12-31-2022

Twitter sentiment analysis: applications in healthcare and finance

Jiali Wang

New Jersey Institute of Technology

ABSTRACT

TWITTER SENTIMENT ANALYSIS:
APPLICATIONS IN HEALTHCARE AND FINANCE

By
Jiali Wang

This research explores the influence of Twitter sentiment on healthcare and finance industries. It assesses how Twitter sentiment and culture measure influence COVID-19 statistics, and it investigates the impact of Twitter sentiment on S&P 1500 stock mispricing. Furthermore, it examines how tweet sentiment predicts major industry returns.

The first part examines how Hofstede's Culture Dimensions (HCD) and Twitter economic uncertainty index (TEU) relate to COVID-19 infection rate and death rate. The results show certain aspects in HCD, such as power distance index (PDI) and masculinity (MAS) both are negatively and significantly associated with the infection rate, while indulgence (IVR) and long-term orientation (LTO) exhibit negative statistical significance to the death rate. TEU based in USA is relevant to COVID-19 death rate in short run (up to 3 months). Some practical strategies are proposed for public health officials to help mitigate COVID-19 spread.

The second part bridges a research gap by exploring the relation between aggregated tweet contents and stock market mispricing. In short, tweet features affect future stock mispricing, in different directions and magnitudes. For overvalued stocks, tweet variables including proportion of external links, average number of words, percentage of retweets, likes and replies are negatively associated with mispricing of S&P 1500 stocks. Average number of words possibly reduces mispricing by reducing idiosyncratic volatility, while proportion of external links can mitigate mispricing via

channels other than liquidity or idiosyncratic volatility. For undervalued stocks, only average number of words is positively related to mispricing; average number of words affect mispricing via channels other than liquidity or idiosyncratic volatility.

Additionally, this study investigates how tweet sentiment from S&P 1500 firms predicts major industry returns by constructing multiple sentiment indices. The robustness tests show highly consistent results, proving such indices can predict the returns from three out of five major industries, including Consumables, High Technology and Healthcare. In general, the sentiment index type and prediction length do not matter much.

In conclusion, this research shows tweet sentiment is more than some meaningless noise. Instead, it has beneficial applications in both healthcare and finance fields, such as COVID-19 pandemic prediction and possible investment reference.

TWITTER SENTIMENT ANALYSIS: APPLICATIONS IN HEALTHCARE AND FINANCE

by
Jiali Wang

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Business Data Science

Martin Tuchman School of Management

December 2022

APPROVAL PAGE

TWITTER SENTIMENT ANALYSIS:
APPLICATIONS IN HEALTHCARE AND FINANCE

Jiali Wang

Doctor Zhipeng (Allen) Yan, Dissertation Advisor Date
Director, Beilaide Jianxin Licai Company, China

Professor Jim (Junmin) Shi, Committee Chair Date
Associate Professor of Supply Chain Management and Finance,
Martin Tuchman School of Management, NJIT

Professor Michael Ehrlich, Committee Member Date
Associate Professor of Finance and Director of Leir Institute for Business, Technology and Society
Martin Tuchman School of Management, NJIT

Professor Dantong Yu, Committee Member Date
Associate Professor of Business Data Science,
Martin Tuchman School of Management, NJIT

Professor Xinyuan (Stacie) Tao, Committee Member Date
Assistant Professor of Finance, Martin Tuchman School of Management, NJIT

Professor Chase Wu, External Committee Member Date
Professor of Computer Science, Ying Wu College of Computing, NJIT

BIOGRAPHICAL SKETCH

Author: Jiali Wang

Degree: Doctor of Philosophy

Date: December 2022

Undergraduate and Graduate Education:

- Doctor of Philosophy in Business Data Science,
  New Jersey Institute of Technology, Newark, NJ, 2022

- Master of Science in Financial Mathematics,
  University of Chicago, Chicago, IL, 2016

- Bachelor of Science in Mathematics,
  Pennsylvania State University, State College, PA, 2014

Major: Business Data Science

Publication:

Wang, J., Bandera, C., & Yan, Z. (2022). Culture Measure, Twitter Sentiment and COVID-19 Statistics: A Regression Analysis. *JMIR Public Health and Surveillance* (in review). https://doi.org/10.2196/44506

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my dissertation advisor Doctor Zhipeng (Allan) Yan for his guidance and dedicated support.

I would also like to thank the rest of my dissertation committee members, including Professor Jim (Junmin) Shi, Professor Dantong Yu, Professor Michael Ehrlich, Professor Xinyuan (Stacie) Tao, and Professor Chase Wu for providing insightful comments.

Meanwhile, I appreciate Professor Xuewu (Wesley) Wang for his instructions during our regular meetings.

In addition, I am very grateful to Martin Tuchman School of Management for the financial support.

In the meantime, I am very thankful to my peers, Shaoqing Zhang and Wuji Liu for their assistance.

Finally, I am deeply indebted to my family members for their understanding and tremendous support for years.

TABLE OF CONTENTS

Chapter **Page**

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF TABLES
(Continued)

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Objective

This dissertation aims to answer three questions, with the first one being healthcare related, and the last two being financially concentrated. More specifically, the first question attempts to link culture measure and Twitter sentiment to COVID-19 pandemic to see if there is any significance to the infection rate or death rate. The second part aims to answer the other question: how does tweet sentiment relate to stock mispricing? The last part seeks to unravel the relation between tweet sentiment and major industry returns.

1.2 Background

Social media serve as a digital channel to facilitate communication of user-generated contents by its interactive nature. Unlike traditional media whose creators are usually professionals, such as reporters, columnists and analysts, users mostly drive social media, and many of them are amateurs. These users do not necessarily have momentary incentives to share their contents. Instead, the user-generated contents often serve as a channel to fulfill their desire of self-expression (Lietsala & Sirkkunen, 2008). Some most well-known examples include You Tube, Facebook, and Twitter. Twitter is a micro-blogging system which allows users to compile messages up to 280 characters in length. There are many studies whose research results are based on datasets extracted from Twitter, either from company tweets or personal tweets. Despite its relatively short existence (the site launched in 2006), it has become a prevalent way of communication nowadays globally. There are around 340 million monthly active users across various countries. [1] We use tweets for our research because Twitter (64%) is the most popular data source to perform surveillance research using social media text data (Gupta & Katarya, 2020). However, there is more than that and researchers both in healthcare and finance industries can take advantage of them to extrapolate something meaningful. Tweets are helpful for medical professionals to fight pandemic by conveying valuable statistics when properly analyzed. For example, an Epidemic Sentiment Monitoring System has been developed via tweets to help public health officials strategize disease actions (Ji et al., 2013). Another benefit is Twitter remarkably facilitates medical collaboration, which is an extremely valuable way to potentially revolutionize public health efforts (Pershad et al., 2018).

There is ample literature on the relationship between infectious disease and natural factors such as environmental conditions, biological bases, and comorbidities. However, few studies investigate sociological or economic attributes, and even fewer attempt to relate pandemics to culture dimensions. This dissertation aims to fill the gap by linking spread of the ongoing COVID-19 pandemic to social factors proxied by the Geert Hofstede's Culture Dimensions (HCD) that include power distance index (PDI), indulgence (IVR), long-term orientation (LTO), uncertainty and avoidance (UAI), individualism (IDV) and masculinity (MAS).

COVID-19 has spread globally and evolved into a pandemic with far-reaching impacts. The Coronavirus Resource Center of John Hopkins University of Medicine reports over 628 million confirmed infection cases worldwide and over 6 million deaths as of October 2022 (*COVID-19 Map*, 2022). The recommended safety policies result in social

[1] Retrieved on November 12, 2022, from https://www.businessofapps.com/data/twitter-statistics/

isolation and significant national, organizational, and individual economic disruption. The global economic growth is expected to slow from 6% in 2021 to 2.7% in 2022 and 2.7% in 2023, partially due to COVID-19 (*World Economic Outlook*, 2022). The International Labor Organization estimates a wipeout of 6.7% working hours in the second quarter in 2020, which equals 195 million full-time workers, and the estimated unemployment number is 30 million, in comparison to 25 million during the 2008 financial crisis (*ILO*, 2020). The cumulative GDP loss globally over 2020 and 2021 from this pandemic could amount to around 9 trillion dollars, although there is a projected global growth of 5.8% next year (Gopinath, 2020). Other social issues include increased bankruptcies and unemployment rate. Chapter 11 filing, a form of bankruptcy record, has increased almost by 200% from large corporations, from January to August in 2020, as compared to 2019 (J. Wang et al., 2020). The data from U.S. Bureau of Labor Statistics shows that the civilian unemployment rate nationwide has been going up significantly in March 2020, peaking at around 15% one month later (*Civilian Unemployment Rate*, n.d.).

It is reported that public health cooperation at global level can effectively alleviate such issues, yet due to conflicted interests, politicians keep ignoring possible cooperation and intensifies contradictions among countries (McKibbin & Fernando, 2021). Recovery from a pandemic is accelerated by compliance with protocols recommended by public health authorities such as the World Health Organization, the Centers for Disease Control and Prevention, and the relevant institutes with the National Institutes of Health. However, culture influences compliance, which is why pandemic hot spots are often localized regions that are culturally homogeneous and culturally different from their surroundings, i.e., geographies whose health and cultural measures are easily segmented from their surroundings. Public health researchers have realized the importance of culture in public health campaigns (Bavel et al., 2020). Therefore, it is important to quantify a nation's culture. Geert Hofstede's seminal work of culture has become the most prevalent model among social scientists evaluating cultural dependencies. Hofstede Culture Dimension (HCD) uses six constructs with values ranging from 1 to 100 to describe culture: Power Distance Index (PDI), Individualism vs Collectivism (IDV), Masculinity vs. Femininity (MAS), Uncertainty Avoidance (UAI), Long Term Orientation vs. Short Term Normative Orientation (LTO) and Indulgence vs Restraint (IVR) (Hofstede, 2011). HCD has may applications. For example, the Global Entrepreneurship Monitor uses HCD to evaluate the state of entrepreneurship in countries across the world (150,000 participants in 50 economies). Since 1998, policy makers use the annual GEM data to advance the regional economic impact of entrepreneurship (Bosma et al., 2019). In general, cultures with high individualism and low uncertainty avoidance exhibit the strongest entrepreneurial activity (Mueller & Thomas, 2001). While HCD is most frequently used in economic studies, we propose that it might also provide insight into the impact of culture on compliance with public health recommendations and help explain disparities between the COVID-19 statistics of different countries. We investigate how PDI, LTO, MAS, IDV, UAI and IVR of a country correlate with its COVID-19 infection rate and death rate.

The studies connecting tweets and finance can be classified in two categories, which are information dissemination and tweet implications. The former focuses on how, why, and when corporate executives use tweets as a channel to diffuse pivotal information and the mechanism of such information spread and they are commonly seen in accounting and are often related to information asymmetry, disclosure, and tone interpretation. The latter involves what financial statistics can be inferred from tweets, including stock price prediction, earning forecast and cost of capital estimation. This study is more relevant to the second group, as we are more interested in learning how sentiment proxied by tweets relates to financial market performance, both quantitatively and qualitatively.

Twitter information is equally beneficial in finance industry if applied strategically. For instance, independent studies prove financial Twitter data have statistical significance on stock returns (Ranco et al., 2015). Besides, gathering and analyzing tweets is also an effective way to indirectly promote industry growth, as they provide meaningful feedbacks for companies to act accordingly. For example, in manufacturing industry, automotive companies can enhance their marketing strategies by analyzing brand sentiments from consumer tweets (Shukri et al., 2015).

As technology develops, managers have new ways to communicate the information about their firms, including Twitter, Facebook, LinkedIn, Pinterest, YouTube, Instagram, Google+ and so forth. At the same time, increasing mobility and declining cost of devices with Internet connectivity makes information more accessible to a larger population. These changes affect the way how information about a firm is produced, disseminated, and processed by managers and consequently generate impacts on the financial markets. We seek to explore and understand the way of information communication and how it impacts mispricing. Unlike conventional media, social media have several new features. They allow message posting in a real time as well as instant interactions among readers. There is a variety of formats, including text message and pictures, links to external references, videos and emoji expressions are also available. In addition, there is less restriction on the capacity, unlike the space limitation for newspapers and time concern for conference calls.

Given those features, information communication through social media seems to be more convenient, accessible and with less costs, benefiting the information environment of financial markets. On the other hand, the messages released on social media are voluntary and fully in charge of the managers. Managers decide when, what and how to disclosure this information. Likely, they are going to make the decision on behalf of their firms. Firms with less sophisticated investors or more social media audience have stronger incentive to strategically disseminate financial information via tweets, and similarly, firms with high litigation risk tend to exhibit strategical dissemination (Jung et al., 2017).

However, the overall impact on the information environment is under-explored. Therefore, our study intends to answer the question of asset pricing. More specifically, whether the usage of social media will increase or decrease mispricing. Our study focuses on the usage of tweets on Twitter. Twitter is arguably one of the most popular social media websites currently, disseminating information via messages called tweets. One tweet may contain text up to 280 characters in length, one link, one picture or video, and emoji expression. We expect that if the usage of tweets overall improves the information, the mispricing in the next period will be reduced. We collect the official Twitter account from the list of S&P 1500 firms. First, we show that the usage of tweets with more attention and more embedded links can reduce mispricing, which suggests tweets serve as a channel for information disclosure and benefit the information environment, consistent with the previous study that firm tweets reduce information asymmetry (Blankespoor et al., 2013). The tone of tweets plays a significant role in influencing mispricing. Specifically, sentiment score and mispricing are positively associated, which is consistent with the report that investor sentiment has a positive and significant relation with abnormal stock return

(McGurk et al., 2020). As the tweets posting is decided by managers, most messages have a positive tone. However, there might be polar reactions from readers to such negative messages. While some people consider the posting of negative news as honesty, which makes them trust the firm more, others worry that it could be ominous and have even worse news underneath. Second, we find that the impact of tweets' usage on mispricing is persistent. This finding suggests that the effect of information communication through social media might be more far-reaching than some people think, which is worth more future research. Third, we attempt to understand by which channel the usage of tweets affects mispricing. It turns out there are different influencing mechanisms. For example, average number of words can reduce mispricing by lowering idiosyncratic volatility.

The motivation of the last part rests on the belief that sentiment data available from social media provides valuable information for financial analysis. More specifically, we assume that the aggregate Twitter sentiment conveys useful information in financial performance prediction. The assumption is supported by relevant studies mentioned in the literature review section. The primary research question is whether major players in a financial market, which are some large and medium firms in S&P 1500 in this case, divulge beneficial clues to financial market. Based on our belief and assumption, we construct a sentiment index from company tweets available from S&P 1500 firms and attempt to predict several performance measures in financial markets at monthly levels. We focus on the period between 2009 and 2021, as it takes Twitter a few years to gain popularity among the users since its 2006 creation, which means more available data for our analysis. Then we propose a few ways to aggregate company sentiment from individual firms to market level. After implementing multiple regression analyses, we discover that both our sentiment indices and sentiment dispersion can successfully predict future industry returns.

CHAPTER 2

CONTRIBUTIONS

There are multiple contributions, including fields of both healthcare and finance.

In the healthcare field, this study bridges the gap between culture literature and pandemic research, as there are few works exploring how culture and pandemic response are related. We reveal that Twitter sentiment is relevant to COVID-19 death rate in short run (up to 3 months). The study has social, policy and economic implications. The study informs policy makers on their country's cultural attributes that impact compliance. Public health officials can utilize HCD for public health campaigns. The goal is similar in both cases: increase participation and compliance with recommendations that appeal to regional culture. We find that IVR and death rate are positively related, which is in line with the finding that a high indulgence score with individualistic culture exacerbates fatality (Oey & Rahardjo, 2021). The importance of enhancing social distancing cannot be overestimated, which will generate massive economic benefit if implemented properly. In addition to saving lives, effective social distancing brings economic benefit worth \$5.16 trillion for the U.S., which accounts for more than 24% of its GPD in 2020 (Thunström et al., 2020). Besides, we reveal that the economic uncertainty reflected from Twitter sentiment is linked to COVID-19 statistics. TEU is an effective predictor to forecast short term COVID-19 death rate. This conclusion may enlighten healthcare policy makers, based on available Twitter sentiment.

The contributions are also applicable to finance, both in asset mispricing and investment return. In the asset mispricing aspect, we find that tweets affect mispricing via different channels, such as idiosyncratic volatility and liquidity. This finding contributes to literature related to social media and information disclosure. Our article also contributes to crowdsourcing and market sentiment, as we show these tweets contain useful sentiment which links to the financial market. Additionally, our unique data collection likely benefits relevant research. We have filtered more than 2000 firms over 10 gigabytes. These raw datasets alone have wide applications in many disciplines, especially in business fields, such as finance and accounting. There are many possible topics that can be explored based on our datasets, such as information dissemination, information asymmetry or disclosure, earning forecast, stock price prediction, cost of capital equity and insider trading. The data files collected date back to the creation time of Twitter in 2006, spanning 14 years, which allows more comprehensive analysis over time.

There are contributions to investment return as well. First, these simple yet effective sentiment indices can successfully predict future returns of multiple industries up to a 12-month horizon, including Consumables, High Technology and Healthcare, which is valuable for practitioners such as traders and portfolio managers who engage in any of the three industries, which could serve as a useful investment reference to possibly shield against the current inflation. Second, this paper enriches the sentimental analysis literature investigating how to predict financial market performance by gauging company sentiments and sentiment analysis is worthy to reveal meaningful information for financial prediction. Third, in a broader sense, the evidence from this research manuscript supports the behavioral finance theory in that many players in financial markets are often irrational and their aggregated sentiment is something to be reckoned with. This discovery emphasizes the fact that emotion has remarkable influence on financial investment decisions, showing inconsistency with standard financial theory that market quickly absorbs information including tweets, as the sentiment-based index exhibits significant prediction power over time.

CHAPTER 3

LITERATURE REVIEW

3.1 Culture Measure, Twitter Sentiment, and COVID-19 Pandemic

Culture difference is a contributing factor in the severity of COVID-19, and countries with sociable cultures such as Italy and Spain suffer more in this pandemic, while it is easier for Japan to adopt social distancing due to the lack of close contact in Japanese culture (Baniamin et al., 2020). It is also not surprising that East Asian culture emphasizes collectivism, for which infringing policies against individual freedom are more prevalent, while in western culture individualism usually dominates collectivism even in a crisis (An & Tang, 2020).

While HCD is prevalent in cross-culture research, many researchers also realize its limitations. Some scholars argue that Hofstede's uni-level analysis neglects interactions between macroscopic and microscopic cultural levels (McSweeney, 2002). Others blame the theory for having "ecological fallacy" whereby there is a correlation inconsistency at national (ecological), individual and organizational level (Brewer & Venaik, 2014). In other words, even if HCD accurately describe a nation's culture, it does not mean the citizens from such country behave exactly as the theory suggests. In addition, dividing cultures into "stereotypes" may be misleading (Jain, 2020). Therefore, some more elegant methodology is needed for culture measure and analysis. However, Hofstede knows such limitations and reminds HCD users that "the concept of a common culture applies to societies, not to nations" (Hofstede, Hofstede et al 2010, p. 21), and "one of the weaknesses of much cross-cultural research is not recognizing the difference between analysis at the societal level and at the individual level" (Hofstede, 2011).

Multiple studies show HCD is relevant to COVID-19 statistics. It is questionable whether Americans will comply with government's quarantine decision, due to their ingrained values, including adamant individualism (IDV), self-reliance, nonconformity and independence, and it is unimaginable for them to experience a cordon sanitaire exhibited in Wuhan city of China (Calandrillo, 2004). IVR is related to other infectious diseases and health outcomes. Countries with high UAI, low LTO, low IVR and high IDV have more stock piling behavior after the World Health Organization declared COVID-19 a pandemic (Ahmadi et al., 2022). IVR is mostly correlated with both health outcomes and health behaviors (Mackenbach, 2014). People in countries with higher IVR scores, such as U.S. and other western countries excluding Germany tend to disobey stay-at-home orders, requiring authorities to enforce such order implementation, while Chinese culture only scoring about a quarter of U.S. IVR values endurance and patience, which is a constructive response to the pandemic (Travica, 2020). PDI is a measure of deference to authority figures, which might impact compliance with centralized public health guidance. There can also be a feedback loop between policy and compliance: to avoid losing popularity, some politicians in low PDI cultures might advocate policies they believe will be well received by most of their constituents even if such policies may diverge from those recommended by healthcare professionals. A previous study shows IVR and IDV positively predict COVID-19 cases across nations, while PDI negatively forecasts these cases nationally (Dheer et al., 2020). Scholars also find that both IDV and IVR have significant and positive effects on total COVID-19 cases per million in selected European countries, while PDI has a significant and negative effects (Gokmen et al., 2021).

Similarly, the sentiment analysis can unravel meaningful information of COVID-19 trend from social media, such as Twitter, which is an excellent channel to make the pandemic more understandable for the public (Boon-Itt & Skunkan, 2020). By analyzing English tweets, it turns out people mostly express their fear to the unknow nature of COVID-19 during its outbreak between late January and early March in 2020 (Xue, Chen, Chen, et al., 2020), whereas most users located in USA show neutral sentiment in April, 2020 (Yeasmin et al., 2022). By applying Extra Trees Classifiers, tweets provided by an IEEE dataset up to May 2020 become an effective predictor for COVID-19 (Rustam et al., 2021). Another Twitter analysis related to COVID-19 shows that negative sentiment is harmful, although it also plays a pivotal role in adjusting public sentiment. Therefore, a proactive tactic is needed to balance such emotion out in a timely manner during this pandemic crisis (Naseem et al., 2021). In conclusion, researchers come to consensus that Twitter is a rich medium to potentially enhance real time public awareness of COVID-19 (Medford et al., 2020).

However, there is limited literature exploring how Twitter economic uncertainty indices relates to major pandemic statistics, which explains why it is worthy studying. Some US and UK based economic uncertainty measures, including TEU, are sensitive indicators for the pandemic between January and April in 2020 (Altig et al., 2020). Economic indicators constructed from tweets, also show potential in predicting financial investment. For example, A relevant finding shows TEU is a robust predictor for oil volatility during the pandemic (Lang et al., 2022). In addition, there is a convincing causal link between TEU and cryptocurrency returns (Aharon et al., 2022). Other economic indices based on tweets correlate well with Dow Jones, NASDAQ, VID and S&P 500 (Zhang et al., 2011). Therefore, it is a valid attempt to integrate TEU into a new tweet measure for financial prediction, which is part of what is done in this dissertation.

3.2 Twitter Sentiment and Mispricing

One focus of finance research studies investor sentiment and how tweets reflect such sentiment. There is a study investigating how tweet sentiment affects mispricing on the Indonesian stock market. The researchers have concluded that arbitrage opportunities may be available, and investors can reap profit by taking long position on overvalued stocks following low sentiment, or when there is a positive movement in sentiment (Indra & Husodo, 2020). In contrast to standard financial theory, behavioral finance literature shows that social media has explanatory power to equity return, and more specifically, investor sentiment has a significant and positive relation with abnormal stock returns (McGurk, Nowak, and Hall, 2019). Similarly, there is evidence documenting retail investor sentiment plays a pivotal role in the process of equity pricing, and it is possible to develop trading strategies to take advantage of it to acquire excess returns (Burghardt et al., 2008). Investor sentiment is negatively related to the subsequent quarter returns for both public and private real estate markets, and it also causes mispricing in private real estate markets in multiple periods (Ling et al., 2010). There is evidence that investor sentiment influences asset pricing, especially in real estate markets, even after taking account for control variables such as equity risk premiums and lagged adjustments (Clayton et al., 2009). Additionally, a survey measure built on investor sentiment can explain deviations from intrinsic value of stocks, with a horizon in the next one to three years (Cliff & Brown, 2001). Despite being viewed as irrational, usually investor sentiment along with risk aversion and time preference exhibit clear behavioral patterns in business cycle, when applied by behavioral theory. Such sentiment is also strongly associated with the interaction between risk and return (Barone-Adesi et al., 2017).

However, it has been reported that sentiment-induced mispricing in stock market can be mitigated by quality accounting information. High sentiment is related to more favorable recommendations to hard-to-value firms, even with signs of overpricing and negative subsequent abnormal returns. In contrast, with inadequate account information, such behavior tends to exhibit more frequently (Cornell et al., 2017).

There is much literature about the relation between investor sentiment and asset valuation. However, there are limited studies connecting social media and mispricing. Tone from certain tweets provides valuable yet hidden information which predicts asset fundamentals in the U.K. betting exchange. Even for investment novices, strategies primarily relying on the judgement of social media activity can yield rewarding return for very short-term investments. Meanwhile, the tone of tweets can stabilize market sentiments and prevent investor overreactions (Brown et al., 2016). Postings from the Internet Stock Message Board capture the sentiment of firm-specific investors well and are relevant to temporary stock mispricing (Xiong et al., 2020). Accounting transparency and investor base jointly influence expected mispricing of financial professionals, with a positive correlation (Elliott et al., 2010). There is a positive association between Market-to-Book ratios and overpricing (Bloomfield & Michaely, 2002). Press coverage of annual earnings announcement mitigates cash flow mispricing, while it has little contribution to accrual mispricing. It is because the press disseminates the information more broadly (Drake et al., 2014). There is a positive relation between tweet-based product information disclosure and Tobin's Q (Majumdar & Bose, 2019). A firm’s idiosyncratic risk difference can be gauged from the tone, and it is proven that the tone affects mispricing and arbitrage limits (Liu & Han, 2020). A recent study shows Twitter sentiment is important around earning announcement. More specifically, sentiment-driven short-term mispricing and subsequent return reversals around earnings announcements (Karampatsas et al., 2022). Through Twittersphere, sentiment-induced mispricing is asymmetric, i.e., commodities with low (high) sentiment shifts tend to be overvalued (undervalued) when the aggregate market is in backwardation (contango); the observed premium arises almost entirely from commodities with the most retweet activities, while retweets and likes themselves do not exhibit stronger predictive ability compared to non-influential tweets (Fan et al., 2022). Twitter dissemination seems to be economically important even for sophisticated bond and CDS investors, as well as information intermediaries (Bartov et al., 2022).

There are a lot of known mispricing factors. By combining stock rankings with 11 anomalies and averaging rankings related to these anomalies with prominent co-movement of long-short profits, the resulting long-short return spreads are constructed as mispricing factors, both of which are significantly related to lagged investor sentiment. A four-factor model is developed, after combining the two factors with market and size factors, which outperforms the four-factor model (Hou et al., 2017) and the five-factor model (Fama & French, 2015), in terms of accommodation ability to various anomalies (Stambaugh et al., 2015). Probability of manipulation has been neglected as a contributing variable to earning forecasting models related to accrual mispricing, while this factor significantly weakens accrual mispricing (Beneish & Nichols, 2005). Started from psychological biases of overconfidence and limited attention, a financing factor and a post-earnings announcement drift factor are constructed to reflect long-term (> 1 year) and short-term (< 1 year) mispricing, correspondingly, both with positive relations (Daniel et al., 2019). Mispricing errors only occur when there is an exalted sentiment, while trading frictions and information uncertainty both go up in a distress and are of equal increment regardless sentiment polarity (Avramov et al., 2019). Assets with high beta are usually related to divergent valuation opinions and tend to be overvalued (Hong & Sraer, 2016). Based on a firm's fundamental value, a factor called "value-to-price ratio" is constructed and it yields abnormal-return after risk factor adjustment relevant to portfolio style differences (Kubota et al., 2009). However, many of these factors mentioned above are constructed from financial perspective, while ours are based on user-generated contents from social media.

3.3 Twitter Sentiment and Industry Return

Another focus is investor sentiment and how tweets reflect such sentiment and whether these tweets are meaningful for return prediction. Twitter is a valuable source to analyze financial dynamics in the retail sector and the information could be as prized as these from established sources such as the *Wall Street Journal* and *Dow Jones* Newswires, if analyzed properly (Souza et al., 2015). By forming a financial community from critical tweet nodes, scholars find that Twitter sentiment is a superior proxy to predict financial market and it has significant correlation with the Dow Jones Industrial Index price and volatility series (Yang et al., 2014). Although the connection between tweet sentiment and financial markets remains equally strong whenever a major news event influences financial markets, the prediction for stock index return is tricky and ephemeral even at directional level. There is a strong, positive, and contemporaneous correlation between tweet sentiment and daily market returns, in addition to a weaker negatively correlation with next day's market returns (Liew & Budavári, 2016). Consisting with efficient market theory, these researchers also find that market returns Granger-cause next day's sentiment movements and the tweet sentiments Granger-caused the market to move in a more recent period of 2015. Sentiment indices constructed via tweet collection can predict daily stock returns and volatility jumps (Sanford, 2022). Additionally, there is evidence that positive sentiment shocks strengthen consumption, output, and interest rate, while they weaken inflation (Shapiro et al., 2022) . When there is a low sentiment proxy at the beginning of a measure period, it is to expect higher following returns for small stocks, young stocks, high volatility stocks, unprofitable stocks, non-dividend-paying stocks, extreme growth stocks and distressed stocks; while when there is a high sentiment, there should be lower subsequent returns (Baker & Wurgler, 2006). To answer the question on how to quantify investor sentiment effects, a "top down" macroeconomic approach based on behavior finance assumptions is proposed, showing that investor sentiment come in waves with nonnegligible, perceivable and regular impacts on individual firms as well as the stock market in its entirety. In accordance with their previous paper conclusion, they discover that sentiment mostly affects hard-to-arbitrage stocks or stocks with controversial values (Baker & Wurgler, 2007). There is a finding that relative sentiment correlates with relative prices of dual-listed firms, while global sentiment forecasts return at a nation level. Furthermore, sentiment and future returns of hard-to-arbitrage or hard-to-value stocks are negatively related and there is a report that private capital flows result in sentiment spreads across markets and the formation of global sentiment (Baker et al., 2012). Despite the minimal economic significance and inadequate out-of-sample prediction performance, there is statistical significance of the co-movement between intra-day volatility and Twitter sentiment and activity. High-frequency Twitter information is not particularly beneficial for active investors to predict a stock's intra-day volatility or value (Behrendt & Schmidt, 2018). Likewise, there is a finding that investor sentiment based on Twitter has negligible impact on the spread of S&P 500 Index (Guijarro et al., 2019). However, a contrary study reports that firm-initiated tweets can positively impact returns and trading volume across multiple industries included from Fortune 500 companies (Ganesh & Iyer, 2021).

