256
Obs.	504	503	503	503	501	499	483

Notes: Columns represent an OLS regression on the Fama & French (1993) excess return (Mkt-Rf), size (SMB) and value (HML) risk factors with different θ, cutoff points, of the DTC model predictions. Returns of the predictions are aggregated to the daily level. This trading strategy is longing positive movement predictions and shorting negative movement predictions; the average return, given that the short portfolio returns was multiplied by -1 will then be aggregated to the daily dimension. The number of observations, representative of the trading days that the models trade on during the period January 2022-May 2024. The standard errors are shown in parentheses under their respective coefficient. The stars present after the coefficients represent the following significance levels: * p<0.1, ** p<0.05 and *** p<0.01.

5.2.2 RFC Results

The intercept increases as θ is increased across all three strategies showing that model confidence is also correlated with more reliable movement predictions which translate to higher returns. The long portfolio intercepts found in Table 7 increase steadily in magnitude until the portfolio with $\theta = 0.65$, where the intercept begins to fall and become insignificantly negative by the portfolio with $\theta = 0.80$ with a negative intercept of -0.000720. This can be explained by the relatively small data sample (28 observations) and due to the tight window of valid observations which translates to using overfitted data. The long-short portfolios (Table 6) show a similar pattern, with increasing θ a higher intercept follows but remains insignificant and always positive. Intercepts of the long and long-short strategies remain insignificant as the standard error also increases with θ. The short portfolios found in Table 8, for low values of θ, have a negative insignificant intercept but increases to positive and significant for θ values above 0.7. The best performing portfolio over the two years has a significant intercept of 0.001560 significant to the 0.05 level.

Figure 3 show the cumulative returns of the best θ performing portfolio in the long, short and long-short categories. The long portfolio maximizes returns with a θ value of 0.65 and the short portfolio with a value 0.7. The absolute best long-short portfolio, *positive movement* predictions should be constrained at a confidence interval of 0.65 and *negative movement* predictions should be constrained at a θ value of 0.70, but due to the mirror effect of the θ on both classes, the best long-short portfolio has a θ of 0.7. Cumulative returns on the RFC predictions are also a lot more consistent than the DTC best strategy portfolios from Figure 2 which may highlight the extra edge the RFC (52%) has over the DTC (51%) in predicting stock movement.

