Portfolio Construction with News Sentiment Using Large Language Model [1]

Qi Zhang[2]

University of Technology Sydney

Abstract (Extended)

Most existing text-based sentiment measures in finance are lexicon-based which are effectively based on word counts of positive and negative sentiment dictionaries, and naturally lose most information. We measure news sentiment using BERT, a state-of-the-art large language model, which reads and comprehends the whole text, and explore return predictability based on Refinitiv Machine Readable News. The resulting portfolio achieves annualized Sharpe ratios of 2.79, 3.09, and 3.87 when considering news alerts, news alerts and articles' headlines, and article body contents, respectively, significantly higher than passive investment as proxied by S&P 500 index's Sharpe ratio of 0.32 and dictionary method of 1.59, 2.94, and 0.04, suggesting that large language models are much better at capturing sentiment, and dictionary methods struggle to extract information from complicated texts. Our results also imply that reacting too fast on incomplete textual news information may yield suboptimal performance. An interesting finding is that news of positive sentiment is tailored to fewer audiences, contain fewer topics, and are generally shorter.

[1] Previous Title: Portfolio Construction with News Sentiment. We thank participants' comments from FIRN 2022 and AFBC 2022. This version is subject to frequent updates.

[2] Qi.zhang-9@student.uts.edu.au

1 Introduction

In recent years, as more and more textual data are recorded digitally and made available for research, we see a surge in textual analysis in finance (Gentzkow, Kelly, and Taddy, 2019) applied in multiple fields of finance and economics. Finance and economics literature, in trying to understand textual information, typically relies on traditional dictionary-based bag-of-word methods and use limited data sources, such as front pages of Wall Street Journal, because of lack of datasets and machine learning techniques. Dictionary-based methods are simple to use and understand, and they avoid researchers' subjectivity once the dictionary is selected; we can apply dictionary-based method to any lengths of texts; once dictionaries are made public, we can easily replicate other people's studies (Loughran and McDonald, 2016).

However, dictionary-based methods are overly simplistic. To let the machine understand text, early studies represent the whole text to word counts using dictionaries, such as positive and negative sentiment dictionary in sentiment analysis, and effectively, they throw away almost all information from the text, including the context, word orders, inter-connections of words, grammars, and structures. In 2017, Vaswani et al introduced the Transformer architecture based on self-attention mechanism, leading to BERT of Devlin et al (2018), a milestone of language model in machine learning . It significantly outperforms previous models in natural language processing tasks, including sentiment classification. Effectively, for each word, BERT asks it to pay attention to each word (including itself) in the text, hence taking into account grammar, context, inter-connections of words (such as 'not'), word orders, et cetera. By utilizing all information in the corpus, it is a much more accurate representation of text.

In this essay, we overcome the limits of dataset by using Refinitiv Machine Readable News (MRN) database, which contains all American company news from 2001 to 2019 from Refinitiv (Thomson Reuters), and we choose FinBERT of Huang et al (2021), which is a fine-tuned version of the original BERT model using financial texts, to

classify news sentiment, to investigate how state-of-the-art large language models can be used in portfolio analysis.

While early studies in sentiment and opinion mining in computer science literature rely on simple lexicon-based models, they were quickly replaced by newer models, even if one tries to improve lexicon models by introducing rule-based models to simple lexicons. For example, while 'good' carries positive sentiment, 'not good' negates the positive meaning of the word 'good'. However, natural language is too complicated to fit in a rule-based model, unless the rule is endless. A breakthrough in word representation is Word2Vec of Mikolov et al (2013a,b), which represents word by high-dimensional vectors and captures each word's semantic meanings. However, it has two main drawbacks: 1) it is incapable of handling words that are not seen in training sample; 2) words semantically similar would still be given two completely different encodings. For studies using Word2Vec on sentiment analysis, see, for example, Zhang et al (2015).

Global Vectors for Word Representation (GloVe) of Pennington, Socher, and Manning (2014) from Stanford University is much better at handling words not seen in the training set by considering the whole corpus. Generally, word embedding tries to represent each word in a high-dimensional space, and in the process, words that semantically similar are close to each other. Unsurprisingly, generally the higher dimension, the better semantic meaning we can capture, but the more expensive training we face. Word embeddings are also the basis of encoder-decoder frameworks including BERT.

Before BERT, arguably the best algorithm in word embeddings is Embeddings from Language Model (ELMo) of Peter et al (2018), which tries to incorporate context into word embeddings. In ELMo, the same word would be given different embeddings depending on the context. This greatly improves its performance because now we can capture the same word's different meanings.

More recently, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al, 2018), based on transformer architecture of self-attention mechanism of Vaswani et al (2017) outperforms all previous models in NLP tasks including sentiment analysis. To our knowledge, the latest and most comprehensive comparison of transformer-based models against lexicon-based models in financial sentiment analysis is Mishev et al (2020). In this paper, the author compare performance of a large variety of textual analysis algorithms, including dictionary, Word2Vec, ELMo, GloVe, BERT, and other transformer-based models. The authors find that transformer-based models significantly outperform other models. Specifically, while LM dictionary achieves an accuracy of around 65%, while transformer-based models' accuracies are around 30% higher.

In this study, we use FinBERT model of Huang, Wang, and Yang (2021), which is BERT model fine-tuned using financial texts, to predict sentiment score of news pieces from Refinitiv MRN. We know that we should further fine tune a generic language model (including BERT) to domain-language because the specific terms used in any field is different from generic language. As training is extremely computationally expensive, we choose FinBERT, a pre-trained model using financial documents. The authors show that FinBERT, by 'speaking' finance language, achieves higher accuracy than the original BERT model and LM dictionary in financial sentiment classification.

While many previous studies rely on headlines only, we experiment portfolio performance when considering 1) news alerts only; 2) news alerts and articles' headlines; and 3) articles' body contents only. The comparison would be valuable to both finance academics and practitioners. If portfolio performance using only alerts achieves at least comparable results compared with news articles for finance academics, then researchers can continue using headlines only in their future research, which significantly reduces research cost, as databases containing news articles tend to be extremely expensive; for practitioners, news alerts are shorter and easier to process,

allowing one to react faster to news, hence they may choose to use headline only as a guide in making investment decisions if results. We do not have a definitive a priori expectation on which model should perform better, as we have reasons to believe either way: headlines are designed to be concise but precise, capturing the most important and relevant aspect of an event and importantly, is less noisy, hence models using only headlines may achieve better results; article bodies are much more informative, and may be especially important in complicated cases where we can't summarise the event using a one-sentence headline, hence we should believe that models using article bodies should achieve better results, and that the performance should be higher with longer articles.

We form zero-cost long-short portfolios by going long the most positive-sentiment stocks and short the most negative sentiment stocks. In the simplest (and what appears to be the most popular) form, we have three labels: positive, neutral, and negative. We note that there are at least two FinBERT models, the older version of Araci (2019) appears to be more widely known, but the newer FinBERT model performs better, likely due to the significantly larger dataset in training their model.

We achieve annual Sharpe ratios of 2.79, 3.09, and 3.87 respectively under the 3 strategies. Our results suggest that reacting too fast on incomplete information as they arrive may not always be a good idea as this may reduce portfolio performance. If we believe that market prices reflect and incorporate any information and event that already occurred and is known to the market, then financial academics and practitioners should only react to news alerts, which are timely reflections of events that just occurred. This sheds new light to optimal trading for high frequency traders who seek to react to new information as quickly as possible as the new information hit the market, where doing so may unexpectedly reduce their performance. By considering news pieces with body contents only, we also have fewer pieces of news and hence number of stocks to work with, further reducing transaction cost. Previous studies which rely on headlines only

are often forced to do so because of lack of data and lack of appropriate machine learning methods to extract sentiment information in long news bodies. Our study, by formally comparing portfolio performance based on headlines and articles bodies, fill the gap in literature.

We also document an interesting pattern in news of positive, negative, and neutral sentiments: news pieces of positive sentiment are tailored to fewer audiences, contain fewer topics, and are generally shorter compared with neutral and negative sentiment news. This suggests that news does not just differ in their sentiment and informational contents, and merely the way the news is presented may contain valuable information.

The rest of this paper is organized as follows: section 2 presents related literature, section 3 describes our data and data cleaning procedures, section 4 presents methodology on BERT. Section 5 presents results and robustness checks,and section 6 presents conclusions and suggestions for future research.

2 Related literature

In this section, we give a short lookback of investor sentiment research in finance and economics. We note that investor sentiment proxies in literature include but are not limited to:

1. Investor and consumer confidence surveys. For example, Charoenrook (2005) and Lemmon and Portniaguina (2006). They usually find negative relationships between investor sentiment and future stock market return, i.e., stock price reversals. Some studies do find no statistically significant results or positive relationships between investor sentiment and stock price, such as Solt and Statman (1988), Lee et al (2002), and Brown and Cliff (2004).
2. Proxies for investor sentiment using market-wide variables, such as Baker and Wurgler (2006, 2007).
3. Use news and social media to proxy for investor sentiment. This is a large and

growing literature, and we also use financial news to extract investor sentiment. We note, however, it is less clear if we extract just investor sentiment or information embedded in texts. This is a common issue found in this strand of literature. See, for example, Chen et al (2013), Ke et al (2019), Ghiassi et al (2013), Bollen et al (2011), and Li (2021).

4. Online messages boards. In the early days, Yahoo! Finance and Raging Bull tends to be popular sources of small investor sentiment, now Twitter is the main source. See, for example, Das and Chen (2007) and Tumarkin and Whitelaw (2001).

We have long known that investor sentiment, based not on (at least not entirely on) rational information, possesses predictive power and moves the financial market (Baker and Wurgler, 2006). Interestingly, while textual analysis in finance appears to be a new idea, efforts in exploring how textual information could help with financial decision making may be dated back to almost 90 years ago. In 1933, Cowles (1933) tried to predict stock market by subjectively categorizing articles of William Peter Hamilton, editor of Wall Street Journal, into bullish, bearish, and doubtful sentiments. At around the same period, academics including Keynes (1936) have realized that investor sentiment affects stock market behavior and causes asset prices to deviate from their fundamental values.

After the pioneer studies, a developed version of bag-of-words based natural language processing model as applied to finance is experimented by a few researchers, an important one is Tetlock (2007). In this paper, the author uses General Inquirer (GI), a classical content analysis tool first developed in the 1960s, to analyze the contents of *Abreast of the Market* section of the *Wall Street Journal*. This technique counts words in 77 pre-determined GI categories from the Harvard psychosocial dictionary. Tetlock then uses Principle Component Analysis (PCA) to collapse the 77 categories into a single media factor, which is highly correlated with pessimistic words in the media, hence he calls it ‘pessimism factor’. This measure is predictive of market price drops

which is followed by reversion to fundamental value. Unusually high or low pessimism also predicts market turnover. As Tetlock himself points out, GI is only able to distinguish between positive and negative words, while in his study, he uses only negative words. GI itself is unable to capture contexts and more complex semantic meanings beyond each word themselves. In a follow-up paper, Tetlock et al (2008) shows that language used in company-specific words can predict firms' earnings and stock returns. The author's choice of dataset is popular in literature where researchers are (often forced) to use a small dataset, such as front page or certain column of *Wall Street Journal* or other news media, because: 1) alternative dataset is unavailable; 2) there are limited techniques for extracting information from full text; and 3) it is computationally expensive to use full texts. With better computation power, state-of-the-art machine learning technique, and Refinitiv MRN, which is a comprehensive dataset, we are able to overcome such challenges.

Ke et al (2019) is a recent study in textual sentiment analysis. In this paper, the authors use data from Dow Jones Newswire, which contains all historical news for the US companies from 1986 to April 2020. The authors start from the view that news simultaneously affect investor sentiment and market return and propose a three-step framework: 1) as positive sentiment drives up return and negative sentiment drives down return, we can use market reaction as a guide to automatically create dictionary of positive and negative sentiment; 2) use a two-topic model to estimate positive and negative sentiment scores; 3) predict sentiment scores of news articles. With the predicted sentiment score, the authors then go long the 50 most positive stocks and go short the 50 most negative sentiment stocks each day to forma a zero-cost portfolio. The authors achieve an annualised Sharpe ratio of 4.3 overall. This method can be considered an 'advanced' dictionary-based technique while our method is large language model based. We note that the dataset used in this paper is more comprehensive than ours and contains third-party news. While we intend to follow a similar approach and use market reaction as guide to fine-tune a sentiment classification

model, we were restricted by computation power and could not perform a similar analysis. Future researchers who have sufficient computation power may follow this approach and see if and how using market reaction in additional to large language model may further improve portfolio performance.

3 Data

We obtain American company's historical news from Refinitiv Machine Readable News database. The sole provider for our dataset is Refinitiv (formerly Thomson Reuters). Our sample contains 24 years' news from January 1st, 1996, to December 31st, 2019. Consistent with Ke et al (2019), we keep only news with one company tag, because sentiment content and information contents of individual companies are unclear when one piece of news relates to multiple companies.

Refinitiv MRN contains two types of news: title-only alerts, and articles that have body contents. The dataset contains 59.2% alerts and the rest are articles.

3.1 Data cleaning

We obtain each company's intraday, open, and close prices from Refinitiv (Thomson Reuters) Tick History (TRTH) database. Risk-free rate is approximated by T-bill rates which is obtained from Fed St Louis website. We align our portfolio analysis with risk-free T-bill rates by adjusting starting period of portfolio analysis to 2001. When a piece of news relates to specific companies, the news would be tagged with the companies' Reuters Instrument Code (RIC), and we use RICs to match and identify companies. Companies' RICs are in the format of "Ticker.Exchange", where the suffix identifies the exchange. For example, "AAPL.O" stands for Apple traded on NASDAQ. To identify and filter for US stocks (i.e., find the 'Exchange' part of the RICs), we use the latest US exchange suffix provided by Refinitiv. Unfortunately, Refinitiv does not maintain a historical list of US exchange suffix, and we note that we would lose a small number of exchanges that existed some time during the 26-year period and disappeared

at some point in time. There are a total of 4,807,623 observations from 14,214 companies, where 4,389 firms still exist today, and 7,864 companies are historical and were delisted at some point in time. 1,961 companies are not found in TRTH, but they only account for 17,518 pieces (or 0.36%) of news. Such news may include failed IPOs, data error, Refinitiv service alerts and maintenance, et cetera. After excluding companies not identified in TRTH, we have a total sample of 4,790,104 pieces of news from 12,253 companies over a 24-year period. The original time stamp in Refinitiv MRN is in UTC time, and we convert to NYSE exchange time to align with trading hours.

Figure 1 presents boxplot and distribution of total amount of news of each company and Table 1 shows its distribution. While on average, one company has 390 pieces of news, the distribution is extremely dispersed: at least 25% of companies have at most 4 pieces of news, and at least 50% of companies have at most 27 pieces of news. A large number of news is NYSE order imbalance information automatically issued at around 3:40pm and 3:50pm each trading day. A typical such 'news' reads: "*NYSE ORDER IMBALANCE <F.N> 98700 SHARES ON SELL SIDE*". While they are posted as articles, their titles and bodies are identical. Such information was first available on October 12th, 2015, and account for 647,975 (or 13.52%) of total observations. While they are potentially useful information, they merely contain numerical order imbalance information and there is little room for sentiment analysis. We therefore remove news of NYSE order imbalance information. Now we are left with 4,142,129 pieces of news from 11,968 companies, where 59.16% are headline-only and the rest are articles with body contents.

Table 2 shows the distribution of article's lengths in number of words. The distribution is extremely dispersed with a very long right tail: while the mean number of words is 151.5, its standard deviation is 204.5, and this is stretched by a large number of articles with very long lengths. The very short articles are polluted by the way Refinitiv posts

its news. We first filter for news articles by applying the same filtering rules for alerts-only news pieces and have a total of 2,339,138 news pieces. We note that some articles are exceptionally long and others exceptionally short. In Figure 2, which shows the distribution of articles' lengths, we note that the distribution of articles' number of words is extremely skewed. There are too many articles with unreasonably short body contents; specifically, at 35 percentile, articles' length is only 11 words. To have a sense of what the articles are about, so as to see if there are any 'patterns' or 'series' of news articles.

