Informational Value of Daily Firm-Specific Twitter Sentiment

Analysis on Short-Term Returns of Large-Cap U.S. Companies

Master's Thesis

for attainment of the academic degree
Master of Science (WU)

in partial fulfilment of the requirements of the programme
Quantitative Finance

submitted by
Alissia Hruštšova
(12011283)

supervised by
Univ.Prof.Dipl.-Ing.Dr.techn. Kurt Hornik

Vienna, Austria
11.08.2022

Declaration

I hereby declare that:

1. I have written this master's thesis myself, independently and without the aid of unfair or unauthorized resources. Whenever content has been taken directly or indirectly from other sources, this has been indicated and the source referenced. I am familiar with the regulations specified in the Directive on Plagiarism and Other Types of Academic Fraud in Academic Theses.

2. This master's thesis has not been previously presented as an examination paper in this or any other form in Austria or abroad.[1]

3. This master's thesis is identical with the thesis assessed by the examiner.

11.08.2022

Date

Signature

[1] This does not apply to master's theses written as part of WU cooperation programs (joint or double degrees).

Acknowledgements

First and foremost, I would like to extend my sincere gratitude to my thesis supervisor, Univ.Prof.Dipl.-Ing.Dr.techn. Kurt Hornik, for his unwavering guidance, not only throughout the duration of the master thesis project, but also in the course of the Quantitative Finance programme. I am especially thankful for his unparalleled philosophical approach to teaching, delivering complex concepts with a sense of simplicity by virtue of starting from the first principles, and his cordiality in all interactions with fellow students.

I would like to express my deepest appreciation to my mother Jelena, who through her courage and tireless work ethic provided me with all the necessary resources required to realise my full potential. I am forever indebted to my grandparents, responsible for nurturing my great fondness for numbers and, consequently, the discipline of Mathematics.

Last but not least, I am beyond grateful for sharing the QFin journey with my partner Max, who so generously welcomed me into his life, cured my homesickness and filled the countless intense study days with love and joy.

Abstract

This master's thesis examines firm-level contemporaneous and lagged relationships between Tweet-based features (i.e., Twitter activity and sentiment polarity) and stock market characteristics (i.e., holding period return and trading volume) at daily frequency. Our findings support the existence of contemporaneous association between Twitter mood and returns (turnover), together with the relation of Tweet volume and returns (turnover). Furthermore, we established that investor agreement and Tweet volume could be of use in predicting abnormal trading volume on the following day. In contrast to previous closely related research, we found no indication of Twitter-based investor sentiment and interest being helpful in predicting returns, which was subsequently reflected in the underwhelming performance of the corresponding long-short portfolios. Nevertheless, we observed that aggregation of firm-level Twitter sentiment can be beneficial in forecasting the performance of ETFs, stock indices and portfolios.

Contents

List of Tables

List of Figures

List of Acronyms

API application programming interface

BNB Bernoulli naïve Bayes

CV cross-validation

DTM document-term matrix

ETF exchange traded fund

HPR holding period return

ML machine learning

MNB multinomial naïve Bayes

NB naïve Bayes

NH negation handling

NLP natural language processing

SA sentiment analysis

WLOG without loss of generality

1 Introduction

From the early days of trading, prior to the onset of the modern financial market era, institutional and retail investors alike have paid close attention to the public opinion surrounding stock market events and participants, in the hope of extracting valuable information on company fundamentals. As an illustration, on November 18 1783, the British prime minister William Cavendish-Bentinck (also known as Fox) made a public statement concerning the British East India Company (EIC), the stock of which traded in both London and Amsterdam (Koudijs, 2016). He claimed that the finances of the company were in "deplorable state", leading to a dramatic decline in the stock price in London. With a long delay of 10 days due to unfavourable conditions at sea, the stock price in Amsterdam adjusted accordingly with the arrival of the first ship.

The technological revolution has had crucial consequences on the speed of information transmission, limiting the possibilities for exploiting arbitrage opportunities based on asymmetric information. Moreover, while the communications of the European Central Bank (ECB) and the Federal Open Market Committee (FOMC) continue to spread waves through the markets (Schmeling et al., 2019), the evolution of social media platforms, such as Twitter, Reddit and StockTwits, has amplified the voices of amateur investors and stock market enthusiasts, making it possible for seemingly irrational views to rapidly gain in momentum and subsequently result in substantial shifts in the stock prices. For instance, in early 2021 the fanatic discussions on the r/WallStreeBets subreddit brought about the sensational short-squeeze of the Gamestop stock, with the stock price surging more than 700% in one week (Long et al., 2021).

While the aforementioned cases emphasise the impact of influential figures on share price movements, researchers have long pondered over the value of extracting and aggregating sentiment from messages published on stock microblogging forums, disregarding the followership of the authors. More recently, Gu et al. (2020) made use of the Twitter sentiment measure provided by Bloomberg, in order to examine whether Tweets provide information not already reflected in the stock prices. The authors found that firm-specific Twitter sentiment predicts interday stock prices without subsequent reversals and provides new information about analyst recommendations, analyst price targets and quarterly earnings. Similarly, Duz Tan et al. (2021) found that both Twitter sentiment and activity have significant predictive power with respect to stock returns and abnormal trading volume at firm level.

Motivated by the works of Gu et al. (2020) and Duz Tan et al. (2021), we decided to make use of their methodology in order to investigate whether firm-level Twitter sentiment and volume could convey valuable private signals about firm prospects and, therefore, possess significant predictive power with respect to daily returns and trading volumes of large-cap U.S. stocks. The cross-section of stocks, selected for our study, consisted of all S&P500 constituents during the sampling period beginning on 1 June 2021 and ending on 30 November 2021. Seeing that the Bloomberg sentiment score construction algorithms are proprietary in nature, one of our key intentions was to enhance the transparency of the approach by training our own sentiment polarity classification model and disclosing all relevant details.

The results of our analysis provide evidence in support of the existence of contemporaneous associations between the Tweet-based and stock market features discussed above, as also documented by Duz Tan et al. (2021), Sprenger, Tumasjan, et al. (2014), Antweiler et al. (2004) and Li et al. (2018). Nonetheless, neither Twitter sentiment nor activity seem to have significant predictive power with respect to stock returns at inter-day level, contradicting the abovementioned findings of Gu et al. (2020) and Duz Tan et al. (2021). Furthermore, our results suggest that firm-level Twitter sentiment can be aggregated into a single measure of sentiment, which could be beneficial in forecasting the performance of ETFs, stock indices and portfolios.

The thesis is organised in the following manner. Section 2 provides further details regarding the related work and research questions. Section 3 describes the data retrieval process in conjunction with the applicable sources. Section 4 walks the reader through the training of the sentiment polarity classification model and the experiments conducted with the resulting sentiment scores. Section 5 discusses the interpretations and implications of the empirical results, together with the limitations of the study and recommendations for further research. Finally, the conclusions are drawn in Section 6.

2 Related Work and Research Questions

In 1990, De Long et al. (1990) presented an asset market model, where sophisticated arbitrageurs and noise traders, guided by fallacious pseudosignals, enter trades, based on their fundamental value expectations. The actions of noise traders, driven by their fluctuating sentiment, cause the prices to diverge from the levels reflective of company fundamentals, bearing a risk to rational investors, who are unable to foresee the price reversion to the mean. This theory introduces the initiative of trading profitably on the basis of public opinion, by either undertaking the contrarian standpoint or jumping on the bandwagon of the noise traders.

Wysocki (1998) is acclaimed for his pioneering work in investigation of the internet stock message boards, inferring that message volume could be helpful in predicting the trading volume of frequently discussed firms. The findings of Antweiler et al. (2004) and Sprenger, Tumasjan, et al. (2014) support the hypothesis, with the authors taking the matter a step further and introducing firm-focused sentiment into the discussion. Their results also suggest that increased bullishness (agreement) of stock microblogs is contemporaneously associated with higher (lower) returns, while both investor bullishness and disagreement may be valuable in predicting trading volume.

With the vast majority of studies, concerned with the stock market impact of Twitter sentiment, either focusing on aggregate return effects (see e.g. Mao et al., 2015; Bollen et al., 2011) or examining return predictability around specific events, such as FOMC meetings (Azar et al., 2016), earnings announcements (Bartov et al., 2018) or abnormal trading volume spikes (Ranco et al., 2015), little is known about information content of firm-specific Twitter sentiment, as highlighted by Gu et al. (2020).

In their recent studies, Gu et al. (2020) and Duz Tan et al. (2021) found that firm-specific Twitter sentiment predicts interday stock prices, along with abnormal turnover amounts. This prompted us to revisit the question of whether firm-level Twitter sentiment and volume can explain daily returns and trading volumes of large-cap U.S. stocks and, if so, to subsequently examine whether some Twitter characteristics could contain fundamental information which is not incorporated into the stock prices, considering a 1-day delay.

The significance of our research lies in provision of up-to-date results in an area of limited research (owing to the challenge of retrieving large amounts of firm-level data), by means of classifying relevant texts with a sentiment polarity classifier, trained using a large dataset of labelled data (in contrast to Antweiler et al., 2004; Sprenger, Tumasjan, et al., 2014, who only used 1000 and 2500 texts for training their classifiers, respectively), without reliance on external proprietary products (e.g. Bloomberg Twitter and news sentiment measures, applied by Gu et al., 2020; Duz Tan et al., 2021). The evolution of social media platforms (e.g. omnipresence of bots and spammers), together with the growing reliance of automated trading algorithms on the microblog chatter (believed to be at the center of the 2010 flash-crash, as discussed by Cresci et al., 2019), highlights the necessity of periodic re-examination of established "truths" in relation to connection of the media and markets.

We posed the following hypotheses in line with our research question, with the applicable results presented in Section 4 and further discussed in Section 5. For the following hypotheses, both contemporaneous and lagged relationships between features were considered:

H1: *The more "bullish" the sentiment reflected on stock microblogs is, the higher are the returns.*

As explained by Duz Tan et al. (2021), one could assume Tweets of "bullish" ("bearish") sentiment to implicitly assign "buy" ("sell") labels to the relevant stocks, encouraging fellow stock microbloggers to follow up on the advice. Therefore, one would expect the pressure on the "buy" side of the order-book to increase together with the optimism of Twitter's stock enthusiasts, driving up the stock prices and, therewith, the returns.

H2: *Increased message volume on stock microblogs is associated with an increase in returns.*

Sprenger, Tumasjan, et al. (2014) argue that an increase in Tweet count may be viewed as a signal implying arrival of new information. Taking into account that investors tend to share more bullish than bearish messages in online forums (Sprenger, Sandner, et al., 2014), an increase in Tweet count could be associated with higher returns.

H3: ***a.*** *The more "bullish" the attitude of stock microblogs is, the more active is the trading activity.*

b. *Increased message volume on stock microblogs is associated with increased abnormal turnover.*

In 2003 Van Bommel (2003), aiming to shed light on the role of rumours in stock market movements, brought forward a rumour-mongering strategy. The author argued that small informed investors bear rumours in order to increase their information-driven profits, by inspiring their followers to act in accordance with their recommendations and, thereby, gaining a reputation. Through their collaborative action, the influenced individuals may induce a surge in demand for the given stock, together with an accompanying price increase. Assuming that the associated change in demand does not lead to a pronounced imbalance within the order-book, one may expect the trading volumes and stock prices to increase, along with increasing message volume, in light of this theory.

H4: *Increased disagreement among users of stock microblogs is associated with higher abnormal trading volume.*

Intuitively, high level of disagreement would signify the investors being evenly divided in their views (i.e., they assign different values to the same asset), therewith balancing the supply and demand for the stock in question. In traditional financial theory, it is commonly hypothesised that the divergence of opinions can therefore explain the upswings in trading volumes of applicable stocks (Harris et al., 1993).

3 Data Collection

The six-month period, beginning on 1 June 2021 and ending on 30 November 2021, was chosen as the sampling period for the study. This time interval incorporated 128 trading days. The research concerned itself with daily stock characteristics. We selected all stocks included in S&P 500 for the entire duration of the analysis, i.e. the stocks added to or removed from the index in the course of the experimental period were not considered. Additionally, same companies listed on different exchanges were removed. As a result of this filtering procedure, 498 stocks were included in the sample. Consequently, we had 63744 stock-day observations in total, before accounting for non-missing observations of firm-specific Twitter sentiment.

3.1 Sentiment Polarity Labelled Data

As previously stated, transparency of sentiment polarity classification approach applied to Tweets was of utmost importance in the investigation. On account of this and further reasons outlined in Section 4, we prioritised learning with supervision when choosing machine learning (ML) classification techniques.

As highlighted by Renault (2020), supervised classification models pose a challenge for researchers, since they necessitate retrieval of data labelled with classes of interest. Generally, a considerable sample of texts, manually annotated by human experts, is preferred, which can be relatively expensive or time-consuming, as mentioned by Ranco et al. (2015). Nonetheless, the authors acknowledge some accompanying benefits, including the ability to establish an upper bound on the classifier performance (given that several annotators label the texts), the domain- and language- specificity of the annotation process and the possibility to automise the classifier construction once sufficient labelled data is available.

Due to feasibility concerns we did not manually annotate any sample texts in the course of the study. Instead, we decided to implement a selection of domain-appropriate open-source datasets in the training process.

