Informational Content of CEO Postings and Stock Market Predictability

Kang-Pyo Lee* Suyong Song†

Abstract

We show that CEO postings contain informational content on the U.S. stock markets. We create a unique sample of CEO users on Twitter, extract hashtags and sentiments that can be used as features for prediction from large, unstructured tweet text, and construct time series data. To prove their stock market predictability using machine learning, we predict three numeric stock market indicators as a regression problem and the direction of stock prices as a classification problem. Findings confirm that the select list of hashtags and sentiments have predictive power on the stock return, trading volume, volatility, and stock price direction.

JEL classification: C45, C53, C55, G12, G17

Keywords: text analysis, machine learning, CEO tweets, stock market predictability

*Department of Analytics, Fairfield University, klee1@fairfield.edu

†Department of Economics and Department of Finance, University of Iowa, suyong-song@uiowa.edu

1. Introduction

Elon Musk, the founder and CEO of Tesla Inc. and SpaceX as well as the new owner of Twitter, or now X, is undoubtedly the most followed CEO user on Twitter with more than 100 million followers as of August 2022. He is known to be very popular, influential, and occasionally controversial on Twitter, and what he mentions online draws substantial attention from the public and investors. He posts numerous tweets on Twitter covering a variety of subjects, most of which are related to the companies he is currently involved in. Through what he posts on Twitter, his followers can not only get to know better about him but also obtain unique information about his businesses, which otherwise could not be accessible through the other channels.

Musk is not the only CEO actively using social media. Many of the incumbent or former star CEOs are taking advantage of social media to communicate with people in the world who are interested in the CEOs themselves or the companies they represent. As most CEOs are using their own personal accounts, their tweets can cover a wide range of topics, whether they are business related or unrelated, including their personal lives, interests, and opinions. Malhotra and Malhotra (2016) classify CEOs based on their use of Twitter into four different groups: Generalist CEOs (sharing a wide range of content), Expressionist CEOs (not trying to share business-related content but sharing their opinions and giving their followers insights), Information Maven CEOs (sharing links to information, news, and other happenings), and Business Maven CEOs (sharing business-related content).

The main idea of this study is that what CEOs post on Twitter can be a powerful signal of the economy, especially the stock market movement, based on our beliefs that CEOs are the ones who would be more attentive to business factors and environments than anyone else and that they are believed to be insightful leaders in their industries. While each CEO may have a different opinion on an economic situation, their collective voice can predict how the stock markets move over time as a whole. Their voice can be expressed in their tweets in different forms such as thoughts, opinions, questions, concerns, criticisms, etc. The most

recent example of this is their reactions to the ongoing COVID-19 pandemic. They have expressed deep concerns about the worldwide pandemic in their tweets, focusing on how the pandemic could impact their businesses and also the entire economy. They have also shared their prospects for the future at each turning point and breakthrough during the pandemic, whether it is positive or negative.

To be specific, we show that select keywords from CEO tweets have predictive power of stock market indicators such as stock return, trading volume, volatility, and stock price direction. To that end, we identify 4,714 CEO users in Twitter from our Big Data pool of unstructured text (approximately 4.7 billion tweets) and collect their tweets posted from June 2009 to December 2021 (approximately 6.3 million CEO tweets). Based on the idea that the hashtags and sentiment words used in CEO tweets serve as keywords that can be used as predictors in our models, we construct hashtag and sentiment time series data from CEO tweets using tweet count and users count, and utilize them as features to predict stock market indicators from four major stock indexes: S&P 500, Dow 30, Nasdaq, and Russell 2000. We find that hashtags and sentiment words do have predictive power of the stock market indicators in all stock indexes. We also find that aggregated sentiment features built on CEOs' tweet text provide informational content on the stock market indicators beyond what macroeconomic and financial variables considered in the literature do (e.g., Fama and French, 1988, 1989; Cochrane, 1991, 2007; Pesaran and Timmermann, 1995, 2000; Welch and Goyal, 2008; Van Binsbergen and Koijen, 2010; Golez and Koudijs, 2018).

We utilize the 1,000 most popular hashtags and seven predefined sentiment categories (Negative, Positive, Uncertainty, Litigious, Strong Modal, Weak Modal, and Constraining) as features for regression and classification, respectively, in our daily and weekly predictive models. Each of the three numeric stock market indicators – close, volume, and volatility – is used as a dependent variable to predict in regression, whereas the stock price direction is used as a target to predict in classification. For the regression task, as the number of hashtag regressors in our model is much larger than the number of time series observations,

the ordinary least squares (OLS) estimator is inconsistent. As a means of relevant variable selection, therefore, we employ one of the commonly-used machine learning methods for feature selection, the least absolute shrinkage and selection operator (lasso) approach introduced by Tibshirani (1996), and its variant adaptive group lasso (Yuan and Lin, 2006; Zou, 2006), which considers the hashtags from the same group. For the adaptive group lasso that considers the group information of regressors, we develop our own hashtag clustering method. In general, both the lasso and the adaptive group lasso based on hashtag clustering have predictive power of stock market indicators. We believe that the results are promising considering the widely-known difficulty of stock market prediction.

To evaluate the out-of-sample performances of our predictive models, we build models on the training set and predict the target variable with the models learned on the test set. We employ the rolling windows scheme for both regression and classification. We compare the sample loss function by calculating the mean squared errors (MSE) for regression and the prediction accuracy for classification, respectively, over different sizes of models (0.1, 0.2, ..., 0.9) and different horizons (from 1 to 5). The results present high predictive power of the models and interesting heterogeneous patterns in forecasting stock market indicators across different stock indexes.

For the classification task, we build models on the training set by applying six classification algorithms such as k-Nearest Neighbors, Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and Deep Neural Network. We then predict the direction of stock prices in the test set and compare prediction performances using classification accuracy. We find surprisingly high accuracy of the classification performances, which confirms that a variety of hashtags are informative of the direction of stock prices.

A large number of studies in finance literature have shown that a variety of macroeconomic and financial variables can predict stock returns (e.g., Welch and Goyal, 2008). We check whether the information from the CEO sentiment features is subsumed by macroeconomic and financial variables by controlling for each macroeconomic or financial variable

in the regression of stock indicators on CEO sentiment features using monthly data. The results show that the CEO sentiments still provide predictive power of the stock indicators beyond what the existing macroeconomic and financial variables do. Therefore, we argue that there is evidence of stock market predictability.

Recent studies have investigated the importance of CEOs' characteristics and roles in corporate decisions and outcomes. It is found that the managerial styles, influences, and personality traits of top executives can affect a wide range of corporate decisions, policies, and outcomes (Bertrand and Schoar, 2003; Adams, Almeida, and Ferreira, 2005; Malmendier and Tate, 2005; Kaplan, Klebanov, and Sorensen, 2012; Gow, Kaplan, Larcker, and Zakolyukina, 2016). More recent studies focus on the active use of social media not only by firms but also by CEOs. For example, Chen, Hwang, and Liu (2018) report the emergence of "social executives" directly communicating with investors through social media. Elliott, Grant, and Hodge (2018) discuss the benefits of direct communication of managers on social media. Men and Tsai (2016) take a public relation standpoint on why the public wants to engage with CEOs on social media and why that matters.

Another strand of the literature which is closely related to our paper is the studies that employed textual analysis to examine whether text messages on social media, stock message boards, analyst reports, firm financial statement, and conference calls have informational content on stock market. The literature shows that stock messages are predictive of market volatility and trading volume (Antweiler and Frank, 2004) and stock index levels (Das and Chen, 2007) and also that negative words in media reports serve as a proxy for investor sentiment (Tetlock, 2007; Chen, De, Hu, and Hwang, 2014) and convey negative information about firm earnings (Tetlock, Saar-Tsechansky, and Macskassy, 2008). It is also shown that investor sentiment predicts returns in the cross-section (Baker and Wurgler, 2006; Huang, Jiang, Tu, and Zhou, 2015) and that manager sentiment predicts future aggregate stock market returns and cross-section of stock returns (Jiang, Lee, Martin, and Zhou, 2019). Most recently, Wolfskeil (2023) explores the link between firms' voluntary disclosure strategies on

Twitter and their equity returns.

This paper makes several contributions to the extant literature. First, through our data-driven automated approach to CEO discovery followed by a thorough manual investigation, we create a large, unique sample of CEOs by identifying 4,714 actual CEO users in Twitter who have never been targeted at in the previous research.[1] Second, to translate large, unstructured text data from social media into structured machine-readable data, we apply text analysis techniques such as natural language processing (NLP), regular expressions, word clustering, and sentiment analysis. Third, we apply state-of-the-art machine learning techniques to high-dimensional social media data and show that they are useful in predicting stock market indicators. Fourth, we conduct comprehensive time series prediction analyses with the hashtags and sentiment features for various stock market indicators (return, trading volume, volatility, and stock price direction), stock indexes (S&P 500, Dow 30, Nasdaq, and Russell 2000), and time frequencies (daily, weekly, and monthly) optimizing the prediction performance for different model sizes and horizons. We also find that the predictive power of CEO sentiments still stands after controlling for well-known, macroeconomic and financial variables. The work done by Wolfskeil (2023) is very similar to ours in that the three aspects, timing, tone, and content, that she analyzes can translate to the horizon, sentiments, and hashtags in our approach, while there is a significant distinction between the two approaches in terms of textual data used: tweets posted by the firms' official Twitter accounts versus by the CEOs.

The rest of this paper proceeds as follows. Section 2 outlines the background of this study. Section 3 provides the details on the data we collect and analyze. Section 4 presents the findings from the stock market prediction. Section 5 contains conclusions of this study and discussion of possible extensions in the future. Lastly, the Supplementary Online Appendix contains detailed or additional information.

[1]We make all data used in this paper publicly available at the following website: https://ceo-attention.herokuapp.com/. This website also provides a dashboard for CEO hashtags and sentiments that are updated on a daily basis by automated NLP.

2. Background and Related Literature

2.1. CEOs and Social Media

With the growing popularity of some best-performing CEOs, extensive work has been done primarily in the economics and finance literature over the last two decades to investigate the importance of CEOs' characteristics and roles in critical corporate decisions and outcomes. Bertrand and Schoar (2003) argue that manager fixed effects or managerial styles affect a wide range of corporate decisions. They utilize the Forbes 800 files and ExecuComp data to identify approximately 600 firms and 500 managers including CEOs, CFOs, and COOs. Adams et al. (2005) focus on the firms run by powerful CEOs, based on the hypothesis that top executives can impact firm outcomes only if they have influence over critical corporate decisions. They identify 336 firms from the Fortune 500 and their CEOs from the ExecuComp data and find that firm performance is more variable for the firms run by powerful CEOs. Malmendier and Tate (2005) focus on the negative effects of overconfident managers, arguing that managerial overconfidence can lead to distortions in corporate investment policies. They analyze a sample of 477 large publicly traded U.S. firms and identify their CEOs from the COMPUSTAT database. Kaplan et al. (2012) pay attention to CEOs' 30 specific individual characteristics in five general categories including Leadership, Personal, Intellectual, Motivational, and Interpersonal. They identify 316 CEO candidates considered for positions in 224 companies funded by private equity investors and find that subsequent performance is positively related to general ability and execution skills. The work done by Gow et al. (2016) is another study that focuses on CEO personality. They show that the Big Five personality traits (agreeableness, conscientiousness, extraversion, neuroticism, and openness to experience) of 119 CEOs identified from U.S. technology and public firms are associated with crucial financial choices and firm operating performance. Pan, Siegel, and Wang (2019) investigate the role of cultural heritage in shaping the attitudes of CEOs towards uncertainty. They take a unique approach to cultural origin inference by comparing

CEO's last name with the same last name of the passengers arriving in the port of New York between 1820 and 1957.

Research on CEOs relying on social media is relatively new compared to research on firms taking advantage of social media. It has been generally accepted that firms can benefit from social media in a number of respects, as presented by Blankespoor, Miller, and White (2014), Miller and Skinner (2015), Jung, Naughton, Tahoun, and Wang (2018), and Lee, Hosanagar, and Nair (2018). On the other hand, there have been conflicting views among top executives over the use of social media directly run by themselves, primarily due to the CEO users' indifference to its positive effects or misconceptions about its negative effects, i.e., the costs and risks of using social media. Porter, Anderson, and Nhotsavang (2015) report results from a qualitative content analysis of Fortune 500 CEOs' use of Twitter, which present that senior mangers using Twitter tend to engage in one-sided conversation even though Twitter is a two-way medium and that most of them are using more formal language than general Twitter users. They argue that this naturally results in low credibility and value of social media by senior managers. Capriotti and Ruesja (2018) report that the presence of Fortune Global 500 CEOs in five major social media platforms including Facebook, Instagram, Linkedin, Twitter, and YouTube is as low as 26.0% and that most of them are not adequately using Twitter as 25% of the accounts are inactive and 90% of the active accounts have a low activity.

Recent studies consistently emphasize the importance of social media use by executives. Chen et al. (2018) analyze a sample of 155 S&P 1500 CEOs and CFOs and argue that the emergence of "social executives" who directly communicate with investors through social media and grant investors access to value-relevant corporate information, which can increase investor participation and improve stock market liquidity. They also find that the tone in their personal tweets is useful in predicting future earnings surprises. Elliott et al. (2018) discuss the benefits of managers' communicating via their personal Twitter account, arguing that they can utilize social media to mitigate investors' loss of trust and negative reaction to

negative earnings surprises. Gao (2019) analyze a sample of 226 CEOs from the ExecuComp database and show that positive words in CEO tweets can be used to predict positive future abnormal returns. From a public relations standpoint, Men and Tsai (2016) discusses why the public wants to engage with corporate CEOs on social media and why such engagement matters in terms of organizational reputation. Chatterji and Toffel (2016), presenting the example of Apple CEO Tim Cook, provide evidence that CEO activism on social media can influence public opinion and consumer attitudes, even if the issues are unrelated to their core business.

2.2. Informational content

There have been a number of studies employing textual analysis to examine whether messages on social media, stock message boards, analyst reports, firm financial statements, and conference calls have informational content on stock market. Bollen, Mao, and Zeng (2011) is known as one of the early papers that present promising results. They investigate whether large-scale, collective mood states are correlated to Dow Jones Industiral Average (DJIA) over time. Xing, Cambria, and Welsch (2018) review the recent efforts on natural language-based financial forecasting. The work done by Gjerstad, Meyn, Molnár, and Næss (2021) is a recent work that utilizes tweets in financial market prediction. In order to study how financial markets respond to the live statements posted by the former U.S. president Donald Trump on Twitter, they take an interesting approach by relating the precise timestamp information of each tweet to high-frequency financial data. Antweiler and Frank (2004) examine the level of message activity on Internet stock message boards, and find that the bullishness of the stock messages are predictive of market volatility and disagreement among the messages is associated with more trading volume. Baker and Wurgler (2006) propose a novel investor sentiment index that aggregates the information from six proxies using principal components analysis, and find that high investor sentiment predicts strongly low returns in the cross-section. Das and Chen (2007) develop algorithms to extract sentiment from

stock message boards, and find that the net of positive and negative opinion is related to stock index levels, volumes, and volatility. Tetlock (2007) study the relationship between the content of media reports and daily stock market activity using principal components analysis, and find that the fraction of negative words in media content serve as a proxy for investor sentiment. Tetlock et al. (2008) extend Tetlock (2007)'s analysis to examine the impact of negative words on individual firm's performances, and find that negative words convey negative information about firm earnings and aspects of firm's fundamentals are embedded in linguistic media content. Similarly, Chen et al. (2014) analyze the negative words in the articles and comments posted on Seeking Alpha[2] to find the fraction of those negative words negatively predict stock returns. To study the information content of analyst report text, Huang, Zang, and Zheng (2014) use a naive Bayes machine learning approach to extract textual opinions from analyst reports, and show that these provide useful information to investors. Huang et al. (2015) use the partial least squares method to exploit the information of Baker and Wurgler (2006)'s six sentiment proxies in a more efficient manner, and find that the sentiment index can predict the aggregate stock market well. Wolfskeil (2023) investigates the link between firms' voluntary disclosure strategies on social media in terms of timing, tone, and content and their equity returns. This work is close to ours in that the timing aspect is reflected as the horizon in our approach, the tone aspect as sentiments, and the content aspect as hashtags. Her work is different from ours, however, as it is interested in tweets posted by firms' official Twitter accounts rather than those by CEOs.

A growing body of research in economics and finance underscores the importance of developing word classification categories which gauge tone in economic and financial applications. Loughran and McDonald (2010) provide evidence that a commonly-used Harvard-IV-4 TagNeg list substantially misclassifies words when gauging tone in financial applications. They expand the word classification categories to six word lists (negative, positive, uncertainty, litigious, strong modal, and weak modal). They find that their word lists have significant

[2]See https://seekingalpha.com/.

relations with file date returns, trading volume, subsequent return volatility, standardized unexpected earnings, and two separate samples of fraud and material weakness. Bodnaruk, Loughran, and McDonald (2015) create a constraining word list to measure the level of financial constraints of publicly-traded companies, in addition to Loughran and McDonald (2010)'s six-word lists. They provide the word list by examining tens of thousands of words that appear in at least 5% of all annual reports and selecting only words that would be most likely considered constraining by other researchers. Jiang et al. (2019) construct a manager sentiment index based on the aggregated textual tone in firm financial statements and conference calls. They measure textual tone as the difference between the number of positive and negative words in the disclosure scaled by the total word count of the disclosure. They find that the proposed manager sentiment significantly and negatively predicts future aggregate stock market returns and cross-section of stock returns. Ke, Kelly, and Xiu (2019) propose a supervised learning framework which measures a sentiment score that is adapted to the return prediction and study the extent to which business news predicts observed variation of asset price. We also refer to Gentzkow, Kelly, and Taddy (2019) and Loughran and McDonald (2020) for excellent reviews on textual analysis in economics and finance.

3. Data

We begin this section by presenting formal modeling of the data space that we analyze throughout this study. Our Twitter data space is noted as $\mathcal{U} \times \mathcal{P} \times \mathcal{W}$, where $\mathcal{U}$ is a set of CEO users on Twitter, $\mathcal{P}$ is a set of tweets, or postings, created by CEO users, and $\mathcal{W}$ is a set of words used in CEO tweets. Accordingly, for a certain time $t \in \mathcal{T}$ where $\mathcal{T}$ is a set of timestamps such as days, weeks, and months, the snapshot of the data space can be noted as $\mathcal{U}_t \times \mathcal{P}_t \times \mathcal{W}_t$, where $\mathcal{U}_t \subset \mathcal{U}$ is a set of CEO users in Twitter existing at time t, $\mathcal{P}_t \subset \mathcal{P}$ is a set of tweets created at time t, and $\mathcal{W}_t \subset \mathcal{W}$ is a set of words used in the tweets at time t. This implies that, at time t, a CEO user $u_{i,t} \in \mathcal{U}_t$ creates a tweet $p_{j,t} \in \mathcal{P}_t$

using a set of words $\mathcal{W}_{i,j,t} \subset \mathcal{W}_t$, where user $i \in \{1, ..., |\mathcal{U}|\}$ and tweet $j \in \{1, ..., |\mathcal{P}|\}$. Some of the words in $\mathcal{W}$ can take the form of hashtags by using a hash symbol (#) as a prefix, which are denoted as $\mathcal{H} \subset \mathcal{W}$, where $\mathcal{H}$ is a set of hashtags. Replacing $\mathcal{W}$ and $\mathcal{W}_t$ with $\mathcal{H}$ and $\mathcal{H}_t$, respectively, the aforementioned data space can be transformed to $\mathcal{U} \times \mathcal{P} \times \mathcal{H}$ and $\mathcal{U}_t \times \mathcal{P}_t \times \mathcal{H}_t$ at time t, respectively, when we are only interested in hashtags instead of all words.

3.1. Discovering CEO Users in Twitter

As the first step to identify CEO users $\mathcal{U}$ on Twitter, we adopt the approach proposed by Lee and Song (2022). We begin by collecting a large pool of random users. Twitter opens part of its user-created data to the public via Application Programming Interface (API), called Twitter API.[3] As the Twitter API does not allow users to search for Twitter users by querying over their bio text, we instead use Twitter Streaming API[4] to collect random tweets written in English, each of which contains the author information. In other words, we collect a set of random users from a set of random tweets. The Streaming API allow users to retrieve real-time tweets from Twitter and is known to provide at most 1% sample of all the tweets created on Twitter at a given time. The 1% sample at a given time may sound too small to be used in a study, but it could form a large volume of tweets when collected over a long period of time. The Streaming API allows users to filter real-time tweets on a set of keywords of interest to the user. As a means to collect random tweets written in English, we use a set of extremely general words such as 'a', 'and', and 'the', instead of actual keywords, which is a commonly-used trick to collect random or general tweets. Table 1 presents the monthly statistics of the data collected for 3 years from January 2019 to December 2021, which totals approximately 4.7 billion (specifically 4,704,883,067) unique English tweets created by approximately 177 million (specifically 176,850,729) unique users.

In order to discover CEO users from the pool of random users, we examine the bio text of

[3]See https://developer.twitter.com/en/docs/twitter-api.