We manually investigate short articles and found that:

- Most of the unreasonably short articles with one or two words are NYSE indication information, such as the last, bid, and ask price of a stock, where the body only contains one or two numbers.

- For articles of length around 11 words, they are predominately NYSE order imbalance information automatically posted at end of the trading day and NYSE indications, which contains very similar information to one and two words 'articles', where the body just restates their headlines.

- For articles of length between 10 to 20 words, they are mainly reminders of upcoming events, such as corporate news announcements.

- For articles of length 20-30 words, they are often brief notes to notify the reader there have been duplicated news items and certain news should hence be disregarded. They also contain links to news press they Reuters terminal users may access. Such news also includes initial public offerings (IPOs) and seasoned equity offerings (SEOs) where the body is unstructured and simply contains, for example, issuer and issued price.

- Articles of length 30-59 words also contain predominantly bonds and equity issue information, notices of press release available in Reuters terminals, and short news

items from media. For the latter, the body contains links, contact numbers, and identity of the media source. There are also some brief news where the body merely restates the headline, sometimes with a rephrasing. The significantly longer length in the body compared with the headline is brought by some fixed formats in the body, such as date and source. We hence include media and briefs but only include them in headline-only analysis. We exclude the words 'brief' and 'media' before running classification.

- For articles of around 57 words long, they mainly contain Refinitiv's (Reuter's) legal declarations on news reposting.

For such articles, while they often do contain valuable information on market and stock fundamentals, trades and quotes activities, et cetera, they are insufficient for textual analysis. We therefore exclude all articles with fewer than 60 words. This leaves us with 730,590 unique news articles. We note, however, we do exclude some potentially useful articles in this way. For example, many articles are briefs. Briefs start with 'BRIEF' in their body contents, and we include all briefs regardless of the article's length in preliminary screening. A closer look at briefs reveals that body contents of briefs are usually unstructured, bullet points-like sentences capturing key points in some events, typically financial states. A comparison of a typical brief and an ordinary, non-brief news' body contents is displayed in Figure 10. While readers can indeed form sentiments after reading such briefs, algorithms are typically not able to capture sentiment contents as sentiments are more likely derived from performance themselves (i.e., numerical information) instead of the use of words and phrases (textual information). By contrast, an ordinary non-brief news is typically well-structured articles with proper use of grammar and semantics. We hence use briefs only for title-only analysis.

We have dealt with the short articles and now we turn to long articles. As there is no guidance on appropriate thresholds on article length, we use a subjective threshold and remove all articles beyond 95 percentile (434 words).

We see a drastic decline in number of articles, and we have only a total of 631,521 articles after cleaning. We perform the same set of analysis as before with FinBERT model. We note that processing time for articles are much longer than title-only news pieces.

3.2 Descriptive Statistical Analysis

We now ask: what companies attract reporter attention and have the most number of news, and what the news are about. To answer the first question, we plot the top 30 firms' amount of news. To second the latter question, we explore topic codes. Each piece of news is tagged with a long list of topic codes covering the company's geographic location, industry, asset, events, et cetera.

Figure 3 plots the top thirty firms' amount of news. Five companies have over 20,000 pieces of news: General Motors, Boeing, Ford Motor, Citi Group, and General Electric. Interestingly, all of the thirty companies are from NYSE. 21 companies have more than 10,000 pieces of news. Not surprisingly, companies attracting the most number of news are the large ones.

Figure 4 plots the top 30 topics. Because topic descriptions can be long, we keep only topic codes in this plot. The top 4 categories are US, America, North America, and company news. These 4 topics are not very informative to us as this is how we cleaned our data. We see that the news are about very different topics: corporate events, mergers and acquisitions, financial events, consumer cyclicals, market events, and industrials (TRBC level 1), et cetera.

We now consider how the number of news evolve over the years, across months, and intraday. The patterns by time are largely consistent with intuition, although there are indeed some surprises.

Figure 5 plots the average amount of news by hour and by minute of day. Trading hours (9:30am to 4:00pm) are shown in blue line while non-trading hours are shown in red. There is a very clear pattern intraday: news arrivals are more intense just before market open and close and is calm during the day. There is a small flush of news during middle of the day after which news arrivals are low again. This pattern is consistent with Dow Jones Text Feed & Archive database shown in Ke et al (2019).

Figure 6 plots the number of news pieces by year, percentage of news arrivals by day of week, and number of news arrivals by day of year. While there is a generally increasing trend in news arrivals, market was especially turbulent and produced exceptionally high volume of news during the 2008 Global Financial Crisis. During the beginning of the millennium, the market sees exceptionally low news arrivals. The month before reporting period (December, March, July, and September) see low levels of news arrivals while the month of reporting period see peak in news arrivals. This is as expected as we only have company-specific news in our sample and considerable number of news reports are from or are about company events. It is unsurprising that weekends see minimal news arrival. News arrivals increases from Monday to Thursday and drops on Friday where people appear to start weekend mode and produces less news. We also note very strong seasonality and holiday effects, where around report dates each quarter, market produces large amount of news, and would get calm after reporting days. New year, Christmas, and mid-year holiday seasons see lowest level of news arrivals.

4 Methodology

This section describes methodology used in this paper. We start with the Transformer framework and then show how we build our model.

4.1 Transformer, Attention, and BERT

This section provides a very high-level overview of Attention mechanism and BERT model. In *attention mechanism*, each token (i.e., word or pieces of word) is given an *attention score*, and each hidden state of each word directly considers each word (including itself) in a given corpus. The *attention score* determines how relevant each word is to a given word for each hidden state. *Attention* provides a solution to information bottleneck problem of previous Sequence-to-Sequence models because now we establish direct connection from the decoder to each hidden state of the encoder to focus on a particular part of the source sequence, where the exact part of source sequence to focus on is given by the attention score. It also allows parallel computing, greatly improving performance.

To define attention, we consider encoder hidden states h_i (*values*) and decoder hidden state s_t (*query*). Then attention score is:

$$e^t = [s_t^T h_1, \dots, s_t^T h_N]$$

And attention distribution for each step is:

$$\alpha = softmax(e)$$

The attention output is:

$$a_t = \sum_{1}^{N} \alpha_i^t h_i$$

We note that there are several versions to compute attention score, e_t, such as :

1. Basic dot-product attention:$e_i = s^T h_i$;
2. Multiplicative attention: $e_i = s^T W h_i$, where W is a weighting matrix;
3. Additive attention: $e_i = v^T \tanh (W_1 h_i + W_2 s)$, where W_i is weighting matrix.

Araci (2017) provides an excellent summary of *Transformer framework* and we reproduce here: '*The encoder consists of multiple identical Transformer layers. Each layer has a multi-headed self-attention layer and a fully connected feed-forward network. For one self-attention layer, three mappings from embeddings (key, query and value) are learned. Using each token's key and all tokens' query vectors, a similarity score is calculated with dot product. These scores are used to weight the value vectors to arrive at the new representation of the token. With the multi-headed self-attention, these layers are concatenated together, so that the sequence can be evaluated from varying "perspectives"*.'

The way *Transformers-based models* work is: we input documents (or sentences, or parts of sentences, et cetera, depending on the task) into the encoder part of *Transformer*, which converts the input to text embeddings (or features) and tries to understand the text through multiple transformer blocks, where each block contains *multi-head attention* for it to understand connections between words. For this reason, encoder-only models are typically good at tasks where it is crucial to understand the text, such as classification. The word embedding in transformer models typically use very large dimensional vectors to represent each word. For example, in BERT, each word is represented by a 768-dimensional vector. Large dimensions would increase precision but also increases computing burden. The *decoder* then takes as input the output of encoder. Therefore, the *decoder* would have all information from the input sentence and work sequentially (or piece-by-piece) when translating into the target sentence.

BERT is designed specifically for langue understanding. Previous models only consider a whole sequence left-to-right or right-to-left, but language understanding is bi-directional; or more precisely, it is all-directional, as we consider each word's relation to each other word in order to understand natural language. Previous models tend to build on a uni-directional framework because: 1) we need a direction to generate well-formed probability distribution; 2) in uni-directional framework, we build each token's

representation incrementally, while in a bi-directional framework, words can ‘see themselves’. BERT’s solution is to mask out 15% of input words and let the model predict the masked word that it never sees; to learn relationship between sentences, BERT’s second task is next-sentence-prediction. BERT is trained on Wikipedia (2,500 million words) and BookCorpus (800 million words).

Because even the same word in financial documents may have different meanings, it is crucial that we consider the *context*, it is therefore a great improvement over previous lexicon-based models. For example, the word ‘bank’ may mean the organisation where we borrow and lend money, but in industry-specific texts, even if it is from a financial news vendor. Consider the sentence ‘*Hundreds of thousands of Hindu worshippers flocked to the banks of the Ganges in India’s West Bengal state Friday, braving a surge in Covid-19 infections to bathe in the waters of the holy river*’ from the news titled ‘*Thousands take holy dip in India's Ganges River amid Covid surge’*. The meaning of bank in this context is obviously different from the more popular meaning of *bank* in financial texts. By ignoring contexts and such double meanings of single words, lexicon-based models are only capable of giving a coarse impression of a sentence’s meaning.

4.2 FinBERT

We know that we need to adapt any general model to domain-specific language to achieve higher performance. In other words, we need to make sure the model ‘speaks’ finance language, because the terms, jargons, et cetera used in financial documents are different from general language. In our baseline model, we use FinBERT of Huang et al (2021) which fine-tunes BERT using a huge dataset of financial documents, including:

1) 60,490 Form 10-Ks and 142,622 form 10-Qs of Russell 3000 firms during 1994 and 2019 from SEC website (2.5 billion tokens);

2) 136,578 earnings conference call transcripts of 7,740 public firms between 2004 and 2019 (1.3 billion tokens);

3) 488,494 analyst reports in the Investext database issued for S&P firms during the 1995-2008 (1.1 billion tokens).

The authors then train their model using labelled data from open sources for sentiment learning, including: Financial PhraseBank (4,845 sentences) + AnalystTone (10,000 sentences) + FiQA (1,111 sentences). These sentences are human labelled into positive, negative, and neutral labels and represent sentiment content as understood by human. We use this model in our baseline result.

Before continuing, we note an important difference in literature in studies using sentiment-implied returns. If we believe that after a news announcement, market participants react by doing two things:1) form sentiment; 2) react to the news, then market reaction is a natural guide to investor sentiment, and we can use market reaction as target variable to label news into positive, negative, and neutral sentiments. While this strand of study is interesting (such as Ke et al, 2019), this method intrinsically cannot identify investor reaction to information in the news and investor reaction to pure sentiment, and for this reason, it's probably safer to call this method 'soft information-based' study instead of sentiment based. Investor sentiment in computer science literature (including FinBERT of choice in this study) typically uses human-labeled texts, which is a more accurate indicator of investor sentiment. We tried to carry out analysis by fine-tuning BRET model using market reaction of news as guide but eventually could not do so due to hardware restrictions. Future studies may attempt to investigate how large language models can be used to directly explain and predict stock return directly, instead of through sentiment.

5 Results and robustness check

In this section, we present results of portfolio construction strategy together with robustness checks.

5.1 Baseline model results

We present baseline results using FinBERT of Huang et al (2021) using alerts. After cleaning, we have 2,287,700 unique news alerts. We note that in this analysis, out dataset is significantly smaller than Ke et al (2019) which uses Dow Jones Newswire and after cleaning, has 6,301,532 news pieces from 2004.

We note that none of the training data in FinBERT model contain Refinitiv Machine Readable News, and in this sense, the labels are all predicted labels as none of the news texts are seen in the model's training phase. To form portfolio, we follow the following procedure: for each trading day, we consider news arrivals from 9:00 am the previous trading day (Day t-1) to 9:00 am of the current trading day (Day t) and use this set of news to assess each company's sentiment for portfolio formation. The trading day is aligned to the NYSE trading days. FinBERT model has two outputs: a label (positive, neutral, or negative), and probability of the predicted label. We convert factor label of FinBERT output to numerical as follows: if the predicted label is neutral, we assign a score of 0; if the predicted sentiment is positive, we assign a score that is equal to the probability; if the predicted sentiment is negative, we assign a score that is equal to negative of the probability. For example, if the probability is 0.85 with a label of positive (negative), we assign as core of 0.85 (-0.85). This is easy to use and fits our purpose: we wish to ask, for each news, how likely is it positive? If a company attracts more than one news for a given trading day, we then calculate a simple average of the scores.

Table 2 shows the number and percentage of trading days with at least 50, 80, and 150 (X) news alerts and Figure 7 shows the distribution of news alerts each trading day.

Technically, we do not have a ‘daily’ portfolio; instead, we ask: for days with at least m number of news, form a zero-cost portfolio by going long the top n1 most positive sentiment stocks and going short n2 most negative sentiment stocks, where n1 is the smaller of N or the number of positive or negative news if we have so few news on the day that we cannot find N companies for the day. For example, we form portfolios by considering only days with at least 50 news alerts for the day. Then for each day, we consider forming portfolios by going long the 30 most positive stocks that have a positive sentiment label and going short the 30 most negative stocks with negative sentiment labels. Apparently, there will be days where we do not have 30 stocks that are positive (negative), where we choose all available positive (negative) stocks in forming long (short) legs of the portfolio. Such a portfolio is called 50-30 portfolio and we experiment with different combinations of X-N and find the most optimal portfolio in terms of Sharpe ratio.

Table 3, column1 shows Sharpe ratio by year of the best-performing (50-30) news alerts portfolio, which has an overall Sharpe ratio of 2.79. The portfolio’s performance is volatile over the years. Figure 8 shows cumulative log-returns of the portfolio over the same period and shows a significant drop on 2016/12/23 where cumulative log-return dropped from 7.89 to 7.71 in one day. In the same graph, we also show the cumulative return of S&P500 index and other sentiment portfolios over the same period. S&P 500 index represents a passive investment that is still used actively in funds management, such as Vanguard. The baseline sentiment portfolio significantly outperforms a passive investment in S&P 500 index. We also note that the correlation between daily S&P 500 index return and sentiment portfolio return is very weak at -0.040, suggesting that portfolio return does not seem to originate from market movement but instead, from sentiment of the specific companies covered in news after investors read the news and form their sentiment. We note that the S&P 500 index has an overall annualized Sharpe ratio of 0.312 during this period, again suggesting the superior performance of the sentiment portfolio.

Figure 9 shows the number of stocks in long (red) and short (blue) legs of the portfolio for each trading day. We see that throughout the sampling period, long legs tend to have more stocks than short legs. Specifically, in the long leg, we see 30 stocks (maximum) in 74.5% of the time while in the short leg, only 26.6% of the time. While not immediately obvious in the graph, year-end periods around New Year's Eve tend to attract minimal number of stocks. This is because days leading to New Year's Eve are not public holidays and are hence included in the sample; however, this period unavoidably attracts little investor attention and media coverage, and people are in holiday mood. In the early part of the sampling period, both long leg and short leg see a large number of days with much fewer than 30 stocks in each leg. This is also observed in Ke et al (2019) where in their early periods, because of the small number of news pieces, we simply cannot find enough stocks for portfolio construction.

Figure 10 shows the mean score of stocks in long (red) and short (blue) legs each trading day. We see a generally increasing trend in the long leg and a generally decreasing trend in the short leg's sentiment score. There are two possible effects to this observed phenomena: 1) because the FinBERT (2021) model was trained on relatively recent financial documents, they naturally apply more to recent years' news, because the jargons, vocabularies, et cetera, used in financial news are changing over the years, even if the language is in the same domain of finance; 2) over time, we see a larger number of news released each trading day. In the early days, as there was too few news, we are forced to choose stocks whose sentiment scores are not high (or low) enough. The latter explanation coincides with Figure 9 and appears to be more plausible. Both legs do exhibit fair variation, indicating that news tones are simply more extreme in some days than others. We also note that positive and negative legs' mean sentiment score have a correlation of -0.45, which is moderately negative. This seems to suggest that the tone, or investors' sentiment after reading the news, are generally more extreme than moderate. In the long leg, we see significantly more days with sentiment scores closer to 1 than short leg with sentiment scores closer to -1.

We note that the correlation between daily portfolio performance and the total number of stocks in portfolio, number of stocks in long leg of portfolio, and number of stocks in short leg of portfolio are very weak at -0.025, -0.056, and -0.028, respectively, suggesting that increasing number of stocks does not seem to improve performance.