Twitter sentiment is also relevant to returns of certain industry portfolios. Some news-based measure can explain stock returns of industries grouped by SIC classification (Sprenger & Welpe, 2011). In the financial services industry, only negative tweet sentiment predicts future stock prices (He et al., 2016). Similarly, tweet sentiment score significantly relates to the excess log return of sport industry, especially football clubs (Derouiche & Frunza, 2020). Tweet format does not equally influence industries and digital-sensitive industries benefit more if more hashtag and video or picture URL are used, while there are other industries such as software are digital insensitive (Han et al., 2019).

CHAPTER 4

HYPOTHESES

4.1 Hypotheses for COVID-19 Statistics

Based on the previous literature review, it is reasonable to expect some dimension in HCD is associated with essential COVID-19 statistics with statistical significance, which is the alternative hypothesis below. The null hypothesis is that none of them is relevant.

H_0: None of LTO, IVR, MAS, IDV, UAI and PDI in HCD is significantly related to the confirmed case rate or death rate of COVID-19. ($\beta_{IVR} = 0$, $\beta_{PDI} = 0$, $\beta_{LTO} = 0$, $\beta_{MAS} = 0$, $\beta_{UAI} = 0$, $\beta_{IDV} = 0$ in each regression)

H_1: At least one element among LTO, IVR, MAS, IDV, UAI and PDI is significantly related to the confirmed case rate or death rate in COVID-19.

Likewise, some Twitter sentiment index may be a meaningful predictor for COVID-19 fatality or confirmed case rate. Hence, we proposed the following:

H_2: Twitter sentiment is meaningful to predict death rate or infection rate of COVID-19.

4.2 Hypotheses for Mispricing

To see whether a stock is overvalued or undervalued, we compare its mispricing index to the median of NYSE stocks. If the index is less than the median, it is an undervalued stock. Otherwise, we call it an overvalued one. We propose the following hypotheses based on the literature review above:

We also would like to hypothesize the following statement for mispricing research:

H_0: With other conditions remaining the same, the total number of retweets, replies and likes of a tweet are negatively associated with mispricing.

This is intuitively understandable, since if a firm has much attention which is proxied by the sum of retweets, replies, and likes, then it is bound to be scrutinized by crowd wisdom from various investors, and there is less chance for the stock to be overvalued or undervalued, hence the less mispricing.

H_1: With other conditions remaining the same, the percentage of tweets with external links and mispricing are negatively related.

We assume that tweets with high proportion of links may help reduce mispricing. This is because many of these links will direct readers to product purchase pages, while the reviews and other information there could help users understand the products and the firm itself well. In other words, by making readers more informed about a firm's products, these links may indirectly help investors have a more accurate understanding of the company's market performance, hence there will be less mispricing.

H_2: With other conditions remaining the same, tweet tone positively correlates with mispricing.

H_3: With other conditions remaining the same, tweet sentiment polarity and liquidity are positively related.

Common sense of fear and greed can explain the last two hypotheses. When there are lots of very positive firm-initiated news, some investors might become too optimistic and are inclined to buy too many shares, which will lead to an overvalued price; when too much pessimism comes out of firm tweets, investors could panic and start selling off their stocks, making the market price go below the intrinsic price. Therefore, tweets with either too positive or negative sentiment may exacerbate mispricing regardless of its deviation direction. Either way, the liquidity will increase since there will be many investors selling off or purchasing this stock.

4.3 Hypotheses for Industry Return

Similarly, the above-mentioned literature shows that the sentiment index constructed from Twitter often contains valuable information, including significance, correlation, and predictive power for financial performance. Therefore, we propose the following hypotheses.

H_0: With other conditions remaining the same, sentiment index aggregated from firm tweets is a significant proxy to predict industry returns.

H_1: With other conditions remaining the same, the dispersion of sentiment aggregated from firm tweets is a significant proxy to predict industry returns. In addition to these variables above, there are other Twitter sentiment measures in literature, such as TEU derived from USA tweets, and we also would like to construct a new hybrid variable which combines both to test its prediction power for these industry returns. Based on the literature review above, we are optimistic that our new variable is an ideal predictor for the industry returns. Therefore, there could be another hypothesis.

H_2: With other conditions remaining the same, there are other possible sentiment proxies to predict industry returns.

CHAPTER 5

DATA AND METHODOLOGY

5.1 COVID-19 Statistics

Because there is no consolidated data source that integrates COVID-19 health statistics and HCD values, this study combines multiple data sources. The first data source is the most recent (December 2015) HCD values from Hofstede's website (Hofstede, 2015). There are two issues with the dataset. First, some countries are not uniquely presented. For example, Canada has two versions, a "traditional" Canada, and French Canada composed of the provinces that predominately speak French. Belgium and Switzerland are similarly subdivided in the HCD dataset. Second, Hofstede's website defines some countries in vague geographical terms. For example, the dataset cites Africa East, African West, and Arab countries. We complete this dataset with data from the Country Comparison tool on Hofstede Insights (Hofstede, 2017). The third dataset is the Coronavirus Resource Center of John Hopkins University of Medicine that provides COVID-19 infection, death, and vaccine data as of 10/25/2022 (*COVID-19 Map*, 2022). We introduce control variables, such as GDP per capita, median age, temperature, net migration rate, population density, urbanization, based on risk factors listed by Centers for Disease Control and Prevention (CDC) (CDC, 2020) and environmental health journal (Eisenberg et al., 2007). We use the logarithm of GDP per capita to mitigate possible nonlinearity issue. The GPD per capita, population density, urban population (urbanization), net migration rate, for each country is also available from the World Bank (*World Bank Group - International Development, Poverty, & Sustainability*, n.d.), and the median age by country is accessible from the

Wikipedia (“List of Countries by Median Age,” 2022). The yearly average temperature by each country is downloadable as well (*Climate Change*, 2017). There are 73 valid entries left after matching with HCD data. The legal origin is also a meaningful control variable because of its significance to a country’s economics, which in turn is possibly associated with COVID-19 statistics (La Porta et al., 2007). In fact, legal origin turned out to be statistically significant to the economic outcome, with common law being more economically promising than French civil law (J. Wang et al., 2020).

The legal origin Weighted Aggregate Score divided into four groups in terms of country and region: UK, France, Germany, and Scandinavia. To avoid dummy variable trap, we set UK as the base group in our regression analysis. In Table 1.2 we listed the available control variables used in this study. All these variables are stored in a table in the Appendix. The main statistical tool used in this study are a multilinear regression model. The culture measure included in HCD are PDI, LTO, MAS, UAI, IDV and IVR. The significance levels are set at 1%, 5% and 10%. Since each country has different population, we divide the total infection number by the population. Technically, this term is the confirmed case rate, but we use the two terms interchangeably. The definitions of variables are listed in Table 1.1. To learn how each variable perform individually and collaboratively with other variables, we implement univariate regression and multivariate regression in the following equations:

Univariate Regression

$$\text{Infection Rate} = \beta_0 + \beta_1\text{Vaccine} + \beta_2 \text{ Age} + \beta_3 \text{ GDP} + \beta_4 \text{ Temperature} + \beta_5 \text{ Migration} + \beta_6 \text{ Population} + \beta_7 \text{ Urban} + \beta_8 \text{ SC} + \beta_9 \text{ FR} + \beta_{10} \text{ DE} + \beta_{11} \text{ Urban} + \beta_{11}\text{HCD} \quad (1.1)$$

$$\text{Death Rate} = \beta_0 + \beta_1 \text{ Vaccine} + \beta_2 \text{ Age} + \beta_3 \text{ GDP} + \beta_4 \text{ Temperature} + \beta_5 \text{ Migration} + \beta_6 \text{ Population} + \beta_7 \text{ Urban} + \beta_8 \text{ SC} + \beta_9 \text{ FR} + \beta_{10} \text{ DE} + \beta_{11} \text{ Urban} + \beta_{11} \text{ HCD} \quad (1.2)$$

Where HCD represents PDI, IVR, MAS, IDV, LTO or UAI.

Multivariate Regression

$$\text{Infection Rate} = \beta_0 + \beta_1\text{Vaccine} + \beta_2 \text{ Age} + \beta_3 \text{ GDP} + \beta_4 \text{ Temperature} + \beta_5 \text{ Migration} + \beta_6\text{Population} + \beta_7 \text{ Urban} + \beta_8 \text{ SC} + \beta_9 \text{ FR} + \beta_{10} \text{ DE} + \beta_{11} \text{ Urban} + \beta_{11} \text{ PDI} + \beta_{12} \text{ IVR} + \beta_{13} \text{ MAS} + \beta_{14} \text{ IDV} + \beta_{15} \text{ LTO} + \beta_{16} \text{ UAI} \quad (1.3)$$

$$\text{Death Rate} = \beta_0 + \beta_1 \text{ Vaccine} + \beta_2 \text{ Age} + \beta_3 \text{ GDP} + \beta_4 \text{ Temperature} + \beta_5\text{Migration} + \beta_6 \text{ Population} + \beta_7 \text{ Urban} + \beta_8 \text{ SC} + \beta_9 \text{ FR} + \beta_{10} \text{ DE} + \beta_{11} \text{ Urban} + \beta_{11} \text{ PDI} + \beta_{12} \text{ IVR} + \beta_{13} \text{ MAS} + \beta_{14} \text{ IDV} + \beta_{15} \text{ LTO} + \beta_{16} \text{ UAI} \quad (1.4)$$

To avoid collinearity, we first discard any variable that exceeds 0.8 in the correlation matrix. This step eliminates a few variables such as life expectancy as well as water and sanitation. Second, we run the variance inflation factor (VIF) tests with a threshold of 5 to check HCD, since it does not matter if the variables with high VIF are control variables. These criteria mitigate multicollinearity. Both the correlation matrix and VIF test show there is no multicollinearity.

TEU-USA data is also used as a predictor to forecast the monthly death rate. We use TEU-USA as a proxy to represent Twitter sentiment and economic condition and regress both COVID-19 death rate and infection rate on this variable along with other control variables. We examine the regressions both with and without the presence of ARIMA to make sure the robustness. The equations are shown below:

$$\text{Death Rate}_{t+k} = \beta_1 \text{ Unemployment}_t + \beta_2 \text{ Real GDP}_t + \beta_3 \text{ VIX Rate}_t + \beta_4 \text{ TEU-USA} \tag{1.5}$$

Where k = 1, 2, 3, 4, 5 or 6 month(s)

$$\text{Death Rate}_{t+k} = \beta_1 \text{ Unemployment}_t + \beta_2 \text{ Real GDP}_t + \beta_3 \text{ VIX Rate}_t + \beta_4 \text{ TEU-USA} + \beta_5 \text{ ARIMA (Death Rate)}_t \tag{1.6}$$

Where k = 1, 2, 3, 4, 5 or 6 month(s)

5.2 Mispricing

Our sample period is from 2007 to 2020, and the firm characteristics are from COMPUSTAT, while the daily and monthly stock information is from CRSP, and the mispricing index is downloaded from the Stambaugh and Yuan (2017) website.

The firm list is from the S&P 1500 index. Out of 1500 firms, there are 1206 firms with official twitter accounts. There are two types of samples in our study. The main sample is acquired by matching with CRSP stock data and the mispricing score, which leaves our final sample at 938 firms from 2007 to 2016. The extended sample is acquired by following

Stambaugh, Yu, and Yuan (2015) methodology, and we self-construct the mispricing index for firms after 2016 till 2020.

The study combines multiple data resources. Our primary data source is from Twitter, on which we carefully examine over 2200 firms and more than 5500 executives. There are two types of Twitter files, one contains firm account information, and the other one has tweet information. After extracting the tweet files above, we construct a panel data table which incorporates important information from these files extracted. Each firm is represented by its Twitter ID, along with gvk and permno as keys for reference if combination with other databases is needed later. The Account Age is the number of months from a firm's first to its last tweet (as of February 2021 in this study), and each company's statistics are recorded monthly, including the number of tweets, retweets, replies, likes, URLs, photos, and videos. An industry column categorizes all firms into 48 different industries based on the Standard Industrial Classification (SIC) codes.

To examine how tweet variables relate to mispricing, we implement a set of regressions for each table as below, with the inclusion of control variables such as firm size (Size), idiosyncratic volatility (Ivol), Amihud Illiquidity (Illiq), percentage of institutional ownership (IOR) and daily max return (Max).

Regression 2.1 checks how tweet variables individually influence future mispricing. (2.1)

$$\text{Mispricing}_{t+k} = \alpha + \beta_1 \text{TV} + \beta_2 \text{Size}_t + \beta_3 \text{Illiq}_t + \beta_4 \text{Ivol}_t + \beta_5 \text{Max}_t + \beta_6 \text{IOR}_t + \beta_7 \text{ CGO}_t$$

Where TV represents one of these tweet variables, including Observation, Link, Link Percentage, Picture, Picture Percentage, Average Word, Total Word, RLR, RLR Percentage and Average Score at time t.

Regression 2.2 tests how these tweet variables collectively influence future mispricing.

$$\text{Mispricing}_{t+k} = \alpha + \beta_1 \text{ Observation}_{t} + \text{Link}_{t} + \beta_2 \text{ Link Percentage}_{t} + \beta_3 \text{ Picture}_{t} + \beta_4 \text{ Picture Percentage}_{t} + \beta_5 \text{ Average Word}_{t} + \beta_6 \text{ Total Word}_{t} + \beta_7 \text{ RLR}_{t} + \beta_8 \text{ RLR Percentage}_{t} + \beta_9 \text{ Average Score}_{t} + \beta_{10} \text{ Size}_{t} + \beta_{11} \text{ Illiquidity}_{t} + \beta_{12} \text{ Ivol}_{t} + \beta_{13} \text{ Max}_{t} + \beta_{14} \text{ IOR}_{t} + \beta_{15} \text{ CGO}_{t} \tag{2.2}$$

Regression 2.3 tests how these tweet variables collectively influence future idiosyncratic volatility.

$$\text{Idiosyncratic Volatility}_{t+k} = \alpha + \beta_1 \text{ Observation}_{t} + \text{Link}_{t} + \beta_2 \text{ Link Percentage}_{t} + \beta_3 \text{ Picture}_{t} + \beta_4 \text{ Picture Percentage}_{t} + \beta_5 \text{ Average Word}_{t} + \beta_6 \text{ Total Word}_{t} + \beta_7 \text{ RLR}_{t} + \beta_8 \text{ RLR Percentage}_{t} + \beta_9 \text{ Average Score}_{t} + \beta_{10} \text{ Size}_{t} + \beta_{11} \text{ Illiq}_{t} + \beta_{12} \text{ Ivol}_{t} + \beta_{13} \text{ Max}_{t} + \beta_{14} \text{ IOR}_{t} + \beta_{15} \text{ CGO}_{t} \tag{2.3}$$

Regression 2.4 tests how these tweet variables collectively influence future complex liquidity.

$$\text{Complex Liquidity}_{t+k} = \alpha + \beta_1 \text{ Observation}_{t} + \text{Link}_{t} + \beta_2 \text{ Link Percentage}_{t} + \beta_3 \text{ Picture}_{t} + \beta_4 \text{ Picture Percentage}_{t} + \beta_5 \text{ Average Word}_{t} + \beta_6 \text{ Total Word}_{t} + \beta_7 \text{ RLR}_{t} + \beta_8 \text{ RLR Percentage}_{t} + \beta_9 \text{ Average Score}_{t} + \beta_{10} \text{ Size}_{t} + \beta_{11} \text{ Illiq}_{t} + \beta_{12} \text{ Ivol}_{t} + \beta_{13} \text{ Max}_{t} + \beta_{14} \text{ IOR}_{t} + \beta_{15} \text{ CGO}_{t} \tag{2.4}$$

Regression 2.5 tests how these tweet variables collectively influence future inverse liquidity.

$$\text{Inverse Liquidity}_{t+k} = \alpha + \beta_1 \text{Observation}_t + \text{Link}_t + \beta_2 \text{Link Percentage}_t + \beta_3 \text{Picture}_t + \beta_4 \text{Picture Percentage}_t + \beta_5 \text{Average Word}_t + \beta_6 \text{Total Word}_t + \beta_7 \text{RLR}_t + \beta_8 \text{RLR Percentage}_t + \beta_9 \text{Average Score}_t + \beta_{10} \text{Size}_t + \beta_{11} \text{Illiq}_t + \beta_{12} \text{Ivol}_t + \beta_{13} \text{Max}_t + \beta_{14} \text{IOR}_t + \beta_{15} \text{CGO}_t \tag{2.5}$$

Regression 2.6 tests how these tweet variables collectively influence future turnover.

$$\text{Turnover}_{t+k} = \alpha + \beta_1 \text{Observation}_t + \text{Link}_t + \beta_2 \text{Link Percentage}_t + \beta_3 \text{Picture}_t + \beta_4 \text{Picture Percentage}_t + \beta_5 \text{Average Word}_t + \beta_6 \text{Total Word}_t + \beta_7 \text{RLR}_t + \beta_8 \text{RLR Percentage}_t + \beta_9 \text{Average Score}_t + \beta_{10} \text{Size}_t + \beta_{11} \text{Illiq}_t + \beta_{12} \text{Ivol}_t + \beta_{13} \text{Max}_t + \beta_{14} \text{IOR}_t + \beta_{15} \text{CGO}_t \tag{2.6}$$

Regression 2.7 tests how these tweet variables collectively influence future inverse spread.

$$\text{Inverse Spread}_{t+k} = \alpha + \beta_1 \text{Observation}_t + \text{Link}_t + \beta_2 \text{Link Percentage}_t + \beta_3 \text{Picture}_t + \beta_4 \text{Picture Percentage}_t + \beta_5 \text{Average Word}_t + \beta_6 \text{Total Word}_t + \beta_7 \text{RLR}_t + \beta_8 \text{RLR Percentage}_t + \beta_9 \text{Average Score}_t + \beta_{10} \text{Size}_t + \beta_{11} \text{Illiq}_t + \beta_{12} \text{Ivol}_t + \beta_{13} \text{Max}_t + \beta_{14} \text{IOR}_t + \beta_{15} \text{CGO}_t \tag{2.7}$$

5.3 Industry Return

The study shares a part of data from our previous Section 5.2. All the datasets in this research can be categorized into three groups: predicted variables, control variables and our proposed sentiment measures. To collect predicted variables, we first acquire five industry portfolio data set from Kenneth French's research website. The csv data set is originated from CRSP database between July 1926 and October 2021 on monthly basis.

We are interested in predicting the future return of each industry. The five industries are Consumables, Manufacturing, High Technology, Healthcare and Other, which cover the majority of everybody’s daily lives. Missing data are indicated by -99.99 or -999, which has little impact to our analyses, because of an irrelevant time frame. The data set weighs these industries in two ways: market value and equal value.

We include multiple control variables in different regression equations. Following Welch and Goyal in 2007, we include net equity expansion, inflation, and stock variance as control variables. Welch and Goyal (2007) offer the following definitions:

> **Net Equity Expansion (ntis)** This is the ratio of 12-month moving sums of net issues by NYSE listed stocks divided by the total end-of-year market capitalization of NYSE stocks.
>
> **Inflation (infl)** It refers to the Consumer Price Index, acquired from the Bureau of Labor Statistics from 1919 to 2005.
>
> **Stock Variance (svar)** It is the sum of squared daily returns on the S&P 500 from CRSP.
>
> **Industrial Growth Rate (igr)** This measure is calculated from data provided by the Federal Reserve Economic Data (FRED), from 1919 to 2021. There is an obvious increase trend over time. Therefore, we use the monthly change rate instead to avoid spurious regression results.
>
> **News Sentiment (ns)** This measure is constructed by Buckman et al. in 2020 and the data set ranges from January 1980 to August 2021 on a daily increase. We include this variable as control, as it will help us differentiate our predictor index’s influence on industry return prediction.

We also download the TEU USA data[2] to construct our hybrid sentiment variable, and only TEU USA is chosen because these tweets are USA based.

[2] The data is retrieved on November 3, 2022, from: https://www.dropbox.com/s/o4ddj33odyyz4v6/Twitter_Economic_Uncertainty.xlsx?dl=0

The methodology is described in several steps. The first step is to collect the raw data. We try to extract the S&P 1500 company tweets from Twitter and find 938 matched firm with valid tweets. We make two versions of the data for possible future implementation, one with a separate file for each firm which is named isolated firms and the other is an integrated file including all these firms called combined firms. In addition to the combined firm dataset, we collect historical data of some major market performance measures. To further reduced bias, we set the span from January in 2009 to February 2021, although there is available tweet data as early as 2006. This is because there are few firm tweets available near the inception of Twitter's creation, while the number of firms posting tweets has increased drastically since 2009. Since more firms leads to a larger sample space, the bias is likely lowered.

The second step is to compute sentiment scores. To parse a sentiment score for each tweet, we adopt a Python package called pysentiment2. Such a package has a library for sentiment analysis in dictionary framework, including McDonald Financial Sentiment Dictionary, which is a canonical dictionary for financial sentiment analysis. The score generated via this package is what we concern the most, while other parameters such as polarity and subjectivity are not so important in our study. To save storage space and run the program more efficiently, we take away some insignificant features from the combined firms. The sentiment score characterizes both the direct and magnitude of each tweet. A positive score indicates some good news, while a negative score represents the opposite. The intensity is represented by the absolute value of a score. The range of sentiment score is from -1 to 1 and an intense sentiment corresponds to a score whose absolute value is closer to 1, while a mild one is closer to 0.

Third, it is essential to acquire aggregate sentiment scores at two different levels, which are firm level and market level and there are multiple ways to aggregate sentiment scores as well. Firm level means computing sentiment index grouped by firm name. There are two ways to compute at the firm level, which we call Individual and Aggregate. For example, there are only three firms A, B and C all sending tweets in a day. Firm A sent three tweets whose scores are 0.1, 0.2 and 0.3, respectively. The sentiment score of Firm A is represented by the average of all the three tweets that day, which equals 0.2. The same approach is done for Firm B and Firm C. In Other words, each tweet represents only itself, which is why called this Individual. To work out an Aggregate, we combine all three tweets from Firm A into a single tweet, then compute the combined sentiment score. Intuitively, the individual way treats each tweet separately, while the aggregate way combined all tweets from a firm as a whole and then measures the combined sentiment of the firm that day. Similarly, there are two ways to compute the index at the market level, which we call equal weighted and size weighted. If we assume each company has the same influence on the market and the market size matters not, then it is the equal weighted approach. Otherwise, each firm's score weight is strictly determined by its size. This is a more realistic estimation of each firm's impact on the financial market, because large corporations such as Microsoft, Google, Apple, and Amazon usually are influential on the market and may cause significant shocks in the stock market, even if their stock prices only fluctuate slightly, while smaller to medium firms do not have comparable impact. To acquire a sentiment index, we first need to aggregate the data at firm level first, then at market level and each level has two methods. Therefore, by combination, there are four ways to construct the sentiment index in our study shown in the table below.

We implement a set of regressions to predict the financial index, including basic firm characteristics or other sentiment indices as the control variables mentioned previously. All these predicted returns from each industry are value weighted since it is a more realistic representation than equal value weighted method. The first regression is to find out how our sentiment index perform:

$$R_{t+k} = \alpha + \beta_1 \text{Inflation}_t + \beta_2 \text{Net Equity Expansion}_t + \beta_3 \text{Stock Variance}_t + \beta_4 \text{Industry Growth Rate}_t + \beta_5 \text{Sentiment}_t + \beta_6 \text{News Sentiment}_t \tag{3.1}$$

Where R_{t+k} represents the future return of a certain industry at time t + k, and Inflation_t, $\text{Net Equity Expansion}_t$, Stock Variance_t, $\text{Industry Growth Rate}_t$, Sentiment_t, and News Sentiment_t for Inflation, Net Equity Expansion, Stock Variance, Industry Growth Rate, one of our four sentiment proxies introduced earlier and News Sentiment, all at time t. We try to find out whether our sentiment measure is significantly relevant to the return at different time horizons.

The above methods all address aggregation structure, while there could be other ways to incorporate and interpret tweet information for industry return. Sentiment volatility measured by standard deviation of aggregate tweet sentiment score is another potentially valid answer to address our question regarding industry returns. Therefore, we attempt to use sentiment volatility as the other predictor for modeling industry returns.

The second regression aims to test how our second explanatory variable (monthly market sentiment dispersion) does in predicting next month return:

$$R_{t+k} = \alpha + \beta_1 \text{ Inflation}_t + \beta_2 \text{ Net Equity Expansion}_t + \beta_3 \text{ Stock Variance}_t + \beta_4 \text{ Industry Growth Rate}_t + \beta_5 \text{ STD}_t + \beta_6 \text{ News Sentiment}_t \quad (3.2)$$

The annotations are very similar to Regression 1, except the replacement of Sentiment $_t$ with STD $_t$, which stands for the sentiment volatility index at time t. The correlation matrix in Table 3.3 shows the correlations are only low to modest. Thus, we will not worry about multicollinearity much.

To make sure our variables are robust, we build additional models, including AutoRegressive Integrated Moving Average with eXogenous variables (ARIMAX), Long Short-Term Memory (LSTM), MultiLayer perceptron (MLP). We also consider how economic recession plays a part in our industry return prediction. Therefore, we include monthly US recession indicator. [3] These results are tabulated in the Robustness Test section.

$$R_{t+k} = \alpha + \beta_1 \text{ Inflation}_t + \beta_2 \text{ Net Equity Expansion}_t + \beta_3 \text{ Stock Variance}_t + \beta_4 \text{ Industry Growth Rate}_t + \beta_5 \text{ Sentiment}_t + \beta_6 \text{ News Sentiment}_t + \beta_7 \text{ Recession}_t + \beta_8 \text{ Recession}_t * \text{Sentiment}_t \quad (3.3)$$

$$R_{t+k} = \alpha + \beta_1 \text{ Inflation}_t + \beta_2 \text{ Net Equity Expansion}_t + \beta_3 \text{ Stock Variance}_t + \beta_4 \text{ Industry Growth Rate}_t + \beta_5 \text{ Sentiment}_t + \beta_6 \text{ News Sentiment}_t + \beta_7 \text{ Hybrid}_t \quad (3.4)$$

Where Hybrid $_t = \frac{1}{2}\left(\frac{x_t - \bar{x}_t}{\sigma x_t} + \frac{k_t - \bar{k}_t}{\sigma k_t}\right)$

[3] The dataset is available here: https://fred.stlouisfed.org/series/USREC

The new variable, Hybrid, is the average of normalized value of our sentiment variable (X) and the TEU (k), and t is measured monthly.

CHAPTER 6

RESULTS

6.1 Predicting COVID-19 Statistics

Tables 1.4 and 1.5 present the results of regressions with respect to infection rate and death rate. There are the most important findings below:

In the multivariate regression, MAS and PDI exhibit strong negative significances (P value = 0.001, 0.011 respectively) for infection rate across nations, both in statistical and economic aspects. Both variables perform well in the univariate regression (P value = 0.000, 0.003 for MAS and PDI, respectively).

In the multivariate regression, IVR and LTO show negative statistical significances (P value = 0.027, 0.090 respectively) for death rate across nations. On the other hand, in the univariate regression, IVR only shows mild statistical significance (P value = 0.094), whereas LTO alone has no significance of any kind.

These findings partially echo the conclusion that indulgence can predict COVID-19 death per capita, while PDI, LTO and IVR predict the impact of the pandemic (Lajunen et al., 2022). It is understandable that MAS associates with the infection rate, possibly because countries with masculine leaders likely take decisive measures, such as quick border closure and unilateral executive measures for emergencies (Windsor et al., 2020). Our results are also consistent with the finding that PDI is an influencing factor to ensure conformity to prescribed behaviors, which facilitates containment of COVID-19 cases (Kumar, 2021). Our results are also relevant to the conclusion that LTO and IVR affect social distancing during the pandemic (Y. Wang, 2021).

Our findings contradict some previous study results, possibly due to them being completed in earlier stage of the pandemic. For example, a study shows societies with high IDV and high PDI experience a slowed rate of pathogen multiplication, while population density is negatively related to outbreak (Messner, 2020).

The adjusted r square values mostly vary between 0.6 and 0.7 in Table 1.4, indicate the variables fit the model well, whereas most of them range between 0 and 0.04 in the death rate regression.

