London School of Economics and Political Science

Masters Thesis

Transitory, Or Not Transitory? That Is The Question: Inflation Forecasting Using Twitter and Reddit

Author:

Candidate 202003825

Supervisor:

Dr. Siân Brooke

Word Count:

9,982

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in

Applied Social Data Science

in the

Department of Methodology

August 17, 2022

LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Abstract

Department of Methodology

Master of Science

Transitory, Or Not Transitory? That Is The Question: Inflation Forecasting Using Twitter and Reddit

by Candidate 202003825

This paper aims to investigate the relationship between inflation-related narratives and changes in US inflation expectations in the social media age, given both the proliferation of social media in the last decade and the lack of discussion in the literature about the impact that this exponentially increased interconnectedness between individual economic agents can have on inflation expectations. On this topic, the existing literature finds that the media do have a strong influence on consumer inflation expectations. However, there is little consideration for the impact *social* media in particular, a key driver of actions many people take within the real-world, has. Hence, I seek to add to the literature in this way by generating sentiment and attention indicators relating to inflation from posts on both Twitter and Reddit. These are then employed in similar methodologies applied, for example, in investigations into how tweets have moved certain equity and cryptocurrency markets recently, in order to generate inflation expectation forecasts. Testing various model architectures out-of-sample indicates that incorporating social media content allows for increases in forecast accuracy. As such, the metrics generated from social media data provide a new-age resource for policymakers in their quest to accurately forecast inflation in the 21st century.

Acknowledgements

The author would like to express their immense gratitude to Dr. Siân Brooke for her constant and unwavering encouragement and invaluable insight throughout the duration of this project.

Contents

List of Figures

List of Tables

List of Abbreviations

VAR	**V**ector **A**uto**R**egression
LASSO	**L**east **A**bsolute **S**hrinkage and **S**election **O**perator
SVR	**S**upport **V**ector **R**egression
XGBoost	e**X**treme **G**radient **Boost**ing
BEIR	**B**reak **E**ven **I**nflation **R**ate
GDELT	**G**lobal **D**atabase of **E**vents, **L**anguage and **T**one
GISI	**G**oogle **I**nflation **S**earch **I**ndex
VADER	**V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner
LIWC	**L**ingustic **I**nquiry and **W**ord **C**ount
OSS	**O**verall **S**entiment **S**core
OAS	**O**verall **A**ttention **S**core
AIC	**A**kaike **I**nformation **C**riteria
OLS	**O**rdinal **L**east **S**quares
RSS	**R**esidual **S**um of **S**quares
PRSS	**P**enalised **R**esidual **S**um of **S**quares
RBF	**R**adial **B**asis **F**unction
RMSE	**R**oot **M**ean **S**quared **E**rror

Chapter 1

Introduction

As we have emerged from the COVID-19 Pandemic, inflation has become a central issue of concern for consumers and policymakers alike. Acting as a regressive tax, hurting those on lower incomes more, its effects have spread far and wide, causing food, fuel, and energy prices to rocket. Given the huge effect this has on the standard of living of many, this inflationary pressure is of significant interest to policymakers.

Existing research into US inflation forecasting has typically looked to three main sources of data: the semiannual Livingston Survey, the quarterly Survey of Professional Forecasters, and the quarterly Michigan Surveys of Consumers. The Livingston Survey is the oldest continuous survey of US inflation dating back to 1946 (Fed, 2022a), summarising the forecasts of various economists in both industry and academia, and is conducted bi-annually. The Michigan Surveys of Consumers consist in a minimum of 500 phone interviews with a random sample of households, asking each one 50 core questions. These households are asked to provide forecasts within three key areas: personal finance, business conditions, and buying conditions. Example questions relating strictly to inflation include 'During the next 12 months, do you think that prices, in general, will go up, or go down, or stay where they are now?'. The surveys are conducted every month. The Survey of Professional Forecasters is the oldest quarterly survey of macroeconomic forecasts in the United States, with participants mainly constituted by professional forecasters within industry (Fed, 2022b). This is beneficial as the participants make forecasts for a living, thus incentivising them to maximise the accuracy of their predictions. However, these surveys are costly in various ways. Firstly, surveys in general are costly in purely practical terms. However, more importantly, in the context of policymaking, it could be argued that there is a time-related cost as the information is insufficiently frequent. The fact that they occur only quarterly or semiannually means that they may lack granularity, such that policymakers are not able to react quickly enough to rapid changes in the economic environment.