5.2 Articles

We now consider news articles in addition to headline-only news. We have noted that daily number of news is often a restricting factor especially during early days where we see significantly fewer news produced, this problem may be more severe for news articles because we have fewer news articles (even fewer after our cleaning process). We first look at a comparison between news alerts and news articles' predicted labels. Table 6 shows the percentage of predicted positive, neutral, and negative labels for news articles and news alerts. We note that there is no economically significant difference between predicted labels in news articles and news alerts and that most news are predicted to be of neutral tone. This is consistent with previous studies where neutral sentiments dominate.

Next, we ask: is informational contents of news articles reasonably captured in their titles? We perform two additional sets of analysis: one with all news alert headlines and news articles' body contents, and one with articles' body contents only.

We compare the same 50-30 portfolio based on news alerts only versus news alerts and news articles put together. We achieve overall Sharpe ratios of 2.79 (headline-headline), 3.09 (headline-article), and 3.87 (article body only) respectively, a significant increase from pure alerts-based results. We note that as with other deep learning models, it's slow to run[3] and running speed grows almost exponentially with text length, and this

[3] Articles' body contents after cleaning took around one full week to run on High Performance Computing to get their predicted labels.

may justify practitioners' decisions if they rely on headlines only. However, the performance improvements when using articles' body contents justify researchers' and practitioners' investments into computing power. The results are presented in Table 3, Figures 8, 9, and 10. While the mean score of long and short legs of portfolio exhibit the same trend under the 3 portfolio strategies, there are significantly fewer stocks in each leg in portfolio relying on article body only because of much fewer news articles. However, we see much higher portfolio performance with fewer stocks, which translates to lower transaction costs, under article body-only model.

Our results have important implications for investors and algorithm traders: while it is believed that faster reaction to information is desirable and investors hence try to improve their speed of reaction, it is probably not the case with textual data. The significantly higher Sharpe ratio achieved using articles' body contents while disregarding alerts suggest that with textual news data, it is advisable to wait until more information is released, and reacting too fast based on incomplete information may reduce one's profitatbility.

5.3 Breaking down portfolio return

Table 9 shows portfolio Sharpe ratios by breaking down Sharpe ratio to daily excess returns and standard deviations in Panel A and profitability by portfolio's long and short legs in Panel B. As we move from forming portfolio using alerts to using article body contents, portfolio's volatility stays relatively constant while portfolio return improves, leading to improved Sharpe ratio. Interestingly, the profitable days do not see an improvement: using alerts only yields 67.97% profitable days, while using article body contents only yields 63.73% profitable days. However, relying on article body contents allows us higher overall profitability despite fewer days with positive excess returns. From Panel B, we see that while the negative legs of portfolio yields higher returns than

positive legs (3 to 4.5 times higher), they are also more volatile, where standard deviation of short legs' returns are about twice as large.

Table 10 shows Fama-French 5-factor model results of the 3 FinBERT models using daily returns. We see that the only significant factor is HML and the three models consistently earn significantly positive alpha, suggesting that the sentiment portfolio's returns are explained by the return differences in value and growth stocks, and the consistently negative coefficient of HML suggests that the portfolio is sensitive to growth stocks.

As a guide, when daily transaction cost is 0.7%, 0.8, and 1.0%, the portfolio ceases to be profitable under alerts, alerts-headline, and article body strategies.

5.4 Robustness check: dictionary methods

In this section, we compare FinBERT's model's performance with dictionary methods. To be parsimonious, we perform sentiment analysis on headlines. Specifically, we use 1) the Harvard-IV4 dictionary, a general-purpose dictionary and 2) the Loughran-McDonald dictionary specifically designed for financial data. Sentiment scores based on dictionary method gives a continuous number from -1 (negative sentiment) to 1 (positive sentiment) based on polarity, where:

$$Polarity = \frac{Pos - Neg}{Pos + Neg}$$

Polarity is a simple but widely used metric in lexicon-based natural language processing models where Pos and Neg represent word counts of positive and negative words, respectively. Following standard pre-processing procedures, the headlines are stemmed, converted to lower cases, and removed of stop words before conducting dictionary-based models.

We start by noting that sentiment scores from the LM dictionary, Harvard-IV4 dictionary, and FinBERT models are very low in Table 4. The general dictionary, HIV4,

has very low correlation with FinBERT model and moderate correlation with the LM dictionary. The highest correlation is between the FinBERT model and LM dictionary at 0.327, but it is still just moderately positive.

We then perform the same portfolio construction exercise as before. Using LM dictionary, we achieve Sharpe ratios of 1.59, 2.94, and 0.04 under news alerts, news alerts and article headlines, and article body contents, significantly lower than FinBERT model. Cumulative returns under LM dictionary are plotted in Figure 8. A striking feature is that when using only article body contents, LM dictionary and FinBERT gives the lowest and highest performance, respectively. While including news articles' headlines to news alerts significantly improves portfolio performance, the performance drastically drops when using article body contents only, right the opposite of what we observe in FinBERT model performance. This suggests that dictionary methods are not good at identifying sentiment and information contents from long and complicated texts. Using article body contents under LM dictionary would even give inferior performance than the market.

5.5 Heterogeneity in news of different sentiments

In this section, we ask: is there heterogeneity in news? If news only differs in their informational and sentiment contents, then there should be no difference in anything but their contents. We test using a smaller sample in this section[4]. Specifically, we consider 2014-2019 only. We require, on average, a stock to have at least 30 company-days per year to include it into our analysis. This leaves us 44 companies in articles and 190 companies in headlines. We use the 44 companies that appear both in news articles and news alerts for analysis. 3 companies are now delisted: Raytheon Co (RTN.N), Twitter (TWTR.N), and United Technologies Corp (UTX.N). The 44

[4] This section entails the use of high-frequency data, and we would easily exceed our faculty's download limit if considering all stocks.

companies are large companies in financial, manufacturing, automobile and airline, retail, food, consumer goods, high-tech, and financial industries and the list is shown in Table 7.

We table the differences in length of article body, number of subjects, and number of audiences for alerts, news articles' headlines, and news articles' bodies together with their F-statistic in Table 8. All p-values are lower than 0.00 where we refrain from showing in the table to be parsimonious. We note that news does appear to exhibit heterogeneity beyond their contents, as there are economically and statistically significant differences in positive, neutral, and negative sentiment news' length, number of subjects, and number of audiences the news is tailored to. Specifically, news of positive sentiment is tailored to fewer audiences, contain fewer topics, and are generally shorter.

We also check if they differ in volatility. To do so, we compute 5-minute realized volatilityfor each stock-day and ask the percentile of the volatility measure from 20 days before and 20 days after a specific date. This hence can be considered a 'local percentile' of volatility. We see that there is also heterogeneity in volatility percentiles. It hence appears that positive, neutral, and negative news are intrinsically different and warrant investigation on their own. Previous studies which consider only sentiment-carrying news, i.e., positive and negative news only, potentially gives up valuable information. While the result on realized volatility is probably intuitive, it's probably surprising that there are statistically different length of article, number of news audience, and number of subjects in positive, neutral, and negative news.

5.6 Can news heterogeneities explain sentiment portfolio returns?

In this section, we ask: given: 1) the sentiment portfolio's return remains unexplained by most Fama-French 5-factor model risk factors, and 2) there are heterogeneity in non-content perspective of news, can we find alternative ways to explain the portfolio return? To do so, we include the news topics, audiences,

5.7 What is the news talking about?

In this section, we present word cloud of news used in positive and negative legs of portfolio in Figure 11. Larger fonts indicate more frequent words. Note, however, this is different from sentiment words in traditional dictionary methods: words in dictionary-based methods are sentiment-charged, which are the only words the algorithm looks for when assessing documents' sentiment contents; the word cloud we present here concerns all words in a document and are not 'sentiment-charged'. Instead, they give us a general impression of *what the news talks about*. We perform usual pre-processing step by removing stop words, digits, turning words to lower-cases, and manually remove some too common but meaningless words (such as 'Reuters').

We note that overwhelmingly many of the news pieces used in portfolio constructions talk about share prices and companies' profitability, which is intuitive. They carry clean identification of news sentiment and is usually what moves stock market. Investor sentiments based on investor reactions to such news are carried forward to the next trading day, making it a profitable opportunity to trade on previous days' sentiment. This, however, again suggests market inefficiencies: market doesn't absorb all information into asset prices adequately and timely, leading to profitable opportunities using past news.

6 Conclusions and further research

In this essay, we explored how transformer-based algorithms, state-of-the-art technique in natural language processing, helps in sentiment extraction and portfolio construction. Deviating from previous textual analysis methodologies in finance, which are generally dictionary-based and reduces news (or generally, documents) to word lists of positive, negative, and neutral words, transformer-based models are capable of understanding

the semantic meaning of each word in a sentence; it also considers context, hence greatly improving the algorithm's ability to understand natural language. We use Refinitiv MRN database which consists of all historical North American company news from Refinitiv (Thomson Reuters) from 2001 to 2019. This contains much more information than previous studies, which typically use headlines of a small number of stocks' news or certain columns of Wall Street Journal to extract sentiment information.

The comprehensive dataset allows us to explore features of news arrivals, which is also important for cleaning and structuring dataset. We observe strong seasonality of news arrivals where there are four 'waves' of news arrivals throughout the year, consistent with four reporting periods. News production is higher around seasonal reporting periods and significantly drop after that. During the trading day, news arrivals are significantly concentrated around market close, and news production during the day in other periods are (probably surprisingly) lower than market close period.

Our results show that sentiment portfolio significantly outperforms a passive investment in S&P 500 index with a Sharpe ratio at least eight times higher in the baseline model and 12 times higher when considering news articles. The daily performance of our portfolio and market return exhibit very low correlations, suggesting that the sentiment portfolio's return is not from general market movements. Including body contents significantly improves portfolio performance. While BERT models (including FinBERT) are slow to run even on High-Performance Computers (HPCs), it's worthwhile to consider body contents of news pieces instead of using headlines only. Probably surprisingly, reacting too fast on incomplete textual information may be a bad idea as this reduces portfolio performance. This is likely good news to smaller investors: not having access to fast algorithms and forming portfolios based on a small number of complete news yields significantly better portfolio performance than reacting to every single piece of news that arrive to the market, hence

significantly reducing their cost of trading while improving portfolio performance. Not surprisingly, BERT-based model significantly outperforms dictionary-based method.

An interesting finding is that there are economically and statistically different number of topics, audiences, and news length among news of different sentiment. In future research, researchers may attempt to explore causes of this finding and attempt to investigate topic information in news.

We attempted to use market reaction as guide on news sentiment and train an in-house BERT model in addition to FinBERT model we employ in this study. However, hardware limitations prevented us from carrying out this practice. Future researchers who have sufficient computing power may attempt to train a regression model using market reaction of the stocks following the news as target variable, and this may potentially improve portfolio performance.

7 Appendix

7.1 Appendix A. Transformer Models and BERT

7.1.1 Encoder Models

Many models, including BERT, uses only the encoder part of transformer. Such auto-encoder models typically tackle tasks where it is crucial to understand they language. They are '*bi-directional'* because the attention layer at each state has access to information from all inputs, and it is perhaps more appropriate to call them 'all-directional', as opposed to older models where they go *sequentially* from left to right or from right to left, making it a much more accurate language representation. Other examples include ALBERT (Lan et al, 2019) and RoBERTa (Liu et al, 2019). Many such models contain 'BERT' in their names because they are improvements over the original BERT model.

7.1.2 Decoder Models

Decoder models only use decoder part of the transformer architecture. Each hidden state at each word has access to only words generated so far, and they are typically good at generative tasks, such as next sentence prediction. Influential models include GPT of Radford et al (2018), GPT-2 of Radford et al (2019), and Transformer XL of Dai et al (2019).

7.1.3 Encoder-Decoder Models

Lastly, we have encoder-decoder models with use both encoder and decoder part of the transformer architecture. Both encoder and decoder use self-attention, but each word's hidden state in the encoder can 'see' all other words in the corpus, while in the decoder, each hidden state of each word only has access to all preceding words only. Because of this, encoder-decoder models are best suited for generative tasks where input acts as guides, such as summarization. Important encoder-decoder style sequence-to-sequence models include BART of Lewis et al (2019) and Google's T5 of Raffel et al (2019).

7.1.4 BERT

BERT (Devlin et al 2018) is effectively formed by layers of Transformers stacked together. It is an unsupervised word representation model by using two new pre-training objectives: masked language model (MLM) and next sentence prediction (NSP). Now both words before and after a central word are explicitly accounted for. BERT has two versions: BERT-base, with 12 encoder layers, hidden size of 768, 12 multi-head attention heads and 110M parameters in total; and BERT-large, with 24 encoder layers, hidden size of 1024, 16 multi-head attention heads and 340M parameters. This again illustrates the importance of sufficiently large dataset: deep learning models (and generally, all statistical models) are only as good as their training sets; because of the huge number of parameters, previous studies using human-labeled financial texts are far from sufficient for good results, and this may (at least partly) explain why in some previous studies, transformers failed to outperform lexicon-based models. BERT is pre-trained on English Wikipedia.

7.2 Appendix B. Literature Review on Sentiment Analysis Methods

The formal study of sentiment on financial market appears to begin after the Efficient Market Hypothesis of Fama (1970), one of the most important works in traditional finance research, where behavioral economists began to realise that the EMH could not explain observed market dynamics, and that market does not timely incorporate information into asset prices. For example, Shiller (1980) finds that financial market shows excessive volatility when new information arrives; Summers (1986) then finds that most empirical studies have little statistical power in testing the EMH, and that stock prices may not reflect rational, fundamental value. De Bondt and Thaler (1985) and Cutler et al (1989) were among the first studies to establish how news affect market prices. Specifically, Cutler et al (1989) finds that macroeconomic news explains only less than 1/3 of the total variance in prices, and that market reaction to major political and world events are small. Barberis et al (2005) further establishes a model of investor

sentiment and show that news can cause both over-reaction and under-reaction depending on how we measure sentiment and the timeframe considered. Ke et al (2019) argues that market is inefficient and prices do not reflect all information, hence several days after news announcement, we still observe profitable opportunities.

As early as the early 2000's, academics started to learn that information posted on the internet affects stock prices, either because the postings contain new information or because they represent successful attempts to manipulate stock prices (Tumarkin and Whitelaw, 2001). However, extracting useful information from texts is difficult and early studies used coarse methodologies that unavoidably affected their results, as we are unable to tell if an insignificant result reveals the true relationship of interest or because the methodology fails.

One of the earliest studies of financial and accounting research using textual data is Ingram and Frazier (1980). In this study, the authors try to analyse the contents of firms' environmental disclosures by counting word frequencies. Another early study on financial sentiment analysis using textual analysis is Klibanoff, Lamont, and Wizman (1998). The authors simply ask if there is country-specific news appearing on front page of *The New York Times* and show that weeks with news reports do exhibit higher market reaction, and that investor sentiment is indeed affected by news arrivals. In an interesting study, Huberman and Regev (2001) studies the 'non-event' that a Sunday article on *The New York Times* on cancer drug caused EntreMed's stock price to rise from $12 at the Friday close, to open at $85 and close near $52 on Monday. This price effect appears to be permanent as it closed at above $30 in the following three weeks. It is called non-event because the drug had already been reported in *Nature* and other popular newspapers as far as five months before that time. Therefore, the price effect was driven merely by investor sentiment instead of solid material information. This study vividly shows how 'hard' information is not timely incorporated into asset prices

and how investor sentiment moves the market. The fever even spread to other biotechnology firms and investors were chasing other firms in the same sector.

With the advent of internet, researchers began to explore information from the internet. One pioneer study is Antweiler and Frank (2004). The authors investigate messages on internet bulletin boards on 45 DJIA stocks and find that these messages help predict stock volatility. The authors use a simple Baye's classifier and assumes words are independent of each other and classify messages into buy, sell, and hold signals. This is the coarsest form of so-called 'bag-of-words' based models in natural language processing, where we rely on dictionaries (more precisely, word lists of, for example, positive and negative tones) to assess overall sentiment of a sentence or longer document. By using this method, we essentially ask: how many positive and negative words does each sentence (or document) contains, after accounting for document lengths? While this method appears to be coarse, it is very simple to use and in early days, they achieve high precision and subsequently, high model performance for different tasks, and were favored for a long time.

