7-8-2023

Mining Firm-level Uncertainty in Stock Market: A Text Mining Approach

Yukun Liu
Baruch College, yliu373@fordham.edu

Yilu Zhou
Fordham University, yzhou62@fordham.edu

Zhihan Yang
Fordham University, zyang173@fordham.edu

Junhua Chen
Naperville North High School, daniele20051218@gmail.com

Mining Firm-level Uncertainty in Stock Market: A Text Mining Approach

Short Paper

Yukun Liu
Baruch College
1 Bernard Baruch Way
New York, NY 10010
yukun.liu@baruch.cuny.edu

Yilu Zhou
Fordham University
140 W 62nd St
New York, NY 10023
yzhou62@fordham.edu

Zhihan Yang
Fordham University
140 W 62nd St
New York, NY 10023
zyang173@fordham.edu

Daniel Junhua Chen
Naperville North High School
899 N Mill St
Naperville, IL 60563
danielc20051218@gmail.com

Abstract

The traditional finance paradigm seeks to understand uncertainty and their impact on stock market. However, most previous studies try to quantify uncertainty at macro-level such as the EPU index. There are few studies tapping into firm-level uncertainty. In this paper, we address this empirical anomaly by integrating text mining tools to measure the firm-level uncertainty score from news content. We focus on companies listed in S&P 1500. We crawled a total of 2,196,975 news articles from LexisNexis database from April 2007 to July 2017. We extracted uncertainty related information as features by using named entity extraction, LM dictionary, and other linguistic features. We employed nonlinear machine learning models to investigate the impact on stocks future returns by uncertainty-related features. To address the theoretical problem, we use traditional asset pricing techniques to test the relationship among information derived uncertainty and the financial market performance.

Keywords: Text Mining, Uncertainty, S&P1500, Machine learning, Random Forest

Introduction

Big data is rapidly changing the way financial markets work. Banks use big data analytics as a tool in credit risk management. Investment companies use big data processing and machine learning capabilities to process countless data points every day, helping them construct profitable stock portfolios. Insurance companies use big data in pricing, underwriting, and risk selection.

The traditional finance paradigm seeks to understand financial markets using models which assume agents are “rational”. Because of rational beliefs, people will immediately update their prospective when new information arrives. Furthermore, people’s decisions are always consistent with Subjective Expected Utility. However, in recent years, because of the massive information explosion and the limitation of cognitive substantially mitigate the ability of people to process new information to update their beliefs in a short-term period. Especially during times when the level of uncertainty increase, people’s perception on common information can largely diverge. Investors develop pessimistic expectation under an increasing uncertainty circumstance. On one hand, an uncertainty-related shock leads the increase of people's risk aversion, which makes them more likely to draw on other's outcomes as signals to infer the real state of economic. On the other hand, because of the expectation bias, investors process their decision largely rely on current situation.

Uncertainty is "the conditional volatility of a disturbance that is unforecastable from the perspective of economic agents" (Jurado et al., 2015). It is widely studied in finance and economics literature. To capture uncertainty in markets, researchers perform textual analysis to calculate readability and sentiment of corporate disclosures and measure political uncertainty at the market level. For example, Baker et al. (2016) introduce the Economic Policy Uncertainty (EPU) index to capture financial crisis, partisan policy dispute and serial crisis in Eurozone. In addition, Bernake (1983) point out that uncertainty increase the value of waiting for new information, retards the current rate of investment. Pastor et al (2012) and Gilchrist el al (2014) emphasize the increasing uncertainty cutback household spending and upward pressure of finance. However, most previous studies try to quantify uncertainty at macro-level. They demonstrate that macro-level uncertain has a significant impact on corporation growth and investment. There are few studies tapping into firm-level uncertainty.

Previous studies have suggested the correlation between news articles and stock price (Schumaker et al., 2009; Yu et al., 2013). They mostly rely on topics mining and sentiment analysis. In this study, we address this empirical anomaly by integrating information extraction tools to measure the firm-level uncertainty score from news content. We focus on companies listed in S&P 1500. We crawled a total of 2,196,975 news articles from LexisNexis database from April 2007 to July 2017. We employed nonlinear machine learning models to investigate the impact on stocks future returns by uncertainty-related features. To address the theoretical problem, we use traditional asset pricing techniques to test the relationship among information derived uncertainty and the financial market performance.