3.1.1 Training Set

The following collections of short texts were mixed together in fitting proportions (once the relevant pre-processing steps have been applied) to create a balanced training set for the model. Since we considered only two classes (positive and negative) in the study and the labels used differed between the sets, we changed some labels accordingly (for example, "bullish" and "strongly positive" were changed to "positive") and removed texts corresponding to unused tags (such as "neutral") from the sample.

Financial Phrase Bank A dataset of 5,000 sentences, selected at random from news reports on the components of the OMX Helsinki index. Financial Phrase Bank was constructed by Malo et al. (2014) and subsequently used by Araci (2019) to implement the pre-trained FinBERT model, designed to handle natural language processing (NLP) tasks in finance. Sixteen individuals with adequate knowledge on financial markets labelled the sentences and the strength of inter-annotator agreement was documented.

StockTwits posts This sample of StockTwits[1] (a social media platform for investors) posts was collected by two Carnegie Mellon University (CMU) students and stored in a public GitHub repository[2]. The dataset contains microblog messages regarding fourteen popular stocks, some of which have been labelled as "Bullish" or "Bearish" by the authors of the posts. As StockTwits did not accept any new registrations during the study period, this source was highly valuable, as it supplied reliably labelled Tweet-like messages related to the stock market.

COVID-19 Tweets As mentioned by Sharif et al. (2020), "news regarding oil prices and the COVID-19 outbreak seem to be the irresistible drivers of the US stock market", which underlines the importance of including relevant texts in the sample. This sample of over 40,000 Tweets related to the COVID-19 pandemic was collected by A. Miglani, who self-tagged the Tweets and published the resulting dataset on Kaggle[3].

Bitcoin Tweets A collection of more than 50,000 Bitcoin-focused Tweets, published on data.world[4] (an enterprise data catalog), which contributed a vast supply of domain-specific Tweets. Alarmingly, the publisher failed to specify the details of the annotation method used, undermining the trust in the quality of the data. For this reason, only a portion of the sample was utilised.

Sentiment140 Constructed by Go et al. (2009), Sentiment140 is one of the commonly used datasets in sentiment analysis (SA) research, accounting for the bulk of the train-

[1] https://stocktwits.com/

[2] https://github.com/tdrussell/stocktwits_analysis

[3] https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification

[4] https://data.world/

ing set with its contribution of aproximately 150,000 weakly labelled Tweets. The messages were automatically tagged using a simple rule, where all Tweets containing positive (negative) emoticons such as ":)" (":(") were assumed to be of positive (negative) sentiment. The purpose of including this dataset in the analysis was to achieve a sample size of 100,000-250,000 texts, as recommended by Renault (2020).

3.1.2 Validation Set

Furthermore, we combined two corpora to create a validation set of almost 2,000 Tweets. As the training data comprised datasets from a variety of domains, the objective of the validation set was to indicate, whether the performance on Tweets concerning stocks was comparable to that suggested by the 5-fold cross-validation results on the training set.

Sanders Twitter sentiment corpus The Sanders corpus consists of close-to 5,000 Tweets about Amazon, Apple, Facebook and Twitter, all of which were tagged by N. Sanders, the founder of Sanders Analytics. As the company no longer exists, the dataset can only be retrieved using external sources such as GitHub[5].

Stock market Tweets Taborda et al. (2021) collected over 900,000 Tweets containing references to S&P500, the top 25 stocks in the index and/or the hashtag "#stocks". Two annotators independently labelled 1,300 of these Tweets, releasing the ready dataset to the research community.

3.2 Firm-Level Tweets

The classifier demonstrating optimal performance (among the candidate models) was subsequently applied to firm-specific Tweets, extracted using one of the official Twitter application programming interfaces (APIs). In particular, we submitted an application to avail of the Academic Research access features, which included a Tweet cap of 10,000,000 Tweets per month (in comparison to 100,000 Tweets per month with the basic developer account) and access to the "full-archive search" endpoint, allowing users to make queries for historical Tweets, dating all the way back to 2006.

Once the special access was approved, `academictwitteR`, a specialised CRAN package written by Barrie et al. (2021), facilitated smooth interaction with the relevant endpoints, resulting in over 20,000,000 Tweets being extracted. The collection rules were limited and simple: each Tweet had to contain a cashtag of one of the stocks considered, while all non-English Tweets, Retweets and Quotes were excluded.

[5] `https://github.com/zfz/twitter_corpus`

3.3 Stock Market & Financial Data

Lastly, we required some stock market and financial data for computation of response and control variable values. The opening, high and low prices, dividend amounts and ex-dividend dates were obtained using Yahoo Finance website[6] (via `quantmod`, designed by Ryan et al., 2022), Alpha Vantage API[7] (via `alphavantager`, put together by Dancho et al., 2020) and S&P Capital IQ database[8]. Similarly, we retrieved the trading volume information using Yahoo Finance and Alpha Vantage, while S&P Capital IQ provided the numbers of shares outstanding for the stocks.

[6] `https://finance.yahoo.com/`

[7] `https://www.alphavantage.co/`

[8] `https://www.spglobal.com/marketintelligence/en/`

4 Analysis

4.1 Sentiment Polarity Classification

As pointed out in Section 3.1, the learning of the sentiment polarity classifier was conducted with supervision. The rationale underlying the decision is straightforward. James et al. (2013) explain that while supervised learning is a well-understood area with a clear understanding of the classes and performance assessment methods, unsupervised learning tends to be more challenging due to the subjectivity surrounding the produced results and absence of universally accepted mechanisms for quality evaluation. Evidently, the supervised approach was more closely aligned with our research objective of detailing the sentiment score construction process.

Some of the commonly used supervised classification techniques include classification trees, naive Bayes, random forests, neural networks and support vector machines. The following considerations were made when designating an approach for use in the study. Renault (2020) found that the use of more complex algorithms did not lead to any significant improvements in performance when classifying the sentiment polarity of StockTwits messages, while the training set size proved itself to be of crucial importance. By virtue of these findings, algorithms of lower time and space complexities were favoured in the selection procedure. Moreover, the possibility of probabilistic interpretation of results served as an advantage, as it allowed the sentiment scores to be computed as confidence-weighted averages of individual Tweet scores, analogously to Cui et al. (2016). One of the methods satisfying all of the aforementioned criteria is naïve Bayes (NB), which was ultimately selected to perform the classification task on firm-level Tweets.

4.1.1 Naive Bayes Algorithm

We begin with a description of the algorithm and refer the reader to Jurafsky et al. (2022) for a detailed discussion. The NB model is a supervised probabilistic learning method, based on the Bayes' theorem, which makes a "naïve" simplifying assumption about the nature of interaction between the features.

Within our framework, D and C denote random variables, representing the document (Tweet) and document class (positive or negative), acquired when making a random selection from a pool of documents (Tweets). Using the Bayes' theorem, we can derive the proportionality

$$P(C|D) \propto P(C)P(D|C),$$

where $P(C)$ and $P(C|D)$ are known as the prior and posterior probabilities, respectively, and $P(D|C)$ is referred to as the likelihood. Let Ω_C denote the sample space of all classes. Given some document d which we want to classify, NB classifier outputs $\hat{c}$ as its estimated class, where

$$\begin{aligned}\hat{c} &= \arg\max_{c\in\Omega_C} P(C=c|D=d)\\ &= \arg\max_{c\in\Omega_C} P(C=c)P(D=d|C=c).\end{aligned}$$

As the probability distributions of C and $D|C$ are not known, we require a method to estimate $P(C=c)$ and $P(D=d|C=c)$ for each class $c \in \Omega_C$.

Without loss of generality (WLOG), we assume that every document can be represented as a set of features (for example, words or unigrams, which we assume for simplicity). Let $F = \{f_1, f_2, ..., f_{N_F}\}$ be the set of all unique features, remaining after feature selection, which correspond to our collection of documents used to train the classifier. If the representation of d contains none of the features in F, the class corresponding to the highest estimated prior probability is assigned as the estimated class of d. Let F_i denote a random variable, which takes on counting values representing the frequency of f_i in d and let the feature counts in d be given by $n_{f_1}, n_{f_2}, \ldots, n_{f_{N_F}}$.

To reduce the dimensionality of the problem, we assume that there are no features directly specifying the positioning of words in the documents (the so-called "bag-of-words" assumption). Furthermore, we "naïvely" assume independence of features and the underlying distribution to be multinomial, implying that

$$\begin{aligned}\hat{c} &= \arg\max_{c\in\Omega_C} P(C=c)P(D=d|C=c)\\ &= \arg\max_{c\in\Omega_C} P(C=c)P(F_1=n_{f_1}, F_2=n_{f_2}, \ldots, F_{N_F}=n_{f_{N_F}}|C=c)\\ &= \arg\max_{c\in\Omega_C} P(C=c)\frac{n_d!}{\prod\limits_{1\le i\le N_F} n_{f_i}!}\prod_{i=1}^{N_F} P(F_i=n_{f_i}|C=c)\\ &= \arg\max_{c\in\Omega_C} P(C=c)\prod_{i=1}^{N_F} P(F_i=n_{f_i}|C=c)\end{aligned}$$

$$= \underset{c \in \Omega_C}{\arg\max} \left[\log P(C = c) + \sum_{i=1}^{N_F} \log P(F_i = n_{f_i} | C = c) \right],$$

where n_d is the total count of features from F in d and log scale is applied to avoid underflow and increase calculation speed.

A common way to estimate the priors is to use the maximum likelihood estimates, i.e. if N_c is the number of documents of class c and N_D is the number of documents in the training set D, we let $\hat{P}(C = c) = \frac{N_c}{N_D}$. As for the conditional probabilities, for each feature $f_i \in F$, where $i \in \{1, 2, \ldots, N_F\}$, and class $c \in C$, we let $N_{f_i,c}$ denote the total count of occurrences of f_i within D_c, the collection of all documents of class c in D. Denoting the total count of features from F in D_c by $N_{F,c}$, we take $\hat{P}(F_i = n_{f_i} | C = c) = \left(\frac{N_{f_i,c}}{N_{F,c}}\right)^{n_{f_i}}$.

However, maximum likelihood estimation for the conditional probabilities can be problematic. Suppose that some feature occurs in a positive-class document but not in any negative-class documents in D . The probability corresponding to the "negative" class would then be estimated as 0. If the respective feature is present in d, its likelihood for the "negative" class would be estimated as 0. On that account, one would often apply Laplace smoothing, where counts of all features are increased by a positive number (not necessarily an integer) α, such that $\hat{P}(F_i = n_{f_i} | C = c) = \left(\frac{N_{f_i,c}+\alpha}{N_{F,c}+\alpha N_F}\right)^{n_{f_i}}$.

Algorithms 1 and 2 outline the procedures for training and applying a MNB classifier, as described above. Jurafsky et al. (2022) suggest that in the case of SA and text classification tasks, performance can generally be improved through the use of Boolean values, rather than counts, to record presence or absence of features in documents. This can be achieved by changing the assumed distribution of features to the corresponding multivariate Bernoulli distribution, where $n_{f_1}, n_{f_2}, \ldots, n_{N_F}$ would take on one of two values, namely 1, if the corresponding feature appears in d, or 0, otherwise, which would give us a BNB model.

Algorithm 1 Training MNB Classifier

```
Input: $D$, $\Omega_C, \alpha$
Output: $F$, $LogPriors$, $LogCondProbs$
  $F \leftarrow$ extract and select features from $D$
  $N_F \leftarrow$ count features in $F$
  $N_D \leftarrow$ count documents in $D$
  for each $c$ in $\Omega_C$ do
      $N_c \leftarrow$ count documents of class $c$ in $D$
```

```
        $LogPriors[c] \leftarrow \log\left(\frac{N_c}{N_D}\right)$
        $D_c \leftarrow$ select documents of class $c$ in $D$
        $N_{F,c} \leftarrow$ count instances of features from $F$ in $D_c$
        for each $f$ in $F$ do
            $N_{f,c} \leftarrow$ count instances of $f$ in $D_c$
            $LogCondProbs[f][c] \leftarrow \log\left(\frac{N_{f,c}+\alpha}{N_{F,c}+\alpha N_F}\right)$
        end for
    end for
    return $F$, $LogPriors$, $LogCondProbs$
```

Algorithm 2 Applying MNB Classifier

```
Input: $F$, $\Omega_C$, $LogPriors$, $LogCondProbs$, $d$
Output: $c$
    $F_d \leftarrow$ extract features present in both $d$ and $F$
    for each $c$ in $\Omega_C$ do
        $Scores[c] \leftarrow LogPriors[c]$
        for each $f$ in $F_d$ do
            $Scores[c] \mathrel{+}= LogCondProbs[f][c]$
        end for
        $c \leftarrow \arg\max_{c \in \Omega_C} Scores[c]$
    return $c$
```

4.1.2 Considered Features

We now discuss the features which were involved in the experimental stage of the training process. Conventionally, n-grams (combinations of n words in their original order) constitute a major part of the feature space, with mixtures of unigrams, bigrams and trigrams considered most frequently. Nevertheless, in the context of microblogs, the inclusion of trigrams seems to be of no significant benefit, as documented by Bermingham et al. (2010) and Renault (2020), which led us to exclude trigrams from the analysis.

For selection of unigrams, we used `tm` (Feinerer et al., 2020), a text mining CRAN package, which allowed us to remove all "sparse" (or rare) terms by adjusting the threshold for the relative document frequency of terms. Different principles were applied when identifying valuable bigrams. In particular, we used likelihood ratio tests to pinpoint the unique phrases, whose terms were more likely to be found together than apart.