Tables 1.6 and 1.7 show that TEU-USA is a significant predictor for COVID-19 death rate in short run (1-3 months). However, such conclusion does not hold for the infection rate. This is not too hard to understand, since TEU-USA roughly represents the nationwide economic uncertainty, and a higher uncertainty usually have heavier impact on the elder or people with lower income, which are hit hard by the pandemic. Thus, it makes sense to use TEU-USA only for COVID-19 death rate prediction. The test results also relate well to the conclusions that Twitter users tweeting in USA are more sensitive to the change of COVID-19 daily death rate (Dyer & Kolic, 2020), and people feel strong feeling of fear when discussing COVID-19 deaths (Xue, Chen, Hu, et al., 2020).

Table 1.1 Variable Description

Variable	Definition
	Panel A: Predicted Variables
Infection	short for infection rate, which equals total confirmed cases divided by total population
Death	short for death rate, which equals total deaths divided by total confirmed cases.
	Panel B: Control Variables
Vaccine	vaccine per capita administered in a country
GDP	the logarithm of a countries GDP per capita, measured in US dollars
Migration	net migration rate, contributes to the overall level of population change in a country
Population	short for population density, measured by the number of human inhabitants per square kilometer
Age	the median age that divides a population into two numerically equally sized groups
Urban	the percentage of people living in urban without percentage sign
Temperature	yearly average temperature
UK	UK legal origin
FR	French legal origin
DE	German legal origin
SC	Scandinavian legal origin
	Panel C: Predictor Variables
PDI	This dimension expresses the degree to which the less powerful members of a society accept and expect that power is distributed unequally.
MAS	The Masculinity side of this dimension represents a preference in society for achievement, heroism, assertiveness, and material rewards for success.
IVR	Indulgence stands for a society that allows relatively free gratification of basic and natural human drives related to enjoying life and having fun.
LTO	Societies who score low on this dimension, for example, prefer to maintain time-honored traditions and norms while viewing societal change with suspicion.
UAI	The Uncertainty Avoidance dimension expresses the degree to which the members of a society feel uncomfortable with uncertainty and ambiguity. Countries exhibiting strong UAI maintain rigid codes of belief and behavior, and are intolerant of unorthodox behavior and ideas.
IDV	The high side of this dimension, called Individualism, can be defined as a preference for a loosely-knit social framework in which individuals are expected to take care of only themselves and their immediate families.
TEU-USA	Twitter Economic Uncertainty index derived from tweets generated with USA IP address

Table 1.1 elaborates the meaning of each variable. This table is divided into three panels, each one represents a particular type of variable. The sample period of each variable varies between 2007 and 2022, depending on the availability of the most recent data.

Table 1.2 Summary Statistics

	Mean	P5	P25	P50	P75	P95	STD	Min	Max	Observations
Infection	23.30	0.17	5.98	19.90	38.80	54.88	18.37	0.05	60.32	73
Death	1.75	0.14	0.54	0.86	1.79	2.71	5.17	0.08	44.56	73
Vaccine	1.74	0.55	1.37	1.82	2.30	2.56	0.68	0.19	3.25	73
Population	269.42	12.60	36.00	99.00	201.00	523.20	941.15	3.00	7919.00	73
Temperature	15.95	3.60	9.69	14.17	24.24	27.73	8.17	-2.26	29.39	73
Migration	0.34	-2.74	-0.30	0.27	1.45	3.11	2.28	-11.38	7.62	73
Urban	71.41	38.20	58.00	74.00	84.00	93.40	16.87	31.00	100.00	73
GDP	4.17	3.27	3.81	4.23	4.64	4.90	0.53	2.96	5.13	73
Age	35.99	20.16	29.30	38.40	42.80	45.06	8.29	16.90	48.60	73
LTO	45	13	26	44	62	84	24	4	100	73
PDI	61	29	44	64	74	94	21	11	100	73
IVR	47	15	29	46	66	83	23	0	100	73
MAS	49	12	40	50	63	79	20	5	100	73
IDV	43.90	15	25	37	63	80	22.97	12	91	73
UAI	67.10	33	50	68	85	95	21.03	8	100	73

Table 1.2 reports the summary statistics for independent variables, control variables and HCD related to our sample of 73 observations. The four categorical variables SC, UK, DE and FR are not included here due to their meaningless statistics in this context. P5, P25, P50, P75 and P95 represent percentiles. For example, P5 is the 5th percentile. STD is standard deviation. The exact definition of each variable is in Table 1.1. Table 1.2 shows the COVID-19 infection is widespread across nations, with a mean near a quarter and the median of roughly 20% of a country's total population on average, while roughly 1 or 2 out of 100 infected people die of the disease on average, despite the protection from vaccines.

Table 1.3 Correlation Matrix

	Infection	Death	Vaccine	Age	GDP	Temperature	Migration	Population	Urban	SC	FR	DE	UAI	IDV	MAS	PDI	IVR	LTO
Infection	1.000																	
Death	-0.245	1.000																
Vaccine	0.474	-0.193	1.000															
Age	0.731	-0.232	0.508	1.000														
GDP	0.766	-0.220	0.648	0.782	1.000													
Temperature	-0.595	0.179	-0.319	-0.706	-0.655	1.000												
Migration	0.296	-0.057	0.357	0.187	0.393	-0.281	1.000											
Population	0.078	-0.049	0.200	0.000	0.139	0.214	0.108	1.000										
Urban	0.417	-0.238	0.593	0.373	0.638	-0.345	0.229	0.132	1.000									
SC	0.216	-0.073	0.217	0.149	0.335	-0.390	0.140	-0.066	0.265	1.000								
FR	-0.156	0.161	0.015	-0.134	-0.235	0.165	-0.235	-0.105	0.130	-0.260	1.000							
DE	0.414	-0.093	0.021	0.470	0.315	-0.311	0.102	-0.073	0.001	-0.144	-0.508	1.000						
UAI	0.048	0.115	-0.025	0.207	-0.094	-0.046	-0.170	-0.305	0.046	-0.323	0.455	0.115	1.000					
IDV	0.545	-0.165	0.351	0.503	0.665	-0.631	0.370	-0.124	0.367	0.279	-0.301	0.277	-0.243	1.000				
MAS	-0.268	-0.003	-0.110	-0.094	-0.069	0.258	0.025	0.017	-0.104	-0.499	-0.020	0.217	0.029	-0.007	1.000			
PDI	-0.599	0.117	-0.470	-0.413	-0.604	0.508	-0.357	0.087	-0.333	-0.410	0.336	-0.227	0.256	-0.707	0.209	1.000		
IVR	0.073	-0.221	0.312	-0.056	0.319	0.079	0.131	-0.009	0.465	0.219	-0.081	-0.177	-0.184	0.192	0.044	-0.325	1.000	
LTO	0.427	-0.232	0.175	0.622	0.369	-0.506	0.121	0.173	0.088	-0.084	-0.118	0.436	0.073	0.209	0.005	-0.044	-0.436	1.000

Table 1.3 reports the pairwise Spearman correlation coefficient. Variables with multiple high correlation coefficients with others (at least 0.8) have been removed, including life span and water and sanitation. Variance inflation factors are calculated for each HCD, and none exceeds the acceptable VIF threshold of 5. The exact definition of each variable is in Table 1.1.

Table 1.4 Infection Rate Regressions

	Panel A: Univariate Regression						Panel B: Multivariate Regression
UAI	0.016						0.041
	(0.171)						(0.477)
LTO		-0.006					0.061
		(-0.064)					(0.708)
IVR			-0.027				-0.034
			(-0.309)				(-0.422)
IDV				0.067			0.007
				(0.740)			(0.074)
PDI					-0.273***		-0.255**
					(-3.147)		(-2.627)
MAS						-0.308***	-0.270***
						(-4.056)	(-3.513)
Vaccine	0.105	0.077	0.146	0.118	-2.464	-0.577	-2.635
	(0.035)	(0.026)	(0.048)	(0.039)	(-0.849)	(-0.216)	(-0.966)
Age	0.357	0.389	0.349	0.423	0.596*	0.203	0.231
	(0.924)	(0.986)	(0.927)	(1.151)	(1.729)	(0.623)	(0.624)
GDP	19.705***	19.320***	20.270***	17.605**	13.899**	24.636***	21.561***
	(3.040)	(2.994)	(2.965)	(2.628)	(2.290)	(4.318)	(3.369)
Temperature	-0.135	-0.138	-0.100	-0.062	0.008	0.061	0.281
	(-0.468)	(-0.447)	(-0.326)	(-0.206)	(0.029)	(0.235)	(0.954)
Migration	0.368	0.375	0.355	0.327	0.277	0.525	0.353
	(0.522)	(0.530)	(0.501)	(0.464)	(0.422)	(0.837)	(0.574)
Population	0.001	0.001	0.001	0.001	0.002	0.001	0.001
	(0.752)	(0.693)	(0.577)	(0.882)	(1.429)	(0.351)	(0.614)
Urban	-0.106	-0.103	-0.090	-0.102	-0.060	-0.118	-0.068
	(-0.831)	(-0.815)	(-0.683)	(-0.815)	(-0.512)	(-1.055)	(-0.579)
SC	6.253	6.010	6.049	6.745	3.102	-5.423	-5.535
	(0.872)	(0.834)	(0.852)	(0.946)	(0.466)	(-0.784)	(-0.791)
FR	6.458	6.754	6.448	7.299*	8.509**	6.589*	7.086*
	(1.472)	(1.664)	(1.549)	(1.778)	(2.237)	(1.831)	(1.826)
DE	10.560**	10.782**	10.392**	10.919**	9.759**	13.330***	10.707**
	(2.048)	(2.103)	(2.008)	(2.164)	(2.074)	(2.937)	(2.258)
Adjusted R squared	0.603	0.602	0.603	0.606	0.658	0.687	0.706
Observations	73	73	73	73	73	73	73

Statistical significance is reported where * is 10% significance, ** is 5% significance and *** is 1% significance or better. The t score for each variable is in the parentheses, with its regression coefficient above. UK legal origin is the base group. The exact definition of each variable is in Table 1.1. PDI and MAS are both significant in the univariate regression and multivariate regression. It is understandable that GDP plays a pivotal role in the regressions, while a society whose people know and stick to their places (high PDI) is important to fight COVID-19. Some presumably influential factors, such as temperature or vaccine, are not actually as critical as we may think, which is also in line with the fact we have learned that the constant evolution of COVID-19 and insensitivity to climate change make it a formidable foe of all humans. On the other hand, social factors may have more weight, as law origin related to DE exhibits its significance.

Table 1.5 Death Rate Regressions

	Panel A: Univariate Regression						Panel B: Multivariate Regression
UAI	0.045						0.033
	(1.120)						(0.769)
LTO		-0.052					-0.075*
		(-1.310)					(-1.725)
IVR			-0.064*				-0.093**
			(-1.704)				(-2.275)
IDV				-0.004			-0.011
				(-0.101)			(-0.238)
PDI					-0.011		-0.022
					(-0.265)		(-0.452)
MAS						-0.019	-0.002
						(-0.497)	(-0.055)
Vaccine	-0.295	-0.443	-0.204	-0.347	-0.449	-0.386	-0.464
	(-0.021)	(-0.332)	(-0.154)	(-0.257)	(-0.320)	(-0.286)	(-0.334)
Age	-0.276	-0.123	-0.284*	-0.214	-0.202	-0.222	-0.227
	(-1.605)	(-0.703)	(-1.718)	(-1.291)	(-1.213)	(-1.349)	(-1.205)
GDP	3.760	2.028	4.983	3.038	2.702	3.250	5.104
	(1.306)	(0.710)	(1.662)	(1.007)	(0.921)	(1.129)	(1.566)
Temperature	0.026	2.028	0.114	0.034	0.044	0.050	0.046
	(0.206)	(-0.201)	(0.854)	(0.250)	(0.335)	(0.382)	(0.303)
Migration	0.098	0.132	0.067	0.113	0.106	0.120	0.072
	(0.313)	(0.422)	(0.217)	(0.356)	(0.336)	(0.378)	(0.228)
Population	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.331)	(0.569)	(-0.578)	(0.003)	(0.080)	(-0.032)	(0.215)
Urban	-0.116**	-0.107*	-0.076	-0.106*	-0.104*	-0.107*	-0.069
	(-2.053)	(-1.928)	(-1.310)	(-1.873)	(-1.831)	(-1.892)	(-1.154)
SC	2.804	1.623	2.231	2.282	2.200	1.610	1.090
	(0.880)	(0.509)	(0.716)	(0.710)	(0.684)	(0.461)	(0.306)
FR	2.845	3.764**	2.964	3.645*	3.751**	3.670**	2.202
	(1.460)	(2.097)	(1.623)	(1.971)	(2.041)	(2.021)	(1.115)
DE	1.792	2.754	1.468	2.273	2.245	2.446	1.328
	(0.782)	(1.214)	(0.647)	(1.000)	(0.987)	(1.069)	(0.550)
Adjusted R squared	0.009	0.016	0.035	-0.011	-0.010	-0.007	0.035
Observations	73	73	73	73	73	73	73

Statistical significance is reported where * is 10% significance, ** is 5% significance and *** is 1% significance or better. The t score for each variable is in the parentheses, with its regression coefficient above. UK legal origin is the base group. Panel A represents a univariate regression for each culture measure in HCD, along with control variables, while Panel B represents the regression including all 6 culture measures in HCD. The exact definition of each variable is in Table 1.1. IVR is significant both in the univariate regression and multivariate regression. Interestingly, GDP is no longer significant for the death rate, while urbanization percentage has taken its place to some extent.

Table 1.6 Death Rate Regressions and Twitter Economic Uncertainty excluding ARIMA variables

	k = 1	**k = 2**	**k = 3**	**k = 4**	**k = 5**	**k = 6**
Unemployment	0.001	0.002	0.004***	0.005***	0.005***	0.004***
	(0.800)	(1.033)	(2.906)	(4.256)	(3.959)	(3.380)
Real GDP	0.000	0.002	0.002	0.002	0.004**	0.005**
	(0.005)	(0.902)	(0.902)	(1.049)	(2.352)	(2.829)
VIX Rate	0.001	0.002	0.001	0.003	0.003	0.002
	(0.585)	(0.791)	(0.230)	(1.447)	(1.690)	(0.778)
TEU-USA	0.013***	0.009***	0.003	0.000	0.002	0.002
	(5.031)	(3.514)	(1.102)	(0.019)	(0.721)	(0.872)
Adjusted R-squared	0.828	0.743	0.691	0.747	0.791	0.764
Observations	25	24	23	22	21	20

***, ** and * denote significance at 1, 5 and 10 percent respectively. The t statistics are in parentheses, with its regression coefficient above. k refers to prediction length measured in month. This model does not consider possible ARIMA variables. VIX rate is the change of monthly Volatility Index downloaded from Yahoo! Finance. Real GDP has taken into consideration of inflation, which is converted from quarterly to monthly. Unemployment rate is recorded monthly from FRED. TEU-USA is the Twitter Economic Uncertainty index derived from tweets recorded with USA IP address.

Table 1.7 Death Rate Regressions and Twitter Economic Uncertainty including ARIMA variables

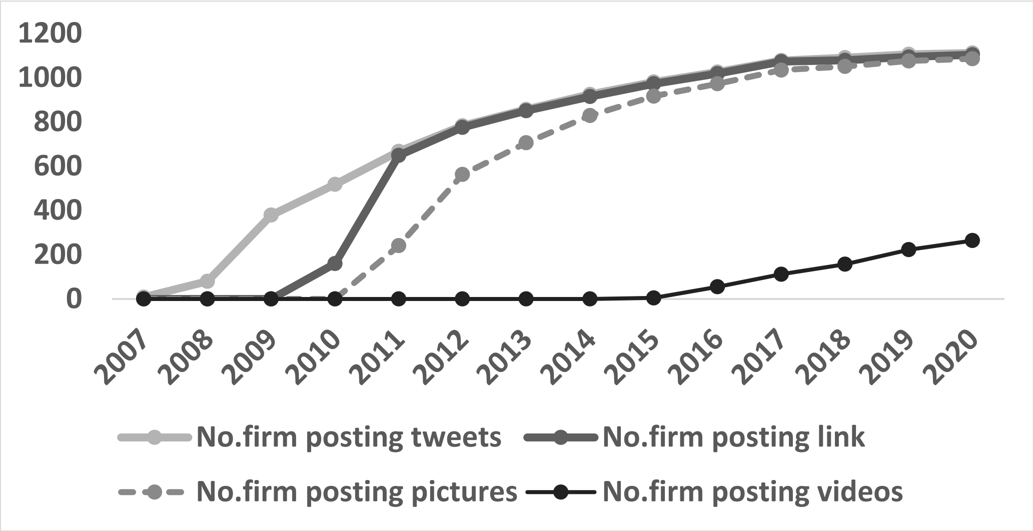
	k = 1	**k = 2**	**k = 3**	**k = 4**	**k = 5**	**k = 6**
Unemployment	0.003	0.000*	-0.008***	0.000	0.001	-0.002
	(1.048)	(-2.019)	(-3.006)	(-0.055)	(0.410)	(-0.488)
Real GDP	0.002	-0.001	-0.004**	-0.001	0.002	0.002
	(0.949)	(-0.574)	(-2.345)	(-0.446)	(1.083)	(0.881)
VIX Rate	0.001	-0.001	-0.003*	0.001	0.001	0.000
	(0.597)	(-0.649)	(-1.781)	(0.786)	(0.951)	(0.043)
TEU-USA	0.009***	0.010***	0.006**	0.000	0.001	0.003
	(4.014)	(4.016)	(2.761)	(0.129)	(0.315)	(1.018)

ARIMA-Death	-0.005***	0.000	0.004**	0.000	-0.001	0.001
	(-3.090)	(-0.135)	(2.242)	(0.301)	(-0.277)	(0.573)
ARIMA-Infection	0.005***	0.001	-0.001	0.000	0.001	0.000
	(3.575)	(0.896)	(-1.495)	(0.073)	(0.577)	(-0.223)
Adjusted R-squared	0.907	0.864	0.881	0.808	0.816	0.810
Observations	25	24	23	22	21	20

***, ** and * denote significance at 1, 5 and 10 percent respectively. The t statistics are in parentheses, with its regression coefficient above. k refers to prediction length measured in month. This model considers the possible ARIMA variables with respect to the previous death rate and infection rate. VIX rate is the change of monthly Volatility Index downloaded from Yahoo! Finance. Real GDP has taken into consideration of inflation, which is converted from quarterly to monthly. Unemployment rate is recorded monthly from FRED. TEU-USA is the Twitter Economic Uncertainty index derived from tweets recorded with USA IP address.

6.2 Forecasting Stock Mispricing

Figure 2.1 Trends of Various Contents from Firm Tweets



The above chart shows how each content develops over time. Figure 2.1 illustrates the trends for the usage of tweets with different features, including tweet texts, links, pictures, and videos. Only tweet texts were available near the creation of Twitter, and links became prevalent in 2009, followed by pictures prevalence one year later. These features all have similar trends a few years after their adoptions, which make sense, as the tweets of many firms combine texts, links, and pictures for better promotional or customer engagement purposes. However, the embedding of videos Weighted Aggregate Score unavailable until 2015. The usage of videos has been increasing steadily each year since its creation; however, it is still not nearly as prevalent as links, texts, or pictures. This is perhaps because videos are more time-consuming to make and upload, or some customers do not bother watching them due to limited bandwidth.

Compared to other statistics, the number of videos is much lower. It is possibly because it requires more effort to create an original video than some textual information.

In addition, we try to regress mispricing on these different tweet categories. However, the statistics are insignificant. Unfortunately, none of these tweet variables offer predictability for future stock return, which is consistent with previous findings that sentiment does not predict short term stock return well (Cliff & Brown, 2001).

Table 2.1 Variable Description

Variable	Description
Panel A: Monthly tweet-related variables	
Observation	Total number of tweets
Link	Total number of tweets with external links
Link Percentage	The percentage of tweets with external links, which equals Links divided by Observations
Picture	Total number of tweets with pictures
Picture Percentage	The percentage of tweets with pictures, which equals Picture divided by Observations
Video	Total number of tweets with videos
Video Percentage	The percentage of tweets with videos, which equals Video divided by Observations
Average Word	Average number of words per tweet
Total Word	Total number of words, which equals Observation multiplied by Average Word and divided by 1000
Average Score	Average positive-negative score of tweets, ranging from -1 to +1, where score is based on 2020 Master Dictionary and Sentiment Word Lists on a software repository from University of Notre Dame
Average Retweet	Average number of retweets per tweet
Average Reply	Average number of replies per tweet
RLR	To calculate RLR, first sum retweet, like, and reply, then divide it by 1000
RLR Percentage	To calculate RLR Percentage, first sum retweet, like, and reply, then divide it by Observation
Average Like	Average number of like per tweet
Panel B: Firm/stock variables	
Size	Firm size
Ivol	Idiosyncratic volatility from CRSP
Illiq	Amihud illiquidity constructed from daily CRSP
Max	Monthly max return for a given firm
IOR	Monthly institutional holding ratio

CGO	Monthly capital gain overhang, following Grinblatt and Han (2005) equation (11)
Mispricing	Mispricing score constructed by Stambaugh and Yuan (2017), downloaded from authors' website
Turnover	Monthly total number of trading shares divided by outstanding shares
Bid-ask Spread	Bid-ask Spread = (Ask price minus bid price)/monthly closing price

Table 2.1 elaborates each variable. The firm list is from the S&P 1500 index. Out of 1500 firms, there are 1206 firms with official twitter accounts.

Main sample: after matching the CRSP stock data with the mispricing score, our final sample is 938 firms from 2007 to 2016.

Extended sample: following Stambaugh, Yu, and Yuan (2015) methodology, we construct a mispricing index for firms after 2016 till 2020.

Table 2.1 consists of Panel A and Panel B. Panel A analyzes tweet variables, mostly based on monthly statistics. We aggregate different features in tweets on monthly basis in Panel A, including text, link, picture, video, word, polarity, retweet, reply and like. Panel B addresses variables related to firms or stocks, including control variables, such as size, idiosyncratic volatility, illiquidity measures, along with a few financial performance indicators. The sample period is from 2007 to 2020. Firm characteristics are from COMPUSTAT, daily and monthly stock information is from CRSP, mispricing index is downloaded from the Stambaugh and Yuan (2017) website.

In our study, the sentiment polarity characterizes the direction of each tweet, which is measured by a score. A positive score indicates an optimistic opinion, or that something is desirable, while a negative score represents the opposite sentiment. The intensity is represented by the absolute value of a score. The range of a sentiment score is from -1 to 1, and an intense sentiment corresponds to a score whose absolute value is closer to 1, while a mild sentiment is closer to 0. Thus, tweets with strong polarity either express a very positive sentiment or a considerably negative opinion of a financial asset, and either one may signal investors to buy or sell it, which facilitates liquidity.

Table 2.2 Summary Statistics

	Mean	P25	Median	P75	Standard Deviation
Panel A: Tweet-related variables					
Observation	164.989	16.086	36.321	86.029	870.184
Link	57.295	9.653	21.663	49.589	180.594
Picture	21.454	6.611	13.385	26.422	33.866
Video	4.102	1.500	2.427	4.365	5.788
Word	16.909	15.087	16.895	18.624	3.053
Score	0.172	0.114	0.178	0.237	0.108
Average Retweet	5.177	0.578	1.123	2.322	44.632
Average Reply	1.108	0.122	0.222	0.467	10.941
Average Like	16.526	1.222	2.220	5.316	192.251
Panel B: Firm variables					
Size	22.244	21.090	22.039	23.273	1.524
Ivol	0.015	0.011	0.014	0.018	0.005
Illiq	0.002	0.000	0.000	0.002	0.005
Max	0.041	0.031	0.039	0.049	0.013
CGO	0.065	-0.005	0.117	0.211	0.259
Score	46.071	39.067	45.118	52.603	9.627
Turnover	2.035	1.239	1.664	2.393	1.310
Bid-ask Spread	0.065	0.028	0.045	0.081	0.061

Table 2.2 reports the summary statistics of tweet-related variables and firm characteristics. We compute the mean of each variable for each firm from March 2007 to December 2020 and then report the distribution across 938 firms. P25 and P75 refer to the 25th percentile and 75th percentile respectively. The total number of monthly tweets has the highest variance, and there is a significant difference between the 1st and 3rd quartile. This is not surprising, since firms weigh Twitter differently when it comes to promotional and customer engagement purposes. Some firms only have a few hundred followers, and they are less likely to spend much effort on customer interactions or promotions on Twitter; whereas, large corporations, such as Amazon and Microsoft, typically have many more followers who tend to use Twitter as a separate information dissemination channel along with their own official websites, which results in many more tweets. This effect is even more significant for many airline companies, such as United Airlines and Delta with over a million tweets each. The same idea is likely true for other variables, including total number of links, replies and likes. Consistent with Figure 2.1, the usage of video is very limited, compared to the total number of tweets, links, and pictures monthly, regardless of measures. In Panel B, the mispricing has much less variation across quartiles, compared to these measures in Panel A, and so are turnover and spread.

Table 2.3 Correlation Matrix

	Mispricing	Idiosyncratic Volatility	Complex Liquidity	Observation	Link	Link Percentage	Picture	Picture Percentage	Video	Video Percentage	Average Word	Total Word	Average Retweet	Average Reply	Average Like	Average Score
Mispricing $_{t+1}$	1.000															
Idiosyncratic Volatility$_{t+1}$	0.142	1.000														
Complex Liquidity$_{t+1}$	-0.025	0.191	1.000													
Observation	-0.058	0.001	0.123	1.000												
Link	-0.079	-0.019	0.106	0.769	1.000											
Link Percentage	-0.005	-0.054	-0.074	-0.136	0.112	1.000										
Picture	-0.130	-0.026	0.105	0.279	0.436	0.045	1.000									
Picture Percentage	-0.055	-0.009	0.019	-0.067	0.016	0.127	0.492	1.000								
Video	-0.051	-0.007	0.065	0.152	0.144	-0.026	0.295	0.150	1.000							
Video Percentage	-0.025	-0.006	0.031	-0.011	0.000	0.002	0.105	0.154	0.606	1.000						
Average Word	0.065	-0.029	-0.005	0.022	-0.039	-0.223	-0.129	-0.142	-0.019	-0.025	1.000					
Total Word	-0.050	-0.004	0.117	0.982	0.762	-0.136	0.251	-0.070	0.156	-0.012	0.075	1.000				
Average Retweet	-0.039	-0.004	0.173	-0.006	0.010	0.016	0.124	0.154	0.075	0.079	-0.056	-0.010	1.000			
Average Reply	-0.027	0.011	0.195	0.017	0.020	-0.003	0.079	0.073	0.047	0.044	-0.027	0.013	0.804	1.000		
Average Like	-0.023	0.007	0.138	0.002	0.011	0.006	0.145	0.185	0.132	0.127	-0.060	-0.003	0.880	0.687	1.000	
Average Score	-0.015	0.000	0.028	0.019	-0.003	-0.056	0.019	0.025	0.017	0.020	0.089	0.023	-0.007	0.000	0.005	1.000

Table 2.3 reports correlations of pairwise tweet variables. Most variables are not significantly correlated. However, the correlation matrix shows that Average Retweet and Average Like are highly correlated (0.880). Also, Average Retweet and Average Reply are very relevant (0.804), as well as Link and Total Word (0.762). These high correlations are not surprising, as readers tend to retweet and reply more, if a tweet is very popular, and there will be fewer retweets or replies otherwise. Similarly, a tweet with many likes is likely to receive more replies. In addition, since many firms created their Twitter accounts for promotional and customer service purposes, it is expected that there will be many external links embedded in their tweets. Again, it makes sense that tweets with many likes tend to gain much attention and are a discussion topic for a while, while ordinary tweets without exciting news usually will not attract attention from many users. Most significantly, Observation and Total Word have a nearly perfect correlation (0.982), which can be explained by their exact mathematical relation. This is because we define the function under Table 2.4.

Table 2.4 Univariate Regression

	Panel A: Overvalued Stocks									
Observation	-0.001***									
	(-4.641)									
Link		-0.007***								
		(-3.663)								
Link Percentage			-0.951***							
			(-2.798)							
Picture				-0.016						
				(-0.952)						
Picture Percentage					1.202					
					(0.903)					
Average Word						-0.002				
						(-0.065)				
Total Word							-0.092***			
							(-3.620)			
RLR								-0.366***		
								(-3.406)		
RLR Percentage									-30.413***	
									(-3.648)	
Average Score										0.824**
										(2.157)
Size	-0.516***	-0.517***	-0.603***	-0.541***	-0.578***	-0.589***	-0.516***	-0.496***	-0.510***	-0.582***
	(-5.735)	(-5.511)	(-6.532)	(-6.399)	(-6.191)	(-6.541)	(-5.712)	(-5.529)	(-5.602)	(-6.285)
Illiq	-47.451	-47.577	-51.316	-50.326	-48.638	-50.388	-46.847	-44.993	-43.736	-48.405
	(-1.493)	(-1.485)	(-1.621)	(-1.563)	(-1.570)	(-1.597)	(-1.478)	(-1.380)	(-1.409)	(-1.559)
Ivol	75.532***	74.846***	68.322***	71.391***	69.165***	69.351***	74.611***	78.789***	74.238***	68.961***
	(6.923)	(6.942)	(6.331)	(6.654)	(6.249)	(6.416)	(6.777)	(7.220)	(6.794)	(6.474)
Max	2.135	2.467	2.893	2.874	2.667	2.544	2.148	2.063	2.285	2.637
	(0.519)	(0.608)	(0.717)	(0.721)	(0.667)	(0.632)	(0.520)	(0.511)	(0.569)	(0.654)
IOR	-1.060***	-1.027***	-1.165***	-1.121***	-1.216***	-1.195***	-1.079***	-1.158***	-1.163***	-1.215***
	(-4.212)	(-4.004)	(-4.929)	(-4.591)	(-4.868)	(-5.098)	(-4.300)	(-4.580)	(-4.689)	(-4.996)
CGO	-0.722	-0.750	-0.746	-0.755	-0.735	-0.729	-0.724	-0.732	-0.703	-0.739
	(-1.256)	(-1.311)	(-1.327)	(-1.356)	(-1.309)	(-1.300)	(-1.257)	(-1.275)	(-1.232)	(-1.319)
Constant	68.127***	68.215***	70.594***	68.798***	69.492***	69.769***	68.133***	67.686***	68.021***	69.486***
	(34.062)	(32.952)	(34.569)	(36.439)	(33.508)	(33.519)	(33.861)	(34.224)	(34.280)	(33.751)
Observations	14,905	14,905	14,905	14,905	14,905	14,905	14,905	14,905	14,905	14,905
R squared	0.065	0.067	0.064	0.064	0.063	0.062	0.064	0.065	0.064	0.064
Months	66	66	66	66	66	66	66	66	66	66

Table 2.4 report univariate regression for each explanatory Twitter variable constructed. Observation, Link, Link Percentage, Total Word, RLR, RLR Percentage and Average Score all exhibit statistical significance to overvalued stocks, most of these variables have a negative association, indicating they are helpful to mitigate overvalued mispricing.
To calculate RLR, first sum retweet, like, and reply, then divide it by 1000, To calculate RLR Percentage, first sum retweet, like, and reply, then divide it by Observation, Total Word = Observation * Average Word/1000, all sum variables are winsorized at 1% on large side at each month. We apply Fama-Macbeth cross sectional regression from 2011 July, with t-statistics adjusted for Newey West with lag of 6.