The most popular generic dictionary designed for finance domain is introduced by Loughran and McDonald (2011). The authors note that 'In a large sample of 10-Ks during 1994 to 2008, almost three-fourths of the words identified as negative by the widely used Harvard Dictionary are words typically not considered negative in financial contexts.' Using 10-K files, the authors examined all words that occur in at least 5% of all documents and designed their own dictionary consisting of 2,707 unique words in six categories (negative, positive, uncertainty, litigious, strong modal, and weak modal). The authors argue that some words unexpectedly apply more to certain sectors than the others, and they raise the overall precision of generic dictionaries. For example, words like 'cancer', 'hospital' relate more to health and medical sector and they actually proxy for industry effect rather than tone. The LM dictionary subsequently became the most popular dictionary in finance research that is still widely used today. As the authors

argue, their dictionary is very large, extensive, and highly relevant, as they only consider words that are used by managers in 10K filings, hence were created with business communication in mind. We note that their dictionary is highly imbalanced: only 354 words are positive and 2,329 words are negative. Many researchers further examine and modify the LM dictionary and devise their own dictionaries to apply to their specific research area. For example, Larcker and Zakolyukina (2012) create their own dictionary to detect managers' deceptive language during earnings conference calls. They created word lists measuring hesitations (hmmm, huh, and umm), extreme negative emotions (idiot, slimy, and disgraceful), and extreme positive emotion (tremendous, smashing, and swell).

7.3 Appendix C. A Comprehensive Literature Review of Sentiment Analysis in NLP and Finance

In this section, we review sentiment and opinion mining literature in computer science and how finance and economics literature adopt such methods in financial sentiment analysis. We also briefly extend the scope beyond sentiment analysis to show that natural language processing is a large and evolving sector in finance and economics literature.

7.3.1 C1. Sentiment, textual analysis, and asset returns

Finding ways to extract information from texts and making use of them has become increasingly important as new techniques in NLP make it possible to extract information from financial texts. While we review some important applications in this section, it is far from exhaustive. A few recent papers give excellent reviews on NLP in finance. See, for example, Loughran and McDonald (2016, 2020), Gentzkow et al (2019).

Return prediction using textual data perhaps attracts the most attention from researchers. Prior to research using NLP methods, considerably many early studies investigate the

relationship between information and asset prices. Many studies find that asset prices should reflect public and private information and demand shocks through rational and irrational trading. See, for example, Daniel et al (1998), De Lon et al (1990), Glosten et al (1985), among others.

In early days, measuring sentiment in finance literature tends to rely on numerical instead of textual data. One of the most widely cited literature in finance and economics for sentiment analysis is Baker and Wurgler (2007), which reviews developments in sentiment analysis in finance and provides a sentiment index. The authors note that early studies, which date back to the 1980s, tend to investigate sentiments' effects on aggregate stock market. However, research at this stage is very pre-mature and they do not try to explicitly state or investigate the role of sentiment. With advances in behavioral finance after De Long, Shleifer, Summers, and Waldmann (1990), researchers typically assume we have two types of investors: rational arbitrageurs who are free from sentiment, and irrational traders who are prone to sentiment trading. Prices deviating from fundamental value can then come from either limits to arbitrage or irrational, sentiment-driven trading. The sentiment measures they propose and survey are the basis of early studies, including: investor survey (such as Brown and Cliff, 2005), investor mood as measured by, for example, cold seasons (Kamstra, Kramer, and Levi, 2003), investor age (such as Greenwood and Nagel, 2009), mutual fund flows, which signal what stocks are favored by mutual funds and hence may proxy investor sentiment (such as Frazzini and Lamont, 2007), trading volume, which proxies which stocks are favored by investors especially with short selling constraints (such as Scheinkman and Xiong, 2003), et cetera. Constructing their sentiment index based on Baker and Wurgler (2006), which considers six indicators, including: trading volume, the dividend premium, the closed-end fund discount, the number and first-day returns on IPOs, and the equity share in new issues, the authors find that 'stocks of low capitalization, younger, unprofitable, high-volatility, non–dividend paying, growth companies or stocks of firms in financial distress' are more sensitive to investor sentiment. To gauge

a single index, the authors performed Principal Component Analysis (PCA) on the six components. One possible explanation for their finding is that such firms are more difficult to arbitrage and are more prone to valuation errors, or at least disagreements in valuation (Miller, 1977).

Researchers have long suspected that information on the internet and the availability of internet itself change investor behavior and have attempted to quantify and investigate how such information affect financial markets. One early study is Choi et al (2002). In this study, the authors try to answer the question: how internet affects investor behavior. The authors note that back then, the internet was considered a negative shock to financial markets and was often blamed to be the cause of excessive trading, excessive herding, higher volatility in the stock market, excessive risk-taking, the Internet "bubble" of the late 1990s, and the bursting of this bubble in 2000. However, much of the blame was pure suspect and there was little evidence backing up these claims. To contribute empirical evidence to this debate, the authors specifically tackle the issue: how allowing online trading affects trading volume and investor performance. The authors first investigate investor characteristics that make one more likely to participate in internet trading instead of traditional phone trading. They find that young, wealthy, male investors are the early adopters. Controlling for trends in stock trading, the authors then find that 18 months after the introduction of internet trading, internet trading nearly doubles traditional phone trading, but trading size is considerably smaller. The authors find no statistically significant evidence suggesting that internet and phone trading have any differences in performance. The authors note, however, while in the sample they consider, namely 401(k) plan, there's no direct transaction costs to investors, the more frequent trading does incur higher transaction costs to the funds level, and this is eventually born by all investors in the 401(k) plan.

Other than Choi et al (2002), some other early studies had also realized the link between small investor behavior and stock market activity (Das and Chen, 2007). For example,

Wysocki (1998) simply uses message counts and finds that variation in daily message posting volume is related to news and earnings announcements. Tumarkin and Whitelaw (2001) investigates messages posted on Raging Bull with self-reported investor sentiment measure about how positive or negative they are about a particular stock. The authors find that on days with abnormally high message activity, changes in investor opinion correlated with abnormal industry-adjusted returns. These event days also coincided with abnormally high trading volume, which persisted for a second day. However, we found that message board activity did not predict industry-adjusted returns or abnormal trading volume, consistent with market efficiency. Tetlock (2007) and Tetlock (2008) find that negative sentiments do predict downward movements in stock prices.

In Das and Chen (2007), the authors develop a method for extracting small investor sentiment from stock message boards. The authors note that internet message boards contain a variety of information including investor sentiment, investor insights, and investors' reactions to other sources of news. The messages posted online are not necessarily information but may also contain rumors and messages intended for market manipulation. As a result, internet message boards attract the attention of investors, corporate management, and regulators. The authors explicitly acknowledge that the way they define 'sentiment' is unavoidably noisy: they define 'sentiment' as net of positive and negative opinions, but it would they include sentiment, information, and measurement errors. They used techniques that were available back then in classifying messages into bearish, bullish, and neutral sentiments, such as support vector machines and Naive Bayes classifiers, which have accuracy of only 50%, close to a pure guess. They focused on tech stocks as they are actively discussed in message boards. They find that the aggregated sentiment tracks the index returns while such effect is quite weak for individual stocks.

Early studies including Tetlock (2007) use general dictionaries from psychology literature, but they are unable to capture semantic meaning in finance (or 'domain-specific language' in computer science literature). Consider a simple example: the word 'liability' tends to carry negative sentiment in generic language, but it is merely used to describe a company's financial position when discussing a company's assets and liabilities. Therefore, while early studies shed light on textual analysis in finance, their results and model performances tend to be poor. Also, generic dictionaries tend to correlate more with negative sentiment in finance, hence early studies tend to use negative words only (Loughran and McDonald, 2011; Tetlock, 2007).

Not all research on sentiment analysis in finance finds predictive power of sentiments in relation to stock market movements, though. For example, Kim and Kim (2014) investigates Yahoo! Finance messages boards' predictive power on stock return, volatility, and trading volume. The authors did not find online messages board's predictive power on stock market. Instead, they find that stocks' past performance predicts message board messages' sentiment contents.

While traditional research in information economics also studies return prediction, sentiment and information theory differ drastically because price impact from information is permanent while sentiment effects are transitory. As noted by Tetlock (2007), 'The sentiment theory predicts short-horizon returns will be reversed in the long run, whereas the information theory predicts they will persist indefinitely'. Empirical research in sentiment analysis tends to find reversals in return following the initial reaction. For example, Hillert et al (2014) uses 2.2 million newspaper articles from 45 US newspapers and find that stocks with higher media coverage exhibit higher momentum and the momentum reverses in the long run. The momentum continues for up to 12 months after forming portfolio based on media coverage, where high coverage portfolio outperforms low coverage portfolio by about 40 basis points per month but

drops thereafter. In 2 and 3 years after portfolio formation, high coverage portfolio underperforms low coverage portfolio by, on average, 22 basis points per month.

As researchers in economics and finance started to borrow from NLP methodologies, we see a growing literature using text and media as data sources to measure sentiment. An early study in this strand is Antweiler and Frank (2004). In this paper, the authors start with media's view that online forums move the market and systematically investigate if this is true. To do so, the authors collected 1.5 million messages from Yahoo! Finance and Raging Bull, which were the most popular online forums back then. As online messages are not labeled, the authors manually labeled 1,000 messages as 'Buy', 'Hold', or 'Sell' signals. The authors then implemented two simple machine learning algorithms: Naive Bayes Classifier, and Support Vector Machine to train their data. Their sample is highly unbalanced: of the 1,000 messages, 69.3% are hold, 25.2% are buy, and only 5.5% are sell signals. A feature of intraday trading in their sample is that trading around market open and close, especially for the first and last half hour, is significantly more than other periods. They argue that this is because small traders and investors think about trading strategy after work and there are news arrivals after market close, hence small investors place considerable number of orders for market open. Mutual funds and day traders close their positions near market close, driving up market close trading activities. This feature is still present today. The authors then used OLS, realized volatility models, and GARCH to study the relationship between online forum posting, sentiment, and stock market behavior. The authors claim that they are the first to report a negative relationship between forum posting and next-day return, although this is economically negligible especially after accounting for transaction costs. Consistent with intuition, the authors find that disagreements in message sentiment causes more trading and higher volatility.

Research in this field typically uses short window of a few days after the relevant news release. Shiller (2000) argues that investors follow the printed word even though much

of it is pure hype, suggesting that market sentiment is driven by news' content. Following Shiller (2000)'s argument and the formal evidence from Tetlock (2007) that negative words in Wall Street Journal predicts daily stock returns, Garcia (2013) revisits the issue by studying financial news from New York Times from 1905 to 2005. He shows that in hard times as proxied by recession, investors are more sensitive to news. Specifically, during recessions (expansions), a one standard deviation change in sentiment measure predicts 12 (3.5) basis points change in daily average of DJIA, where sentiment measure is the classic bag-of-word approach based on the number of positive and negative words in financial columns of New York Times. He also finds that both positive and negative words help predict returns while previously, Tetlock (2007) shows that only negative words have predictive power.

Ke et al (2019) is a recent study in textual sentiment analysis. In this paper, the authors use data from Dow Jones Newswire, which contains all historical news for the US companies from 1986 to April 2020. The authors start from the view that news simultaneously affect investor sentiment and market return and propose a three-step framework: 1) as positive sentiment drives up return and negative sentiment drives down return, we can use market reaction as a guide to automatically create dictionary of positive and negative sentiment; 2) use a two-topic model to estimate positive and negative sentiment scores; 3) predict sentiment scores of news articles. With the predicted sentiment score, the authors then go long the 50 most positive stocks and go short the 50 most negative sentiment stocks each day to forma a zero-cost portfolio. The authors achieve an annualised Sharpe ratio of 4.3 overall.

Hoberg and Phillips (2021) specifically deals with the issue of industry momentum using textual data. Drawing from previous literature, the authors suggest that industry momentum likely stems from investor inattention.

A handful of research use readily available sentiment data from commercial vendors. Groß-Klußmann and Hautsch (2011) empirically examines high-frequency market

reactions to an intraday stock-specific news flow. The authors wish to analyze to which extent high-frequency movements in returns, volatility and liquidity can be explained by the underlying mostly nonscheduled news arrivals during a day. To do so, they rely on Reuters NewsScope Sentiment Engine, which is a black-box engine that automatically analyses news when they arrive and produces a sentiment label (positive, neutral, and negative), novelty, and relevance indicator. The authors use 29,497 news headlines for 40 stocks from January 2007 to June 2008. The authors find that high frequency trading does react significantly to relevant intraday company-specific news arrivals, as expected; among other measures, volatility and cumulative trading volume are the most significant responders.

In another study, Uhl (2014) uses a much larger dataset of sentiment data from 3.6 million Reuters news articles from January 2003 to December 2010. The dataset again uses the black-box sentiment data from Reuters. The author acknowledges that there are two issues central to their study: 1) prior studies had no consensus on sentiment measure, hence one needs to carefully choose the sentiment measure; 2) timeframe also greatly affect stock prices post news arrivals. To tackle the first issue, the author chooses to use Reuters sentiment which gives sentiment score of positive, neutral, or negative for each news piece. To tackle the second issue, the author first note that prior studies either look at investor sentiment, such as Tetlock (2007, 2011) and Tetlock et al (2008), or investor sentiment, such as Brown and Cliff (2005). Studies on news sentiment typically consider sentiment effects at short intervals up to a few days while studies on investor sentiment typically consider longer timeframe of monthly sentiment effects. Other asset classes may see ‘sentiment effects’ lasting a few years. For example, Menkhoff and Rebitzky (2008) finds that sentiment effects in the foreign exchange market may last up to two years. The author then chooses to form monthly sentiment index using all the news sentiment available. The author then uses VAR model to assess the dynamics of sentiment and finds that positive and negative Reuters sentiments do affect stock returns, although negative sentiment’s effects are larger; fundamental

factors, such as the Conference Board Leading Economic Indicator, do not have a measurable effect on stock returns and the author proposes that this is because market participants can quickly incorporate fundamental information into asset prices hence they are not significant in analysis spanning months.

Sentiments in financial market are also interesting to policy makers and market participants rely on a range of hard (such as unemployment rate, price index, et cetera) and soft variables (such as survey-based methods) in forecasting future economic conditions. For the latter, consumer sentiment by the University of Michigan and the Conference Board appears to be the most popular used by practitioners and policy makers (Shapiro, Sudhof, and Wilson, 2020). A recent sentiment analysis study in finance and economics literature using 'soft' variables is Shapiro et al (2020). In this paper, the authors experimented with different sentiment analysis tools and propose their own sentiment score measure to extract sentiment information from news. The authors purchased 238,685 economic and financial news from LexisNexis from 16 major newspapers (Atlanta Journal-Constitution, Boston Globe, Chicago Tribune, Detroit Free Press, Houston Chronicle, Los Angeles Times, Memphis Commercial Appeal, Miami Herald, Minneapolis Star Tribune, New Orleans Times-Picayune, New York Times, Philadelphia Inquirer, San Francisco Chronicle, Seattle Times, St. Louis Post-Dispatch, and The Washington Post) from January 1980 to April 2015. The authors purchased only news with sufficient contents (news not labeled as 'brief', 'summary', or 'digest') and are long enough (longer than 200 words) to exclude articles that appear elsewhere and reduce noise, because 'very short articles are likely to be more noisy'. As an additional step, they only include articles that include 'said, says, told, stated, wrote, reported', because they consider such articles reporting opinion of someone or some group of people, hence carrying strong sentiment signal. As news articles are by construction unlabeled, the authors asked 15 research assistants at the Federal Reserve Bank of San Francisco to hand-label 800 news articles into: Very Negative (1), Negative (2), Neutral (3), Positive (4), and Very Positive (5), and the 800

labeled data forms the basis of their training sample. We note that their labeled sample is very small compared with their sample size, accounting for only 0.335% of their total sample. The authors very explicitly distinguishes sentiment and information contents. They state that:

"By sentiment, we mean the tone/feeling/emotion expressed of the article rather than the economic substance of the article. For example, If the writer is talking about a report of very high GDP growth but is expressing concern that this reflects overheating of the economy and monetary policy being behind the curve, then this could be the writer expressing negative sentiment even though he/she was talking about high growth."