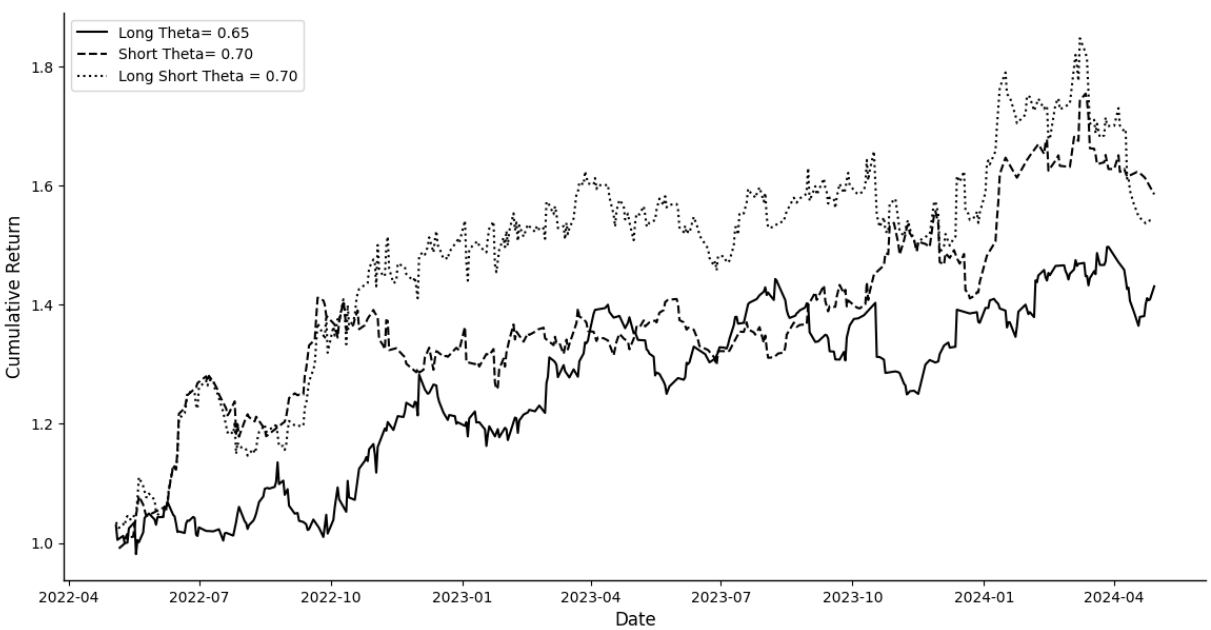

Figure 3: RFC best θ portfolios for long, short and long-short strategies

Table 7. RFC Fama – French portfolio regressions with different θ values – long strategy

	RFC - Long portfolios						
	0.50	0.55	0.60	0.65	0.70	0.75	0.80
Intercept	0.000150	0.000296	0.000480	0.000870	0.000567	-0.000607	-0.000720
	(0.000263)	(0.000337)	(0.000444)	(0.000646)	(0.000993)	(0.001216)	(0.003130)
Mkt-Rf	0.005488***	0.005839***	0.005455***	0.005963***	0.007123***	0.008971***	0.007824***
	(0.000253)	(0.000324)	(0.000425)	(0.000597)	(0.000908)	(0.001212)	(0.002311)
SMB	-0.000673*	-0.000750	-0.000181	0.001141	0.002056	0.001710	0.001283
	(0.000405)	(0.000518)	(0.000689)	(0.001000)	(0.001557)	(0.001893)	(0.004207)
HML	0.002238***	0.002606***	0.003160***	0.003862***	0.006111***	0.008785***	0.007388**
	(0.000318)	(0.000408)	(0.000537)	(0.000773)	(0.001237)	(0.001523)	(0.003149)
Adj. R2	0.498359	0.405890	0.273083	0.248161	0.256969	0.323584	0.292962
Obs.	505	503	477	356	223	150	28

Notes: Columns represent an OLS regression on the Fama & French (1993) excess return (Mkt-Rf), size (SMB) and value (HML) risk factors and a constant with different θ, cutoff points, of the RFC positive movement model predictions. Returns of the predictions are aggregated to the daily level. This trading strategy buys when the model signals a positive movement signal. The number of observations, representative of the trading days that the models trade on during the period January 2022-May 2024. The standard errors are shown in parentheses under their respective coefficient. The stars present after the coefficients represent the following significance levels: * $p<0.1$, ** $p<0.05$ and *** $p<0.01$.

Table 8. RFC Fama – French portfolio regressions with different θ values – short strategy

	RFC - Short portfolios						
	0.50	0.55	0.60	0.65	0.70	0.75	0.80
Intercept	-0.000026	-0.000406	0.000170	0.000489	0.001560**	0.002111*	0.003733*
	(0.000306)	(0.000368)	(0.000489)	(0.000583)	(0.000749)	(0.001104)	(0.002152)
Mkt-Rf	-0.005885***	-0.006298***	-0.006310***	-0.005826***	-0.006013***	-0.006146***	-0.006303***
	(0.000289)	(0.000348)	(0.000469)	(0.000561)	(0.000702)	(0.001061)	(0.002121)
SMB	0.000748	-0.000449	-0.000867	-0.001049	-0.001201	-0.000766	0.005470*
	(0.000468)	(0.000563)	(0.000751)	(0.000891)	(0.001097)	(0.001610)	(0.003246)
HML	-0.002882***	-0.003234***	-0.004402***	-0.004289***	-0.004788***	-0.005333***	-0.003980
	(0.000371)	(0.000447)	(0.000592)	(0.000713)	(0.000887)	(0.001303)	(0.002497)
Adj. R2	0.463594	0.423440	0.315388	0.246115	0.236577	0.171739	0.151882
Obs.	507	504	475	409	299	207	55

