ORIE 4741 – Project Final Report – Portfolio of ETFs

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

Introduction

The goal of this project is to identify the degree of impact that financial news has on individual financial sectors, by considering the behavior of the most-traded ETF in each, as a proxy. Based on this, the project explores whether that information can be incorporated to provide investors with an additional measure of risk, if they are keen on exposure in a sector. Normally, volatility is determined by looking at the standard deviation of the returns of an instrument. For this project, the degree of correlation of news data with the price of the instrument (in this case ETFs), will give a measure of the correlation between news data and the sector. An assumption that will be considered is that most of the volatility in any instrument's price is due to incoming news which has not already been incorporated in the price.

Data Collection

Sectors chosen, Representative ETFs and ETF Holdings

For this project we are using 10 sectors namely – Consumer Discretionary, Consumer Staples, Energy, Financials, Healthcare, Industrials, Materials, Technology, Telecom and Utilities. For each sector there are multiple ETFs which track the sector, however we only looked at the ETFs with the highest average volume. These ETFs with highest average volume for each sector have been referred to as representative ETFs in the rest of the report. Now each of these representative ETFs have positions in multitudes of stocks and the companies whose stock is included in the ETF are known as holdings. The names of these companies included in the holdings are important to identify articles relevant to the representative ETF and thus the sector.

Based on the ETF data we obtained the following are the representative ETFs:

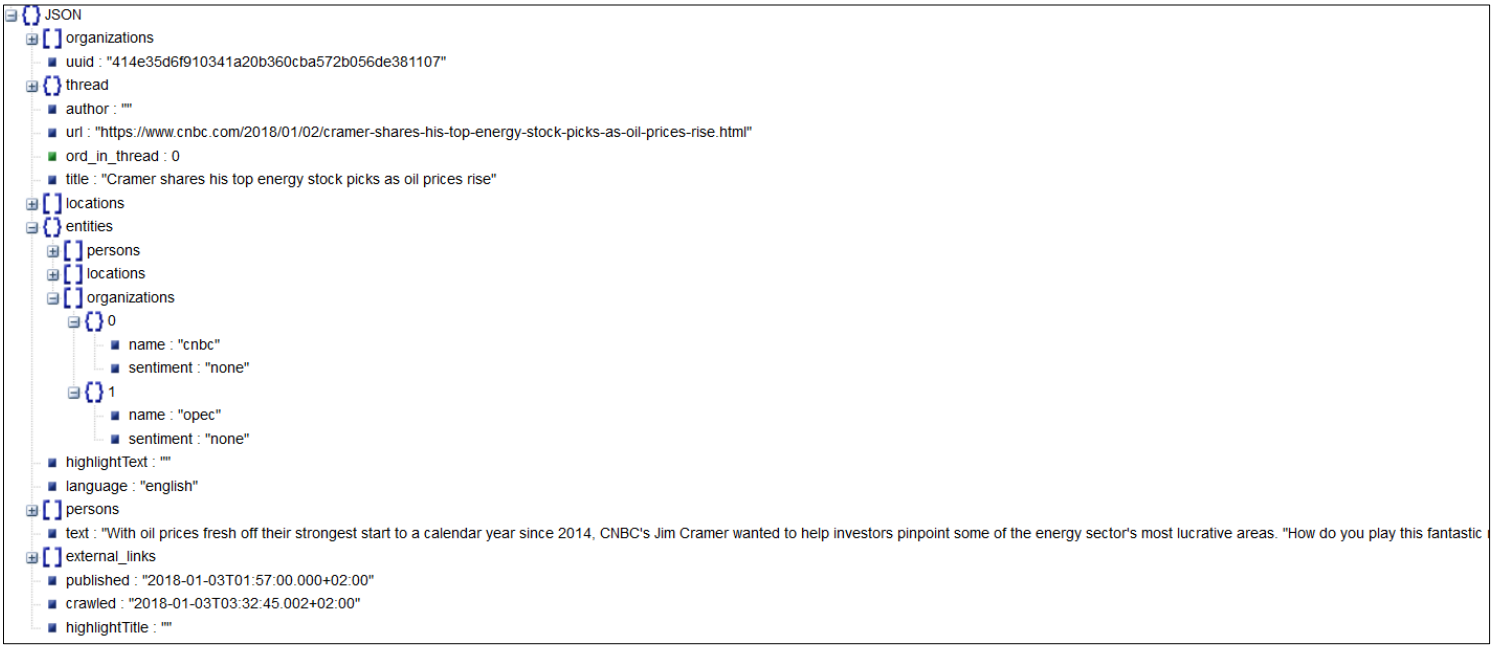
Sector	Representative ETF
Consumer Discretionary	XRT
Consumer Staples	XLP
Energy	XOP
Financials	XLF
Healthcare	XLV
Industrials	XLI
Materials	GDX
Technology	QQQ
Telecom	IYZ
Utilities	XLU