Before proceeding further, we briefly summarise the aforementioned likelihood ratio test approach and refer the reader to Manning and Schutze (1999) for details. We let w_1w_2 denote a bigram, where w_1 and w_2 represent the first and second constituting words. Under the null hypothesis (H_0), we assume the occurrence of w_2 to be independent of whether w_1 occurs before it in a selected bigram, i.e. $P(w_2|w_1) = p = P(w_2|\neg w_1)$. The alternative hypothesis (H_A) is then given by $P(w_2|w_1) = p_1 \neq p_2 = P(w_2|\neg w_1)$. We represent the count of bigrams with w_1 (w_2) as their first (second) word by c_1 (c_2) and the count of $w_1 w_2$ within the corpus by c_{12}. The estimated probabilities are computed as

$$\hat{p} = \frac{c_2}{N}, \; \hat{p_1} = \frac{c_{12}}{c_1}, \; \hat{p_2} = \frac{c_2 - c_{12}}{N - c_1},$$

where N gives the total number of bigrams. We assume the counts to be binomially distributed, meaning that the probability mass function takes the form

$$f(k; n, x) = \binom{n}{k} x^k (1-x)^{n-k},$$

with n standing for the number of all occurrences from which a selection is made, while k and x describe the observed count for instances of interest and the associated probability, respectively. The implied log-likelihood ratio takes the form

$$\log \lambda = \log \frac{f(c_{12}; c_1, p) f(c_2 - c_{12}; N - c_1, p)}{f(c_{12}; c_1, p_1) f(c_2 - c_{12}; N - c_1, p_2)},$$

such that $-2 \log \lambda$ is asymptotically χ^2 distributed and H_0 is rejected for significantly high values of this test statistic.

Apart from n-grams, some additional features were incorporated into the model. Jurafsky et al. (2022) mention that on occasion it is beneficial to include dense lexicon features, especially when sparse document-term matrices (DTMs) are used to train text classifiers. For this reason, we decided to make use of the two field-specific polarity lexicons, which Renault (2017) constructed from 750,000 StockTwits messages, following the automated procedure of Oliveira et al. (2016), based on several statistical measures. According to Kolchyna et al. (2015), features describing document length, counts of special symbols and negations can serve a similar purpose to lexicon features, which we put to a test during the training phase.

4.1.3 Pre-Processing Methods

As documented by Uysal et al. (2014) and Renault (2020), it is possible to significantly improve classification accuracy by choosing an appropriate pre-processing procedure. To make the search for a suitable step sequence feasible, we divided the methods into two categories - mandatory and optional steps. While the mandatory steps were undertaken with certainty, performance impact of the optional steps was brought into question. In regard to the latter category, the focus was placed on the potential effects of including punctuation marks, emojis and emoticons, in addition to exclusion of stop-words (as per pre-compiled list), stemming and lemmatisation of tokens and negation handling.

Punctuation Symeonidis et al. (2018) highlight that although punctuation removal is a standard technique in the fields of data mining and information retrieval, oftentimes the presence of punctuation marks points to existence of some sentiment. Specifically, punctuation can be used to intensify the sentiment captured by alternative sentence features. Nevertheless, there is no general recommendation regarding the method, as indicated by differing results in the literature, with some authors strongly advocating for punctuation inclusion (see eg. Renault, 2020; Renault, 2017) and others taking the opposite standpoint (see eg. Symeonidis et al., 2018). Our punctuation mark set included the following symbols: !, ?, %, +, -, =, :, ;,), (and].

Emojis and emoticons The definitions of Kralj Novak et al. (2015) describe an emoticon as a "shorthand for facial expression", a non-verbal aid in conveying moods and feelings in written messages, capturing attention and enhancing mutual understanding. Emojis can be thought of as the new generation of emoticons, characterised as graphical symbols representing a wider variety of concepts, including professions, cultures and celebrations. Both are used extensively either as features (see e.g. Mahmoudi et al., 2018; Da Silva et al., 2014) or as noisy labels (see e.g. Go et al., 2009) in sentiment polarity classification. All emoticons and emojis were extracted with the help of `qdapRegex` (Rinker, 2022) and `emoji` (Hvitfeldt, 2021) CRAN packages, respectively.

Stopwords As discussed by Saif et al. (2014), in order to partially filter out noise from textual data, researchers commonly reduce feature space dimensionality by removing non-discriminative (or uninformative in regard to sentiment polarity) words. The results documenting the effect on performance are contradictory, with some works favouring the procedure (see e.g. Haddi et al., 2013) and others warning of potential detrimental consequences (see e.g. Symeonidis et al., 2018). We employed a combination of dynamic stopword removal techniques in an effort to avoid memory usage issues. In particular, we eliminated all terms consisting of less than three characters, as well as rare terms, as mentioned in Section 4.1.2, at all times. Removal of singletons, an analogue of the latter method, evidently provides "the best trade-off" between feature count and classification accuracy (see Saif et al., 2014). Furthermore, the impact of using a pre-compiled list was examined by additionally removing words found in the English stopword list constructed by the Snowball Stemmer project[1].

Stemming and lemmatisation These two mutually exclusive techniques share the common goal of transforming terms into their base form. Manning, Raghavan, et al. (2010) describe stemming as a "crude heuristic process", removing word endings without

[1] `https://snowballstem.org/projects.html`

any knowledge of the context. Lemmatisation on the other hand involves the use of vocabularies and morphological word analysis, attempting to arrive at the dictionary form (lemma) of a given word. As for the abovementioned methods, the findings are not unilateral. For instance, Symeonidis et al. (2018) identified stemming as one of the most critical among sixteen pre-processing techniques when applied to Tweets, while Kolchyna et al. (2015) recorded a decrease in classification accuracy associated with both stemming and lemmatisation. The Snowball project implementation of Porter's (Van Rijsbergen et al., 1980) stemmer and the lemmatizer in `textstem` (Rinker, 2018) based on the morphological analyser of `hunspell` (Ooms, 2020) were utilised in our study.

Negation handling Kiritchenko et al. (2014) define negation as a "contextual sentiment modifier", the reason being that in a negated context many words change either their polarity or their sentiment intensity. They recognise the key aspects of automatic negation handling (NH) to be the identification of negating words, negation scope and negation impact. Referring to Taboada et al. (2011), they state that negating words are generally selected from small hand-crafted lists, which led us to construct our own list, consisting of the words "no", "not", "never", "without", as well as those ending in "n't". Regarding the scope of negation, Symeonidis et al. (2018) mention that researchers often assume the words between the negating word and the next punctuation mark to be affected. Others (see e.g. Farooq et al., 2016) capture negation scope using static windows of words (fixed number of words following a negating word). We employed three related methods, particularly static windows of one (SW1) and two (SW2) words and negation until next punctuation mark (PM). Lastly, the "NEG_" prefix was added to the affected words to account for the potential impact, analogously to Haddi et al. (2013), Symeonidis et al. (2018) and Renault (2017).

We summarise the complete pre-processing procedure with the help of the flowchart, depicted in Figure 4.1. We use "remove" as a shorthand for "replace by whitespaces". Lastly, we note that certain special symbols were occasionally replaced by their corresponding "codewords" (for example, intra-word contractions and dashes were temporarily changed to "A" and "H", respectively), to ensure that the intermediate steps did not result in their omission.

4.1.4 Classifier Training

To facilitate model comparison, a benchmark model was initially selected and subsequently updated, as the classifier was progressively optimised. The sparsity level was fixed at 0.9999 and the Laplace smoothing parameter α was set equal to 1, frequently used as its default value. As our training set was balanced with respect to sentiment class and the classes carried equal importance (i.e. correct guesses for the two classes

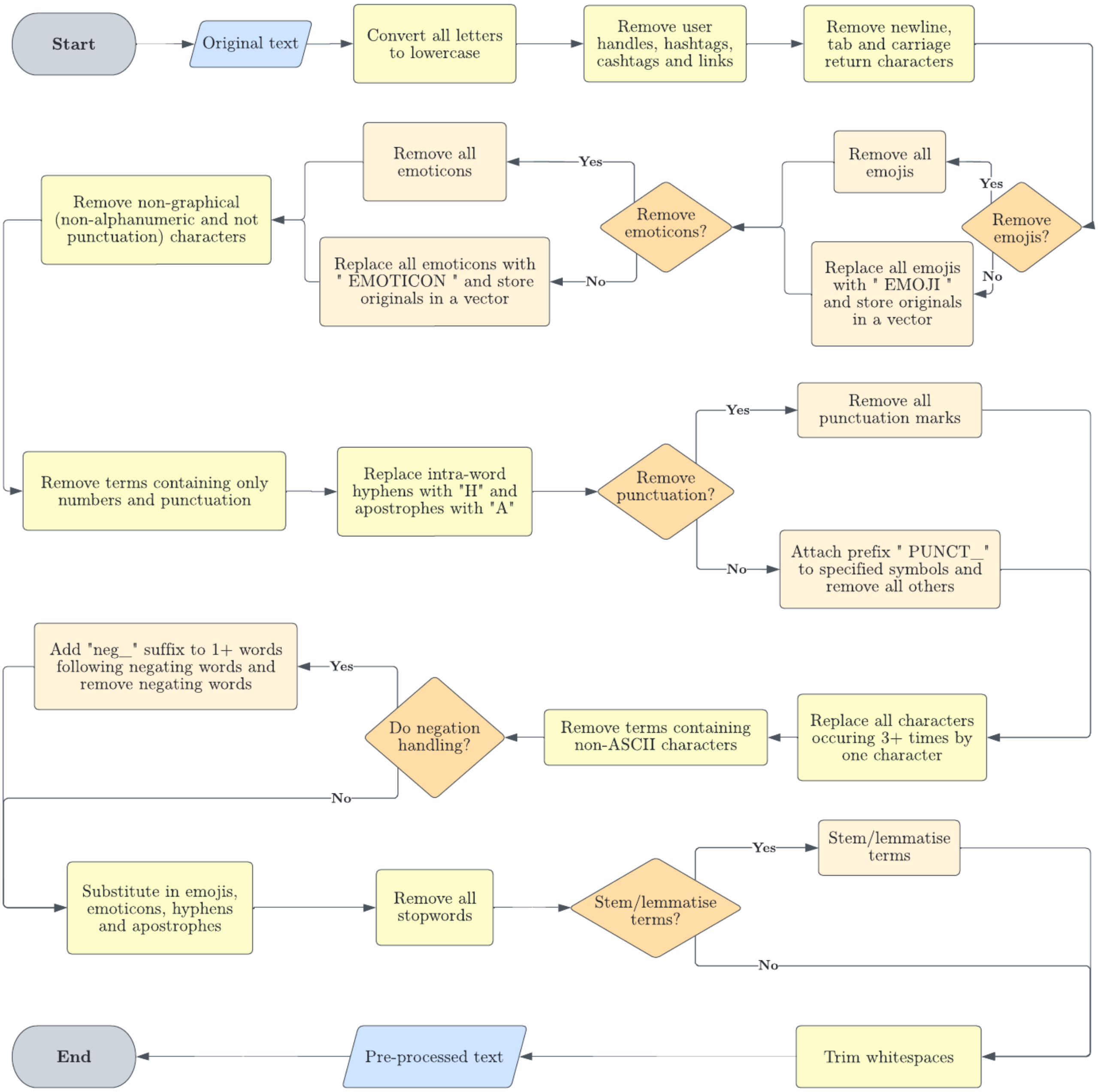
Start
Original text
Convert all letters to lowercase
Remove user handles, hashtags, cashtags and links
Remove newline, tab and carriage return characters
Remove emojis?
Yes
Remove all emojis
No
Replace all emojis with " EMOJI " and store originals in a vector
Remove emoticons?
Yes
Remove all emoticons
No
Replace all emoticons with " EMOTICON " and store originals in a vector
Remove non-graphical (non-alphanumeric and not punctuation) characters
Remove terms containing only numbers and punctuation
Replace intra-word hyphens with "H" and apostrophes with "A"
Remove punctuation?
Yes
Remove all punctuation marks
No
Attach prefix " PUNCT_" to specified symbols and remove all others
Replace all characters occuring 3+ times by one character
Remove terms containing non-ASCII characters
Do negation handling?
Yes
Add "neg_" suffix to 1+ words following negating words and remove negating words
No
Substitute in emojis, emoticons, hyphens and apostrophes
Remove all stopwords
Stem/lemmatise terms?
Yes
Stem/lemmatise terms
No
Trim whitespaces
Pre-processed text
End

Figure 4.1. Pre-processing steps applied to texts prior to feature extraction

were viewed as equally valuable), the performance metrics used were accuracy, sensitivity and specificity, given by

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}},$$

$$\mathrm{S}^{+} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

$$\mathrm{S}^{-} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}},$$

respectively. TP (FP) stands for the number of true (false) positives, which in our case was the count of documents of positive polarity classified correctly (incorrectly). TN and FN were analogously defined as the counts of true and false negatives, respectively. Following their conversion to percentages, the average values of these performance metrics, resulting from 5-fold cross-validation (CV) on the training set, as well as those computed from applying the trained classifier to the validation set, were used to compare the models against each other and the benchmark models.

We first compared the impact of incorporating the optional pre-processing steps, discussed in the previous section, one-by-one into the training of both BNB and MNB models, the results for which are presented in Table 4.1 and Table 4.2, respectively. We denote the benchmark model by B1. In its training, punctuation, emojis and emoticons were removed, no pre-compiled stopword list was used and none of the stemming, lemmatisation and negation handling procedures were applied. One can observe that MNB models seemingly outperformed the BNB counterparts in the majority of the cases. As a result, our focus for the upcoming feature selection stages was placed on MNB classifiers.

