News and Expected Volatility in the Stock Market

Mengchen Shi
Computational Social Science, University of Chicago

Background & Aims

VIX is a measure of this expectation in the stock market. It was commonly used to inform investment strategy since 1980s. Information obtained from news influence investors' behaviors as well.

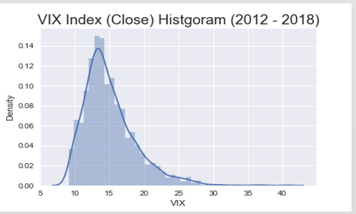
Several researchers analyzed news using textural analysis methods and have discovered relationship between the news and the stock market. Some of them successfully used sentiment, readability, and topics of news to predict stock returns, trading volume and stock volatility.

This research aims to predict VIX with daily news using Natural Language Processing (NLP). Specifically, the aim is to predict VIX and VIX percent change by news context represented by NLP models

Data

News articles are scraped from Wall Street Journal website from January 1, 2012 to May, 2018.
VIX data are downloaded from Chicago Board Options Exchange website.

	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



VIX, VIX Percentage Change and Word2Vec OLS Regression Model

Google's Word2Vec model treats words as an N-vectors of meaning. Words can therefore be added up, reducing the total number of features to whatever number of dimensions is desired.

News on each day into a 300-dim vector based on the average Word2Vec word in the news.

$$h_j = \frac{1}{N_j}\sum_i^{N_j} v_i$$

where N_j is the number of words in the j_{th} news.

The effect of any word can be approximated by: $e_i = \hat{\beta} v_i$

Perform an OLS estimation of the effects of Word2Vec news onto the VIX index. The model in-sample:

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

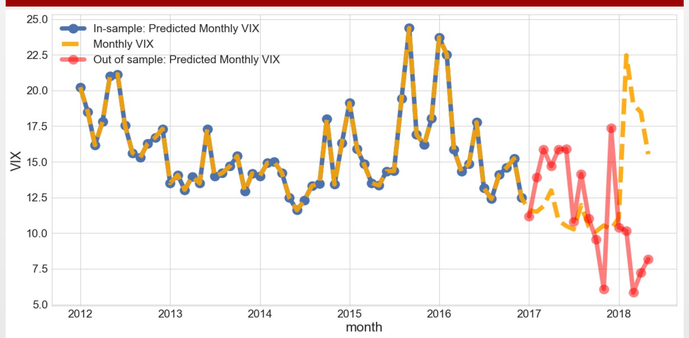
where t is each day, and X_t is the Word2Vec representation for news.

It could be possible that news influence the percent change of VIX. Perform an OLS estimation of the effects of Word2Vec news onto the percent change of VIX index:

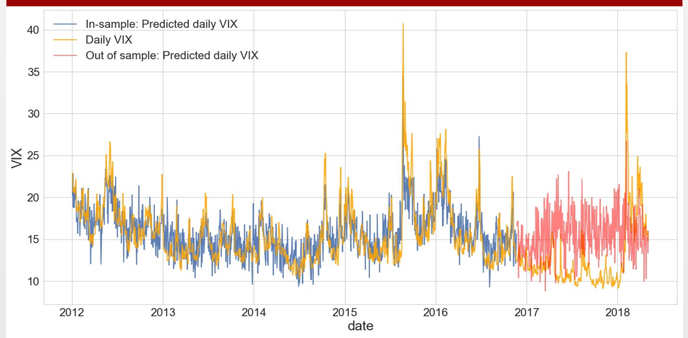
$$VIX_t - VIX_{t-1} = \beta X_t + \epsilon$$

Results

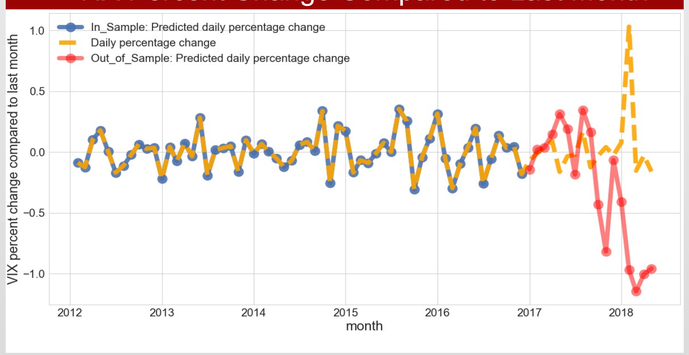
Word2Vec Estimation Performance (Monthly)



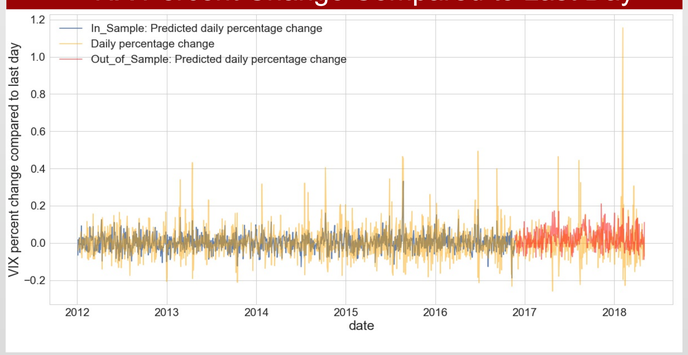
Word2Vec Estimation Performance (Daily)



Word2Vec Estimation Performance -- VIX Percent Change Compared to Last Month



Word2Vec Estimation Performance -- VIX Percent Change Compared to Last Day



Results

We estimate the effects of certain words by taking their vector representation and multiplying by beta. This captures the average effect of a word on the volatility levels.

Top 200 words that influence volatility the most

Limitations & Future Work

The predictions suffer from overfitting problems and perform not very well in the out-of-sample data. It can be fixed by adjusting parameters such as number of dimensions.

The model is like a black-box that it is hard to extract conclusions from it. To construct more models containing sentiments and topics of news can help understand how the content of news influence the volatility of the stock market.

More news data from different news websites should be included to improve the model by erasing opinion biases and covering more information.

Special Thanks

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

Please scan the QR code in the upper right corner of this poster for Github repository of this study.