[4]See https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/overview.

each user, which can be found in the Description field of each User object in the tweet data. This is based on our observation that many CEO users on Twitter tend to indicate their role as CEO in their bio. For example, Apple CEO Tim Cook (@tim_cook) describes himself in his Twitter bio as "Apple CEO" and Google CEO Sundar Pichai (@sundarpichai) describes himself as "CEO, Google and Alphabet". Some CEO users link the Twitter account of the company they work for to their bio. For example, Arianna Huffington (@ariannahuff) describes herself as "@HuffPost Founder. Founder & CEO of @Thrive Global", which gives us an idea of how to identify the account of their company as additional information to extract.

As there are many different ways to indicate their CEO role and those expressions can exist with some irrelevant text, we employ a text pattern matching technique called regular expressions that are designed to catch only the CEO-indicating expressions in the bio text. In addition to leveraging regular expressions to find CEO users, we also check if the found account is verified by Twitter. A verified account indicates that Twitter has verified that the account of public interest is authentic.[5] Verified accounts have a blue verified badge on Twitter. We utilize this unique feature of Twitter as a means to guarantee the user's popularity or authority.

From the 177 million users in our pool, 5,226 users are identified by the aforementioned process. We filter out 512 accounts that are suspended or closed for some reason using the Twitter API[6] or that prove to be not CEO accounts after thorough manual investigation, which results in the set $\mathcal{U}$ of 4,714 users as the final sample of CEO users, i.e., $|\mathcal{U}| = 4714$. Many of the false positives are attributed to ambiguous descriptions of CEO roles, e.g., "Follow our CEO: @TonyPorterACTM" or untrue statements, e.g., "CEO of coffee." Table 2 presents the top-20 users identified, sorted by the number of followers they have. Elon Musk (@elonmusk) tops the ranking, followed by Paris Hilton (@parishilton) and Dwayne

[5]See https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts.

[6]See https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/api-reference/get-users-lookup.

Johnson (@therock). The late Kobe Bryant (@kobebryant), who was once an NBA super star and later the CEO of a production company called Granity Studios, ranks number four, as his Twitter account is still active as of August 2022, although there have been no new tweets posted since his death.

What is unique about this sample of CEO users is that it covers a wide range of CEOs from famous star CEOs of large firms to lesser-known CEOs of smaller businesses. Table 3 presents the bottom-20 users in the ranking who have only several hundreds or even dozens of followers but nevertheless are Twitter verified users. Previous literature on CEO heavily relies on proprietary data sets such as Standard and Poor's ExecuComp database that focuses on executives of large firms. For example, Chen et al. (2018) and Gao (2019) identify 155 and 226 Twitter accounts, respectively, of CEOs and/or CFOs from ExecuComp. They start with a list of executives in the data set to search Twitter for any matching accounts run by the executives. In contrast, we employee a data-driven, automated approach by examining a large amount of historical tweet data to discover CEO users. Our data set covers a unique and wide range of 4,714 CEOs who have never been identified and analyzed in the previous work but are expected to represent the collective voice of different levels of CEOs. We believe that we have significantly overcame the limitation from the low rate of CEO presence on social media reported in Capriotti and Ruesja (2018), which allows us to minimize the sampling bias.

One issue to note regarding finding CEOs from their bios is that Twitter users can update their bios at any time they would want. For that reason, it is common that our CEO users indicate their CEO role in their bio at some point, but they do not at other points for their own reasons. In other words, there is no guarantee that the users who we previously identified as CEO are currently still holding their CEO position. For example, John Legere (@johnlegere) was identified as CEO in our data set, but he stepped down as T-Mobile CEO in April 2020 and updated his bio accordingly, so he is no longer a CEO. To address this issue, we assume that, once a user is identified as CEO in our data set, the user is considered

to be a CEO whether or not their current bio actually indicates their CEO role, unlike the approach adopted by Gao (2019) who exclude the tweets posted outside CEO tenure. This assumption arises from the fact that our data set does not allows us to track their term as CEO. We believe, however, that our assumption is acceptable as the users are still important and influential on social media, even if they currently no longer act as official CEO.

Another issue is that some of the CEO users do not use their Twitter accounts solely for their companies. In other words, they also use Twitter for their own good, not necessarily for the company they represent. Chen et al. (2018) reports that top executives use their own personal Twitter accounts not only for breaking company news and describing their work related day-to-day activities but also for sharing their unrelated-to-work personal interests. We, again, do not exclude the users from our data set simply because they do not talk about their companies and, for the same reason, do not exclude any of their tweets. We assume that everything they mention on Twitter matters in one way or another, even if what they say on Twitter appears irrelevant, personal, or trivial. Unlike our approach, Gao (2019) exclude tweets about firm operations because their study focuses on the information content of CEOs' personal activities.

The last issue to note is that we do not consider the locations of CEO users. We only consider 1) whether their bio text has any expression indicating CEO and 2) whether the account in question is verified by Twitter. This means that CEOs located outside the United States can be included in our data. We do not attempt to identify and exclude those non-U.S. CEO users because, according to a web report from a market and consumer data firm called Statista,[7] U.S. Twitter users represent approximately 75% of all Twitter users in English speaking countries such as United Kingdom and Canada, so we believe that our data set mostly captures the voice of U.S.-based CEOs. We also believe that even the voice from non-U.S.-based CEOs can be helpful in our stock market prediction, given the interconnected world economy.

[7]See https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.

3.2. Collecting and Processing CEO Tweets

With the 4,714 CEO users found, we collect the set $\mathcal{P}$ of 62,754,212 tweets posted by the users, i.e., $|\mathcal{P}| = 62754212$. The Twitter API[8] allows users to retrieve up to 3,200 most recent tweets of a user in a structured manner, as long as the account is set to public. As CEO users typically post much more than 3,200 tweets, however, we use web scraping to retrieve up to 200,000 tweets from each account. The mean and the standard deviation of tweet count per user are 13,272 and 19,582, respectively. Figure 1 presents the histogram of tweet count per user. The histogram is skewed right, and tweet count of up to 15K account for approximately 75% of the total count.

Figure 2 demonstrates the change of CEO tweet count over time, represented with a solid line on the left primary Y-axis, and the change of CEO user count over time, represented with a dotted line on the right secondary Y-axis. The earliest tweet found in our data was posted on June 29, 2006 by Kevin Systrom (@kevin), which accordingly becomes the start date of our data set, and the last date is December 31, 2021, which we determine for this study. As presented in the figure, both the tweet and user counts increase at a fast rate until 2012, after which the user count continues to grow gradually, while the tweet count plateaus until 2020 but increases exponentially in 2020 and 2021 during the COVID-19 pandemic.

3.2.1. Mining Hashtags

Now that we have collected necessary data, the next step is to mine keywords from tweet text that will be used as predictors in our predictive models. Here, we focus on hashtags, rather than on all words in the text, as hashtags are users' active expressions of interest that facilitate search and aggregation of messages related to the same topic (Laniado and Mika (2010)). In other words, a hashtag can serve as an indicator of something popular or wide-spread on social media. Although not all tweets have hashtags, hashtag sparsity can

[8]See https://developer.twitter.com/en/docs/twitter-api/v1/tweets/timelines/api-reference/get-statuses-user_timeline.

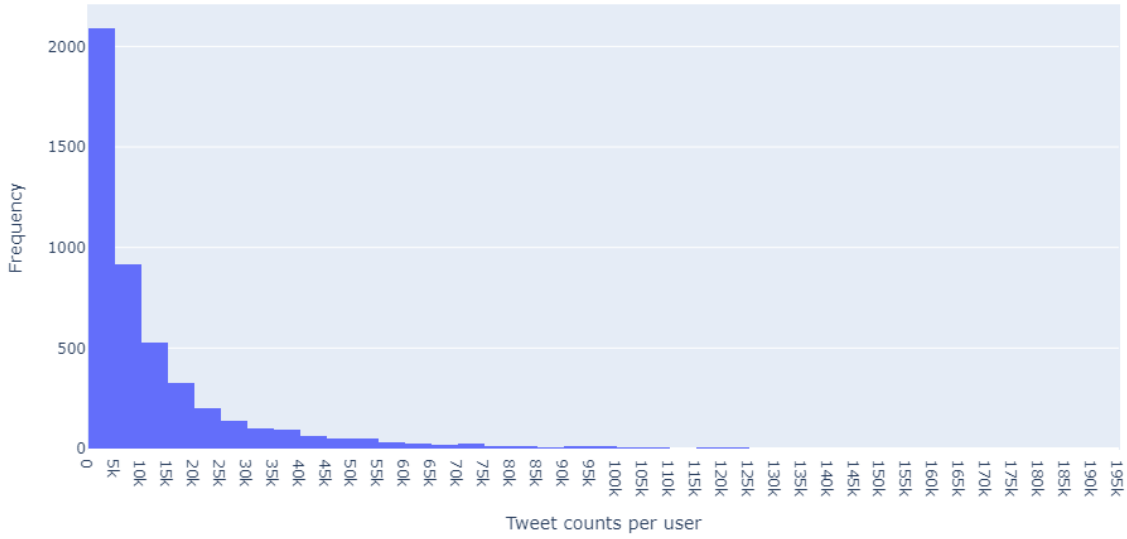

Figure 1. Histogram of tweet count per user

The histogram of tweet count per user is plotted.

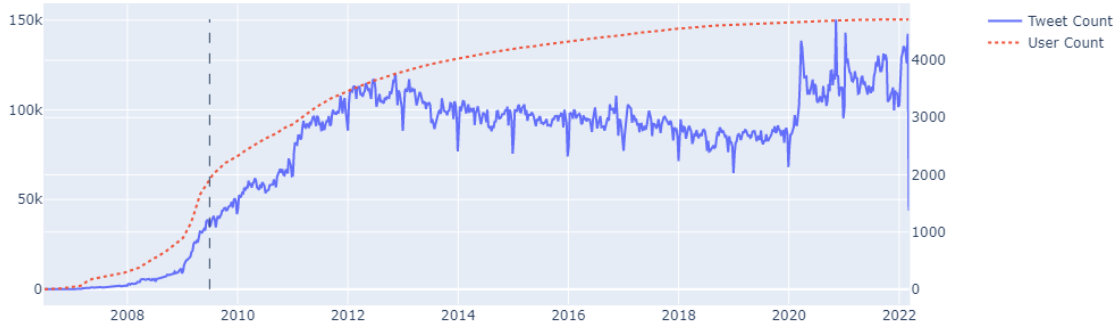

Figure 2. CEO tweet count and user count over time

The change of CEO tweet count over time (a solid line on the left, primary Y-axis) and the change of CEO user count over time (a dotted line on the right, secondary Y-axis) are plotted.

be overcome when a large number of tweets and their hashtags are aggregated.

We first identify all 1,861,649 English hashtags used in the CEO tweets and then calculate the frequency for each of all the hashtags. When calculating hashtag frequencies, there are two approaches. The first approach calculates how many 'tweets' contain the hashtag. Formally, for a hashtag $h \in \mathcal{H}$, the tweet count function $TC(h) = |\{(p_j, h)|j \in \{1, ..., |P|\}\}|$. The other approach calculates how many 'users' mention the hashtag. Formally, the user count function $UC(h) = |\{(u_i, h)|i \in \{1, ..., |U|\}\}|$. As there are 4,714 users in total in our data set, the user count cannot exceed 4714. The first tweet count metric focuses on the resource aspect of social media data, whereas the second user count metric emphasizes the social aspect. While we equally utilize both metrics, we present mostly with the user count metric throughout this paper, as it is novel and significantly reduces the numbers in our data. A performance comparison of the two metrics is presented later in Section 4.4.

As not all hashtags in $\mathcal{H}$ have frequencies higher enough to be used as regressors in our predictive models, we select only the 1,000 most popular hashtags, which leads to the reduced final set $\mathcal{H}$ of 1,000 hashtags, i.e., $|\mathcal{H}| = 1000$. In terms of user count, the number one hashtag #covid19 has frequency of 2,568, while the 1000th hashtag #ethics has frequency of 283, which we believe is still high enough to avoid the data sparsity problem. In terms of tweet count, the number one hashtag is #quote, which has frequency of 121,410, while the 1000th hashtag is #payments, which has frequency of 1989.

Table 4 lists the top-100 popular hashtags, out of 1,000 hashtags, sorted by user count (the complete list of all 1,000 most popular hashtags is available on the Supplementary Online Appendix). The #covid19 hashtag tops the ranking used by more than 2,500 CEO users, along with #coronavirus (number four) and #covid (number 19), which is a clear indication of CEOs' significant interest in the virus and the pandemic. The ranking includes many business-related hashtags such as #business, #startup (#startups), #jobs, #entrepreneur, #marketing, and #work. It also has the hashtags that represent technologies, tech services, or tech companies, such as #tech, #twitter, #facebook, #podcast, #socialmedia, #tech-

nology, #periscope, #apple, #ai, #google, #instagram, #iphone, #bitcoin, #digital, and #youtube.

The ranking also includes city names such as #nyc, #london, #newyork, and #paris, most of which indicate where they tweeted, and the names of countries or international organizations such as #usa, #china, #brexit, and #india, which indicate CEOs' interest in international affairs. The ranking also shows CEOs' interest in sports, represented by #superbowl, #worldcup, #olympics, and #nfl, and in politics represented by #vote and #trump. CEOs also show interest in social issues using such hashtags as #blacklivesmatter, #neverforget, #women, climatechange, #metoo, and #diversity. It is also interesting to see that the CEO users like to follow the trend on Twitter that people use specific hashtags on different days of the week or to celebrate something, e.g., #tbt (meaning Throwback Thursday), #ff (meaning Follow Friday), #internationalwomensday, #mondaymotivation, #christmas, #halloween, #thanksgiving, #happynewyear, #mothersday, #throwbackthursday, #valentinesday, #blackfriday, #followfriday, #fathersday, #merrychristmas, #givingtuesday, and #earthday.

Table 5 lists the top-100 popular hashtags sorted by tweet count, as opposed to user count (the complete list of all 1,000 most popular hashtags is available on the Supplementary Online Appendix). This tweet count-based ranking provides slightly different perspectives from the previous user count-based ranking, mainly because there can be biases when some users create a large number of tweets using a specific set of hashtags. For example, the number one hashtag #quote, which has frequency of more than 121,000, does not appear in the previous ranking. It turns out that approximately 34% of the tweet count is attributed to a single user named Tim Fargo (@tim_fargo).

3.2.2. Mining Sentiments

In addition to the hashtags used by the CEO users, we believe that the sentiments expressed in their tweets can also be useful for our stock prediction for the following reasons.

First, we consider all sentiment words in tweets, so that the word coverage of sentiments is more comprehensive than that of hashtags. Second, we use discipline-specific word lists focusing on financial contexts to construct sentiment time series. It is, therefore, expected to predict financial events well. Third, the analysis based on sentiments improves the interpretability of the regressors, as only a small number of sentiment features are included as predictors. In order to identify sentiments in CEO tweets, we employ the Sentiment Word Lists defined in Loughran and McDonald (2010) and Bodnaruk et al. (2015) (available from Bill McDonald's website). The sentiment lexicon covers seven sentiment categories: Negative, Positive, Uncertainty, Litigious, Strong Modal, Weak Modal, and Constraining, each of which has a comprehensive list of sentiment words specifically designed for the language used in the business domain.

Specifically, the seven sentiment categories above have 2,355, 354, 297, 904, 19, 27, and 184 words, respectively. The Negative and Positive lists include words that appear in negative and positive contexts, respectively. For example, the Negative category starts with ABANDON, ABANDONED, and ABANDONING, while the Positive category starts with ABLE, ABUNDANCE, and ABUNDANT. The Uncertainty list categorizes words reflecting uncertainty or imprecision, such as CONTINGENCY and INDEFINITE. The litigious list includes words denoting a propensity for legal contest, such as CLAIMANT, TESTIMONY, and LEGISLATION. The Strong Modal and Weak Modal lists categorize words expressing levels of confidence, such as ALWAYS and STRONGLY for Strong Modal and ALMOST and COULD for Weak Modal. Lastly, the Constraining list developed in Bodnaruk et al. (2015) includes words related to financial constraints and starts with ABIDE, ABIDING, and BOUND. The seven word lists were originally developed specifically in the context of 10-K filings, not of social media data, so the application of the lexicon to tweet text may not be ideal. As modifying the foundational words in each list is beyond the scope of this paper, however, we follow their recommendation on the word lists.

Since any of the words listed in each of the seven sentiment categories can match a word

in $\mathcal{W}$ defined earlier, we denote $\mathcal{S}_c \subset \mathcal{W}$ as a set of sentiment words in category $c \in \mathcal{C}$, where $\mathcal{C}$ is a set of the seven sentiment categories, and thus $|\mathcal{C}| = 7$. Accordingly, our data space $\mathcal{U} \times \mathcal{P} \times \mathcal{W}$ and $\mathcal{U}_t \times \mathcal{P}_t \times \mathcal{W}_t$ at time $t \in \mathcal{T}$ can be transformed to $\mathcal{U} \times \mathcal{P} \times \mathcal{S}_c$ and $\mathcal{U}_t \times \mathcal{P}_t \times \mathcal{S}_{c,t}$ with respect to $c \in \mathcal{C}$, respectively, when we are only interested in sentiment words instead of all words.

3.3. Creating Time Series Data

3.3.1. CEO hashtag features

Due to the small number of CEO users and their tweets in the early period of data, we need to trim the data set to avoid the data sparsity problem, which otherwise could have a negative impact on prediction. We remove all the tweets posted before Monday, June 29, 2009 (represented with a vertical dashed line in Figure 2), based on the two facts that 1) June 2009 was determined by the Business Cycle Dating Committee of the National Bureau of Economic Research (NBER)[9] to be the end of the recession that had begun in December 2007 and 2) the number of CEO users had increased exponentially until that month as presented in Figure 2, which provides a significant amount of tweets to be used for prediction. The last date in our data set is December 31, 2021. This results in the reduced final set $\mathcal{T}$ of timestamps: 3,152 days for daily analysis, i.e., $|\mathcal{T}| = 3152$, and 653 weeks for weekly analysis, i.e., $|\mathcal{T}| = 653$.

We next construct time series data for each hashtag, such that the number of tweets containing each hashtag is counted with the tweet count metric and also the number of users mentioning each hashtag is counted with the user count metric. When calculating the count, two different time frequencies, day and week, are applied to make daily and weekly predictions of the stock market indicators. Here, daily counts are calculated first, based on which weekly counts are aggregated by week. Formally, for a hashtag $h \in \mathcal{H}$ at a time

[9]We refer to the Business Cycle Dating Committee Announcement on September 20, 2010, https://www.nber.org/news/business-cycle-dating-committee-announcement-september-20-2010.

$t \in \mathcal{T}$, the tweet count function $TC(h,t) = |\{(p_{j,t}, h)|j \in \{1, ..., |P|\}\}|$ and the user count function $UC(h,t) = |\{(u_{i,t}, h)|i \in \{1, ..., |U|\}\}|$. For daily time series data, as there would be no corresponding stock market data for weekends and holidays when the stock markets were closed, the counts for those closed days are aggregated with the count for the last week day when the market was open. For example, the counts for Saturday and Sunday are aggregated with the count for Friday. This issue does not apply to the weekly time series data, as all the daily counts are simply aggregated by week.

Another issue with these raw hashtag counts is that, as more people use Twitter over time, the numbers of CEO users and their tweets and hashtags grow as well, as already shown in Figure 2. To address this bias, we divide the raw tweet count by the total number of tweets for that day or week with the tweet count metric and, likewise, divide the raw user count by the total number of users for that day or week with the user count metric. Each User object from Twitter API provides information of when the account was created, which allows us to calculate the number of CEO users existing at a certain point. As there would be many zeros in the raw counts, we add one to all values before dividing them by tweet or user count. We define hashtag features as these normalized counts, which can be formally denoted as follows:

$$NormTC(h,t) = \frac{TC(h,t) + 1}{|\mathcal{P}_t|}, \tag{1}$$

$$NormUC(h,t) = \frac{UC(h,t) + 1}{|\mathcal{U}_t|}, \tag{2}$$

for hashtag $h \in \mathcal{H}$ and time $t \in \mathcal{T}$.

3.3.2. CEO sentiment features

Using the sentiment lexicon introduced in Section 3.2.2, we count the number of tweets containing any of the matching words in each sentiment category with the tweet count metric

and also the number of users mentioning the matching words with the user count metric, and then create time series data for the seven categories by day, week, and month in the same way as with hashtags. Formally, for a sentiment category $c \in \mathcal{C}$ at a time $t \in \mathcal{T}$, the tweet count function $TC(c,t) = |\{(p_{j,t}, s_{c,t}) | j \in \{1, ..., |P|\}\}|$ and the user count function $UC(c,t) = |\{(u_{i,t}, s_{c,t}) | i \in \{1, ..., |U|\}\}|$. We finally normalize the raw counts by dividing them by the total number of tweets for that timestamp with the tweet count metric and divide them by total number of users for that timestamp with the user count metric. We define sentiment features as these normalized counts, which can be formally denoted as follows:

$$NormTC(c,t) = \frac{TC(c,t)+1}{|\mathcal{P}_t|}, \tag{3}$$

$$NormUC(c,t) = \frac{UC(c,t)+1}{|\mathcal{U}_t|}, \tag{4}$$

for sentiment category $c \in \mathcal{C}$ and time $t \in \mathcal{T}$.