We begin our empirical analyses by investigating the impact of firm-level uncertainty on stock return. Our study contributes to the literature in several ways. First, we quantify the effects of economic linkages on the frequency of news co-occurrence. Second, we show that in the context of news cooccurrence, it is not "in-the-news" per se but the surprise component that attracts more investor attention. Third, we show that stock return co-movements increase with news co-occurrences, and such increased co-movements cannot be explained away by economic linkages, well-known stock characteristics, and after accounting for persistence in return correlations.

Literature Review

Uncertainty and Stock Market

In previous literatures, Baker, Bloom and Davis (2016) construct a series of Economic Policy Uncertainty (EPU) indexes by extracting specific keywords related to economic policy uncertainty from leading newspapers, referring the number of federal tax code provisions set to expire, and forecasting disagreement over future inflation and government purchase. As a result, they demonstrate the EPU index negatively impact on economic growth and labor market performance. We consider their methodology as a baseline to identify the potential uncertainty across and within sectors. First, the aggregating news information on firm-level draw massive attention from investors. Second, their perception on uncertainty shocks possibly be incompleteness. Afterward, agents do not update their beliefs in a "rational" manner.

Nevertheless, Baker et al (2016) document that the EPU case the market turbulence by showing the high correlation among EPU and US VIX. Similarly, Chiang (2019) shows the raising EPU negatively impact the business operation and cause the market expectation to deteriorate. Thus, an induced sell-off aggravate market volatility. Nevertheless, Gozgor, Lau, and Bilgin (2016) use the GJR-GARCH model estimation to price volatility transmission on commodity market. Bilgin, Gozgor, Lau and Sheng (2018) expand their previous research to measure the impact on gold price from economic uncertainty. Their nonlinear Autoregression-distributed Lag model investigate on the asymmetric effect of uncertainty measures on gold price. In addition, Matkovskyy, Jalan, Dowling (2020) analyze the interdependence between traditional financial markets and Cryptocurrency markets and their reaction to selected policy shocks.

Moreover, uncertainty and stock price are both stochastic with large temporal shocks. However, some research proves the expectation bias among people do exits. Chang, Huang and Wang (2017) find that daily air pollution levels have a significant effect on the decision to purchase or cancel health insurance in a manner inconsistent with a rational choice theory. Based on the standard economic theory, they demonstrate that the importance decisions that people make have lasting consequence. As such, they require people to predict the utility they will receive in the future from decision they make today. Hirshleifer and Teoh (2003) address the firms' choices between alternative means of presenting information and the effects of different presentations on market prices when investors have limited attention and processing power. Meanwhile, Peng and Xiong (2006) are motivated the idea of limited attention. They model investors' attention allocation in learning and study the effects of this on asset-price dynamics. Both research demonstrate cognitive-overloaded investors pay attention to only a subset of publicly available information.

All above mentioned among others research indicate the investors' pessimistic expectation under an increasing uncertainty circumstance. On one hand, an uncertainty-related shock leads the increase of people's risk aversion, which makes them more likely to draw on other's outcomes as signals to infer the real state of economic. On the other hand, because of the expectation bias, investors process their decision largely rely on current situation. Especially the continues increasingly negative sentiment cause the divergent perception of new information about uncertainty.

Text Mining

Text mining is a powerful analytics tool in leveraging information from news articles (Chen, Chiang et al. 2012). It is capable of discovering hidden knowledge from large volume of data. Text mining research deals with a variety of problems including text summarization, document and information retrieval, text categorization, authorship identification, entity extraction and relation extraction (Witten, 2014). It is capable of discovering hidden knowledge from large volume of data.

Previous studies have suggested the correlation between news articles and stock price (Schumaker et al., 2009; Yu et al., 2013). However, they mostly rely on topics mining and sentiment analysis. For example, Schumaker et al. (2009) extracted noun phrases from news articles to detect breaking news. This information is then combined with regression analysis to improve stock price prediction accuracy. Yu et al. (2013) and Schumaker et al. (2013) both extracted sentiment in news articles to correlate with stock price.