Some twitter variables can affect future mispricing, however, there are asymmetric effects for overvalued (Panel A) and undervalued stocks (Panel B). By applying univariate regression, the number of tweets, the number and percentage of links, total number of words, the tweet population can help reduce the mispricing index for overvalued stocks, for instance, reduce mispricing. By contrast, the average sentiment score can increase the mispricing index to make stocks more overvalued.

For undervalued stocks, results are mixed. Only the average number of words per tweet can increase the mispricing index, for example, reduce mispricing. Other variables even worsen the already mispriced stocks, such as the number of tweets, the number of links, and total number of words.

We experiment with univariate models with the corresponding explanatory variable from Model 1 to 10, and find that Observation, Link, Total Word show strong statistical significances for both overvalued and undervalued stocks. However, some predictors have different prediction power. Link Percentage and Average Score are only statistically significant for overvalued stocks, while word is only significant for undervalued stocks. Observation shows strong significance both for undervalued and overvalued stocks, which is understandable, since when there are many tweets relevant to a stock, it is possible that some investors find its value mispriced, either being overvalued or undervalued. However, total number of tweets has very weak economic significance in both mispricing cases, since the corresponding coefficients have small magnitudes. This is possibly because total number of tweets is a very broad measure which does not correlate with mispricing very much. Similarly, the total number of tweets with external links also exhibit its significance at both aspects of mispricing, with stronger economic significance. The same explanation possibly holds since link and mispricing are negatively related, despite their weak magnitude. Intuitively, more tweets with external links lead to even worse mispricing, possibly due to these external links being promotional related, possibly with unhelpful or even misleading information to investors. Because Total Word and Observation are related via a precise mathematical function, it is not surprising that the former has similar significance results in both mispricing directions. Another significant variable RLR helps decrease mispricing, possibly because a stock with many likes, retweets and replies will bring much attention of the investors, which helps reduce mispricing because the stock will be scrutinized by the wisdom of crowds.

	Panel B: Under-valued Stocks									
Observation	-0.001***									
	(-3.128)									
Link		-0.003**								
		(-2.518)								
Link Percentage			0.069							
			(0.286)							
Picture				-0.008						
				(-0.442)						
Picture Percentage					-4.668					
					(-1.100)					
Average Word						0.131***				
						(5.286)				
Total Word							-0.058**			
							(-2.631)			
RLR								-0.122*		
								(-1.794)		
RLR Percentage									-1.928	
									(-0.495)	
Average Score										0.182
										(0.769)
Size	-0.537***	-0.525***	-0.556***	-0.499***	-0.551***	-0.565***	-0.538***	-0.522***	-0.566***	-0.556***
	(-8.807)	(-8.273)	(-9.280)	(-8.435)	(-9.213)	(-9.213)	(-8.897)	(-9.911)	(-10.134)	(-9.169)
Illiq	-19.739	-20.039	-19.974	-23.491	-20.979	-20.720	-19.608	-18.076	-19.881	-18.641
	(-1.413)	(-1.439)	(-1.385)	(-1.653)	(-1.514)	(-1.444)	(-1.408)	(-1.315)	(-1.433)	(-1.346)
Ivol	-37.679**	-36.180**	-40.841**	-33.702*	-41.223**	-38.644**	-37.716**	-37.749**	-41.284**	-40.609**
	(-2.226)	(-2.019)	(-2.302)	(-1.739)	(-2.300)	(-2.160)	(-2.264)	(-2.147)	(-2.342)	(-2.284)
Max	12.504***	12.334***	12.803***	11.899**	12.953***	12.914***	12.470***	12.594***	12.890***	12.830***
	(2.846)	(2.758)	(2.960)	(2.540)	(2.911)	(2.941)	(2.867)	(2.953)	(2.994)	(2.941)
IOR	0.612	0.567	0.597	0.632	0.582	0.552	0.614	0.603	0.619	0.607
	(1.183)	(1.107)	(1.131)	(1.221)	(1.093)	(1.053)	(1.181)	(1.153)	(1.172)	(1.177)
CGO	-1.658***	-1.697***	-1.638***	-1.702***	-1.654***	-1.649***	-1.663***	-1.644***	-1.666***	-1.662***
	(-3.062)	(-3.084)	(-2.978)	(-3.250)	(-2.993)	(-2.995)	(-3.082)	(-3.035)	(-3.046)	(-3.032)
Constant	49.123***	48.994***	49.456***	48.469***	49.623***	48.012***	49.149***	48.770***	49.726***	49.487***
	(31.256)	(30.245)	(33.347)	(31.240)	(32.271)	(27.360)	(31.540)	(34.617)	(33.718)	(31.314)
Observations	25,766	25,766	25,766	25,766	25,766	25,766	25,766	25,766	25,766	25,766
R squared	0.046	0.049	0.046	0.053	0.048	0.048	0.046	0.046	0.045	0.046
Months	66	66	66	66	66	66	66	66	66	66

Table 2.5 Multivariate Regression

Variable	**Forecast Length**					
	Panel A: Overvalued Stocks					
	k = 1	**k = 2**	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Link	-0.001	-0.002	-0.003	-0.006	-0.005	-0.002
	(-0.475)	(-0.471)	(-0.748)	(-1.063)	(-0.846)	(-0.306)
Link Percentage	**-1.805*****	**-1.906*****	**-1.877*****	**-1.771****	**-1.643***	**-2.085****
	(-5.565)	**(-4.910)**	**(-4.027)**	**(-2.599)**	**(-1.873)**	**(-2.067)**
Picture	-0.071	-0.082	-0.085	-0.088	-0.091	-0.127
	(-1.587)	(-1.349)	(-1.230)	(-1.202)	(-1.113)	(-1.462)
Picture Percentage	2.64	3.134	2.941	3.317	4.648	5.766
	(-0.918)	(-0.827)	(-0.689)	(-0.675)	(-0.788)	(-0.884)
Average Word	**-0.065****	**-0.053***	-0.05	-0.037	-0.007	-0.013
	(-2.303)	**(-1.816)**	(-1.504)	(-0.885)	(-0.124)	(-0.195)
Total Word	-0.092	-0.116*	-0.120*	-0.118	-0.126	-0.158
	(-1.434)	(-1.863)	(-1.838)	(-1.383)	(-1.167)	(-1.190)
RLR	-0.07	-0.062	-0.001	0.212	0.331	0.296
	(-0.557)	(-0.493)	(-0.008)	(-0.905)	(-0.805)	(-0.583)
RLR Percentage	**-24.904****	**-28.903*****	**-33.160*****	**-39.083*****	**-35.592****	-23.973
	(-2.410)	**(-3.258)**	**(-5.048)**	**(-4.456)**	**(-2.598)**	(-1.298)
Average Score	**0.843****	**1.100*****	**1.160*****	**1.356****	1.106	0.538
	(-2.278)	**(-2.831)**	**(-2.867)**	**(-2.476)**	-1.631	-0.557
Size	-0.425***	-0.387***	-0.339***	-0.229	-0.222	-0.208
	(-4.794)	(-3.727)	(-2.755)	(-1.270)	(-1.018)	(-0.905)
Illiq	-48.471	-65.608*	-73.5	-67.834	-77.202	-26.273
	(-1.563)	(-1.756)	(-1.619)	(-1.122)	(-1.108)	(-0.430)
Ivol	78.897***	127.102***	134.218***	118.562***	112.058***	104.224***
	(-7.197)	(-9.393)	(-8.142)	(-4.784)	(-3.897)	(-3.762)
Max	3.023	-15.148***	-18.055***	-15.512***	-15.035***	-19.633***
	(-0.792)	(-3.370)	(-4.088)	(-2.888)	(-3.486)	(-3.470)
IOR	-1.125***	-1.246***	-1.330***	-1.688***	-1.605***	-1.849***
	(-4.383)	(-4.629)	(-4.424)	(-4.494)	(-3.087)	(-3.027)
CGO	-0.76	-0.564	-0.239	0.673	1.474***	2.515***
	(-1.354)	(-1.067)	(-0.486)	(-1.581)	(-3.333)	(-6.372)
Constant	68.198***	67.006***	65.594***	61.997***	60.114***	59.370***
	(-32.84)	(-25.812)	(-21.204)	(-14.787)	(-12.17)	(-11.128)
Observations	14,905	14,565	14,241	13,306	12,436	11,579

R-squared	0.102	0.101	0.098	0.089	0.092	0.098
Months	66	65	64	61	58	55

We also would like to stretch the previous one-month duration up to a twelve-month horizon with more predictors being accounted for, to check if these predictors altogether have more persistent influence on mispricing. Therefore, Table 2.5 reports the mispricing regression results of next-k-months mispricing with controls for firm characteristics, where k goes up to 12 months. The sample period is from March 2007 to December 2016. The control variables are Size, Illiq, Ivol, Max, IOR and CGO. Again, we discuss mispricing in two directions, including overvalued stocks in Panel A and undervalued stocks in Panel B.

For overvalued stocks, Link Percentage and RLR Percentage can reduce mispricing index, i.e., reduce mispricing. Word can reduce mispricing as well, but its effect is only for the near future, where k =1 and 2. Sentiment score is significantly positive, i.e., makes overvalued stock further from the intrinsic value. We apply Fama-Macbeth cross sectional regression from 2011 July, with t-statistics adjusted for Newey West with lag of 6.

Overall, we show that the impact of percentage tweet with external links has persistent significance throughout our test horizon. RLR Percentage performs similarly to Link Percentage, except for its lack of significance at k = 12. However, the predictive power of total number of words is short-lived. Intuitively, for overvalued stocks, more tweets with embedded external links help decrease mispricing. More retweets, replies and likes along with average number of words have similar but weakened influences. On the contrary, sentiment score has the opposite effect, where a positive sentiment tends to worsen already overvalued prices. What is worse, as the predicted period lengthened, we find an increasing trend. This could be due to investor judgement being carried away by too positive comments of an already overvalued stock.

Variable	Forecast Length					
	Panel B: Undervalued Stocks					
	k = 1	k = 2	k = 3	k = 6	k = 9	k = 12
Link	0.001	0.000	-0.000	-0.003	-0.002	-0.004*
	(0.310)	(0.077)	(-0.131)	(-1.175)	(-1.362)	(-1.765)
Link Percentage	0.011	0.177	0.273	0.642*	0.765**	0.659
	(0.038)	(0.588)	(0.886)	(1.858)	(2.030)	(1.660)
Picture	0.002	0.010	0.003	0.020	-0.017	0.006
	(0.066)	(0.326)	(0.097)	(0.500)	(-1.541)	(0.229)
Picture Percentage	-4.381	-3.361	-0.296	0.867	8.557	8.372
	(-1.039)	(-1.104)	(-0.582)	(0.512)	(0.960)	(0.969)
Average Word	**0.113*** **	**0.121*** **	**0.126*** **	**0.123*** **	**0.134*** **	**0.151*** **
	(5.974)	**(5.301)**	**(4.640)**	**(3.254)**	**(3.044)**	**(3.465)**
Total Word	-0.070	-0.059	-0.050	-0.024	-0.033	-0.057
	(-0.974)	(-0.820)	(-0.729)	(-0.408)	(-0.643)	(-1.148)
RLR	-0.105	-0.133	-0.132	-0.077	-0.021	0.038
	(-1.165)	(-1.119)	(-1.079)	(-0.835)	(-0.328)	(0.451)
RLR Percentage	1.851	6.356	7.378	8.438	6.274	7.604
	(0.400)	(1.220)	(1.301)	(1.279)	(0.993)	(1.449)
Average Score	0.125	0.204	0.323	0.315	0.596	1.028**
	(0.508)	(0.699)	(0.966)	(0.785)	(1.380)	(2.405)
Size	-0.515***	-0.544***	-0.560***	-0.661***	-0.752***	-0.912***
	(-9.755)	(-9.103)	(-8.035)	(-5.795)	(-5.221)	(-5.550)
Illiq	-27.558*	-33.718**	-48.118**	-69.556***	-71.064**	-88.044***
	(-1.795)	(-2.041)	(-2.616)	(-2.715)	(-2.555)	(-2.822)
Ivol	-30.830	39.135*	52.836***	36.875*	43.805	25.554
	(-1.642)	(1.940)	(2.717)	(1.748)	(1.432)	(0.731)
Max	11.403**	-12.518**	-11.914**	-3.396	-0.202	4.007
	(2.591)	(-2.596)	(-2.539)	(-0.808)	(-0.035)	(0.652)
IOR	0.678	0.655	0.612	0.422	0.501	0.587
	(1.346)	(1.230)	(1.091)	(0.679)	(0.742)	(0.926)
CGO	-1.769***	-1.644***	-1.273**	-0.083	1.040	2.155**
	(-3.395)	(-3.096)	(-2.372)	(-0.129)	(1.376)	(2.293)
Constant	47.405***	48.045***	48.337***	51.036***	52.957***	56.811***
	(30.493)	(27.652)	(24.449)	(16.390)	(13.768)	(13.271)
Observations	25,766	25,279	24,781	23,309	21,835	20,398
R-squared	0.072	0.073	0.072	0.073	0.076	0.081
Months	66	65	64	61	58	55

Generally, for undervalued stocks, tweets have less impact. Therefore, there are fewer findings. This conclusion is consistent, regardless of univariate or multivariate regressions. Only average number of words can significantly increase mispricing index.

Only the average number of words makes a difference here throughout the prediction horizon, while the other variables have no significant effect. One possible explanation is that, unlike overvalued stocks, there is relatively less attention to undervalued stocks, which means there are fewer related retweets, likes, replies, links, or other comments to cloud investor judgement, resulting in less mispricing. However, tweets with lower average number of words tend to make undervalued stocks even less appreciated.

Table 2.6 Idiosyncratic Volatility and Liquidity Regression

	Panel A: Overvalued Stocks		**Panel B: Undervalued Stocks**	
	Idiosyncratic Volatility	**Liquidity**	**Idiosyncratic Volatility**	**Liquidity**
Link	0.000	0.001	0.000	-0.000**
	(0.890)	(1.248)	(1.118)	(-2.060)
Link Percentage	-0.000	-0.019	0.010	-0.020
	(-0.004)	(-0.907)	(0.496)	(-1.282)
Picture	-0.001	0.001	**0.002****	**0.003*****
	(-0.323)	(0.545)	**(2.138)**	**(2.771)**
Picture Percentage	0.143	0.006	0.347	-0.572
	(0.888)	(0.095)	(1.139)	(-0.853)
Average Word	**-0.011*****	**-0.006*****	-0.001	**-0.003***
	(-4.389)	**(-3.105)**	(-0.542)	**(-1.979)**
Total Word	-0.005	-0.009	0.006*	0.008
	(-0.972)	(-1.389)	(1.733)	(1.447)
RLR	0.041	0.012	0.010	0.004
	(1.203)	(0.894)	(1.228)	(0.383)
RLR Percentage	**4.025***	**7.588*****	1.453	2.624
	(1.760)	**(2.842)**	(1.142)	(1.178)
Average Score	0.036	0.091***	0.064***	0.060***
	(1.025)	(6.903)	(3.170)	(3.330)
Size	-0.126***	0.230***	-0.127***	0.279***
	(-18.952)	(40.272)	(-24.334)	(83.813)
Illiq	-0.436	-10.375***	2.879*	0.189
	(-0.168)	(-5.317)	(1.950)	(0.260)
Ivol	36.073***	20.671***	34.119***	18.413***
	(12.947)	(12.205)	(15.436)	(15.197)
Max	-0.656	-0.997**	-2.457***	-1.538***
	(-0.783)	(-2.154)	(-5.024)	(-6.599)
IOR	0.201***	0.226***	-0.037	0.207***
	(5.938)	(6.710)	(-1.394)	(8.107)
CGO	-0.213***	-0.008	-0.071	0.121***
	(-3.411)	(-0.458)	(-1.639)	(4.144)
Constant	3.698***	-5.505***	3.802***	-6.564***
	(26.611)	(-45.745)	(25.019)	(-79.456)
Observations	14,905	14,863	25,766	25,687
R-squared	0.301	0.491	0.251	0.562
Months	66	66	66	66

Table 2.6 examines if our predictors explain mispricing via idiosyncratic volatility or liquidity. Complex liquidity is the equal weighted average of three standardized liquidity measures, including Turnover, 1/Amihud illiquidity, and 1/(Bid-ask Spreads).

We have the following findings for overvalued stocks:

A higher word count average per tweet reduces idiosyncratic volatility and mispricing, while it is documented that there is a significant positive relationship between absolute idiosyncratic volatility and mispricing (Aabo et al., 2017). Idiosyncratic volatility measures unsystematic risk which should be positively related to mispricing, which is consistent with our finding. Therefore, word count possibly reduce mispricing via lessening idiosyncratic volatility.

RLR Percentage increases both idiosyncratic volatility and liquidity, however, the benefit from liquidity increase is stronger than the cost from volatility increase. Link Percentage can reduce mispricing through channels other than liquidity or idiosyncratic volatility since it does not show significant effects.

There are some findings regarding the undervalued stocks:

We already show that average number of words increases mispricing in Table 2.5, and Table 2.6 shows word decreases liquidity. It implies liquidity and mispricing are negatively related, and such relation is consistent with the common sense that more liquid stocks are often priced more accurately, as market can adjust prices more quickly for these stocks with higher liquidity. Average Word can increase mispricing index (i.e., reduce mispricing) but slightly reduce liquidity here. This finding suggests that Average Word affects mispricing through channels other than liquidity or idiosyncratic volatility. Picture can increase both liquidity and volatility. However, the cost and benefit seem to offset each other.

Table 2.7 Robustness Test

	Panel A: Overvalued Stocks			Panel B: Undervalued Stocks		
	Inverse Liquidity	Turnover	Inverse Spread	Inverse Liquidity	Turnover	Inverse Spread
Link	0.015	0.000	0.000	0.010	-0.000**	-0.019***
	(1.506)	(1.482)	(0.004)	(1.236)	(-2.182)	(-2.911)
Link Percentage	0.987*	-0.036***	3.427***	-0.079	-0.025***	2.972***
	(1.952)	(-4.651)	(2.800)	(-0.160)	(-5.294)	(4.044)
Picture	0.097	-0.001	0.033	-0.057	0.000	0.231*
	(1.305)	(-0.509)	(0.530)	(-1.039)	(1.092)	(1.818)
Picture Percentage	-3.403	0.027	4.560	-17.909	0.384	-61.511
	(-1.647)	(0.658)	(1.538)	(-1.030)	(1.054)	(-0.890)
Average Word	-0.008	-0.001**	-0.237***	-0.124*	0.000	-0.154
	(-0.142)	(-2.377)	(-3.678)	(-1.857)	(0.838)	(-1.655)
Total Word	-0.123	-0.007	0.300*	-0.182**	0.003	0.566*
	(-0.817)	(-1.373)	(1.915)	(-2.217)	(1.609)	(1.698)
RLR	-0.321	0.017	-0.978	1.011***	-0.004	-0.650
	(-0.590)	(1.553)	(-1.546)	(3.169)	(-0.847)	(-1.541)
RLR Percentage	210.830***	1.867**	108.479*	57.474	1.058	13.813
	(3.059)	(2.120)	(1.672)	(1.515)	(1.151)	(0.535)
Average Score	2.354***	0.025***	0.240	1.267	0.025***	-0.757
	(5.896)	(4.257)	(0.192)	(1.128)	(9.908)	(-1.017)
Size	8.716***	-0.002	10.356***	13.150***	-0.009***	10.089***
	(20.188)	(-0.358)	(17.228)	(31.433)	(-6.695)	(27.980)
Illiq	813.024***	-11.347***	133.838***	1,109.426***	-7.306***	-57.352
	(5.846)	(-6.327)	(3.045)	(10.348)	(-10.201)	(-1.336)
Ivol	64.069***	10.416***	-28.493	256.611***	8.190***	-170.050***
	(5.034)	(8.784)	(-1.021)	(5.478)	(19.944)	(-3.904)
Max	11.353**	-0.448	-32.487***	-33.716***	-0.513***	2.899
	(2.203)	(-1.659)	(-4.112)	(-3.622)	(-6.677)	(0.268)
IOR	-1.887***	0.089***	7.327***	-4.912***	0.089***	9.272***
	(-5.986)	(6.783)	(5.274)	(-8.545)	(8.599)	(9.238)
CGO	-1.644***	-0.040***	8.454***	-1.011	-0.023***	17.154***
	(-3.399)	(-2.977)	(7.913)	(-1.153)	(-3.843)	(13.233)
Constant	-186.234***	0.102	-199.313***	-280.570***	0.242***	-194.227***
	(-18.342)	(0.855)	(-14.707)	(-30.151)	(7.057)	(-25.575)
Observations	14,905	14,905	14,863	25,766	25,766	25,687
R-squared	0.601	0.347	0.396	0.633	0.292	0.327
Months	66	66	66	66	66	66

Table 2.7 reports robust test with controls. Sample period is from 2007 Mar to 2020 December. We adopt three individual liquidity measures as the variables in this robustness test. There is no clear pattern regarding these predictors, which means the significances are situational. For overvalued stocks, Link Percentage and RLR Percentage are significant predictors throughout all these three measures, while word and score are only significant for turnover and inverse spread, as well as inverse liquidity and turnover respectively. For undervalued stocks, Link and Link Percentage are the only significant predictors for turnover and inverse spread. To predict inverse liquidity effectively, we still need other predictors including Average Word, Total Word and RLR Percentage.

Inverse Liquidity = 1 / (Amihud illiqudidy * 1000)
Turnover = trading volume/outstanding shares
Inverse Spread = 1 / Spread, where Spread = (ask price - bid price) / close price

6.3 Forecasting Industry Return

There is a largely consistent pattern of statistical significance for Consumables, High Technology and Healthcare industries. These results show our sentiment indices, sentiment dispersion and hybrid variable are effective predictors for the three industry returns. Our results are consistent with the discovery that Twitter sentiment is relevant to some industry portfolios, including High Technology (Oliveira, Cortez, and Areal, 2017). The various robustness tests support such conclusion. This is understandable and consistent with common sense, as there are many influential technological, healthcare and food firm accounts on Twitter, such as Microsoft, PepsiCo, and CVS. Our finding is in accordance with researchers from Mayo Clinic claiming that Twitter is currently the most popular form of social media used for healthcare communication, and Twitter is an informative channel for all parties in the healthcare industry, including patients, providers and researchers (Pershad et al., 2018).

In addition, it makes sense for Consumables companies to promote their products on Twitter, since they know many consumers are also Twitter users. It has been reported that certain food industry actors actively use Twitter to influence food and health policy in debates (Hunt, 2021) signaling consumables companies might be more aware of the importance of Twitter as a promotion strategy for their business. Therefore, it may help explain the stronger connection and significance between consumables industry return and the tweets from these companies in comparison to some other industries.

It is also not surprising that technology and health companies might be more familiar with modern communication technologies including Twitter, compared to some traditional industry such as manufacturing. Besides, many manufacturing firms acquire stable contracts from their government or other downstream firms whose official communication is unlikely via Twitter, rendering our tweet-based prediction less effective. Our models cannot predict the last category called Other, perhaps because it blends multiple industry characteristics and lacks any distinct industry feature, finally becoming too general for valid prediction.

There are other less noticeable patterns. First, the measure typically has the strongest statistical and economic significances for the first and second quarter. In other words, the prediction power is strongest both in statistical and economic aspects for the return prediction in the next three to six months, regardless of industry types. If we create a coefficient chart, it will be in an inverted U shape. This is possible due to either insufficient or redundant information for return prediction whose time frame is too early or too late and either one is ideal for optimal prediction. Second, the health industry is the most robust one out of the three valid predictions and it does not matter which method or prediction month to choose, it always has statistical significance. That is, the healthcare industry has the most consistent prediction performance. A study argues that social media is far reaching and acts as a game changer for health community to greatly promote health related behaviors and tackle problems, especially in crisis time (Gupta et al., 2013), in which many company executives in this industry may already realize. Such awareness could act as a reinforcement of Twitter usage, pumping out more tweets for us to analyze, which appeals to strengthen the relation between their financial performance and tweets.

We find that the aggregate method has little impact on industry return predictions, which is consistent with common sense. If we read the entire passage at one time, or read these individual sentences in the same passage, but in different days. However, it is a bit surprising that firm size weight does not play a significant role in the return predictions. Compared to equal weighted aggregate approaches, we do not observe any additional pattern after factoring in each firm's market capital. This is perhaps because the total market size is too large for any firm to be dominant across all the industries.

As these sentiment aggregate methods (AS, WAS, IS, WAS) are highly similar, we only list one of them (AS) to represent the rest. The sentiment volatility measure is different compared to these aggregation methods. Therefore, we include tables related to this variable in the robustness test section. The robustness tests show our sentiment indices are often effective in predicating three out of the five industry returns.

Table 3.1 Variable Description

Variable	Description
Panel A: Explanatory Variables	
Aggregation Score (AS)	Monthly aggregate tweet score with equal weight
Weighted Aggregation Score (WAS)	Monthly aggregate tweet score with firm size weight
Individual Score (IS)	Monthly individual tweet score with equal weight
Weighted Individual Score (WIS)	Monthly individual tweet score with firm size weight
Sentiment Volatility (STD)	Monthly sentiment dispersion measured by standard deviation with equal firm size weight, i.e., sentiment volatility
Panel B: Control Variables	
Inflation (infl)	Inflation: Consumer Price Index (All Urban Consumers) from 1919 to 2005 from the Bureau of Labor Statistics.
Net Equity Expansion (ntis)	Net Equity Expansion: the ratio of 12-month moving sums of net issues by NYSE listed stocks divided by the total end-of-year market capitalization of NYSE stocks.
Stock Variance (svar)	Stock Variance: computed as sum of squared daily returns on the S&P 500.
News Sentiment (ns)	Monthly conversion from the Daily News Sentiment Index, which is a high frequency measure of economic sentiment based on lexical analysis of economics-related news articles. The index is described in Buckman, Shapiro, Sudhof and Wilson (2020) and based on the methodology developed in Shapiro, Sudhof and Wilson (2020).
Industry Growth Rate (igr)	The industrial production (IP) index measures the real output of all relevant establishments located in the United States, regardless of their ownership, but not those located in U.S. territories. Industry Growth Rate is short for industry growth rate, as its name

	suggests, is the monthly industrial production percentage increase.
Recession	This time series is an interpretation of US Business Cycle Expansions and Contractions data provided by The National Bureau of Economic Research (NBER). A value of 1 is a recessionary period, while a value of 0 is an expansionary period. For this time series, the recession begins the first day of the period following a peak and ends on the last day of the period of the trough.

Table 3.1 reports variables involved in the sentiment regression analyses, all converted to monthly statistics. Panel A analyzes explanatory variables. Panel B addresses control variables related to macroeconomic indicators or known sentiment indices.

Table 3.2 Summary Statistic

	Min	**Q1**	**Q2**	**Q3**	**P90**	**Max**	**Mean**	**STD**	**Months**
Standard Deviation	-0.027	-0.003	0.001	0.005	0.008	0.044	0.001	0.009	144
Aggregate Score	0.073	0.238	0.264	0.277	0.288	0.313	0.253	0.037	144
Weighted Aggregate Score	0.134	0.286	0.323	0.337	0.351	0.378	0.306	0.049	144
Individual Score	0.038	0.143	0.156	0.177	0.216	0.245	0.162	0.036	144
Weighted Individual Score	0.041	0.158	0.177	0.201	0.228	0.264	0.177	0.040	144

Table 3.2 reports the summary statistics of the four sentiment indices and the sentiment dispersion from January 2009 to December 2020 and then reports the distribution across 938 firms. The columns from left to right are minimum, first quartile, median, third quartile, 90th percentile, maximum, mean, standard deviation and duration measured in month, respectively. The summary statistics is consistent with our observation that company tweets generally have a positive tone, which corresponds to positive sentiment scores. Therefore, the overall sentiment of each index is modestly positive. However, when using the sentiment dispersion as the market sentiment proxy, we see a significant increase in these magnitudes.