3.3.3. Stock market indicators

In order to combine the hashtag and sentiment time series data described above with stock market data, we retrieve the historical data of the four major stock indexes including S&P 500 (called SP500 hereafter), Dow Jones Industrial Average (Dow30), NASDAQ Composite (Nasdaq), and Russell 2000 (Russell2000) from the Yahoo! Finance web site[10] with three different frequency options, daily, weekly, and monthly. Each of the historical stock market data sets includes six columns of values including Open, High, Low, Close, Adj Close, and Volume. We use Close and Volume as they are to refer to return and trading volume, respectively, while deriving a new variable called Volatility being the difference between High and Low to capture the uncertainty ranging from High to Low. In addition, we derive

[10]See https://finance.yahoo.com/.

another variable called Direction, which has one when the previous Close value goes up at the current timestamp or zero when the Close value goes down or stay. Those four variables Close, Volume, Volatility, and Direction will be used as stock market indicators to predict in the following section.

4. Stock Market Prediction

In this section, we present the results and findings from predicting the three numeric stock market indicators Close, Volume, and Volatility, which can be classified as a regression problem, and also the Direction of stock prices, which is a binary classification problem. We also compare the performance of our approaches with that of other baselines.

4.1. Predicting Numeric Stock Market Indicators

We predict various stock market indicators using hashtags used by CEOs on Twitter and their sentiment words. This is a linear regression problem, as we aim to forecast the stock market indicators as numeric values. In particular, we consider the following predictive model for the stock market indicators:

$$y_t = \alpha + \boldsymbol{x}_{t-1}^{'}\boldsymbol{\beta} + \varepsilon_t, \quad t = 2, ..., |\mathcal{T}| \tag{5}$$

where y_t is the stock market indicator, $\boldsymbol{x}_{t-1}$ is $m \times 1$ vector of regressors at $t-1$, ε_t is the idiosyncratic error term, α is a constant, $\boldsymbol{\beta}$ is $m \times 1$ vector of parameters of interest, and t is a timestamp of either daily or weekly frequency. For daily regression, the number of observations is $|\mathcal{T}| = 3152$, and for weekly regression, $|\mathcal{T}| = 653$.

The hashtag time series data for the 1,000 hashtags in $\mathcal{H}$ and the sentiment time series data for the seven sentiment categories in $\mathcal{S}$ are used as regressors $\boldsymbol{x}_{t-1}$ in our predictive models, while the stock market time series data for the three stock market indicators (Close, Volume, and Volatility) from the four stock indexes (SP500, Dow30, Nasdaq, and Russell2000)

are used as dependent variables to forecast. We therefore define 12 dependent variables from the combinations of the three stock indexes and the four stock market indicators and name them SP500_Close, Dow30_Close, Nasdaq_Close, Russell2000_Close, SP500_Volume, Dow30_Volume, Nasdaq_Volume, Russell2000_Volume, SP500_Volatility, Dow30_Volatility, Nasdaq_Volatility, and Russell2000_Volatility. In addition, we add the first-order lagged values of the dependent variables to the set of regressors for both the daily and weekly data in order to take into account possible serial correlation in the error term. We put '_L1' after each stock market indicator name as a suffix, e.g., 'SP500_Close_L1'. Note that we do not combine the hashtag features and sentiment features into a single model, as the two sets of features are heterogeneous and thus should be treated separately, as supported by Cookson, Lu, Mullins, and Niessner (2024).

In predictive models for stock returns, there have been concerns on the spurious regressions or regressions with persistent lagged regressors (e.g., Granger and Newbold, 1974; Stambaugh, 1999; Ferson, Sarkissian, and Simin, 2003). To address the issue, we proceed to unit-root test to check whether the time series for each regressor and dependent variable is either stationary or non-stationary and transform the variable by taking the percentage change if it is non-stationary. Then, as the last step, we standardize all variables by removing the mean and scaling to unit variance in order to facilitate comparison and interpretation across regressors. Figures 3 and 4 present the standardized percentage changes of the twelve business-related hashtag time series as examples and the standardized percentage changes of seven sentiment time series, respectively.

It is well-known that the OLS estimator is inconsistent when the number of regressors m is larger than the number of observations $|\mathcal{T}|$ in time series (note that this only applies to our 1,000 hashtag features, not to the seven sentiment features). Thus, when 1,000 hashtag features are used as the regressors, we estimate the model using the least absolute shrinkage and selection operator (**lasso**) approach with L_1 penalty introduced by Tibshirani (1996) and select relevant regressors. As a comparison, we have estimated the model by the OLS

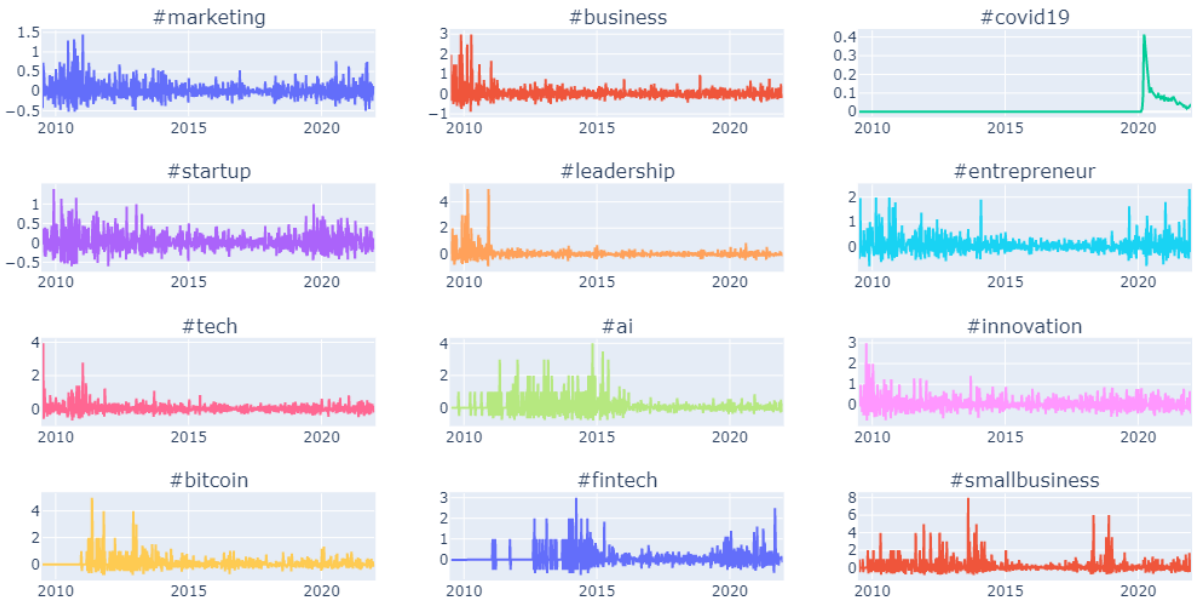

Figure 3. Standardized percentage changes of the twelve business-related hashtag time series
We standardize the percentage changes of the hashtag series by removing the mean and scaling to unit variance. The standardized percentage changes of the twelve business-related hashtag time series over time are plotted.

Figure 4. Standardized percentage changes of the seven sentiment time series
We standardize the percentage changes of the sentiment time series by removing the mean and scaling to unit variance. The standardized percentage changes of the sentiment time series over time are plotted.

with the most-frequently mentioned regressors of which the number is smaller than $|\mathcal{T}|$ and confirmed that the lasso regression outperforms the OLS. The lasso approach estimates coefficients of the linear model by minimizing the sum of squared residuals subject to a penalty term. The lasso estimate $(\widehat{\alpha}, \widehat{\boldsymbol{\beta}})$ is defined by

$$(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}) = \arg\min \left\{ \sum_{t=1}^{|\mathcal{T}|} \left(y_t - \alpha - \boldsymbol{x}_{t-1}^{'} \boldsymbol{\beta} \right)^2 + \lambda \sum_{i}^{m} |\beta_i| \right\}, \tag{6}$$

where $\lambda \geq 0$ is a tuning parameter and β_i is a scalar coefficient for i-th regressor in $\boldsymbol{x}_{t-1}$. Because of the penalty function the lasso tends to generate some coefficients that are exactly zero, so that we can select the non-zero coefficients as relevant regressors.

It has been shown in the recent studies that the lasso could be inconsistent for variable selection (Zou, 2006). Furthermore, if variables have a grouped structure, it is more desirable to proceed prediction based on a subset of important groups (Yuan and Lin, 2006). Thus, we also utilize the **adaptive group lasso** method to overcome drawbacks of the original lasso. We assume that $\boldsymbol{x}_t$ can be grouped into l factors as $\boldsymbol{x}_t = (\boldsymbol{x}_{t1}^{'}, ..., \boldsymbol{x}_{tl}^{'})^{'}$, where $\boldsymbol{x}_{ti} = (\boldsymbol{x}_{ti1}, ..., \boldsymbol{x}_{tid_i})^{'}$ is a group of d_i variables. The adaptive group lasso estimate $(\widehat{\alpha}, \widehat{\boldsymbol{\beta}})$ is defined by

$$(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}) = \arg\min \left\{ \sum_{t=1}^{|\mathcal{T}|} \left(y_t - \alpha - \boldsymbol{x}_{t-1}^{'} \boldsymbol{\beta} \right)^2 + \lambda \sum_{i}^{l} w_i ||\beta_i|| \right\}, \tag{7}$$

where $w_i \geq 0$ is a weight for i-th group in $\boldsymbol{x}_{t-1}$ and where $||\beta_i|| = (\beta_i^{'} \beta_i)^{1/2}$. Below we provide a detailed procedure for grouping hashtags. Note that if the number of variables in each group is one (i.e., $d_i = 1$ for all groups) and the weight is one (i.e., $w_i = 1$ for all groups), the equation (7) essentially goes back to the lasso equation (6). In both lasso and adaptive group lasso, users are required to select the tuning parameter λ. We find the optimal parameter by searching over $\lambda = [0.001, 0.003, 0.01, 0.03, 0.1, 0.3]$. Throughout the paper, we estimate the standard errors using heteroskedasticity– and autocorrelation–robust standard errors (Newey and West, 1987).

We perform out-of-sample forecasting by splitting the entire time series data set into an training set and a test set. We estimate models using the lasso and adaptive group lasso on the training set and predict the dependent variable at the very next timestamp in the test set based on the estimated coefficients when hashtags are used as the regressors. Similarly, OLS is applied to the model when sentiment features are used in place of hashtags. Two dimensions are considered for evaluation: model size and horizon. For the model size dimension, we evaluate how the training set size affects the forecasting performance by trying nine different sizes ranging from 0.1 (i.e., 10% of the entire time series) to 0.9 (i.e., 90% of the entire time series). Following the literature on forecasting, we consider one of the most commonly-used schemes, rolling windows scheme. As well described in Elliott and Timmermann (2016), the rolling windows scheme for forecasting refers to the practice of adopting an equal-weighted window of the most recent $\bar{\omega}$ observations to estimate the parameters of the model and dropping older observations as new observations are added. Specifically, we use the following estimator based on the rolling windows for OLS when sentiment features are of interest:

$$(\widehat{\alpha}_t, \widehat{\boldsymbol{\beta}}_t) = \arg\min \left\{ \sum_{s=t-\bar{\omega}+1}^{t} \left(y_s - \alpha - \boldsymbol{x}_{s-1}^{'} \boldsymbol{\beta} \right)^2 \right\}. \tag{8}$$

On the other hand, we have the following estimators for lasso:

$$(\widehat{\alpha}_t, \widehat{\boldsymbol{\beta}}_t) = \arg\min \left\{ \sum_{s=t-\bar{\omega}+1}^{t} \left(y_s - \alpha - \boldsymbol{x}_{s-1}^{'} \boldsymbol{\beta} \right)^2 + \lambda \sum_i |\beta_i| \right\}. \tag{9}$$

The rolling windows scheme can be similarly applied to the adaptive group lasso.

Let $f_{t+q|t}$ be q-step-ahead forecast of the dependent variable given the estimates $(\widehat{\alpha}_t, \widehat{\boldsymbol{\beta}}_t)$ at time t with the forecast horizon q based on the rolling windows scheme. Then the mean

squared errors is calculated as the sample loss function as follows:

$$MSE = (|\mathcal{T}| - q)^{-1} \sum_{t=1}^{|\mathcal{T}|-q} e_{t+q|t}^2, \tag{10}$$

with q-step-ahead forecast errors, $e_{t+q|t} = y_{t+q} - f_{t+q|t}$, over a sample $t \in \{1, ..., |\mathcal{T}| - q\}$. We then find the best model that yields the smallest MSE in each stock indicator.

For the horizon dimension, we define *horizon* as the time length of the future to be predicted by the model. For example, if the horizon is set to 1 for weekly analysis, a model is built with the data of up to a certain week and it attempts to predict the target value of the very next week; if the horizon is set to 2, the model attempts to predict the target value of the week after next week. We evaluate how far the model predicts well for the future by considering different horizons ranging from 1 to 5 for both weekly and daily analysis. Formally, horizon $q \in \mathcal{Q} = \{1, 2, 3, 4, 5\}$.

Regarding the feature selection algorithms for the hashtag features, specifically, we compare the simple lasso technique with the adaptive group lasso technique, based on the idea that grouping of similar hashtags could help to improve the model performance. As there is no controlled vocabulary for hashtags on Twitter, some hashtags can refer to the same thing. For example, the four hashtags, #covid, #covid19, #covid2019, #covid_19, and #coronavirus share exactly the same meaning and thus can be used interchangeably. Examples include different names of an entity as with #coronavirus and #covid19, different representations of an entity as with #covid19 and #covid_19, singular and plural nouns as with #startup and #startups, variations of a word as with #happy (adjective) and #happiness (noun), or two words that substantially overlap in their meanings as with #quarantine and #stayhome. Based on our belief that it would make more sense to consider those hashtags to be one rather than separated, we group the hashtags that are so similar that they can be used interchangeably.

In order to find those very similar hashtags, we first apply some heuristics: a) find the

pairs of hashtags with and without an underscore, e.g., #covid_19 and #covid19, b) find the pairs of hashtags with and without trailing numbers or with different trailing numbers, e.g., #iwd and #iwd2021 or #iwd2021 and #iwd2020, c) find the pairs of hashtags that have the same word stem in common using the word stemming technique, e.g., #inspiration and #inspiring. In addition to the simple heuristics, we also consider correlation between hashtags in the time series data. Specifically, we include the pairs of hashtags if the correlation coefficient between the hashtags is greater than 0.95 or less than -0.95, which means extremely highly positive/negative correlation, e.g., #thanksgiving and #blackfriday (correlation coefficient of 0.987). As the weekly and daily time series differ, there are both weekly and daily versions of highly correlated hashtags.

Once all the pairs of similar hashtags have been identified, we build an undirected graph in which nodes represent hashtags and edges represent that the two hashtags are similar. We then discover all of the connected components in the graph, in which any two nodes are connected to each other by paths. Figure 5 demonstrates three examples of the connected components found. In the first hashtag cluster, #stayathome can be connected to #quarantine via #stayhome, which results in all those three hashtags being in the same cluster. The second hashtag cluster in the figure comprises three hashtags, #blackfriday, #thanksgiving, and #happythanksgiving, all directly connected to each other. The third hashtag cluster demonstrates a larger cluster of six hashtags that are partially interconnected to each other. Table 6 presents all of the 56 identified clusters of similar hashtags from the weekly time series data. We manually examined the clustering results and found no errors at least from the precision perspective. This hashtag clustering information is finally passed to the adaptive group lasso algorithm as the group index information.

Figure 6 demonstrates how the MSE from weekly forecasting changes with different model sizes for each of the hashtag time series and sentiment time series and for each of the four stock markets. Here, the lower MSE, the better performances. The cross (x) on each line indicates the minimum MSE achieved. The figure shows that the MSE tend to go up as the

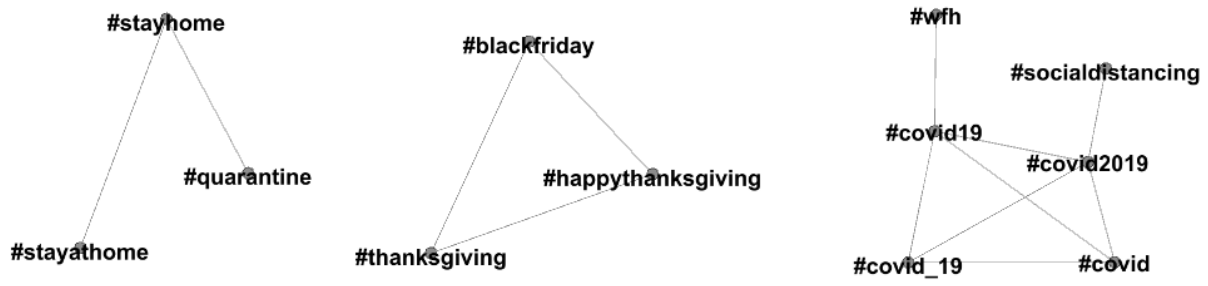

Figure 5. Three examples of clusters of similar hashtags

Three examples of clusters of similar hashtags are plotted in a graph. The middle cluster comprises three hashtags that are fully-connected, while the other two comprise hashtags that are partially-connected.

model size increases until it drops when the model size is the largest, i.e., 0.9. Most of the minimum MSE are achieved when the model size is either small or large. The forecasting based on sentiment features always outperforms two lasso approaches with hashtag features. The hashtag feature selection by the adaptive group lasso performs better than the one by the simple lasso in terms of MSE, which indicates that our hashtag clustering works well in weekly forecasting. Figure 7 presents the MSE from daily forecasting with different model sizes. The sentiment features perform better than hashtags features based on both lasso approaches. The MSE from the simple lasso are more stable than those from the adaptive group lasso in daily forecasting. Overall, the minimum MSE from all three methods are close to each other.

Figure 8 provides a slightly different perspective: how the MSE change with different horizons in weekly forecasting. The sentiment-based forecast again performs the best in most cases over different horizons except for Russell2000_Close where the adaptive group lasso with hashtag features slightly perform better than the forecasting with sentiment features. Among the figures for sentiment features, the results for Volume and Volatility show that the lowest MSE are mostly achieved when the horizons are small, which is in line with our expectation. On the other hand, results for Close indicate that MSEs are relatively flat over different horizons. The MSE from the adaptive group lasso are more stable than those from the simple lasso over different horizons in weekly forecasting. Figure 9 presents MSE change

Figure 6. Mean Square Errors by different model sizes from weekly forecasting

For weekly forecasting, we calculate the Mean Squared Errors by different model sizes ranging from 0.1 to 0.9. The lower MSE, the better.

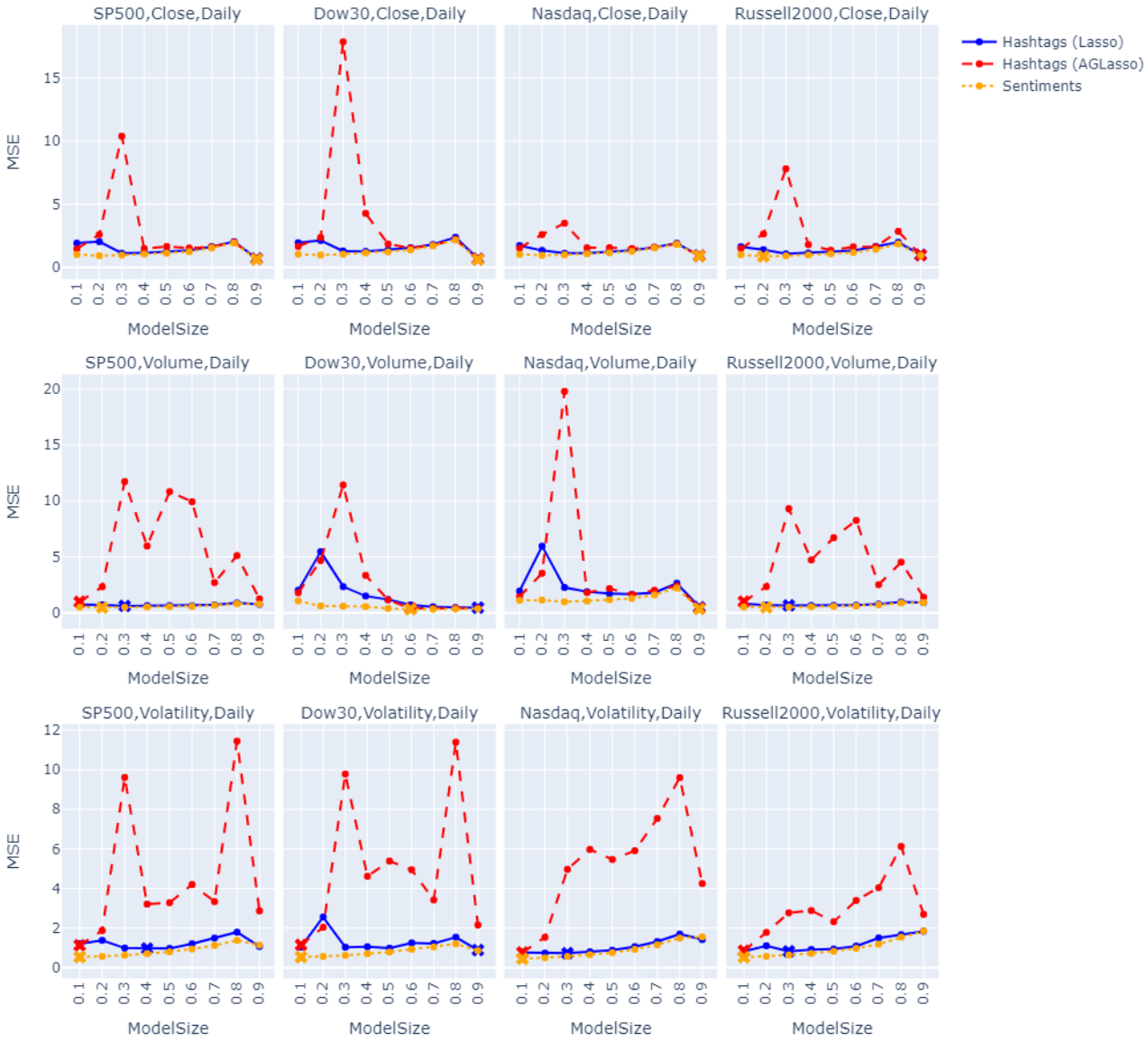

Figure 7. Mean Square Errors by different model sizes from daily forecasting

For daily forecasting, we calculate the Mean Squared Errors by different model sizes ranging from 0.1 to 0.9. The lower MSE, the better.