Sample of holdings for Consumer Discretionary - XRT

Apple Inc.	Microsoft Corp	Amazon.com Inc
PayPal Holdings Inc	Texas Instruments Inc	T-Mobile US Inc

Obtained data, its source and description

1. ETF prices
   The closing price and volume is obtained from Yahoo Finance and is for each trading day between January 02, 2018 and June 01, 2018. 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.
2. News articles
   The news dataset contained over 300,000 news articles from various sources such as CNBC, Reuters and Bloomberg and was obtained via Kaggle. As this dataset 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.
   An example of the JSON file is:



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.

Dataset Creation

VADER Sentiment Analyzer

VADER is a sentiment analyzer based on lexicons of sentiment-related words. In this approach, each of the words in the lexicon is rated as to whether it is positive or negative, and in many cases, how positive or negative. VADER analyses a piece of text by checking if any of the words in the text are present in the lexicon. VADER doesn’t just do simple matching between the words in the text and in its lexicon. It also considers

certain things about the way the words are written as well as their context. The result is a composite score between -1 and +1 with -1 being most negative and +1 being most positive. Given the composite score, we have used [-1, -0.2] as the range for negative articles, [-0.2, +0.2] for neutral articles and [+0.2, +1] for positive articles.

Creating the dataset for each sector

Using the datasets described above we created data tables for each of the sectors individually. The following steps were taken for the same:

1. From the news data we first identify articles relevant to each sector. To do this we checked if the organizations section of the JSON file mentions any company name belonging to the holdings of the representative ETF. 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 sector.
2. For each article was identified as relevant, the text section of the JSON file was passed to the VADER sentiment analyzer and the result was noted. The result along with the date of publishing was stored in the data table.
3. The data table was stored as a csv for future use.

Once the data was processed for each sector, some additional features were added. The additional features were autoregressive ones based on the news data. The first few lines of the final dataset for consumer discretionary sector looked like:

	date	positive	negative	none	total_articles	Close	Volume	returns	returns_1	volume_1	positive_1	negative_1	none_1	positive_7	negative_7	none_7	returns_class
8	2018-01-09	19	1	0	20	45.810001	7205200.0	-0.010583	0.004121	6099800.0	14	1	1	94	12	6	2
9	2018-01-10	13	4	3	20	46.139999	3302300.0	0.007204	-0.010583	7205200.0	19	1	0	87	13	8	1
10	2018-01-11	24	4	5	33	47.279999	8989100.0	0.024707	0.007204	3302300.0	13	4	3	87	14	11	3
11	2018-01-12	39	5	5	49	47.820000	8708100.0	0.011421	0.024707	8989100.0	24	4	5	111	15	14	4
12	2018-01-13	7	0	1	8	47.820000	8708100.0	0.000000	0.011421	8708100.0	39	5	5	116	15	15	3

The column names like returns_1 or positive_1 mean the previous days value and columns like returns_7 or positive_7 are the total trailing past 7 days values. Returns mean the change in price divided by the previous day's price.

Preliminary Analysis

This section has been taken from the midterm report just to give an idea about our preliminary analysis. 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 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. 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. More drastic price changes saw higher 'Interest'. However, due to challenges in obtaining the data at scale for 10 sectors in an automated manner – we decided to not use this filed in this round.