The authors compare the performances of various lexicons with advanced ML models. To our surprise, the authors find that BERT models perform almost as good as lexicon-based models which combines LM and HL lexicon in predicting news sentiment, and that LM+HL lexicon performs almost as good as VADAR. However, their findings are not too surprising. As is with other deep learning models (Marcus 2018), BERT requires a large training sample due to the large (usually several millions) number of parameters to learn. The authors' very small training set is hence unable to allow BERT to work to its full capacity. As is with other lexicon-based sentiment analysis, the overall sentiment for a text is simply the difference between its proportions of positive and negative words. Because of the theoretical advantage of transformers and the fact that BERT's superiority in sentiment classification has been confirmed in many other studies, we believe the lower performance of transformers is largely because the BERT model the authors are using are not designed specifically for finance domain, and their very small training sample is especially problematic for deep learning models including BERT.

There are many other applications of NLP in finance but they are not the focus of this study: Manela and Moreira (2017) constructs a news-implied volatility measure from Wall Street Journal front page and their findings are consistent with recent theoretical

advances suggesting disaster risk is an important source of volatility; Jeon et al (2021) studies stock price jumps and they find that news frequency, tone, and uncertainty and they find that news flows can explain jump intensity and jump-size distributions and explain an important fraction of variations in the jumps across individual companies; Huang et al (2020) finds that institution trading on stocks tend to concentrate on the first release of a series of news and such trading predicts returns over weeks and suggests that institutional investors facilitates price efficiency by quickly interpreting public information and incorporating public information into asset prices; Engle et al (2020) studies climate news and shows how to use a synthetic portfolio to hedge climate risk.

Apart from empirical research, some studies also attempt to give theoretical grounds to news trading. For example, Foucault et al (2016) is a theoretical paper on news trading. The authors constructs a model where the speculator's private signals can be used to forecast both short-run price reactions to news arrival and long-term price changes.

A handful of research also investigates volatility and information. In an early study, French and Roll (1986) examines information arrival during market open and market close. Based on the notion that volatility may come from private information, which is revealed through trading during trading hours, or public information, which may arrive either during trading hour or non-trading hour, or irrational trading, the authors find that return volatility is mainly from rational trading driven by private information. Their conclusion appears to be supported by later research. See, for example, Ito et al, 1998; Chordia et al (2011). However, their findings are subject to different interpretations. For example, Hong and Stein (2003) notes that ‘Roll (1984, 1988) and French and Roll (1986) demonstrate in various ways that it is hard to explain asset price movements with tangible public information’.

In a recent study, Boudoukh et al (2019) revisits this issue using textual data. The authors wish to investigate the saliency of news instead extract sentiment contents from news and they wish to match companies with events such as such as new product launches, lawsuits, analyst coverage, news on financial results, and mergers. Because of computation limitations, the authors only consider S&P 500 companies that have at least 20 trading days in the sampling period. To do so, they use two methods: 1) visual information extraction platform (VIP), which uses a mixture of a rule-based information extraction platform and a trained support vector machine classifier, to identify event instances for companies and measure sentiment from text contained in financial news. The authors apply VIP to Dow Jones Newswire from January 1, 2000 to December 31, 2015. 2) A commercial product, RavenPack. When news articles is released, RavenPack would automatically process the news, with the algorithm being a black box, and produce 16 fields including timestamp, company identifier, relevance score, et cetera. RavenPack recommends a relevance score of at least 90 and such news are typically highly relevant of company events. The authors are then able to assign variance in stock prices that are due to arrival of firm-specific events, or 'fundamental information in news'. The authors find that 49.6% (12.4%) overnight (trading hour) idiosyncratic volatility is due to fundamental information related to company events. The authors also document a large negative correlation over time (i.e., −0.50) between average idiosyncratic volatility during overnight hours and the contribution of identified news to overnight return volatility, arguing that the benefit to private information production has decreased due to the increase in publicly available information.

Market participants and policy makers rely on very different indexes of investor and consumer sentiment. While labor-intensive and very time-consuming and expensive to construct, survey-based methods, notably, Michigan Consumer Sentiment index and the Conference Board's Consumer Confidence index, are widely used in economics. The most popular method in finance and economic literature to extract sentiment from

texts is lexicon-based method. This method relies on pre-defined ‘dictionaries’, which are ‘sentiment-charged’ lists of words, and by giving each word a score (such as 1 for positive, 0 for neutral, and -1 for negative), it effectively asks: in each corpus, how many words belong to each dictionary? In early studies, generic dictionaries from psychology and sociology literature are used, such as Bollen et al (2011). However, researchers quickly realized that context, or domain-specific dictionaries are necessary for accurate sentiment classification, and tailoring dictionaries to specific needs of each research greatly improves performance, because words have very different meanings in different contexts. For example, the word ‘liability’ does not always carry negative sentiment in financial news, but in general texts, they are very negative words (Loughran and McDonald, 2011). Other words that appear frequently in finance documents but do not necessarily carry negative meaning include tax, cost, capital, board, and depreciation. Dictionaries may contain only single words (‘unigrams’) or phrases containing several words (n-grams), although unigrams are much more prevalent than n-grams. This method belongs to ‘bag-of-word’ representation of texts, which ignores the context, inter-relatedness of word, sentence structure, et cetera, and considers each word in isolation. One popular generic dictionary designed for finance research is Loughran and McDonald (2011). While this dictionary is 10 years old, it is still widely used today. See, for example, Shapiro and Wilson (2021), which uses meeting transcripts of the Federal Open Market Committee to study the committee’s loss function. Common dictionaries include:

1. Harvard General Inquirer (GI) Dictionary. This dictionary is mostly seen in early studies where dictionaries catered specifically for finance and economic research was not available, although some recent studies do use this dictionary (such as Heston and Sinha, 2017). This is also used as the basis of Loughran and McDonald (LM) dictionary. It consists of 3,626 words labeled ‘positive’ and ‘negative’.

2. Loughran and McDonald (LM) (2011) dictionary, which was subsequently updated in 2014 and 2020. It consists of 2,707 words in 2014 version. The words are labeled 'positive' and 'negative' and are sourced from 10-K files of publicly traded companies.

3. Hu and Liu (HL) (2014) dictionary. This dictionary is very popular (with over 8,500 citations) in general sentiment analysis literature but has limited use in finance and economics literature, because while it is very large (with 6,786 words), it is sourced from online movie reviews where reviewer themselves label their review as 'positive' and 'negative'.

The dictionaries can be very different from each other. Shapiro et al (2020) directly compares GI, LM, and HL dictionaries. They find that 58% of LM dictionary words are not found in the other dictionaries, and only 31% of GI words are not found in the other two dictionaries. For the common words that appear in two dictionaries, the different dictionaries tend to agree on their sentiment contents: the disagreement rate of HL-LM dictionaries is only 0.9%, with HL-GI and LM-GI being 1.4% and 2.7% respectively. While LM dictionary is tailored to finance and economic literature, its small size means very often, news may use words that are not found in it (LM, GI, and HL dictionaries only classify 2.8, 6.4, and 4.4% words in their sample corpus). The authors hence suggest combining the different lexicons.

We are interested in studying sentiment in financial markets because they ultimately move the market. For example, Baker and Wurgler (2007) sentiment index is successful in explaining cross-section of stock returns and is one of the most popular sentiment index by far. A few studies investigate stock behavior in high and low sentiment periods where the classification is given by Baker and Wurgler (2007) index. For example, Stambaugh, Yu, and Yuan (2012) investigates how investor sentiment explains stock market anomalies. The authors find that high sentiment follows high anomalies, and that the short leg is the profitable one. Yu and Yuan (2011) show that the positive relationship between market return and volatility is gone with high sentiment period.

More recently, Jiang, Wu, and Zhou (2018) show that many anomalies are sensitive to investor sentiment. Pástor, Stambaugh, and Taylor (2017) show that funds trade more when sentiment is high, and that there is a positive relation between fund turnover and return.

It is also interesting to consider the types of texts inputs used in financial studies. The exact types of financial documents to use in financial sentiment analysis differ in different studies and novel dataset in recent years greatly facilitates sentiment extraction. Broadly, sentiment may be extracted from corporate reports (see, for example, Li 2006), social media such as twitter (such as Bollen, Mao, and Zeng, 2011), partial texts in financial media (such as Tetlock 2007), and in more recent years, full texts of historical financial news (such as Ke, Kelly, and Xiu, 2019). Because of data availability, social media such as Twitter and StockTwits, which could be considered Twitter for investors, were the main source of data in early studies, as such data tends to be freely available and covers sufficient large number of stocks. Bollen et al (2011) is the most widely cited study to use Twitter to predict stock market movements. The authors note that traditional asset pricing theories and studies under the efficient market hypothesis (Fama, 1996; Fama, Fisher, Jensen, and Roll, 1969) imply that stock market prices should follow random walks and should be unpredictable because new information is unpredictable, but multiple strands of literature, including socioeconomic theory of finance and behavioral finance, have been challenging this view of stock market. See, for example, Prechter and Parker (2007), Smith (2003), and Nofsinger (2005). Based on research from psychology (such as Dolan, 2002) and behavioral finance (such as prospect theory of Kahneman and Tversky, 1979) that emotions do affect behavior and decision-making, they authors try to extract public sentiment from a collection of tweets from February 2008 to December 2008. To do so, they used two sets of tools: OpinionFinder, which classifies daily moods into positive and negative sentiment, and GPOMS, which classifies daily moods into six categories: Calm, Alert, Sure, Vital, Kind, and Happy. Both classifiers are based on lexicons and potentially asks how many

words in the tweets belong to each category's dictionary. The dictionaries are pre-defined from psychology research. The authors find that sentiments do predict Dow Jones Industrial Average in coming days.

Recent studies also attempt to use advances in machine learning beyond textual information. For example, Obaid and Pukthuanthong (2021) constructs sentiment index from news photos and they find that their Photo Pessimism index predicts market return reversals and volume.

7.3.2 C2 Developments in Sentiment Analysis in Computer Science

Apart from finance and economics literature, a large and growing number of machine learning literature attempt to measure and classify financial sentiment analysis. Machine learning methods are especially suitable for this task because financial sentiment analysis at is core is a classification problem. In early studies of this strand of literature, researchers typically use simple supervised classifiers to assess sentiment contents of texts and are typically reliant more on bag-of-words approach with significant effort in feature engineering. See, for example, Turney and Pantel (2010). Ghiassi, Skinner, and Zimbra (2013) introduces a Twitter-specific lexicon and uses Dynamic Architecture for Artificial Neural Networks (DAN2) which outperforms a simple Support Vector Machine (SVM). Wang et al (2015) experiments with common classifiers in classifying financial sentiment using StockTwits sample into 'bearish' and 'bullish' and finds that SVM outperforms Naive Bayes and Decision Trees with an accuracy of 76.2%. Later, deep learning methods became popular because of their superior classification performance.

While early studies in sentiment and opinion mining in computer science literature also relies on simple lexicon-based models, they were quickly replaced by newer models, even if one tries to improve model performance by introducing rule-based models in addition to simple word counts. For example, while 'good' carries positive sentiment, 'not good' negates the positive meaning of the word 'good'.

However, natural language is too complicated to fit in a rule-based model, unless the rule is endless. More recently, researchers using lexicon-based methods attempt to account for contexts in addition to words, such as VADER of Hutto and Gilbert (2014). VADER is a simple rule-based sentence-level classifier comprising of: 1) a large dictionary, which assigns each word a score from -4 to 4 for most negative to most positive; and 2) a set of heuristic rules to determine each word's context in the sentence. The heuristic rules are very simple and accounts for common sentence structures, such as 'but' reverses sentence's meaning, words like 'very', 'a bit' modifies intenseness of a sentence's sentiment, et cetera.

A breakthrough *word representation* method is Word2Vec of Mikolov et al (2013a,b). This model has two versions: Continuous Bag-of-Words, which predicts central word based on surrounding contextual words, and Skip-Gram, where central words are used to predict surrounding, contextual words. The authors experimented the performance of Word2Vec on a range of NLP tasks including sentiment analysis. However, it has two main drawbacks: 1) it is incapable of handling words that are not seen in training sample; 2) words semantically similar would still be given two completely different encodings. For studies using Word2Vec on sentiment analysis, see, for example, Zhang et al (2015).

One important development after *bag-of-words* is Global Vectors for Word Representation (GloVe) of Pennington, Socher, and Manning (2014) from Stanford University, which is one way of word embeddings. Generally, word embedding tries to represent each word in a high-dimensional space, and in the process, words that semantically similar are close to each other. Unsurprisingly, generally the higher dimension, the better semantic meaning we can capture, but the more expensive training we face. Word embeddings are also the basis of encoder-decoder frameworks including BERT. This strand of method is also called distributional semantic model, (DSM), vector space model of meaning, and semantic space model of meaning. In GloVe, each

word is still represented by a dense vector that incorporates its semantic characteristic, hence words of similar meaning are close to each other in the high-dimensional vector space, such as ‘very’ and ‘extremely’. Lexicon-based models typically account for document length and words that appear only in certain documents through ‘Term Frequency - Inverse Document Frequency’ (TF-IDF). For research using GloVe on sentiment analysis, see, for example, Rezaeinia, Rahmani, Ghodsi, and Veisi (2019).

One breakthrough in *word embeddings* is a pre-trained model called Embeddings from Language Model (ELMo) of Peter et al (2018). Previously, word embeddings would produce the same static regardless of different texts. In ELMo, the same word would be given different embeddings depending on the context. This greatly improves its performance because now we can capture the same word’s different meanings.

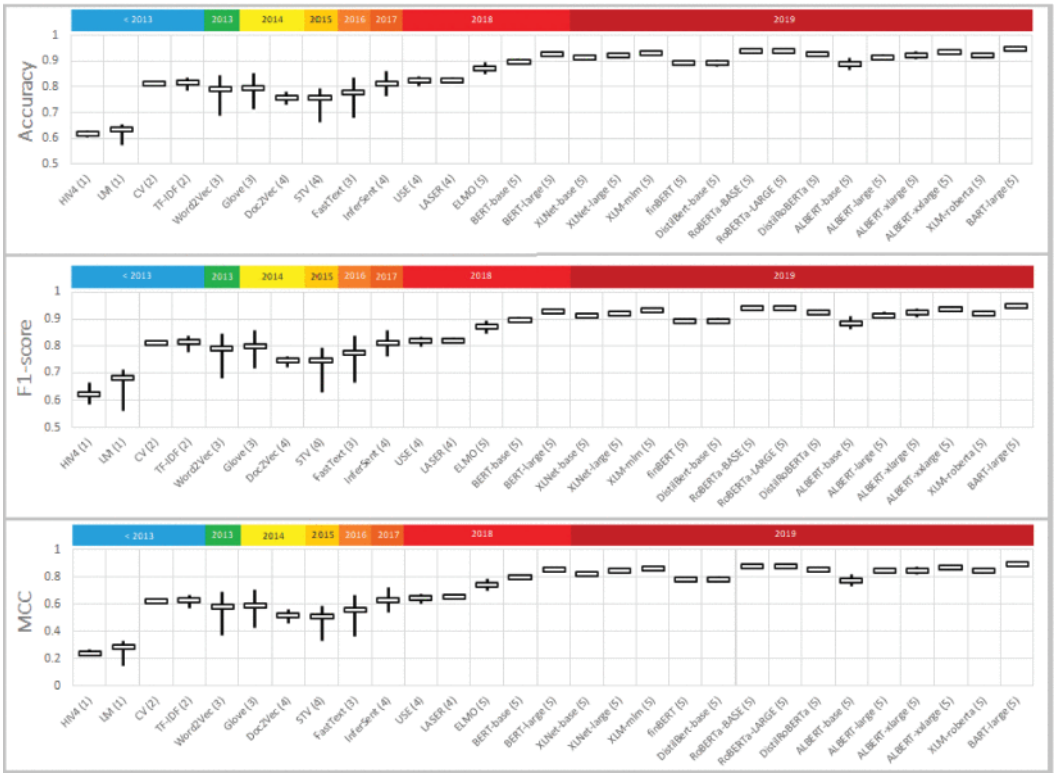
Recent developments in machine learning, especially deep learning methods, attempt to let the algorithm learn the context, and they typically achieve much better performance than bag-of-words and rule-based models. This is not surprising: each word has a meaning, but the meaning is deeply affected by all words in front of and after the word. By ignoring the context, it is not surprising that traditional language models perform poorly in understanding a corpus, and hence poor performance of models relying on the language model. In plain vanilla deep learning models, each node, or hidden state, takes input as the output of its previous state and performs a non-linear transformation. The exact non-linear function may vary, such as ReLU, sigmoid, and Adam.

