ORIE 4741 – Project Midterm Report – Portfolio of ETFs

By Indraneel Bhanap (ijb37), Sukrut Nigwekar (sn639) and Vidhisha Nakhwa (vn83)

1. Introduction

The goal of this project is to identify the degree of impact that financial news has on ETF prices, and if we can combine that information to create portfolios of ETFs, adjusted for investors' needs. The adjustment can be done based on the user's risk tolerance which is a representation of how willing the person is to invest in some high-volatility financial instrument. Normally, volatility is determined by looking at the standard deviation of the returns of an instrument but for this project we will be looking at the impact of news on the price of the instrument (in this case ETFs). We assume that most of the volatility in any instrument's price is due to incoming news which has not already been incorporated in the price. At this stage of the project we have not been able to perform the analysis for all the ETFs and are focusing on only the technology ETFs.

2. Description of the dataset

2.1 ETF Prices

The first data set contains a list of the top ten most liquid ETFs in the technology sector and their corresponding closing price and volume, each in its own column, for each trading day between January 02, 2018 and June 01, 2018. The data was sourced from Yahoo Finance. The closing price is the market price at the end of the day for each share of the ETF. Volume represents the number of ETF shares that were bought/sold during that trading day. The reason for selecting the aforementioned trading dates will be explained further in Section 2.2.

2.2 News Articles

The second data set used contained over 300,000 news articles from various sources such as CNBC, Reuters and Bloomberg. The source for this file is Kaggle. As this file only contained articles from January 02, 2018 to June 01, 2018, the ETF prices had to be constrained to this period. The articles were given in JSON format and included fields for companies mentioned in the article and the sentiment associated with each company based on the article. The sentiment had three possible values – positive, negative or none. From this dataset we had to extract what impact did an article have on the technology sector. To do this, a list of technology companies was created by looking at the holdings of the ten ETFs described above. It is important to note that the company names mentioned in the articles may be different than the company names in the ETF holdings. For example, Apple Inc. could be mentioned in the article as just Apple or as AAPL. To solve this, an exhaustive list of possible names for each company was drawn up. By referring to the exhaustive list and the company names given in the articles, it could be determined if an article was related to the technology sector. Once an article was identified as being relevant to the technology sector, the sentiment data was extracted and tabulated. The table had five columns – date, number of positive articles, number of negative articles, number of none articles and total number of articles. This allowed an understanding of the daily picture of the sentiment of news articles for the technology sector.

2.3 Google Trends

The third and final data set is a measure of the interest in the topic “technology”, sourced from Google Trends. This was extracted for dates from January 02, 2018 to June 01, 2018. The values range from 0 to 100, with 100 representing peak interest. The idea was to check if interest and ETF price are somewhat correlated

These three datasets were combined to form the input space. Additional features, which include the previous day’s trading volume, the price change, and the change in the trading volume, were added to the dataset. As such, there were no missing or corrupted data points.

3. Preventing overfitting and underfitting

To prevent overfitting, the primary strategy that will be employed is to have a dataset with a large enough number of examples. In addition, cross-validation techniques such as the k-fold cross validation technique, which splits the dataset into ‘k’ number of train/test splits can also be used to ensure that the model will be able to generalize well on data not in the training model.

4. Analysis Performed

Based on the dataset created (described in the previous section), a time series plot of the closing price and the fraction of negative articles was generated. In addition, a second plot shows the change in the ETF price change overlaid atop a scatter plot of the Google Trends “interest” data in the topic “technology”. These plots show the feasibility of the project’s scope and goal.