Figure 8. Mean Square Errors by different horizons from weekly forecasting

For weekly forecasting, we calculate the Mean Squared Errors by different horizons ranging from 1 to 5. The lower MSE, the better.

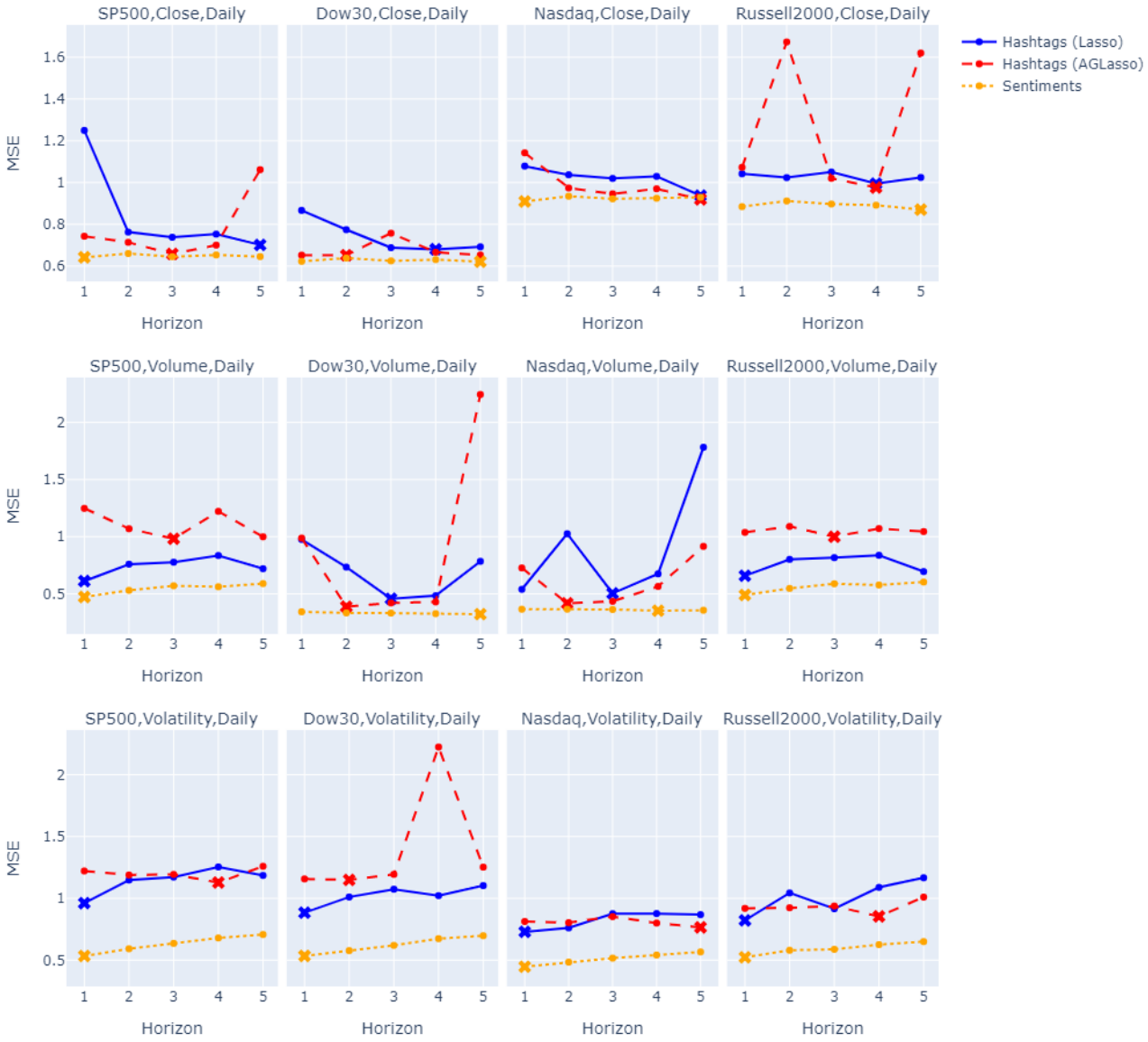

Figure 9. Mean Square Errors by different horizons from daily forecasting

For daily forecasting, we calculate the Mean Squared Errors by different horizons ranging from 1 to 5. The lower MSE, the better.

with different horizons in daily forecasting. Similarly, the forecasting based on sentiment features outperform two lasso approaches.

Tables 7 and 8 present the parameter setting that yields the best forecasting performance for each of the combinations of two frequencies, four stock markets, three targets, three feature types, nine model sizes, and five horizons. The lowest MSE achieved from the hashtag and sentiment features are 0.7 and 0.548, respectively, both from SP500_Volume. We believe that these forecasting performances are promising, considering the widely-known difficulty of stock market prediction.

It also shows that the optimal forecasting model is varying across different target variables. This finding confirms that which factors matter in explaining stock market movement varies by stock market indicators and stock indexes of interest. SP500 index tracks 500 large U.S. companies and the stocks in the index represents roughly 75% of all stocks on the New York Stock Exchange (NYSE). Dow30 index follows the 30 largest U.S. companies which represents approximately 25% of all publicly traded stocks and it gauges the overall health of financial markets for some of biggest "blue chip" companies. Nasdaq index tracks about 3,000 companies traded on the Nasdaq Exchange and measures the performances of innovative and tech-sector companies. Russell2000 index tracks 2,000 of the smallest-capitalization companies in the stock market and it is frequently used as a benchmark for small-cap investors. Since these indexes track different types of stocks and indicators measures different characteristics of stock movement, it makes sense to see differential predictive power of hashtags and sentiment features across different stock indexes and indicators.

In order to learn which hashtags are informative of predicting stock markets, we estimate the equation (5) using the full sample period. For brevity, Table 9 only reports the top-5 most significant hashtag features from lasso regression for weekly frequency if there are more than five significant hashtag features identified, sorted by absolute coefficient. It turns out that there are lots of hashtags which are statistically significant. This confirms that many hashtags have significant predictive power on the stock market indicators and that

the estimation methods are capable of capturing such relationships in the model where the number of regressors is larger than the number of observations.

We find that various CEO hashtags predict stock returns, which is in line with those papers on stock return predictability such as Baker and Wurgler (2006) who find that high investor sentiment predicts strongly low returns in the cross-section. Interestingly, trading volume and volatility in addition to stock returns are also predicted by many hashtags. The result is similar to that in Das and Chen (2007) who find that the net of positive and negative opinion from stock message boards predicts stock index levels, volumes, and volatility and that in Antweiler and Frank (2004) who find that the bullishness of the stock messages are predictive of market volatility and disagreement among the messages is associated with more trading volume.

4.2. Predicting Stock Price Directions

The hashtag time series data for the 1,000 hashtags in $\mathcal{H}$ and the sentiment time series data for the seven sentiment categories in $\mathcal{S}$ are again be used as features in our predictive models, while the stock market time series data for the stock price directions from the four stock indexes are used as targets to predict. We therefore define four target variables named SP500_Direction, Dow30_Direction, Nasdaq_Direction, and Russell2000_Direction. This is a binary classification problem, as opposed to the regression problem that we dealt with in Section 4.1, as we attempt to classify the stock price direction as either 0 (going down or staying) or 1 (going up). In order to make the original time series data simple enough to fit into classification algorithms, we do not add the first-order lagged values of the target variables for this classification task, unlike the previous forecasting. We also keep the original features as predictors without transforming them (Park and Phillips, 2000). We, again, do not combine the hashtag features and sentiment features into a single model.

The rest of the process is basically the same as the previous forecasting except that it is a classification task. We perform out-of-sample prediction by splitting the entire time

series data set into a training set and a test set. We build models on the training set by applying six classification algorithms ranging from traditional classification algorithms such as k-Nearest Neighbors, Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines to state-of-the-art Deep Neural Network. These algorithms are listed in a survey by Jiang (2021) as widely-used classification algorithms for stock market prediction. We then predict the Direction variable at the timestamp in the test set that is horizon-steps away. Formally, we consider the following predictive model for the stock price direction:

$$y_{t+q} = \begin{cases} 1 & \text{if } p(x_t) > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where y_{t+q} is the q-step-ahead stock price direction over a sample $t \in \{1, ..., |\mathcal{T}| - q\}$ and $q \in \mathcal{Q}$, and $p(x_t)$ is the estimated probability at time $t \in \mathcal{T}$ that the stock price would go up at $t + q$ with the given features. Different classification algorithms determine different p functions depending on their underlying ideas. When building a model based on a selected classification algorithm, we find the optimal set of hyper-parameter values whenever possible by cross validation, i.e., further splitting the training set into a training set and a validation set.

As with the previous forecasting, two dimensions are considered for evaluation: model size and horizon. For the model size dimension, we evaluate how the training set size affects the prediction performance by taking nine different sizes ranging from 0.1 to 0.9. We again employ the rolling windows scheme for prediction. For the horizon dimension, we consider five different horizons ranging from 1 to 5 for both weekly and daily analyses. To compare prediction performances model by model, we employ classification accuracy, which is one of the most commonly-used and intuitive classification metrics, ranging from 0 to 1. Formally, accuracy is defined as the number of correct predictions divided by the number of all samples

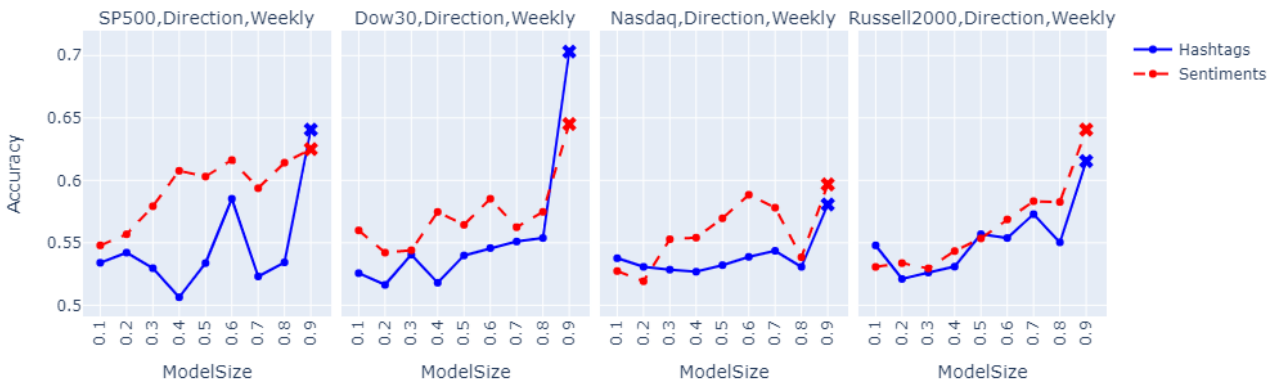

Figure 10. Prediction accuracy by different model sizes from weekly prediction

For weekly prediction, we calculate the prediction accuracy by different model sizes ranging from 0.1 to 0.9. The higher accuracy, the better.

as follows:

$$Accuracy = (|\mathcal{T}| - q)^{-1} \sum_{t=1}^{|\mathcal{T}|-q} r_{t+q|t}, \tag{12}$$

with q-step-ahead prediction results,

$$r_{t+q|t} = \begin{cases} 1 & \text{if } y_{t+q} = f_{t+q|t} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

over a sample $t \in \{1, ..., |\mathcal{T}| - q\}$ and $q \in \mathcal{Q}$, where y_{t+q} is the true target value (either 0 or 1) at time $t + q$ from the training set and $f_{t+q|t}$ is q-step-ahead prediction (0 or 1) of the target variable at time t with the prediction horizon q from the test set based on the rolling windows scheme. We then find the best model that yields the highest accuracy.

Figures 10 and 11 demonstrate how the weekly and daily prediction accuracy, respectively, changes with different model sizes for each of the hashtag time series and sentiment time series and for each of the four stock markets. Here, in contrast to the MSE previously, the higher classification accuracy, the better. The cross (x) indicates the maximum accuracy score achieved. The figures show that the accuracy tends to go up as the model size increases and

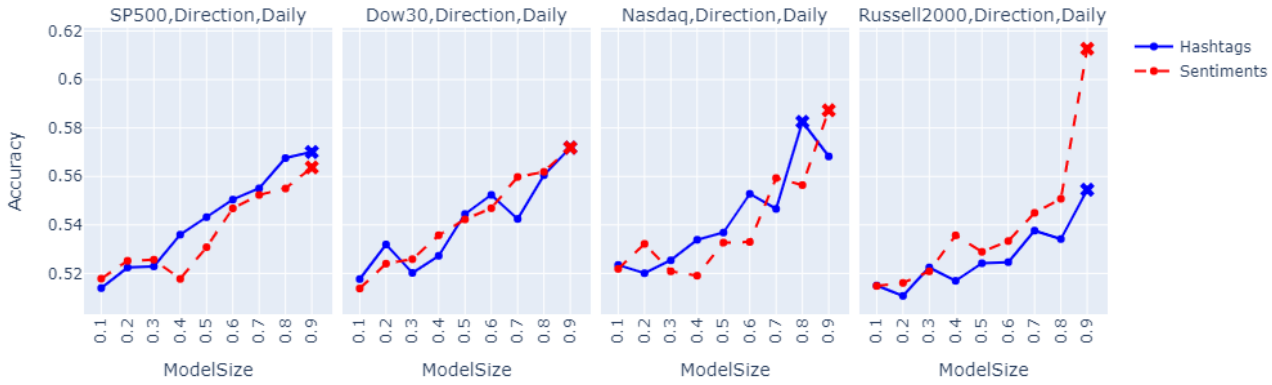

Figure 11. Prediction accuracy by different model sizes from daily prediction

For daily prediction, we calculate the prediction accuracy by different model sizes ranging from 0.1 to 0.9. The higher accuracy, the better.

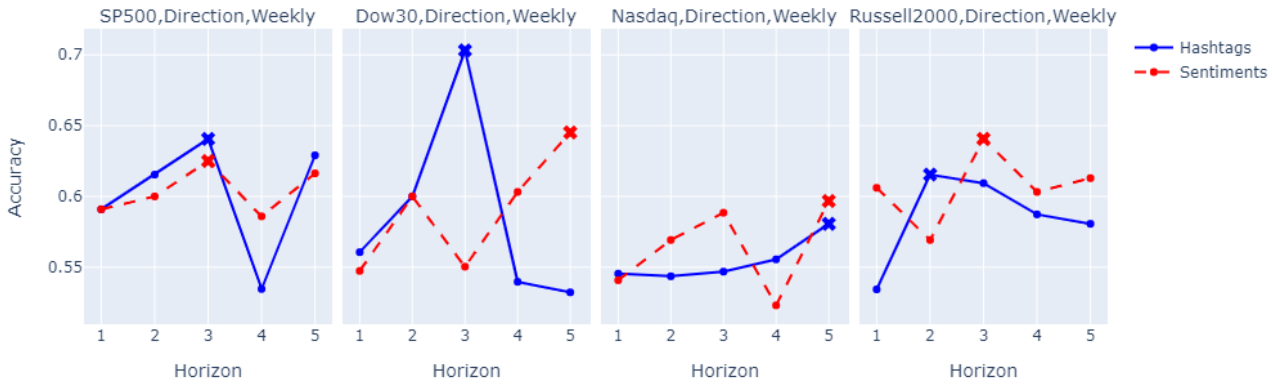

Figure 12. Prediction accuracy by different horizons from weekly prediction

For weekly prediction, we calculate the prediction accuracy by different horizons ranging from 1 to 5. The higher accuracy, the better.

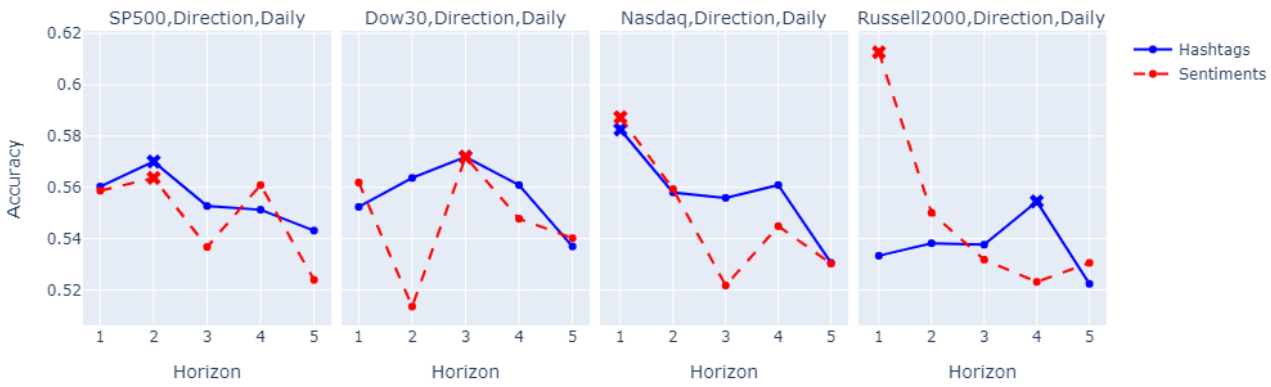

Figure 13. Prediction accuracy by different horizons from daily prediction
For daily prediction, we calculate the prediction accuracy by different horizons ranging from 1 to 5. The higher accuracy, the better.

thus most of the maximum accuracy scores are achieved when the model size is the largest, i.e., 0.9. The sentiment features generally outperform the hashtag features, although which type outperforms with respect to the highest accuracy varies case by case. Figures 12 and 13 present how the weekly and daily accuracy, respectively, changes with different horizons. Contrary to our expectation that the farther the model predicts, the worse the prediction would be, the accuracy does not always go down as the horizon increases, and the highest accuracy scores are achieved when the horizons are larger than one in most cases. This implies that it takes some time for what CEO mentions on Twitter to get reflected in the stock markets in terms of stock price direction.

Table 10 presents the parameter setting that yields the best prediction performance along with the contributing classification algorithm for each of the combinations of frequency, stock market, feature type, model size, horizon, and classification algorithm. The highest accuracy scores achieved from the hashtag and sentiment features are 70.3% and 64.5%, respectively, both from weekly Dow30, while the lowest accuracy scores achieved are 55.4% from daily Russell2000 and 56.4% from daily SP500, respectively. We believe that these prediction performances are surprisingly good, considering the widely-known difficulty of stock market prediction. As stated by Nguyen and Shirai (2015), the classification accuracy of 56% or higher is generally reported as satisfying results for stock predictions. Interestingly, weekly

predictions clearly outperform daily predictions, which reflects the difficulty of short-term stock market prediction with high variability. A variety of classification algorithms prove to contribute the best performances, ranging from the simplest k-Nearest Neighbors to the state-of-the-art Deep Neural Network.

As with the previous forecasting, we, again, identify the hashtags that are informative of stock price direction prediction over the full sample period. To that end, we calculate the importance score of each hashtag feature, which indicates how much a feature contributes to the classification outcome. For Logistic Regression, we can simply use the coefficients from the model as the importance scores; The Decision Trees algorithm supports native feature importance scores; for k-Nearest Neighbors, Support Vector Machines, and Deep Neural Network, as they do not offer native feature importance scores, we instead use the permutation feature importance scores, which can be acquired by calculating relative importance scores that are independent of the model used (Altmann, Toloşi, Sander, and Lengauer, 2010). When selecting the classification algorithm out of five for each of the target, we select the algorithm that yields the best performance, i.e., the highest accuracy, when the model size and the horizon are set to 0.9 and 1, respectively. Table 11 reports the top-5 most important hashtags features from each of the algorithm selected for weekly frequency only, sorted by absolute importance score. It turns out that a variety of hashtags have predictive power of the direction of stock prices. Some hashtags have positive importance scores for SP500 Direction, e.g., #brexit and #superbowl, whereas other hashtags have negative scores, e.g., #ff, #internationalwomensday, and #fb.