Table 3.2 lists how our sentiment variables behave in a snapshot of basic statistics. Overall, all these variables are close to even distribution across 144 months, with standard deviations only ranging from 0.037 to 0.049, and there is not much difference between Q1 and P90. When tweet sentiments are measured individually, they generally have a lower mean and std (around 0.16 and 0.038 respectively) in comparison to aggregate methods whose values are around 0.27 and 0.043 respectively. This is probably because each sentiment stands out when computed individually, while they tend to cancel each other when calculated in aggregation.

Table 3.3 Correlation Matrix

	infl	ntis	svar	igr	ntis	std	as	was	is	wis
infl	1.000									
ntis	-0.066	1.000								
svar	-0.132	-0.036	1.000							
igr	0.227	0.124	-0.331	1.000						
ns	0.000	-0.068	-0.369	0.066	1.000					
std	-0.036	-0.213	0.180	-0.021	-0.273	1.000				
as	-0.132	-0.077	-0.315	0.159	0.470	0.283	1.000			
was	-0.120	-0.191	-0.292	0.025	0.479	0.176	0.906	1.000		
is	-0.094	-0.239	-0.134	0.112	0.222	0.743	0.832	0.723	1.000	
wis	-0.067	-0.334	-0.168	0.041	0.333	0.596	0.829	0.855	0.927	1.000

Abbreviations are used due to space limitation to represent the following each variable: inflation (infl), net equity expansion (ntis), stock variance (svar), industry growth rate (igr), news sentiment (ns), sentiment volatility (std), aggregate score (as), weighted aggregate score (was), individual score (is) and weighted individual score (wis)

Table 3.3 reports the pairwise correlations among control variables and explanatory variables. The correlation matrix shows that only these last five sentiment indices in the table are highly correlated with each other, with magnitudes above 0.7. Therefore, we will only use one of them along with other variables in each regression, to avoid multicollinearity. This is intuitively understandable, as these sentiment indices are constructed similarly.

The following tables share the most data sources below:

Federal Reserve Bank of San Francisco. (2020). *Daily News Sentiment Index*. Retrieved November 3, 2022, from https://www.frbsf.org/economic-research/indicators-data/daily-news-sentiment-index/

Federal Reserve Economic Data. (2021, August 17). *Industrial production: Total index*. Federal Reserve Economic Data | FRED | St. Louis Fed. Retrieved November 3, 2022, from https://fred.stlouisfed.org/series/INDPRO

Goyal, A. (2008, July). *A Comprehensive Look at The Empirical Performance of Equity Premium Prediction*. Google Sites. Retrieved November 3, 2022, from https://sites.google.com/view/agoyal145
All the regression terms have been standardized, so we can directly compare the magnitudes of these coefficients for economic significance.

Table 3.4 Predictions with Aggregate Score

Panel A: Consumables

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	0.053	-0.073	0.049	0.444	0.065
	(0.161)	(-0.215)	(0.141)	(1.292)	(0.185)
Net Equity Expansion	-0.191	0.029	-0.092	0.182	-0.117
	(-0.579)	(0.087)	(-0.267)	(0.490)	(-0.302)
Stock Variance	1.194	0.064	-0.374	-0.046	0.257
	(3.268)	(0.171)	(-0.971)	(-0.104)	(0.634)
Industry Growth Rate	0.442	-0.757	0.305	-0.275	0.541
	(1.252)	(-2.082)	(0.822)	(-0.622)	(1.380)
News Sentiment	-0.619	-0.611	-0.821	-0.498	-0.151
	(-1.783)	(-1.708)	(-2.228)	(-1.312)	(-0.359)
Aggregate Score	0.093	0.205***	0.202***	0.181**	0.150*
	(1.464)	(3.134)	(3.011)	(2.447)	(1.905)
R-squared	0.224	0.169	0.144	0.131	0.118
Observations	143	141	138	135	132

Panel B: Manufacturing

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-0.120	0.096	0.012	0.409	-0.167
	(-0.299)	(0.238)	(0.029)	(1.022)	(-0.412)
Net Equity Expansion	-0.134	-0.198	-0.306	0.113	-0.303
	(-0.333)	(-0.493)	(-0.763)	(0.260)	(-0.678)
Stock Variance	0.878	0.061	-0.236	-0.004	0.245
	(1.964)	(0.136)	(-0.530)	(-0.008)	(0.524)
Industry Growth Rate	0.201	-0.138	0.170	-0.284	0.529
	(0.465)	(-0.318)	(0.398)	(-0.553)	(1.168)
News Sentiment	-0.458	-0.566	-0.909	-0.185	-0.071
	(-1.078)	(-1.330)	(-2.134)	(-0.419)	(-0.146)
Aggregate Score	0.060	0.095	0.093	0.097	0.069
	(0.777)	(1.222)	(1.196)	(1.130)	(0.761)
R squared	0.086	0.056	0.067	0.041	0.042
Observations	143	141	138	135	132

Panel C: High Technology

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-0.065	-0.138	0.251	0.262	0.101

	(-0.168)	(-0.350)	(0.632)	(0.661)	(0.254)
Net Equity Expansion	-0.452	-0.234	-0.309	0.006	-0.474
	(-1.170)	(-0.593)	(-0.774)	(0.014)	(-1.075)
Stock Variance	1.022	0.261	-0.430	-0.177	0.377
	(2.384)	(0.595)	(-0.969)	(-0.343)	(0.817)
Industry Growth Rate	0.148	-0.584	0.187	-0.588	0.715
	(0.357)	(-1.379)	(0.438)	(-1.153)	(1.600)
News Sentiment	-0.782	-0.518	-0.787	-0.390	0.129
	(-1.921)	(-1.243)	(-1.856)	(-0.889)	(0.268)
Aggregate Score	0.131*	0.223***	0.211***	0.233***	0.144
	(1.755)	(2.926)	(2.732)	(2.722)	(1.605)
R squared	0.206	0.160	0.141	0.125	0.126
Observations	143	141	138	135	132

Panel D: Healthcare

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	0.128	-0.109	-0.228	0.318	0.347
	(0.362)	(-0.307)	(-0.647)	(0.900)	(0.979)
Net Equity Expansion	0.081	0.047	-0.198	0.131	0.047
	(0.231)	(0.134)	(-0.557)	(0.342)	(0.119)
Stock Variance	0.813	-0.106	-0.316	0.104	0.013
	(2.077)	(-0.270)	(-0.802)	(0.225)	(0.031)
Industry Growth Rate	-0.001	-0.229	0.373	-0.146	0.404
	(-0.002)	(-0.605)	(0.982)	(-0.320)	(1.018)
News Sentiment	-0.149	-0.254	-0.601	-0.446	-0.098
	(-0.399)	(-0.682)	(-1.594)	(-1.140)	(-0.230)
Aggregate Score	0.144**	0.221***	0.207***	0.168**	0.162**
	(2.122)	(3.242)	(3.022)	(2.203)	(2.031)
R squared	0.132	0.107	0.121	0.110	0.107
Observations	143	141	138	135	132

Panel E: Other

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-0.266	-0.010	0.019	0.736	-0.396
	(-0.584)	(-0.022)	(0.043)	(1.708)	(-0.910)
Net Equity Expansion	-0.679	-0.572	-0.519	-0.051	-0.360
	(-1.490)	(-1.282)	(-1.187)	(-0.109)	(-0.749)
Stock Variance	0.684	-0.147	-0.221	0.111	0.484
	(1.350)	(-0.296)	(-0.455)	(0.199)	(0.962)
Industry Growth Rate	0.291	-0.351	0.455	-0.212	0.552
	(0.595)	(-0.734)	(0.974)	(-0.384)	(1.135)
News Sentiment	-0.635	-0.747	-0.939	-0.210	0.036
	(-1.319)	(-1.586)	(-2.022)	(-0.441)	(0.068)

Aggregate Score	0.091	0.135	0.121	0.102	0.115
	(1.039)	(1.563)	(1.437)	(1.095)	(1.178)
R-squared	0.103	0.088	0.090	0.074	0.072
Observations	143	141	138	135	132

Table 3.4 shows the prediction performance of Aggregate Score for selected duration k month(s) in major industries. K represents the prediction length measured in months.

Asterisk annotation throughout this manual:
***: p value < 0.01
**: p value between 0.01 and 0.05
*: p value between 0.05 and 0.1

Table 3.4 reports the regressions for predictions with macroeconomic and sentiment control variables, including Inflation, Net Equity Expansion, Industrial Growth Rate and News Sentiment. The sample period is from January 2009 to December 2020. These prediction results above show that Aggregate Score usually can predict returns from three out of five industries, including Consumables, High Technology and Healthcare. The predictions are relatively robust, as we have checked multiple periods on a quarterly basis. The coefficients are all positive, which makes sense, as an overall positive market sentiment usually leads to increased return regardless of industry type.

The explanatory variable is Aggregate Score from our sentiment indices. Each t score is attached inside the parentheses below the regression coefficient. The same format holds across all tables.

Table 3.4 shows the prediction performance of the aggregate score for the return of each industry. This measure generally shows statistical significance for Consumables, High Technology and Healthcare.

There are different ways to aggregate tweet information into sentiment scores, depending on various understandings of content accumulation. Intuitively, we can separately analyze each tweet, or combine all into one block for a comprehensive analysis.

These various evaluation methods may impact the prediction power of our index. Therefore, we check other aggregate methods to see if the impact is deterministic in Tables 3.5 to 3.17.

Table 3.5 Predictions with Individual Score

Panel A: Consumables

	k = 1	k = 3	k = 6	k = 9	k = 12
Inflation	0.078	-0.056	0.056	0.440	0.062
	(0.239)	(-0.165)	(0.164)	(1.291)	(0.179)
Net Equity Expansion	-0.190	0.079	-0.033	0.243	-0.059
	(-0.564)	(0.228)	(-0.093)	(0.642)	(-0.149)
Stock Variance	1.209	0.060	-0.393	-0.081	0.254
	(3.297)	(0.158)	(-1.018)	(-0.181)	(0.635)
Industry Growth Rate	0.444	-0.768	0.297	-0.293	0.521
	(1.252)	(-2.108)	(0.802)	(-0.665)	(1.331)
News Sentiment	-0.611	-0.590	-0.814	-0.513	-0.156
	(-1.756)	(-1.646)	(-2.213)	(-1.353)	(-0.372)
Individual Score	0.122	0.298***	0.296***	0.276**	0.226**
	(1.293)	(3.098)	(3.072)	(2.563)	(2.017)
R squared	0.222	0.168	0.146	0.135	0.121
Observations	143	141	138	135	132

Panel B: Manufacturing

	k = 1	k = 3	k = 6	k = 9	k = 12
Inflation	-0.084	0.124	0.035	0.420	-0.150
	(-0.208)	(0.308)	(0.087)	(1.055)	(-0.372)
Net Equity Expansion	-0.159	-0.201	-0.303	0.127	-0.302
	(-0.388)	(-0.488)	(-0.741)	(0.286)	(-0.662)
Stock Variance	0.908	0.079	-0.225	-0.001	0.273
	(2.025)	(0.175)	(-0.503)	(-0.002)	(0.588)
Industry Growth Rate	0.212	-0.134	0.174	-0.274	0.544
	(0.488)	(-0.310)	(0.406)	(-0.533)	(1.200)
News Sentiment	-0.457	-0.559	-0.908	-0.190	-0.062
	(-1.074)	(-1.312)	(-2.127)	(-0.430)	(-0.127)
Individual Score	0.061	0.121	0.120	0.137	0.086
	(0.528)	(1.055)	(1.069)	(1.088)	(0.665)
R squared	0.084	0.053	0.065	0.041	0.041
Observations	143	141	138	135	132

Panel C: High Technology

	k = 1	k = 3	k = 6	k = 9	k = 12
Inflation	-0.040	-0.124	0.252	0.259	0.096
	(-0.103)	(-0.315)	(0.640)	(0.656)	(0.243)

Net Equity Expansion	-0.437	-0.175	-0.238	0.083	-0.415
	(-1.109)	(-0.435)	(-0.588)	(0.190)	(-0.922)
Stock Variance	1.032	0.251	-0.456	-0.220	0.371
	(2.399)	(0.572)	(-1.028)	(-0.426)	(0.812)
Industry Growth Rate	0.145	-0.598	0.176	-0.610	0.693
	(0.349)	(-1.410)	(0.414)	(-1.200)	(1.551)
News Sentiment	-0.770	-0.494	-0.780	-0.409	0.122
	(-1.885)	(-1.185)	(-1.843)	(-0.934)	(0.255)
Individual Score	0.181	0.328***	0.315***	0.354***	0.219*
	(1.635)	(2.929)	(2.840)	(2.847)	(1.719)
R squared	0.204	0.160	0.145	0.129	0.129
Observations	143	141	138	135	132

Panel D: Healthcare

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	0.161	-0.079	-0.206	0.338	0.361
	(0.459)	(-0.223)	(-0.585)	(0.957)	(1.023)
Net Equity Expansion	0.090	0.086	-0.155	0.156	0.085
	(0.251)	(0.239)	(-0.428)	(0.398)	(0.213)
Stock Variance	0.830	-0.100	-0.321	0.107	0.037
	(2.110)	(-0.254)	(-0.809)	(0.232)	(0.090)
Industry Growth Rate	-0.001	-0.237	0.370	-0.129	0.405
	(-0.003)	(-0.622)	(0.974)	(-0.283)	(1.021)
News Sentiment	-0.136	-0.233	-0.596	-0.454	-0.092
	(-0.363)	(-0.623)	(-1.576)	(-1.159)	(-0.217)
Individual Score	0.194*	0.312***	0.292***	0.236**	0.227**
	(1.923)	(3.101)	(2.946)	(2.125)	(1.997)
R squared	0.127	0.102	0.118	0.108	0.106
Observations	143	141	138	135	132

Panel E: Other

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-0.217	0.031	0.046	0.754	-0.376
	(-0.477)	(0.069)	(0.107)	(1.758)	(-0.868)
Net Equity Expansion	-0.710	-0.576	-0.512	-0.045	-0.347
	(-1.525)	(-1.263)	(-1.149)	(-0.095)	(-0.705)
Stock Variance	0.723	-0.121	-0.209	0.124	0.517
	(1.422)	(-0.244)	(-0.428)	(0.220)	(1.036)
Industry Growth Rate	0.304	-0.347	0.460	-0.192	0.567
	(0.619)	(-0.721)	(0.982)	(-0.347)	(1.161)
News Sentiment	-0.632	-0.738	-0.936	-0.213	0.046
	(-1.310)	(-1.561)	(-2.013)	(-0.447)	(0.088)
Individual Score	0.098	0.171	0.158	0.137	0.152
	(0.749)	(1.344)	(1.298)	(1.015)	(1.090)
R squared	0.099	0.084	0.087	0.073	0.070
Observations	143	141	138	135	132

K is the prediction length measured in month. These tables report the regressions for predictions with macroeconomic and sentiment control variables, including Inflation, Net Equity Expansion, Industry Growth Rate and News Sentiment. The Sample period is from January 2009 to December 2020. The explanatory variable is Individual Score from our sentiment indices, which usually can predict returns from three out of five industries, including Consumables, High Technology and Healthcare. The predictions are robust, as we have checked multiple periods on a quarterly basis. The coefficients are all positive, which makes sense, as an overall positive market sentiment usually leads to increased return regardless of industry type.

Table 3.5 shows very similar conclusion in comparison to Table 3.4. The measure exhibits prediction power for Consumables, High Technology and Healthcare. The third and sixth months are the most reliable for prediction, both with p values less than 0.01 and higher magnitudes compared to other months, while the next month and the next year predictions tend to be less robust. For example, there is no significance for the next month for the Consumables and High Technology industries.

Table 3.6 Predictions with Weighted Aggregate Score

Panel A: Consumables					
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	0.046	-0.075	0.026	0.431	0.068
	(0.138)	(-0.220)	(0.076)	(1.256)	(0.195)
Net Equity Expansion	-0.175	0.047	-0.053	0.210	-0.115
	(-0.529)	(0.137)	(-0.153)	(0.561)	(-0.295)
Stock Variance	1.192	0.073	-0.385	-0.046	0.259
	(3.269)	(0.195)	(-1.004)	(-0.103)	(0.635)
Industry Growth Rate	0.453	-0.728	0.330	-0.254	0.574
	(1.288)	(-2.005)	(0.895)	(-0.585)	(1.485)
News Sentiment	-0.621	-0.617	-0.829	-0.498	-0.153
	(-1.789)	(-1.723)	(-2.259)	(-1.315)	(-0.364)
Weighted Aggregate Score	0.101	0.213***	0.221***	0.192**	0.152*
	(1.512)	(3.103)	(3.168)	(2.538)	(1.860)
R squared	0.225	0.168	0.150	0.134	0.117
Observations	143	141	138	135	132
Panel B: Manufacturing					
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-0.119	0.101	-0.019	0.397	-0.163
	(-0.294)	(0.249)	(-0.047)	(0.993)	(-0.401)
Net Equity Expansion	-0.131	-0.197	-0.266	0.133	-0.306
	(-0.323)	(-0.484)	(-0.660)	(0.306)	(-0.680)
Stock Variance	0.883	0.070	-0.260	-0.010	0.250

	(1.976)	(0.157)	(-0.585)	(-0.019)	(0.532)
Industry Growth Rate	0.210	-0.122	0.176	-0.279	0.548
	(0.488)	(-0.283)	(0.413)	(-0.551)	(1.226)
News Sentiment	-0.460	-0.569	-0.913	-0.186	-0.070
	(-1.082)	(-1.337)	(-2.147)	(-0.422)	(-0.145)
Weighted Aggregate Score	0.061	0.096	0.114	0.106	0.068
	(0.755)	(1.170)	(1.404)	(1.200)	(0.724)
R squared	0.086	0.055	0.071	0.043	0.041
Observations	143	141	138	135	132
Panel C: High Technology					
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-0.077	-0.141	0.224	0.248	0.108
	(-0.199)	(-0.357)	(0.565)	(0.626)	(0.271)
Net Equity Expansion	-0.428	-0.214	-0.264	0.038	-0.477
	(-1.103)	(-0.539)	(-0.660)	(0.088)	(-1.076)
Stock Variance	1.018	0.270	-0.445	-0.172	0.385
	(2.380)	(0.617)	(-1.007)	(-0.336)	(0.829)
Industry Growth Rate	0.163	-0.552	0.212	-0.558	0.751
	(0.395)	(-1.307)	(0.500)	(-1.111)	(1.704)
News Sentiment	-0.785	-0.524	-0.796	-0.390	0.129
	(-1.929)	(-1.257)	(-1.884)	(-0.890)	(0.269)
Weighted Aggregate Score	0.142*	0.232***	0.233***	0.245***	0.143
	(1.824)	(2.902)	(2.899)	(2.805)	(1.540)
R squared	0.207	0.159	0.147	0.128	0.125
Observations	143	141	138	135	132
Panel D: Healthcare					
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	0.128	-0.109	-0.255	0.313	0.333
	(0.363)	(-0.307)	(-0.723)	(0.884)	(0.942)
Net Equity Expansion	0.092	0.064	-0.154	0.148	0.070
	(0.257)	(0.179)	(-0.432)	(0.385)	(0.178)
Stock Variance	0.822	-0.094	-0.330	0.114	-0.013
	(2.099)	(-0.241)	(-0.842)	(0.249)	(-0.032)
Industry Growth Rate	0.020	-0.197	0.397	-0.118	0.420
	(0.054)	(-0.521)	(1.055)	(-0.263)	(1.076)
News Sentiment	-0.153	-0.260	-0.610	-0.445	-0.111
	(-0.410)	(-0.698)	(-1.625)	(-1.138)	(-0.261)
Weighted Aggregate Score	0.149**	0.228***	0.230***	0.174**	0.176**
	(2.086)	(3.192)	(3.209)	(2.234)	(2.130)
R squared	0.131	0.105	0.128	0.111	0.110
Observations	143	141	138	135	132
Panel E: Other					
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**

Inflation	-0.270	-0.022	-0.017	0.722	-0.400
	(-0.592)	(-0.050)	(-0.039)	(1.677)	(-0.919)
Net Equity Expansion	-0.668	-0.548	-0.471	-0.028	-0.351
	(-1.455)	(-1.221)	(-1.074)	(-0.059)	(-0.726)
Stock Variance	0.685	-0.151	-0.247	0.103	0.475
	(1.355)	(-0.304)	(-0.511)	(0.185)	(0.940)
Industry Growth Rate	0.303	-0.335	0.464	-0.209	0.570
	(0.621)	(-0.702)	(0.999)	(-0.383)	(1.188)
News Sentiment	-0.637	-0.750	-0.944	-0.211	0.030
	(-1.324)	(-1.593)	(-2.038)	(-0.444)	(0.057)
Weighted Aggregate Score	0.097	0.147	0.146	0.111	0.121
	(1.051)	(1.621)	(1.656)	(1.172)	(1.197)
R squared	0.103	0.089	0.094	0.076	0.072
Observations	143	141	138	135	132

K is the prediction length measured in months. Table 3.6 reports the regressions for predictions with macroeconomic and sentiment control variables, including Inflation, Net Equity Expansion, Industry Growth Rate and News Sentiment. The sample period is from January 2009 to December 2020.

The explanatory variable is Weighted Aggregate Score from our sentiment indices. These prediction results above show that Weighted Aggregate Score usually can predict returns from three out of five industries, including Consumables, High Technology and Healthcare. The predictions are relatively robust, as we have checked multiple periods on a quarterly basis. The coefficients are all positive, which makes sense, as an overall positive market sentiment usually leads to increased return regardless of industry type.

Again, when we switch to weighted aggregate scores in Table 3.6, the results barely change. The model can mostly predict the returns of Consumables, High Technology and Healthcare. The coefficients have lower magnitudes for the next month's prediction, usually peaking at the third or sixth month, then dropping to a similar level at the end.

Table 3.7 Predictions with Weighted Individual Score

Panel A: Consumables					
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	0.068	-0.066	0.032	0.416	0.069
	(0.208)	(-0.195)	(0.094)	(1.218)	(0.198)
Net Equity Expansion	-0.174	0.099	0.005	0.281	-0.070
	(-0.513)	(0.285)	(0.013)	(0.738)	(-0.176)
Stock Variance	1.206	0.067	-0.399	-0.088	0.272
	(3.295)	(0.177)	(-1.037)	(-0.197)	(0.677)
Industry Growth Rate	0.452	-0.743	0.320	-0.281	0.566
	(1.280)	(-2.042)	(0.868)	(-0.646)	(1.461)
News Sentiment	-0.623	-0.620	-0.845	-0.534	-0.166
	(-1.792)	(-1.731)	(-2.302)	(-1.411)	(-0.394)
Weighted Individual Score	0.125	0.296***	0.307***	0.288***	0.215*
	(1.336)	(3.075)	(3.171)	(2.686)	(1.881)
R squared	0.222	0.167	0.150	0.139	0.118
Observations	143	141	138	135	132

Panel B: Manufacturing

	k = 1	k = 3	k = 6	k = 9	k = 12
Inflation	-0.085	0.126	0.001	0.399	-0.141
	(-0.210)	(0.311)	(0.001)	(0.999)	(-0.349)
Net Equity Expansion	-0.157	-0.200	-0.256	0.159	-0.315
	(-0.378)	(-0.481)	(-0.623)	(0.356)	(-0.685)
Stock Variance	0.911	0.087	-0.251	-0.018	0.289
	(2.033)	(0.194)	(-0.563)	(-0.035)	(0.621)
Industry Growth Rate	0.217	-0.122	0.176	-0.280	0.569
	(0.503)	(-0.282)	(0.412)	(-0.553)	(1.268)
News Sentiment	-0.464	-0.572	-0.921	-0.203	-0.061
	(-1.090)	(-1.342)	(-2.162)	(-0.459)	(-0.126)
Weighted Individual Score	0.059	0.115	0.144	0.151	0.076
	(0.516)	(1.005)	(1.285)	(1.201)	(0.577)
	0.084	0.053	0.069	0.043	0.040
Observations	143	141	138	135	132

Panel C: High Technology

	k = 1	k = 3	k = 6	k = 9	k = 12
Inflation	-0.053	-0.131	0.220	0.230	0.109
	(-0.137)	(-0.332)	(0.558)	(0.582)	(0.273)
Net Equity Expansion	-0.416	-0.158	-0.190	0.129	-0.434
	(-1.046)	(-0.389)	(-0.466)	(0.294)	(-0.957)
Stock Variance	1.030	0.264	-0.469	-0.226	0.397
	(2.398)	(0.600)	(-1.061)	(-0.440)	(0.865)
Industry Growth Rate	0.158	-0.568	0.199	-0.592	0.743
	(0.382)	(-1.342)	(0.469)	(-1.181)	(1.681)
News Sentiment	-0.787	-0.528	-0.813	-0.436	0.116
	(-1.931)	(-1.265)	(-1.927)	(-0.997)	(0.241)
Weighted Individual Score	0.184*	0.322***	0.333***	0.368***	0.203
	(1.677)	(2.871)	(2.986)	(2.972)	(1.561)
R squared	0.204	0.158	0.150	0.134	0.125
Observations	143	141	138	135	132

Panel D: Healthcare

	k = 1	k = 3	k = 6	k = 9	k = 12
Inflation	0.153	-0.082	-0.240	0.327	0.346
	(0.434)	(-0.232)	(-0.682)	(0.925)	(0.979)
Net Equity Expansion	0.105	0.098	-0.104	0.174	0.105
	(0.289)	(0.269)	(-0.287)	(0.440)	(0.260)
Stock Variance	0.834	-0.086	-0.337	0.118	0.021
	(2.121)	(-0.217)	(-0.854)	(0.255)	(0.052)
Industry Growth Rate	0.015	-0.207	0.390	-0.104	0.425
	(0.040)	(-0.545)	(1.032)	(-0.231)	(1.084)
News Sentiment	-0.155	-0.265	-0.626	-0.469	-0.118

	(-0.416)	(-0.708)	(-1.665)	(-1.196)	(-0.276)
Weighted Individual Score	0.193*	0.303***	0.312***	0.238**	0.237**
	(1.919)	(3.012)	(3.139)	(2.144)	(2.050)
R squared	0.127	0.098	0.125	0.108	0.108
Observations	143	141	138	135	132

Panel E: Other

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-0.218	0.013	0.009	0.733	-0.379
	(-0.478)	(0.029)	(0.020)	(1.704)	(-0.873)
Net Equity Expansion	-0.706	-0.549	-0.460	-0.014	-0.343
	(-1.503)	(-1.195)	(-1.027)	(-0.029)	(-0.693)
Stock Variance	0.727	-0.128	-0.236	0.108	0.516
	(1.432)	(-0.257)	(-0.485)	(0.193)	(1.032)
Industry Growth Rate	0.313	-0.336	0.464	-0.198	0.587
	(0.639)	(-0.701)	(0.997)	(-0.362)	(1.219)
News Sentiment	-0.642	-0.754	-0.954	-0.226	0.033
	(-1.333)	(-1.599)	(-2.055)	(-0.474)	(0.063)
Weighted Individual Score	0.095	0.179	0.185	0.151	0.153
	(0.730)	(1.412)	(1.509)	(1.116)	(1.074)
R squared	0.099	0.085	0.091	0.075	0.070
Observations	143	141	138	135	132

Table 3.7 reports the regressions for predictions with macroeconomic and sentiment control variables, including Inflation, Net Equity Expansion, Industry Growth Rate and News Sentiment. The sample period is from January 2009 to December 2020.The explanatory variable is Weighted Individual Score from our sentiment indices. These prediction results above show that Weighted Individual Score usually can predict returns from three out of five industries, including Consumables, High Technology and Healthcare. The predictions are relatively robust, as we have checked multiple periods on a quarterly basis. The coefficients are all positive, which makes sense, as an overall positive market sentiment usually leads to increased return regardless of industry type.

Lastly, we check the weighted aggregate score and report the same pattern compared to previous aggregate methods. This finding concludes our robustness test, and now it is safe to say that aggregation structure rarely influences valid prediction results.