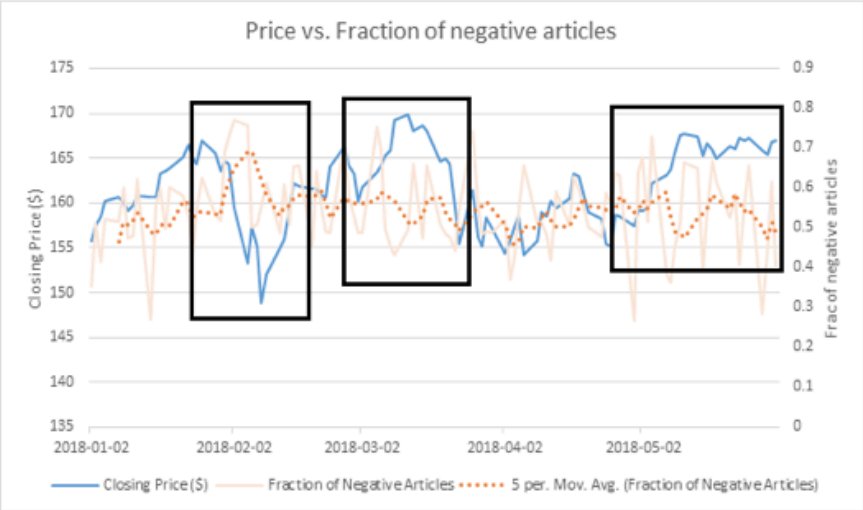

Figure 1: ETF Price vs. Fraction of negative articles

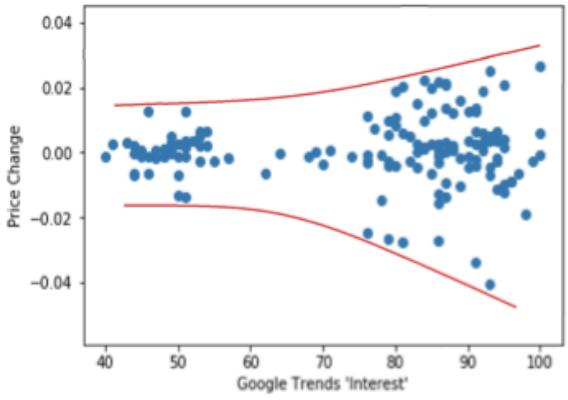

Figure 2: ETF Price Change vs. Google Trends 'Interest'

4.1 Effect of negative articles on ETF price

Figure 1 shows a strong negative correlation between 5 day moving average of fraction of negative articles (defined as total number of negative articles on a day / total number of articles on a day). The

correlation seems particularly strong during the periods highlighted. Moving average was used because it reduced the fluctuations in the fraction of negative articles and allowed us to focus on the trend.

4.2 Effect of Google Trends 'Interest' on ETF Price Change

A plot between Google Trends 'Interest' and the ETF closing prices show very little correlation. However, Figure 2 shows there is some correlation between Google Trends 'Interest' and changes in the ETF closing prices. The hypothesis therefore, is that when the Google Trends 'Interest' is high, drastic changes in the ETF prices can be expected. This hypothesis will be tested going further.

5. Preliminary Model

The preliminary model fits a regression model to get a sense of whether the features being explored make directional sense. Below are the details of the model:

Input features: Previous Day's closing price, Previous Day's volume, No. of negative sentiment articles, Total number of news articles, Google Trends 'Interest', Previous Day's price change

Output Space: Changes in the ETF closing price

Feature scaling on input features was followed by fitting a linear regression model. Directionally, the number of negative articles had a negative coefficient indicating a negative effect on price change, while total number of articles, Google Trends 'Interest' and previous day's price change had a positive coefficient.

The corresponding R-squared value obtained was 0.05 which does indicate under-fitting. What can be said in the current state, with the available data in this feature space, is that the model can only explain 5% of the variability in the predicted change in ETF prices. This calls for using more features, performing more rigorous feature transformation, and fitting a more complex model to improve the fit.

6. Next Steps

The sentiment analysis which was included in the data was restrictive as it only featured nominal values - {positive, negative, none}. Moving forward, the idea is to use a better sentiment analyzer that yields numeric values ranging from -1 to 1 or determines the sentiment of the article as a whole and not just for each company. Another improvement could be to look for words which are highly relevant to technology sector such as cloud computing, 5G etc. which could add depth to the sentiment analysis. As the original goal was to somehow create a portfolio of ETFs, the focus on ETFs in the technology sector will be broadened to include other sectors as well.

7. Data Sources

1. https://www.kaggle.com/jeet2016/us-financial-news-articles - For the news articles in JSON format.
2. https://etfdb.com/etfs/industry/broad-technology/ - For list of technology ETFs.
3. https://finance.yahoo.com/ - For ETF price and volume data.
4. https://trends.google.com/trends/ - For Google Trends data.