4.3. Comparison of Hashtag/Sentiment Features with Full Text Features

As elaborated earlier, we select the 1,000 most popular hashtags and predefined sentiment words for stock market prediction, as opposed to all words in the CEO tweet text. In order to compare our hashtag and sentiment features with full text features, we apply the Bidirectional Encoder Representations from Transformers, or BERT (Devlin, Chang, Lee, and

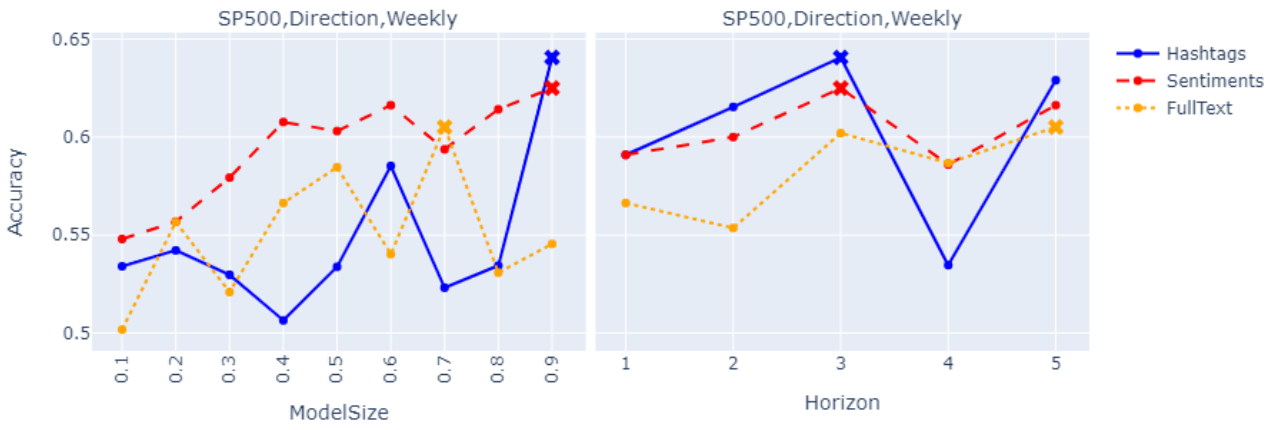

Figure 14. Prediction accuracy comparison: hashtag/sentiment features versus full text features

The figure compares the prediction accuracy from the models with the hashtag/sentiment features and the accuracy from the model with the full text features. The higher accuracy, the better.

Toutanova, 2019). Everything is the same as the previous stock price direction classification task, except that full text is used as features for classification. We perform this classification only for SP500 due to the considerable computation cost for the BERT algorithm. As shown in Figure 14, our hashtag and sentiment features outperform the full text features in terms of not only general performance but also the highest accuracy, i.e., 64.1% from the hashtag features versus 62.5% from the sentiment features versus 60.5% from the full text features. We believe that it is a better idea to utilize a select list of hashtags and sentiment words than considering all words in terms of both performance and computational cost.

4.4. Comparison of User Count with Tweet Count

As described in Section 3.2, we rely on both the user count and tweet count metrics when creating time series data but mostly present the user count-driven outcome in this paper. Recall that both metrics have different sets of top-1,000 hashtags as shown in Tables 4 and 5, whereas they have the same set of predefined sentiment words for the sentiment features. Figure 15 presents the pairwise lowest MSE (the lower MSE, the better) from the user count and tweet count metrics for each of the three weekly stock market indicators and for each

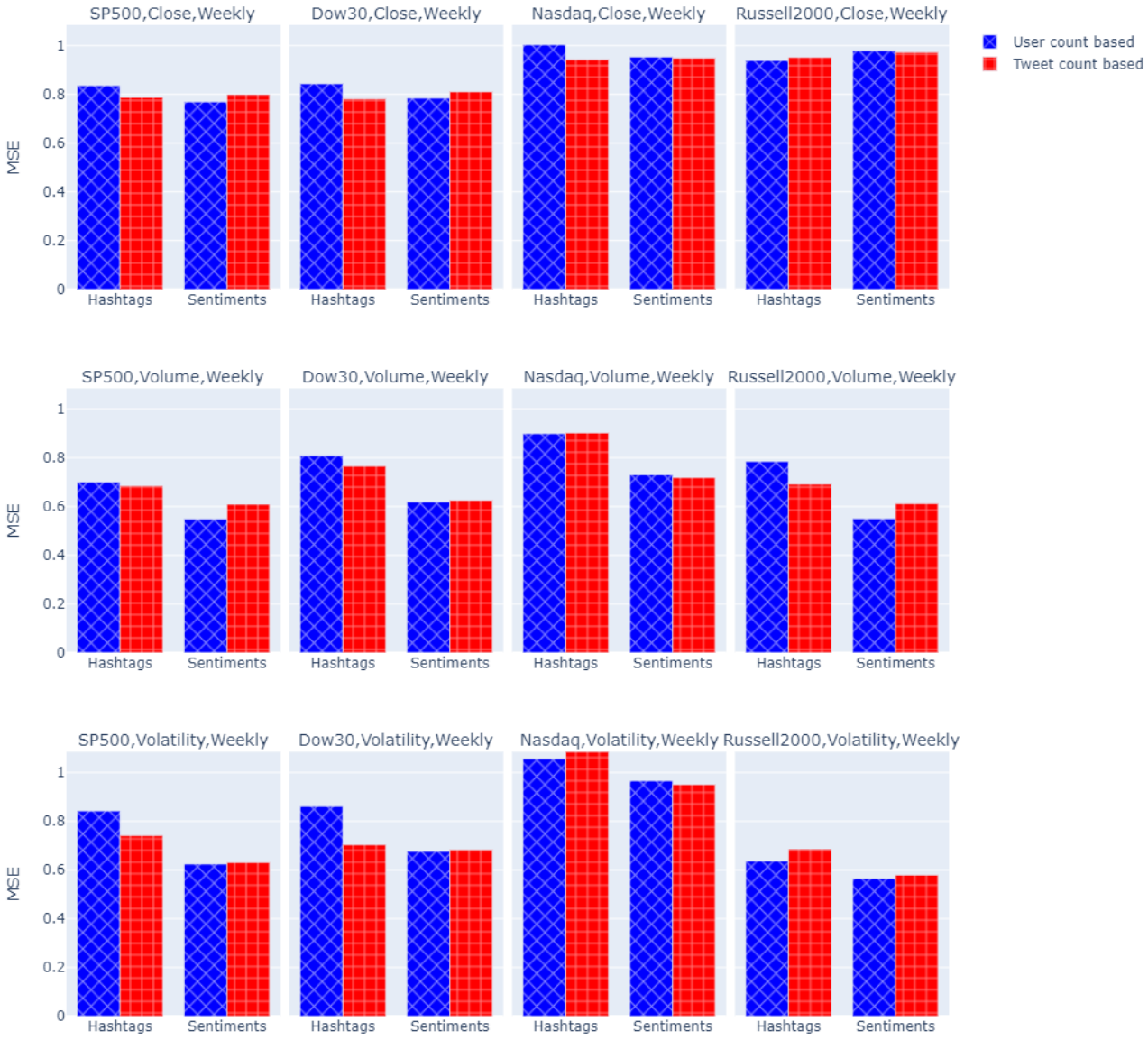

Figure 15. Mean Squared Errors comparison: user count-based versus tweet count-based
The figure compares the Mean Squared Errors from the user count-based models with those from the tweet count-based models. The lower MSE, the better.

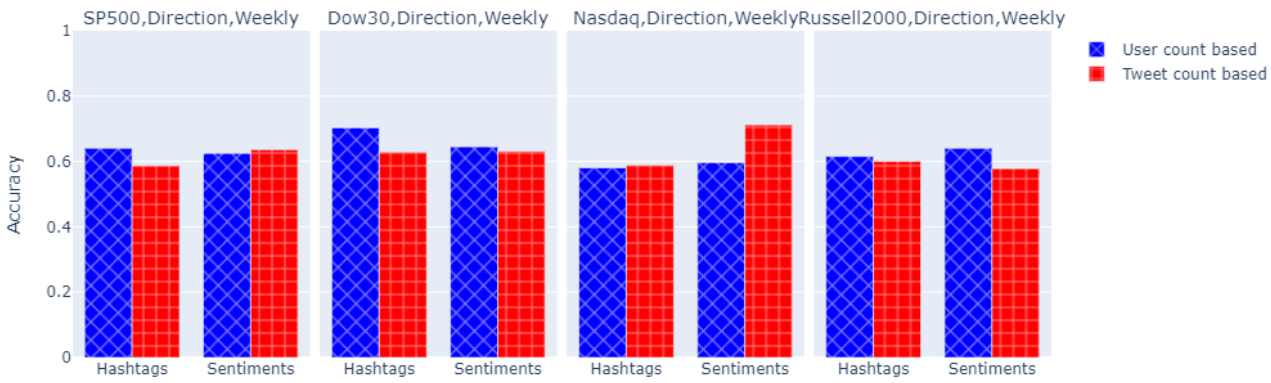

Figure 16. Prediction accuracy comparison: user count-based versus tweet count-based
The figure compares the prediction accuracy from the user count-based models with that from the tweet count-based models. The higher accuracy, the better.

of the four stock markets, and Figure 16 exhibits the pairwise highest accuracy scores (the higher accuracy, the better) from the two metrics for the weekly stock price direction. While it is hard to tell from Figure 15 alone which metric is better for regression, the overall classification performance of the user count metric slightly outweighs that of the tweet count metric as shown in Figure 16. This implies that our novel user count metric works better for classification or at least performs as well as the common tweet count metric. The tweet count-driven results are available on the Supplementary Online Appendix.

4.5. Comparison with Macroeconomic Variables

A large number of studies in finance literature have shown that a variety of macroeconomic and financial variables can predict stock returns (Fama and French, 1988, 1989; Cochrane, 1991, 2007; Pesaran and Timmermann, 1995, 2000; Welch and Goyal, 2008; Van Binsbergen and Koijen, 2010; Golez and Koudijs, 2018). If these variables are sufficient statistics for the relationship with stock indicators, our proposed CEO sentiment features would not have predictive power after controlling for them. In order to check such a possibility, we now include macroeconomic and financial variables considered by Welch and Goyal (2008) in the models for the stock market indicators (available from Amit Goyal's

website) and estimate the relationship between stock indicators and sentiment features. In that way, we can test if the information from the CEO sentiment features is subsumed by the macroeconomic and financial variables. As these macroeconomic and financial variables are available at most as monthly data, we reconstruct our sentiments on a monthly basis. We first present the list of dependent variables and control variables.

- Dependent variables: Four variables are considered. The first three variables are the same as before; Close, Volume, and Volatility. The fourth dependent variable is the equity premium (**Excess Return**), that is, the total rate of return on the stock market SP500 minus the prevailing short-term interest rate given as

$$r_t = \frac{P_t + D_t - P_{t-1}}{P_{t-1}} - TB_t$$

  where P_t is the stock price (SP500 index), D_t is dividends, TB_t is the return from holding the risk-free Treasury-bill.

- Control variables: Total 13 variables are included.

  1. Dividend Price Ratio (**d/p**): is the difference between the log of dividends paid on the SP 500 index and the log of prices.
  2. Dividend Yield (**d/y**): is the difference between the log of dividends and the log of lagged prices.
  3. Earnings Price Ratio (**e/p**): is the difference between the log of earnings on the SP 500 index and the log of prices.
  4. Dividend Payout Ratio (**d/e**): is the difference between the log of dividends and the log of earnings on the SP 500 index.
  5. Stock Variance (**svar**): is computed as sum of squared daily returns on the SP 500 index.
  6. Book-to-Market Ratio (**b/m**): is the ratio of book value to market value for the Dow Jones Industrial Average.

7. Net Equity Expansion (**ntis**): 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.
8. Treasury-bill rates (**tbl**): are interest rate on the 3-Month Treasury Bill; Secondary Market Rate from the economic research data base at the Federal Reserve Bank at St. Louis (FRED).
9. Long Term Rate of Returns (**ltr**): are returns on long-term government bonds.
10. Term Spread (**tms**): is the difference between the long-term yield on government bonds and the Treasury-bill.
11. Default Yield Spread (**dfy**): is the difference between BAA- and AAA-rated corporate bond yields.
12. Default Return Spread (**dfr**): is the difference between long-term corporate bond and long-term government bond returns.
13. Inflation (**infl**): is the Consumer Price Index (all urban consumers) from the Bureau of Labor Statistics.

As before, we perform unit-root tests on each control variable and transform the variable by calculating the percentage change if it is non-stationary. Before estimating control-augmented models, we regress each dependent variable on a constant and the first-order lags of seven sentiment time series on a monthly basis and find significant sentiment indexes in each stock indexes. Table 12 summarizes the set of sentiment features which are statistically significant at 10% significance level. It also reports the estimated coefficients and Newey–West heteroskedasticity– and autocorrelation–robust standard errors inside parentheses.

Interestingly, we find different set of significant sentiment features in the monthly analyses. 'Constraining' is significant predictors for stock return (i.e., Close) or excess return in all stock indexes. 'StrongModal' and 'Litigious' are also commonly detected as a significant predictor for stock return. 'Positive' is statistically significant predictors of trading volume in SP500, Nasdaq, and Russell2000, while 'Negative' is significant for trading volume in

Dow30. 'Uncertainty' is the most significant predictor of stock volatility in SP500. On the other hand, no sentiment index explains stock volatility in Nasdaq.

We next estimate control-augmented models by adding one of the macroeconomic or financial variables as a control variable to the basic model described above. For this, we only consider the models that obtain at least one significant sentiment feature in Table 12 and estimate them by regressing the stock indicator on all significant sentiments and one of the control variables, as follows:

$$y_t = \phi + \boldsymbol{z}_{t-1}^{'}\boldsymbol{\theta} + w_{t-1}\pi + \epsilon_t, \quad t = 1, ..., |\mathcal{T}| \tag{14}$$

where y_t is a stock market indicator, $\boldsymbol{z}_{t-1}$ is $r \times 1$ vector of significant sentiment features at $t-1$ from Table 12, w_{t-1} is a scalar macroeconomic or financial variable at $t-1$ and ϵ_t is the error term; where ϕ is a constant, $\boldsymbol{\theta}$ is $r \times 1$ vector of parameters of interest, and π is the parameter of the control variable; and where t is monthly frequency ranging from June 2009 to December 2019.

Tables 13-23 report the estimation results for each stock market indicator. The findings confirm that the proposed CEO sentiment features still provide predictive power of the stock indicators such as stock returns beyond what the existing macroeconomic and financial variables do. We observe that including sentiment features significantly improves the adjusted R^2 in all models. For example, Table 13 shows that the multivariate regression of the excess return on 'Constraining' and 'Litigious' in addition to a macroeconomic/financial variable significantly improves R^2, compared to the univariate regression of the excess return on a macroeconomic/financial variable.

Next, we consider a "Kitchen Sink" regression that includes all the macroeconomic/financial variables in the equation 14. Thus, w_{t-1} in the equation 14 is now a vector of all the macroeconomic/financial variables. Table 24 reports the estimation results. Interestingly, in all stock indicators, we observe that the regression model with the sentiment features as well

as the macroeconomic/financial variables significantly improves the adjusted R^2, compared to the regression model with only macroeconomic/financial variables. These findings confirm that the sentiment features from the CEO tweets contain predictive power on the stock movement.

4.6. Possible structural break since 2020

The world economy has been experiencing an unprecedented challenges imposed by the pandemic since early 2020. The Business Cycle Dating Committee of the NBER has announced that a peak in monthly economic activity occurred in the U.S. economy in February 2020 which marks the end of the expansion that began in June 2009 and the beginning of a recession. Although it is not straightforward to claim a structural break and it is generally challenging to detect breaks in financial forecasting models (Kim, Morley, and Nelson, 2005; Timmermann, 2018), it would be reasonable to be concerned about a structural break in the relation between stock market indicators and CEO hashtag (or sentiment) features since 2020 during our sample period. To address this concern, we re-estimate the weekly regressions in the previous sections (Tables 9 and 11) using the same algorithms and settings as before, but the data only from June 2009 to December 2019.

Tables 25-26 report the most significant hashtag features from the regression model for numeric stock market indicators in Section 4.1 and the predictive model for the stock price directions in Section 4.2, respectively. As expected, there are changes in the ranking of the most popular hashtag features. For example, the hashtags related to COVID-19 no loner show up, since these begin to appear after January 2020. However, we observe that many hashtag features are still statistically significant and that some of them are the same as those from the full-sample estimation in Tables 9 and 11. These results confirm that our main findings are robust to the possible structural break in that CEO tweets still contain informational content on the U.S. stock market.

5. Concluding Remarks

In this study, we demonstrate how the collective voice of CEOs on Twitter can be analyzed in predicting the economy, especially the most prominent stock indexes in the U.S. Specifically, we create a large, unique sample of CEOs by identifying 4,714 CEO users on Twitter and translate the unstructured tweet text data into structured time series by applying text analysis techniques. We perform comprehensive time series prediction analyses in a wide range of settings and show that the select list of hashtags and sentiment words in CEO tweets have significant predictive power of the stock market indicators such as return, trading volume, volatility, and stock price direction for each of the four stock market indices – S&P 500, Dow 30, Nasdaq, and Russell 2000. We provide experimental evidence that CEOs' language expressed in their tweets has informational content on various stock market indicators.

We believe this study could be a stepping stone for promising extensions as future research. First, we do not take user weighting into account in this study when constructing time series data. In other words, all CEO users and their tweets are equally weighted. Considering the different amounts of influence of CEOs, however, it would make more sense to measure each CEO's influence and weigh their tweets accordingly. For example, we may simply take the number of followers as the influence score of a user and use that measure to weigh the tweets posted by the user, rather than counting every tweet as just one as done in this study. Second, we do not consider the network of CEO users in our study. Each CEO user can have follow relationships on Twitter not only with the other CEOs in our sample but also with any other general users on Twitter. By looking at the network structure of CEO users, we can advance user weighting schemes mentioned above by calculating CEO network centrality (El-Khatib, Fogel, and Jandik, 2015) or explore how executive peer networks can affect corporate decisions (Shue, 2013). Third, in addition to being able to identify the 4,714 CEO users who describe themselves as CEOs in their bios, we are also able to identify the Twitter accounts of the firms they represent for approximately one thirds of the CEOs, as

discussed earlier in Section 3.1. This unique information on CEO-firm matching provides us a great opportunity to investigate how CEOs are interacting with or taking advantage of the Twitter accounts of the firms they currently work for. Lastly, the aggregate-level analysis of this study can be extended further to a firm-level analysis. In order to do this, we need to map the CEO users and their firms virtually existing on social media to the corresponding real world entities of which characteristics are publicly available.

References

Adams, R. B., Almeida, H., Ferreira, D., 2005. Powerful CEOs and Their Impact on Corporate Performance. Review of Financial Studies 18, 1403–1432.

Altmann, A., Toloşi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26, 1340–1347.

Antweiler, W., Frank, M. Z., 2004. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. Journal of Finance 59, 1259–1294.

Baker, M., Wurgler, J., 2006. Investor Sentiment and the Cross-Section of Stock Returns. Journal of Finance 61, 1645–1680.

Bertrand, M., Schoar, A., 2003. Managing with Style: The Effect of Managers on Firm Policies. Quarterly Journal of Economics 118, 1169–1208.

Blankespoor, E., Miller, G., White, H., 2014. The Role of Dissemination in Market Liquidity: Evidence from Firms' Use of Twitter. The Accounting Review 89, 79–112.

Bodnaruk, A., Loughran, T., McDonald, B., 2015. Using 10-K Text to Gauge Financial Constraints. Journal of Financial and Quantitative Analysis 50, 623–646.

Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. Journal of Computational Science 2, 1–8.

Capriotti, P., Ruesja, L., 2018. How CEOs use Twitter: A comparative analysis of Global and Latin American companies. International Journal of Information Management 39, 242–248.

Chatterji, A. K., Toffel, M. W., 2016. Do CEO Activists Make a Difference? Evidence from a Field Experiment. Working paper.

Chen, H., De, P., Hu, Y. J., Hwang, B.-H., 2014. Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. The Review of Financial Studies 27, 1367–1403.

Chen, H., Hwang, B.-H., Liu, B., 2018. The Emergence of 'Social Executives' and Its Consequences for Financial Markets. Working paper.

Cochrane, J. H., 1991. Production-based asset pricing and the link between stock returns and economic fluctuations. Journal of Finance 46, 209–237.

Cochrane, J. H., 2007. The Dog That Did Not Bark: A Defense of Return Predictability. Review of Financial Studies 21, 1533–1575.

Cookson, J. A., Lu, R., Mullins, W., Niessner, M., 2024. The Social Signal. Journal of Financial Economics 158.

Das, S. R., Chen, M. Y., 2007. Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. Management Science 53, 1375–1388.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.

El-Khatib, R., Fogel, K., Jandik, T., 2015. CEO Network Centrality and Merger Performance. Journal of Financial Economics 116, 349 – 382.

Elliott, G., Timmermann, A., 2016. Economic Forecasting. Princeton University Press.

Elliott, W. B., Grant, S. M., Hodge, F. D., 2018. Negative News and Investor Trust: The Role of $Firm and #CEO Twitter Use. Journal of Accounting Research 56, 1483–1519.

Fama, E. F., French, K. R., 1988. Dividend yields and expected stock returns. Journal of Financial Economics 22, 3–25.

Fama, E. F., French, K. R., 1989. Business Conditions and Expected Returns on Stocks and Bonds. Journal of Financial Economics 25, 23–49.

Ferson, W. E., Sarkissian, S., Simin, T. T., 2003. Spurious Regressions in Financial Economics? Journal of Finance 58, 1393–1413.

Gao, X., 2019. The Information Content of CEOs' Personal Social Media: Evidence from Stock Returns and Earnings Surprise. Working paper.

Gentzkow, M., Kelly, B., Taddy, M., 2019. Text as Data. Journal of Economic Literature 57, 535–74.