Table 3.8 Predictions with Sentiment Volatility

Panel A: Consumables

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	0.008	-0.030	0.000	0.103	0.009
	(0.098)	(-0.348)	(0.001)	(1.171)	(0.097)
Net Equity Expansion	-0.045	0.014	-0.016	0.050	-0.027
	(-0.542)	(0.162)	(-0.186)	(0.531)	(-0.276)
Stock Variance	0.291	0.002	-0.109	-0.024	0.049
	(3.166)	(0.023)	(-1.109)	(-0.212)	(0.469)
Industry Growth Rate	0.112	-0.187	0.081	-0.071	0.135
	(1.275)	(-2.045)	(0.867)	(-0.633)	(1.376)

News Sentiment	-0.140	-0.124	-0.179	-0.106	-0.026
	(-1.613)	(-1.369)	(-1.916)	(-1.108)	(-0.252)
Sentiment Volatility	0.014	0.031***	0.029***	0.024**	0.019*
	(1.503)	(3.225)	(3.119)	(2.459)	(1.929)
R squared	0.225	0.172	0.148	0.131	0.119
Observations	143	141	138	135	132

Panel B: Manufacturing

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-0.026	0.014	-0.003	0.083	-0.042
	(-0.300)	(0.158)	(-0.032)	(0.929)	(-0.471)
Net Equity Expansion	-0.029	-0.038	-0.064	0.030	-0.062
	(-0.341)	(-0.437)	(-0.722)	(0.313)	(-0.637)
Stock Variance	0.184	0.006	-0.058	-0.011	0.043
	(1.930)	(0.058)	(-0.588)	(-0.094)	(0.407)
Industry Growth Rate	0.044	-0.029	0.039	-0.067	0.112
	(0.485)	(-0.308)	(0.413)	(-0.590)	(1.137)
News Sentiment	-0.090	-0.110	-0.188	-0.031	-0.012
	(-0.997)	(-1.185)	(-2.000)	(-0.324)	(-0.113)
Sentiment Volatility	0.007	0.013	0.012	0.012	0.008
	(0.719)	(1.321)	(1.253)	(1.212)	(0.852)
R squared	0.086	0.058	0.068	0.043	0.043
Observations	143	141	138	135	132

Panel C: High Technology

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-0.020	-0.041	0.044	0.047	0.017
	(-0.243)	(-0.469)	(0.503)	(0.532)	(0.191)
Net Equity Expansion	-0.093	-0.045	-0.062	0.007	-0.103
	(-1.120)	(-0.525)	(-0.702)	(0.070)	(-1.063)
Stock Variance	0.212	0.044	-0.107	-0.054	0.072
	(2.276)	(0.457)	(-1.091)	(-0.466)	(0.686)
Industry Growth Rate	0.034	-0.123	0.045	-0.131	0.157
	(0.381)	(-1.340)	(0.478)	(-1.171)	(1.609)
News Sentiment	-0.152	-0.085	-0.147	-0.063	0.038
	(-1.720)	(-0.930)	(-1.574)	(-0.661)	(0.368)
Sentiment Volatility	0.017*	0.029***	0.026***	0.027***	0.016
	(1.810)	(2.998)	(2.822)	(2.750)	(1.590)
R squared	0.207	0.163	0.144	0.126	0.126
Observations	143	141	138	135	132

Panel D: Healthcare

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	0.032	-0.034	-0.064	0.075	0.083
	(0.366)	(-0.385)	(-0.725)	(0.851)	(0.928)
Net Equity Expansion	0.015	0.013	-0.047	0.030	0.009

	(0.178)	(0.152)	(-0.526)	(0.316)	(0.090)
Stock Variance	0.196	-0.036	-0.088	0.022	-0.005
	(2.036)	(-0.367)	(-0.894)	(0.187)	(-0.046)
Industry Growth Rate	0.005	-0.051	0.097	-0.030	0.105
	(0.056)	(-0.543)	(1.038)	(-0.269)	(1.068)
News Sentiment	-0.019	-0.033	-0.122	-0.092	-0.010
	(-0.211)	(-0.357)	(-1.296)	(-0.957)	(-0.096)
Sentiment Volatility	0.019*	0.031***	0.028***	0.020**	0.019*
	(1.901)	(3.167)	(2.980)	(2.066)	(1.901)
R squared	0.126	0.104	0.119	0.106	0.104
Observations	143	141	138	135	132

Panel E: Other

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-0.052	-0.009	-0.002	0.150	-0.081
	(-0.599)	(-0.106)	(-0.022)	(1.697)	(-0.911)
Net Equity Expansion	-0.127	-0.106	-0.101	-0.015	-0.075
	(-1.480)	(-1.222)	(-1.146)	(-0.156)	(-0.768)
Stock Variance	0.125	-0.037	-0.051	0.025	0.094
	(1.307)	(-0.380)	(-0.519)	(0.217)	(0.893)
Industry Growth Rate	0.056	-0.067	0.093	-0.036	0.114
	(0.617)	(-0.718)	(0.997)	(-0.322)	(1.170)
News Sentiment	-0.109	-0.130	-0.175	-0.034	0.016
	(-1.206)	(-1.405)	(-1.867)	(-0.351)	(0.149)
Sentiment Volatility	0.010	0.016	0.014	0.009	0.011
	(1.005)	(1.652)	(1.481)	(0.945)	(1.090)
R squared	0.102	0.090	0.091	0.072	0.070
Observations	143	141	138	135	132

Table 3.8 reports the regressions for predictions with macroeconomic and sentiment control variables, including Inflation, Net Equity Expansion, Industry Growth Rate and News Sentiment. The sample period is from January 2009 to December 2020. The explanatory variable is Sentiment Volatility (STD) from our sentiment indices. These prediction results above show that STD usually can predict returns from three out of five industries, including Consumables, High Technology and Healthcare. The predictions are relatively robust, as we have checked multiple periods on a quarterly basis. The coefficients are all positive, which makes sense, as the sentiment dispersion measured by standard deviation represents the volatility risk, and higher risk usually corresponds to a higher return rate for any industry type. K represents the prediction length measured in months.

Table 3.8 tells us that sentiment volatility is another effective predictor for the three industry returns. In general, this is a reliable variable whose p values are significant regardless of which prediction month we choose, except the first month and the last one for Consumables and High Technology firms respectively. The inverted U shape finding still holds across these industries.

Table 3.9 Predictions with ARIMAX Aggregate Score

Panel A: Consumables					
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	18.041	-24.655	16.341	136.101	22.402
	(0.150)	(-0.178)	(0.133)	(0.905)	(0.165)
Net Equity Expansion	-13.000	1.997	-6.204	29.126	-7.695
	(-0.615)	(0.086)	(-0.231)	(1.062)	(-0.234)
Stock Variance	179.037**	9.577	-55.104	-34.212	85.239
	(2.480)	(0.082)	(-0.608)	(-0.428)	(0.543)
News Sentiment	-3.562	-3.503	-4.843**	-3.738*	-1.016
	(-1.578)	(-1.565)	(-2.182)	(-1.762)	(-0.381)
Industry Growth Rate	30.299	-51.641	21.177	-82.633	95.147
	(1.236)	(-1.269)	(0.644)	(-1.344)	(1.359)
Aggregate Score	2.356	5.197***	5.059***	4.434	3.785*
	(1.429)	(3.248)	(2.711)	(0.350)	(1.756)
Observations	143	141	138	135	132

Panel B: Manufacturing					
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-40.925	32.468	3.951	141.340	-57.828
	(-0.242)	(0.188)	(0.029)	(0.843)	(-0.372)
Net Equity Expansion	-9.143	-13.456	-20.522	7.513	-15.020
	(-0.316)	(-0.463)	(-0.642)	(0.233)	(-0.849)
Stock Variance	131.651	9.078	-34.818	-0.588	82.966
	(1.369)	(0.084)	(-0.298)	(-0.005)	(0.396)
News Sentiment	-2.636	-3.246	-5.368**	-1.237	-2.425
	(-0.783)	(-1.070)	(-2.042)	(-0.377)	(-0.799)
Industry Growth Rate	13.776	-9.381	11.845	-43.316	91.754
	(0.301)	(-0.164)	(0.257)	(-0.544)	(1.115)
Aggregate Score	1.529	2.411	2.326	2.450	1.714
	(0.708)	(1.199)	(0.958)	(1.133)	(0.938)
Observations	143	141	138	135	132

Panel C: High Technology					
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-20.392	-48.174	84.526	90.730	35.016

	(-0.144)	(-0.416)	(0.599)	(0.587)	(0.259)
Net Equity Expansion	-20.832	-23.310	-20.723	0.415	-38.771*
	(-1.431)	(-1.360)	(-0.705)	(0.013)	(-1.902)
Stock Variance	151.785	152.271**	-63.376	-26.135	117.578
	(1.478)	(2.046)	(-0.672)	(-0.234)	(0.687)
News Sentiment	-2.834	-1.799	-4.648*	-2.602	0.246
	(-1.179)	(-0.949)	(-1.778)	(-0.843)	(0.090)
Industry Growth Rate	14.800	-24.391	12.998	-89.617	121.582*
	(0.414)	(-0.589)	(0.317)	(-1.254)	(1.840)
Aggregate Score	3.790***	4.181***	5.287**	5.854***	3.286*
	(3.716)	(2.689)	(2.445)	(2.743)	(1.696)
Observations	143	141	138	135	132

Panel D: Healthcare

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	43.479	-40.587	-76.937	110.118	119.681
	(0.355)	(-0.320)	(-0.595)	(0.817)	(0.887)
Net Equity Expansion	5.556	4.020	-13.267	8.729	3.080
	(0.229)	(0.171)	(-0.527)	(0.285)	(0.104)
Stock Variance	121.941*	-2.725	-46.597	15.300	4.166
	(1.855)	(-0.037)	(-0.280)	(0.136)	(0.022)
News Sentiment	-0.855	-1.473	-3.550	-2.974	-0.657
	(-0.355)	(-0.549)	(-1.348)	(-1.106)	(-0.230)
Industry Growth Rate	-0.055	-14.128	25.905	-22.166	71.014
	(-0.001)	(-0.298)	(0.886)	(-0.380)	(1.104)
Aggregate Score	3.660*	5.483***	5.201**	4.223**	4.084**
	(1.906)	(2.955)	(2.384)	(2.241)	(2.037)
Observations	143	141	138	135	132

Panel E: Other

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-90.431	-3.286	6.228	254.444	-136.567
	(-0.474)	(-0.018)	(0.044)	(1.412)	(-0.925)
Net Equity Expansion	-46.387	-38.757	-34.805	-3.382	-23.732
	(-1.378)	(-1.250)	(-1.118)	(-0.101)	(-0.797)
Stock Variance	101.680	-21.829	-32.570	16.423	160.740
	(0.924)	(-0.166)	(-0.260)	(0.152)	(0.825)
News Sentiment	-3.633	-4.284	-5.542*	-1.399	0.061
	(-1.030)	(-1.357)	(-1.938)	(-0.407)	(0.019)
Industry Growth Rate	21.320	-23.965	31.638	-32.361	97.189
	(0.353)	(-0.554)	(0.548)	(-0.436)	(1.167)

Aggregate Score	2.287	3.413	3.044	2.555	2.797
	(0.906)	(1.466)	(1.151)	(1.055)	(1.189)
Observations	143	141	138	135	132

Table 3.9 reports the regression results in ARIMAX model with our aggregate score variable. The Aggregate Score remains effective in predicting High Technology and Healthcare industry returns throughout these prediction intervals. Between the two extremities there is Consumable return. The Weighted Individual Score, Weighted Aggregate Score and is ARIMAX models show very similar results. K represents the prediction length measured in months. K represents the prediction length measured in months.

Table 3.10 Predictions with ARIMAX Sentiment Volatility

Panel A: Consumables

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	11.070	-40.426	0.155	155.732	-19.339
	(0.092)	(-0.288)	(0.001)	(1.119)	(-0.158)
Net Equity Expansion	-12.223	3.739	-4.344	17.496	-10.739
	(-0.571)	(0.162)	(-0.165)	(0.658)	(-0.876)
Stock Variance	175.325**	1.328	-63.524	-24.670	-38.233
	(2.456)	(0.010)	(-0.626)	(-0.263)	(-0.294)
News Sentiment	-3.253	-2.826	-4.182*	-3.161	-3.572**
	(-1.455)	(-1.245)	(-1.835)	(-1.086)	(-2.422)
Industry Growth Rate	30.780	-50.523	22.237	-71.249	12.875
	(1.259)	(-1.257)	(0.697)	(-1.124)	(0.223)
STD	1.499	3.313***	3.246***	10.214	2.774***
	(1.486)	(3.315)	(2.848)	(1.502)	(3.599)
Observations	143	141	138	135	132

Panel B: Manufacturing

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-41.553	21.863	-4.302	130.408	-66.698
	(-0.243)	(0.125)	(-0.031)	(0.765)	(-0.406)
Net Equity Expansion	-9.435	-11.994	-19.505	9.085	-18.803
	(-0.323)	(-0.414)	(-0.616)	(0.282)	(-0.525)
Stock Variance	130.908	3.878	-39.034	-7.381	65.419
	(1.359)	(0.034)	(-0.316)	(-0.057)	(0.281)
News Sentiment	-2.462	-2.911	-5.059*	-0.952	-0.367
	(-0.732)	(-0.941)	(-1.865)	(-0.291)	(-0.105)
Industry Growth Rate	14.340	-9.050	12.286	-46.298	90.481
	(0.322)	(-0.159)	(0.270)	(-0.585)	(1.085)
STD	0.879	1.616	1.512	1.638	1.242

	(0.659)	(1.294)	(1.023)	(1.246)	(0.777)
Observations	143	141	138	135	132

Panel C: High Technology

	k = 1	k = 3	k = 6	k = 9	k = 12
Inflation	-32.268	-63.455	67.964	74.207	25.442
	(-0.219)	(-0.433)	(0.474)	(0.473)	(0.171)
Net Equity Expansion	-29.640	-14.094	-18.850	2.029	-37.647**
	(-1.125)	(-0.518)	(-0.652)	(0.065)	(-2.258)
Stock Variance	147.807	30.105	-72.026	-36.227	103.216
	(1.552)	(0.231)	(-0.704)	(-0.332)	(0.566)
News Sentiment	-4.066	-2.237	-3.958	-1.930	-0.185
	(-1.474)	(-0.870)	(-1.467)	(-0.632)	(-0.073)
Industry Growth Rate	10.779	-38.603	14.127	-91.133	121.047*
	(0.313)	(-1.093)	(0.354)	(-1.301)	(1.840)
STD	2.118*	3.591	3.384	3.689***	2.019*
	(1.748)	(2.977)	(2.590)	(2.892)	(1.763)
Observations	143	141	138	135	132

Panel D: Healthcare

	k = 1	k = 3	k = 6	k = 9	k = 12
Inflation	44.567	-50.472	-87.312	106.016	115.268
	(0.357)	(-0.391)	(-0.656)	(0.778)	(0.853)
Net Equity Expansion	4.327	4.282	-12.601	8.148	2.342
	(0.175)	(0.181)	(-0.503)	(0.264)	(0.079)
Stock Variance	121.273*	-9.302	-52.594	13.020	-6.425
	(1.824)	(-0.117)	(-0.309)	(0.115)	(-0.033)
News Sentiment	-0.457	-0.796	-2.904	-2.495	-0.273
	(-0.191)	(-0.294)	(-1.094)	(-0.950)	(-0.096)
Industry Growth Rate	1.462	-12.794	27.357	-18.693	74.581
	(0.033)	(-0.267)	(0.945)	(-0.324)	(1.143)
STD	2.040*	3.334***	3.187**	2.478**	2.432*
	(1.727)	(2.882)	(2.290)	(2.128)	(1.935)
Observations	143	141	138	135	132

Panel E: Other

	k = 1	k = 3	k = 6	k = 9	k = 12
Inflation	-93.973	-16.195	-3.255	257.127	-138.346
	(-0.511)	(-0.087)	(-0.022)	(1.390)	(-0.929)
Net Equity Expansion	-46.308	-37.105	-33.736	-4.887	-26.806
	(-1.437)	(-1.197)	(-1.095)	(-0.146)	(-0.891)
Stock Variance	100.338	-28.311	-37.528	18.347	152.492
	(0.913)	(-0.203)	(-0.281)	(0.168)	(0.735)

News Sentiment	-3.373	-3.824	-5.145*	-1.113	0.302
	(-0.950)	(-1.181)	(-1.740)	(-0.321)	(0.094)
Industry Growth Rate	20.635	-23.377	32.291	-27.241	97.639
	(0.358)	(-0.551)	(0.565)	(-0.369)	(1.147)
STD	1.391	2.237	1.947	1.378	1.626
	(0.893)	(1.559)	(1.203)	(0.931)	(1.104)
Observations	143	141	138	135	132

Table 3.10 show similar conclusions to that of ARIMAX model with Aggregate Scores, in Healthcare returns regardless the prediction length, followed by Consumables and High Technology industries. However, the variable still fails to predict Manufacturing or Other returns at any time listed. K represents the prediction length measured in months.

Table 3.11 Prediction Comparisons with Aggregate Score

Table 3.11 compares prediction performances of three methods (LSTM, MLP and ARIMAX) across these industries, with roughly a quarterly increment in forecast duration measured in months ahead. ARIMAX is often the most performant one. The predictor is Aggregate Score. K represents the prediction length measured in months.

		k = 1			**k = 3**			**k = 6**			**k = 9**			**k = 12**		
Aggregate Score		**ARIMAX**	**LSTM**	**MLP**	**ARIMAX**	**LSTM**	**MLP**	**ARIMAX**	**LSTM**	**MLP**	**ARIMAX**	**LSTM**	**MLP**	**ARIMAX**	**LSTM**	**MLP**
Consumables	MSE	1.242	19.476	93.520	1.290	19.622	56.193	1.682	10.173	35.644	1.159	14.017	63.206	1.356	21.367	79.656
	MAE	0.866	3.446	8.506	0.771	3.100	6.510	0.859	2.604	4.965	0.801	2.723	6.656	0.868	3.460	7.386
	MAPE	1.667	1.808	6.460	1.226	1.882	7.304	1.338	1.354	3.501	6.333	4.559	12.053	3.225	4.096	6.442
Manufacturing	MSE	1.117	23.986	112.743	1.248	22.144	89.872	0.951	17.979	41.564	1.241	25.616	71.172	1.513	31.808	89.973
	MAE	0.804	3.726	9.132	0.819	3.404	8.204	0.793	3.399	5.373	0.767	3.369	6.608	0.857	3.966	7.742
	MAPE	2.051	2.157	13.644	1.298	1.601	7.701	8.919	2.725	27.743	3.494	3.460	9.626	3.939	3.519	8.008
High Technology	MSE	1.197	22.832	85.510	1.381	23.815	61.363	1.368	17.469	36.905	0.924	22.337	66.380	1.368	24.443	78.417
	MAE	0.879	3.832	7.735	0.942	3.767	6.792	0.841	3.457	4.936	0.714	3.514	6.828	0.867	3.763	7.470
	MAPE	2.113	2.983	9.035	1.353	1.287	3.733	1.143	1.212	3.300	1.187	1.419	3.709	1.402	1.276	3.587
Healthcare	MSE	1.354	20.423	87.043	1.233	17.485	41.314	1.487	19.198	28.061	0.842	15.702	59.759	1.269	19.621	60.670
	MAE	0.902	3.478	8.278	0.847	3.124	5.547	0.958	3.376	4.128	0.721	3.122	6.508	0.947	3.636	6.666
	MAPE	2.855	2.224	13.069	1.480	1.555	7.243	1.509	1.455	4.263	2.297	2.213	13.591	1.570	1.651	4.545
Other	MSE	1.185	30.111	152.686	1.108	27.988	125.536	1.013	15.950	54.609	1.015	23.694	81.912	1.604	32.327	105.070
	MAE	0.851	4.473	10.907	0.799	3.947	9.553	0.760	3.150	6.224	0.672	3.141	7.367	0.876	4.055	8.262
	MAPE	1.987	1.535	7.354	1.152	1.310	5.579	1.405	1.246	5.920	2.260	2.342	6.009	3.079	1.744	7.310

Table 3.12 Prediction comparisons with Sentiment Volatility

Table 3.12 compares prediction performances of three methods (LSTM, MLP and ARIMAX) across these industries, with roughly a quarterly increment in forecast duration measured in months ahead. ARIMAX is often the most performant one. The predictor is Sentiment Volatility.

		k = 1			**k = 3**			**k = 6**			**k = 9**			**k = 12**		
		ARIMAX	**LSTM**	**MLP**	**ARIMAX**	**LSTM**	**MLP**	**ARIMAX**	**LSTM**	**MLP**	**ARIMAX**	**LSTM**	**MLP**	**ARIMAX**	**LSTM**	**MLP**
Consumables	MSE	1.244	20.673	81.685	1.254	19.425	51.116	1.578	14.024	33.063	1.167	13.591	60.017	1.751	22.214	68.944
	MAE	0.868	3.586	7.882	0.767	3.151	6.110	0.847	3.009	4.688	0.759	2.651	6.164	1.019	3.533	6.844
	MAPE	1.665	1.994	5.982	1.228	1.826	6.666	1.325	1.299	3.434	4.458	4.057	10.243	2.826	4.953	6.484
Manufacturing	MSE	1.116	25.567	109.867	1.256	22.746	92.277	0.940	15.502	38.350	1.428	28.355	62.630	1.664	32.933	80.538
	MAE	0.806	3.915	9.069	0.841	3.426	8.293	0.788	3.207	5.237	0.851	3.425	6.014	0.914	4.015	7.279
	MAPE	2.005	3.019	12.370	1.828	1.397	8.069	9.334	3.585	29.969	5.236	4.779	10.628	4.518	3.640	5.911
High Technology	MSE	1.134	24.299	74.892	1.295	23.491	54.749	1.307	16.774	38.985	1.104	21.383	62.640	1.420	22.778	67.331
	MAE	0.867	3.922	7.128	0.905	3.777	6.274	0.837	3.361	5.121	0.804	3.358	6.316	0.885	3.684	6.883
	MAPE	2.211	3.233	8.127	1.219	1.226	3.366	1.147	1.226	3.404	1.230	1.337	3.419	1.302	1.196	3.201
Healthcare	MSE	1.353	21.874	70.671	1.239	17.837	38.839	1.448	24.262	30.171	0.840	16.388	50.896	1.544	23.753	57.693
	MAE	0.901	3.601	7.429	0.844	3.200	5.437	0.951	3.668	4.273	0.722	3.108	5.817	1.043	3.881	6.606
	MAPE	2.958	2.747	11.345	1.302	1.612	6.906	1.477	1.728	4.293	2.217	2.051	12.201	1.597	1.672	4.485
Other	MSE	1.115	31.240	119.379	1.080	26.799	123.719	1.001	17.955	50.281	0.920	25.201	70.915	1.579	34.757	93.857
	MAE	0.815	4.551	9.383	0.790	3.877	9.363	0.757	3.373	6.000	0.647	3.222	6.597	0.869	4.216	7.709
	MAPE	1.954	1.866	5.745	1.274	1.051	5.602	1.407	1.082	5.729	1.765	2.306	5.110	2.902	1.427	7.229

Tables 3.11, 3.12, and 3.13 all show ARIMAX model usually outperforms LSTM which mostly outperforms MLP regardless prediction length or measure. The measures used include mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). This finding is not surprising, since in lower dimensions traditional statistical models, such as autoregression methods often are more performant than more complex machine learning models.

Table 3.13 Predictions with Recession Indicator and Aggregate Score

K represents the prediction length measured in months.

Panel A: Consumables					
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Recession	-0.467	1.158	1.788	-0.071	0.5491
	(-0.435)	(1.044)	(1.579)	(-0.063)	(0.443)
Aggregate Score	0.112*	0.227***	0.217***	0.187**	0.132
	(1.709)	(3.335)	(3.130)	(2.455)	(1.599)
Inflation	-0.023	-0.204	-0.072	0.420	0.144
	(-0.068)	(-0.575)	(-0.201)	(1.175)	(0.397)
Net Equity Expansion	-0.078	0.179	0.031	0.218	-0.246
	(-0.224)	(0.497)	(0.084)	(0.554)	(-0.614)
Stock Variance	0.801*	-0.067	-0.292	-0.151	0.413
	(1.823)	(-0.146)	(-0.624)	(-0.286)	(0.900)
News Sentiment	-0.489	-0.440	-0.673*	-0.461	-0.278
	(-1.315)	(-1.148)	(-1.706)	(-1.104)	(-0.643)
Industry Growth Rate	0.745*	-0.567	0.344	-0.235	0.429
	(1.809)	(-1.334)	(0.799)	(-0.503)	(1.024)
Recession*Aggregate Score	1.224	-0.645	-1.614	0.273	-1.152
	(1.069)	(-0.546)	(-1.339)	(0.237)	(-1.042)
Adj. R-squared	0.192	0.130	0.109	0.077	0.080
Observations	143	141	138	135	132
Panel B: Manufacturing					
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Recession	-1.995	2.458	1.380	0.324	0.163
	(-1.532)	(1.883)	(1.048)	(0.247)	(0.113)
Aggregate Score	0.077	0.133	0.107	0.100	0.045
	(0.961)	(1.659)	(1.327)	(1.125)	(0.468)
Inflation	-0.140	-0.143	-0.093	0.385	-0.065
	(-0.336)	(-0.342)	(-0.223)	(0.923)	(-0.155)
Net Equity Expansion	-0.063	0.066	-0.195	0.141	-0.450
	(-0.150)	(0.156)	(-0.457)	(0.309)	(-0.969)
Stock Variance	0.272	-0.084	-0.213	0.005	0.457
	(0.506)	(-0.155)	(-0.391)	(0.008)	(0.859)
News Sentiment	-0.374	-0.265	-0.778	-0.138	-0.216
	(-0.831)	(-0.586)	(-1.694)	(-0.284)	(-0.431)

Industry Growth Rate	0.594	0.151	0.235	-0.250	0.379
	(1.190)	(0.302)	(0.469)	(-0.459)	(0.779)
Recession*Aggregate Score	2.944**	-1.664	-1.157	-0.240	-0.859
	(2.122)	(-1.197)	(-0.825)	(-0.179)	(-0.671)
Adj. R-squared	0.067	0.032	0.019	-0.019	-0.005
Observations	143	141	138	135	132
Panel C: High Technology					
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Recession	-0.272	2.077	1.664	-0.173	-0.680
	(-0.217)	(1.619)	(1.272)	(-0.133)	(-0.482)
Aggregate Score	0.161**	0.257***	0.225***	0.229**	0.100
	(2.089)	(3.254)	(2.817)	(2.606)	(1.071)
Inflation	-0.195	-0.346	0.135	0.286	0.275
	(-0.485)	(-0.844)	(0.326)	(0.692)	(0.668)
Net Equity Expansion	-0.272	-0.002	-0.190	-0.025	-0.688
	(-0.666)	(-0.006)	(-0.448)	(-0.055)	(-1.512)
Stock Variance	0.508	0.118	-0.364	-0.142	0.756
	(0.978)	(0.222)	(-0.673)	(-0.233)	(1.449)
News Sentiment	-0.575	-0.254	-0.646	-0.433	-0.085
	(-1.322)	(-0.571)	(-1.416)	(-0.897)	(-0.172)
Industry Growth Rate	0.576	-0.322	0.232	-0.625	0.447
	(1.195)	(-0.655)	(0.467)	(-1.159)	(0.938)
Recession*Aggregate Score	1.352	-1.360	-1.480	0.043	-0.364
	(1.010)	(-0.995)	(-1.062)	(0.032)	(-0.290)
Adj. R-squared	0.178	0.132	0.100	0.070	0.094
Observations	143	141	138	135	132
Panel D: Healthcare					
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Recession	-1.886	1.130	1.291	0.110	0.545
	(-1.655)	(0.982)	(1.109)	(0.095)	(0.437)
Aggregate Score	0.157**	0.250***	0.226***	0.166**	0.136
	(2.243)	(3.528)	(3.169)	(2.116)	(1.642)
Inflation	0.124	-0.269	-0.350	0.323	0.456
	(0.340)	(-0.730)	(-0.949)	(0.877)	(1.253)
Net Equity Expansion	0.129	0.236	-0.062	0.122	-0.124
	(0.349)	(0.632)	(-0.165)	(0.301)	(-0.308)
Stock Variance	0.289	-0.344	-0.375	0.154	0.233
	(0.613)	(-0.722)	(-0.781)	(0.282)	(0.505)
News Sentiment	-0.091	-0.038	-0.441	-0.451	-0.266
	(-0.232)	(-0.094)	(-1.088)	(-1.048)	(-0.612)
Industry Growth Rate	0.329	0.052	0.503	-0.155	0.246
	(0.753)	(0.117)	(1.137)	(-0.323)	(0.584)
Recession*Aggregate Score	2.677**	-0.388	-0.903	-0.182	-1.350
	(2.205)	(-0.316)	(-0.728)	(-0.153)	(-1.214)
Adj. R-squared	0.115	0.070	0.077	0.054	0.079
Observations	143	141	138	135	132
Panel E: Other					
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Recession	-0.447	3.294**	2.743*	-1.035	0.714
	(-0.302)	(2.302)	(1.932)	(-0.736)	(0.463)
Aggregate Score	0.130	0.189**	0.132	0.105	0.094
	(1.436)	(2.153)	(1.521)	(1.102)	(0.914)
Inflation	-0.430	-0.347	-0.108	0.762	-0.301
	(-0.910)	(-0.759)	(-0.241)	(1.703)	(-0.670)
Net Equity Expansion	-0.450	-0.194	-0.408	-0.070	-0.517

	(-0.937)	(-0.417)	(-0.887)	(-0.142)	(-1.039)
Stock Variance	0.007	-0.400	0.105	-0.094	0.670
	(0.012)	(-0.677)	(0.178)	(-0.143)	(1.176)
News Sentiment	-0.370	-0.316	-0.804	-0.279	-0.118
	(-0.721)	(-0.638)	(-1.623)	(-0.534)	(-0.220)
Industry Growth Rate	0.848	0.087	0.345	-0.242	0.420
	(1.495)	(0.159)	(0.638)	(-0.414)	(0.805)
Recession*Aggregate Score	1.851	-2.097	-2.918*	1.148	-1.445
	(1.173)	(-1.376)	(-1.930)	(0.794)	(-1.052)
Adj. R-squared	0.076	0.083	0.062	0.021	0.031
Observations	143	141	138	135	132