The observed negative impacts of punctuation inclusion and removal of stopwords from a pre-compiled list are consistent with Symeonidis et al. (2018) and Saif et al. (2014), respectively. Both techniques resulted in considerable drops in classification accuracy of 0.35 and 0.85 p.p. (in the case of MNB), respectively. On the other hand, retaining emoticons and emojis slightly improved the performance scores, with the positive effect of the latter method also documented by Mahmoudi et al. (2018).

Although neither stemming nor lemmatisation seemed to improve the classification accuracy, we considered stemming in the further steps, since the fixed sparsity value reduced the corresponding DTM size by 1000-2000 terms (when compared to DTMs of classifiers related to the alternative methods), the reduced dimensionality implied increased flexibility with regard to hyperparameter tuning (i.e. sparsity could be significantly increased) and the results for the validation set were noteworthy. The positive impact of SW2 negation handling was most impressive, as also noted by Hogenboom et al. (2011), and linked to a 0.45 p.p. increase in accuracy (for MNB).

	Training set			**Validation set**		
Model	ACC	S^+	S^-	ACC	S^+	S^-
Benchmark (**B1**)	75.83	73.41	78.19	71.34	70.60	72.16
B1 + Punctuation	75.04	72.59	77.46	70.60	66.67	75.00
B1 + Emojis	76.00	73.49	78.45	70.86	70.10	71.71
B1 + Emoticons	75.97	73.64	78.24	72.16	71.76	72.61
B1 - Stopwords	75.20	74.30	76.07	70.20	68.15	72.51
B1 + Stemming	75.66	73.11	78.15	72.66	72.40	72.94
B1 + Lemmatisation	75.62	73.07	78.11	72.55	72.10	73.05
B1 + NH (SW1)	75.85	73.53	78.12	71.02	69.70	72.49
B1 + NH (SW2)	76.22	74.01	78.38	72.44	71.50	73.50
B1 + NH (PM)	76.13	75.57	76.69	71.23	72.10	70.27

Table 4.1. Pre-processing methods and their effects on performance of BNB sentiment polarity classifiers

	Training set			**Validation set**		
Model	ACC	S^+	S^-	ACC	S^+	S^-
Benchmark (**B1**)	76.13	74.48	77.73	71.55	71.70	71.38
B1 + Punctuation	75.78	75.05	76.50	71.23	69.74	72.89
B1 + Emojis	76.23	74.42	77.99	71.39	71.80	70.94
B1 + Emoticons	76.25	74.68	77.77	72.37	72.75	71.94
B1 - Stopwords	75.28	74.76	75.79	70.26	68.77	71.93
B1 + Stemming	75.97	74.39	77.50	73.08	73.30	72.83
B1 + Lemmatisation	75.95	74.46	77.39	71.55	71.70	71.38
B1 + NH (SW1)	76.29	75.19	77.37	71.18	71.10	71.27
B1 + NH (SW2)	76.58	75.53	77.62	72.13	73.00	71.16
B1 + NH (PM)	76.37	76.77	75.99	71.55	74.10	68.71

Table 4.2. Pre-processing methods and their effects on performance of MNB sentiment polarity classifiers

In the next step we considered combinations of the pre-processing techniques, which we identified above as beneficial with respect to performance. The results for this round are shown in Table 4.3. Three combinations were deemed optimal, specifically

- **B1** + Emoticons + NH (SW2),

- **B1** + Emojis + Emoticons + NH (SW2),
- **B1** + Emojis + Emoticons + NH (SW2) + Stemming.

We favoured the last combination due to the comparability of its scores to those of the direct competitors, the superiority regarding the performance on the validation set, as well as the large space for potential improvement (sparsity level could be increased more than for others due to stemming). This model, which we denote by B2, became the benchmark for the next stage.

	Training set			**Validation set**		
Model	ACC	S^+	S^-	ACC	S^+	S^-
Benchmark (**B1**)	76.13	74.48	77.73	71.55	71.70	71.38
B1 + Emojis + Emoticons	76.08	74.27	77.84	72.42	72.46	72.38
B1 + Emojis + NH (SW2)	76.56	75.29	77.79	71.92	72.30	71.49
B1 + Emojis + Stemming	75.96	74.19	77.69	73.08	72.80	73.39
B1 + Emoticons + NH (SW2)	76.69	75.68	77.68	72.79	72.55	73.05
B1 + Emoticons + Stemming	75.99	74.39	77.55	74.32	74.65	73.94
B1 + NH (SW2) + Stemming	76.34	75.26	77.39	72.71	72.90	72.49
B1 + Emojis + Emoticons + NH (SW2)	76.72	75.50	77.92	72.68	73.25	72.05
B1 + Emojis + Emoticons + Stemming	75.86	73.98	77.68	73.84	74.55	73.05
B1 + Emojis + NH (SW2) + Stemming	76.37	75.00	77.71	73.23	74.50	71.83
B1 + Emoticons + NH (SW2) + Stemming	76.49	75.37	77.59	73.42	74.25	72.49
B1 + Emojis + Emoticons + NH (SW2) + Stemming	76.61	75.29	77.90	73.79	74.65	72.83

Table 4.3. Combinations of pre-processing techniques and their effects on performance of MNB sentiment polarity classifiers

Thereafter, we experimented with including the supplementary features, discussed in Section 4.1.2. Here, we let LEX^+ and LEX^- denote the positive and negative lexicon features, respectively. The word, negating word and negated word counts are denoted by $\mathrm{N}_{\mathrm{words}}$, $\mathrm{N}_{\mathrm{negating}}$ and $\mathrm{N}_{\mathrm{negated}}$, respectively, while $\mathrm{N}_!$ and $\mathrm{N}_?$ are the exclamation and question mark counts, respectively. Table 4.4 demonstrates our findings, with the combination of LEX^+, LEX^-, $\mathrm{N}_{\mathrm{negating}}$ and $\mathrm{N}_!$ being associated with the highest scoring classifier.

We then used likelihood ratio tests, described in Section 4.1.2, to select the bigram features for our model. The pre-processed texts used to train the benchmark model (B2) were converted into a DTM of unigrams and all bigrams, whose test statistic values exceeded 7.879[2]. We required the selected bigrams to make a minimum number of appearances, which we denote by $\mathrm{N}_{\mathrm{min}}$. The results can be found in Table 4.5, indicating that in our case the inclusion of bigrams did not lead to any significant improvements in classification accuracy.

	Training set			**Validation set**		
Model	ACC	S^+	S^-	ACC	S^+	S^-
Benchmark (**B2**)	76.61	75.29	77.90	73.79	74.65	72.83
B2 + LEX^+ + LEX^-	76.64	75.37	77.88	73.89	74.85	72.83
B2 + $\mathrm{N}_{\mathrm{words}}$	76.60	75.21	77.96	73.79	74.45	73.05
B2 + $\mathrm{N}_{\mathrm{negated}}$	76.47	75.63	77.30	73.26	75.05	71.27
B2 + $\mathrm{N}_{\mathrm{negating}}$	76.88	75.41	78.32	74.37	75.85	72.72
B2 + $\mathrm{N}_{!}$	76.60	75.44	77.74	74.21	74.55	73.83
B2 + $\mathrm{N}_{?}$	76.56	75.33	77.77	73.68	74.05	73.27
B2 + LEX^+ + LEX^- + $\mathrm{N}_{\mathrm{negating}}$	76.91	75.48	78.30	74.53	76.05	72.83
B2 + LEX^+ + LEX^- + $\mathrm{N}_{!}$	76.60	75.50	77.68	74.42	74.85	73.94
B2 + LEX^+ + LEX^- + $\mathrm{N}_{!}$ + $\mathrm{N}_{\mathrm{negating}}$	76.92	75.64	78.16	75.11	76.25	73.83

Table 4.4. Addition of supplementary features and the effects on performance of MNB sentiment polarity classifiers

	Training set			**Validation set**		
Model	ACC	S^+	S^-	ACC	S^+	S^-
Benchmark (**B2**)	76.61	75.29	77.90	73.79	74.65	72.83
B2 + Bigrams ($\mathrm{N}_{\mathrm{min}} = 1$)	76.48	75.17	77.76	74.16	75.45	72.72
B2 + Bigrams ($\mathrm{N}_{\mathrm{min}} = 5$)	76.53	75.22	77.81	73.68	75.15	72.05
B2 + Bigrams ($\mathrm{N}_{\mathrm{min}} = 10$)	76.57	75.25	77.86	73.58	74.85	72.16
B2 + Bigrams ($\mathrm{N}_{\mathrm{min}} = 25$)	76.61	75.30	77.89	73.79	74.85	72.61
B2 + Bigrams ($\mathrm{N}_{\mathrm{min}} = 50$)	76.60	75.31	77.87	73.63	74.55	72.61
B2 + Bigrams ($\mathrm{N}_{\mathrm{min}} = 100$)	76.61	75.31	77.89	73.74	74.55	72.83

Table 4.5. Addition of bigram features and the effects on performance of MNB sentiment polarity classifiers

[2] The critical value for our tests, as we are dealing with a χ^2 distributed random variable with 1 degree of freedom and set the significance level equal to 0.5%.

In the last step of the training process we carried out hyperparameter tuning by means of grid-search CV. The free parameters for our problem included sparsity level and α. We constructed a small grid of paired values and ran 5-fold CV for each pair and our best model specification. The characteristics and performance of the classifier elected for the upcoming experiments are shown in Table 4.6.

Optional pre-processing methods	**Features**	**Sparsity**	α
✗ Punctuation ✓ Emojis ✓ Emoticons ✗ Pre-compiled stopword list ✓ Stemming ✗ Lemmatisation ✗ Negation handling (SW1) ✓ Negation handling (SW2) ✗ Negation handling (PM)	Unigrams LEX^+ LEX^- $\text{N}_{\text{negating}}$ $\text{N}_!$	0.99993	0.01

Training set			**Validation set**		
ACC	S^+	S^-	ACC	S^+	S^-
77.00	75.71	78.26	74.79	76.05	73.39

Table 4.6. Characteristics and performance of the MNB sentiment polarity classifier selected for regressions

4.2 Linking Stock Returns & Trading Activity to Twitter Sentiment & Volume

Once the training of the sentiment polarity classifier was complete, the collected firm-specific Tweets were pre-processed appropriately and labelled with the help of the resultant model. The next stages required us to aggregate the information carried by the 20,000,000 sentiment-tagged Tweets to produce firm-level measures of Twitter sentiment and activity and subsequently investigate if they could be of potential value to investors. The methodology applied closely resembles the one undertaken by Gu et al. (2020) and Duz Tan et al. (2021).

4.2.1 Twitter Sentiment & Volume Measures

Motivated by Bloomberg (Cui et al., 2016), we constructed the daily sentiment score of a company for trading day t based on the company-specific Tweets posted between 9:20 a.m. on trading day t and 9:20 a.m. on trading day $t+1$ (we refer to this period simply as "day t"). Since all of the stocks considered are traded on either NASDAQ, NYSE or BATS stock exchanges, the Eastern Time (ET) Zone was used[3].

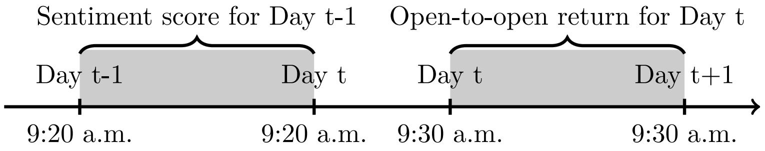

Figure 4.2. Construction method behind daily sentiment score and return values

For a given stock, we denote the number of Tweets classified as positive on day t by N_t^+. Moreover, for the Tweets labelled as positive and published on day t, we let C_t^+ represent the vector of confidence scores or probabilities of the tagged Tweets being of positive class, as approximated by the classifier. Assuming that all Tweets from day t labelled as positive are assigned numbers from 1 to N_t^+, inclusive, $C_{i,t}^+$ is the confidence score of the i-th Tweet given the positive tag. For the Tweets labelled as negative and posted on day t, N_t^-, C_t^- and $C_{i,t}^-$ are defined analogously.

We decided to experiment with multiple sentiment polarity measure forms, particularly

- **Average Sentiment Score** (used by Renault, 2020)

$$AVG_t = \frac{1 \cdot N_t^+ + (-1) \cdot N_t^-}{N_t^+ + N_t^-},$$

[3] Either Eastern Daylight Time (EDT) or Eastern Standard Time (EST), depending on the time of the year.

- **Average Weighted Sentiment Score** (motivated by Cui et al., 2016)

$$AVGw_t = \frac{\sum_{i=1}^{N_t^+} 1 \cdot C_{i,t}^+ + \sum_{j=1}^{N_t^-} (-1) \cdot C_{j,t}^-}{N_t^+ + N_t^-},$$

- **Agreement** (considered by Antweiler et al., 2004; Sprenger, Tumasjan, et al., 2014)

$$AG_t = 1 - \sqrt{1 - \left(\frac{N_t^+ - N_t^-}{N_t^+ + N_t^-} \right)^2}.$$

While the average sentiment scores aim to measure the polarity as well as the intensity of the sentiment towards a particular firm on day t, the goal of the agreement measure is vastly different. In essence, the more imbalanced the investor views are (for example, if everybody considers some stock to be a buy/sell), the higher the agreement measure values become. However, if investors are divided and express their opposing viewpoints through the use of appropriate tone on Twitter, the agreement measure would fall towards zero due to the high degree of disagreement.