Textual analysis in finance is not limited to English. In recent years, as China attracts attention from almost all fields in economics and finance, a growing number of literature studies textual analysis in the Chinese financial market. A recent work in progress is Fan, Xue, and Zhou (2021). Deviating from previous studies, which typically use topic models or dictionaries to reduce dimensionality and ignores semantic sequences and structures, the authors propose a method to account for the

whole corpus. The authors propose a 'Factor-Augmented Regularized Model for Prediction (FarmPredict) on stock returns by extracting the hidden topics (factors) from all words with consideration of structure and interactions of phrases or words'. The authors use a three-step framework: 1) use PCA to convert articles into vectors of hidden features consisting of multiple factors and idiosyncratic components; 2) screen the idiosyncratic variables by their correlations with beta-adjusted returns; 3) use a simple LASSO model to predict return using hidden factors and the screened idiosyncratic components. The authors use data from Sina Finance and predict returns from 2015 to 2019. The authors then long the 50 most positive sentiment (return) companies and short the 50 most negative companies with daily rebalance. They find that for the Chinese market, positive leg's returns is much more important than negative leg, likely due to China's short-sale constraints; and that 7 days before news publish, returns already started to react, likely due to information leakage. We believe part of the pre-news reaction is brought about by the feature of their dataset: Sino Finance does not always provide timely information on the latest news, a large number of news articles are in-depth analysis of events already occurred some time ago, hence market reaction was already in place. It is, however, interesting to note that market reaction can last up to 7 days before Sino News, a major public, online, and freely available news provider in China, publishes news articles. Still, return prediction is overwhelming: FarmPredict achieves a Sharpe ratio of a stunning 9.37, higher than all previous models.

More recently, Bidirectional Encoder Representations from Transformers (BERT) (Devlin, Chang, Lee, and Toutanova, 2018), based on transformer architecture of self-attention mechanism of Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, and Polosukhin (2017), outperforms all previous models in almost all NLP tasks including sentiment analysis. Transformer is an encoder-decoder architecture. Models based on transformer may use encoder only, decoder only, or both encoder and decoder.

To our knowledge, the latest and most comprehensive comparison of transformer-based models against lexicon-based models in financial sentiment analysis is Mishev, Gjorgjevikj, Vodenska, Chitkushev, and Trajanov (2020). In this paper, the author acknowledges that the problem of unavailability of large, labeled dataset and the lack of domain-specific model for financial sentiment analysis, and uses publicly available labeled dataset to assess the performance of lexicon and transformers-based models. Their results are opposite to Sharpiro et al (2020) and show that transformer-based models significantly outperform lexicon models. We reproduce their Figure 9 here:



Clearly, Transformer-based models significantly outperform previous models including all lexicon-based models across accuracy, F1 score, and MCC: the authors experimented with different versions of lexicons and introduced machine learning techniques in addition to plain vanilla lexicon model, such as TF-IDF, and cross validation. Lexicon-based models are around 60%-80% accurate with F1 score also around 60%-80%, while different Transformer-based models achieve accuracy and F1 scores both over 90%. The best-performing model is BART-large, which achieves

accuracy and F1 of 94.7%. The model of choice in this paper is ALBERT-xlarge, the second-best performing model with an accuracy of 93.6% and F1 scoe of 93.5%, While BART performs slightly better, it is much more resource-consuming with 406 million parameters. ALBERT-xlarge is designed to be light on resources with only 58 millions parameters in its xlarge-V2 version. As our dataset contains all historical news of US companies from 1996, model performance would be an issue and the next-best model which is cheap to train is critical for our study.

Before BERT, the best-performing sentiment classification model relies heavily on convolutional neural networks (CNN) and LSTM, which introduces a forget gate at each hidden state to teach the model how much previous information to keep and to drop. See, for example, Wang, Huang, Zhu, and Zhao (2016). GloVe provides pretrained word embeddings for each word and largely ignores context. As context does affect each word's meaning, especially for domain-specific tasks, BERT-based models, which directly incorporates context, is expected to outperform almost all previous models. However, as with other deep learning models, we need a large training set.

Traditionally, language models are trained to predict the next word in a given text and more recently, researchers in NLP apply language models to general downstream tasks including sentiment classification. Normally, such models are pre-trained on a large sample and then fine-tuned using domain-specific texts (Kant et al, 2018). BERT is one such model that could be fine-tuned using finance texts. As is with other NLP tasks (see, for example, McCann et al 2017), pre-training models specifically on financial text greatly improves performance. A version of BERT specifically trained for finance domain is FinBERT of Araci (2019). The author uses an open-source dataset from Reuters TRC2 comprising of 1,800,370 news stories covering the period from 2008-01-01 to 2009-02-28 or 2,871,075,221 bytes. The author only uses a subset of TRC which consists of '46,143 documents with more than 29M words and nearly 400K sentences' after filtering for 'some financial keywords' to limit sample size for training

and make their sample more relevant. This dataset is unlabeled. They also use a labeled dataset of Financial PhraseBank from Malo et al (2014), where the authors ask 16 annotators with adequate business education to hand-label around 4,845 phrases and sentences sampled from financial news texts sourced from LexisNexis into ‘positive’, ‘neutral’, and ‘negative’. The annotators were asked to label based on how they think the sentence would affect the company’s stock price. The authors use Financial PhraseBank to run their main set of analysis and set 60% as training set, 20% as validation, and 20% as test set. As is with routine procedures, re-training model on domain-specific texts greatly improves the model’s performance (see, for example, Howard and Ruder, 2018). The author then trains his model by adding a dense layer on top of the usual BERT model, as is the suggested method for arbitrary downstream NLP tasks by the original author of BERT (Devlin et al 2018). The author achieves an accuracy of 86%, cross entropy loss of 37%, and F1 score of 84%. While their performance is not too eye-catching, this is probably because there are indeed different ways to interpret the same piece of news. For the subset of Financial PhraseBank with agreement rate of 100% (about 70% of total sample), BERT retrained on financial texts achieves an accuracy of 96%. This represents another advantage of using market reaction to label news sentiment: market reaction represents the average opinion of all market participants; while we cannot observe how many investors think a news is bullish and how many think it is bearish, we can infer the average opinion from market prices, and this is what is important for our purpose: formulating a long-short portfolio consisting of stocks that are most likely and least likely to rise, instead of precisely working out the public’s opinion. In terms of fine-tuning, their best results are achieved with slanted triangular learning rate, gradual unfreezing, and discriminative fine-tuning.

8 References

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

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.

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

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. Journal of economic perspectives, 21(2), 129-152.

Barberis, N., Shleifer, A., & Vishny, R. W. (2005). A model of investor sentiment (pp. 423-459). Princeton University Press.

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

Brown, G. W., & Cliff, M. T. (2004). Investor sentiment and the near-term stock market. Journal of empirical finance, 11(1), 1-27.

Brown, G. W., & Cliff, M. T. (2005). Investor sentiment and asset valuation. The Journal of Business, 78(2), 405-440.

Boudoukh, J., Feldman, R., Kogan, S., & Richardson, M. (2019). Information, trading, and volatility: Evidence from firm-specific news. The Review of Financial Studies, 32(3), 992-1033.

Charoenrook, A. (2005). Does sentiment matter. Unpublished working paper. Vanderbilt University.

Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2013). Customers as advisors: The role of social media in financial markets. Working paper.

Chordia, T., Roll, R., & Subrahmanyam, A. (2011). Recent trends in trading activity and market quality. Journal of Financial Economics, 101(2), 243-263.

Choi, J. J., Laibson, D., & Metrick, A. (2002). How does the Internet affect trading? Evidence from investor behavior in 401 (k) plans. Journal of Financial economics, 64(3), 397-421.

Cowles 3rd, A. (1933). Can stock market forecasters forecast?. Econometrica: Journal of the Econometric Society, 309-324.

Cutler, D. M., J. M. Poterba and L. H. Summers. “What Moves Stock Prices?” Journal of Portfolio Management, 15, (1989), pp. 4–12.

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. The journal of finance, 66(5), 1461-1499.

Daniel Kent, D., Hirshleifer, D., & Subrah-manyam, A. (1998). Investor psychology and security market under-and overreactions. Journal of Finance, 53(5), 1839-1885.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. Management science, 53(9), 1375-1388.

De Bondt, W. F., & Thaler, R. (1985). Does the stock market overreact?. The Journal of finance, 40(3), 793-805.

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. Journal of political Economy, 98(4), 703-738.

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

Dolan, R. J. (2002). Emotion, cognition, and behavior. Science, 298(5596), 1191-1194.

Engle, R. F., Giglio, S., Kelly, B., Lee, H., & Stroebel, J. (2020). Hedging climate change news. The Review of Financial Studies, 33(3), 1184-1216.

Fama, E. F. (1965). The behavior of stock-market prices. The journal of Business, 38(1), 34-105.

Fama, E. F., Fisher, L., Jensen, M., & Roll, R. (1969). The adjustment of stock prices to new information. International economic review, 10(1).

Fama, E. F. (1970). Efficient capital markets a review of theory and empirical work. The Fama Portfolio, 76-121.

Fan, J., Xue, L., & Zhou, Y. (2021). How Much Can Machines Learn Finance From Chinese Text Data?. Available at SSRN.

Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. Review of Accounting Studies, 15(4), 915-953.

Foucault, T., Hombert, J., & Roşu, I. (2016). News trading and speed. The Journal of Finance, 71(1), 335-382.

Frazzini, A., & Lamont, O. A. (2007). The earnings announcement premium and trading volume. NBER working paper, (w13090).

French, K. R., & Roll, R. (1986). Stock return variances: The arrival of information and the reaction of traders. Journal of financial economics, 17(1), 5-26.

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

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3), 535-74.

Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with applications, 40(16), 6266-6282.

Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. Journal of financial economics, 14(1), 71-100.

Groß-Klußmann, A., & Hautsch, N. (2011). When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. Journal of Empirical Finance, 18(2), 321-340.

Greenwood, R., & Nagel, S. (2009). Inexperienced investors and bubbles. Journal of Financial Economics, 93(2), 239-258.

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

Hillert, A., Jacobs, H., & Müller, S. (2014). Media makes momentum. The Review of Financial Studies, 27(12), 3467-3501.

Hoberg, G., & Phillips, G. M. (2018). Text-based industry momentum. Journal of Financial and Quantitative Analysis, 53(6), 2355-2388.

Hong, H., & Stein, J. C. (2003). Differences of opinion, short-sales constraints, and market crashes. The Review of Financial Studies, 16(2), 487-525.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 8, No. 1).

Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177).

Huberman, G., & Regev, T. (2001). Contagious speculation and a cure for cancer: A nonevent that made stock prices soar. The Journal of Finance, 56(1), 387-396.

Huang, A., Wang, H., & Yang, Y. (2020). FinBERT—A Deep Learning Approach to Extracting Textual Information. Available at SSRN 3910214.

Huang, A. G., Tan, H., & Wermers, R. (2020). Institutional trading around corporate news: Evidence from textual analysis. The Review of Financial Studies, 33(10), 4627-4675.

Ingram, R. W., & Frazier, K. B. (1980). Environmental performance and corporate disclosure. *Journal of accounting research*, 614-622.

Ito, T., Lyons, R. K., & Melvin, M. T. (1998). Is there private information in the FX market? The Tokyo experiment. The Journal of Finance, 53(3), 1111-1130.

Jeon, Y., McCurdy, T. H., & Zhao, X. (2021). News as sources of jumps in stock returns: Evidence from 21 million news articles for 9000 companies. Available at SSRN 3318517.

Jiang, L., Wu, K., & Zhou, G. (2018). Asymmetry in stock comovements: An entropy approach. Journal of Financial and Quantitative Analysis, 53(4), 1479-1507.

KAI-INEMAN, D. A. N. I. E. L., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 363-391.

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

Kamstra, M. J., Kramer, L. A., & Levi, M. D. (2003). Winter blues: A SAD stock market cycle. American Economic Review, 93(1), 324-343.

Kant, N., Puri, R., Yakovenko, N., & Catanzaro, B. (2018). Practical text classification with large pre-trained language models. arXiv preprint arXiv:1812.01207.

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

Keynes, J. M. (1936). The general theory of employment, interest, and money. Springer.

Kim, S. H., & Kim, D. (2014). Investor sentiment from internet message postings and the predictability of stock returns. Journal of Economic Behavior & Organization, 107, 708-729.

Klibanoff, P., Lamont, O., & Wizman, T. A. (1998). Investor reaction to salient news in closed-end country funds. The Journal of Finance, 53(2), 673-699.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

Larcker, D. F., & Zakolyukina, A. A. (2012). Detecting deceptive discussions in conference calls. Journal of Accounting Research, 50(2), 495-540.

Lee, W. Y., Jiang, C. X., & Indro, D. C. (2002). Stock market volatility, excess returns, and the role of investor sentiment. Journal of banking & Finance, 26(12), 2277-2299.

Lemmon, M., & Portniaguina, E. (2006). Consumer confidence and asset prices: Some empirical evidence. The Review of Financial Studies, 19(4), 1499-1529.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

Li, F. (2006). Do stock market investors understand the risk sentiment of corporate annual reports?. Available at SSRN 898181.

Li, X. (2021). Does Chinese investor sentiment predict Asia-pacific stock markets? Evidence from a nonparametric causality-in-quantiles test. Finance Research Letters, 38, 101395.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

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

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

Loughran, T., & McDonald, B. (2020). Textual analysis in finance. Annual Review of Financial Economics, 12, 357-375.

Maia, M., Freitas, A., & Handschuh, S. (2018, January). Finsslx: A sentiment analysis model for the financial domain using text simplification. In 2018 IEEE 12th International Conference on Semantic Computing (ICSC) (pp. 318-319). IEEE.

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

Manela, A., & Moreira, A. (2017). News implied volatility and disaster concerns. Journal of Financial Economics, 123(1), 137-162.

Marcus, G. (2018). Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631.

McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in translation: Contextualized word vectors. arXiv preprint arXiv:1708.00107.

Menkhoff, L., & Rebitzky, R. R. (2008). Investor sentiment in the US-dollar: Longer-term, non-linear orientation on PPP. Journal of Empirical finance, 15(3), 455-467.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013b). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Miller, E. M. (1977). Risk, uncertainty, and divergence of opinion. The Journal of finance, 32(4), 1151-1168.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of sentiment analysis in finance: from lexicons to transformers. IEEE Access, 8, 131662-131682.

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

Obaid, K., & Pukthuanthong, K. (2021). A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news. Journal of Financial Economics.

Pástor, Ľ., Stambaugh, R. F., & Taylor, L. A. (2017). Do funds make more when they trade more?. The Journal of Finance, 72(4), 1483-1528.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

Prechter Jr, R. R., & Parker, W. D. (2007). The financial/economic dichotomy in social behavioral dynamics: the socionomic perspective. The Journal of Behavioral Finance, 8(2), 84-108.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.

Rezaeinia, S. M., Rahmani, R., Ghodsi, A., & Veisi, H. (2019). Sentiment analysis based on improved pre-trained word embeddings. Expert Systems with Applications, 117, 139-147.

Scheinkman, J. A., & Xiong, W. (2003). Overconfidence and speculative bubbles. Journal of political Economy, 111(6), 1183-1220.

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020). Measuring news sentiment. Journal of Econometrics.

Shapiro, A. H., & Wilson, D. (2021, January). Taking the fed at its word: A new approach to estimating central bank objectives using text analysis. Federal Reserve Bank of San Francisco.

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

Shiller, Robert J., 2000, Irrational Exuberance (Princeton University Press, Princeton, NJ)

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

Solt, M. E., & Statman, M. (1988). How useful is the sentiment index?. Financial Analysts Journal, 44(5), 45-55.