Research Questions

In this paper, we aim to address the real uncertainty impact on people's decision from firm-level news articles. First, we apply different benchmark to test the uncertainty magnitude based on textual analysis and information extraction. For example, sentiments conduct the news watchers' initial subjective tendency. With a monthly basis, the increasingly negative sentiment cumulate investors' panic on future business performance. On the opposite side, positive sentiment lifts the expectation on the firms' growth outlook. However, the sentiment analysis is insufficient to represent the uncertainty impacts. Second, we follow the dictionary from Loughran and McDonald Master Dictionary (2020) (LM), which is constructed on firms' annual reports. By detecting the keywords term frequency help us to more comprehensively measure the uncertainty level for each news content. Third, we expand the dictionary by adding the antonym. The uncertainty-related keywords have capability to attract news watchers' attention, but the certainty-related keywords yield the same functionality. If an article contains high volume of both uncertainty-related and certainty-related keywords, this extremely increase the complexity of cognitive. As a result, it enhances investors' disagreements on the specific event. Telling examples include the suddenly obtained patents among firms, but the profitability of patents remain uncertainty. Rationally, the technology progress enhances companies' competitive and economic growth estimation. Behaviorally, the remained uncertainty on future cash flow immediately mitigates the positive outlook. As a result, people are facing the challenge to distinct the real uncertainty in a positive sentiment and real certainty in a negative sentiment. Fourth, since the reading preference adjusts people's weighted attention, we use the position strength approach to allocate the magnitude for each keyword.

For empirical analysis, we employ nonlinear machine learning models to investigate the impact on stocks future returns by uncertainty-related features. To address the theoretical problem, we use traditional asset pricing techniques to test the relationship among information derived uncertainty and the financial market performance.

Methodology

Our methodology which includes five components: 1) Data Collection, 2) Dictionary Building, 3) Dictionary Expansion, 4) Building Uncertainty Score, 5) Random Forest Model and 6) Validation. ***Normal or Body***

Data Collection

The news article data comes from Nexis Uni (Previous named LexisNexis). The Nexis Uni started to provide legal and journalistic documents accessible electronically since 1970s. As of 2006, the company incorporates the largest electronic database for legal and public-records information. Due to the limitation of manually buck download on news articles, we utilize the selenium to simulate human actions for avoid authentication. Selenium is a web browser automation tool created to automate web applications for testing. It is now used for various other purposes, including automating web-based administrative tasks, interacting with platforms that do not provide an API, and web crawling. The crawling algorithm searches news articles that contain each company name within each year. In this paper, our news data time horizontal covers from April 2007 to July 2017, and the variant news resources including “Plastic News”, “Business Matters”, “Right Version News”, “The Mecklenburg Times” etc.

Because the same news article may return from different searches, we remove redundant articles by checking the title, date and author of the article. A total of 2,671,004 news articles that covers years 2007 to 2017 were retained after the crawling process. These articles cover a total of 1,434 companies. The missing companies are those not generating results from news search. For the stock price performance data, we incorporate the monthly basis data from CRSP. In addition, for better comprehensively analyze the uncertainty impact, we remove stocks with prices less than $5 at the observation time, and focus on traditional exchanges NYSE, AMEX, and NASDAQ.

Number of Articles	2,196,975
Time Period	April 2007 – July 2017
The total Number of unique Companies	1,195
Stock Exchange	NYSE, AMEX, NASDAQ
Frequency	Monthly

Table 1 Basic Information about News Dataset (after filtering)