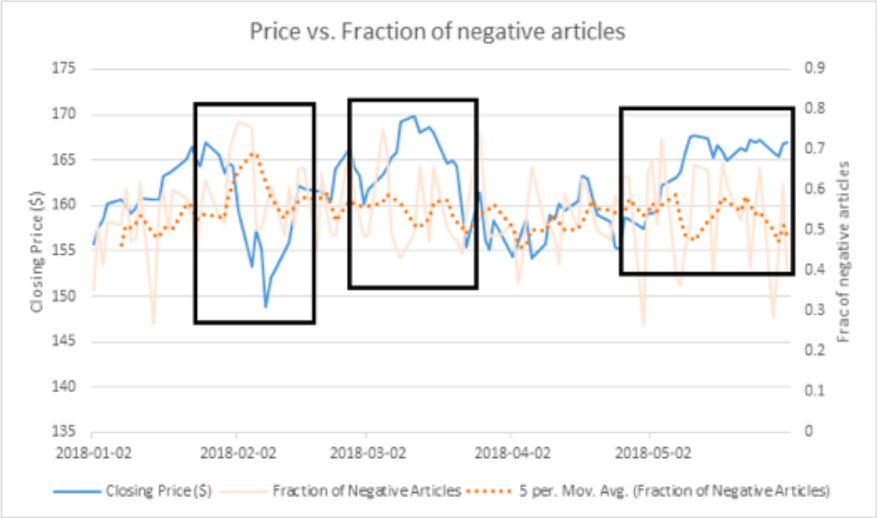


Figure 1: ETF Price vs. Fraction of negative articles

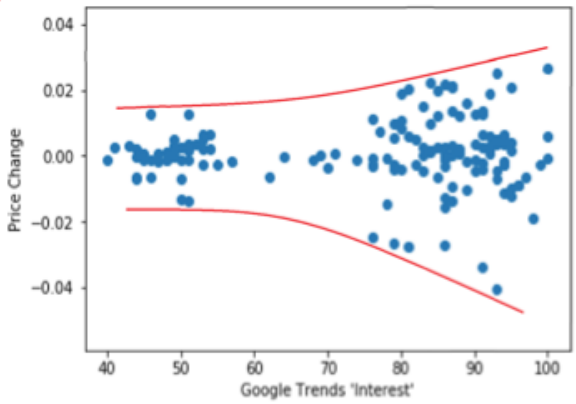


Figure 2: ETF Price Change vs. Google Trends 'Interest'

Modelling

Model fitting

The overall idea of the model is to use the available ETF prices and news data to predict the expected returns. Modeling has been done at sector level, to arrive at how different sectors respond to news. The problem has been solved as a multi-class classification problem. Based on the historical variance in the returns for individual sector, 5 classes of returns are created. These classes may be different for different sectors depending on the intrinsic level of volatility that each sector has observed historically.

Feature Engineering

Creating the final dataset involved combining the sentiment analysis of news articles with the closing ETF prices data. Since ETF prices for weekends and public holidays are not available, the prices for the previous day are assumed to be held constant.

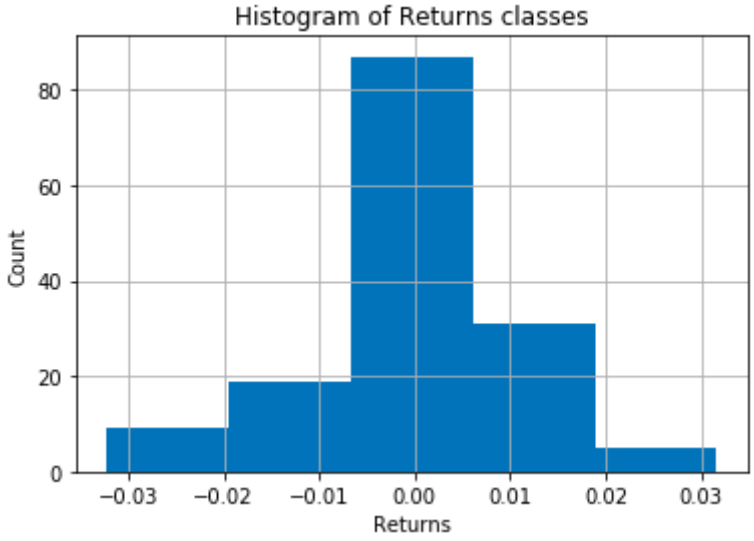
Features for Auto-Regressive Modeling:

For predicting the 'Returns class' of the next day, two models with following input features were used:

- Case 1: Returns of previous day, Volume of previous day
- Case 2: Returns of previous day, Volume of previous day, Number of positive, negative & neutral sentiment articles on the previous day, Total Trailing Number of positive, negative & neutral sentiment articles in the past 7 days

Up-sampling for creating a balanced dataset:

The figure below shows the classes of Returns for a certain sector. In order to create a balanced dataset, we up-sampled the 4 minority classes to match the size of the majority class. For example, in this case, for each of class 1, 2, 4, and 5 we sampled with replacement till we had 85 observations in each class.

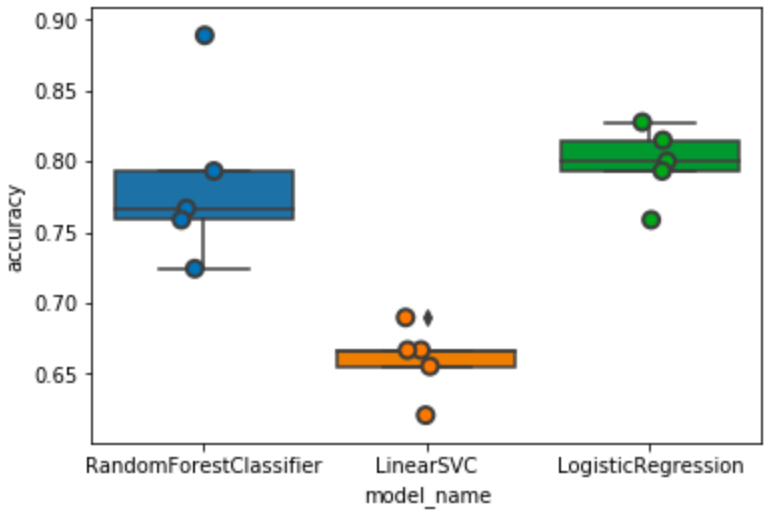


Models Considered

The following models were used to solve the classification problem:

- Random Forest
- SVM
- Logistic Regression

Cross Validation technique was used to arrive at the best fitting model, which was defined to be the one that generated the highest test-set accuracy. The figure below shows a sample of the cross-validation results for one sector.