Stambaugh, R. F., Yu, J., & Yuan, Y. (2012). The short of it: Investor sentiment and anomalies. Journal of Financial Economics, 104(2), 288-302.

Smith, V. L. (2003). Constructivist and ecological rationality in economics. American economic revSummers, L. H. (1986). Does the stock market rationally reflect fundamental values?. The Journal of Finance, 41(3), 591-601.iew, 93(3), 465-508.

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

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

Tetlock, P. C. (2011). All the news that's fit to reprint: Do investors react to stale information?. The Review of Financial Studies, 24(5), 1481-1512.

Tumarkin, R., & Whitelaw, R. F. (2001). News or noise? Internet postings and stock prices. Financial Analysts Journal, 57(3), 41-51.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 37, 141-188.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

Yu, J., & Yuan, Y. (2011). Investor sentiment and the mean–variance relation. Journal of Financial Economics, 100(2), 367-381.

Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016, November). Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 conference on empirical methods in natural language processing (pp. 606-615).

Wang, G., Wang, T., Wang, B., Sambasivan, D., Zhang, Z., Zheng, H., & Zhao, B. Y. (2015, February). Crowds on wall street: Extracting value from collaborative investing platforms. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (pp. 17-30).

Wysocki, P. D. (1998). Cheap talk on the web: The determinants of postings on stock message boards. University of Michigan Business School Working Paper, (98025).

Zhang, D., Xu, H., Su, Z., & Xu, Y. (2015). Chinese comments sentiment classification based on word2vec and SVMperf. Expert Systems with Applications, 42(4), 1857-1863.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4)

9 Tables and Graphs

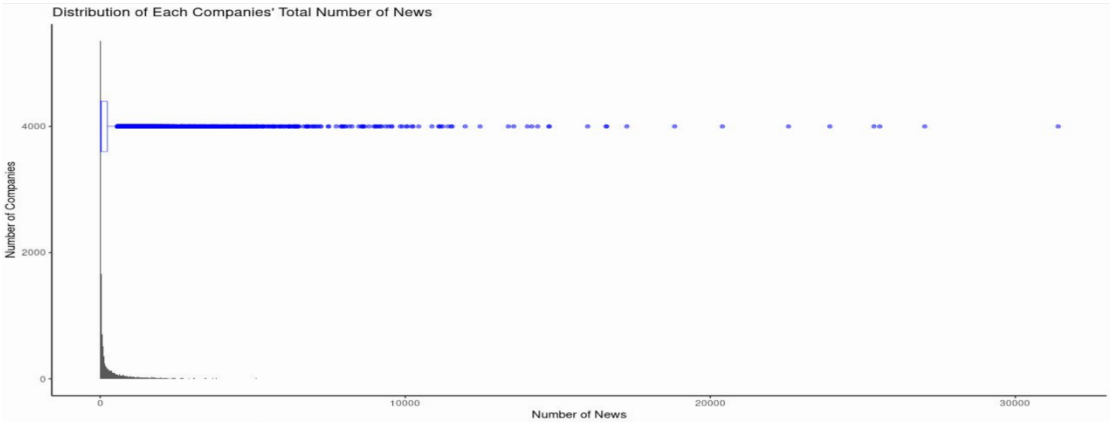


Figure 1. Boxplot and Distribution of Each Company's Total Number of News. This graph shows the distribution and corresponding box plot of total amount of news of each company in our raw sample.

Minimum	25% Quantile	Median	Mean	75% Quantile	Max
1	4	27	390	229	31403

Table 1. Distribution of Total Number of News per Company. This table shows the distribution of total amount of news of each company in our raw sample.

Minimum	25%	Median	Mean	75%	Max	SD
1	57	80	151.5	176	10793	204.55

Table 2. Distribution of article's lengths in number of words. This table shows the distribution of all articles' body lengths before cleaning.

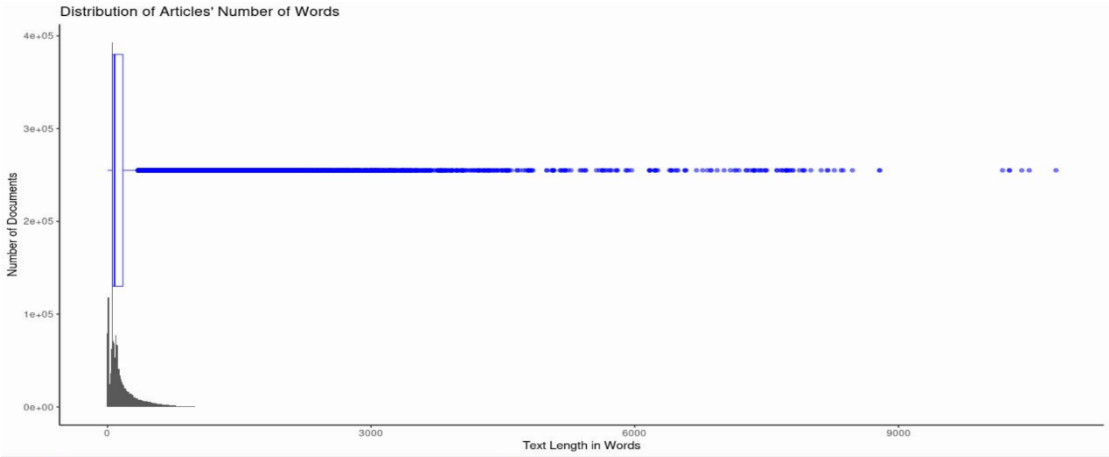


Figure 2. Distribution of Article's Lengths in Number of Words.

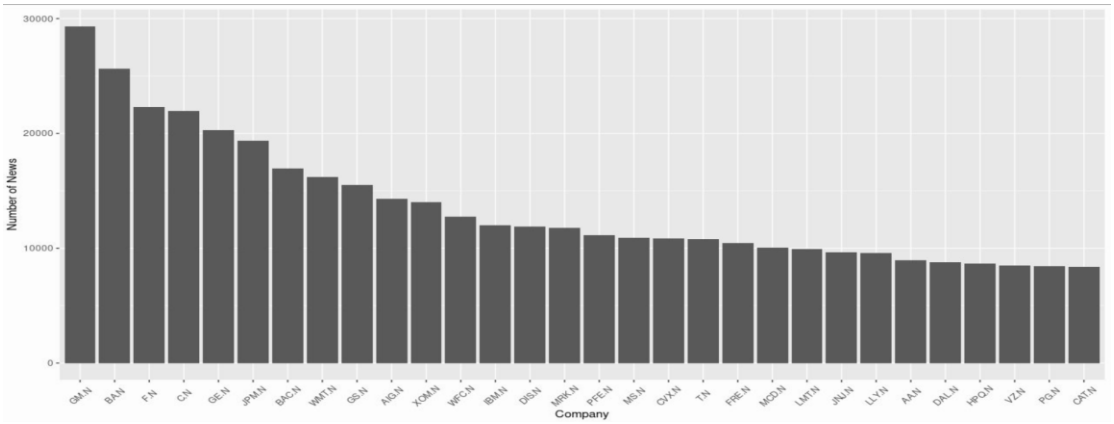


Figure 3. Top 30 companies' number of news over full sample. This figure shows the number of news pieces from top 30 companies by their RIC.

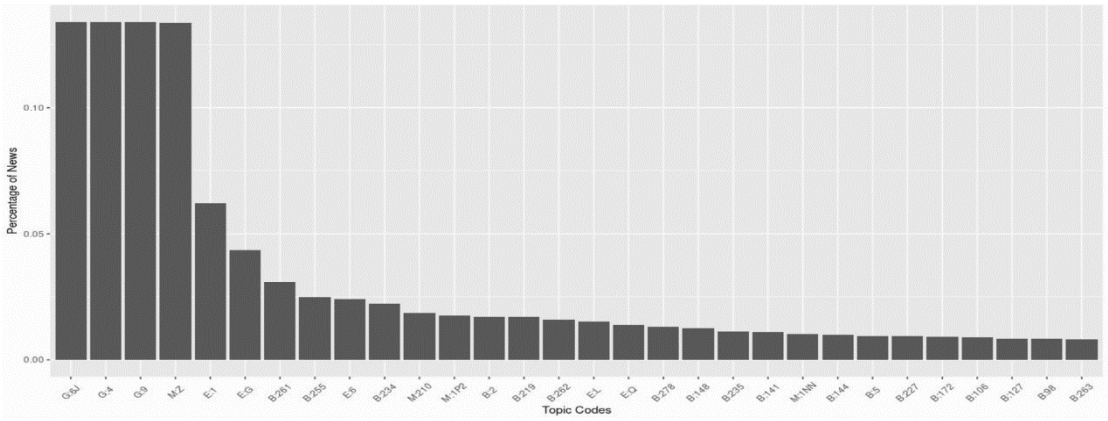


Figure 4. Top 30 topics in full sample. This figure shows the top 30 topics that appear in our sample after cleaning.

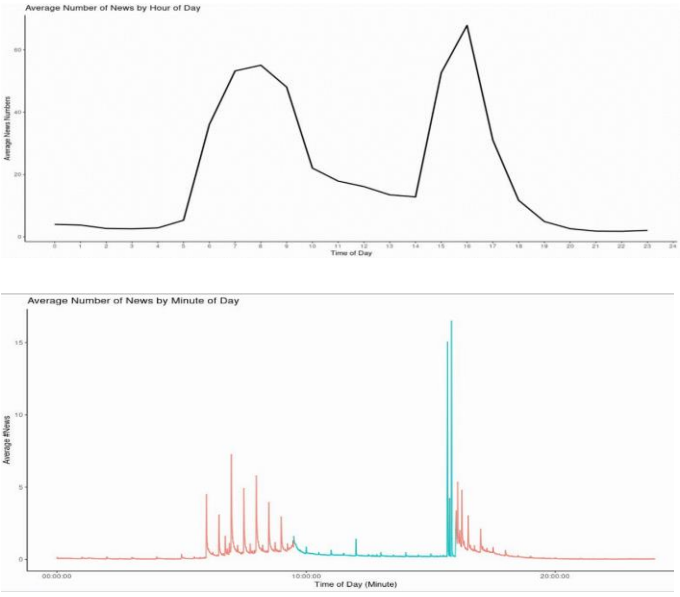


Figure 5. Average Number of News per hour (top) and per minute (bottom).

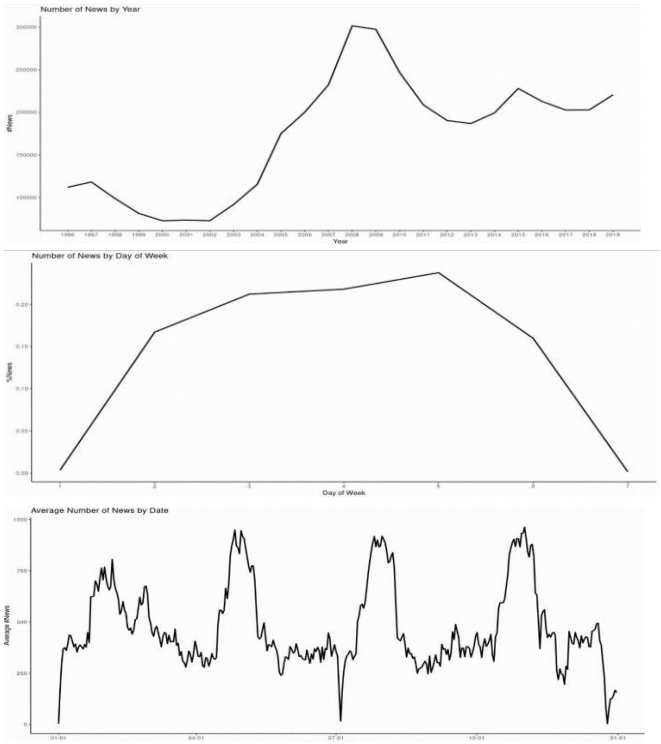


Figure 6. Number of News by Year (top), percentage of news arrivals by day of week(middle, where 1 indicates Sunday, 2 indicates Monday, et cetera), and number of news arrivals by day of year (bottom)

Minimum News (N)	Days	%Sample
40	4534	98.50%
50	4517	98.13%
80	4409	95.79%
100	4297	93.35%
150	3992	86.73%

Table 2. Days with at least 40, 50, 80, 100, and 150 News Alerts and as a percentage of total sampling

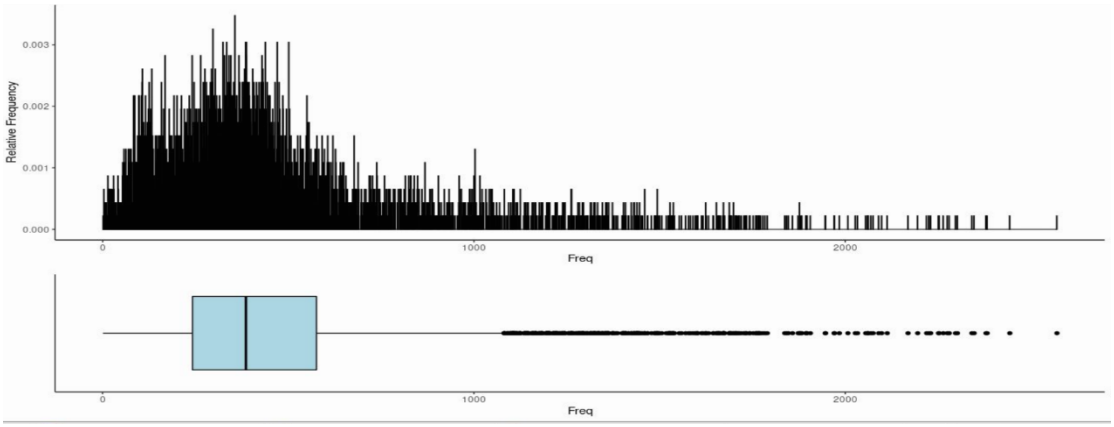


Figure 7. Distribution of News Alerts by Trading Day

	Alerts	Alerts-headline	Article Body
2002	1.85	2.78	2.50
2003	1.05	4.93	5.58
2004	0.91	3.32	2.85
2005	2.26	8.01	6.83
2006	-0.23	5.78	5.51
2007	0.10	4.17	3.97
2008	1.29	4.71	5.23
2009	2.35	2.25	5.39
2010	2.51	7.21	7.47
2011	1.38	0.07	4.02
2012	2.12	6.19	4.82
2013	2.24	3.87	2.16
2014	2.03	5.59	5.91
2015	2.64	5.78	2.23
2016	-1.26	7.19	3.01
2017	4.33	7.8	4.66
2018	1.24	5.98	8.42
2019	-0.45	4.65	3.39
Overall	**2.79**	**3.09**	**3.87**

Table 3. Annual Sharpe ratio under different models. This graph shows annual Sharpe ratio for each year and average Sharpe ratio under different models. Column 1 applies FinBERT on alerts only, column 2 applies FinBERT on all alerts and news articles' headlines, and column 3 applies FinBERT on news articles' body contents only.

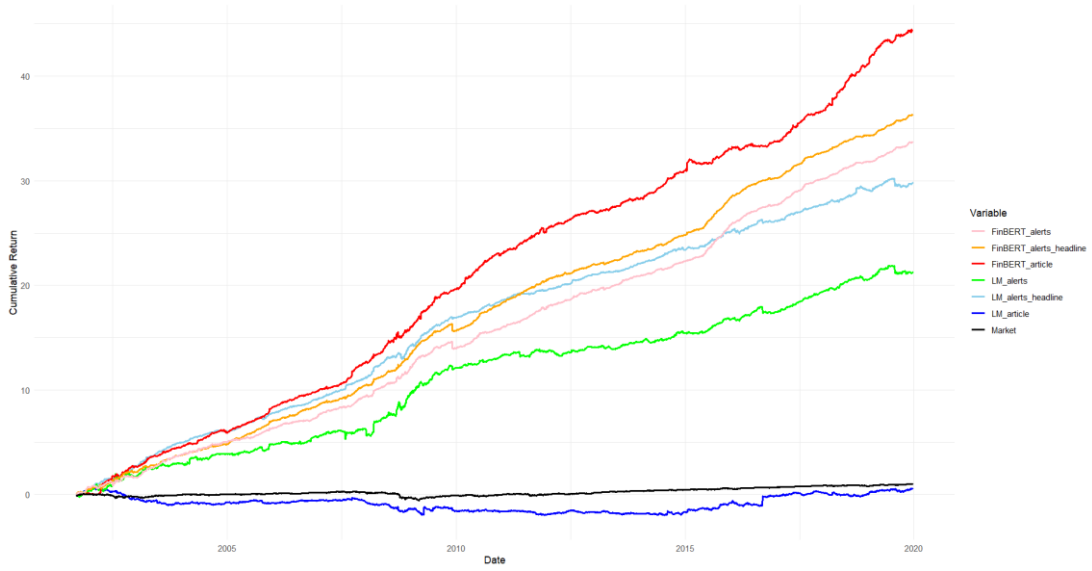


Figure 8. Cumulative daily log Returns of S&P 500 index and sentiment portfolio. This graph shows cumulative log returns using different models starting in 9/2001 and end in 12/2019. Portfolios are constructed based on LM dictionary and FinBERT model using alerts only, alerts and articles' headlines, and articles' body contents only. S&P 500 is included as benchmark and proxies for a passive investment in the market.