Gjerstad, P., Meyn, P. F., Molnár, P., Næss, T. D., 2021. Do president trump's tweets affect financial markets? Decision Support Systems 147, 113577.

Golez, B., Koudijs, P., 2018. Four centuries of return predictability. Journal of Financial Economics 127, 248–263.

Gow, I. D., Kaplan, S. N., Larcker, D. F., Zakolyukina, A. A., 2016. CEO Personality and Firm Policies. Working Paper 22435, National Bureau of Economic Research.

Granger, C. W., Newbold, P., 1974. Spurious Regressions in Economics. Journal of Econometrics 4, 111–120.

Huang, A. H., Zang, A. Y., Zheng, R., 2014. Evidence on the Information Content of Text in Analyst Reports. The Accounting Review 89, 2151–2180.

Huang, D., Jiang, F., Tu, J., Zhou, G., 2015. Investor Sentiment Aligned: A Powerful Predictor of Stock Return. Review of Financial Studies 28, 791–837.

Jiang, F., Lee, J., Martin, X., Zhou, G., 2019. Manager Sentiment and Stock returns. Journal of Financial Economics 132, 126–149.

Jiang, W., 2021. Applications of deep learning in stock market prediction: Recent progress. Expert Systems with Applications 184, 115537.

Jung, M. J., Naughton, J. P., Tahoun, A., Wang, C., 2018. Do Firms Strategically Disseminate? Evidence from Corporate Use of Social Media. The Accounting Review 93, 225–252.

Kaplan, S. N., Klebanov, M. M., Sorensen, M., 2012. Which CEO Characteristics and Abilities Matter? Journal of Finance 67, 973–1007.

Ke, Z. T., Kelly, B., Xiu, D., 2019. Predicting returns with text data. Journal of Econometrics (Forthcoming) .

Kim, C.-J., Morley, J. C., Nelson, C. R., 2005. The Structural Break in the Equity Premium. Journal of Business & Economic Statistic 23, 181–191.

Laniado, D., Mika, P., 2010. Making sense of twitter. In: *The Semantic Web – ISWC 2010*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 470–485.

Lee, D., Hosanagar, K., Nair, H., 2018. Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook. Management Science 64.

Lee, K.-P., Song, S., 2022. Developing Insights from the Collective Voice of Target Users in Twitter. Journal of Big Data 9, 75.

Loughran, T., McDonald, B., 2010. When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. Journal of Finance 66, 35–65.

Loughran, T., McDonald, B., 2020. Textual Analysis in Finance. Annual Review of Financial Economics 12, 357–375.

Malhotra, C. K., Malhotra, A., 2016. How CEOs Can Leverage Twitter. MIT Sloan Management Review 57, 72–79.

Malmendier, U., Tate, G., 2005. CEO Overconfidence and Corporate Investment. Journal of Finance 60, 2661–2700.

Men, L. R., Tsai, W.-H. S., 2016. Public Engagement with CEOs on Social Media: Motivations and Relational Outcomes. Public Relations Review 42, 932 – 942.

Miller, G., Skinner, D., 2015. The Evolving Disclosure Landscape: How Changes in Technology, the Media, and Capital Markets Are Affecting Disclosure. Journal of Accounting Research 53.

Newey, W. K., West, K. D., 1987. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica 55, 703–708.

Nguyen, T. H., Shirai, K., 2015. Topic modeling based sentiment analysis on social media for stock market prediction. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Beijing, China, pp. 1354–1364.

Pan, Y., Siegel, S., Wang, T. Y., 2019. The Cultural Origin of CEOs' Attitudes toward Uncertainty: Evidence from Corporate Acquisitions. Review of Financial Studies 33, 2977–3030.

Park, J. Y., Phillips, P. C. B., 2000. Nonstationary Binary Choice. Econometrica 68, 1249–1280.

Pesaran, M. H., Timmermann, A., 1995. Predictability of Stock Returns: Robustness and Economic Significance. Journal of Finance 50, 1201–1228.

Pesaran, M. H., Timmermann, A., 2000. A Recursive Modelling Approach to Predicting UK Stock Returns. Economic Journal 110, 159–191.

Porter, M. C., Anderson, B., Nhotsavang, M., 2015. Anti-social Media: Executive Twitter "Engagement" and Attitudes about Media Credibility. Journal of Communication Management 19, 270–287.

Shue, K., 2013. Executive Networks and Firm Policies: Evidence from the Random Assignment of MBA Peers. Review of Financial Studies 26, 1401–1442.

Stambaugh, R. F., 1999. Predictive Regressions. Journal of Financial Economics 54, 375–421.

Tetlock, P. C., 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. Journal of Finance 62, 1139–1168.

Tetlock, P. C., Saar-Tsechansky, M., Macskassy, S., 2008. More Than Words: Quantifying Language to Measure Firms' Fundamentals. Journal of Finance 63, 1437–1467.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58, 267–288.

Timmermann, A., 2018. Forecasting Methods in Finance. Annual Review of Financial Economics 10, 449–479.

Van Binsbergen, J. H., Koijen, R. S., 2010. Predictive Regressions: A Present-value Approach. Journal of Finance 65, 1439–1471.

Welch, I., Goyal, A., 2008. A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. Review of Financial Studies 21, 1455–1508.

Wolfskeil, I., 2023. Tweeting in the Dark: Corporate Communication and Information Diffusion. Social Science Research Network (SSRN) .

Xing, F. Z., Cambria, E., Welsch, R. E., 2018. Natural language based financial forecasting: a survey. Artificial Intelligence Review 50, 49–73.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society. Series B (Methodological) 68, 49–67.

Zou, H., 2006. The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association 101, 1418–1429.

Table 1: Monthly data statistics

This table reports the monthly statistics of the data collected for 36 months from January 2019 to December 2021 with respect to user count and tweet count.

Month	User Count	Tweet Count
12/2021	22,569,110	133,387,546
11/2021	21,876,935	129,462,997
10/2021	22,175,272	133,334,050
09/2021	21,446,941	127,009,377
08/2021	21,708,191	133,447,209
07/2021	21,979,242	133,358,039
06/2021	21,611,226	128,414,906
05/2021	22,651,068	133,741,215
04/2021	22,138,958	129,235,713
03/2021	22,441,309	133,544,952
02/2021	21,529,017	120,703,913
01/2021	22,317,570	133,754,300
12/2020	21,107,115	120,627,976
11/2020	21,950,691	129,635,445
10/2020	21,889,317	133,221,211
09/2020	22,344,474	128,867,950
08/2020	22,643,060	133,302,754
07/2020	22,930,209	133,609,303
06/2020	22,419,694	128,885,150
05/2020	23,554,291	133,389,857
04/2020	23,420,878	129,330,138
03/2020	23,400,803	133,474,640
02/2020	21,260,800	125,029,995
01/2020	22,275,681	133,666,464
12/2019	22,176,797	133,683,030
11/2019	21,905,451	129,514,576
10/2019	22,236,154	132,998,513
09/2019	22,070,159	129,350,458
08/2019	22,349,152	133,653,777
07/2019	22,218,823	133,510,690
06/2019	22,046,523	129,321,270
05/2019	22,474,111	133,517,533
04/2019	21,961,307	129,453,370
03/2019	22,276,114	133,600,069
02/2019	21,103,458	120,796,157
01/2019	22,060,698	133,048,524
Total (Unique)	176,850,729	4,704,883,067

Table 2: Top-20 CEO users sorted by follower count

The user screen names, follower counts, and descriptions of the top-20 users are presented, sorted by follower count.

Rank	User Screen Name	Followers	Description
1	elonmusk	72,497,051	Former CEO of Dogecoin
2	parishilton	16,879,195	Text PH to 833-240-3728 for updates Hear my new spinoff series #Dominated, hosted by @CindyPinkCEO on @ThisIsParisPod
3	therock	15,752,600	CEO Seven Bucks Companies
4	kobebryant	15,023,019	@Granity Studios - CEO, Writer, Producer
5	tim_cook	13,232,112	Apple CEO Auburn Duke National Parks "Life's most persistent and urgent question is, 'What are you doing for others?"" - MLK. he/him
6	50cent	12,535,470	CEO G-UNIT FILM & TV, SIRE SPIRITS, G-UNIT RECORDS & THE G-UNITY FOUNDATION, 2X NY Times Best Selling Author, Award Winning Director & Rapper
7	werevertumorro	8,687,262	CEO @Timbersesports y @ace1_gg Somos uno, para siempre. Negocios: contactowerever@gmail.com
8	jacksepticeye	7,499,889	Artist, Entrepreneur, CEO http://topofthemornincoffee.com
9	sudhirchaudhary	6,941,980	Editor In Chief & CEO , Zee News, WION, Zee Business. Hosts India's No.1 News Show DNA.Ramnath Goenka Awardee https://t.co/mR8AzsmCw2
10	jeffreestar	6,651,465	Self-Made. Makeup Magician. CEO. Owner of Jeffree Star Cosmetics Mom of 7 Pomeranians Living in Wyoming raising yaks & writing my autobiography.
11	johnlegere	5,995,821	Lover of all things slow cookers, golf, NFTs, Batman, and magenta! Ex-CEO @TMobile
12	billsimmons	5,674,187	@ringer (CEO) + The BS Podcast + @BookoBasketball 2.0 pod + @therewatchables ... Past Life: @grantland33 @30for30
13	souljaboy	5,479,229	CEO of #SouljaBoyApparel — Platinum Recording Artist — Bitcoin Investor #BTC — http://SouljaBoyApparel.com
14	cz_binance	5,100,429	#BNB, #bitcoin hodler CEO @binance
15	llcoolj	5,003,555	CEO of @rockthebells (917)540-5512
16	tonyhawk	4,555,521	Pro skater, husband, child rearer, videogame character, CEO, food/spirit glutton and public skatepark advocate. Old AF and still skating. Have ramp, will travel
17	sundarpichai	4,203,860	CEO, Google and Alphabet
18	peterpsquare	3,730,922	Musician/Entertainer/Businessman/Dr&CEO @zoomupyourlife @pclassicgroup @ziprepublic @theokoyes @aphrospirit Management info@one1mgt.com +2348121110000
19	nelly_mo	3,727,025	https://t.co/l8dkNt6I4H New single #FreakyWithYou NELLY's OFFICIAL TWITTER...CEO of Nelly Inc, Derrty Ent, NELLYVILLE on BET ...!!!
20	tommyinnit	3,557,222	CEO @realmeIndia & @realmeeurope

Table 3: Bottom-20 CEO users sorted by follower count

The user screen names, follower counts, and descriptions of the bottom-20 users are presented, sorted by follower count.

Rank	User Screen Name	Followers	Description
4695	secchambersin	400	Secretary of Commerce for the State of #Indiana and CEO of the Indiana Economic Development Corporation (@Indiana_EDC) #AStateThatWorks
4696	zacharymoses	399	CEO of https://t.co/Nkyju6l9jA, https://t.co/9USbdSC9a2, & https://t.co/EVrdclmpUi, Comedian, Travel Guru, former candidate for UT Governor
4697	swayanchaudhuri	396	Urban Enthusiast —— Trying to make use of Technology for the Common Good —— Formerly Mission Director & MD & CEO - Smart City & AMRUT, Goa
4698	fairtrade_ceo	359	Interim @Fairtrade Global CEO — Born and raised in Kenya to a farmer, I am dedicated to improving the livelihoods of small-scale farmers around the world.
4699	aumaldives	357	Entrepreneur Founder & CEO of @FaseyhaRecharge
4700	davidawilliams	346	President + CEO @GenesysWorks. Love to play #basketball and #tennis. New grandfather. #nonprofitleadership
4701	sbirdabrdn	273	CEO @abrdn_plc
4702	erikmunderwood	256	Tech Innovator. Colorado. CEO of My24. Dog lover, , buff, runner, GQ, & Netflix. @My24Erica.Com @KillerPolitics @illmaticv1 . All of my views are my own.
4703	goshantanu	256	IAS 2004 — CEO & Principal Secretary, BTC, Assam— Princeton— Passionate about Sustainability, Green Growth and Energy.
4704	mhedengren	253	CEO @Readly — Bringing the magic of magazines into the future
4705	jsrung	252	President/CEO of Shaw Media (Illinois and Iowa). Avid reader, lover of history, family guy, Bears fan, former triathlete.
4706	davidhillok	231	Husband. Father of 6. CEO. Business Owner. Manufacturer. Proud Oklahoman
4707	davidsilveroak	225	Proprietor and Chairman/CEO of @SilverOak, @Twomey, OVID Napa Valley and Timeless Napa Valley
4708	yiehsin	219	CEO for New York Life Investment Management, the asset management business of @NewYorkLife Insurance Company. Securities distributed by NYLIFE Distributors LLC.
4709	karenjhanrahan	214	Pres. & CEO of @GLIDEsf. Global human rights lawyer. Social innovator. Former @StateDept for Democracy + Security. Passionate about equity for all. Proud mom.
4710	kellycoffeycnb	207	CEO @CityNational
4711	kevin_makely	198	Actor / Producer / CEO Papa Octopus Productions https://t.co/gT2GR9NuDj https://t.co/KZ2CfpCgGE
4712	drvinaynair	196	A serial entrepreneur, investor and academic. Founder, CEO and Chairman of @TIFINfintech: @investpositivly • @DiscoverMagnifi • @content_clout • @TotumRisk
4713	brianpocrass	159	CEO/Founder/Producer, 22 Vision
4714	mihaelhp	51	Genetics and Medicine, CEO @vandapharma, tweets are my own ideas and opinions

Table 4: Top-100 popular hashtags and their user counts sorted by user count

The top-100 popular hashtags and their user counts are presented, sorted by user count.

Rank	Hashtag	Frequency	Rank	Hashtag	Frequency
1	#covid19	2,568	51	#proud	1,127
2	#tbt	2,113	52	#startups	1,113
3	#ff	2,079	53	#health	1,112
4	#coronavirus	1,835	54	#entrepreneur	1,105
5	#nyc	1,707	55	#google	1,091
6	#superbowl	1,681	56	#women	1,091
7	#internationalwomensday	1,677	57	#mothersday	1,089
8	#love	1,602	58	#china	1,084
9	#tech	1,524	59	#throwbackthursday	1,084
10	#rip	1,478	60	#art	1,066
11	#twitter	1,477	61	#brexit	1,066
12	#facebook	1,455	62	#awesome	1,065
13	#innovation	1,454	63	#marketing	1,046
14	#respect	1,453	64	#instagram	1,035
15	#leadership	1,445	65	#trump	1,032
16	#usa	1,442	66	#valentinesday	1,016
17	#winning	1,398	67	#blackfriday	1,012
18	#blacklivesmatter	1,391	68	#fb	1,011
19	#covid	1,384	69	#video	1,000
20	#business	1,376	70	#success	997
21	#fail	1,344	71	#legend	995
22	#truth	1,330	72	#life	991
23	#mondaymotivation	1,306	73	#iphone	987
24	#oscars	1,290	74	#followfriday	987
25	#podcast	1,280	75	#fathersday	985
26	#family	1,261	76	#nowplaying	985
27	#christmas	1,258	77	#merrychristmas	984
28	#london	1,258	78	#live	983
29	#newprofilepic	1,256	79	#bitcoin	983
30	#socialmedia	1,255	80	#climatechange	977
31	#music	1,247	81	#newyork	966
32	#breaking	1,237	82	#givingtuesday	964
33	#halloween	1,229	83	#selfie	958
34	#education	1,228	84	#happy	952
35	#sxsw	1,222	85	#justsaying	951
36	#worldcup	1,222	86	#nfl	950
37	#startup	1,219	87	#india	949
38	#olympics	1,215	88	#icymi	948
39	#thanksgiving	1,203	89	#news	947
40	#technology	1,198	90	#paris	946
41	#vote	1,197	91	#digital	945
42	#periscope	1,193	92	#amazing	945
43	#apple	1,190	93	#win	943
44	#inspiration	1,185	94	#travel	941
45	#neverforget	1,176	95	#youtube	939
46	#jobs	1,161	96	#work	921
47	#happynewyear	1,154	97	#metoo	917
48	#ai	1,146	98	#grateful	917
49	#blessed	1,139	99	#diversity	916
50	#thankyou	1,133	100	#earthday	916

Table 5: Top-100 popular hashtags and their tweet counts sorted by tweet count

The top-100 popular hashtags and their tweet counts, as opposed to user counts, are presented, sorted by tweet count.

Rank	Hashtag	Frequency	Rank	Hashtag	Frequency
1	#quote	121,410	51	#mobile	19,397
2	#ff	80,518	52	#music	18,609
3	#socialmedia	77,461	53	#israel	18,482
4	#marketing	77,064	54	#mondaymotivation	18,236
5	#startup	62,086	55	#instagram	17,872
6	#covid19	61,491	56	#justhaves	17,447
7	#leadership	61,055	57	#portraitdestartuper	17,351
8	#bitcoin	44,994	58	#startups	17,144
9	#entrepreneur	44,251	59	#quotes	17,045
10	#ai	41,975	60	#breaking	16,768
11	#tech	41,922	61	#nfl	16,599
12	#business	41,594	62	#education	16,364
13	#np	39,947	63	#celtics	16,273
14	#askfft	36,913	64	#salute	15,815
15	#sales	36,635	65	#quoteoftheday	15,813
16	#fashion	35,743	66	#digital	15,791
17	#travel	33,325	67	#fintech	15,653
18	#nowplaying	31,981	68	#cdnpoli	14,693
19	#success	30,972	69	#iran	14,649
20	#fb	28,467	70	#news	14,180
21	#tbt	27,939	71	#cleantech	14,171
22	#periscope	26,764	72	#entrepreneurship	14,152
23	#getrealchat	26,373	73	#ad	14,138
24	#cio	25,900	74	#yeg	14,121
25	#nyc	25,763	75	#nba	13,840
26	#tcot	25,525	76	#mufc	13,827
27	#syria	25,238	77	#video	13,799
28	#repost	25,114	78	#smm	13,721
29	#contentmarketing	24,123	79	#maga	13,632
30	#twitter	23,803	80	#ux	13,501
31	#blockchain	23,323	81	#crypto	13,498
32	#facebook	23,228	82	#china	13,489
33	#inspiration	23,176	83	#artdesign	13,420
34	#motivation	23,067	84	#london	13,091
35	#healthcare	22,293	85	#coast2coast	13,040
36	#soundcloud	22,254	86	#fail	12,909
37	#lifestyle	22,221	87	#entrepreneurs	12,816
38	#innovation	21,956	88	#stem	12,658
39	#cloud	21,837	89	#jobs	12,651
40	#iot	21,670	90	#climate	12,646
41	#rt	21,369	91	#india	12,586
42	#trump	21,252	92	#women	12,525
43	#sxsw	21,167	93	#superbowl	12,405
44	#egypt	20,918	94	#technology	12,240
45	#podcast	20,811	95	#truth	11,984
46	#hyc	20,194	96	#brexit	11,928
47	#seo	19,928	97	#txlege	11,910
48	#love	19,826	98	#wednesdaywisdom	11,898
49	#coronavirus	19,512	99	#oscars	11,849
50	#mambomseto	19,507	100	#health	11,833

Table 6: List of 58 clusters of similar hashtags among the top-1,000 popular hashtags

All the 58 clusters of similar hashtags found among the top-1,000 popular hashtags are listed.

#wfh, #covid, #covid19, #covid2019, #covid_19, #socialdistancing
#innovation, #innovative
#winning, #win
#oscar, #oscars
#podcasts, #podcast
#london, #london2012
#worldcup2014, #worldcup
#startups, #startup
#olympic, #olympics
#thanksgiving, #blackfriday, #happythanksgiving
#inspirational, #inspiration, #inspire, #inspired, #inspiring
#jobs, #job
#happynewyear, #newyearseve
#blessed, #blessings
#entrepreneur, #entrepreneurs
#mothersday, #happymothersday
#valentinesday, #happyvalentinesday
#legend, #legends
#fathersday, #happyfathersday
#happy, #happiness
#facts, #fact
#debate, #debates, #debates2020
#quotes, #quote
#thanks, #thankful
#beauty, #beautiful
#power, #powerful
#goodfriday, #easter, #happyeaster
#quarantine, #stayathome, #stayhome
#sports, #sport
#sustainability, #sustainable
#election, #election2020, #election2016
#hero, #heroes
#books, #book
#creativity, #creative
#leader, #leaders
#hashtags, #hashtag
#holidays, #holiday
#brand, #branding, #brands
#apps, #app
#nft, #nfts
#learn, #learning
#tokyo, #tokyo2020
#icebucketchallenge, #alsicebucketchallenge
#gaming, #game
#dream, #dreams
#iwd2016, #iwd2021, #iwd2020, #iwd2019, #iwd2018, #iwd2017, #iwd
#investment, #investing
#movie, #movies
#dogs, #dog
#trending, #trends
#parents, #parenting
#champions, #champion
#drones, #drone
#positive, #positivity
#robotics, #robots
#euro2020, #euro2016
#mandela, #nelsonmandela
#event, #events

Table 7: The parameter settings that yield the best weekly forecasting performances

This table reports the parameter settings that yield the best weekly forecasting performances, i.e., the lowest Mean Squared Errors, for the weekly frequency, four markets, three targets, three feature types, nine model sizes, and five horizons.