Notes: Columns represent an OLS regression on the Fama & French (1993) excess return (Mkt-Rf), size (SMB) and value (HML) risk factors with different θ, cutoff points, of the RFC negative movement model predictions. Returns of the predictions are aggregated to the daily level and multiplied by -1 for simplicity of comparison across the other strategies. This trading strategy is thus limited to only selling when the model signals a negative movement signal. The number of observations, representative of the trading days that the models trade on during the period January 2022-May 2024. The standard errors are shown in parentheses under their respective coefficient. The stars present after the coefficients represent the following significance levels: * $p<0.1$, ** $p<0.05$ and *** $p<0.01$.

Table 9. RFC Fama – French portfolio regressions with different θ values – long-short strategy

	RFC – Long-Short portfolios						
	0.50	0.55	0.60	0.65	0.70	0.75	0.80
Intercept	0.000277	0.000008	0.000695	0.000980	0.001184	0.001125	0.00267
	(0.000249)	(0.000376)	(0.000528)	(0.000676)	(0.000813)	(0.001046)	(0.002060)
Mkt-Rf	-0.000599**	-0.000637*	-0.000793	-0.000154	-0.000371	-0.000334	-0.000044
	(0.000235)	(0.000355)	(0.000508)	(0.000645)	(0.000761)	(0.001021)	(0.001833)
SMB	0.000241	-0.001042*	-0.001197	-0.000451	-0.000205	0.000199	0.004691
	(0.000383)	(0.000576)	(0.000810)	(0.001041)	(0.001241)	(0.001584)	(0.003004)
HML	-0.000666**	-0.000685	-0.001130*	-0.000922	-0.000970	-0.000582	0.000340
	(0.000302)	(0.000455)	(0.000641)	(0.000829)	(0.000978)	(0.001256)	(0.002278)
Adj. R2	0.010799	0.013150	0.009679	-0.003315	-0.004831	-0.008849	-0.004800
Obs.	512	507	500	483	405	313	78

Notes: Columns represent an OLS regression on the Fama & French (1993) excess return (Mkt-Rf), size (SMB) and value (HML) risk factors with different θ, cutoff points, of the RFC model predictions. Returns of the predictions are aggregated to the daily level. This trading strategy is longing positive movement predictions and shorting negative movement predictions; the average return, given that the short portfolio returns was multiplied by -1 will then be aggregated to the daily dimension. The number of observations, representative of the trading days that the models trade on during the period January 2022-May 2024. The standard errors are shown in parentheses under their respective coefficient. The stars present after the coefficients represent the following significance levels: * p<0.1, ** p<0.05 and *** p<0.01.

6. Discussion

6.1 Hypothesis 1

The findings of this research provide evidence supporting the first hypothesis, that media-expressed sentiment tone has explanatory power in predicting stock movement. The data analysis, employing both the DTC and RFC, demonstrates an average prediction accuracy of 51% and 52%, respectively, across the sampled *cases*. This suggests a slight edge over a coin flip, which has a 50% probability of predicting stock direction. Although the prediction accuracy appears marginally better than random chance, a small predictive edge can be significant on average in financial markets, where consistent small gains can accumulate over time.

The results align with the work of Boudoukh et al. (2013) and Ho & Huang (2021), who also found that sentiment indicators could provide early signals of stock price movements through traditional means of OLS. Comparing results to Khan et al. (2020) that also use classifiers to predict stock market movement their accuracy results range between 44% and 58% for a 1 day holding period after feature selection. This research has a slightly smaller range of results, but this can be attributed to a larger data set given the additional dimension of the *publisher* and more lags, and the inclusion of hyperparameters that control for overfitting, as Khan et al. (2020) use 1 as their minimum sample leaf which can alter results quite significantly as shown in Table 2 and 3. In conclusion, there are patterns that are found at the *publisher* sentiment dimension in predicting stock direction.