Table 3.14 Predictions with Recession Indicator and Sentiment Volatility

			Panel A: Consumables		
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Recession	-11.334**	3.548	9.346*	-7.034	26.420
	(-2.389)	(0.712)	(1.762)	(-0.960)	(1.050)
STD	0.076**	0.128***	0.112***	0.110***	0.065
	(2.008)	(3.269)	(2.984)	(2.636)	(1.618)
Inflation	-0.029	-0.223	-0.093	0.380	0.123
	(-0.084)	(-0.627)	(-0.260)	(1.055)	(0.338)
Net Equity Expansion	-0.109	0.182	0.058	0.200	-0.236
	(-0.320)	(0.506)	(0.158)	(0.510)	(-0.591)
Stock Variance	-0.129	0.101	0.295	-0.959	0.428
	(-0.214)	(0.160)	(0.445)	(-0.973)	(0.928)
News Sentiment	-0.263	-0.386	-0.685	-0.509	-0.241
	(-0.701)	(-0.981)	(-1.718)	(-1.184)	(-0.560)
Industry Growth Rate	1.717***	-0.806	-0.380	-0.180	0.360
	(2.948)	(-1.319)	(-0.588)	(-0.378)	(0.863)
Recession*STD	13.174**	-3.357	-10.003*	7.619	-27.055
	(2.507)	(-0.609)	(-1.701)	(0.974)	(-1.077)
Adj. R-squared	0.221	0.132	0.119	0.084	0.083
Observations	143	141	138	135	132
			Panel B: Manufacturing		
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Recession	-10.629*	5.084	6.392	-2.482	32.699
	(-1.807)	(0.863)	(1.031)	(-0.290)	(1.124)
STD	0.051	0.072	0.054	0.060	0.028
	(1.072)	(1.569)	(1.225)	(1.234)	(0.590)
Inflation	-0.167	-0.131	-0.099	0.359	-0.101
	(-0.398)	(-0.311)	(-0.237)	(0.854)	(-0.240)
Net Equity Expansion	-0.088	0.061	-0.180	0.147	-0.424
	(-0.207)	(0.143)	(-0.421)	(0.321)	(-0.920)
Stock Variance	-0.313	0.039	0.174	-0.374	0.443
	(-0.420)	(0.052)	(0.225)	(-0.325)	(0.831)
News Sentiment	-0.165	-0.279	-0.804*	-0.160	-0.199
	(-0.354)	(-0.599)	(-1.726)	(-0.318)	(-0.400)
Industry Growth Rate	1.365*	-0.100	-0.241	-0.254	0.324
	(1.891)	(-0.138)	(-0.320)	(-0.458)	(0.671)
Recession*STD	12.277*	-4.518	-6.698	2.740	-33.378

	(1.885)	(-0.693)	(-0.975)	(0.300)	(-1.149)
Adj. R-squared	0.059	0.025	0.021	-0.017	0.003
Observations	143	141	138	135	132
		Panel C: High Technology			
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Recession	-10.706*	7.040	10.064	-7.5576	22.572
	(-1.916)	(1.220)	(1.643)	(-0.892)	(0.791)
STD	0.104**	0.140***	0.116***	0.135***	0.053
	(2.324)	(3.094)	(2.676)	(2.796)	(1.155)
Inflation	-0.208	-0.355	0.109	0.235	0.239
	(-0.521)	(-0.861)	(0.262)	(0.565)	(0.578)
Net Equity Expansion	-0.302	-0.003	-0.159	-0.036	-0.678
	(-0.750)	(-0.007)	(-0.377)	(-0.079)	(-1.497)
Stock Variance	-0.381	0.483	0.313	-1.037	0.721
	(-0.538)	(0.660)	(0.410)	(-0.911)	(1.376)
News Sentiment	-0.331	-0.241	-0.659	-0.482	-0.055
	(-0.749)	(-0.529)	(-1.430)	(-0.969)	(-0.113)
Industry Growth Rate	1.503**	-0.798	-0.571	-0.587	0.418
	(2.192)	(-1.127)	(-0.766)	(-1.070)	(0.884)
Recession*STD	12.787**	-6.898	-10.820	7.841	-23.628
	(2.066)	(-1.080)	(-1.593)	(0.867)	(-0.828)
Adj. R-squared	0.196	0.133	0.111	0.077	0.099
Observations	143	141	138	135	132
		Panel D: Healthcare			
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Recession	-10.910**	3.638	6.125	-3.451	24.664
	(-2.119)	(0.701)	(1.117)	(-0.455)	(0.971)
STD	0.0920**	0.134	0.113***	0.088**	0.061
	(2.220)	(3.278)	(2.933)	(2.022)	(1.499)
Inflation	0.116	-0.270	-0.355	0.319	0.462
	(0.315)	(-0.726)	(-0.958)	(0.857)	(1.253)
Net Equity Expansion	0.078	0.215	-0.065	0.088	-0.135
	(0.211)	(0.573)	(-0.173)	(0.216)	(-0.335)
Stock Variance	-0.330	-0.116	0.029	-0.264	0.298
	(-0.506)	(-0.177)	(0.043)	(-0.259)	(0.639)
News Sentiment	0.141	0.016	-0.419	-0.468	-0.226
	(0.345)	(0.039)	(-1.017)	(-1.052)	(-0.520)
Industry Growth Rate	1.125*	-0.204	0.029	-0.124	0.181
	(1.780)	(-0.320)	(0.043)	(-0.253)	(0.429)
Recession*STD	12.411**	-3.283	-6.334	3.557	-25.523
	(2.177)	(-0.571)	(-1.042)	(0.439)	(-1.006)
Adj. R-squared	0.107	0.065	0.077	0.052	0.075
Observations	143	141	138	135	132
		Panel E: Other			
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Recession	-7.638	7.360	8.214	-10.409	35.029
	(-1.147)	(1.139)	(1.217)	(-1.132)	(1.120)
STD	0.078	0.101**	0.066	0.067	0.041
	(1.462)	(1.992)	(1.378)	(1.270)	(0.811)
Inflation	-0.439	-0.328	-0.083	0.737	-0.298
	(-0.924)	(-0.711)	(-0.183)	(1.633)	(-0.656)

Net Equity Expansion	-0.485	-0.206	-0.406	-0.119	-0.526
	(-1.011)	(-0.442)	(-0.874)	(-0.242)	(-1.060)
Stock Variance	-0.516	-0.167	0.406	-1.057	0.743
	(-0.611)	(-0.204)	(0.482)	(-0.855)	(1.294)
News Sentiment	-0.184	-0.339	-0.866*	-0.367	-0.085
	(-0.350)	(-0.663)	(-1.706)	(-0.680)	(-0.160)
Industry Growth Rate	1.483*	-0.300	-0.173	-0.098	0.355
	(1.814)	(-0.379)	(-0.211)	(-0.164)	(0.685)
Recession*STD	9.634	-6.553	-8.846	11.033	-35.793
	(1.306)	(-0.916)	(-1.182)	(1.124)	(-1.145)
Adj. R-squared	0.076	0.075	0.046	0.024	0.032
Observations	143	141	138	135	132

K represents the prediction length measured in months.

Tables 3.13 and 3.14 both show that our aggregate sentiment score and sentiment standard deviation are informative indicators to predict Consumables, High Technology and Healthcare industry returns up to the third quarter. We introduce the recession indicator along with the interaction term. However, std alone is more performant. None of the variables can successfully predict any return after one year.

Table 3.15 Predictions with TEU and Aggregate Score

Panel A: Consumables					
	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-0.160	0.025	-0.223	0.392	0.256
	(-0.429)	(0.062)	(-0.563)	(0.993)	(0.643)
Net Equity Expansion	-0.094	0.229	0.084	0.039	-0.143
	(-0.260)	(0.596)	(0.218)	(0.095)	(-0.340)
TEU- Aggregate Score	0.061	0.120***	0.134***	0.104**	0.095*
	(1.474)	(2.780)	(3.167)	(2.111)	(1.748)
Stock Variance	1.249***	0.019	-0.518	-0.101	-0.050
	(3.142)	(0.044)	(-1.222)	(-0.191)	(-0.117)
Industry Growth Rate	0.516	-0.743*	0.514	-0.106	0.488
	(1.335)	(-1.810)	(1.259)	(-0.216)	(1.205)
News Sentiment	-0.555	-0.423	-0.505	-0.648	-0.440
	(-1.467)	(-1.055)	(-1.247)	(-1.532)	(-1.024)
Adj. R-squared	0.184	0.107	0.100	0.079	0.083
Observations	114	112	109	106	103
Panel B: Manufacturing					

	k = 1	k = 3	k = 6	k = 9	k = 12
Inflation	-0.275	0.099	-0.204	0.171	0.108
	(-0.612)	(0.212)	(-0.460)	(0.389)	(0.246)
Net Equity Expansion	-0.258	-0.049	-0.407	-0.379	-0.455
	(-0.593)	(-0.109)	(-0.946)	(-0.831)	(-0.985)
TEU- Aggregate Score	0.015	0.049	0.044	0.029	0.015
	(0.295)	(0.970)	(0.921)	(0.527)	(0.249)
Stock Variance	0.931*	0.074	-0.320	-0.073	0.210
	(1.949)	(0.149)	(-0.676)	(-0.125)	(0.446)
Industry Growth Rate	0.117	-0.046	0.328	-0.189	0.344
	(0.252)	(-0.094)	(0.719)	(-0.345)	(0.772)
News Sentiment	-0.380	-0.233	-0.595	-0.276	-0.383
	(-0.835)	(-0.495)	(-1.314)	(-0.585)	(-0.809)
Adj. R-squared	0.034	-0.024	0.001	-0.021	-0.003
Observations	114	112	109	106	103

Panel C: High Technology

	k = 1	k = 3	k = 6	k = 9	k = 12
Inflation	-0.289	-0.070	-0.059	0.043	0.356
	(-0.675)	(-0.156)	(-0.134)	(0.099)	(0.830)
Net Equity Expansion	-0.255	0.145	-0.130	-0.296	-0.519
	(-0.615)	(0.337)	(-0.305)	(-0.665)	(-1.150)
TEU-Aggregate Score	0.087*	0.141***	0.143***	0.133**	0.074
	(1.848)	(2.929)	(3.034)	(2.480)	(1.260)
Stock Variance	0.958**	0.196	-0.591	-0.444	0.313
	(2.104)	(0.413)	(-1.257)	(-0.774)	(0.680)
Industry Growth Rate	0.137	-0.477	0.432	-0.655	0.397
	(0.310)	(-1.036)	(0.953)	(-1.225)	(0.911)
News Sentiment	-0.645	-0.218	-0.541	-0.633	-0.329
	(-1.488)	(-0.485)	(-1.204)	(-1.374)	(-0.711)
Adj. R-squared	0.157	0.106	0.116	0.106	0.120
Observations	114	112	109	106	103

Panel D: Healthcare

	k = 1	k = 3	k = 6	k = 9	k = 12
Inflation	0.058	-0.052	-0.392	0.257	0.537
	(0.146)	(-0.126)	(-0.949)	(0.625)	(1.287)
Net Equity Expansion	0.241	0.476	0.101	-0.026	-0.015
	(0.622)	(-0.126)	(0.250)	(-0.061)	(-0.034)
TEU-Aggregate Score	0.094**	0.153***	0.141***	0.106**	0.099*
	(2.131)	(3.405)	(3.181)	(2.071)	(1.745)
Stock Variance	0.915**	-0.266	-0.367	-0.071	-0.043
	(2.149)	(-0.605)	(-0.827)	(-0.130)	(-0.097)
Industry Growth Rate	-0.061	-0.202	0.650	-0.152	0.258
	(-0.147)	(-0.472)	(1.522)	(-0.297)	(0.609)
News Sentiment	-0.122	-0.048	-0.313	-0.786	-0.443

	(-0.302)	(-0.116)	(-0.739)	(-1.784)	(-0.986)
Adj. R-squared	0.109	0.072	0.094	0.079	0.080
Observations	114	112	109	106	103

Panel E: Other

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-0.394	0.019	-0.0874	0.553	-0.231
	(-0.803)	(0.038)	(-0.184)	(1.179)	(-0.494)
Net Equity Expansion	-0.344	-0.023	-0.220	-0.278	-0.370
	(-0.725)	(-0.047)	(-0.477)	(-0.573)	(-0.752)
TEU-Aggregate Score	0.064	0.100*	0.096*	0.062	0.074
	(1.189)	(1.853)	(1.888)	(1.048)	(1.159)
Stock Variance	0.529	-0.218	-0.338	0.089	0.205
	(1.015)	(-0.412)	(-0.664)	(0.142)	(0.409)
Industry Growth Rate	0.202	-0.125	0.630	-0.141	0.605
	(0.399)	(-0.243)	(1.285)	(-0.242)	(1.276)
News Sentiment	-0.448	-0.448	-0.640	-0.540	-0.498
	(-0.902)	(-0.891)	(-1.316)	(-1.073)	(-0.989)
Adj. R-squared	0.036	0.017	0.050	0.036	0.054
Observations	114	112	109	106	103

K represents the prediction length measured in months.

Table 3.16 Predictions with TEU and Sentiment Volatility

Panel A: Consumables

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-0.159	0.013	-0.230	0.380	0.257
	(-0.424)	(0.032)	(-0.582)	(0.963)	(0.643)
Net Equity Expansion	-0.081	0.288	0.142	0.089	-0.121
	(-0.218)	(0.733)	(0.361)	(0.215)	(-0.281)
TEU-STD	0.116	0.238***	0.253***	0.171**	0.138*
	(1.438)	(2.851)	(3.212)	(2.185)	(1.706)
Stock Variance	1.229***	-0.042	-0.582	-0.179	-0.074
	(3.047)	(-0.098)	(-1.355)	(-0.332)	(-0.169)
Industry Growth Rate	0.525	-0.726*	0.538	-0.125	0.491
	(1.358)	(-1.772)	(1.320)	(-0.255)	(1.212)
News Sentiment	-0.508	-0.322	-0.412	-0.620	-0.416
	(-1.324)	(-0.795)	(-1.012)	(-1.468)	(-0.970)
Adj. R-squared	0.183	0.110	0.103	0.081	0.082
Observations	114	112	109	106	103

Panel B: Manufacturing

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-0.261	0.097	-0.200	0.175	0.115

	(-0.581)	(0.207)	(-0.451)	(0.396)	(0.262)
Net Equity Expansion	-0.284	-0.031	-0.402	-0.380	-0.468
	(-0.639)	(-0.067)	(-0.914)	(-0.815)	(-0.990)
TEU-STD	0.017	0.096	0.077	0.043	0.016
	(0.174)	(0.973)	(0.876)	(0.488)	(0.181)
Stock Variance	0.945*	0.053	-0.331	-0.078	0.224
	(1.951)	(0.105)	(-0.689)	(-0.130)	(0.463)
Industry Growth Rate	0.121	-0.038	0.336	-0.183	0.350
	(0.262)	(-0.079)	(0.737)	(-0.334)	(0.785)
News Sentiment	-0.378	-0.193	-0.568	-0.268	-0.377
	(-0.821)	(-0.406)	(-1.246)	(-0.568)	(-0.798)
Adj. R-squared	0.033	-0.024	0.000	-0.022	-0.004
Observations	114	112	109	106	103

Panel C: High Technology

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	-0.290	-0.081	-0.072	0.030	0.360
	(-0.675)	(-0.181)	(-0.164)	(0.070)	(0.837)
Net Equity Expansion	-0.229	0.208	-0.056	-0.238	-0.510
	(-0.540)	(0.473)	(-0.129)	(-0.525)	(-1.102)
TEU-STD	0.169*	0.279***	0.272***	0.217**	0.104
	(1.831)	(2.978)	(3.127)	(2.540)	(1.199)
Stock Variance	0.925**	0.128	-0.667	-0.536	0.303
	(2.002)	(0.267)	(-1.401)	(-0.914)	(0.640)
Industry Growth Rate	0.150	-0.456	0.457	-0.675	0.402
	(0.338)	(-0.993)	(1.011)	(-1.261)	(0.922)
News Sentiment	-0.575	-0.100	-0.439	-0.597	-0.309
	(-1.310)	(-0.221)	(-0.975)	(-1.298)	(-0.669)
Adj. R-squared	0.157	0.109	0.121	0.108	0.119
Observations	114	112	109	106	103

Panel D: Healthcare

	k = 1	**k = 3**	**k = 6**	**k = 9**	**k = 12**
Inflation	0.076	-0.055	-0.394	0.252	0.547
	(0.188)	(-0.132)	(-0.952)	(0.611)	(1.306)
Net Equity Expansion	0.229	0.525	0.148	0.010	-0.014
	(0.577)	(1.279)	(0.359)	(0.023)	(-0.030)
TEU-STD	0.167*	0.295***	0.261***	0.170**	0.136
	(1.933)	(3.373)	(3.163)	(2.080)	(1.616)
Stock Variance	0.905**	-0.326	-0.424	-0.135	-0.046
	(2.088)	(-0.729)	(-0.943)	(-0.239)	(-0.100)
Industry Growth Rate	-0.045	-0.179	0.675	-0.161	0.268
	(-0.107)	(-0.417)	(1.582)	(-0.314)	(0.631)
News Sentiment	-0.061	0.073	-0.218	-0.757	-0.415
	(-0.147)	(0.172)	(-0.512)	(-1.719)	(-0.924)
Adj. R-squared	0.103	0.071	0.093	0.079	0.076

Observations	114	112	109	106	103

Panel E: Other

	k = 1	k = 3	k = 6	k = 9	k = 12
Inflation	-0.378	0.015	-0.090	0.559	-0.217
	(-0.768)	(0.030)	(-0.189)	(1.190)	(-0.463)
Net Equity Expansion	-0.363	0.013	-0.185	-0.279	-0.384
	(-0.746)	(0.026)	(-0.391)	(-0.561)	(-0.762)
TEU-STD	0.110	0.195*	0.179*	0.092	0.096
	(1.045)	(1.850)	(1.890)	(0.980)	(1.021)
Stock Variance	0.529	-0.260	-0.380	0.075	0.219
	(0.999)	(-0.484)	(-0.735)	(0.117)	(0.425)
Industry Growth Rate	0.215	-0.110	0.648	-0.132	0.617
	(0.423)	(-0.214)	(1.322)	(-0.225)	(1.300)
News Sentiment	-0.409	-0.368	-0.575	-0.523	-0.475
	(-0.812)	(-0.723)	(-1.175)	(-1.040)	(-0.944)
Adj. R-squared	0.033	0.017	0.050	0.035	0.051
Observations	114	112	109	106	103

K represents the prediction length measured in months.

Tables 3.15 and 3.16 show that the hybrid variables TEU-AS and TEU-STD are mostly useful to predict Consumables, High Technology and Healthcare industry returns, with the former being more informative since it is also helpful to predict Other return in short run and Healthcare return next year. This result is reasonable, as Aggregate Score is a more specific and straightforward measure compared to the standard deviation component, which should translate to stronger prediction power. TEU-IS, TEU-WIS and TEU-WAS show very similar results to TEU-AS.

CHAPTER 7

CONCLUSIONS

7.1 Limitations

The part related to COVID-19 statistics includes several limitations. The actual number of the infection is very likely to be underestimated in many countries. Likewise, the number of deaths may not be entirely reliable. Due to omitted cases, different ways used to calculate these statistics, and even intentionally misleading information, it is difficult to confirm the accuracy of the raw data. Due to the lack of data availability, many countries or districts are not covered. This may lead to bias in our estimations. Some attributes do not chronologically match. For example, the population and pandemic statistics are not synchronized. The date of the webpage displaying country population is 2020; that is not necessarily the date of the population measurements, and the pandemic statistics are up to 10/25/2022. However, since usually a country does not rapidly change the population within a few months, this should not cause significant bias.

Due to the shared dataset between mispricing and industry return sections, they have limitations in common. There are only around 63% firms with valid tweets. Therefore, our sample cannot cover every firm in the S&P 1500 list. This research only focuses on English tweets, while other languages have been ignored. Third, certain information from a tweet, such as emojis and symbols, have not been incorporated in our study. This is because most emojis are special symbols, which are filtered out by our program. Also, current textual analysis has not overcome technical difficulties in understanding verbal expression fully. There are existing challenges to be solved, such as how to accurately interpret sarcasm, negation, word ambiguity and multipolarity. As much as we would like to fix such issues, there is extensive work to be done, which is beyond the scope of this study.

7.2 Implications

For over two decades, economic leaders have been using the Global Entrepreneurship Monitor (GEM), which is built on the HCD, to guide regional economy policy (Al-Kadi, 2017), and the latest GEM report presents the entrepreneurship policy roadmap for each of the fifty participating economies (Bosma et al., 2019). We recommend a similar culture-specific policy roadmap for pandemic response. Because political borders are more porous to pandemic effects than to economic effects, such roadmaps may be even more important to public health policy than to economic policy. Because IVR along with LTO show statistical significance with the death rate, and PDI and MAS have significance for the infection rate, there may be causal relations. Changing some relevant behaviors of IVR and may help reduce COVID-19 spread. However, public health officials may find it helpful to practice certain strategies, based on some relevant culture shifts from HCD. It may be worthy to attempt a MAS shift. Let people know being masculine has nothing to do with not wearing mask, as identification with norms of masculinity has a significant influence on affective responses toward mask wearing (Palmer & Peterson, 2020). For many people, words or behaviors from their idols usually override opinions from the rest. Public health marketing could thus cooperate with celebrities in the entertainment industry, political and business leaders, and other social idols, to ameliorate the pandemic. Policy makers may have to enforce laws to restrict crowd indulgence, such as parties, parades, and demonstrations, which in essence is equivalent to increasing a society's PDI and decreasing its IVR. Adjusting LTO may be beneficial to fight COVID-19. While many people may take pride in their time-honored traditions and norms, it is advisable to embrace societal change when facing this unprecedented pandemic. More pragmatic approaches shall be taken to thrift in this difficult time. For example, it is sensible to encourage people to adopt new lifestyles, such as working at home.

Our models show that certain tweet variables and stock mispricing are relevant. Therefore, tweets do convey valuable information other than some meaningless noise as some investors think. These tweet variables we have explored can generally predict mispricing direction and magnitude in a timeframe. In general, our tweet variables have less influence on undervalued stocks, compared to overvalued stocks. However, the reason is unknown, which could be another concentration for future research. Our study suggests that there might be other channels related to mispricing, in addition to volatility and liquidity, and further relevant findings could be made. Although we have not discovered the prediction power of return from these tweet variables so far, there could be other tweet variables capable of predicting return related to a particular financial market or sector, and it is still worth delving into this field.

Despite the simplicity, our sentiment indices can predict three out of five industry categories and have implications. Although our models currently are not applicable to the other industry categories, it does not necessarily mean that tweet information cannot predict them. Our work relates tweet sentiment to industry returns, and possibly more could be achieved if new sentiment models are explored. Our research also can serve as a reference for practitioners of financial investment, such as financial analysts whose specialties include such industries. Furthermore, our dataset could be a cornerstone for future relevant studies.

REFERENCES

Aabo, T., Pantzalis, C., & Park, J. C. (2017). Idiosyncratic Volatility: An Indicator of Noise Trading? *Journal of Banking & Finance*, *75*, 136–151. https://doi.org/10.1016/j.jbankfin.2016.11.003

Aharon, D. Y., Demir, E., Lau, C. K. M., & Zaremba, A. (2022). Twitter-Based Uncertainty and Cryptocurrency Returns. *Research in International Business and Finance*, *59*, 101546. https://doi.org/10.1016/j.ribaf.2021.101546

Ahmadi, I., Habel, J., Jia, M., Lee, N., & Wei, S. (2022). Consumer Stockpiling Across Cultures During the COVID-19 Pandemic. *Journal of International Marketing*, *30*(2), 28–37. https://doi.org/10.1177/1069031X211037590

Al-Kadi, F. (2017). Entrepreneurship and Culture: The Role of National Culture in Entrepreneurship: A Study of 51 Countries: The Role of National Culture in Entrepreneurship: A Study of 51 Countries. Morrisville, North Carolina: Lulu Press. Retrieved on November 3, 2022. https://books.google.com/books?hl=en&lr=&id=3iauDgAAQBAJ&oi=fnd&pg=PA1&dq=Global+Entrepreneurship+Monitor+Al-Kadi&ots=tU6wmGSVUq&sig=CUQYm9rk5cLgnkNFoG36AqhJHVo#v=onepage&q=Global%20Entrepreneurship%20Monitor%20Al-Kadi&f=false

Altig, D., Baker, S., Barrero, J. M., Bloom, N., Bunn, P., Chen, S., Davis, S. J., Leather, J., Meyer, B., Mihaylov, E., Mizen, P., Parker, N., Renault, T., Smietanka, P., & Thwaites, G. (2020). Economic Uncertainty before and during the COVID-19 Pandemic. *Journal of Public Economics*, *191*, 104274. https://doi.org/10.1016/j.jpubeco.2020.104274

An, B., & Tang, S. Y. (2020). *Lessons from COVID-19 Responses in East Asia: Institutional Infrastructure and Enduring Policy Instruments* (SSRN Scholarly Paper No. 3602375). https://doi.org/10.2139/ssrn.3602375

Avramov, D., Chordia, T., Jostova, G., & Philipov, A. (2019). Bonds, Stocks, and Sources of Mispricing. *George Mason University School of Business Research Paper*, (18-5). https://doi.org/10.2139/ssrn.3063424

Baker, M., & Wurgler, J. (2006). Investor sentiment and the Cross-section of Stock Returns. *The Journal of Finance*, *61*(4), 1645-1680. Hoboken, New Jersey: Wiley-Blackwell. Retrieved on November 28, 2022. https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2006.00885.x

Baker, M., & Wurgler, J. (2007). Investor Sentiment in the Stock Market. *Journal of Economic Perspectives*, *21*(2), 129–152. https://doi.org/10.1257/jep.21.2.129

Baker, M., Wurgler, J., & Yuan, Y. (2012). Global, Local, and Contagious Investor Sentiment. *Journal of Financial Economics*, *104*(2), 272–287. https://doi.org/10.1016/j.jfineco.2011.11.002

Baniamin, H. M., Rahman, M., & Hasan, M. T. (2020). *The COVID-19 Pandemic: Why Are Some Countries More Successful Than Others?* (SSRN Scholarly Paper No. 3575251). https://doi.org/10.2139/ssrn.3575251

Barone-Adesi, G., Mancini, L., & Shefrin, H. (2017). *Estimating Sentiment, Risk Aversion, and Time Preference from Behavioral Pricing Kernel Theory* (SSRN Scholarly Paper No. 2060983). https://doi.org/10.2139/ssrn.2060983

Bartov, E., Faurel, L., & Mohanram, P. S. (2022). *The Role of Social Media in the Corporate Bond Market: Evidence from Twitter* (SSRN Scholarly Paper No. 4059704). https://doi.org/10.2139/ssrn.4059704

Bavel, J. J. V., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., Drury, J., Dube, O., Ellemers, N., Finkel, E. J., Fowler, J. H., Gelfand, M., Han, S., Haslam, S. A., Jetten, J., … Willer, R. (2020). Using Social and Behavioural Science to Support COVID-19 Pandemic Response. *Nature Human Behaviour*, *4*(5), Article 5. https://doi.org/10.1038/s41562-020-0884-z

Behrendt, S., & Schmidt, A. (2018). *The Twitter Myth Revisited: Intraday Investor Sentiment, Twitter Activity and Individual-level Stock Return Volatility—ScienceDirect*. https://doi.org/10.1016/j.jbankfin.2018.09.016

Beneish, M. D., & Nichols, C. (2005). *Earnings Quality and Future Returns: The Relation between Accruals and the Probability of Earnings Manipulation* (SSRN Scholarly Paper No. 725162). https://doi.org/10.2139/ssrn.725162

Blankespoor, E., Miller, G. S., & White, H. D. (2013). The Role of Dissemination in Market Liquidity: Evidence from Firms’ Use of Twitter^{TM}. *The Accounting Review*, *89*(1), 79–112. https://doi.org/10.2308/accr-50576

Bloomfield, R. J., & Michaely, R. (2002). *Risk or Mispricing? From the Mouths of Professionals* (SSRN Scholarly Paper No. 319240). https://doi.org/10.2139/ssrn.319240

Boon-Itt, S., & Skunkan, Y. (2020). Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. *JMIR Public Health and Surveillance*, *6*(4), e21978. https://doi.org/10.2196/21978

Bosma, N., Hill, S., Ionescu-Somers, A., Kelley, D., Levie, J., & Tarnawa, A. (2019). *GEM Global Report 2019/2020*. 232. Global Entrepreneurship Monitor. Retrieved on November 3, 2022. https://www.gemconsortium.org/file/open?fileId=50443

Brewer, P., & Venaik, S. (2014). The Ecological Fallacy in National Culture Research. *Organization Studies*, *35*(7), 1063–1086. https://doi.org/10.1177/0170840613517602

Brown, A., Rambaccussing, D., Reade, J. J., & Rossi, G. (2016, February). *Using Social Media to Identify Market Inefficiencies: Evidence from Twitter and Betfair* [Monograph]. Birkbeck, University of London. http://www.sportbusinesscentre.com/research/research-papers/

Burghardt, M., Czink, M., & Riordan, R. (2008). *Retail Investor Sentiment and the Stock Market* (SSRN Scholarly Paper No. 1100038). https://doi.org/10.2139/ssrn.1100038

CDC. (2020, February 11). *Cases, Data, and Surveillance*. Centers for Disease Control and Prevention. Retrieved on November 3, 2022, from https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/assessing-risk-factors.html

Civilian Unemployment Rate. (n.d.). Washington, D.C.: Bureau of Labor Statistics. Retrieved on October 26, 2022, from https://www.bls.gov/charts/employment-situation/civilian-unemployment-rate.htm

