News and Expected Volatility in the Stock Market *

Mengchen Shi†

June 2018

Abstract

Often referred to as the investor fear gauge, the market volatility index VIX expresses a consensus view about expected future stock market volatility; the higher the VIX, the greater the fear in the market. Prior work has considered the relationship between the implied volatility index and stock market returns. The goal of this study is to uncover some potential signals embedded in unstructured text data. We employ a few methods to model text data and attempt to capture signals that may help describe or predict the movement of the VIX index. Text data is unstructured, since there lacks a clear and defined way to understand it in intuitive, quantitative terms. Hence we introduce the Word2Vec model to analyze text. The model performs well in in-sample data, and successfully predicts general trends of out-of-sample VIX.

keywords: stock market volatility, VIX, news articles, textual analysis, NLP.

*I would like to express my sincere gratitude to Dr.Rick Evans and Dr.Ben Soltaff for your teaching, guidance, and support throughout the quarter.

†University of Chicago, Computational Social Science, mshi17@uchicago.edu.

1 Introduction

This study explores the relationship between news and expected volatility in the stock market. Expected volatility, intuitively, is the stock markets expectation of volatility in the next period of time.VIX is a popular measure of the stock market's expectation of volatility implied by SP 500 index options, calculated and published by the Chicago Board Options Exchange (CBOE). It is colloquially referred to as the fear index or the fear gauge. Prior work has considered the relationship between the implied volatility index and stock market returns. A significant negative and asymmetric contemporaneous relationship between stock returns and changes in implied volatility (Whaley, 2000; Giot, 2005).

Asset pricing theory holds that fluctuation in options implied volatility can predict stock market returns to some degree. Merton (1973) found that implied volatility measures fluctuation in expected stock market volatility. Drechsler and Yaron (2011) found that implied volatility measures variance risk premium. Some researcher also discovered that the option implied volatility can predict large economics disasters (Gabaix, 2012; Wachter, 2013; Gourio, 2008, 2012). These researches illustrate the importance of predicting VIX one of the target of my project.

Some financial economics research works have evaluated the impact of news on stock returns as well as volatility of the returns. Researchers have found that the arrival of firm-specific news have relationship with both stock prices and volatility. Earlier researchers focused on news that are known in advance, such as dividend announcements, earnings results in annual reports.

More recently, analyzing context and quality of news content has driven the identification of a broader range of news. The relationships between the news and stock returns and stock volatility have been confirmed by these researches.

There are some studies about the relationship between news and stock market volatility as well. These researches are the most inspiring part of the literature reviews. They provide ideas about where to find data sources, how to analyze the content and sentiment of news, as well as how to analyze the relationship between stock volatility and news. For instance, Smales (2014) used Ravenpacks Multi-Classifier for Equities sentiment indicator to confirm the significant negative relationship between 2000 - 2010 news releases and market volatility, measured by the implied volatility index. Manela and Moreira (2016) constructed a text-based measure of uncertainty starting in 1890 using front-page articles of the Wall Street Journal. They found that News implied volatility peaks during stock market crashes, times of policy-related uncertainty, world wars and financial crises.

2 Data

2.1 Source of Data

2.1.1 News

Using BeautifulSoup4 (a package in Python), I build a website crawler to scrape news article data on Wall Street Journal website from January 1, 2012 to May 1, 2018. Headline text, abstracts and date of articles are saved into a structured dataset. In total, 326,000 observations are collected. Headlines and abstract are available for free for everybody, but people need to pay for the whole articles on Wall Street Journal. That is the reason why we only use headlines and abstracts in this project.

2.1.2 Expected volatility in the stock market

VIX is a popular measure of the stock market's expectation of volatility implied by S&P 500 index options, calculated and published by the Chicago Board Options Exchange (CBOE). It expresses a consensus view about expected future stock market volatility; the higher the VIX, the greater the fear in the market. It is colloquially referred to as the fear index or the fear gauge. I download daily VIX data from January 1, 2012 to May 1, 2018 from Yahoo Finance.

2.2 Exploratory Data Analysis

2.2.1 VIX

VIX is available on 1595 days in the observation duration. Table 1 is the summary of VIX over six years (2012 to 2018) . Figure 1 below is the general performance over the years. Figure 2 is a histogram of the distribution of VIX.