Evidence for the potential inadequacy of these surveys can arguably be seen from the lack of reactivity from policymakers across the world in responding to the emerging inflationary pressures in 2021, instead choosing to constantly try to reassure businesses and households that such pressures were merely transitory (Giles, 2021; Politi and Smith, 2021), something which unfortunately has not come to pass. The consequences of this inability of policymakers to deal with these pressures have been devastating and look set to worsen. Significant increases in the prices of essential goods have persisted throughout 2022, leading to rocketing costs for consumers and consequently one of the worst cost-of-living crises in the UK in recent memory (Francis-Devine et al., 2022), for example. As such, these established sources of survey data could be considered as outdated and we should therefore look to supplement them with more modern, higher-frequency real-time data. This is especially the case given the utility we have seen this kind of data, such as social media data, contribute to forecasts in similar contexts, such as financial markets. This contribution seems to make sense in light of recent events, most prominently the surge in involvement of retail traders in the markets for GameStop and AMC shares. The power of the conversations going on within the WallStreetBets subreddit community sent shockwaves through these markets, much to the surprise and even detriment of huge traditional investors, such as hedge funds. Therefore, it seems almost negligent to completely ignore these online voices given the great power they demonstrably hold. I believe it would be both important and extremely useful to apply this principle and investigate the influence of these voices in more conventional economic contexts, such as within the realm of inflation forecasting. As such, in this research I set out to discover whether online conversation can meaningfully effect changes in inflation expectations in the United States and whether accurate forecasts can be generated as a result.

The central research questions of this analysis are as follows:

RQ1 - Do social sentiment towards and attention to US inflation impact inflation expectations?

RQ2 - How can public expectations be operationalised?

The data used to construct the independent variables of interest consisted in social media posts, with 60,061 posts from Twitter and 24,848 posts from Reddit, which were then analysed in order to yield 365 days worth of sentiment and attention data points, from 1st June 2021 to 31st May 2022. I hypothesise that both more negative sentiment and increased attention are likely to contribute to increased inflation expectations. I then employ five model architectures to generate forecasts: Vector Autoregression (VAR), Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Regression (SVR), Random Forest Regression, and Extreme Gradient Boosting (XGBoost). The results of this analysis indicate that both social media sentiment and attention are important determinants of US inflation expectations, with models that include these factors providing better forecasts than those that exclude them.

The rest of this paper is structured as follows. Chapter II examines the existing literature both related to and within this space. Chapter III describes the data used in this research. Chapter IV outlines the data analysis strategies and methodologies that were employed. Chapter V speaks on the results of these analyses. Finally, Chapter VI discusses the limitations, potential extensions, and concludes.

Chapter 2

Literature Review

2.1 Criticisms of Existing Survey Data

Thomas (1999) considers some of the most popular sources of inflation expectation survey data mentioned earlier in this paper: the Livingston Survey, the Michigan Survey of Households, and the Survey of Professional Forecasters. The Livingston Survey provides predictions from professional economists, as does the Survey of Professional Forecasters. In contrast, the Michigan Survey consults households directly for their opinion on the future trajectory of the economy. Issues with the Livingston Survey include its historically poor design, which was due to the intention for it to be used for more journalistic purposes than scientific. Conversely, the Michigan Survey and Survey of Professional Forecasters are popular and relatively well designed. However, rather than replace these surveys, I believe they would be complemented by the inclusion of the analysis of social media activity. This is because there are still certain issues with surveys in general, such as the fact that respondents may self-report in a potentially inaccurate way, meaning that they might yield misleading results. Thomas also echoes this idea, noting the fact that the surveys do not necessarily measure informed opinion, in addition to the fact that agents have little incentive to put thought into the forecasts they do provide, or even to provide their genuine thoughts at all. In contrast, on social media, users are voluntarily sharing their thoughts. Therefore, I feel that, rather than asking people directly what they think about these issues, we can instead look to social media as somewhere where individuals volunteer their thoughts which are likely to be more genuine than the responses garnered in self-reported surveys (notwithstanding the presence of trolls and those simply looking to maximise their online interactions, which is an important consideration). We are then able to infer what people's behaviour is likely to be given their online activity, since people's search for or discussion of certain topics online is itself information. Rather than focusing on what people tell us, through surveys for instance, we can instead observe how they behave online and draw inferences from this.

