Tiny Data Science Project: Impact of News Headlines on Stock Indices

Hisho Rajanathan

Introduction

Analysis Domain

A stock market index is a hypothetical portfolio of investment holdings which represents a segment of the financial market [1]. A number of factors influence the stock market from interest rates, economic growth and behavioural economics [2]. There is a possibility of 'beating' the market if an investor has superior information [3]. According to the Efficient Market Hypothesis, the share price reflects all information and neither fundamental nor technical analysis can generate excess returns. [4].

Motivation

As conventional methods of trying to beat the stock market are used by a number of traders, different methods will need to be adopted to have a chance of outperforming the stock market to produce excess returns. Traditional approaches of technical and fundamental analysis may give investors an insight into long term investing, however using an unconventional approach of applying machine learning algorithms may improve the chances of generating excess returns. Investigating the impact of news headlines on stock indices or even incorporating this into a trading strategy may allow an investor to predict price mismatches or market movements.

Data, Questions, Plan

Data Source

News articles were obtained using The Guardians API from 01/01/2004 to 05/12/2019 [5][6]. The news articles include UK, Business and World News which will help predict global indices.

The Guardian is a UK leading media organisation [7], hence there could be a tendency for bias towards news articles related to the UK. Furthermore, The Guardian is considered as a left-wing publisher with support towards the Labour Party and being pro EU [17]. The FTSE 100, a major UK index, consists of the 100 largest market capitalisation companies, will be investigated [8]. The index values were obtained from investing.com from 01/01/2004 to 01/12/2019, using the same date range as the news articles [9]. In addition, the S&P 500 index will be investigated to understand if headlines published by a UK media organisation could have an impact on the index. The S&P 500 consists of 500 leading US companies which covers approximately 80% of market capitalisation [10]. S&P 500 index values were obtained from Yahoo! Finance using yfinance and Pandas Data Reader Library [12][13].

Research Question

The objective is to identify the impact of news headlines on stock market indices.

The following research questions this paper aims to answer:

- Are the news headlines predominantly positive, negative or neutral?
- Can news headlines be classified into topics reliably?
- Is there any correlation between news headlines and stock index price?
- The news articles have been published by The Guardian, but do these news articles have an effect on other stock indices?
- Can news headlines be used to predict the stock index values?

Plan

In order to answer the questions above the following plan has been developed:

1. Connect to The Guardian API and extract data
2. Clean data: to prepare useful and robust data to work with.
3. Create new features e.g.
    a. Sentiment analysis
4. Exploratory data analysis and modelling e.g.

a. LDA (Latent Dirichlet Allocation)

5. Predict stock price

Findings and Reflection

Findings

Initial Exploratory Analysis

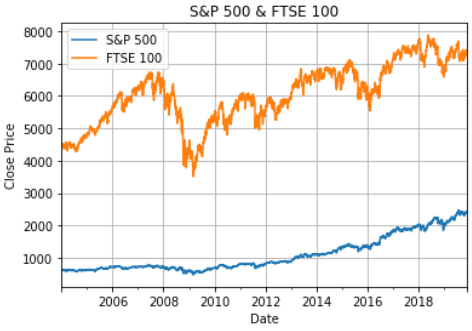

Figure 1 *FTSE 100 and S&P 500 Close Price*

The movement of the FTSE 100 and S&P 500 prices over the years is important to understand how words in news headlines change. This can be split into four distinct periods:

- 2004 – 2007
- 2008 – 2009
- 2010 – 2015
- 2016 – present

Figure 2 *Word clouds of headlines between 2004 and 2019*

There was a significant drop in the FTSE 100 closing price in 2007 due to financial crisis which only started to recover in 2009. With these distinct periods we can identify the most common words from the word clouds. For example, in the period 2016 to present, major events occurred such as Trump becoming president of USA, Brexit in the UK, Facebook appearing in the news due to privacy. 2008 - 2009 when Obama was president, in addition to a lot of focus on the economy crashing with words such as 'bank', 'fall', 'market', 'cut' and 'rating'.

Sentiment analysis and topic modelling was applied to identify if headlines had an impact on the stock indices.

Headline Sentiment

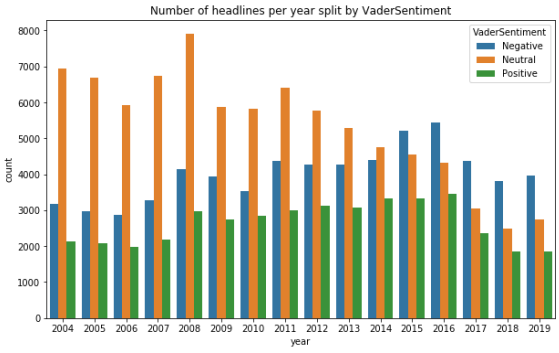

Figure 3 *Count of number of headlines per year and sentiment of headlines*

The Guardian API allows us to extract the body of the headlines, however sentiment analysis was not applied due to it being computationally expensive. According to VADER, a lexicon and rule-based sentiment analysis tool [14], the Guardian publishes a number of neutral headlines. However, more recently there has been a high number of negative articles being published.