Figure 1: Daily VIX from 2012 to 2018

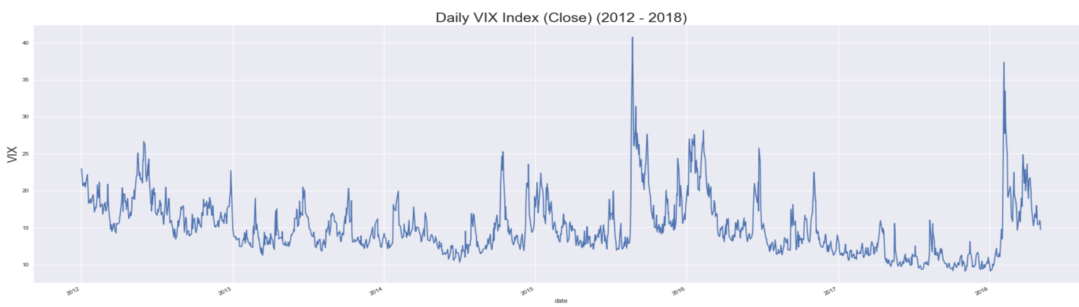


Figure 2: VIX Histogram

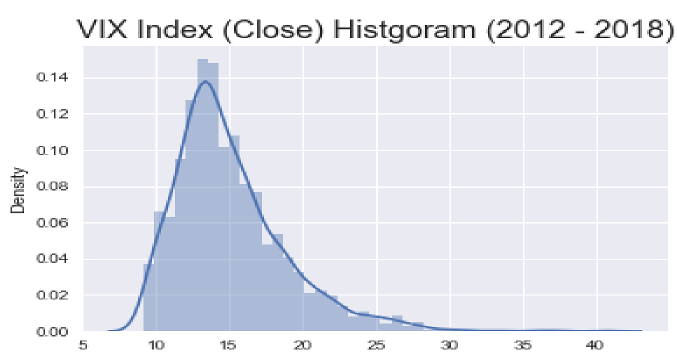


Table 1: VIX: Data Description

	VIX	N	Mean	Std. Dev.	Min	Median	Max
0	Open	1595	15.180188	3.785375	9.01	14.37	9.01
1	High	1595	15.972821	4.335125	9.31	14.97	9.31
2	Low	1595	14.462727	3.393180	8.56	13.79	8.56
3	Close	1595	15.100690	3.818294	9.14	14.23	9.14

2.2.2 News

In total, 326,000 observations are collected. After excluding dates on which less than 50 news articles were collected, we get 316,511 news articles on 1,901 dates. After combining VIX and news, we have 1,556 dates eventually.

Figure 3: Number of Daily News Histogram

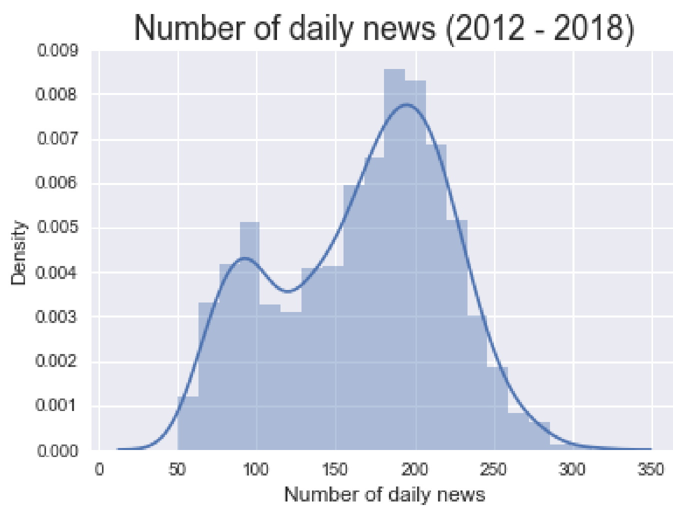


Table 2: Number of Words in Daily News

	Number of Words in Daily News
count	1556
mean	5804.785
std	1407.404
min	1501.000
25%	4982.500
50%	5917.500
75%	6782.500
max	9942.000

3 Analysis and Results

3.1 VIX and Word2Vec OLS Regression

Google's Word2Vec model, developed by Mikolov et al is a useful and commonly model in Natural Language Processing. Instead of representing each word as a unique feature, Word2Vec treat words as an N-vectors of meaning. Words can therefore be added up, reducing the total number of features in our representation from ¿ 10,000 to 50, 100, or whatever number of dimensions is desired. Using the original Wall Street Journal news dataset, we convert news on each day into a 300-dim vector based on the average Word2Vec word in the news. In other words,

$$h_j = \frac{1}{N_j} \sum_{i}^{N_j} v_i$$

where N_j is the number of words in the $j - th$ news. I choose 300-dim because it generates a fairly small Mean squared error compared to others.