Overall, asking households directly about their inflation expectations, as the Michigan Survey does, is an improvement upon asking experts how households are likely to think. This point is especially pertinent given the inability of many experts to accurately gauge public opinion with regards to important political and economic issues in recent times (Greenberg, 2016; Saiidi, 2016). However, in addition to this, I believe that inferring what households really think by observing their behaviour is a further improvement, for the reasons outlined above. Hence, we should supplement these established sources of data with new-age, real-time data, as I am proposing in this paper.

2.2 The Role of Narratives – Real World & Online

A fundamental component of economic discourse today is that which occurs online. Jurgenson (2012) discusses and proposes his idea of an Augmented Reality, where the digital and physical worlds, which are often separated by 'Digital dualists', effectively merge. This is because, while our online activity is influenced by our offline lives (e.g. determining who we follow on Instagram), it is equally as legitimate in this day and age to claim that our offline lives are determined to an extent by our online personas – as Jurgenson notes, social media users now constantly view the world as a potential photo, tweet, or status update. Our online lives augment our offline lives, rather than them being two separate entities. He discusses this concept in the context of protests and revolutions, where, for example, referring to the Arab Spring as a 'Twitter Revolution does not account for how these two worlds have merged and augment each other. However, I believe a similar idea also applies in the economic realm. In other words, it is wrong to separate the economic discussion and activity that occurs online from that which occurs offline.

Baudrillard's concept of hyperreality represents a similar perspective, although from a slightly different angle. Constituted by our representations of reality, including media simulations of reality, virtual reality games, and social networking sites, Baudrillard posits the idea that these so-called 'hyperreal' spaces can have huge influences on individuals, so much so that they come to control common thought and behaviour (Kellner, 2020). Again, I feel that this concept could be applied in economic contexts. It is the hyperrealities of traditional news media and, increasingly, social media that mould and shape our behaviour in the 'real' world.

This idea of narratives playing a role in driving economic activity has been of increased interest in the literature - in particular, Shiller (2017) considers the application of these ideas in an economic context more directly. He considers the role of narratives, which he defines as "mixtures of fact and emotion and human interest and other extraneous detail that form an impression on the human mind" (Shiller, 2017, p.11), in Economics. He does this in terms of how these popular narratives can spread and cause economic fluctuations, instead of purely attributing changes to economic feedback or multiplier effects popular within conventional economic models. He exemplifies this by discussing their influence during the Great Depression and also leading up to the Great Recession. Although not typically a central consideration in the analysis of economic shocks, he argues that we need to recognise how changes in narratives determine how much we spend, whether we decide to start a new business, or hire new employees, also known within Economics as the 'Animal Spirits'. Although he mainly considers this idea in the context of recessions, I believe these same ideas can and should be applied in the realm of inflation.

2.3 Forecasting with News

As part of a Bank of England study, Kalamara et al. (2020) consider whether newspaper text data contains pertinent information about future economic activity to inform policymakers with forward-looking information about various key economic variables. Using text from three popular UK newspapers from 1990 to 2019, representing views across the political spectrum, they compare the forecasting performance of models with and without the inclusion of text, in a wide range of forms including simple word counts and more complex text-derived regressors. They find that text does improve forecasts of macroeconomic variables, including GDP, inflation, and unemployment, when compared to widely used benchmarks, which interestingly they find to be true of forecasts of up to nine months into the future. Furthermore, they observe this effect particularly in times of stress, which implies the use of text in forecasts as being particularly useful in the prediction of recessions, for example.