It's important to acknowledge that predicting stock movements is inherently challenging due to market complexities and various influencing factors. While the classifiers used in this research showed slightly better performance than a coinflip chance, they are not the best predictions. This is the reason why altering θ was used; the first hypothesis was calculated using $\theta = 0.5$ and while it does offer higher performance to a coinflip chance, portfolios with significant abnormal returns were found given $\theta \geq 0.7$. Without calculating the validation accuracies for different θ values, it is obvious that stock movement predictions get more accurate as the intercept does increase in Tables 4-9 with the cutoff point being increased showing that there are instances in which sentiment may be a very valuable predictor of stock movement.

Borrowing answers from both the designated results table of the 1st hypothesis as well as the underlying findings from the results of the 2nd hypothesis, the null hypothesis is rejected of sentiment having no explanatory power in predicting stock market movements. This denotes that stock prices can be predictable to a certain extent, due to the existing relationship between sentiment and tomorrow's stock movement; it is very important to highlight that the effect is not being evaluated. To answer the research question, the classifier models find patterns in the vector of sentiment which yield to correct predictions training set which are then reapplied to a validation set which yields accuracies of 52% in case of the RFC highlighting sentiment's predictive capabilities.

6.2 Hypothesis 2

This research attempted to tune many parameters, and after the extensive data manipulation of transforming news *articles* into sentimental values to predict stock direction, abnormal returns are found under certain characteristics of portfolios. It is important to note that this research uses sentiment to predict stock direction which is then used as a trading strategy whereas the literature below trades stocks based on their sentiment values on that day, longing and shorting accordingly. Thus, there are some slight differences in interpretation due to the nuances of the data and methodology, but the final findings and conclusions of this research resonate with the current literature.

Relating to the work of Boudoukh et al. (2013), they find that all portfolios, except their short leg portfolio, yield abnormal returns while Ahmad et al. (2016) find a trading strategy that yields abnormal predominantly in their short leg. It is interesting to see such differences across Tables 4-9 also present, as there are certain portfolios of the DTC that yield significant abnormal returns, mainly the long-short portfolios and of the RFC that yield significant abnormal returns across certain short portfolios. This highlights that the data used for sentiment as well as the methodology used can yield slightly different results.

Certainly, using different data and different methodologies may yield different results. It is important to note the chance of a reversal effect highlighted by Pedersen (2022) in sentiment-driven trading strategies. Ahmad et al., (2016) highlighted that there are sentiment reversal effects at the company level, as there may be time periods in which sentiment is not correlated with stock market movements. Additionally, Ferguson et al. (2014) show that there are certain time frames when the aggregated market returns are correlated with sentiment highlighted by having significant abnormal returns during certain time frames, followed by others that do not. The time frame used in this research for predictions, May 2022-May 2024 may be subject to a lucky/ unlucky draw but by increasing θ, the model limits those lucky/unlucky predictions which is the reason why significant intercepts are only found for higher values of θ.

The portfolio analysis was extensively done due to the capabilities and the flexibility machine learning output. It is clear from prior research that sentiment analysis and stock returns are related. Given certain circumstances of certain portfolios, this research rejects the null hypothesis of no significant abnormal returns deeming sentiment predicted market movement a successful trading strategy leading to significant abnormal returns. It must be mentioned that the null hypothesis was rejected given the data, method and time frame used, since slightly altering either of those three components may alter the results also. In answering the research question, a conclusion can be stated that while for some portfolios the returns can be explained by the Fama & French (1993) risk factors, following a sentiment-driven strategy can yield additional returns unexplained by the risk, size and value risk factors.

6.3 Limitations and Further Research

This research uses a basket of 29 companies while also trading nearly daily across all portfolios. A limitation is clear that a sentiment driven strategy yields significant abnormal returns but with extremely high trading volume which comes with higher fees. There is an attempt at limiting the number of trades made by altering θ but the portfolios that have significant abnormal returns still trade nearly daily. Further, the trading period is limited to two years, which highlights one important factor in using machine learning methods, the training resources needed to compute valuable predictions.