For every company-day tuple, we required a minimum of ten Tweets labelled for sentiment in order to compute the sentiment scores. If this requirement was not fulfilled, the relevant observation was omitted from the regressions. Consequently, the contemporaneous regressions involving sentiment were based on 30162 observation vectors, while predictive regressions were run with 30115 observations.

To measure company-level Twitter activity, we utilised several proxies for the number of relevant Tweets, namely

- N^{OG} - the original Tweet count,
- N^{OG+RE} - the count of original Tweets and Replies,
- $N^{OG+RT+RE}$ - the count of original Tweets, Retweets and Replies.

For the most part, researchers exploring the nature of relationships between Twitter volume and returns or trading activity do not make a distinction between the various possibilities of computing the Tweet count and decide in favour of the last measure (see e.g. Duz Tan et al., 2021; Ranco et al., 2015).

4.2.2 Response & Control Variables

Seeing that the analysis of Duz Tan et al. (2021) was partly based on the S&P500 constituents, our choice of response and control variables was to a large extent driven by their selection of variables (influenced by Tetlock, 2011; Sprenger, Tumasjan, et al., 2014), owing to the similarity of observation counts for cross-sectional regressions. We let P_t, H_t and L_t represent the opening, high and low prices for day t, respectively, and provide details on the dependent and independent variables.

Holding period return

In our investigation of the relations between stock returns and Twitter sentiment/activity, holding period return (HPR) played the role of the response variable. We computed stock return values using opening prices and adjusted the calculations for dividends. The formula used to compute HPR for day t, Ret_t, is given by

$$Ret_t = \frac{(P_{t+1} - P_t) + Income_t}{P_t},$$

where $Income_t$ refers to the dividend income, which an investor holding the stock through day t would be entitled to receive[4].

Abnormal turnover

Abnormal turnover was used as a measure of trading volume in an effort to make volume comparable across firms[5]. We defined the abnormal trading volume on day t ($AbTurn_t$) as

$$AbTurn_t = LogTurn_t - \frac{\sum_{i=1}^{5} LogTurn_{t-i}}{5},$$

where $LogTurn_t$ denotes the log-transformed trading volume for day t. Positive (negative) values would be indicative of trading in the given stock on day t being more (less) active, in comparison to the 5-day average (Gu et al., 2020).

Cumulative return

Cumulative return values, corresponding to day t, were computed analogously to HPR values, with the holding period comprising the five days prior to the market open of day t. More precisely, cumulative return corresponding to an observation for day t and some given stock ($Ret_{[t-5,t-1]}$) was calculated as

[4] Since we are dealing with daily returns, in order to receive the dividend one would need to buy the relevant stock one day before the ex-dividend date.

[5] For instance, Ford Motor Company's stock was traded most heavily (over the sampling period) with 80,725,740 shares traded daily on average. The stock of NVR, Inc. was the least popular among traders, with an average daily turnover of 18,656 shares.

$$Ret_{[t-5,t-1]} = \frac{(P_t - P_{t-5}) + Income_{[t-5,t-1]}}{P_{t-5}},$$

where $Income_{[t-5,t-1]}$ denotes the dividend income earned by holding the stock throughout the period.

Market capitalisation

As demonstrated by Fama and French (1992), company size is an important return determinant, able to explain much of the cross-sectional variation of stock returns. We used log of market capitalisation from the day preceding the observation day t (denoted by $Size_{t-1}$) as a control variable, which is commonly employed to account for firm size, applying the log-transform as a remedy for possible non-normality of the variable.

Parkinson's volatility

Due to the large body of research providing evidence in support of existence of bilateral links between expected stock market returns and volatility (see e.g. French et al., 1987), a control for volatility was also included. For an observation on day t, Parkinson (1980) estimate of volatility based on extreme values (high and low prices) from the 5-day period preceding day t, denoted by $Vola_{[t-5,t-1]}$, was used as a volatility proxy, with the definition given by

$$Vola_{[t-5,t-1]} = \sqrt{\frac{1}{4\log 2} \frac{\sum_{i=1}^{5} \log(H_{t-i}/L_{t-i})}{5}}.$$

Amihud's illiquidity

Amihud (2002) has previously shown that expected stock returns increase with expected illiquidity, both across stocks and over time, thus highlighting the importance of controlling for its effect. In his seminal paper, the author proposed a novel illiquidity measure, which does not require any microstructure data on quotes and transactions. For observations on day t, we computed Amihud's illiquidity as the 5-day average of the daily absolute return to the daily dollar volume, denoting it by $Illiq_{[t-5,t-1]}$, i.e. we let

$$Illiq_{[t-5,t-1]} = \sum_{i=1}^{5} \frac{|Ret_{t-i}|}{P_{t-i} \cdot Turn_{t-i}},$$

thereby using readily available data to estimate the daily price impact of the order flow.

4.2.3 Cross-Sectional Regressions

To investigate the impact of firm-level Twitter sentiment and activity on daily stock returns and abnormal turnover, a framework similar to those of Tetlock (2011) and Sprenger, Tumasjan, et al. (2014) was considered. Tetlock's methodology has recently been popularised by the publications of Gu et al. (2020) and Duz Tan et al. (2021), which investigated the links between firm-specific Twitter features and stock markets.

As described by Gu et al. (2020), one would initially run daily cross-sectional regressions (as per linear model specification), similar to those of Fama and MacBeth (1973), which would produce a time series of coefficient estimates for each variable. Thereafter, one would estimate the coefficients of interest (such as the change in individual firm's stock return associated with a unit change in the stock's Twitter sentiment, c.p.) by evaluating the time series means. Lastly, when computing the corresponding t-statistics, robust standard errors could be used to account for possible residual serial correlation and heteroskedasticity. We followed the decision of the aforementioned authors and used Newey et al. (1986) standard errors, adjusted up to four[6] lags for heteroskedasticity and autocorrelation.

Stock Returns, Trading Activity & Twitter Sentiment

The following regression specifications were considered to analyse the impact of Twitter sentiment on stock returns and trading volumes at firm level:

$$\begin{aligned} Ret_{i,t} &= a + b \cdot Sent_{i,t/t-1} + c \cdot Size_{i,t-1} + d \cdot Ret_{i,[t-5,t-1]} \\ &\quad + e \cdot AbTurn_{i,t-1} + f \cdot Vola_{i,[t-5,t-1]} + g \cdot Illiq_{i,[t-5,t-1]} + \epsilon_{i,t}, \end{aligned}$$

$$\begin{aligned} AbTurn_{i,t} &= a + b \cdot Sent_{i,t/t-1} + c \cdot Size_{i,t-1} + d \cdot Ret_{i,[t-5,t-1]} \\ &\quad + e \cdot Vola_{i,[t-5,t-1]} + f \cdot Illiq_{i,[t-5,t-1]} + \epsilon_{i,t}, \end{aligned}$$

where i represents a particular stock and $t/t-1$ indicates that both contemporaneous and predictive regressions were run, with $Sent_{i,t}$ ($Sent_{i,t-1}$) used as a regressor in contemporaneous (predictive) regressions. We note that $Sent$ is used as a shorthand to denote the sentiment measure used in regressions. The results are shown in Table 4.7. The superscripts *, ** and *** indicate significance of estimated coefficients at 10%, 5% and 1% levels, respectively.

The results of contemporaneous regressions of returns on sentiment are for the most part consistent with those of Duz Tan et al. (2021) and Sprenger, Tumasjan, et al.

[6] Newey and West recommend to set the number of lags equal to $q = 4(n/100)^{2/9}$, where n is the number of observations in the time series.

	Contemporaneous			**Predictive**		
	AVG_t	$AVGw_t$	AG_t	AVG_{t-1}	$AVGw_{t-1}$	AG_{t-1}
Panel A: Returns and Twitter sentiment						
$Sent_{t/t-1}$	0.0055^{***}	0.0063^{***}	0.0048^{***}	-0.0004	-0.0004	-0.0004
	(16.00)	(15.22)	(5.61)	(-1.41)	(-1.30)	(-0.63)
$Size_{t-1}$	0.0002	0.0002	0.0003^{*}	0.0004^{**}	0.0004^{**}	0.0004^{**}
	(1.30)	(1.30)	(1.94)	(2.39)	(2.42)	(2.31)
$Ret_{[t-5,t-1]}$	0.0135	0.0140	0.0166^{*}	0.0086	0.0084	0.0083
	(1.48)	(1.53)	(1.77)	(0.99)	(0.98)	(0.95)
$AbTurn_{t-1}$	0.0009^{*}	0.0009^{*}	0.0009	0.0008	0.0008	0.0008
	(1.69)	(1.71)	(1.61)	(1.49)	(1.48)	(1.52)
$Vola_{[t-5,t-1]}$	-0.0213	-0.0213	-0.0214	-0.0013	-0.0015	0.0010
	(-0.28)	(-0.28)	(-0.28)	(-0.02)	(-0.02)	(0.01)
$Illiq_{[t-5,t-1]}$	0.0124^{***}	0.0125^{***}	0.0099^{***}	0.0046	0.0047	0.0044
	(3.27)	(3.30)	(2.75)	(1.21)	(1.23)	(1.17)
$Const.$	-0.0071	-0.0071	-0.0087^{*}	-0.0110^{**}	-0.0112^{**}	-0.0109^{**}
	(-1.57)	(-1.55)	(-1.92)	(-2.33)	(-2.36)	(-2.25)
R^2	0.0102	0.0096	0.0038	0.0018	0.0018	0.0018
N	30162	30162	30162	30115	30115	30115
Time periods	128	128	128	128	128	128
Panel B: Abnormal turnover and Twitter sentiment						
$Sent_{t/t-1}$	-0.0714^{***}	-0.1039^{***}	-0.2181^{***}	-0.0028	-0.0119	-0.0754^{***}
	(-9.10)	(-10.95)	(-14.63)	(-0.49)	(-1.56)	(-5.44)
$Size_{t-1}$	-0.0134^{**}	-0.0133^{**}	-0.0166^{***}	-0.0058	-0.0058	-0.0068
	(-2.45)	(-2.44)	(-3.00)	(-1.08)	(-1.09)	(-1.28)
$Ret_{[t-5,t-1]}$	-0.1766	-0.1686	-0.1799	-0.2419^{*}	-0.2351^{*}	-0.2311^{*}
	(-1.33)	(-1.28)	(-1.35)	(-1.82)	(-1.78)	(-1.71)
$Vola_{[t-5,t-1]}$	-10.1481^{***}	-10.2003^{***}	-10.5689^{***}	-9.017^{***}	-9.0401^{***}	-9.2432^{***}
	(-11.67)	(-11.81)	(-12.20)	(-11.91)	(-11.90)	(-12.33)
$Illiq_{[t-5,t-1]}$	0.9716^{***}	0.9664^{***}	1.0427^{***}	0.6687^{***}	0.6664^{***}	0.6814^{***}
	(8.46)	(8.45)	(8.82)	(6.89)	(6.89)	(6.98)
$Const.$	0.4952^{***}	0.4992^{***}	0.5800^{***}	0.2593^{*}	0.2630^{*}	0.2954^{**}
	(3.55)	(3.60)	(4.11)	(1.89)	(1.91)	(2.13)
R^2	0.0395	0.0414	0.0434	0.0241	0.0241	0.0249
N	30162	30162	30162	30115	30115	30115
Time periods	128	128	128	128	128	128

Table 4.7. Results of contemporaneous and predictive regressions of daily returns and abnormal turnover on Twitter sentiment, constant and controls

(2014). All of our three estimated coefficients are positive and significant at 1% level. In the study led by Duz Tan et al. (2021), the estimated change in return resulting from a unit increase in the Bloomberg Twitter sentiment measure (c.p.) was equal to 0.76% for the S&P500 portfolio, compares to the estimated changes of 0.55% and 0.63% in the case of our sentiment polarity measures, AVG and $AVGw$, respectively. The results of the fixed-effect panel regressions of Sprenger, Tumasjan, et al. (2014) also point to a positive contemporaneous linear association of returns and the bullishness index, quantifying the intensity of positive sentiment towards stocks. Sprenger, Tumasjan, et al. (2014) also identified a significant contemporaneous relationship between returns and the agreement measure, AG, however the direction of association contradicts our results.

As for the trading volume, similar to the aforementioned authors we found the relationships between trading activity and our sentiment measures to be significant at 1% level. However, our findings suggest that the direction of association is negative for all three measures, inconsistent with the results of Duz Tan et al. (2021) and consistent with those of Sprenger, Tumasjan, et al. (2014), with respect to both the bullishness index and the agreement. Moreover, the estimated coefficient magnitudes differ substantially from the ones estimated by Duz Tan et al. (2021), with our AVG and $AVGw$ coefficients being on average four times larger than the S&P500 Bloomberg Twitter sentiment coefficient of Duz Tan et al. (2021).

Albeit the contemporaneous links between Twitter sentiment and market features are impressive in their own right, one may wonder whether Tweet-based features are able to anticipate changes in stock prices and trading volumes and if those are permanent in nature, as noted by Sprenger, Tumasjan, et al. (2014). Our findings suggest that Twitter sentiment holds no significant predictive power with respect to returns, contradicting the results of Gu et al. (2020) and Duz Tan et al. (2021). Both estimated the change in return associated with a unit increase in Bloomberg Twitter sentiment (c.p.) to be approximately equal to 10 basis points. Nevertheless, our results are consistent with the works of Sprenger, Tumasjan, et al. (2014) and Renault (2020), who found no evidence of investor sentiment helping to predict stock returns.