Similarly, Tilly and Livan (2021) attempt to forecast whether the 10-year Break Even Inflation Rate (BEIR – the difference between nominal and inflation linked bond yields) will increase or decrease in the next five days for eight different countries, using market-based variables in conjunction with narrative-based features gathered from the Global Database of Events, Language and Tone (GDELT), which monitors world media and extracts themes from them. The authors instantiate five separate Machine Learning models, using market-based variables alone as a baseline. They then test the performance of these models with and without the aforementioned narrative-based variables, which leads to the finding that for seven out of eight countries, those features derived from GDELT themes improve one day ahead predictions of five day BEIR movements. These examples highlight the benefits of including narrative-related variables in forecasting models.

2.4 Forecasting with Internet Search Data

Some authors use an intermediate form of data between traditional newspaper text and social media activity to make predictions: internet search data. Given the time lag associated with the release of monthly government data, Varian and Choi (2009) hypothesise that Google search query data is correlated with current economic activity and so can help us in predicting future data releases. One example they apply this to is the prediction of the number of Ford vehicles sold in a given month. To show this, they use seasonal autoregression models, predicting the number of sales based on the sales of the previous month and the same month in the previous year. They then add a Google search 'query index' for the term 'Ford' during the first week of the month of interest. Overall, they observe a positive coefficient on this variable, indicating a positive association between the volume of 'Ford' searches and Ford sales, and that models including the relevant Google Trends variables outperformed those excluding these predictors. They also achieve similar success in the case of predicting initial claims for unemployment benefits in this way.

Guzmán (2011) sees similar power in internet search data, arguing that it can be used as a measure of consumers' revealed expectations. The author therefore looks to apply this to the prediction of inflation, given that changes in the number of search queries concerning inflation can be used to measure households' revealed expectations for future changes in inflation. A further advantage is that internet search data could be seen as more objective than survey data, given that there is no potential for framing effects as there is in the case of surveys, resulting from the very specific ways in which questions are often posed to participants. Using a Granger Causality test, Guzmán tests the predictive power of various inflation expectation indicators, including the Google Inflation Search Index (GISI) formulated by the author, which corresponds to the volume of searches for the term 'inflation'. The results of this analysis indicate that the GISI anticipates the inflation rate without bilateral feedback, meaning that inflation expectations derived from internet search queries are predictive of future inflation, but are interestingly also not influenced by the actual current rate of inflation.

2.5 Forecasting with Social Media Data

Here, I consider the use of social media data in forecasting within financial markets, as the methods discussed could potentially be applied in this research. Using event study methodology, Ante (2022) seeks to calculate the share of the identified returns and trading volume of various cryptocurrencies attributable to Elon Musk's Twitter activity. Initially, the expected return is calculated over a given period before an unexpected event.

This is then compared to the true observed return around the event. The difference between the two is the abnormal return that can be attributed to the event (Brown and Warner, 1985). Overall, they find large positive effects in the first two minutes after tweeting, leading to the conclusion that the market reacts quickly and significantly to Musk's tweets. Studies like these show us how social media can move markets. However, there is limited research on how this impact might translate in the context of traditional economic metrics.

The literature covering inflation forecasting with social media data is sparse. Aromi and Llada (2020) come the closest in attempting to do this, using Twitter data in Argentina. To do so, they build an autoregressive forecasting model including lagged monthly inflation and lagged currency devaluation as a baseline. This model is then extended to include an indicator of lagged Twitter content, in the form of the 'attention' given to the subject of inflation – this model indicates that social media content adds information regarding future inflation levels. Namely, a one standard deviation increment in the indicator of attention anticipates a mean increment of approximately 0.4% in monthly inflation. Therefore, their research indicates that Twitter content does play a role in anticipating macroeconomic outcomes. However, one shortcoming of the analysis is the attention index they use, which is given by the frequency of the terms 'inflation' or 'inflationary' divided by the number of tweets in their corpus. This lends itself to overrepresentation of these terms in the corpus and therefore the overinflation of the attention index. Instead, the frequency of the relevant terms divided by the number of words in the corpus would be a more accurate metric, which is something I implement in 3.3.

2.6 Summary

In summary, the existing literature relevant to this research includes the role of narratives, both real world and online, in impacting our day-to-day lives. While Jurgenson and Baudrillard consider them in isolation, Shiller demonstrates how we can take these ideas and apply them in an economic setting. More directly related to this study, we see the success in forecasting with a wide variety of data, ranging from conventional data to social media data, although the latter is not considered so directly in the context of Economics. Finally, the criticisms of the popular existing inflation forecasting survey data leads me to propose the use of social media data directly within this context.