Given the data available and gaps in the current literature, this research attempted at maximizing internal validity. To limit noisy predictions the method above filters the output variable through θ, rather than the input variables. Future research would benefit greatly from additional filtering methods to limit noise. The aggregation method used yields valuable results, but questions arise surrounding the idea of its additional benefit compared to the more classical methods (Ahmad et al., 2016) with only one aggregation dimension, time. The use of 10 lags and multiple sentiment values may lead to the models in finding spurious noisy coincidences rather than causal patterns and effects. This research is the first to use this aggregation technique and for further research, it is important to find first, if having more sentiment scores increases explanatory power, and second, which lags are relevant in predicting movement or returns. By finding the most valuable data, predictions are bound to get more accurate, which would also solve the problem of trading volume as well.

7. Conclusion

The purpose of this study was to explore the relationship between media-expressed sentiment tone and short-term stock returns. To answer the research question two hypotheses were employed: whether sentiment derived from news articles could predict stock movement and if a trading strategy based on these predictions would yield significant abnormal returns. By examining the sentiment from various publishers over a 10-year timeframe, this research uses the potential of advanced machine learning models to uncover patterns and correlations not readily observed by traditional methods such as OLS for panel fixed effects models.

The research employed a robust methodology that integrated a sophisticated method for sentiment analysis using an NLP and machine learning classifiers for stock movement prediction. Using an original aggregation technique of the publisher when aggregating news articles, this research opens a branch in existing opinion mining and stock returns literature, machine learning models, specifically Decision Tree Classifier and Random Forest Classifier. The findings in this research are attributable to using complex data manipulation, top of the line supervised machine learning classifiers and the attention given to internal validity of providing means to possible issues. The multi-step process concludes significantly that there are patterns at the sentiment level data correlated with tomorrow's stock price movement which yields significant abnormal returns.

While the predictive accuracy of these models was just above a coin toss chance, the portfolios constructed based on sentiment predictions yielded significant abnormal returns under certain conditions. This highlights the potential for sentiment-driven trading strategies to exploit market inefficiencies. The study concluded that while sentiment analysis can provide valuable insights and predictive power, the complexity of market dynamics and the noise in data necessitate careful consideration and further refinement of models and methodologies. This research contributes to the understanding of media sentiment aggregated at the publisher level and stock returns, suggesting that even minor but consistent predictive advantages can be translated into financial gains.

8. References

Ahmad, K., Han, J., Hutson, E., Kearney, C., & Liu, S. (2016). Media-expressed negative tone and firm-level stock returns. *Journal of Corporate Finance*, *37*(C), 152–172.

Alexander, Sidney S. 1961. Price Movements in Speculative Markets: Trends or Random Walks. *Industrial Management Review 2*, 7–21.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, *9*(1), 3–18.

Basari, A. S., Hussin, B., Pramudya, G., & Zeniarja, J. (2013). Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. *Procedia Engineering*, *53*, 453–462.

Boudoukh, J., Feldman, R., Kogan, S., & Richardson, M. P. (2013). *Which News Moves Stock Prices? A Textual Analysis* (SSRN Scholarly Paper 2207241).

Business Insider. (2024, March 8). "Dumb Money" follows Keith Gill and the GameStop stock surge — here's where the financial analyst and YouTuber is now. Business Insider. https://www.businessinsider.com/where-is-keith-gill-now-dumb-money-movie-gamestop-2023-9?international=true&r=US&IR=T#:~:text=Gill%20said%20he%20didn't,This%20lawsuit%20was%20dismissed.

De Bondt, W. F. M., & Thaler, R. (1985). Does the Stock Market Overreact? *The Journal of Finance*, *40*(3), 793–805.

Bhandari, L. C. (1988). Debt/Equity Ratio and Expected Common Stock Returns: Empirical Evidence. *The Journal of Finance, 40*(3), 507–528.

Elbannan, M. (2014). The Capital Asset Pricing Model: An Overview of the Theory. *International Journal of Economics and Finance*, *7*, 216.