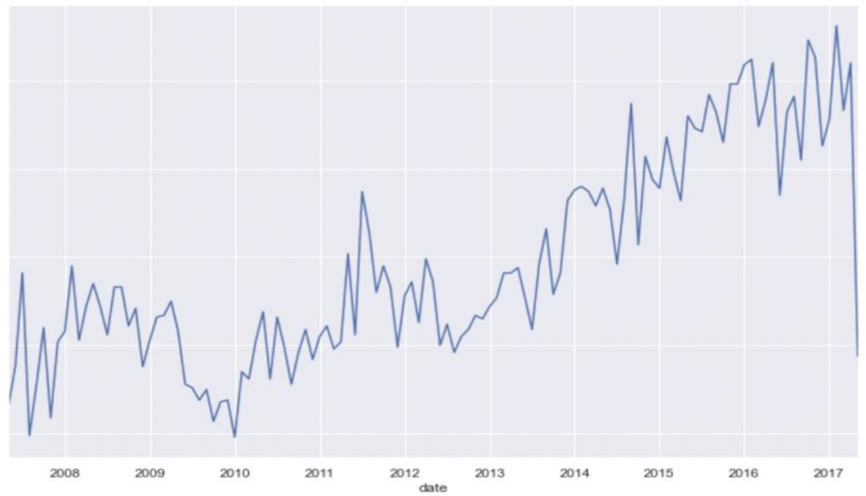


Figure 1. The monthly observations we have covered in our dataset

Dictionary Building

Loughran and Mcdonald (2011) firstly introduce the widespread recognized textual analysis approach in financial market. In traditional empirical analysis, the negative words play an important rule to measures the ton of text. However, Loughran et al (2011) demonstrates nearly three-fourths of negative words are typically not considered negative in financial market. Afterwards, they develop alternative dictionary links to 10-K filling returns, trading volume, return volatility, fraud, material weakness, and unexpected earnings.

In addition, due to the frequently occurring negative words like “loss”, “losses”, “adverse”, and “failure”, Loughran and Mcdonald (2015) shows the faily high correlation among Diction pessimistic and their negative words dictionary. Similarly, Guillamon-Saorin, Isidro, and Marques (2017) integrate LM dictionary to prove the managerial sentiment on non-GAAP earnings quality. Their results indicate that non-GAAP measurement are informative to capital market, but non-GAAP adjustments are more persistent when accompanied by higher levels of impression management. Nevertheless, Allee and Deangelis (2015) employ the “bag-of-word” approach by following Loughran et al (2015) findings. They suggest that tone dispersion both reflects and affects the information that managers convey through their narratives.

In our paper, we incorporate the LM dictionary alternative list “uncertainty” as our first approach to measure the potential uncertainty scores in firm news. The words list consists of 297 unique words which highly correlated the meaning of uncertainty in financial documents. Since the negative words strongly represent a pessimistic tone, that has limitation to measure the perception of uncertainty from news watchers.

Dictionary Expansion

In our empirical analysis, we find the uncertainty words independently yield bias on estimation the uncertainty magnitude. Because it is feasible to measure the negative sentiment about uncertainty problems, but it has limitation to capture the disagreement or divergency on information incompleteness. In this way, we employ Wordnet tool to process the expansionary of words list. A simple way to address the issue is calculating the negatively similarity scores on each word. It helps us to identify the true antonym words, which represents the meaning of “certainty”. We consider the expanded words list as a proxy to better comprehend and classify real uncertainty within the news article.

Named Entity Extraction

These news articles are processed to extract meta information such as publish time, source, author, title and news text. We ran Stanford Named Entity Recognizer (NER) program to extract named entities (https://nlp.stanford.edu/software/CRF-NER.html). NER aims to extract and classify rigid designators (Nadeau and Sekine, 2007). There are many types of named entities in text such as company name, person name, and product name. These named entities are then mapped against our company name list and their variations. Performing NER before mapping company names is necessary to handle generic keywords that appear in company names. For example, Gap Inc is sometimes referred to as Gap. If we map the keyword “Gap” directly, we may mistakenly count the generic keyword “gap” as the occurrence of the company name. We also manually created a name variation table to increase the coverage of our mapping algorithm. A company’s name can appear in multiple forms, a problem referred to as name variation. Since we are only interested in the S&P 1500 company names, a name variation table is the easiest way to handle the problem. For example, Walmart Inc can appear as Wal Mart, Walmart, Wal-Mart Stores, Inc. The appearance of these names will all be aggregated to Walmart Inc.

Building Uncertainty Scores

We develop variant models to approach the uncertainty score measurement. First, the Benchmark-1 is a baseline model that only depends on the term frequency of uncertainty words in LM dictionary. This approach has limitation to estimate the real uncertainty. Then, we implement a seconding approach Benchamark-2. In this practice, we include the characteristics regarding each news article. For example, the number of news observed in time t-1, the position strength of word “uncertainty”, the average sentiment scores for company i at time t-1, the normalized uncertainty score based on LM dictionary only (Benchmark-1), the first mention of LM uncertainty words and the “uncertainty” word. The characteristic features enhance the predictability on measuring the uncertainty level but remain bias on capturing the disagreement. In the Benchmark-3, we expand our textual features by incorporating the “certainty” related words. As a result, our model yield strong prediction on the uncertainty scores with statistically significant.

