The analysis of relationship between US stock market and world news

Chulwoo Kim, U01372124

SUMMARY

Since US stock market is one of the biggest areas in the world, various factors affect US stock market such as foreign exchange rate, interest rate, influential people's speech, national news, world news, and other markets. Thus, we can have questions reasonably; How much does stock market relate to the daily news?, Is it possible that people predict stock market based on daily news?

Of course, world news will not have the greatest impact on US stock market. However, experiments show us there are slight effects between them.

I. INTRODUCTION

Every countries' stock market is affected by various issues. In most cases, they are affected by domestic issues but the stock market of countries with a large economy is likely to be more sensitive to international issues. In particular, US economy responds sensitively to the issues related to the Middle East, North Korea, and Europe. Therefore, it is important to analyze how it is actually affected. I tried various experiments to confirm this hypothesis as well as each data analysis using TF-IDF (Term frequency and inverse document frequency), sentiment analysis, and machine learning algorithm. First, exploratory data analysis was conducted for each data to understand the data. I used the moving average line to determine whether stock price data has a pattern and scatter chart to check outliers. Second, in order to confirm the relationship between stock market and world news, I did experiment to compare basic and advanced models. To be specific, basic and advanced models are separated by the way to vectorize word. Therefore, basic models refer to using word frequency and advanced models mean TF-IDF, N-gram, and sentiment analysis. The vectorized words were used as features to classify stock price up and down as a label. These models were commonly learned by three algorithms; logistic regression, random forest, and support vector machine, and the results were evaluated via accuracy and AUC (Area Under Curve). Lastly, I designed the experiment which uses another market to compare with the stock market. There are various markets, but I chose the foreign exchange market because it is also affected by global issues and has different patterns from U.S. stock market.

II. METHODS

A. Word frequency

A common task in text mining is to look at word frequencies and to compare frequencies across different texts. Analyzing entire news, we can observe the following graph.

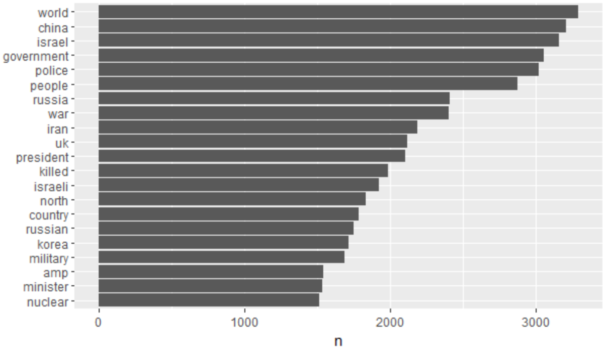

Figure 1: Top word counts except stop words

B. TF-IDF

The term frequency (TF) is one measure of how important a word or how frequently a word occurs in a document. What I mean is that we might take the approach of adding words like “the”, “is”, “of”, and so forth to a list of stop words and removing them before analysis, but it is possible that some of these words might be more important in some documents than others. A list of stop words is not a very sophisticated approach to adjusting term frequency for commonly used words. Another approach is to look at a term’s inverse document frequency (IDF), which decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents. This can be combined with term frequency to calculate a term’s TF-IDF, the frequency of a term adjusted for how rarely it is used. The inverse document frequency for any given term is defined as

$$\text{idf(term)} = \ln\left(\frac{n_{documents}}{n_{documents\ containing\ term}}\right)$$

We can get the following graph if we apply for the news data by year. This graph shows a different result from the figure 1 (word frequency graph), which means that the importance of the word depends on the application method.

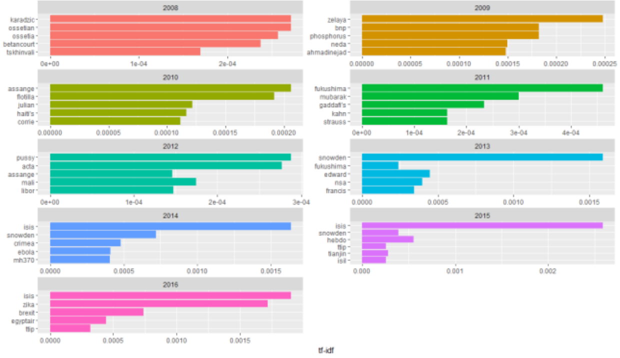

Figure 2: Top 5 words by year using TF-IDF

C. Sentiment analysis

When human readers approach a text, we use our understanding of the emotional intent of words to infer whether a section of text is positive or negative, or perhaps characterized by some other more nuanced emotion like surprise or disgust. The sentiment analysis refers to text mining to approach the emotional content of text programmatically. There are the several ways to approach sentiment analysis, but I use the following three lexicons. The AFINN lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment. The Bing lexicon categorizes words in a binary fashion into positive and negative categories. The NRC lexicon categorizes words in a binary fashion (“yes”/“no”) into categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. If we draw the graph using these lexicons, we can know that most world news deals with the negative topic. Unfortunately, the fact that a topic is biased on the negative side does not help in the analysis.