Lastly, we found Twitter sentiment polarity measures to be unimportant as predictors of trading volume (inconsistent with the results of Duz Tan et al., 2021). However, we did find strong evidence in support of the agreement measure negatively predicting next day's trading volume, which has also been documented by Sprenger, Tumasjan, et al. (2014).

Stock Returns, Trading Activity & Twitter Volume

The regression specifications used to investigate the impact of Twitter activity on stock returns and trading volumes at firm level are given by

$$\begin{aligned} Ret_{i,t} &= a + b \cdot NoTweets_{i,t/t-1} + c \cdot Size_{i,t-1} + d \cdot Ret_{i,[t-5,t-1]} \\ &\quad + e \cdot AbTurn_{i,t-1} + f \cdot Vola_{i,[t-5,t-1]} + g \cdot Illiq_{i,[t-5,t-1]} + \epsilon_{i,t}, \end{aligned}$$

$$\begin{aligned} AbTurn_{i,t} &= a + b \cdot NoTweets_{i,t/t-1} + c \cdot Size_{i,t-1} + d \cdot Ret_{i,[t-5,t-1]} \\ &\quad + e \cdot Vola_{i,[t-5,t-1]} + f \cdot Illiq_{i,[t-5,t-1]} + \epsilon_{i,t}. \end{aligned}$$

For details on the notation we refer the reader to the previous section. We use $NoTweets$ as a shorthand to denote the Tweet volume measure used in regressions. The corresponding results are shown in Table 4.8.

The contemporaneous regression results, concerning returns and Tweet volume, provide strong evidence in favour of the existence of a positive linear relationship between interday returns and Twitter activity, supporting the findings of Duz Tan et al. (2021), Li et al. (2018) and Wysocki (1998). We have estimated the increase in return associated with a 100 unit increase in the original Tweet count (c.p.) to be equal to 3.9 basis points, with smaller changes estimated for the alternative Tweet volume measures.

All estimated coefficients of $NoTweets_t$, resulting from the contemporaneous regressions of trading volume on Twitter activity, were positive and highly significant, providing further evidence in support of the contemporaneous relationship between trading volume and Twitter activity and, thus, agreeing with the findings of Duz Tan et al. (2021), Sprenger, Tumasjan, et al. (2014), Antweiler et al. (2004) and Li et al. (2018).

Similarly to the predictive regressions of returns on Twitter sentiment, no significant predictive power was observed in regard to the predictive regressions of returns on Twitter activity, contrary to the findings of Duz Tan et al. (2021) and consistent with those of Sprenger, Tumasjan, et al. (2014) and Li et al. (2018). However, in prediction of abnormal trading volume all of the estimated coefficients for the Tweet count variables were positive and significant at 1% level, indicating the existence of a lagged relationship between the two variables. Therefore, the result is consistent with the findings of Duz Tan et al. (2021), Antweiler et al. (2004), Sprenger, Tumasjan, et al. (2014) and Li et al. (2018).

	Contemporaneous			**Predictive**		
	N_t^{OG}	N_t^{OG+RE}	$N_t^{OG+RE+RT}$	N_{t-1}^{OG}	N_{t-1}^{OG+RE}	$N_{t-1}^{OG+RE+RT}$
Panel A: Returns and Tweet volume						
$NoTweets_{t/t-1}$	0.0000039***	0.0000029***	0.0000011***	0.0000009	0.0000007	0.0000003
	(4.12)	(3.84)	(3.89)	(1.26)	(1.22)	(1.02)
$Size_{t-1}$	-0.0000385	-0.0000048	0.0000425	0.0001563	0.0001630	0.0001834
	(-0.29)	(-0.04)	(0.34)	(1.24)	(1.30)	(1.53)
$Ret_{[t-5,t-1]}$	0.0048708	0.0048812	0.0048471	0.0052519	0.0052380	0.0050868
	(0.61)	(0.61)	(0.60)	(0.65)	(0.65)	(0.63)
$AbTurn_{t-1}$	0.0005833*	0.0005901*	0.0006125*	0.0006311*	0.0006357*	0.0006534**
	(1.79)	(1.81)	(1.86)	(1.91)	(1.93)	(1.98)
$Vola_{[t-5,t-1]}$	-0.0401223	-0.0374720	-0.0328878	-0.0210664	-0.0206711	-0.0192487
	(-0.55)	(-0.51)	(-0.45)	(-0.29)	(-0.28)	(-0.26)
$Illiq_{[t-5,t-1]}$	-0.0018267	-0.0016980	-0.0016064	-0.0011968	-0.0011575	-0.0010465
	(-1.61)	(-1.50)	(-1.43)	(-1.05)	(-1.01)	(-0.92)
$Const.$	0.0012413	0.0003996	-0.0007824	-0.0036596	-0.0038303	-0.0043452
	(0.36)	(0.12)	(-0.24)	(-1.14)	(-1.20)	(-1.41)
R^2	0.0023629	0.0022243	0.0021848	0.0014097	0.0014089	0.0014059
N	63744	63744	63744	63744	63744	63744
Time periods	128	128	128	128	128	128
Panel B: Abnormal turnover and Tweet volume						
$NoTweets_{t/t-1}$	0.0002982***	0.0002172***	0.0000721***	0.0001715***	0.0001260***	0.0000442***
	(12.95)	(11.76)	(10.65)	(6.92)	(6.22)	(5.67)
$Size_{t-1}$	−0.0157459***	−0.0132083***	−0.0066609*	−0.0078895*	-0.0064554	-0.0028894
	(-3.74)	(-3.16)	(-1.75)	(-1.90)	(-1.57)	(-0.76)
$Ret_{[t-5,t-1]}$	−0.2693800**	−0.2674190**	−0.2599557**	−0.2584549**	−0.2554762*	−0.2563884*
	(-2.07)	(-2.05)	(-1.96)	(-1.98)	(-1.95)	(-1.93)
$Vola_{[t-5,t-1]}$	−9.8757220***	−9.6633100***	−9.0948230***	−9.3105320***	−9.1743380***	−8.8223130***
	(-13.00)	(-12.77)	(-11.95)	(-12.33)	(-12.19)	(-11.58)
$Illiq_{[t-5,t-1]}$	0.1996868***	0.2082174***	0.2278957***	0.2276997***	0.2322465***	0.2428367***
	(4.86)	(5.08)	(5.39)	(5.49)	(5.60)	(5.78)
$Const.$	0.4909727***	0.4274575***	0.2633204***	0.2949714***	0.2588809**	0.1689402*
	(4.62)	(4.05)	(2.71)	(2.86)	(2.53)	(1.75)
R^2	0.0263152	0.0246411	0.0208665	0.0192884	0.0182561	0.0172576
N	63744	63744	63744	63744	63744	63744
Time periods	128	128	128	128	128	128

Table 4.8. Results of contemporaneous and predictive regressions of daily returns and abnormal turnover on Tweet volume, constant and controls

4.3 Twitter Sentiment as Signal in Trading Strategies

In the previous section we observed that Twitter sentiment did not seem to be a significant predictor of next-day's returns. We decided to investigate whether a profitable portfolio could be constructed based on Twitter sentiment alone. This idea was motivated by the works of Gu et al. (2020) and Duz Tan et al. (2021), who viewed the performance of the resulting portfolios as a showcase of the economic importance of their findings.

Similar to the aforementioned authors, we applied a simple long-short trading strategy. The trading began at 9:30 a.m. on 1st June 2021 and ended at 9:30 a.m. on 30th November 2021. At 9:20 a.m., prior to the market open, we made the decision on how the long and short portfolios would be rebalanced, based on the updated Twitter sentiment values. All stocks with the Twitter sentiment values lying in the bottom decile were assigned to the short portfolio, while those with the sentiment values in the top decile made it into the long portfolio. We rebalanced the portfolios at the market open according to this rule. The costs associated with the execution of trades were ignored.

The results for equal-weighted portfolios are presented in Table 4.9 and those for value-weighted portfolios are shown in Table 4.10. In addition to providing the estimated average raw returns, we risk-adjusted the returns, i.e. we computed the abnormal returns as residuals of the CAPM, Fama-French three-factor and Fama-French-Carhart four-factor models. The relevant factor data was downloaded from Kenneth R. French's website[7].

Ultimately, none of our long-short portfolios were seen to generate significant returns, which does not come as a surprise in light of the observations made in Section 4.2.3. On the contrary, Gu et al. (2020) estimated all of the daily alphas to be slightly below 10 basis points, with the abnormal returns of Duz Tan et al. (2021) for the S&P500 portfolio estimated to fall between 5 and 6 basis points, ignoring the transaction costs.

4.4 Stock Market Indices & Constituent-Based Twitter Sentiment

Another aspect of firm-level Twitter sentiment, which to our knowledge has not been explored in the relevant literature, is the potential ability to forecast the performance of exchange traded funds (ETFs) (or the returns of the corresponding stock indices) by means of aggregation of Twitter sentiment values of the constituents. Since our study concerned itself exclusively with S&P500 stocks, we picked out three of the

[7] `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

	Raw	CAPM	Fama-French (3)	Carhart (4)
Panel A: Portfolios based on AVG Twitter sentiment values				
Raw return	0.0076 (0.36)			
Alpha		0.0077 (0.43)	0.0077 (0.42)	0.0073 (0.48)
Market		-0.0075^{**} (-2.11)	-0.0074^{**} (-2.46)	-0.0099^{**} (-2.57)
SMB			-0.0003 (-0.05)	-0.0031 (-0.43)
HML			-0.0003 (-0.12)	-0.0019 (-0.56)
UMD				0.0062 (0.75)
R^2		0.0258	0.0259	0.0357
N$_{obs}$	128	128	128	128
Panel B: Portfolios based on $AVGw$ Twitter sentiment values				
Raw return	0.0102 (0.60)			
Alpha		0.0104 (0.68)	0.0103 (0.69)	0.0099 (0.75)
Market		-0.0073^{**} (-2.21)	-0.0070^{**} (-2.45)	-0.0094^{***} (-2.62)
SMB			-0.0013 (-0.29)	-0.0041 (-0.62)
HML			-0.0004 (-0.14)	-0.0019 (-0.61)
UMD				0.0062 (0.83)
R^2		0.0284	0.0292	0.0404
N$_{obs}$	128	128	128	128
Panel C: Portfolios based on AG Twitter sentiment values				
Raw return	0.0031 (0.29)			
Alpha		0.0033 (0.31)	0.0032 (0.31)	0.0032 (0.31)
Market		-0.0044 (-1.52)	-0.0042^{*} (-1.85)	-0.0044^{*} (-1.86)
SMB			-0.0003 (-0.07)	-0.0005 (-0.10)
HML			-0.0011 (-0.55)	-0.0012 (-0.50)
UMD				0.0005 (0.11)
R^2		0.0193	0.0215	0.0216
N$_{obs}$	128	128	128	128

Table 4.9. Raw and risk-adjusted returns of equal-weighted long-short portfolios based on Twitter sentiment

	Raw	CAPM	Fama-French (3)	Carhart (4)
Panel A: Portfolios based on *AVG* Twitter sentiment values				
Raw return	0.0086 (0.70)			
Alpha		0.0086 (0.87)	0.0085 (0.86)	0.0080 (0.97)
Market		-0.0076^{**} (-2.27)	-0.0072^{**} (-2.46)	-0.0106^{**} (-2.49)
SMB			-0.0017 (-0.35)	-0.0056 (-0.80)
HML			-0.0006 (-0.23)	-0.0028 (-0.94)
UMD				0.0087 (1.16)
R^2		0.0362	0.0381	0.0639
N_{obs}	128	128	128	128
Panel B: Portfolios based on *AVGw* Twitter sentiment values				
Raw return	0.0179 (1.53)			
Alpha		0.0180^{*} (1.80)	0.0178^{*} (1.81)	0.0173^{**} (2.12)
Market		-0.0073^{**} (-2.33)	-0.0066^{**} (-2.35)	-0.0103^{**} (-2.73)
SMB			-0.0028 (-0.59)	-0.0070 (-1.10)
HML			0.0002 (0.09)	-0.0022 (-0.75)
UMD				0.0095 (1.40)
R^2		0.0326	0.0360	0.0664
N_{obs}	128	128	128	128
Panel C: Portfolios based on *AG* Twitter sentiment values				
Raw return	0.0060 (0.72)			
Alpha		0.0063 (0.79)	0.0066 (0.84)	0.0067 (0.89)
Market		-0.0036 (-1.23)	-0.0044^{*} (-1.88)	-0.0032 (-1.54)
SMB			0.0032 (0.85)	0.0046 (1.05)
HML			0.0002 (0.11)	0.0010 (0.43)
UMD				-0.0030 (-0.76)
R^2		0.0130	0.0211	0.0261
N_{obs}	128	128	128	128

Table 4.10. Raw and risk-adjusted returns of value-weighted long-short portfolios based on Twitter sentiment

best-performing (as per Forbes selection for July 2022) S&P500-replicating funds, in particular iShares Core S&P 500 ETF (IVV), SPDR S&P 500 ETF Trust (SPY) and Vanguard 500 Index Fund (VOO).