Feature	Explanation	Datatype
Sentence_count	The number of sentences in an article	int
lm_first_mention	The sentence index of an LM word first mentioned in an article	int
uncert_first_mention	The sentence index of the first mention of “uncertain” related keyword (uncertain, uncertainty, uncertainties, etc.)	int
position_strength	The calculated position strength based on first mentioned uncertainty sentences	float
lm_count	count the number of LM words in an article	int
lm_score	the summation of TF-IDF score of LM dictionary keywords in an article.	float
uncertainty_control	A binary feature to indicate if an article contains uncertain words	binary

Table 2. Features used to calculate “Uncertainty Score”

Random Forest Classification

Random Forest is a supervised learning algorithm. The “Forest” is an ensemble of decision trees, usually trained with the “bagging” method. This method builds multiple decision trees and merges together to get more accurate and stable prediction.
Implementation in Scikit-learn:

$$n_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

For each decision tree, Scikit-learn calculates a nodes importance using Gini importance with binary tree. Where n_j stands for the importance of node j. w_j represents for weighted number of samples reaching node j. C_j represents the impurity value of node j. $left(j)$ represents child node from left split on node j. $right(j)$ represents child node from right split on node j.

$$f_i = \frac{\sum_{j=1}^{T} n_j}{\sum_{k=1}^{N} n_k}$$

The feature importance f_i is calculated by the sum of all node j from feature i divided all importance from node j.

$$norm(f_i) = \frac{f_i}{\sum_{i=1}^{F} f_i}$$

The normalized feature importance is divided the feature importance f_i by all feature importance.

$$RandomForest(f_i) = \frac{\sum_{t=1}^{Tr} norm(f_{i,t})}{Tr}$$

The Random Forest feature importance f_i is calculated by divided the sum of all tree normalized feature importance by number of trees.

Initial Experiment

The RandomForest Classification provide significant result of our approach. In the empirical analysis, we define the model by following: (i) Using the one month ahead stock returns as our target variable to measure the uncertainty scores; (ii) Identity the problem as classification supervised learning by convert stock returns into quintile ranked variables; (iii) Test the rolling window with 5 years training sample and 1 month out of sample testing; (iv) Employ the non-linear machine learning model to estimate the magnitude of uncertainty mentioned in news article per company; (v) Test the predicted variable on portfolio returns difference.

Benchmark 1

$$S_{i,t} \sim w * \eta_{i,t-1}$$

Where the $S_{i,t}$ stands for the uncertainty score for company i at time t (monthly basis). $\eta_{i,t-1}$ is the series to represent the “uncertainty” words term frequency for company i at time $t-1$.

Benchmark 2

$$S_{i,t} \sim w_1 * \eta_{i,t-1} + w_2 * \delta_{i,t-1}$$

Where $\delta_{i,t-1}$ is the series to represent the characteristic features for company i at time $t-1$.

Benchmark 3

$$S_{i,t} \sim w_1 * \eta_{i,t-1} + w_2 * \delta_{i,t-1} + w_3 * \rho_{i,t-1}$$

Where $\rho_{i,t-1}$ is the series to represent the “certainty” words term frequency for company i at time $t-1$. The certainty words list contains 97 antonym unique words, which extracted by WordNet from LM dictionary with word similarity calculation.

With a 5-year training and 1 month out of sample prediction, our model generates the testing period from 2013 to 2017 (60 months observation). The return difference in quintile sort is based on the machine learning weighted probability whereas:

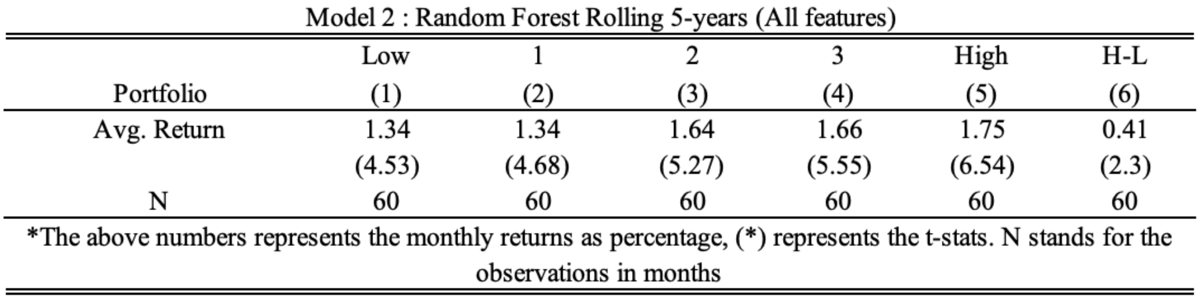
$$P_{i,t} = w_1 * Pr_1 + w_2 * Pr_2 + w_3 * Pr_3 + w_4 * Pr_4 + w_5 * Pr_5$$

$P_{i,t}$ stands for the predicted probability from machine learning for company i at time t; and the $[w_1, w_2, w_3, w_4, w_5]$ is the sample weights [0,1,2,3,4]; $[Pr_1, Pr_2, Pr_3, Pr_4, Pr_5]$ is the predicted probability for each label, respectively.

As a result, our model shows a statistic significant return difference approximately 0.41% for value weighted and 0.33% for equal weighted with t-stat 2.3 and 2.21, respectively. All the t-stat is Newey-west 3 months adjustment.

Model	Features	Rolling Window	Significant
RandomForest	Benchmark – 1	5-yr Train + 1-mo test	No
RandomForest	Benchmark – 2	5-yr Train + 1-mo test	No
RandomForest	Benchmark – 3	5-yr Train + 1-mo test	YES
RandomForest	Benchmark – 3	3-yr Train + 1-mo test	No
RandomForest	Benchmark – 3	5-yr Train + 5-yr test	No

Table 2. Validation Test Result with Random Forest Model

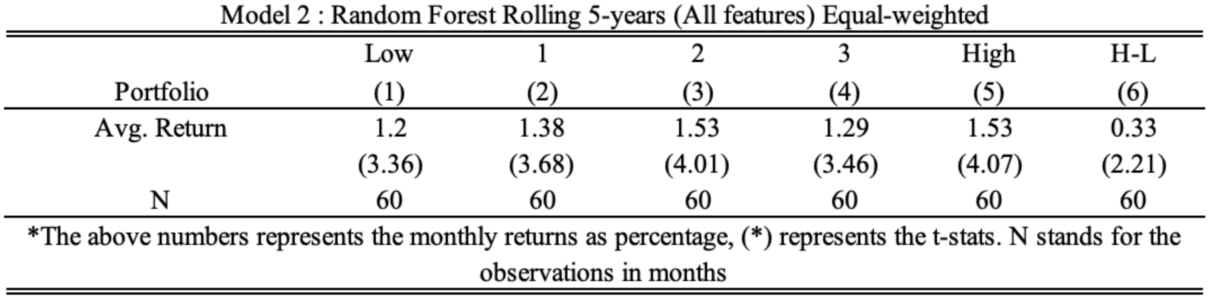


Model 2 : Random Forest Rolling 5-years (All features)

	Low	1	2	3	High	H-L
Portfolio	(1)	(2)	(3)	(4)	(5)	(6)
Avg. Return	1.34	1.34	1.64	1.66	1.75	0.41
	(4.53)	(4.68)	(5.27)	(5.55)	(6.54)	(2.3)
N	60	60	60	60	60	60

*The above numbers represents the monthly returns as percentage, (*) represents the t-stats. N stands for the observations in months

Table 3. Validation Test Result with Value-Weighted



Model 2 : Random Forest Rolling 5-years (All features) Equal-weighted

	Low	1	2	3	High	H-L
Portfolio	(1)	(2)	(3)	(4)	(5)	(6)
Avg. Return	1.2	1.38	1.53	1.29	1.53	0.33
	(3.36)	(3.68)	(4.01)	(3.46)	(4.07)	(2.21)
N	60	60	60	60	60	60

*The above numbers represents the monthly returns as percentage, (*) represents the t-stats. N stands for the observations in months

Table 4. Validation Test Result with Equal-Weighted

Conclusion and Future Work

This study is to design a data-driven text mining framework that discovers uncertainty information embedded in news articles. Using an expanded uncertainty keyword dictionary and linguistic features of word context and position, we measured the intensity of uncertainty. This is one of the pioneer research that quantifies intensity of uncertainty using news articles. We believe that data model can address the limitation of the ability of people to process new information to update their beliefs in a short-term period. Especially during times when the level of uncertainty increase, the model can measure the firm-level uncertainty score from news content and able to achieve better stock portfolio performance by utilizing such information.

