The University of Chicago

MACS 30200 Research Paper Methods and Initial Results: News and Expected Volatility in the Stock Market

Mengchen Shi
May 8th, 2018

1 Research Question

What is the relationship between news and expected volatility in the stock market?

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

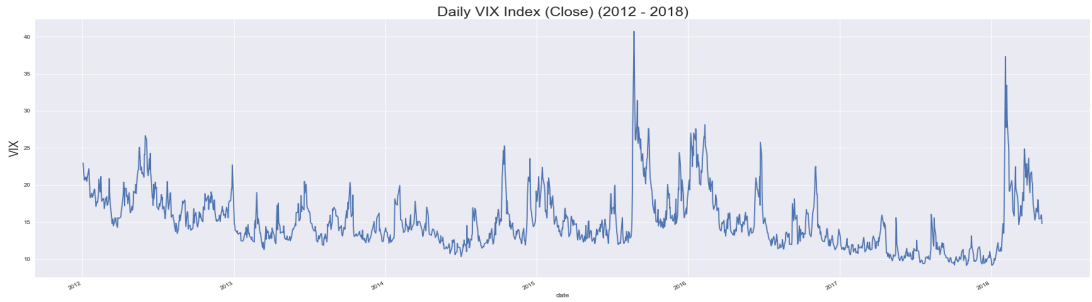


Figure 2: VIX Histogram

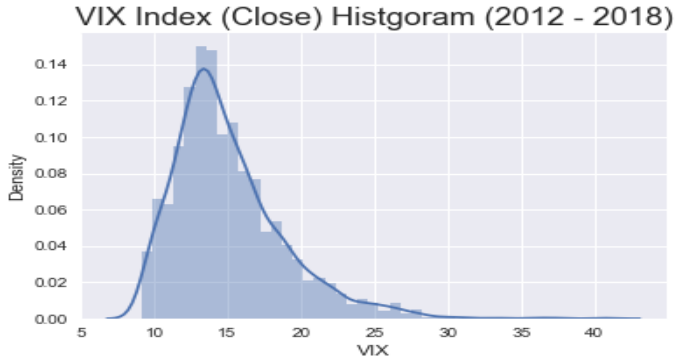


Table 1: VIX: Data Description

	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

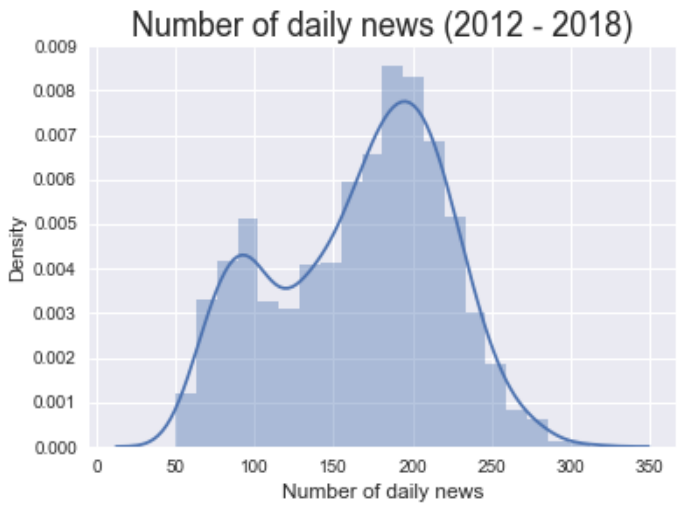


Table 2: Number of Words in Daily News

	Number of Words in News
count	1556.000000
mean	5804.785347
std	1407.404483
min	1501.000000
25%	4982.500000
50%	5917.500000
75%	6782.500000
max	9942.000000