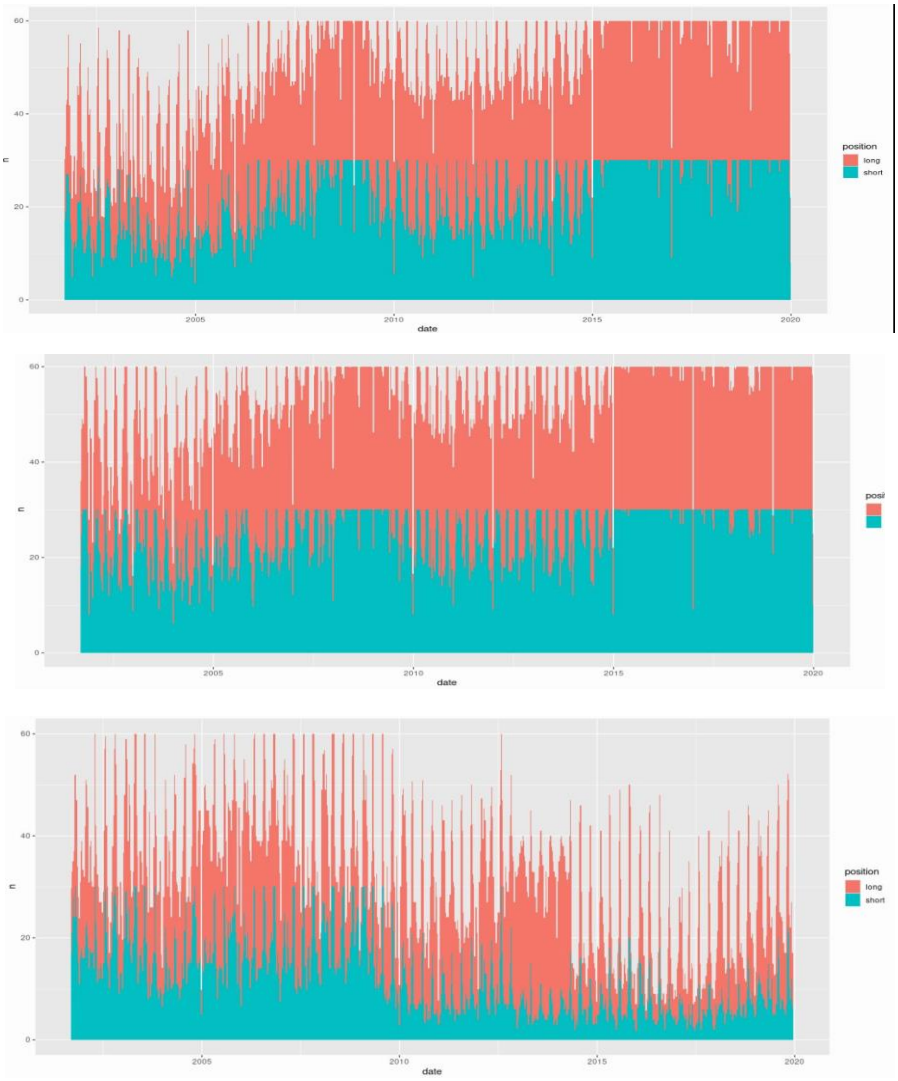


Figure 9. Number of Stocks in Long (red) and Short (blue) Legs (50-30 portfolio). This graph shows the number of stocks in long and short legs of portfolio when considering alerts only (top left), alerts and articles' headlines only (top right), and articles' body contents only (bottom) using FinBERT.

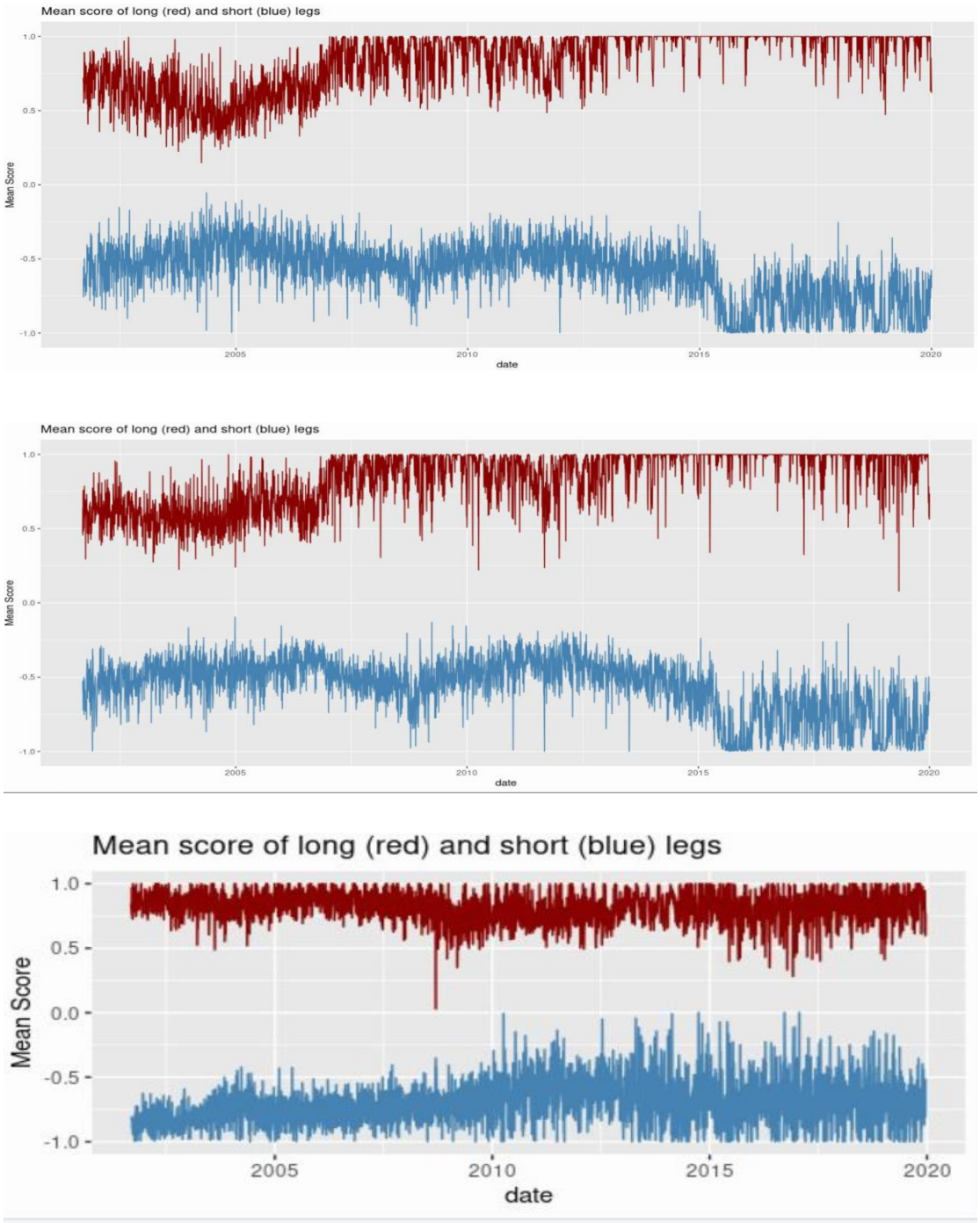


Figure 10. Mean score of long (red) and short (blue) legs by trading day. This graphs shows the mean score of stocks in long and short legs of sentiment portfolio when considering Alerts only (top), Alerts and articles' headlines only (middle), and articles' body contents only (bottom).

Correlation between different models			
	LM	HIV4	FinBERT
LM	1	0.268	0.327
HIV4		1	0.012
FinBERT			1

Table 4. Correlations between LM model, HIV4, and baseline FinBERT model sentiment scores

Label	Negative	Neutral	Positive
%Alerts	0.105	0.686	0.209
%Articles	0.108	0.669	0.222

Table 6 Percentage of Alerts and Articles Body's Predicted Labels. This table shows the percentage of positive, negative, and neutral labels under FinBERT model for news alerts and news articles (body contents).

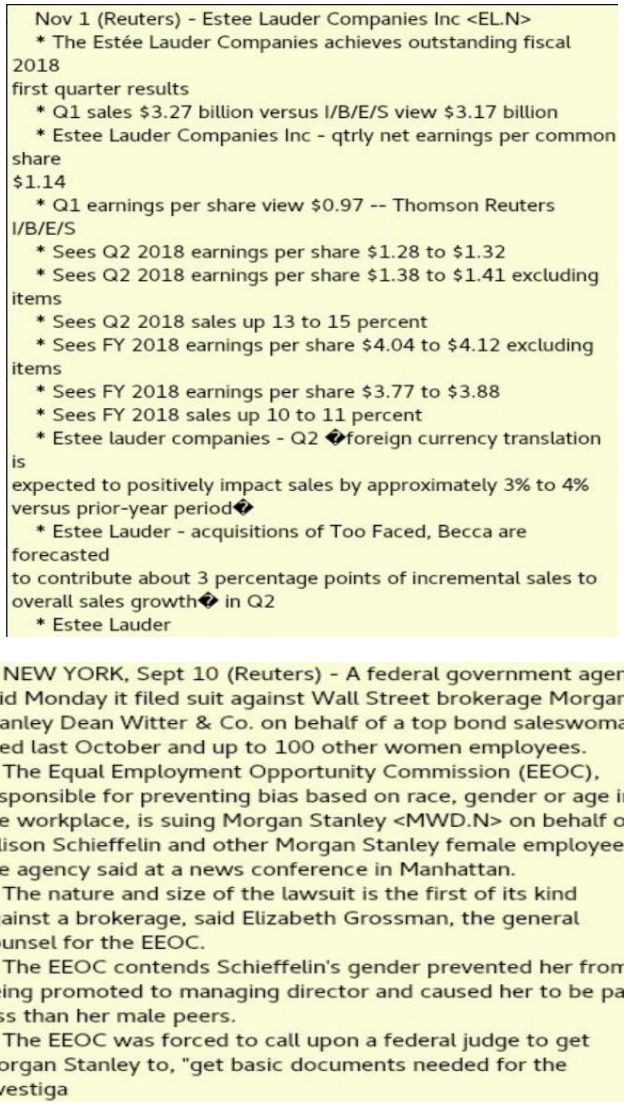


Figure 10. A typical body content of briefs (top) and non-brief ordinary news (bottom)

RIC	n (articles)	n (alerts)	Company
BA.N	2834	5655	Boeing Company
GM.N	1977	3860	General Motors Company
TWTR.N	1874	2804	Twitter
JPM.N	1440	3284	JPMorgan Chase & Co

GS.N	1427	2549	Goldman Sachs Group
F.N	1395	3124	Ford Motor Co
WFC.N	1121	3255	Wells Fargo & Company
WMT.N	1056	2614	Walmart Inc
GE.N	1003	3585	General Electric Company
C.N	1000	3029	Citigroup Inc
XOM.N	991	2511	Exxon Mobil Corporation
BAC.N	822	2201	Bank of America Corporation
LMT.N	762	1988	Lockheed Martin Corporation
BLK.N	749	2405	BlackRock Inc
JNJ.N	688	2390	Johnson & Johnson
DIS.N	676	1636	Walt Disney Co
MRK.N	644	2431	Merck & Co Inc
MS.N	642	1873	Morgan Stanley
PFE.N	632	2550	Pfizer Inc
CVX.N	586	1998	Chevron Corp
ICE.N	566	1872	Intercontinental Exchange Inc
DAL.N	553	2363	Delta Air Lines Inc
T.N	544	2235	AT&T Inc
LLY.N	538	2486	El We Lilly and Company
NYT.N	531	702	New York Times Co
PCG.N	524	1787	PG&E Corp
MCD.N	513	1741	McDonald's Corporation
NKE.N	508	1543	Nike Inc
UAL.N	475	1669	United Airlines Holdings Inc
AIG.N	435	1653	American International Group
BX.N	435	1275	Blackstone Inc
CAT.N	427	2385	Caterpillar Inc
FCX.N	401	1798	Freeport-McMoRan Inc
VZ.N	400	1773	Verizon Communications Inc
TGT.N	383	1715	Target Corp
BMY.N	375	1955	Bristol-Ourers Squibb Company
LUV.N	361	1857	Southwest Airlines Co
KO.N	348	1499	Coca-Cola Co
UTX.N	337	1597	United Technologies Corp
UPS.N	336	1582	United Parcel Service Inc
IBM.N	334	1598	International Business Machines Corporation
MDT.N	317	1774	Medtronic PLC
BK.N	309	1227	Bank of New York Mellon Corp
RTN.N	302	1274	Restaurant Group PLC
Total	**32,571**	**97,102**	

Table 7. List of companies in news heterogeneity analysis. This table shows the list of companies that we use in news heterogeneity analysis with high-frequency data.

		length of article body	Number of audiences	Number of Topics	Percentile of RV
alerts	Negative		3.7	23.3	73.2
	Neutral		3.6	23.9	64.5
	Positive		3.3	21.8	71.3
	F-Stat		**174.2**	**203.1**	**697.1**
article body	Negative	191	8.6	32.5	70.8
	Neutral	175	7.9	31.8	56.8
	Positive	187	7.5	28.1	76.0
	F-stat	**43.35**	**90.6**	**84.3**	**965.2**
article title	Negative		7.6	34.1	66.2
	Neutral		8.1	31.2	57.6
	Positive		7.6	29.3	70.4
	F-stat		**32.8**	**48.3**	**433.2**

Table 8. News heterogeneity

Figure 11. Word Clouds of news articles in long (left) and short (right) leg of portfolio

	Alerts only	Alerts - headline	Article Body
	Panel A		

Mean excess return	0.007	0.008	0.010
sd	0.042	0.041	0.042
Sharpe% (Annualised)	2.79	3.09	3.87
% Profitable days (excess return)	67.97	70.44	63.73
Panel B			
Mean return (long)	0.002	0.002	0.002
sd(return) (long)	0.019	0.019	0.024
Mean return (short)	0.006	0.006	0.009
sd(return) (short)	0.044	0.043	0.042

Table 9. Portfolio Sharpe ratios. Panel A shows daily mean excess return, standard deviation of excess ratio, Sharpe ratio of portfolio, and percentage of profitable days under FinBERT model. Panel B shows average and standard deviation of daily portfolio returns by long and short legs.

	Alerts	**Alerts - Headline**	**Article Body**
(Intercept)	0.007^{***}	0.008^{***}	0.010^{***}
	(0.001)	(0.001)	(0.001)
Market	-0.001	0.000	-0.000
	(0.001)	(0.001)	(0.001)
SMB	-0.001	-0.000	-0.002
	(0.001)	(0.001)	(0.001)
HML	-0.003^{**}	-0.004^{***}	-0.005^{***}
	(0.001)	(0.001)	(0.001)
RMW	0.000	0.002	0.001
	(0.002)	(0.002)	(0.002)
CMA	-0.001	0.000	0.001
	(0.002)	(0.002)	(0.002)
Information Ratio	0.177	0.195	0.245
R^2	0.004	0.004	0.008
Adj. R^2	0.003	0.003	0.007
Num. obs.	4496	4530	4287

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 10. Fama-French 5-factor model results of FinBERT model daily excess returns.