Fama, E. (1970) Efficient Capital Market: A Review of Theory and Empirical Work. *Journal of Finance, 25*, 382-417.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56.

Feldman, R., Rosenfeld, B., Bar-Haim, R., & Fresko, M. (2011). *The Stock Sonar—Sentiment Analysis of Stocks Based on a Hybrid Approach. 2*.

Ferguson, N., Philip, D., Lam, H., & Guo, J. (2015). Media Content and Stock Returns: The Predictive Power of Press. *Multinational Finance Journal*.

Pedersen, L. H. (2022). Game on: Social networks and markets. *Journal of Financial Economics*, *146*(3), 1097-1119).

Guo, J. (1997). Critical Tokenization and its Properties. *Computational Linguistics*, *23*(4), 569–596.

Hu, D., Jones, C. M., Li, S., Zhang, V., & Zhang, X. (2021). *The Rise of Reddit: How Social Media Affects Retail Investors and Short-sellers' Roles in Price Discovery* (SSRN Scholarly Paper 3807655).

Ho, T.-T., & Huang, Y. (2021). Stock Price Movement Prediction Using Sentiment Analysis and CandleStick Chart Representation. *Sensors*, *21*(23), 7957.

Hutto, C. J., & Gilbert, E. (2015). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.

Jegadeesh, N., & Titman, S. (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of Finance*, *48*(1), 65–91.

Kenneth R. French - Data Library (2024). Retrieved 10 June 2024, from https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Khan, W., Ghazanfar, M. Ali, Azam, M. A., Karami, A., Alyoubi, K., & Alfakeeh, A. (2020). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing*, *13*.

Leung, K. M. (2007). Naive Bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering, 2007, 123-156.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In K. Bontcheva & J. Zhu (Eds.), *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60).

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, *27*, 16–32.

Molnar, C. (2024). *Interpretable Machine Learning*. Retrieved 15 June 2024, from https://christophm.github.io/interpretable-ml-book/

Mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis · Hugging Face. (2024). Retrieved 15 June 2024, from https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis

Nippani, S., & Arize, A. C. (2008). U.S. corporate bond returns: A study of market anomalies based on broad industry groups. *Review of Financial Economics*, *17*(3), 157–171

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, *2*, 1–135.

Patel, J., Zeckhauser, R., & Hendricks, D. (1991). The Rationality Struggle: Illustrations from Financial Markets. *The American Economic Review*, *81*(2), 232–236.

Press Releases | U.S. House Committee on Financial Services Democrats. (2021, February 12). U.S. House Committee on Financial Services Democrats. https://democrats-financialservices.house.gov/news/documentsingle.aspx?DocumentID=407152

ProQuest | Better research, better learning, better insights. (2024). Retrieved 15 May 2024, from https://www.proquest.com

Woo, K.-Y., Mai, C., McAleer, M., & Wong, W.-K. (2020). Review on Efficiency and Anomalies in Stock Markets. *In Economies, 8*(1), 20

Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk*. *The Journal of Finance*, *19*(3), 425–442.

Stagner, R. (1940). The Cross-Out Technique as a Method in Public Opinion Analysis. *The Journal of Social Psychology*, *11*(1), 79–90.

Sweeney, E. (2018, December 20). *Cigna closes $67B Express Scripts acquisition, promising affordability and choice | Fierce Healthcare*. https://www.fiercehealthcare.com/payer/cigna-closes-67b-express-scripts-acquisition

Roberts, H. V. (1959). STOCK-MARKET “PATTERNS” AND FINANCIAL ANALYSIS: METHODOLOGICAL SUGGESTIONS. *The Journal of Finance*, *14*(1), 1–10.

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.

Loughran, T., & McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65.

Yahoo Finance—Stock Market Live, Quotes, Business & Finance News. (n.d.). Retrieved 15 May 2024, from https://finance.yahoo.com/

Yu, H., & Hatzivassiloglou, V. (2003). Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 129–136.