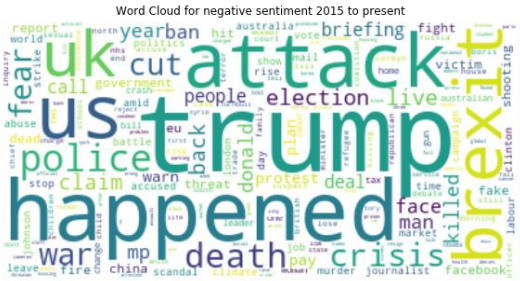

Figure 4 *Word Cloud showing negative sentiment from 2015 to present*

From 2015 onwards, we can see that the most common words were 'Brexit', 'death', 'killed', 'fear' and 'crisis'. The EU referendum happened on 23/06/2016 [15], which created a lot of uncertainty in the market with a number of cuts to jobs as well as

a possibility of another financial crisis. [16]. The Guardian having a left-wing stance, explains why more recently there has been a number of negative articles with events such as Brexit and the Conservative party governing parliament.

Topic Modelling

Manually classifying headlines into topics is a cumbersome task, however there is a more sophisticated approach using an unsupervised algorithm called topic modelling. Topic modelling is a technique which extracts topics from large text data [18]. LDA (Latent Dirichlet Allocation) from the Gensim package will be used to classify headlines into topics. As this is an unsupervised approach the reliability of topic modelling can be questioned and depends on the quality of pre-processing text data.

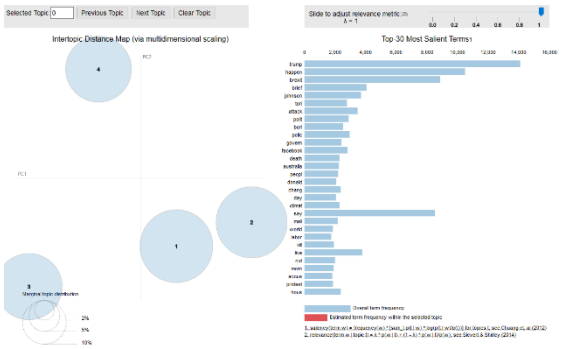

Figure 5 *LDA Topic Modelling of Headlines*

Four topics have been created using LDA, however there is no general topic name that can be assigned to each of the topics. Another approach could be classifying topics on the body of the text instead of the headline.

Correlation and Effect on other indices

In order to determine if news headline sentiment will have an impact on the stock closing price, Spearman's correlation has been applied to the dataset. However, there is a weak correlation between the headline sentiment and the FTSE 100 close price; the same can be seen with S&P 500 close price.

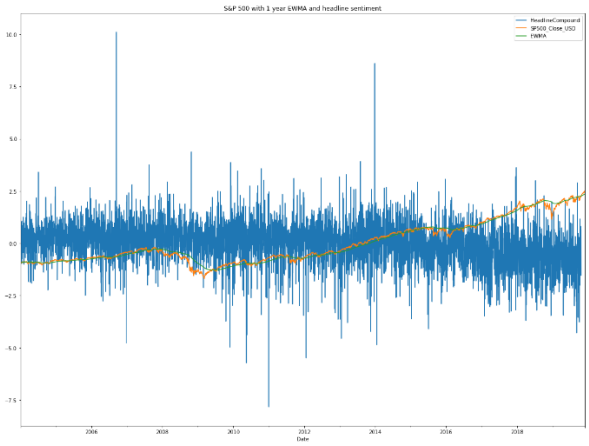

Figure 6 *Effect of Headline Sentiment on S&P 500*

Plotting the headline sentiment against the S&P 500 close price, it can be difficult to assess whether the news headlines have an effect on the S&P 500. There is a lot of noise that can be seen and also there are days where there is a large negative or positive sentiment and the stock price does not move in the way that is expected.

Predicting stock prices

Since the introduction of machine learning and artificial intelligence, there have been a number of hedge funds implementing machine learning algorithms to identify patterns in stock prices that an individual would struggle to detect [19]. Stock indices prices can be predicted by using Multiple Linear Regression and Random Forest Regression. Both models capture the upward trend of the FTSE 100 close price, however it does not consider the sudden drops such as in 2007 due to the financial crisis.