Chapter 3

Data

3.1 Data Collection

3.1.1 Dependent Variable

As my dependent variable, I use the US 10-year Breakeven Inflation Rate (BEIR), as per Tilly and Livan (2021). This is the future inflation rate embedded into the US Treasury securities market, constituted by the difference in the nominal yield on a US 10-year treasury and the real yield on an inflation linked investment of the same maturity (Girola, 2019). I use this as it is one of the few inflation-related measures that is reported almost daily, whereas direct inflation measurements are provided only monthly. This data was collected from the St Louis Federal Reserve website (FRED, 2022).

3.1.2 Independent Variables

As the main independent variables of interest, I constructed both sentiment and attention time series metrics from social media posts. To do this, I gathered posts from both Twitter and Reddit containing the terms 'inflation' or 'inflationary', with a view to then analysing and manipulating the raw data in order to yield the aforementioned metrics. How this was done in particular is discussed later in 3.2.

Tweets were collected via the Academic API, querying for those tweets containing the aforementioned two terms that were also geotagged to the US. This made it more likely that these tweets would be relevant to the discussion of US inflation specifically. I decided to do this given the aim of supplementing the popular US inflation forecasting surveys, as mentioned earlier. However, one issue with this selection strategy is that of selection bias, as less than 1% of tweets are geotagged (Schlosser, Toninelli, and Cameletti, 2021), with some in the literature finding a statistically significant difference between geotagged and non-geotagged users (Karami et al., 2021).

Reddit posts were collected in a similar way using the Pushshift API (Baumgartner, 2022), which acts as a store of copies of Reddit comments and submissions from across the platform, with queries made for posts made containing references to 'US inflation'. These posts were taken from a range of subreddits, which is beneficial in getting a variety of ideological views, however, it was also the case that some posts were not particularly relevant as a result. Other issues with the original Reddit data included the fact that some posts were almost identical, but were not perfect duplicates as they were posted at slightly different times. Social media users often do this as a joke, through copypastas (*Copypasta* 2022), for example, where users intentionally copy and paste certain posts. If not dealt with, these posts would therefore inflate the sentiment and attention scores, making them unrepresentative. Given their broad similarity, I observed that these posts were identical after processing the text for sentiment analysis. Therefore, those posts that were identical after text pre-processing were dropped. This represented almost 9% of the total data initially collected, making its removal important in ensuring the validity of the data set.

There are various benefits to using posts from both platforms. Firstly, Reddit posts do not have character limits unlike tweets which are capped at 280 characters. This allows for greater richness of discussion and expression of sentiment in each post. Secondly, and more significantly, it is usually the case that researchers collect data from one platform in isolation. However, this can be disadvantageous as the average user on each individual platform is likely to have distinctive biases and characteristics. For instance, if we consider the political alignment of the users on each platform, according to the Pew Research Centre, those on Twitter tend to be more left-leaning (Wojcik and Hughes, 2019), whilst the share of right-leaning users is greater on Reddit (Vogels, Auxier, and Anderson, 2021). Therefore, using only one of these platforms as a source of data runs the risk of biasing the analysis. Hence, I feel that combining these platforms in my analysis will mean that these biases are likely to almost cancel each other out, given their somewhat opposing natures, notwithstanding the fact that there will be some overlap in terms of user base. As a result, this strategy is more likely to yield balanced and therefore more generalisable and valid outcomes. In total there 84,909 posts were collected, spanning a 1-year time period from 1st June 2021 to 31st May 2022 - 60,061 posts were from Twitter and 24,848 were from Reddit.

3.1.3 Control Variables

In terms of control variables, I use the interest rate (*interest_rate* - Fed Funds Effective Rate), the Producer Price Index for commodities (*commodities* - the importance of which in this context is highlighted by Cabral, Ribeiro, and Nicolau (2022)), the unemployment rate (*UNRATENSA*), Michigan Survey sentiment score (*UMCSENT*), M2 money supply (*WM2NS*), GDP (*gdp*), and USD/EUR exchange rate (*USDEUR*). The data for these variables was collected from the St Louis Federal Reserve Economic Database (FRED, 2022). Some of the data was provided in a weekly or monthly format and so needed to be repurposed into daily data. Some data was also missing and so needed to be imputed, as was the case with BEIR, described in 3.3.