In order to aggregate the firm-specific Twitter sentiment values on a particular trading day into a measure of sentiment towards the index, we computed the equal- and value-weighted means of all available firm-specific $AVG/AVGw/AG$ values on the given day. Instead of cross-sectional regressions, we ran controlled multiple linear regressions with the following specifications:

$$\begin{aligned} Ret_{i,t} &= a + b \cdot Sent_{t-1} + c \cdot Size_{i,t-1} + d \cdot Ret_{i,[t-5,t-1]} \\ &\quad + e \cdot AbTurn_{i,t-1} + f \cdot Vola_{i,[t-5,t-1]} + g \cdot Illiq_{i,[t-5,t-1]} + \epsilon_{i,t}, \end{aligned}$$

$$\begin{aligned} Ret_{i,t} &= a + \sum_{j=1}^{5} b_j \cdot Sent_{t-j} + c \cdot Size_{i,t-1} + d \cdot Ret_{i,[t-5,t-1]} \\ &\quad + e \cdot AbTurn_{i,t-1} + f \cdot Vola_{i,[t-5,t-1]} + g \cdot Illiq_{i,[t-5,t-1]} + \epsilon_{i,t}, \end{aligned}$$

where i represents one of the three indices and $Sent_{t-1}, \ldots, Sent_{t-5}$ denote aggregated lagged S&P500 sentiment values for the week preceding trading day t.

Gu et al. (2020) mention that the predictive power of Twitter sentiment for individual stock returns (if significant) could be related to its information content. If Twitter sentiment contains useful information about the fundamentals of a company, which has not been incorporated into the prices, its price effect should be permanent, with no reversals observed over the course of the following week. Otherwise, if the sentiment merely reflects the opinions of uninformed investors, the associated price effects would only be temporary and subsequently reversed due to the temporary shifts in demand.

Since the fundamental values of the ETFs can be seen as functions of the fundamental values of the constituents, one could potentially extend this notion to stock portfolios. For this reason, we included several lags of Twitter sentiment in our model specifications. The results for the regressions using equal-weighted (value-weighted) Twitter sentiment measures are demonstrated in Table 4.11 (4.12), with the estimated coefficients for the control variables omitted.

Our findings suggest that aggregation of firm-level Twitter sentiment could be of value in forecasting ETF performance. The coefficients for previous day's sentiment are all positive and significant (mostly at 5% and 1%) and no price reversals are observed

	Predictive (1 day)			**Predictive (5 days)**		
	AVG	*AVGw*	*AG*	*AVG*	*AVGw*	*AG*
Panel A: IVV						
Const.	0.343 (1.56)	0.458** (2.04)	0.248 (1.04)	0.382 (1.56)	0.515* (1.81)	0.282 (1.15)
$Sent_{t-1}$	0.032*** (2.92)	0.035*** (2.62)	0.066** (2.38)	0.030** (2.54)	0.033** (2.35)	0.068** (2.48)
$Sent_{t-2}$				0.005 (0.30)	0.008 (0.42)	-0.045 (-1.05)
$Sent_{t-3}$				-0.001 (-0.09)	-0.007 (-0.42)	0.019 (0.59)
$Sent_{t-4}$				0.015 (1.20)	0.021 (1.42)	0.057* (1.95)
$Sent_{t-5}$				-0.010 (-0.71)	-0.013 (-0.84)	0.011 (0.31)
R^2	0.069	0.083	0.043	0.082	0.083	0.074
N_{obs}	128	128	128	128	128	128
Panel B: SPY						
Const.	0.323 (1.31)	0.434* (1.73)	0.224 (0.84)	0.374 (1.45)	0.513* (1.75)	0.254 (0.96)
$Sent_{t-1}$	0.028** (2.56)	0.031** (2.34)	0.053* (1.90)	0.027** (2.31)	0.029** (2.16)	0.056** (2.12)
$Sent_{t-2}$				0.005 (0.33)	0.009 (0.45)	-0.048 (-1.11)
$Sent_{t-3}$				-0.004 (-0.24)	-0.010 (-0.56)	0.017 (0.55)
$Sent_{t-4}$				0.018 (1.42)	0.024 (1.64)	0.059** (2.17)
$Sent_{t-5}$				-0.010 (-0.71)	-0.012 (-0.80)	0.011 (0.28)
R^2	0.058	0.054	0.035	0.075	0.078	0.068
N_{obs}	128	128	128	128	128	128
Panel C: VOO						
Const.	0.569* (1.82)	0.701** (2.15)	0.462 (1.44)	0.601* (1.82)	0.758** (2.12)	0.498 (1.48)
$Sent_{t-1}$	0.031*** (2.87)	0.035*** (2.74)	0.056** (2.11)	0.029** (2.57)	0.032** (2.40)	0.058** (2.34)
$Sent_{t-2}$				0.006 (0.35)	0.009 (0.48)	-0.048 (-1.09)
$Sent_{t-3}$				-0.002 (-0.11)	-0.008 (-0.44)	0.020 (0.64)
$Sent_{t-4}$				0.015 (1.30)	0.021 (1.54)	0.059** (2.19)
$Sent_{t-5}$				-0.010 (-0.76)	-0.013 (-0.88)	0.011 (0.30)
R^2	0.063	0.063	0.035	0.080	0.084	0.069
N_{obs}	128	128	128	128	128	128

Table 4.11. Results of predictive regressions of ETF returns on equal-weighted Twitter sentiment, constant and controls

	Predictive (1 day)			**Predictive (5 days)**		
	AVG	*AVGw*	*AG*	*AVG*	*AVGw*	*AG*
Panel A: iShares Core S&P 500 ETF (IVV)						
Const.	0.192 (0.91)	0.357* (1.73)	0.148 (0.68)	0.188 (0.82)	0.321 (1.18)	0.153 (0.68)
$Sent_{t-1}$	0.040*** (2.62)	0.044** (2.38)	0.101** (2.62)	0.039** (2.51)	0.043** (2.29)	0.098** (2.46)
$Sent_{t-2}$				-0.001 (-0.07)	0.001 (0.06)	-0.026 (-0.74)
$Sent_{t-3}$				-0.014 (-0.93)	-0.021 (-1.13)	-0.036 (-0.84)
$Sent_{t-4}$				0.022** (2.04)	0.024* (1.90)	0.067** (2.31)
$Sent_{t-5}$				-0.010 (-0.74)	-0.012 (-0.74)	-0.010 (-0.25)
R^2	0.080	0.075	0.071	0.107	0.104	0.101
N$_{obs}$	128	128	128	128	128	128
Panel B: SPDR S&P 500 ETF Trust (SPY)						
Const.	0.190 (0.79)	0.342 (1.44)	0.141 (0.57)	0.189 (0.74)	0.323 (1.06)	0.142 (0.57)
$Sent_{t-1}$	0.033** (2.32)	0.037** (2.16)	0.084** (2.23)	0.033** (2.23)	0.037** (2.07)	0.082** (2.14)
$Sent_{t-2}$				-0.001 (-0.10)	0.001 (0.05)	-0.030 (-0.85)
$Sent_{t-3}$				-0.015 (-0.97)	-0.021 (-1.16)	-0.036 (-0.85)
$Sent_{t-4}$				0.025** (2.53)	0.028** (2.38)	0.076*** (2.81)
$Sent_{t-5}$				-0.009 (-0.64)	-0.010 (-0.61)	-0.007 (-0.15)
R^2	0.063	0.060	0.055	0.095	0.093	0.093
N$_{obs}$	128	128	128	128	128	128
Panel C: Vanguard S&P 500 ETF (VOO)						
Const.	0.420 (1.48)	0.624** (2.15)	0.358 (1.22)	0.442 (1.45)	0.674** (2.08)	0.394 (1.25)
$Sent_{t-1}$	0.038*** (2.73)	0.045*** (2.65)	0.093*** (2.61)	0.038** (2.57)	0.044** (2.49)	0.090** (2.49)
$Sent_{t-2}$				0.000 (0.00)	0.004 (0.26)	-0.028 (-0.78)
$Sent_{t-3}$				-0.013 (-0.81)	-0.017 (-0.88)	-0.034 (-0.77)
$Sent_{t-4}$				0.025*** (2.65)	0.029*** (2.61)	0.075*** (2.98)
$Sent_{t-5}$				-0.011 (-0.78)	-0.011 (-0.71)	-0.014 (-0.32)
R^2	0.071	0.073	0.059	0.102	0.106	0.095
N$_{obs}$	128	128	128	128	128	128

Table 4.12. Results of predictive regressions of ETF returns on value-weighted Twitter sentiment, constant and controls

in the following days. Interestingly, the vast majority of the estimated coefficients of value-weighted $Sent_{t-4}$ variables are significant at 5% level, but as they do not reverse the price effects we do not concentrate on the potential causes driving these observations.

The magnitudes of the estimated coefficients add to the economic importance of the observations. For instance, in the predictive regression of IVV returns on five lags of equal-weighted (value-weighted) $AVGw$ values, one unit increase in $AVGw_{t-1}$ is associated with a 3.3% (4.3%) increase in return, c.p.. Meanwhile, increases in the aggregated AG measure values are associated with the largest increases in returns.

5 Discussion

5.1 Summary, Interpretations & Implications of Results

In summary, we find Twitter features to be helpful in explaining stock returns and turnover at contemporaneous level[1] and of limited use with respect to return predictability[2]. Our results provide strong evidence in support of hypotheses *H1, H2, H3.b* and *H4* holding contemporaneously. We reject hypothesis *H3.a* at 1% significance level, as we found the abnormal trading volume to decrease with rising optimism among investors (consistent with Sprenger, Tumasjan, et al., 2014). One possible explanation could involve Twitter bullishness signalling the intensifying desire of the investor community as a whole to go "long" on a particular stock, resulting in illiquidity due to the supply-demand imbalance.

Considering lagged relationships, our results support the validity of hypotheses *H3.b* and *H4*, thereby providing evidence in support of Tweet volume positively predicting next day's abnormal trading volume and the disagreement among stock microbloggers negatively predicting the trading activity on the following day (see Table 4.7). Our observations therefore add to the mounting evidence supporting these hypotheses (Duz Tan et al., 2021; Sprenger, Tumasjan, et al., 2014; Antweiler et al., 2004; Li et al., 2018). On the other hand, we found no evidence in support of hypotheses *H1, H2* and *H3.a.*

Particularly, we note that our findings failed to attribute significant return predictability to Twitter sentiment (in line with most results in the area), consistent with the efficient market hypothesis and contradicting the results of Gu et al. (2020) and Duz Tan et al. (2021). Therefore, there was no need to include additional lagged values of sentiment in our regressions, in order to spot potential reversals. In effect, we found firm-level Twitter sentiment to be of no informational value with respect to daily returns of large-capitalisation stocks, traded on U.S. stock exchanges.

[1] All coefficients are significant at 1% level, as shown in Table 4.7.

[2] All coefficients in return regressions are insignificant at all common significance levels, as demonstrated in Table 4.7.

5.2 Research Limitations & Avenues for Future Research

We acknowledge several limitations which establish some compelling avenues for future research.

Bots and spammers In a recent study, Cresci et al. (2019) addressed bot and spammer activity omnipresent on Twitter, particularly in the realm of stock microblogs. The researchers utilised a state-of-the-art spambot-detection algorithm with the aim of classifying authors of Tweets, identified as suspicious, into bots and non-bots. As a result, approximately 71% of the authors of the suspicious Tweets were labelled as bots, with further investigation demonstrating that 37% of these users were suspended shortly after the conclusion of the study. Furthermore, the authors raised the issue of "cashtag piggybacking", which refers to the widespread co-occurrence of cashtags of low-value stocks and high-value stocks and is likely to be aimed at promoting the corresponding low-value stocks in this manner. Due to the feasibility concerns and the lack of affordable spambot-detection software for Twitter, we did not adopt any spam- and bot-detection techniques, potentially jeopardising the reliability of results.

Sample size and selection The limits on the Twitter API usage, together with the narrow timeframe of the study, posed a substantial constraint to the size of our sample. In comparison to the cross-sectional regressions of Gu et al. (2020) (Duz Tan et al., 2021), all based on over 247,698 (104,444) observations, the results of our regressions on sentiment rest on just over 30,000 observations. The rules applied in selection of relevant Tweets are another point of concern. For instance, although our queries to Twitter APIs specified our wish to filter out Retweets, some Retweets found their way into the resulting JSON files, which we appropriately removed during the pre-processing stage. Given this observation, how can we be certain that all Tweets matching our requests were returned by the API?

Domains of labelled texts Bloomberg proudly claim, that their team of human experts manually annotate domain-specific Tweets as positive, negative or neutral, based on the following question: “If an investor having a long position in the security mentioned were to read this news or tweet, is he/she bullish, bearish or neutral on his/her holdings?” (Cui et al., 2016). Considering that Renault (2020) found the sample size to be of crucial importance in training of classification models and keeping in mind the time constraints associated with our study, we had no choice but to resort to publicly available texts labelled for sentiment polarity, stemming from a diverse range of domains.

In light of the listed limitations, our suggestions for future research, using user-generated data from Twitter for stock market prediction, would be to adopt spambot-detecting mechanisms for use as a Tweet-filtering sieve, devote a significant portion of time to Tweet data collection to ensure result comparability and select domain-specific annotated texts for sentiment polarity classifier training. Following the proposals of Gu et al. (2020) and Renault (2020), we advocate for an in-depth exploration of intraday associations between Tweet-based and stock market features, as the ever-increasing speed of information transmission may serve as one of the potential reasons for the dissimilarities between our results and those of Gu et al. (2020) and Duz Tan et al. (2021).