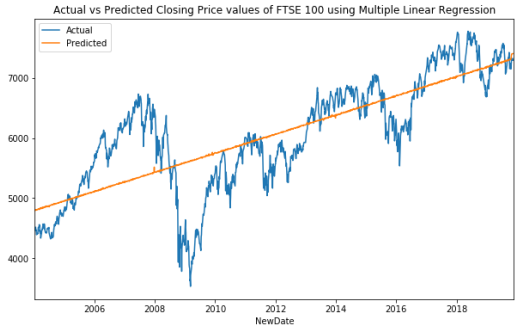

Figure 7 *Multiple Linear Regression to predict FTSE 100 price using News Headline Sentiment*

Tiny Data Science Project: Impact of News Headlines on Stock Indices

Hisho Rajanathan

Critical Reflection

To answer the question of impact of news headlines on the stock market, we do not consider other factors that may affect the stock market such as interest rates, efficient market theory or microeconomic factors. The Guardian was chosen as 5,000 API requests could be made per day, higher than other media publishers. In addition, using news articles published by other news outlets such as Financial Times, The Times as well as publishers abroad, such as The New York Times should be considered as media outlets could have a political stance affecting the headline sentiment. Recently, social media also has an impact on the stock market especially with day traders, hence sentiment analysis should be applied to social media data such as Twitter to look into the effect on the stock market.

Multiple headlines are published daily, however an average sentiment was calculated for the day. This will not capture the full extent of how the stock market may change according to the news headlines published. Moreover, daily price changes have been considered, a more accurate approach may be looking at the hourly price changes as articles can be published at any time of the day. News headlines may not have an immediate impact on the stock market and there may be a lag between when the news headlines are incorporated in the price of the stock which should be considered to provide an accurate prediction. Another approach to this, would be looking at how the body of the news articles affect the stock price instead of using the headline.

The models to predict index prices showed the general upward trend of the FTSE 100, however it did not factor in sudden drops and price surges. LSTM Neural Networks should be used to accurately predict the stock index.

References

[1] Investopedia. (2019). Understanding Market Indexes and Their Uses Helps Investors. [online] Available at: https://www.investopedia.com/terms/m/marketindex.asp [Accessed 27 Oct. 2019].

[2] Pettinger, T. (2019). Factors affecting the Stock Market | Economics Help. [online] Economicshelp.org. Available at: https://www.economicshelp.org/blog/2841/economics/factors-affecting-the-stock-market/.

[3] Investopedia. (2019). *Can Anybody Beat the Market?*. [online] Available at: https://www.investopedia.com/ask/answers/12/beating-the-market.asp.

[4] Investopedia. (2019). *Efficient Market Hypothesis (EMH) Definition*. [online] Available at: https://www.investopedia.com/terms/e/efficientmarkethypothesis.asp.

[5] Open-platform.theguardian.com. (2019). *theguardian / open platform - documentation / overview*. [online] Available at: https://open-platform.theguardian.com/documentation/.

[6] Gist. (2019). *How to use The Guardian's API to download article data for content analysis (in Python 3.x)*. [online] Available at: https://gist.github.com/dannguyen/c9cb220093ee4c12b840.

[7] the Guardian. (2019). *About Guardian Media Group*. [online] Available at: https://www.theguardian.com/gmg/2018/jul/24/about-guardian-media-group.

[8] Share.com. (2019). [online] Available at: https://www.share.com/a-guide-to-investing/before-you-start/what-is-the-ftse-100.

[9] Investing.com UK. (2019). *FTSE 100 Historical Rates - Investing.com UK*. [online] Available at: https://uk.investing.com/indices/uk-100-historical-data.

[10] Us.spindices.com. (2019). *S&P 500® - S&P Dow Jones Indices*. [online] Available at: https://us.spindices.com/indices/equity/sp-500.

[11] Investing.com. (2019). *USD GBP Historical Data - Investing.com*. [online] Available at: https://www.investing.com/currencies/usd-gbp-historical-data.

[12] PyPI. (2019). *yfinance*. [online] Available at: https://pypi.org/project/yfinance/.

[13] Programcreek.com. (2019). *pandas_datareader.data.get_data_yahoo Python Example*. [online] Available at: https://www.programcreek.com/python/example/92135/pandas_datareader.data.get_data_yahoo.

[14] GitHub. (2019). *cjhutto/vaderSentiment*. [online] Available at: https://github.com/cjhutto/vaderSentiment.

[15] Gov.uk. (2019). *EU referendum - GOV.UK*. [online] Available at: https://www.gov.uk/government/topical-events/eu-referendum.

[16] Partington, R. (2019). *Bank of England warns no-deal Brexit could trigger economic shock*. [online] the Guardian. Available at: https://www.theguardian.com/business/2019/jul/11/bank-of-england-warns-of-lending-crisis-for-eu-firms-after-no-deal-brexit.

[17] En.wikipedia.org. (2019). *The Guardian*. [online] Available at: https://en.wikipedia.org/wiki/The_Guardian#Political_stance_and_editorial_opinion.

[18] Machine Learning Plus. (2019). *Topic Modeling in Python with Gensim*. [online] Available at: https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/.

[19] Ft.com. (2019). *Why hedge fund managers are happy to let the machines take over | Financial Times*. [online] Available at: https://www.ft.com/content/338962c0-eeaf-11e9-ad1e-4367d8281195.

[20] Kapadia, S. (2019). Topic Modeling in Python: Latent Dirichlet Allocation (LDA). [online] Medium. Available at: https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0 [Accessed 30 Oct. 2019].