Frequency	Market	Target	Feature Type	Model Size	Horizon	MSE
Weekly	SP500	Close	Hashtags (Lasso)	0.9	2	0.767
Weekly	SP500	Close	Hashtags (AGLasso)	0.9	2	0.836
Weekly	SP500	Close	Sentiments	0.9	4	0.769
Weekly	SP500	Volume	Hashtags (Lasso)	0.2	1	0.841
Weekly	SP500	Volume	Hashtags (AGLasso)	0.1	1	0.7
Weekly	SP500	Volume	Sentiments	0.2	1	0.548
Weekly	SP500	Volatility	Hashtags (Lasso)	0.2	1	0.938
Weekly	SP500	Volatility	Hashtags (AGLasso)	0.1	2	0.842
Weekly	SP500	Volatility	Sentiments	0.2	1	0.624
Weekly	Dow30	Close	Hashtags (Lasso)	0.9	2	0.746
Weekly	Dow30	Close	Hashtags (AGLasso)	0.9	2	0.843
Weekly	Dow30	Close	Sentiments	0.9	4	0.784
Weekly	Dow30	Volume	Hashtags (Lasso)	0.9	2	0.892
Weekly	Dow30	Volume	Hashtags (AGLasso)	0.9	5	0.809
Weekly	Dow30	Volume	Sentiments	0.9	1	0.619
Weekly	Dow30	Volatility	Hashtags (Lasso)	0.2	1	0.897
Weekly	Dow30	Volatility	Hashtags (AGLasso)	0.1	1	0.861
Weekly	Dow30	Volatility	Sentiments	0.2	1	0.676
Weekly	Nasdaq	Close	Hashtags (Lasso)	0.9	5	1.055
Weekly	Nasdaq	Close	Hashtags (AGLasso)	0.3	1	1.003
Weekly	Nasdaq	Close	Sentiments	0.2	5	0.953
Weekly	Nasdaq	Volume	Hashtags (Lasso)	0.9	2	0.919
Weekly	Nasdaq	Volume	Hashtags (AGLasso)	0.3	2	0.899
Weekly	Nasdaq	Volume	Sentiments	0.9	1	0.73
Weekly	Nasdaq	Volatility	Hashtags (Lasso)	0.7	1	1.09
Weekly	Nasdaq	Volatility	Hashtags (AGLasso)	0.1	4	1.056
Weekly	Nasdaq	Volatility	Sentiments	0.7	1	0.966
Weekly	Russell2000	Close	Hashtags (Lasso)	0.9	2	1.032
Weekly	Russell2000	Close	Hashtags (AGLasso)	0.2	1	0.939
Weekly	Russell2000	Close	Sentiments	0.2	5	0.98
Weekly	Russell2000	Volume	Hashtags (Lasso)	0.1	1	0.813
Weekly	Russell2000	Volume	Hashtags (AGLasso)	0.2	1	0.785
Weekly	Russell2000	Volume	Sentiments	0.2	1	0.55
Weekly	Russell2000	Volatility	Hashtags (Lasso)	0.1	1	0.941
Weekly	Russell2000	Volatility	Hashtags (AGLasso)	0.1	1	0.637
Weekly	Russell2000	Volatility	Sentiments	0.1	1	0.564

Table 8: The parameter settings that yield the best daily forecasting performances

This table reports the parameter settings that yield the best daily forecasting performances, i.e., the lowest Mean Squared Errors, for the daily frequency, four markets, three targets, three feature types, nine model sizes, and five horizons.

Frequency	Market	Target	Feature Type	Model Size	Horizon	MSE
Daily	SP500	Close	Hashtags (Lasso)	0.9	5	0.701
Daily	SP500	Close	Hashtags (AGLasso)	0.9	3	0.658
Daily	SP500	Close	Sentiments	0.9	1	0.642
Daily	SP500	Volume	Hashtags (Lasso)	0.3	1	0.611
Daily	SP500	Volume	Hashtags (AGLasso)	0.1	3	0.982
Daily	SP500	Volume	Sentiments	0.2	1	0.472
Daily	SP500	Volatility	Hashtags (Lasso)	0.4	1	0.961
Daily	SP500	Volatility	Hashtags (AGLasso)	0.1	4	1.128
Daily	SP500	Volatility	Sentiments	0.1	1	0.533
Daily	Dow30	Close	Hashtags (Lasso)	0.9	4	0.68
Daily	Dow30	Close	Hashtags (AGLasso)	0.9	2	0.652
Daily	Dow30	Close	Sentiments	0.9	5	0.621
Daily	Dow30	Volume	Hashtags (Lasso)	0.9	3	0.456
Daily	Dow30	Volume	Hashtags (AGLasso)	0.6	2	0.387
Daily	Dow30	Volume	Sentiments	0.6	5	0.321
Daily	Dow30	Volatility	Hashtags (Lasso)	0.9	1	0.884
Daily	Dow30	Volatility	Hashtags (AGLasso)	0.1	2	1.149
Daily	Dow30	Volatility	Sentiments	0.1	1	0.534
Daily	Nasdaq	Close	Hashtags (Lasso)	0.9	5	0.937
Daily	Nasdaq	Close	Hashtags (AGLasso)	0.9	5	0.92
Daily	Nasdaq	Close	Sentiments	0.9	1	0.909
Daily	Nasdaq	Volume	Hashtags (Lasso)	0.9	3	0.504
Daily	Nasdaq	Volume	Hashtags (AGLasso)	0.9	2	0.416
Daily	Nasdaq	Volume	Sentiments	0.9	4	0.352
Daily	Nasdaq	Volatility	Hashtags (Lasso)	0.3	1	0.729
Daily	Nasdaq	Volatility	Hashtags (AGLasso)	0.1	5	0.765
Daily	Nasdaq	Volatility	Sentiments	0.1	1	0.447
Daily	Russell2000	Close	Hashtags (Lasso)	0.9	4	0.995
Daily	Russell2000	Close	Hashtags (AGLasso)	0.9	4	0.976
Daily	Russell2000	Close	Sentiments	0.2	5	0.87
Daily	Russell2000	Volume	Hashtags (Lasso)	0.3	1	0.658
Daily	Russell2000	Volume	Hashtags (AGLasso)	0.1	3	1.0
Daily	Russell2000	Volume	Sentiments	0.2	1	0.489
Daily	Russell2000	Volatility	Hashtags (Lasso)	0.3	1	0.822
Daily	Russell2000	Volatility	Hashtags (AGLasso)	0.1	4	0.854
Daily	Russell2000	Volatility	Sentiments	0.1	1	0.523

Table 9: Top-5 most significant hashtag features from weekly forecasting

This table summarizes the set of top-5 hashtag features from lasso that are statistically significant at 10% significance level and reports the estimated coefficients and Newey–West heteroskedasticity– and autocorrelation–robust standard errors inside parentheses.

Target	Market	Significant Features
Close	SP500	#iwd2020 -0.245(0.016) #pandemic 0.191(0.064) #womensday -0.076(0.043)
Close	Dow30	#iwd2020 -0.278(0.017) #pandemic 0.240(0.072) #womensday -0.077(0.039)
Close	Nasdaq	#corona 1.109(0.421) #covid -0.646(0.304) #stayhome 0.455(0.269) #covid_19 0.367(0.218) #breonnataylor 0.257(0.058)
Close	Russell2000	#iwd2020 -0.185(0.030)
Volume	SP500	#coronavirus 0.291(0.089) #clubhouse 0.250(0.043) #holiday -0.221(0.058) #iwd2020 0.196(0.034) #finalfour -0.184(0.076)
Volume	Dow30	#merrychristmas 0.175(0.062) #random 0.133(0.031) #cybersecurity 0.128(0.038) #happyvalentinesday 0.123(0.040) #sorrynotsorry 0.121(0.037)
Volume	Nasdaq	#merrychristmas 0.191(0.082) #happyvalentinesday 0.164(0.032) #veteransday -0.137(0.020) #game 0.122(0.033) #vmas 0.118(0.022)
Volume	Russell2000	#coronavirus 0.341(0.112) #followfriday 0.187(0.038) #clubhouse 0.170(0.045) #holiday -0.143(0.058) #iwd2020 0.141(0.031)
Volatility	SP500	SP500_Volatility_L1 0.285(0.058) #coronavirus 0.229(0.131) #georgefloyd 0.120(0.018) #safety 0.092(0.032) #car -0.086(0.025)
Volatility	Dow30	#boston -0.757(0.090) #resist 0.646(0.079) #valentinesday 0.626(0.111) #inauguration -0.609(0.128) #nelsonmandela 0.580(0.105)
Volatility	Nasdaq	
Volatility	Russell2000	Russell2000_Volatility_L1 0.225(0.060) #iwd2020 0.136(0.018) #georgefloyd 0.109(0.028) #stem 0.093(0.028) #blackhistorymonth 0.090(0.045)

Table 10: The parameter settings that yield the best prediction performances

This table reports the parameter settings that yield the best weekly prediction performances, i.e., the highest accuracy, for each of the two frequencies, four markets, two feature types, nine model sizes, five horizons, and five classification algorithms.

Frequency	Market	Target	Feature Type	Model Size	Horizon	Algorithm	Accuracy
Weekly	SP500	Direction	Hashtags	0.9	3	Logit	0.641
Weekly	SP500	Direction	Sentiments	0.9	3	Logit	0.625
Weekly	Dow30	Direction	Hashtags	0.9	3	DTrees	0.703
Weekly	Dow30	Direction	Sentiments	0.9	5	Logit	0.645
Weekly	Nasdaq	Direction	Hashtags	0.9	5	k-NNs	0.581
Weekly	Nasdaq	Direction	Sentiments	0.9	5	SVC	0.597
Weekly	Russell2000	Direction	Hashtags	0.9	2	Logit	0.615
Weekly	Russell2000	Direction	Sentiments	0.9	3	DeepNeural	0.641
Daily	SP500	Direction	Hashtags	0.9	2	SVC	0.57
Daily	SP500	Direction	Sentiments	0.9	2	SVC	0.564
Daily	Dow30	Direction	Hashtags	0.9	3	DTrees	0.572
Daily	Dow30	Direction	Hashtags	0.9	3	Logit	0.572
Daily	Dow30	Direction	Sentiments	0.9	3	DeepNeural	0.572
Daily	Dow30	Direction	Sentiments	0.9	3	Logit	0.572
Daily	Nasdaq	Direction	Hashtags	0.8	1	Logit	0.583
Daily	Nasdaq	Direction	Sentiments	0.9	1	Logit	0.587
Daily	Russell2000	Direction	Hashtags	0.9	4	SVC	0.554
Daily	Russell2000	Direction	Sentiments	0.9	1	k-NNs	0.613

Table 11: Top-5 most important hashtag features from weekly forecasting

This table summarizes the set of the most important hashtag features along with their importance scores.

Target	Market	Algorithm	Significant Features
Direction	SP500	Logit	#ff -0.0034 #brexit 0.0024 #internationalwomensday -0.002 #superbowl 0.0018 #fb -0.0017
Direction	Dow30	DTrees	#good 0.0322 #siliconvalley 0.0307 #culture 0.0298 #newprofilepic 0.0296 #cheers 0.0273
Direction	Nasdaq	SVC	#coronavirus 0.0101 #superbowl 0.009 #syria 0.009 #sxsw 0.0078 #grammys 0.0075
Direction	Russell2000	SVC	#ai -0.0207 #blockchain -0.0181 #covid19 0.0145 #fintech -0.0123 #iot -0.0083

Table 12: Significant sentiment regressors from monthly forecasting

This table summarizes the set of sentiment features which are statistically significant at 10% significance level and reports the estimated coefficients and Newey–West heteroskedasticity– and autocorrelation–robust standard errors inside parentheses.

Target	Market	Significant Sentiment Features
Excess Return	SP500	Constraining 0.365(0.133) Litigious 0.209(0.107)
Close	SP500	Constraining 0.378(0.122) StrongModal 0.223(0.130) Litigious 0.205(0.106)
Close	Dow30	Constraining 0.303(0.130) Litigious 0.218(0.107)
Close	Nasdaq	Constraining 0.380(0.128) StrongModal 0.311(0.163)
Close	Russell2000	Constraining 0.219(0.124) Litigious 0.201(0.110)
Volume	SP500	Positive -0.920(0.226) StrongModal 0.899(0.231)
Volume	Dow30	Negative -0.336(0.198)
Volume	Nasdaq	WeakModal 0.808(0.423) Positive -0.350(0.151) StrongModal 0.322(0.157) Constraining -0.286(0.109)
Volume	Russell2000	Positive -0.919(0.227) StrongModal 0.894(0.231)
Volatility	SP500	Uncertainty 0.969(0.577) WeakModal -0.704(0.410)
Volatility	Dow30	StrongModal 1.185(0.186) Positive -1.125(0.208) Negative -0.264(0.141)
Volatility	Nasdaq	–
Volatility	Russell2000	Constraining -0.258(0.106)

Table 13: Estimation results for sentiments, monthly frequency, and SP500 Excess Return

We estimate control-augmented models for monthly Excess Return by adding one of the macroeconomic and financial variables as a control variable and report the estimated coefficients and Newey–West heteroskedasticity– and autocorrelation–robust standard errors inside parentheses.

	Univariate		Multivariate			
	Coef. (S.E.)	Adj. R^2	Coef. (S.E.)	Coef. (S.E.) : Constraining	Coef. (S.E.) : Litigious	Adj. R^2
d/p	-0.111 (0.078)	0.006	-0.042 (0.101)	0.176 (0.103)	0.130 (0.082)	0.061
d/y	-0.105 (0.090)	0.004	-0.112 (0.080)	0.175 (0.081)	0.145 (0.081)	0.072
e/p	-0.061 (0.078)	-0.003	-0.056 (0.078)	0.195 (0.089)	0.121 (0.080)	0.063
d/e	0.146 (0.056)	0.014	0.132 (0.055)	0.197 (0.087)	0.111 (0.077)	0.077
svar	0.235 (0.051)	0.049	0.156 (0.069)	0.100 (0.096)	0.146 (0.084)	0.078
b/m	0.085 (0.069)	0.000	0.001 (0.059)	0.191 (0.089)	0.126 (0.085)	0.060
ntis	-0.115 (0.017)	0.006	-0.102 (0.016)	0.191 (0.089)	0.120 (0.082)	0.070
tbl	-0.114 (0.068)	0.006	-0.084 (0.055)	0.175 (0.087)	0.132 (0.081)	0.067
lty	-0.120 (0.085)	0.008	-0.102 (0.067)	0.174 (0.079)	0.136 (0.080)	0.070
ltr	0.063 (0.091)	-0.003	0.044 (0.092)	0.186 (0.086)	0.128 (0.080)	0.062
tms	-0.224 (0.029)	0.044	-0.139 (0.051)	0.106 (0.099)	0.145 (0.085)	0.074
dfy	0.090 (0.120)	0.001	0.057 (0.086)	0.176 (0.080)	0.133 (0.083)	0.063
dfr	0.036 (0.102)	-0.005	0.156 (0.079)	0.257 (0.115)	0.116 (0.080)	0.081
infl	-0.062 (0.071)	-0.003	-0.107 (0.068)	0.188 (0.086)	0.149 (0.087)	0.071

Table 14: Estimation results for sentiments, monthly frequency, and SP500 Close

We estimate control-augmented models for monthly SP500 Close by adding one of the macroeconomic and financial variables as a control variable and report the estimated coefficients and Newey–West heteroskedasticity– and autocorrelation–robust standard errors inside parentheses.

	Univariate		Multivariate				
	Coef. (S.E.)	Adj. R^2	Coef. (S.E.)	Coef. (S.E.) : Constraining	Coef. (S.E.) : StrongModal	Coef. (S.E.) : Litigious	Adj. R^2
d/p	-0.094 (0.078)	0.002	-0.025 (0.101)	0.166 (0.097)	0.071 (0.078)	0.112 (0.084)	0.054
d/y	-0.092 (0.088)	0.002	-0.092 (0.075)	0.165 (0.077)	0.062 (0.076)	0.126 (0.082)	0.062
e/p	-0.083 (0.063)	0.000	-0.095 (0.064)	0.177 (0.081)	0.087 (0.072)	0.101 (0.081)	0.063
d/e	0.123 (0.055)	0.008	0.140 (0.057)	0.170 (0.081)	0.111 (0.073)	0.093 (0.080)	0.072
svar	0.221 (0.055)	0.042	0.126 (0.070)	0.108 (0.096)	0.047 (0.072)	0.127 (0.086)	0.065
b/m	0.074 (0.066)	-0.001	-0.004 (0.057)	0.176 (0.081)	0.070 (0.077)	0.111 (0.087)	0.054
ntis	-0.111 (0.020)	0.006	-0.105 (0.017)	0.173 (0.081)	0.079 (0.077)	0.104 (0.084)	0.065
tbl	-0.122 (0.068)	0.008	-0.088 (0.052)	0.161 (0.080)	0.063 (0.076)	0.117 (0.082)	0.062
lty	-0.090 (0.092)	0.001	-0.073 (0.069)	0.163 (0.077)	0.072 (0.077)	0.118 (0.083)	0.059
ltr	0.047 (0.094)	-0.005	0.031 (0.095)	0.172 (0.080)	0.071 (0.078)	0.111 (0.082)	0.055
tms	-0.216 (0.029)	0.040	-0.119 (0.051)	0.106 (0.100)	0.057 (0.074)	0.127 (0.088)	0.064
dfy	0.088 (0.116)	0.001	0.049 (0.078)	0.165 (0.078)	0.066 (0.075)	0.117 (0.084)	0.056
dfr	0.044 (0.101)	-0.005	0.166 (0.079)	0.246 (0.105)	0.069 (0.072)	0.100 (0.080)	0.078
infl	-0.005 (0.072)	-0.007	-0.057 (0.072)	0.171 (0.079)	0.078 (0.077)	0.122 (0.087)	0.057

Table 15: Estimation results for sentiments, monthly frequency, and Dow30 Close

We estimate control-augmented models for monthly Dow30 Close by adding one of the macroeconomic and financial variables as a control variable and report the estimated coefficients and Newey–West heteroskedasticity– and autocorrelation–robust standard errors inside parentheses.

	Univariate		Multivariate			
	Coef. (S.E.)	Adj. R^2	Coef. (S.E.)	Coef. (S.E.) : Constraining	Coef. (S.E.) : Litigious	Adj. R^2
d/p	-0.082 (0.079)	0.000	-0.019 (0.103)	0.151 (0.099)	0.151 (0.088)	0.054
d/y	-0.106 (0.087)	0.005	-0.116 (0.076)	0.142 (0.076)	0.170 (0.084)	0.068
e/p	-0.101 (0.064)	0.004	-0.095 (0.062)	0.164 (0.084)	0.140 (0.084)	0.063
d/e	0.151 (0.051)	0.016	0.135 (0.049)	0.164 (0.083)	0.134 (0.083)	0.073
svar	0.198 (0.044)	0.033	0.122 (0.060)	0.086 (0.092)	0.166 (0.089)	0.065
b/m	0.076 (0.063)	-0.001	-0.008 (0.059)	0.158 (0.084)	0.152 (0.091)	0.054
ntis	-0.139 (0.028)	0.013	-0.126 (0.022)	0.158 (0.083)	0.142 (0.087)	0.070
tbl	-0.128 (0.059)	0.010	-0.103 (0.050)	0.139 (0.081)	0.157 (0.085)	0.065
lty	-0.076 (0.089)	-0.001	-0.061 (0.075)	0.147 (0.077)	0.156 (0.085)	0.058
ltr	0.024 (0.096)	-0.006	0.007 (0.098)	0.157 (0.083)	0.150 (0.085)	0.054
tms	-0.186 (0.027)	0.028	-0.103 (0.050)	0.095 (0.096)	0.164 (0.089)	0.062
dfy	0.090 (0.106)	0.001	0.063 (0.080)	0.142 (0.079)	0.158 (0.088)	0.058
dfr	0.045 (0.091)	-0.005	0.157 (0.079)	0.224 (0.113)	0.139 (0.085)	0.075
infl	-0.035 (0.074)	-0.006	-0.081 (0.075)	0.155 (0.082)	0.167 (0.094)	0.061

Table 16: Estimation results for sentiments, monthly frequency, and Nasdaq Close

We estimate control-augmented models for monthly Nasdaq Close by adding one of the macroeconomic and financial variables as a control variable and report the estimated coefficients and Newey–West heteroskedasticity– and autocorrelation–robust standard errors inside parentheses.