Clayton, J., Ling, D. C., & Naranjo, A. (2009). Commercial Real Estate Valuation: Fundamentals Versus Investor Sentiment. *The Journal of Real Estate Finance and Economics*, *38*(1), 5–37. https://doi.org/10.1007/s11146-008-9130-6

Cliff, M. T., & Brown, G. W. (2001). *Investor Sentiment and the Near-Term Stock Market* (SSRN Scholarly Paper No. 282915). https://doi.org/10.2139/ssrn.282915

Climate Change: Earth Surface Temperature Data. (2017). San Francisco, California: Kaggle. Retrieved November 3, 2022, from https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

Cornell, B., Landsman, W. R., & Stubben, S. R. (2017). Accounting Information, Investor Sentiment, and Market Pricing. *Journal of Law, Finance, and Accounting*, *2*(2), 325–345. https://doi.org/10.1561/108.00000017

COVID-19 Map. (2022, October 20). Baltimore, Maryland: Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/map.html

Daniel, K. D., Hirshleifer, D. A., & Sun, L. (2019). *Short- and Long-Horizon Behavioral Factors* (SSRN Scholarly Paper No. 3086063). https://doi.org/10.2139/ssrn.3086063

Derouiche, K., & Frunza, M. (2020). *Impact of Tweets' Sentiment Upon Stock Prices of Sport Companies: Can Fans Inflationuence the Share Price of Their Preferred Sport Brand?* (SSRN Scholarly Paper No. 3655256). https://doi.org/10.2139/ssrn.3655256

Dheer, R., Egri, C., & Treviño, L. (2020). *COVID-19 A Cultural Analysis to Understand Variance in Infection Rate across Nation*. PsyArXiv Preprints. Retrieved November 3, 2022, from https://doi.org/10.31234/osf.io/cbxhw

Drake, M. S., Guest, N. M., & Twedt, B. J. (2014). The Media and Mispricing: The Role of the Business Press in the Pricing of Accounting Information. *The Accounting Review*, *89*(5), 1673–1701. https://doi.org/10.2308/accr-50757

Dyer, J., & Kolic, B. (2020). Public Risk Perception and Emotion on Twitter during the Covid-19 Pandemic. *Applied Network Science*, *5*(1), 99. https://doi.org/10.1007/s41109-020-00334-7

Eisenberg, J. N. S., Desai, M. A., Levy, K., Bates, S. J., Liang, S., Naumoff, K., & Scott, J. C. (2007). Environmental Determinants of Infectious Disease: A Framework for Tracking Causal Links and Guiding Public Health Research. *Environmental Health Perspectives*, *115*(8), 1216–1223. https://doi.org/10.1289/ehp.9806

Elliott, W. B., Krische, S. D., & Peecher, M. E. (2010). Expected Mispricing: The Joint Inflationuence of Accounting Transparency and Investor Base. *Journal of Accounting Research*, *48*(2), 343–381. https://doi.org/10.1111/j.1475-679X.2010.00370.x

Fama, E. F., & French, K. R. (2015). A Five-factor Asset Pricing Model. *Journal of Financial Economics*, *116*(1), 1–22. https://doi.org/10.1016/j.jfineco.2014.10.010

Fan, J. H., Binnewies, S., & De SILVA, S. (2022). *Wisdom of Crowds and Commodity Pricing* (SSRN Scholarly Paper No. 4104888). https://doi.org/10.2139/ssrn.4104888

Ganesh, A., & Iyer, S. (2021). Impact of Firm-Initiated Tweets on Stock Return and Trading Volume. *Journal of Behavioral Finance*, *0*(0), 1–12. https://doi.org/10.1080/15427560.2021.1949717

Gokmen, Y., Baskici, C., & Ercil, Y. (2021). The Impact of National Culture on the Increase of COVID-19: A Cross-country Analysis of European Countries. *International Journal of Intercultural Relations*, *81*, 1–8. https://doi.org/10.1016/j.ijintrel.2020.12.006

Gopinath, G. (2020, April 7). *The Great Lockdown: Worst Economic Downturn since the Great Depression*. IMF. Retrieved November 3, 2022, from https://www.imf.org/en/Blogs/Articles/2020/04/14/blog-weo-the-great-lockdown-worst-economic-downturn-since-the-great-depression

Guijarro, F., Moya-Clemente, I., & Saleemi, J. (2019). Liquidity Risk and Investors' Mood: Linking the Financial Market Liquidity to Sentiment Analysis through Twitter in the S&P500 Index. *Sustainability*, *11*(24), Article 24. https://doi.org/10.3390/su11247048

Gupta, A., & Katarya, R. (2020). Social Media based Surveillance Systems for Healthcare Using Machine Learning: A systematic review. *Journal of Biomedical Informatics*, *108*, 103500. https://doi.org/10.1016/j.jbi.2020.103500

Han, X., Gu, X., & Peng, S. (2019). Analysis of Tweet Form's Effect on Users' Engagement on Twitter. *Cogent Business & Management*, *6*(1), 1564168. https://doi.org/10.1080/23311975.2018.1564168

He, W., Guo, L., Shen, J., & Akula, V. (2016). Social Media-Based Forecasting: A Case Study of Tweets and Stock Prices in the Financial Services Industry. *Journal of Organizational and End User Computing (JOEUC)*, *28*(2), 74–91. https://doi.org/10.4018/JOEUC.2016040105

Hofstede, G. (2011). Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture*, *2*(1). https://doi.org/10.9707/2307-0919.1014

Hofstede, G. (2017). Country Comparison. *Hofstede Insights*. Retrieved November 3, 2022, from https://www.hofstede-insights.com/country-comparison/

Hong, H., & Sraer, D. A. (2016). Speculative Betas. *The Journal of Finance*, *71*(5), 2095–2144. https://doi.org/10.1111/jofi.12431

Hou, K., Xue, C., & Zhang, L. (2017). *A Comparison of New Factor Models* (SSRN Scholarly Paper No. 2520929). https://doi.org/10.2139/ssrn.2520929

ILO: COVID-19 causes devastating losses in working hours and employment. (2020, April 7). [Press release]. International Labour Organization. Retrieved November 3, 2022, from http://www.ilo.org/global/about-the-ilo/newsroom/news/WCMS_740893/lang--en/index.htm

Indra, L., & Husodo, Z. A. (2020, November). Twitter Sentiment on Mispricing in Indonesia Stock Market. *In The Fifth Padang International Conference on Economics Education, Economics, Business and Management, Accounting and Entrepreneurship (PICEEBA-5 2020)* (pp. 501-509). Paris, France: Atlantis Press. Retrieved November 3, 2022, from https://doi.org/10.2991/aebmr.k.201126.056

Jain, T. (2020, April 10). Hofstede’s legacy and separate national responses to the Covid-19 crisis. *LSE Business Review*. Retrieved November 3, 2022, from https://blogs.lse.ac.uk/businessreview/2020/04/10/hofstedes-legacy-and-separate-national-responses-to-the-covid-19-crisis/

Ji, X., Chun, S. A., & Geller, J. (2013). Monitoring Public Health Concerns Using Twitter Sentiment Classifications. *2013 IEEE International Conference on Healthcare Informatics*, 335–344. https://doi.org/10.1109/ICHI.2013.47

Jung, M. J., Naughton, J. P., Tahoun, A., & Wang, C. (2017). Do Firms Strategically Disseminate? Evidence from Corporate Use of Social Media. *The Accounting Review*, *93*(4), 225–252. https://doi.org/10.2308/accr-51906

Karampatsas, N., Malekpour, S., Mason, A., & Mavis, C. P. (2022). Twitter Investor Sentiment and Corporate Earnings Announcements. *European Financial Management*, *n/a*(n/a). https://doi.org/10.1111/eufm.12384

Kubota, K., Suda, K., & Takehara, H. (2009). Common Risk Factors Versus a Mispricing Factor of Tokyo Stock Exchange Firms: Inquiries into the Fundamental Value Derived from Analyst Earnings Forecasts*. *International Review of Finance*, *9*(3), 269–294. https://doi.org/10.1111/j.1468-2443.2009.01091.x

Kumar, R. (2021). Impact of Societal Culture on Covid-19 Morbidity and Mortality across Countries. *Journal of Cross-Cultural Psychology*, *52*(7), 643–662. https://doi.org/10.1177/00220221211025100

Lajunen, T., Gaygısız, E., & Gaygısız, Ü. (2022). Socio-cultural Correlates of the COVID-19 Outcomes. *Journal of Epidemiology and Global Health*, *12*(3), 328–339. https://doi.org/10.1007/s44197-022-00055-3

Lang, Q., Lu, X., Ma, F., & Huang, D. (2022). Oil futures volatility predictability: Evidence based on Twitter-based Uncertainty. *Finance Research Letters*, *47*, 102536. https://doi.org/10.1016/j.frl.2021.102536

La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2008). The economic consequences of legal origins. *Journal of economic literature*, *46*(2), 285-332. https://www.aeaweb.org/articles?id=10.1257/jel.46.2.285

Lietsala, K., & Sirkkunen, E. (2008). *Social Media: Introduction to the Tools and Processes of Participatory Economy*. Kanslerinrinne, Finland: Tampere University Press. Retrieved on November 28, 2022 https://trepo.tuni.fi/bitstream/handle/10024/65560/978-951-44-7320-3.pdf?seq

Liew, J. K.-S., & Budavári, T. (2016). *Do Tweet Sentiments Still Predict the Stock Market?* (SSRN Scholarly Paper No. 2820269). https://doi.org/10.2139/ssrn.2820269

Ling, D., Naranjo, A., & Scheick, B. (2010). *Executive Summary Investor Sentiment and Asset Pricing in Public and Private Markets*. *RERI WP*, *170*. https://www.researchgate.net/profile/David-Ling-2/publication/228316743_Executive_Summary_Investor_Sentiment_and_Asset_Pricing_in_Public_and_Private_Markets/links/54dc3c570cf2a7769d95682d/Executive-Summary-Investor-Sentiment-and-Asset-Pricing-in-Public-and-Private-Markets.pdf

List of Countries by Median Age. (2022). In *Wikipedia*. Retrieved on November 3, 2022, from https://en.wikipedia.org/w/index.php?title=List_of_countries_by_median_age&oldid=1104731832

Liu, S., & Han, J. (2020). Media Tone and Expected Stock Returns. *International Review of Financial Analysis*, *70*, 101522. https://doi.org/10.1016/j.irfa.2020.101522

Mackenbach, J. P. (2014). Cultural Values and Population Health: A Quantitative Analysis of Variations in Cultural Values, Health Behaviours and Health Outcomes among 42 European Countries. *Health & Place*, *28*, 116–132. https://doi.org/10.1016/j.healthplace.2014.04.004

Majumdar, A., & Bose, I. (2019). Do Tweets Create Value? A Multi-period Analysis of Twitter Use and Content of Tweets for Manufacturing Firms. *International Journal of Production Economics*, *216*, 1–11.

https://doi.org/10.1016/j.ijpe.2019.04.008

McGurk, Z., Nowak, A., & Hall, J. C. (2020). Stock Returns and Investor Sentiment: Textual Analysis and Social Media. *Journal of Economics and Finance*, *44*(3), 458–485. https://doi.org/10.1007/s12197-019-09494-4

McKibbin, W., & Fernando, R. (2021). The Global Macroeconomic Impacts of COVID-19: Seven Scenarios. *Asian Economic Papers*, *20*(2), 1–30. https://doi.org/10.1162/asep_a_00796

McSweeney, B. (2002). Hofstede's Model of National Cultural Differences and their Consequences: A Triumph of Faith - a Failure of Analysis. *Human Relations*, *55*(1), 89–118. https://doi.org/10.1177/0018726702551004

Medford, R. J., Saleh, S. N., Sumarsono, A., Perl, T. M., & Lehmann, C. U. (2020). An "Infodemic": Leveraging High-Volume Twitter Data to Understand Early Public Sentiment for the Coronavirus Disease 2019 Outbreak. *Open Forum Infectious Diseases*, *7*(7), ofaa258. https://doi.org/10.1093/ofid/ofaa258

Messner, W. (2020). *The Institutional and Cultural Context of Cross-National Variation in COVID-19 Outbreaks* (p. 2020.03.30.20047589). medRxiv. https://doi.org/10.1101/2020.03.30.20047589

Mueller, S. L., & Thomas, A. S. (2001). Culture and entrepreneurial potential: A Nine Country Study of Locus of Control and Innovativeness. *Journal of Business Venturing*, *16*(1), 51–75. https://doi.org/10.1016/S0883-9026(99)00039-7

Naseem, U., Razzak, I., Khushi, M., Eklund, P. W., & Kim, J. (2021). COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis. *IEEE Transactions on Computational Social Systems*, *8*(4), 1003–1015. https://doi.org/10.1109/TCSS.2021.3051189

Oey, E., & Rahardjo, B. S. (2021). Does Culture Inflationuence Our Ways in Handling COVID-19? *International Journal of Sociology and Social Policy*, *41*(11/12), 1149–1169. https://doi.org/10.1108/IJSSP-02-2021-0051

Palmer, C. L., & Peterson, R. D. (2020). Toxic Mask-ulinity: The Link between Masculine Toughness and Affective Reactions to Mask Wearing in the COVID-19 Era. *Politics & Gender*, *16*(4), 1044–1051. https://doi.org/10.1017/S1743923X20000422

Pershad, Y., Hangge, P. T., Albadawi, H., & Oklu, R. (2018). Social Medicine: Twitter in Healthcare. *Journal of Clinical Medicine*, *7*(6), Article 6. https://doi.org/10.3390/jcm7060121

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The Effects of Twitter Sentiment on Stock Price Returns. *PLOS ONE*, *10*(9), e0138441. https://doi.org/10.1371/journal.pone.0138441

Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021). A Performance Comparison of Supervised Machine Learning Models for Covid-19 Tweets Sentiment Analysis. *PLOS ONE*, *16*(2), e0245909. https://doi.org/10.1371/journal.pone.0245909

Sanford, A. (2022). Does Perception Matter in Asset Pricing? Modeling Volatility Jumps Using Twitter-Based Sentiment Indices. *Journal of Behavioral Finance*, *23*(3), 262–280. https://doi.org/10.1080/15427560.2020.1866573

Shapiro, A., Sudhof, M., & Wilson, D. (2022). *Measuring News Sentiment—ScienceDirect*. https://doi.org/10.1016/j.jeconom.2020.07.053

Shukri, S. E., Yaghi, R. I., Aljarah, I., & Alsawalqah, H. (2015). Twitter Sentiment Analysis: A Case Study in the Automotive Industry. *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, 1–5. https://doi.org/10.1109/AEECT.2015.7360594

Souza, T. T. P., Kolchyna, O., Treleaven, P. C., & Aste, T. (2015). *Twitter Sentiment Analysis Applied to Finance: A Case Study in the Retail Industry* (arXiv:1507.00784). arXiv. https://doi.org/10.48550/arXiv.1507.00784

Sprenger, T. O., & Welpe, I. M. (2011). Tweets and Peers: Defining Industry Groups and Strategic Peers based on Investor Perceptions of Stocks on Twitter. *Algorithmic Finance*, *1*(1), 57–76. https://doi.org/10.3233/AF-2011-006

Stambaugh, R. F., Yu, J., & Yuan, Y. (2015). Arbitrage Asymmetry and the Idiosyncratic Volatility Puzzle. *The Journal of Finance*, *70*(5), 1903–1948. https://doi.org/10.1111/jofi.12286

Steve P. Calandrillo, Vanishing Vaccinations: Why Are So Many Americans Opting Out of Vaccinating Their Children?, 37 U. Mich. J. L. Reform 353 (2004), retrieved on November 28, 2022 https://digitalcommons.law.uw.edu/faculty-articles/138

Thunström, L., Newbold, S. C., Finnoff, D., Ashworth, M., & Shogren, J. F. (2020). The Benefits and Costs of Using Social Distancing to Flatten the Curve for COVID-19. *Journal of Benefit-Cost Analysis*, *11*(2), 179–195. https://doi.org/10.1017/bca.2020.12

Travica, B. (2020). *Containment Strategies for COVID-19 Pandemic* (SSRN Scholarly Paper No. 3604519). https://doi.org/10.2139/ssrn.3604519

Wang, J., Yang, J., Iverson, B. C., & Kluender, R. (2020). *Bankruptcy and the COVID-19 Crisis* (SSRN Scholarly Paper No. 3690398). https://doi.org/10.2139/ssrn.3690398

Wang, Y. (2021). Government Policies, National Culture and Social Distancing during the First Wave of the COVID-19 Pandemic: International evidence. *Safety Science*, *135*, 105138. https://doi.org/10.1016/j.ssci.2020.105138

Windsor, L. C., Reinhardt, G. Y., Windsor, A. J., Ostergard, R., Allen, S., Burns, C., Giger, J., & Wood, R. (2020). Gender in the time of COVID-19: Evaluating national leadership and COVID-19 fatalities. *PLOS ONE*, *15*(12), e0244531. https://doi.org/10.1371/journal.pone.0244531

World Bank Group—International Development, Poverty, & Sustainability. (n.d.). World Bank. Retrieved October 27, 2022, from https://www.worldbank.org/en/home

World Economic Outlook. (2022). International Monetary Fund. Retrieved October 11, 2022, from https://www.imf.org/en/Publications/WEO

Xiong, X., Meng, Y., Joseph, N. L., & Shen, D. (2020). Stock Mispricing, Hard-to-value Stocks and the Influence of Internet Stock Message Boards. *International Review of Financial Analysis*, *72*, 101576. https://doi.org/10.1016/j.irfa.2020.101576

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., & Zhu, T. (2020). Public Discourse and Sentiment during the COVID-19 Pandemic: Using Latent Dirichlet Allocation for Topic Modeling on Twitter. *PLOS ONE*, *15*(9), e0239441. https://doi.org/10.1371/journal.pone.0239441

Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., & Zhu, T. (2020). Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. *Journal of Medical Internet Research*, *22*(11), e20550. https://doi.org/10.2196/20550

Yang, S., Mo, S. Y. K., & Zhu, X. (2014). *An Empirical Study of the Financial Community Network on Twitter | IEEE Conference Publication | IEEE Xplore*. https://ieeexplore.ieee.org/document/6924054

Yeasmin, N., Mahbub, N. I., Baowaly, M. K., Singh, B. C., Alom, Z., Aung, Z., & Azim, M. A. (2022). Analysis and Prediction of User Sentiment on COVID-19 Pandemic Using Tweets. *Big Data and Cognitive Computing*, *6*(2), Article 2. https://doi.org/10.3390/bdcc6020065

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear." *Procedia - Social and Behavioral Sciences*, *26*, 55–62. https://doi.org/10.1016/j.sbspro.2011.10.562

APPENDIX

COVID-19 MAIN DATASET

Table A COVID-19 MAIN DATASET

Table A shows the main variable statistics used in the regression analysis.

Country	Infection	Death	Vaccine	Age	GDP	Temperature	Migration	Population	Urban	SC	FR	DE	UAI	IDV	MAS	PDI	IVR	LTO
Luxembourg	47.01	0.38	2.11	39.50	5.13	9.90	7.62	245.00	92.00	0	1	0	70	60	50	40	56	64
Iceland	55.37	0.10	2.29	37.10	4.83	2.92	0.51	4.00	94.00	1	0	0	50	60	10	30	67	28
Spain	28.50	0.85	2.17	43.90	4.48	14.86	0.42	95.00	81.00	0	1	0	86	51	42	57	44	48
Ireland	33.22	0.48	2.30	37.80	5.00	9.97	2.35	72.00	64.00	0	0	0	35	70	68	28	65	24
Belgium	39.76	0.71	2.30	41.60	4.71	10.06	2.07	381.00	98.00	0	1	0	94	75	54	65	57	82
Singapore	37.97	0.08	2.70	35.60	4.86	27.60	2.48	7919.00	100.00	0	0	0	8	20	48	74	46	72
United States	29.30	1.10	1.84	38.50	4.84	11.30	1.44	36.00	83.00	0	0	0	46	91	62	40	68	26
Italy	39.55	0.76	2.40	46.50	4.55	14.17	1.26	201.00	71.00	0	1	0	75	76	70	50	30	61
Switzerland	48.31	0.33	1.86	42.70	4.97	8.18	2.99	219.00	74.00	0	0	1	58	68	70	34	66	74
United Kingdom	35.76	0.87	2.30	40.60	4.68	9.11	1.94	277.00	84.00	0	0	0	35	89	66	35	69	51
France	54.55	0.43	2.27	41.70	4.64	11.20	0.27	123.00	81.00	0	1	0	86	71	43	35	48	63
Portugal	53.57	0.46	2.54	44.60	4.38	15.89	-0.29	112.00	67.00	0	1	0	99	27	31	63	33	28
Sweden	25.04	0.79	2.37	41.10	4.78	4.20	1.92	25.00	88.00	1	0	0	29	71	5	31	78	53
Netherlands	49.08	0.27	2.06	42.80	4.76	9.45	0.46	518.00	93.00	0	1	0	53	80	14	38	68	67
Germany	42.45	0.43	2.23	47.80	4.71	9.24	3.27	238.00	78.00	0	0	1	65	67	66	35	40	83
Peru	12.45	5.22	2.55	29.10	3.83	19.98	1.48	26.00	79.00	0	1	0	87	16	42	64	46	25
Canada	11.37	1.07	2.40	41.80	4.72	-1.64	3.16	4.00	82.00	0	0	0	48	80	52	39	68	36
Denmark	57.65	0.22	2.25	42.00	4.83	8.52	1.30	146.00	88.00	1	0	0	23	74	16	18	70	35
Austria	60.32	0.39	2.15	44.50	4.73	8.06	3.63	108.00	59.00	0	0	1	70	55	79	11	63	60
Turkey	19.90	0.60	1.79	32.20	3.98	14.03	1.67	110.00	77.00	0	1	0	85	37	45	66	49	46
Russia	14.70	1.81	1.25	40.30	4.09	-2.26	0.64	9.00	75.00	0	1	0	95	39	36	93	20	81
Chile	24.56	1.30	3.25	35.50	4.22	9.88	2.91	26.00	88.00	0	1	0	86	23	28	63	68	31
Norway	27.07	0.29	2.21	39.50	4.95	1.74	2.59	15.00	83.00	1	0	0	50	69	8	31	55	35
Estonia	45.63	0.45	1.62	43.70	4.44	6.96	1.47	31.00	69.00	0	0	1	60	60	30	40	16	82
Iran	8.89	1.91	1.82	31.70	3.44	20.54	-0.32	52.00	76.00	0	1	0	59	41	43	58	40	14
Serbia	35.03	0.72	1.25	43.40	3.96	12.84	0.29	79.00	57.00	0	1	0	92	25	43	86	28	52
Malta	22.30	0.70	2.59	42.30	4.52	19.44	0.87	1610.00	57.00	0	1	0	96	59	47	56	66	47
Saudi Arabia	2.32	1.14	1.92	30.80	4.37	27.74	1.91	16.00	85.00	0	0	0	80	25	60	95	52	36
Finland	24.10	0.48	2.31	42.80	4.73	4.06	1.26	18.00	86.00	1	0	0	59	63	26	33	57	38
Dominican Rep	5.91	0.68	1.46	27.90	3.93	26.29	-1.37	225.00	83.00	0	1	0	45	30	65	65	54	13
Romania	17.18	2.05	0.88	42.50	4.17	11.65	-1.94	84.00	54.00	0	1	0	90	30	42	90	20	52
Brazil	16.25	1.98	2.27	33.20	3.88	25.35	0.05	25.00	87.00	0	1	0	76	38	49	69	59	44

Czechia	38.80	1.00	1.71	43.20	4.42	9.28	1.03	139.00	74.00	0	0	1	74	58	57	57	29	70
Slovenia	58.29	0.56	1.42	44.90	4.47	11.74	0.47	104.00	55.00	0	0	1	88	27	19	71	48	49
Lithuania	45.29	0.74	1.61	44.50	4.37	7.80	-5.86	45.00	68.00	0	1	0	65	60	19	42	16	82
Croatia	31.92	1.37	1.37	43.90	4.24	13.59	-1.03	72.00	58.00	0	0	1	80	33	40	73	33	58
Latvia	50.41	0.64	1.54	44.40	4.31	7.32	-3.94	31.00	68.00	0	0	1	63	70	9	44	13	69
Poland	16.77	1.86	1.52	41.90	4.25	9.26	-0.39	124.00	60.00	0	0	1	93	60	64	68	29	38
Ukraine	12.74	2.11	0.72	41.20	3.68	10.91	0.11	76.00	70.00	0	1	0	95	25	27	92	14	86
Hungary	21.96	2.24	1.70	43.60	4.27	12.30	0.31	107.00	72.00	0	0	1	82	80	88	46	31	58
New Zealand	35.87	0.11	2.30	37.20	4.69	11.09	1.45	19.00	87.00	0	0	1	49	79	58	22	75	33
Albania	11.84	1.08	1.06	34.30	3.81	14.99	-2.49	104.00	63.00	0	1	0	70	20	80	90	15	61
Bulgaria	18.49	2.97	0.66	43.70	4.07	13.41	-0.35	64.00	76.00	0	0	1	85	30	40	70	16	69
Australia	40.19	0.15	2.47	37.50	4.78	22.14	3.07	3.00	86.00	0	0	1	51	90	61	38	71	21
Mexico	5.46	4.65	1.71	29.30	4.00	22.22	-0.23	66.00	81.00	0	1	0	82	30	69	81	97	24
Slovakia	24.77	0.78	0.66	41.80	4.32	9.69	0.07	114.00	54.00	0	0	1	51	52	100	100	28	77
Greece	47.65	0.66	2.03	45.30	4.31	17.15	-0.75	83.00	80.00	0	1	0	100	35	57	60	50	45
Colombia	12.31	2.25	1.74	31.20	3.79	25.58	2.00	46.00	82.00	0	1	0	80	13	64	67	83	13
South Korea	49.00	0.11	2.53	43.20	4.54	13.76	0.11	531.00	81.00	0	0	1	85	18	39	60	29	100
Malaysia	14.90	0.75	2.20	29.20	4.06	26.97	0.76	99.00	78.00	0	0	0	36	26	50	100	57	41
Uruguay	28.40	0.76	2.51	35.50	4.23	16.75	-0.43	20.00	96.00	0	1	0	98	36	38	61	53	26
South Africa	6.71	2.54	0.63	28.00	3.84	17.33	1.21	49.00	68.00	0	0	0	49	65	63	49	63	34
Morocco	3.39	1.29	1.48	29.10	3.54	19.31	-0.69	83.00	64.00	0	1	0	68	46	53	70	25	14
El Salvador	3.10	2.10	1.71	27.70	3.64	25.91	-3.11	313.00	74.00	0	1	0	94	19	40	66	89	20
Pakistan	0.70	1.95	1.39	22.00	3.19	22.32	-0.52	287.00	37.00	0	0	0	70	14	50	55	0	50
Ghana	0.54	0.86	0.64	21.40	3.39	27.72	-0.16	137.00	58.00	0	0	0	65	15	40	80	72	4
Argentina	21.21	1.34	2.40	32.40	4.03	14.46	0.05	17.00	92.00	0	1	0	86	46	56	49	62	20
Japan	17.57	0.21	2.57	48.60	4.59	12.96	0.28	346.00	92.00	0	0	1	92	46	95	54	42	88
Philippines	3.60	1.60	1.50	24.10	3.55	27.42	-0.30	368.00	48.00	0	1	0	44	32	64	94	42	27
Egypt	0.05	44.56	0.95	24.10	3.59	24.51	-0.18	103.00	43.00	0	1	0	80	25	45	70	4	7
Bangladesh	1.22	1.45	1.91	27.90	3.40	25.97	-1.11	1265.00	39.00	0	0	0	60	20	55	80	20	47
Trinidad and Tobago	13.15	2.30	1.13	37.80	4.18	27.09	-0.28	273.00	53.00	0	0	0	55	16	58	47	80	13
Iraq	5.98	1.03	0.47	21.20	3.70	24.24	0.10	93.00	71.00	0	1	0	85	30	70	95	17	25
China	0.21	0.54	2.45	38.40	4.10	9.30	-0.12	150.00	63.00	0	1	0	30	20	66	80	24	87
Jordan	17.01	0.81	0.98	23.50	3.64	21.32	0.50	115.00	92.00	0	1	0	65	30	45	70	43	16
Indonesia	2.34	2.45	1.59	31.10	3.63	26.47	-0.18	146.00	57.00	0	1	0	48	14	46	78	38	62
India	3.20	1.18	1.58	28.70	3.36	25.41	-0.19	464.00	35.00	0	0	0	40	48	56	77	26	51
Thailand	6.70	0.70	2.05	39.00	3.86	27.86	0.14	137.00	52.00	0	0	0	64	20	34	64	45	32
Burkina Faso	0.10	1.79	0.19	17.90	2.96	29.39	-0.58	76.00	31.00	0	1	0	55	15	50	70	18	27
Nigeria	0.13	1.19	0.27	18.60	3.32	28.01	-0.14	226.00	53.00	0	0	0	55	30	60	80	84	13
Venezuela	1.90	1.07	1.32	30.00	4.21	25.91	-11.38	32.00	88.00	0	1	0	76	12	73	81	100	16
Zambia	1.76	1.20	0.61	16.90	3.05	21.20	-0.21	25.00	45.00	0	0	0	50	35	40	60	42	30
Tanzania	0.06	2.12	0.27	18.20	3.06	22.77	-0.33	67.00	36.00	0	0	0	50	25	40	70	38	34