3 Model 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

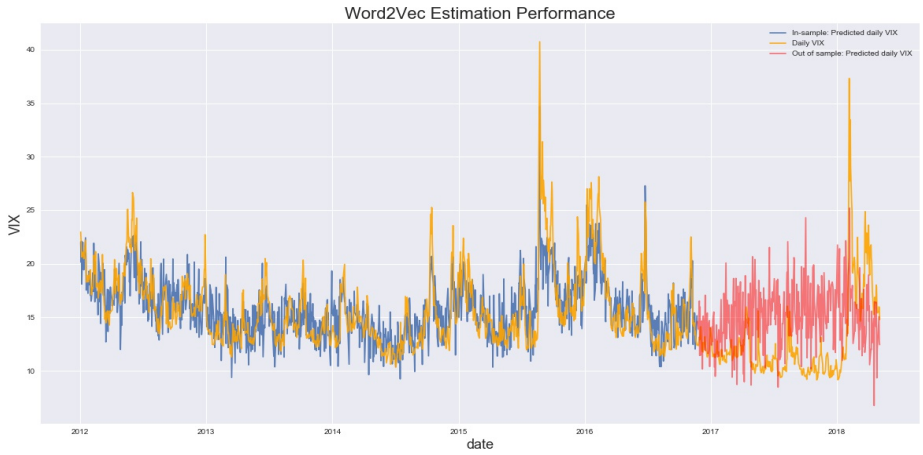
$$e_i = \hat{\beta} v_i$$

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.
The outcomes of our regression are in Figure 4.

Figure 4: Word2Vec Estimation Performance



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.

3.2 VIX Percent Change and Word2Vec OLS

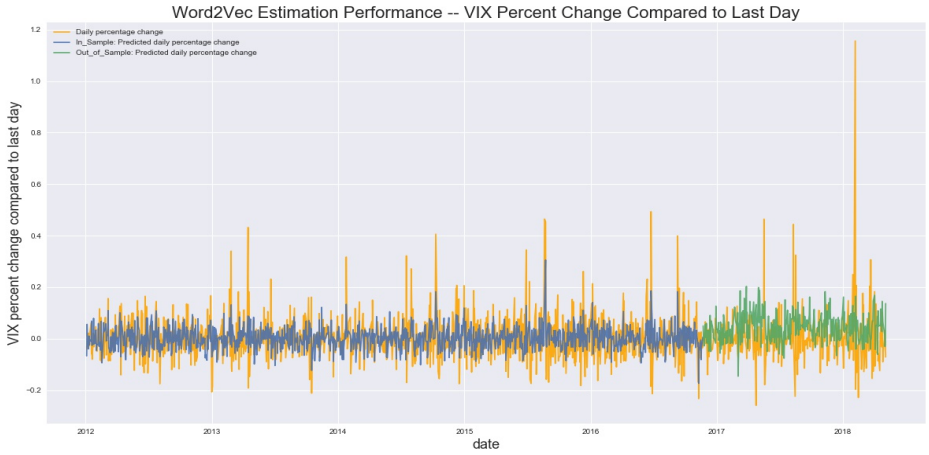
It could be possible that news not only influence absolute value of VIX, but also the percentage change of VIX. Therefore, we would like to perform an OLS estimation of the effects of Word2Vec news onto the percent change of VIX index everyday:

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

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

The outcomes of our regression are in Figure 5.

Figure 5: Word2Vec Estimation Performance – VIX Percent Change Compared to Last Day



Unfortunately, both in-sample and out-of-sample fitting perform poorly.

4 Next steps

There are lots of potential methods to try in the next steps.

4.1 Bag-of-words Model

Bag-of-words model is another commonly used model in NLP. It decomposes any string of text as the instances of certain words. In other words, in a bag-of-words model, there exists a feature for each unique word among all the observations, where

each feature signals the number of instances of the word in the particular observation. We can try this model to fit the news as well, and do a horse racing between it and Word2Vec.

4.2 Time Lag

It is possible that in influence of news on VIX has a lag effect. We can try to add some lag variables into the model.

4.3 Sentiment Analysis

There are lots of sentiment analysis methods in NLP. It is worth a try to extract sentiment in news and find relationship between sentiment and VIX.

4.4 More Data

Last but not the least, we can absolutely scrape more news from various websites such as New York Times and Financial Times. More data can improve the model by erasing opinion biases as well as by covering more information.