Furthermore, one could consider weighing the sentiment scores by importance, with the help of the follower counts of Tweet authors, in addition to the numbers of times a Tweet has been Retweeted or marked as a Favourite. Finally, it could be of interest to investigate the potential benefits of using aggregated firm-level Twitter sentiment in forecasting of returns of the corresponding ETF, stock index or portfolio on a bigger dataset and for alternative stock selections.

6 Conclusion

Inspired by the vast amount of research, exploring the connections between Twitter feeds and stock market happenings, we set off to investigate whether firm-level Twitter sentiment and activity can explain interday returns and trading volumes of large-cap stocks, traded on U.S. stock exchanges. We trained a binary multinomial naïve Bayes sentiment polarity classifier, which was subsequently used to classify ca. 20,000,000 Tweets, related to S&P500 constituents, as positive or negative.

Unlike the works of Gu et al. (2020) and Duz Tan et al. (2021), our results attribute no significant predictive power to either of the Twitter features with respect to future returns, consistent with the efficient market hypothesis and the classic literature on the information content of stock microblogs, including the works of Sprenger, Tumasjan, et al. (2014) and Antweiler et al. (2004). The anti-climactic performance of the portfolios, constructed using a simple long-short strategy and Twitter sentiment as the sole trading signal, further support the lack of valuable informational content in Tweets, related to large-cap U.S. stocks.

On the other hand, our findings provide strong evidence, supporting the existence of contemporaneous associations between the Tweet-based and stock market features, lagged relationship of investor agreement with the following day's abnormal turnover, as well as the significant role of Tweet volume in predicting next day's abnormal trading volume. Finally, we observed that aggregation of Twitter sentiment towards individual constituent firms can be beneficial in forecasting the performance of the corresponding ETFs, stock indices and portfolios.

In retrospect, one cannot deny the existence of an association between the content published on the Twitter platform and the movements and turbulence of the stock markets, considering the ever-growing body of related illuminating research. Nevertheless, investors should be weary of the growing spambot population, clouding the judgements of automated classifiers, as well as the dominance of studies failing to attribute any significant predictive power to Twitter sentiment with regard to interday returns, when using Twitter sentiment as a trading signal.

Bibliography

Amihud, Yakov (2002). “Illiquidity and stock returns: Cross-section and time-series effects”. *Journal of Financial Markets* 5, pp. 31–56.

Antweiler, Werner and Murray Z Frank (2004). “Is all that talk just noise? The information content of internet stock message boards”. *The Journal of Finance* 59, pp. 1259–1294.

Araci, Dogu (2019). “Finbert: Financial sentiment analysis with pre-trained language models”. *arXiv*, e1908.10063.

Azar, Pablo D and Andrew W Lo (2016). “The wisdom of Twitter crowds: Predicting stock market reactions to FOMC meetings via Twitter feeds”. *The Journal of Portfolio Management* 42, pp. 123–134.

Barrie, Christopher and Justin Chun-ting Ho (2021). *academictwitteR: An R package to access the Twitter Academic Research product track v2 API endpoint.* R package version 0.3.1. URL: `https://CRAN.R-project.org/package=academictwitteR`.

Bartov, Eli, Lucile Faurel, and Partha S Mohanram (2018). “Can Twitter help predict firm-level earnings and stock returns?” *The Accounting Review* 93, pp. 25–57.

Bermingham, Adam and Alan F Smeaton (2010). “Classifying sentiment in microblogs: Is brevity an advantage?” *Proceedings of 19th ACM International Conference on Information and Knowledge Management*, pp. 1833–1836.

Bollen, Johan, Huina Mao, and Xiaojun Zeng (2011). “Twitter mood predicts the stock market”. *Journal of Computational Science* 2, pp. 1–8.

Cresci, Stefano, Fabrizio Lillo, Daniele Regoli, Serena Tardelli, and Maurizio Tesconi (2019). “Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on Twitter”. *ACM Transactions on the Web* 13, pp. 1–27.

Cui, Xin, Daniel Lam, and Arun Verma (2016). “Embedded value in Bloomberg news and social sentiment data”. Bloomberg Twitter Data Research Report. URL: `https://developer.twitter.com/content/dam/developer-twitter/pdfs-and-files/Bloomberg-Twitter-Data-Research-Report.pdf`.

Da Silva, Nadia FF, Eduardo R Hruschka, and Estevam R Hruschka Jr (2014). “Tweet sentiment analysis with classifier ensembles”. *Decision Support Systems* 66, pp. 170–179.

Dancho, Matt and Davis Vaughan (2020). *alphavantager: Lightweight R interface to the Alpha Vantage API.* R package version 0.1.2. URL: `https://CRAN.R-project.org/package=alphavantager`.

De Long, Bradford J, Andrei Shleifer, Lawrence H Summers, and Robert J Waldmann (1990). “Noise trader risk in financial markets”. *Journal of Political Economy* 98, pp. 703–738.

Duz Tan, Selin and Oktay Tas (2021). “Social media sentiment in international stock returns and trading activity”. *Journal of Behavioral Finance* 22, pp. 221–234.

Fama, Eugene F and Kenneth R French (1992). “The cross-section of expected stock returns”. *The Journal of Finance* 47, pp. 427–465.

Fama, Eugene F and James D MacBeth (1973). “Risk, return, and equilibrium: Empirical tests”. *Journal of Political Economy* 81, pp. 607–636.

Farooq, Umar, Hassan Mansour, Antoine Nongaillard, and Muhammad A Qadir (2016). “Negation handling in sentiment analysis at sentence level”. *Journal of Computers* 12, pp. 470–478.

Feinerer, Ingo and Kurt Hornik (2020). *tm: Text Mining Package.* R package version 0.7-8. URL: `https://CRAN.R-project.org/package=tm`.

French, Kenneth R, William G Schwert, and Robert F Stambaugh (1987). “Expected stock returns and volatility”. *Journal of Financial Economics* 19, pp. 3–29.

Go, Alec, Richa Bhayani, and Lei Huang (2009). *Twitter sentiment classification using distant supervision.* Tech. rep. Natural Language Processing with Deep Learning (CS224n) Project. Stanford, p. 1.

Gu, Chen and Alexander Kurov (2020). “Informational role of social media: Evidence from Twitter sentiment”. *Journal of Banking & Finance* 121, e105969.

Haddi, Emma, Xiaohui Liu, and Yong Shi (2013). “The role of text pre-processing in sentiment analysis”. *Procedia Computer Science* 17, pp. 26–32.

Harris, Milton and Artur Raviv (1993). “Differences of opinion make a horse race”. *The Review of Financial Studies* 6, pp. 473–506.

Hogenboom, Alexander, Paul Van Iterson, Bas Heerschop, Flavius Frasincar, and Uzay Kaymak (2011). “Determining negation scope and strength in sentiment analysis”. IEEE, pp. 2589–2594.

Hvitfeldt, Emil (2021). *emoji: Data and Function to Work with Emojis.* R package version 0.2.0. URL: `https://CRAN.R-project.org/package=emoji`.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013). *An introduction to statistical learning.* Vol. 2. Springer.

Jurafsky, Daniel and James H Martin (2022). *Speech and language processing.* Vol. 3. Draft. Stanford, CA.

Kiritchenko, Svetlana, Xiaodan Zhu, and Saif M Mohammad (2014). “Sentiment analysis of short informal texts”. *Journal of Artificial Intelligence Research* 50, pp. 723–762.

Kolchyna, Olga, Tharsis TP Souza, Philip Treleaven, and Tomaso Aste (2015). “Twitter sentiment analysis: Lexicon method, machine learning method and their combination”. *arXiv*, e1507.00955.

Koudijs, Peter (2016). “The boats that did not sail: Asset price volatility in a natural experiment”. *The Journal of Finance* 71, pp. 1185–1226.

Kralj Novak, Petra, Jasmina Smailović, Borut Sluban, and Igor Mozetič (2015). “Sentiment of emojis”. *PLoS ONE*, e0144296.

Li, Ting, Jan Van Dalen, and Pieter Jan Van Rees (2018). “More than just noise? Examining the information content of stock microblogs on financial markets”. *Journal of Information Technology* 33, pp. 50–69.

Long, Cheng, Brian M Lucey, and Larisa Yarovaya (2021). “’I just like the stock’ versus ’Fear and loathing on Main Street’: The role of Reddit sentiment in the GameStop short squeeze”. *SSRN*, e3822315.

Mahmoudi, Nader, Paul Docherty, and Pablo Moscato (2018). “Deep neural networks understand investors better”. *Decision Support Systems* 112, pp. 23–34.

Malo, Pekka, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala (2014). “Good debt or bad debt: Detecting semantic orientations in economic texts”. *Journal of the Association for Information Science and Technology* 65, pp. 782–796.

Manning, Christopher, Prabhakar Raghavan, and Hinrich Schütze (2010). *Introduction to information retrieval.* Vol. 16. Cambridge, UK: Cambridge University Press, pp. 100–103.

Manning, Christopher and Hinrich Schutze (1999). *Foundations of statistical natural language processing.* Vol. 2. Cambridge, MA: MIT press, pp. 172–175.

Mao, Huina, Scott Counts, and Johan Bollen (2015). *Quantifying the effects of online bullishness on international financial markets.* Tech. rep. Working Paper in Statistics. European Central Bank, pp. 2–21.

Newey, Whitney K and Kenneth D West (1986). “A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix”. *Econometrica* 55, pp. 703–708.

Oliveira, Nuno, Paulo Cortez, and Nelson Areal (2016). “Stock market sentiment lexicon acquisition using microblogging data and statistical measures”. *Decision Support Systems* 85, pp. 62–73.

Ooms, Jeroen (2020). *hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker.* R package version 3.0.1. URL: `https://CRAN.R-project.org/package=hunspell`.

Parkinson, Michael (1980). “The extreme value method for estimating the variance of the rate of return”. *Journal of Business* 53, pp. 61–65.

Ranco, Gabriele, Darko Aleksovski, Guido Caldarelli, Miha Grčar, and Igor Mozetič (2015). “The effects of Twitter sentiment on stock price returns”. *PLoS ONE*, e0138441.

Renault, Thomas (2017). “Intraday online investor sentiment and return patterns in the US stock market”. *Journal of Banking & Finance* 84, pp. 25–40.

– (2020). “Sentiment analysis and machine learning in finance: A comparison of methods and models on one million messages”. *Digital Finance* 2, pp. 1–13.

Rinker, Tyler W (2018). *textstem: Tools for stemming and lemmatizing text.* version 0.1.4. Buffalo, New York. URL: `http://github.com/trinker/textstem`.

– (2022). *qdapRegex: Regular Expression Removal, Extraction, and Replacement Tools.* 0.7.5. Buffalo, New York. URL: `https://github.com/trinker/qdapRegex`.

Ryan, Jeffrey A and Joshua M Ulrich (2022). *quantmod: Quantitative Financial Modelling Framework.* R package version 0.4.20. URL: `https://CRAN.R-project.org/package=quantmod`.

Saif, Hassan, Miriam Fernández, Yulan He, and Harith Alani (2014). "On stopwords, filtering and data sparsity for sentiment analysis of Twitter". *Proceedings of 9th International Conference on Language Resources and Evaluation*, pp. 810–817.

Schmeling, Maik and Christian Wagner (2019). "Does central bank tone move asset prices?" *SSRN*, e2629978.

Sharif, Arshian, Chaker Aloui, and Larisa Yarovaya (2020). "COVID-19 pandemic, oil prices, stock market, geopolitical risk and policy uncertainty nexus in the US economy: Fresh evidence from the wavelet-based approach". *International Review of Financial Analysis* 70, e101496.

Sprenger, Timm O, Philipp G Sandner, Andranik Tumasjan, and Isabell M Welpe (2014). "News or noise? Using Twitter to identify and understand company-specific news flow". *Journal of Business Finance & Accounting* 41, pp. 791–830.

Sprenger, Timm O, Andranik Tumasjan, Philipp G Sandner, and Isabell M Welpe (2014). "Tweets and trades: The information content of stock microblogs". *European Financial Management* 20, pp. 926–957.

Symeonidis, Symeon, Dimitrios Effrosynidis, and Avi Arampatzis (2018). "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis". *Expert Systems with Applications* 110, pp. 298–310.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede (2011). "Lexicon-based methods for sentiment analysis". *Computational Linguistics* 37, pp. 267–307.

Taborda, Bruno, Ana de Almeida, José Carlos Dias, Fernando Batista, and Ricardo Ribeiro (2021). *Stock market Tweets data.* IEEE Dataport. URL: `https://dx.doi.org/10.21227/g8vy-5w61`.

Tetlock, Paul C (2011). "All the news that's fit to reprint: Do investors react to stale information?" *The Review of Financial Studies* 24, pp. 1481–1512.

Uysal, Alper K and Serkan Gunal (2014). "The impact of preprocessing on text classification". *Information Processing & Management* 50, pp. 104–112.

Van Bommel, Jos (2003). "Rumors". *The Journal of Finance* 58, pp. 1499–1520.

Van Rijsbergen, Cornelis J, Stephen E Robertson, and Martin F Porter (1980). *New models in probabilistic information retrieval.* Tech. rep. British Library R & D Department, Cambridge, UK.

Wysocki, Peter D (1998). “Cheap talk on the web: The determinants of postings on stock message boards”. *SSRN*. University of Michigan Business School Working Paper No. 98025, e160170.