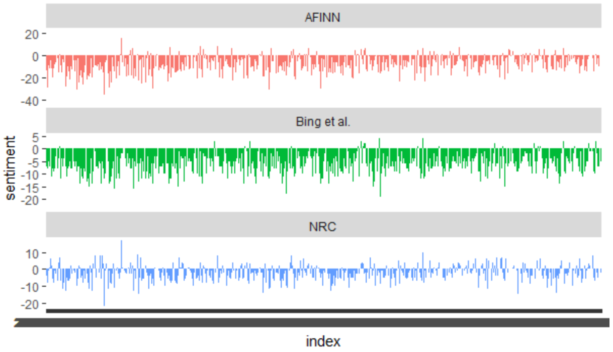

Figure 3: Sentiment analysis for world news using three lexicon

D. N-gram

When applying word frequency or TF-IDF to analysis, I have considered words as individual units. However, many interesting text analyses are based on the relationships between words, whether examining which words tend to follow others immediately, or that tend to co-occur within the same years. Therefore, I use the N-gram, which tokenizes by pairs of adjacent words rather than by individual ones and analyzes their relationship. With N-gram, words became more clear. For example, Edward and Snowden were counted as separate words, but using this method, they were calculated as a single unit and eliminated unnecessary.

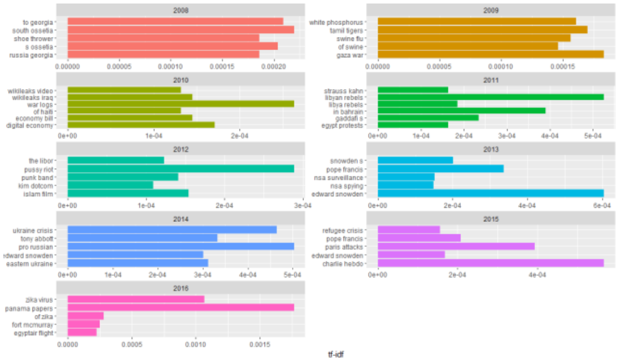

Figure 4: Top 5 words by year using N-gram

III. DISCUSSION

I have tried several experiments to claim that the world news and the stock market are related. The first experiment is to analyze a relationship between world news and US stock prices through basic and advanced models. Word frequency is used as the basic model and TF-IDF and sentiment analysis are used as the advanced model. The results show that the base model is 0.49 and the advanced model is

0.53, which means that the advanced model has better results than the basic model and ultimately the percentage of world news is related to US stock market is about 53%.

Table 1: Comparison Dow Jones Index Average with the basic and advanced models

Model	Evaluation	Logit	Random Forest	SVM
Word Frequency	Accuracy	0.49	0.49	0.49
	AUC	0.53	0.50	0.51
TF-IDF	Accuracy	0.52	0.53	0.52
	AUC	0.50	0.51	0.55
Sentiment Analysis	Accuracy	0.52	0.53	0.52
	AUC	0.50	0.47	0.50
N-gram	Accuracy	0.51	0.50	0.51
	AUC	0.51	0.51	0.54

It can have a question reasonably what kind of relationship other markets have, so I have experimented with the foreign exchange market. The results of this experiment represent similar accuracy, 0.51%. This means that the relationship between the foreign exchange market and world news is not much different from the impact on the US stock market.

Table 2: Comparison EUR/USD with the advanced models

Model	Evaluation	Logit	Random Forest	SVM
TF-IDF	Accuracy	0.51	0.51	0.48
	AUC	0.52	0.50	0.51
N-gram	Accuracy	0.50	0.47	0.49
	AUC	0.51	0.48	0.52

IV. CONCLUSION

In this paper, I did several experiments to confirm the relationship between the world news and US stock market. The reason for doing this experiment is to confirm it based on the premise that economies of countries with large markets are often responsive to global issues. I did the experiment with two markets, US stock market and the foreign exchange market. Before doing this experiment, I estimate over 60% accuracy through previous studies which analyze India stock market and tweets related to India. The results of our experiment are 53%, 51%, respectively. This means that the world news and US stock market or the foreign exchange market have less relationship than I thought. There are many reasons for this. Once the world news deals with issues around the world, its scope is too broad. Also, it is difficult to estimate the relationship because the world news is mostly composed of negative words. Lastly, I used the data for eight years, but there was still insufficient data. Nevertheless, the result of about 50% means that there is an irreducible relationship between them.

REFERENCES

1. Kokoy Siti Komariah, Carmadi Machbub, Ary S. Prihatmanto, Bong-Kee Sin, “A Study on Efficient Market Hypothesis to Predict Exchange Rate Trends Using Sentiment Analysis of Twitter Data,” *Journal of Korea Multimedia Society. 2016. Jul, 19(7): 1107-1115*.
2. Ussama Yaqub, Soon Ae Chun, Vijay Atluri and Jaideep Vaidya, “Sentiment based Analysis of Tweets during the US Presidential Elections,” *18th Annual International Conference on Digital Government Research.*