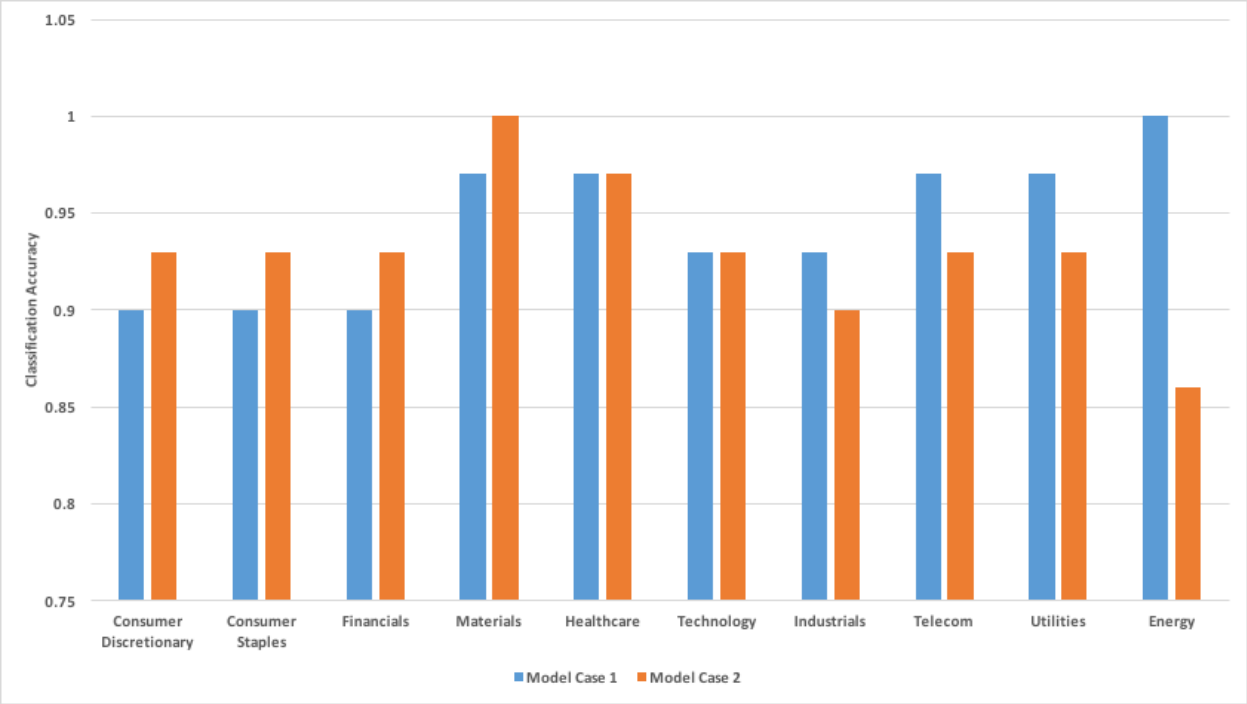
We looked at the precision, recall and F1 score of the test set to ensure that the model accuracies are not misleading. Below is a sample:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2
1	1.00	0.43	0.60	7
2	0.79	1.00	0.88	15
3	1.00	1.00	1.00	4
4	1.00	1.00	1.00	1
accuracy			0.86	29
macro avg	0.96	0.89	0.90	29
weighted avg	0.89	0.86	0.84	29

Results

Figure below has the test set results on the two models.

- Case 1: Returns of previous day, Volume of previous day
- Case 2: Returns of previous day, Volume of previous day, Number of positive, negative & neutral sentiment articles on the previous day, Total Trailing Number of positive, negative & neutral sentiment articles in the past 7 days



The available news data incorporation of sentiment data, in the feature space, showed an improvement in the prediction, and thus the test set accuracy, for the following sectors: Consumer Discretionary, Consumer Staples, Financials and Materials'. On the other hand, it did not seem to show any improvement on the other sectors. Further analysis on the drop in accuracies for 'Telecom', 'Utilities' and 'Energy' needs to be done on the lines of whether the correlation between prices and news articles is high enough to lead to such results.

Weapons of Math Destruction & Fairness

Modeling the effect of news data is not a weapon of math destruction for the following reasons:

- The outcome is easily measurable. The returns of the next day can easily give a sense of how good or bad the prediction was
- Self-fulfilling feedback loop cannot be created in this case because the prices of the previous day are always made available to the model.

Fairness is not important in this case because the data does not include any protected attributes (or its covariates) and thus the model is not biased in any way for this application.

Further Work

1. Determine which features helped in improving the accuracy for the consumer discretionary, consumer staples, financials and materials sectors.
2. Determine why the accuracy dropped for industrials, telecom, utilities and energy sectors after including news sentiment data. An explanation we thought of is that the news sentiment data is already incorporated in the price and volume for these sectors and hence there is high correlation between price and volume and news sentiment. This high correlation might cause the model to give much higher weights to price and volume which in turn causes the accuracy to be adversely affected.
3. Use the model coefficients as an alternative to volatility (as described in the introduction) and performing ETF portfolio optimization.

References

1. https://www.kaggle.com/jeet2016/us-financial-news-articles - For the news articles in JSON format.
2. https://etfdb.com/etfs/ - For list of ETFs.
3. https://finance.yahoo.com/ - For ETF price and volume data.
4. https://trends.google.com/trends/ - For Google Trends data.
5. **Scikit-learn**: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.
6. Hutto, C.J. & Gilbert, E.E. (2014). **VADER**: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.