We add all the vector news representations for each date, and the VIX dataframe by date. The effect of any word can be approximated by

$$e_i = \hat{\beta} v_i$$

since each word has a linear effect onto the VIX index. Hence we would like to perform an OLS estimation of the effects of Word2Vec news onto the VIX index. We compute the model in-sample

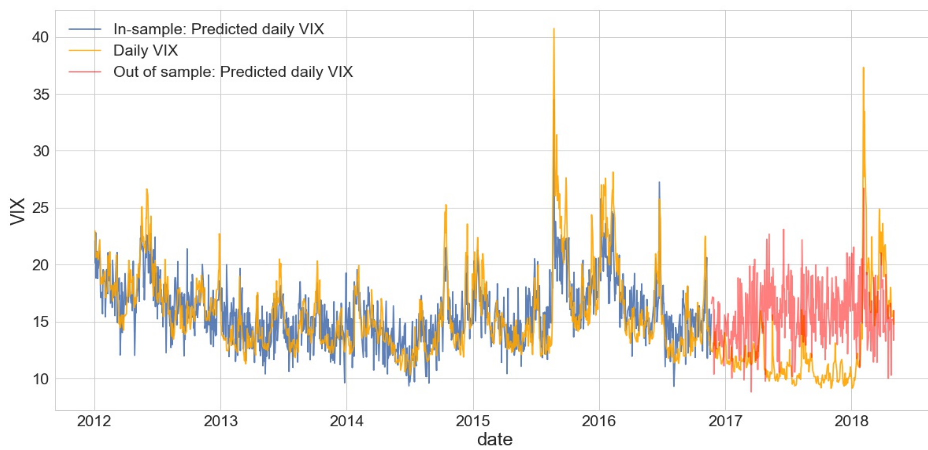
$$VIX_t = \beta X_t + \epsilon$$

where t is each day, and X_t is the Word2Vec representation for news.

To train the Word2Vec model, the main parameter is how many dimensions should be included to describe news. By minimizing mean squared error (MSE) of models as well as to minimize prediction errors, we find that 100 is a reasonable parameter to construct the model.

The outcomes of our regression are in Figure 4 and 5.

Figure 4: Word2Vec Estimation Performance (Daily)



As we notice, the model fails OOS (Out-of-sample). For in-sample, the fit is quite good, since most of the spurious variation is captured by the unwieldy model. However, once OOS, none of the meaningful variation is captured, making this model for prediction quite poor.

As known to all, influence of news has time lags. That is to say, impact of news on volatility of the stock market might not be reflected instantly, but in the long run. We then use news articles in each month to predict average VIX in each month.

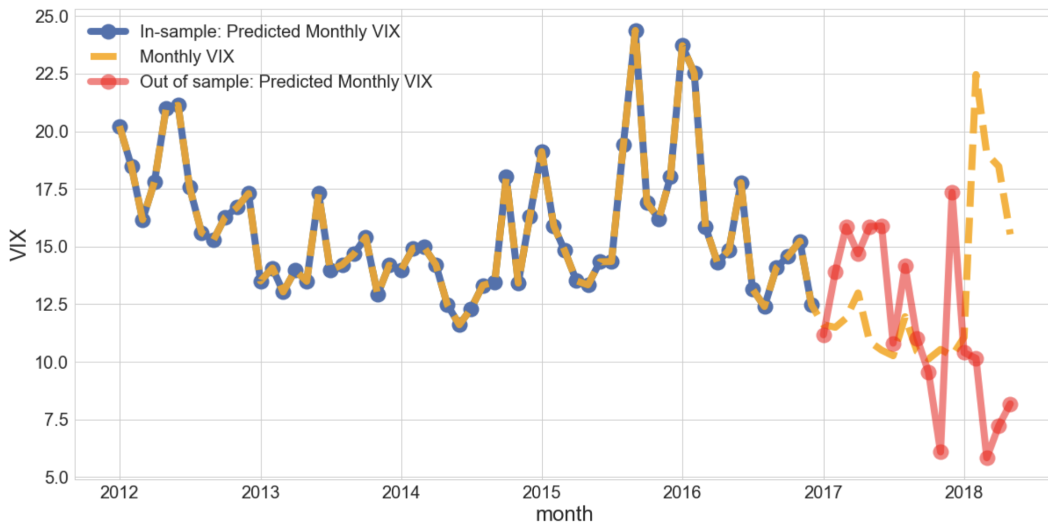
3.2 VIX Percent Change and Word2Vec OLS

It could be possible that news not only influence absolute value of VIX, but also the percentage change of VIX. Therefore, we would like to perform an OLS estimation of the effects of Word2Vec news onto the percent change of VIX index everyday:

$$\Delta VIX_t = \beta X_t + \epsilon$$

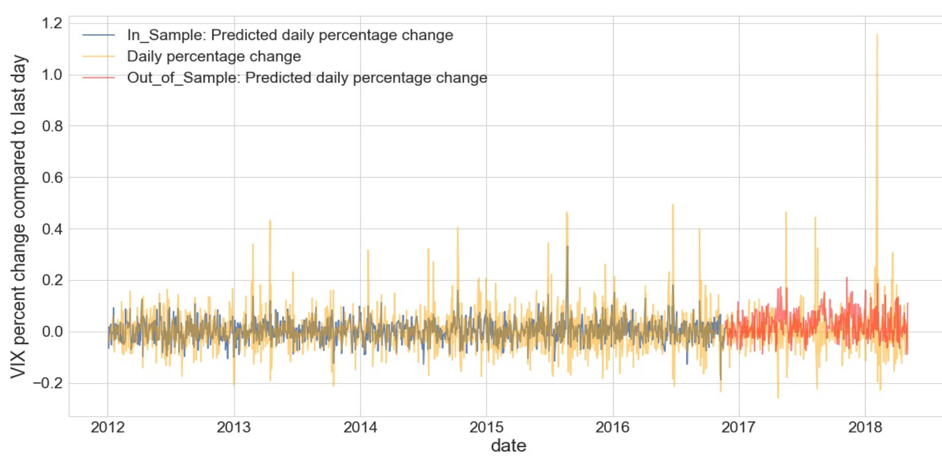
where ΔVIX_t is equal to $VIX_t - VIX_{t-1}$, and Xt is the Word2Vec representation for news.

Figure 5: Word2Vec Estimation Performance (Monthly)



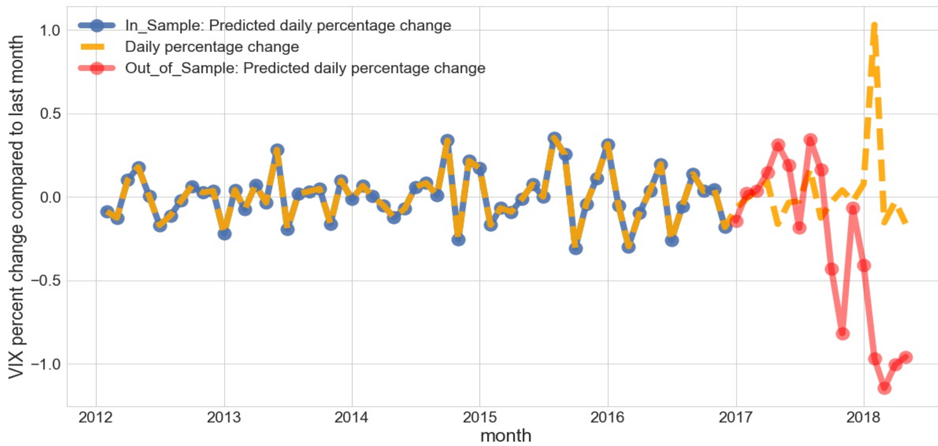
The outcomes of our regression are in Figure 6 and 7.

Figure 6: Word2Vec Estimation Performance – VIX Percent Change Compared to Last Day



Unfortunately, both in-sample and out-of-sample fitting perform poorly. It predict trends of volatility well, even though the absolute differences are large/

Figure 7: Word2Vec Estimation Performance – VIX Percent Change Compared to Last Month



4 Conclusion

From this research, we can learn that Word2Vec is a reasonable model to describe news text. Also, there is a relationship between news articles and volatility of the stock market that can be modelled, and therefore proves that to predict stock market volatility with news is a feasible method. However, it is is hard to balance between over fitting problem and out-of-sample prediction precision.

There are a lot to be done to improve the model.First, we should include more parameters to represent sentiment of news. Although Word2Vec might have included sentiment factors in the model, to include explicit sentiment factors could help us understand the how sentiment and tones of news influence the volatility.

Second, we should considering weighting powers of news. At present, we assign equal weights to every piece of news. But in reality, news articles have different power. For instance, political news and economic news obviously have stronger impact on the stock market then sports news. With more data collected from different news website, we should also take the influence of the websites into consideration.