	Univariate		Multivariate			
	Coef. (S.E.)	Adj. R^2	Coef. (S.E.)	Coef. (S.E.) : Constraining	Coef. (S.E.) : StrongModal	Adj. R^2
d/p	-0.086 (0.075)	0.001	-0.015 (0.093)	0.213 (0.080)	0.077 (0.091)	0.044
d/y	-0.075 (0.086)	-0.001	-0.060 (0.074)	0.216 (0.068)	0.072 (0.089)	0.047
e/p	-0.072 (0.059)	-0.002	-0.091 (0.064)	0.214 (0.070)	0.093 (0.088)	0.052
d/e	0.116 (0.058)	0.007	0.143 (0.064)	0.202 (0.069)	0.118 (0.089)	0.063
svar	0.234 (0.055)	0.049	0.139 (0.070)	0.153 (0.080)	0.052 (0.081)	0.058
b/m	0.069 (0.055)	-0.002	0.016 (0.046)	0.213 (0.073)	0.077 (0.090)	0.044
ntis	-0.062 (0.011)	-0.003	-0.060 (0.016)	0.214 (0.070)	0.082 (0.091)	0.048
tbl	-0.126 (0.073)	0.009	-0.089 (0.058)	0.206 (0.068)	0.070 (0.089)	0.052
lty	-0.145 (0.090)	0.014	-0.122 (0.067)	0.203 (0.064)	0.080 (0.089)	0.059
ltr	0.095 (0.087)	0.002	0.079 (0.089)	0.210 (0.068)	0.080 (0.092)	0.050
tms	-0.239 (0.019)	0.051	-0.146 (0.040)	0.144 (0.079)	0.061 (0.086)	0.059
dfy	0.098 (0.117)	0.003	0.052 (0.077)	0.209 (0.066)	0.072 (0.088)	0.046
dfr	0.035 (0.106)	-0.006	0.152 (0.080)	0.277 (0.089)	0.076 (0.086)	0.064
infl	-0.041 (0.073)	-0.005	-0.075 (0.071)	0.221 (0.069)	0.088 (0.091)	0.049

Table 17: Estimation results for sentiments, monthly frequency, and Russell2000 Close

We estimate control-augmented models for monthly Russell2000 Close by adding one of the macroeconomic and financial variables as a control variable and report the estimated coefficients and Newey–West heteroskedasticity– and autocorrelation–robust standard errors inside parentheses.

	Univariate		Multivariate			
	Coef. (S.E.)	Adj. R^2	Coef. (S.E.)	Coef. (S.E.) : Constraining	Coef. (S.E.) : Litigious	Adj. R^2
d/p	-0.047 (0.087)	-0.005	0.012 (0.111)	0.139 (0.088)	0.139 (0.079)	0.040
d/y	-0.049 (0.080)	-0.004	-0.058 (0.073)	0.127 (0.067)	0.151 (0.082)	0.043
e/p	-0.075 (0.079)	-0.001	-0.069 (0.078)	0.140 (0.069)	0.134 (0.079)	0.044
d/e	0.124 (0.072)	0.009	0.109 (0.069)	0.140 (0.068)	0.128 (0.078)	0.052
svar	0.172 (0.053)	0.023	0.104 (0.070)	0.074 (0.080)	0.154 (0.081)	0.048
b/m	0.089 (0.048)	0.001	0.015 (0.065)	0.134 (0.068)	0.137 (0.085)	0.040
ntis	-0.055 (0.029)	-0.004	-0.044 (0.019)	0.135 (0.069)	0.138 (0.080)	0.041
tbl	-0.162 (0.058)	0.020	-0.142 (0.049)	0.109 (0.065)	0.150 (0.080)	0.060
lty	-0.037 (0.092)	-0.005	-0.024 (0.083)	0.131 (0.067)	0.143 (0.079)	0.040
ltr	0.004 (0.090)	-0.007	-0.010 (0.093)	0.136 (0.070)	0.140 (0.079)	0.040
tms	-0.165 (0.028)	0.021	-0.093 (0.049)	0.079 (0.083)	0.153 (0.082)	0.046
dfy	0.045 (0.109)	-0.005	0.020 (0.083)	0.130 (0.067)	0.143 (0.082)	0.040
dfr	0.042 (0.087)	-0.005	0.141 (0.086)	0.195 (0.097)	0.131 (0.077)	0.057
infl	-0.067 (0.068)	-0.002	-0.110 (0.068)	0.132 (0.067)	0.164 (0.086)	0.051

Table 18: Estimation results for sentiments, monthly frequency, and SP500 Volume

We estimate control-augmented models for monthly SP500 Volume by adding one of the macroeconomic and financial variables as a control variable and report the estimated coefficients and Newey–West heteroskedasticity– and autocorrelation–robust standard errors inside parentheses.

	Univariate		Multivariate			
	Coef. (S.E.)	Adj. R^2	Coef. (S.E.)	Coef. (S.E.) : Positive	Coef. (S.E.) : StrongModal	Adj. R^2
d/p	-0.020 (0.142)	-0.006	-0.043 (0.109)	-0.901 (0.227)	0.878 (0.224)	0.174
d/y	0.056 (0.107)	-0.003	0.034 (0.080)	-0.889 (0.226)	0.874 (0.229)	0.173
e/p	-0.336 (0.089)	0.110	-0.201 (0.127)	-0.688 (0.259)	0.728 (0.248)	0.203
d/e	0.375 (0.074)	0.138	0.302 (0.090)	-0.652 (0.222)	0.741 (0.206)	0.247
svar	0.338 (0.072)	0.111	0.219 (0.103)	-0.743 (0.227)	0.673 (0.234)	0.212
b/m	0.213 (0.078)	0.040	0.203 (0.056)	-0.887 (0.218)	0.863 (0.214)	0.215
ntis	-0.051 (0.037)	-0.004	-0.034 (0.037)	-0.891 (0.227)	0.875 (0.225)	0.173
tbl	-0.077 (0.081)	-0.001	0.008 (0.046)	-0.898 (0.227)	0.879 (0.226)	0.172
lty	-0.123 (0.149)	0.009	-0.071 (0.092)	-0.876 (0.212)	0.859 (0.207)	0.177
ltr	0.103 (0.119)	0.004	0.078 (0.094)	-0.884 (0.219)	0.868 (0.218)	0.178
tms	-0.220 (0.019)	0.043	-0.110 (0.043)	-0.838 (0.216)	0.797 (0.215)	0.183
dfy	0.052 (0.142)	-0.004	0.012 (0.091)	-0.893 (0.227)	0.873 (0.223)	0.172
dfr	-0.056 (0.135)	-0.004	-0.038 (0.091)	-0.893 (0.223)	0.870 (0.218)	0.173
infl	-0.058 (0.073)	-0.003	-0.086 (0.058)	-0.900 (0.219)	0.894 (0.211)	0.179

Table 19: Estimation results for sentiments, monthly frequency, and Dow30 Volume

We estimate control-augmented models for monthly Dow30 Volume by adding one of the macroeconomic and financial variables as a control variable and report the estimated coefficients and Newey–West heteroskedasticity– and autocorrelation–robust standard errors inside parentheses.

	Univariate		Multivariate		
	Coef. (S.E.)	Adj. R^2	Coef. (S.E.)	Coef. (S.E.) : Negative	Adj. R^2
d/p	0.048 (0.076)	-0.004	-0.014 (0.103)	-0.258 (0.042)	0.053
d/y	0.052 (0.063)	-0.004	0.044 (0.056)	-0.253 (0.046)	0.054
e/p	-0.017 (0.051)	-0.006	-0.020 (0.048)	-0.255 (0.043)	0.053
d/e	-0.008 (0.047)	-0.007	0.004 (0.048)	-0.255 (0.044)	0.052
svar	-0.203 (0.067)	0.035	-0.096 (0.077)	-0.205 (0.056)	0.059
b/m	-0.065 (0.057)	-0.002	0.018 (0.066)	-0.261 (0.040)	0.053
ntis	0.069 (0.034)	-0.002	0.062 (0.041)	-0.253 (0.043)	0.056
tbl	0.049 (0.041)	-0.004	0.028 (0.035)	-0.253 (0.044)	0.053
lty	0.191 (0.123)	0.030	0.148 (0.131)	-0.227 (0.048)	0.074
ltr	-0.174 (0.123)	0.024	-0.143 (0.129)	-0.236 (0.040)	0.073
tms	0.104 (0.027)	0.004	-0.050 (0.032)	-0.282 (0.052)	0.054
dfy	-0.122 (0.061)	0.008	-0.075 (0.051)	-0.240 (0.053)	0.058
dfr	0.089 (0.067)	0.001	-0.001 (0.090)	-0.255 (0.044)	0.052
infl	-0.031 (0.068)	-0.006	-0.008 (0.062)	-0.254 (0.040)	0.053

Table 20: Estimation results for sentiments, monthly frequency, and Nasdaq Volume

We estimate control-augmented models for monthly Nasdaq Volume by adding one of the macroeconomic and financial variables as a control variable and report the estimated coefficients and Newey–West heteroskedasticity– and autocorrelation–robust standard errors inside parentheses.

	Univariate		Multivariate					
	Coef. (S.E.)	Adj. R^2	Coef. (S.E.)	Coef. (S.E.) : WeakModal	Coef. (S.E.) : Positive	Coef. (S.E.) : StrongModal	Coef. (S.E.) : Constraining	Adj. R^2
d/p	0.058 (0.083)	-0.003	-0.059 (0.109)	0.163 (0.090)	-0.321 (0.158)	0.289 (0.160)	-0.439 (0.103)	0.091
d/y	0.084 (0.061)	0.000	0.055 (0.064)	0.154 (0.089)	-0.300 (0.153)	0.273 (0.157)	-0.412 (0.085)	0.091
e/p	-0.027 (0.063)	-0.006	0.052 (0.073)	0.156 (0.090)	-0.361 (0.173)	0.315 (0.157)	-0.415 (0.087)	0.089
d/e	-0.004 (0.050)	-0.007	-0.057 (0.060)	0.156 (0.090)	-0.353 (0.158)	0.300 (0.152)	-0.411 (0.087)	0.090
svar	-0.181 (0.075)	0.026	-0.134 (0.085)	0.173 (0.092)	-0.391 (0.150)	0.373 (0.158)	-0.368 (0.088)	0.099
b/m	-0.053 (0.054)	-0.004	-0.004 (0.067)	0.160 (0.091)	-0.308 (0.148)	0.276 (0.154)	-0.416 (0.086)	0.087
ntis	0.101 (0.044)	0.003	0.103 (0.055)	0.165 (0.089)	-0.319 (0.148)	0.277 (0.152)	-0.416 (0.086)	0.098
tbl	0.031 (0.048)	-0.006	0.019 (0.051)	0.160 (0.090)	-0.315 (0.151)	0.284 (0.156)	-0.415 (0.085)	0.088
lty	0.083 (0.084)	0.000	0.087 (0.077)	0.177 (0.092)	-0.330 (0.146)	0.292 (0.152)	-0.420 (0.084)	0.095
ltr	-0.096 (0.083)	0.002	-0.092 (0.081)	0.172 (0.090)	-0.319 (0.144)	0.281 (0.151)	-0.417 (0.083)	0.096
tms	0.095 (0.021)	0.002	-0.029 (0.045)	0.155 (0.090)	-0.296 (0.146)	0.263 (0.152)	-0.428 (0.095)	0.088
dfy	-0.116 (0.050)	0.007	-0.082 (0.052)	0.165 (0.091)	-0.318 (0.159)	0.292 (0.165)	-0.408 (0.085)	0.094
dfr	0.175 (0.080)	0.024	0.076 (0.103)	0.164 (0.090)	-0.306 (0.152)	0.274 (0.155)	-0.390 (0.088)	0.093
infl	-0.058 (0.054)	-0.003	-0.042 (0.054)	0.160 (0.089)	-0.311 (0.144)	0.285 (0.145)	-0.416 (0.085)	0.089

Table 21: Estimation results for sentiments, monthly frequency, and Russell2000 Volume

We estimate control-augmented models for monthly Russell2000 Volume by adding one of the macroeconomic and financial variables as a control variable and report the estimated coefficients and Newey–West heteroskedasticity– and autocorrelation–robust standard errors inside parentheses.

	Univariate		Multivariate			
	Coef. (S.E.)	Adj. R^2	Coef. (S.E.)	Coef. (S.E.) : Positive	Coef. (S.E.) : StrongModal	Adj. R^2
d/p	-0.021 (0.143)	-0.006	-0.044 (0.110)	-0.902 (0.229)	0.875 (0.225)	0.174
d/y	0.056 (0.107)	-0.004	0.033 (0.080)	-0.890 (0.228)	0.871 (0.231)	0.173
e/p	-0.334 (0.090)	0.109	-0.197 (0.130)	-0.693 (0.262)	0.728 (0.250)	0.202
d/e	0.376 (0.074)	0.139	0.302 (0.090)	-0.653 (0.224)	0.737 (0.207)	0.247
svar	0.339 (0.073)	0.112	0.222 (0.103)	-0.741 (0.227)	0.667 (0.233)	0.213
b/m	0.216 (0.078)	0.041	0.206 (0.055)	-0.887 (0.219)	0.859 (0.215)	0.216
ntis	-0.051 (0.037)	-0.004	-0.035 (0.037)	-0.891 (0.229)	0.871 (0.227)	0.173
tbl	-0.076 (0.080)	-0.001	0.009 (0.046)	-0.899 (0.228)	0.877 (0.228)	0.172
lty	-0.131 (0.146)	0.011	-0.079 (0.090)	-0.874 (0.212)	0.854 (0.207)	0.178
ltr	0.109 (0.117)	0.005	0.084 (0.093)	-0.884 (0.220)	0.864 (0.219)	0.179
tms	-0.221 (0.019)	0.044	-0.112 (0.043)	-0.837 (0.217)	0.792 (0.215)	0.183
dfy	0.053 (0.143)	-0.004	0.013 (0.091)	-0.893 (0.228)	0.869 (0.224)	0.172
dfr	-0.057 (0.136)	-0.003	-0.039 (0.091)	-0.894 (0.225)	0.867 (0.219)	0.173
infl	-0.064 (0.072)	-0.002	-0.092 (0.057)	-0.901 (0.220)	0.892 (0.211)	0.180

Table 22: Estimation results for sentiments, monthly frequency, and Dow30 Volatility

We estimate control-augmented models for monthly Dow30 Volatility by adding one of the macroeconomic and financial variables as a control variable and report the estimated coefficients and Newey–West heteroskedasticity– and autocorrelation–robust standard errors inside parentheses.

	Univariate		Multivariate				
	Coef. (S.E.)	Adj. R^2	Coef. (S.E.)	Coef. (S.E.) : StrongModal	Coef. (S.E.) : Positive	Coef. (S.E.) : Negative	Adj. R^2
d/p	-0.265 (0.137)	0.064	-0.321 (0.119)	1.243 (0.207)	-1.175 (0.219)	-0.130 (0.059)	0.377
d/y	-0.096 (0.084)	0.003	-0.119 (0.060)	1.181 (0.184)	-1.121 (0.209)	-0.048 (0.049)	0.293
e/p	-0.178 (0.144)	0.025	0.049 (0.100)	1.210 (0.230)	-1.152 (0.264)	-0.046 (0.052)	0.280
d/e	0.099 (0.126)	0.003	-0.054 (0.093)	1.195 (0.208)	-1.144 (0.238)	-0.041 (0.053)	0.281
svar	0.290 (0.031)	0.078	0.129 (0.050)	1.086 (0.194)	-1.030 (0.214)	-0.101 (0.056)	0.290
b/m	0.114 (0.165)	0.006	0.126 (0.165)	1.191 (0.205)	-1.111 (0.220)	-0.090 (0.056)	0.293
ntis	0.018 (0.020)	-0.006	0.031 (0.020)	1.173 (0.198)	-1.105 (0.219)	-0.044 (0.052)	0.280
tbl	-0.119 (0.067)	0.008	-0.008 (0.029)	1.170 (0.200)	-1.099 (0.219)	-0.046 (0.053)	0.279
lty	-0.181 (0.167)	0.026	-0.130 (0.091)	1.157 (0.173)	-1.075 (0.192)	-0.071 (0.049)	0.295
ltr	0.187 (0.159)	0.028	0.167 (0.121)	1.170 (0.187)	-1.087 (0.202)	-0.070 (0.048)	0.307
tms	-0.236 (0.026)	0.049	-0.126 (0.041)	1.120 (0.182)	-1.056 (0.205)	-0.107 (0.060)	0.290
dfy	0.215 (0.082)	0.040	0.169 (0.051)	1.146 (0.170)	-1.086 (0.194)	-0.073 (0.048)	0.307
dfr	-0.257 (0.135)	0.060	-0.273 (0.113)	1.189 (0.180)	-1.121 (0.200)	-0.142 (0.062)	0.346
infl	-0.009 (0.082)	-0.007	-0.054 (0.056)	1.184 (0.188)	-1.105 (0.212)	-0.044 (0.052)	0.282

Table 23: Estimation results for sentiments, monthly frequency, and Russell2000 Volatility

We estimate control-augmented models for monthly Russell2000 Volatility by adding one of the macroeconomic and financial variables as a control variable and report the estimated coefficients and Newey–West heteroskedasticity– and autocorrelation–robust standard errors inside parentheses.

	Univariate		Multivariate		
	Coef. (S.E.)	Adj. R^2	Coef. (S.E.)	Coef. (S.E.) : Constraining	Adj. R^2
d/p	-0.087 (0.060)	0.001	-0.143 (0.072)	-0.180 (0.050)	0.024
d/y	0.094 (0.061)	0.002	0.089 (0.062)	-0.132 (0.053)	0.013
e/p	-0.050 (0.044)	-0.004	-0.048 (0.044)	-0.135 (0.051)	0.007
d/e	0.031 (0.045)	-0.006	0.033 (0.045)	-0.136 (0.051)	0.006
svar	-0.147 (0.059)	0.015	-0.105 (0.062)	-0.081 (0.069)	0.013
b/m	0.032 (0.052)	-0.006	0.068 (0.069)	-0.151 (0.049)	0.009
ntis	0.031 (0.025)	-0.006	0.027 (0.029)	-0.135 (0.051)	0.006
tbl	0.072 (0.038)	-0.002	0.053 (0.037)	-0.128 (0.051)	0.008
lty	0.045 (0.081)	-0.005	0.031 (0.078)	-0.132 (0.052)	0.006
ltr	0.017 (0.094)	-0.006	0.028 (0.096)	-0.138 (0.049)	0.006
tms	0.077 (0.014)	-0.001	0.006 (0.037)	-0.132 (0.069)	0.005
dfy	-0.043 (0.056)	-0.005	-0.020 (0.052)	-0.132 (0.055)	0.005
dfr	0.056 (0.086)	-0.004	0.003 (0.100)	-0.134 (0.055)	0.005
infl	0.048 (0.066)	-0.004	0.061 (0.062)	-0.141 (0.050)	0.009

Table 24: Estimation results for sentiments, monthly frequency, and all indexes

We estimate control-augmented models for all stock market indexes by adding all macroeconomic and financial variables as control variables and report the adjusted R^2.

Target	Market	Adj. $R^2 - EconVars$	Adj. $R^2 - EconVars + Sentiments$
Excess Return	SP500	0.098	0.135
Close	SP500	0.055	0.091
Close	Dow30	0.049	0.092
Close	Nasdaq	0.058	0.083
Close	Russell2000	0.018	0.045
Volume	SP500	0.263	0.315
Volume	Dow30	0.015	0.059
Volume	Nasdaq	0.016	0.111
Volume	Russell2000	0.266	0.319
Volatility	SP500	0.067	0.070
Volatility	Dow30	0.244	0.389
Volatility	Nasdaq	–	–
Volatility	Russell2000	-0.001	0.007

Table 25: Top-5 most significant hashtag features from the weekly forecasting for the time period ending with December 2019

This table summarizes the set of hashtag features that are statistically significant at 10% significance level from the weekly prediction for the time period ending with December 2019 and reports the estimated coefficients and Newey–West heteroskedasticity– and autocorrelation–robust standard errors inside parentheses.

Target	Market	Significant Features
Close	SP500	#iwd2018 -0.109(0.003)
Close	Dow30	#iwd2018 -0.102(0.003)
Close	Nasdaq	#iwd2018 -0.108(0.003)
Close	Russell2000	
Volume	SP500	#followfriday 0.261(0.037) #goodfriday -0.224(0.119) #holiday -0.177(0.061) #cybermonday 0.167(0.053) #easter 0.164(0.081)
Volume	Dow30	#xmas 0.239(0.046) #random 0.135(0.040) #vmas 0.100(0.039) #skills 0.098(0.042)
Volume	Nasdaq	#merrychristmas 0.204(0.099) #happyvalentinesday 0.154(0.026) #vmas 0.122(0.025) #transparency -0.094(0.037)
Volume	Russell2000	#followfriday 0.233(0.033) #cybermonday 0.148(0.036) #holiday -0.144(0.032) #egypt -0.143(0.030) const -0.132(0.029)
Volatility	SP500	SP500_Volatility_L1 0.290(0.053) const -0.150(0.032) #climatechange 0.074(0.027) #climateaction 0.071(0.030) #orlando 0.069(0.006)
Volatility	Dow30	Dow30_Volatility_L1 0.261(0.060) const -0.119(0.034) #climateaction 0.082(0.031) #blackpanther 0.071(0.013) #parisagreement -0.062(0.016)
Volatility	Nasdaq	#beyonce 0.127(0.075) #skills 0.113(0.044)
Volatility	Russell2000	const -0.288(0.034)

Table 26: Top-5 most important hashtag features from the weekly prediction for the time period ending with December 2019

This table summarizes the set of the most important hashtag features along with their importance scores from the weekly prediction for the time period ending with December 2019.

Target	Market	Algorithm	Significant Features
Direction	SP500	Logit	#ff -0.0029 #internationalwomensday -0.0023 #superbowl 0.0022 #brexit 0.0022 #fb -0.0014
Direction	Dow30	DTrees	#obama 0.0566 #event 0.0381 #barcelona 0.0367 #please 0.0364 #mentor 0.0349
Direction	Nasdaq	SVC	#sxsw 0.0166 #brexit 0.0121 #ff 0.0103 #sotu 0.0072 #blockchain 0.0072
Direction	Russell2000	SVC	#fb -0.0204 #ff -0.0143 #winning 0.0122 #libya 0.01 #sxsw 0.0096