In the future, we plan to extend our study in several directions. First, we plan to incorporate more features in textual data in our machine learning model. The uncertainty score should be capable not only to detect generic uncertainty information, but also fine-grained information of specific uncertainty types. Thus, we plan to use additional text analysis techniques such as topic modeling and sentiment analysis as additional features in our machine learning model. Second, we will continue to expand the dataset to include news articles after 2017. This will allow us to capture more types of uncertainties such as Covid. This will also allow us to perform additional validation tests on a longer period of time. Lastly, we plan to incorporate more comprehensive machine learning models such as LSTM and BERT models to capture more contextual information in news articles.

References

Alfaro, I., Bloom, N., & Lin, X. (2018). The finance uncertainty multiplier (No. w24571). *National Bureau of Economic Research*.

Allee, K. D., & DeAngelis, M. D. (2015). The structure of voluntary disclosure narratives: Evidence from tone dispersion. *Journal of Accounting Research*, 53(2), 241-274

Arısoy, Y. E., Bali, T. G., & Tang, Y. (2019). Anticipated Regret and Equity Returns. Georgetown McDonough School of Business Research Paper, (3195191).

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4), 1593-1636.

Bao, S., Li, R., Yu, Y., & Cao, Y. (2008). Competitor mining with the web. *IEEE Transactions on Knowledge and Data Engineering*, *20*(10), 1297-1310.

Bilgin, M. H., Gozgor, G., Lau, C. K. M., & Sheng, X. (2018). The effects of uncertainty measures on the price of gold. *International Review of Financial Analysis*, 58, 1-7.

Chang, T. Y., Huang, W., & Wang, Y. (2018). Something in the air: Pollution and the demand for health insurance. *The Review of Economic Studies*, 85(3), 1609-1634.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS quarterly*, 1165-1188.

Chiang, T. C. (2019). Economic policy uncertainty, risk and stock returns: Evidence from G7 stock markets. *Finance Research Letters*, 29(C), 41-49.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, 116(1), 1-22.

Gozgor, G., Lau, C. K. M., & Bilgin, M. H. (2016). Commodity markets volatility transmission: Roles of risk perceptions and uncertainty in financial markets. Journal of *International Financial Markets, Institutions and Money*, 44, 35-45.

Hirshleifer, D., & Teoh, S. H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of accounting and economics*, 36(1-3), 337-386.

Hong, H., & Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of finance*, 54(6), 2143-2184.

Jurado, K., Ludvigson, S. C., & Ng, S. (2015). Measuring uncertainty. *American Economic Review*, 105(3), 1177-1216.

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

Loughran, T., & McDonald, B. (2015). The use of word lists in textual analysis. *Journal of Behavioral Finance*, 16(1), 1-11.

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

Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *the Journal of Finance*, 69(4), 1643-1671.

Ma, Z., Pant, G., & Sheng, O. R. (2011). Mining competitor relationships from online news: A network-based approach. *Electronic Commerce Research and Applications*, *10*(4), 418-427.

Matkovskyy, R., Jalan, A., & Dowling, M. (2020). Effects of economic policy uncertainty shocks on the interdependence between Bitcoin and traditional financial markets. *The Quarterly Review of Economics and Finance*, 77, 150-155.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3-26.

Peng, L., & Xiong, W. (2006). Investor attention, overconfidence and category learning. *Journal of Financial Economics*, 80(3), 563-602.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, *27*(2), 12.

Schumaker, R. P., Zhang, Y., Huang, C. N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, *53*(3), 458-464.

Yu, L. C., Wu, J. L., Chang, P. C., & Chu, H. S. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems*, *41*, 89-97.

Witten, I. H., Don, K. J., Dewsnip, M., & Tablan, V. (2004). Text mining in a digital library. *International Journal on Digital Libraries*, *4*(1), 56-59.