References

Das, Sanjiv R., and M. Y. Chen. Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science* 53.9(2007):1375-1388.

Drechsler, Itamar, and A. Yaron. What's Vol Got to Do with It. *Review of Financial Studies* 24.1(2011):1-45.

Dzielinski, M. News Sensitivity and the Cross - Section of Stock Returns. *Ssrn Electronic Journal* (2011).

Gabaix, Xavier. Variable Rare Disasters: An Exactly Solved Framework for Ten Puzzles in Macro-Finance. *Nber Working Papers* 127.2(2008):645-700(56).

Gidofalvi, Gyozo, and Charles Elkan. Using news articles to predict stock price movements. *Department of Computer Science and Engineering, University of California, San Diego* (2001).

Giot, P. Relationships Between Implied Volatility Indexes and Stock Index Returns. *Journal of Portfolio Management* 31.3(2005): 92-100.

Gourio, Franois. Disaster Risk and Business Cycles. *American Economic Review* 102.6(2012):2734-2766.

Gourio, Franois. Time-series predictability in the disaster model. *Finance Research Letters* 5.4(2008):191-203.

Li, Xiaodong, et al. Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information. *Neurocomputing* 142.1(2014):228-238.

Malo, Pekka, et al. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science Technology* 65.4(2014):782-796.

Manela, Asaf, and A. Moreira. News implied volatility and disaster concerns. *Journal of Financial Economics* (2016).

Merton, Robert C. An Intertemporal Capital Asset Pricing Model. *Econometrica* 41.5(1973):867-887.

Schumaker, Robert P., and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)* 27.2 (2009): 12.

Smales, Lee A. News sentiment and the investor fear gauge. *Finance Research Letters* 11.2(2014):122-130.

Smales, Lee A. Non-Scheduled News Arrival and High-Frequency Stock Market Dynamics: Evidence from the Australian Securities Exchange. *Research in International Business Finance* 32(2014):122-138.

Wachter, Jessica A. Can TimeVarying Risk of Rare Disasters Explain Aggregate Stock Market Volatility? *Journal of Finance* 68.3(2013):987-1035.

Whaley, Robert E. The Investor Fear Gauge. *Journal of Portfolio Management* 26.3(2009):12-17.

Woodruff, Catherine S., and A. J. S. Jr. Intradaily PriceVolume Adjustments of NYSE Stocks to Unexpected Earnings. *Journal of Finance* 43.2(2012):467-491.

APPENDIX

A-1 Appendix

Figure 8: Word2Vec Estimation Performance (Monthly) – 10-dimension model

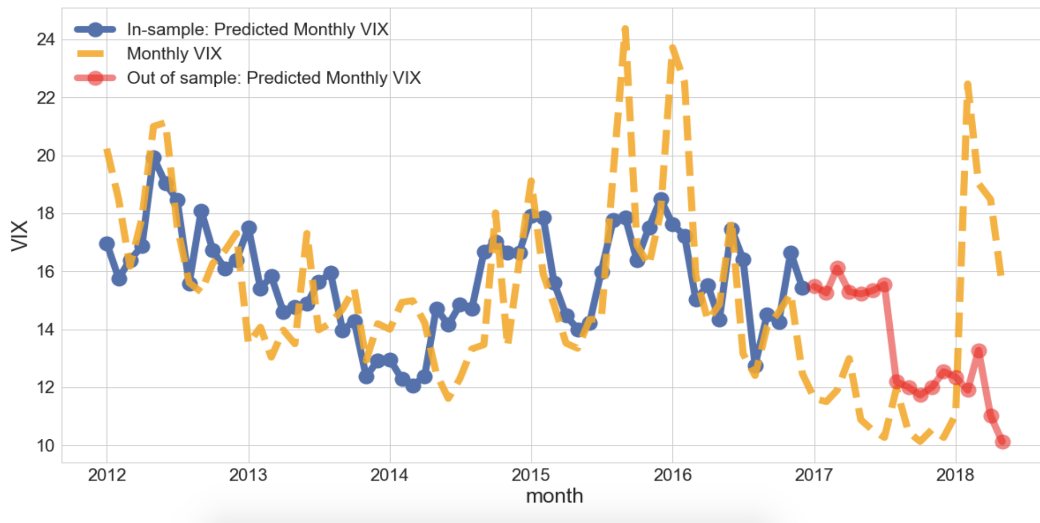


Figure 9: Word2Vec Estimation Performance (Monthly Percent Change) – 10-dimension model