I also incorporate lags in the data, using the interest rate from one year prior, for example, as per Friedman (1972) and Batini and Nelson (2001), who find that this is how long monetary policy takes to make an impact on the wider economy. I also extend this logic to the other broader economic factors used: commodity price index, unemployment rate, M2 money supply, GDP, and Exchange Rate - I do this for simplicity however this logic could be refined in any extensions made to this research. I also use a 20-day lag for the overall sentiment, overall attention, and Michigan sentiment variables, as Varian and Choi (2009) do in the case of predicting future Ford sales, for example, by incorporating a 3-week lag of the query index for searches of 'Ford'.

3.2 Sentiment Analysis

To conduct the analysis of the sentiment contained in each post I use the VADER Sentiment Analysis tool (Hutto and Gilbert, 2014). VADER (Valence Aware Dictionary and sEntiment Reasoner) is a tool, both lexicon and rule based, that has been built specifically to be able to detect the sentiment present in social media activity and to detect both the sentiment polarity and intensity, making it particularly useful in the context of this research.

Certain features of social media content such as the tendency to lack context given the shortness of posts (e.g. Twitter's 280 character limit) and the use of abbreviations and slang mean that detecting sentiment can often be difficult. VADER therefore aims to provide a tool that can generate highly accurate classifications in spite of these features. To do so, the construction of VADER began by building upon existing sentiment word banks such as Linguistic Inquiry and Word Count (LIWC) and Affective Norms for English Words (ANEW) by incorporating three features that are highly common in social media posts: 1) Western-style emoticons such as ':-)' to denote a 'smiley face' (in addition to the popular UTF-8 encoded emojis), 2) acronyms such as 'LOL', and 3) slang such as 'nah' or 'meh'. This process left 9,000 lexical feature candidates, to which sentiment ratings were assigned by 10 human judges - ratings ranged from -4 (Extremely Negative) to 4 (Extremely Positive), with 0 indicating a neutral score. Each token with a non-zero mean and standard deviation of less than 2.5 was kept, leaving 7,500 remaining tokens. Examples of the scores of positive tokens include 'okay' with a score of 0.9, 'good' with a score of 1.9, and 'great' with a score of 3.1. Negative tokens such as 'horrible' scored -2.2, and even unconventional tokens such as a frowning emoticon and slang terms such as 'sux' yield scores of -2.2 and -1.5 respectively. These valence scores are then adjusted and normalised to yield a compound score lying between -1 (extreme negative) and +1 (extreme positive). This normalised, weighted composite score is what I ultimately use to denote sentiment in this research.

In addition to this, to help with determining sentiment intensity, 800 social media posts, in the form of tweets, were analysed to help identify the key features affecting intensity. 400 positive and 400 negative tweets were selected from a random sample of 10,000. Then, two human judges scored the sentiment of each tweet as outlined earlier, on a scale from -4 to 4, in conjunction with highlighting the characteristics of the text that particularly impacted the sentiment intensity. This led to the identification of five key heuristics, outlined in Table 3.1 below, that would not usually be captured in a bag-of-words model, for example. The incorporation of all these features means that the VADER lexicon performs exceptionally well in the social media domain in particular.

Heuristic	Impact on Intensity	Example
Punctuation	Can increase the magnitude of intensity without modifying the semantic orientation	"The food here is **good!!!**" is more intense than "The food here is **good**"
Capitalisation	Using all caps increases the intensity without modifying the semantic orientation	"The food here is **GREAT!**" is more intense than "The food here is **great!**"
Degree Modifiers	Intensifiers or degree adverbs can increase or decrease intensity	"The service here is **extremely** good" is more intense than "The service here is good", whereas "The service here is **marginally** good" reduces intensity
Contrastive Conjunction ("but")	"but" often shifts the sentiment polarity (i.e. from positive to negative or vice versa)	"The food here is great, **but the service is horrible**" has mixed sentiment, the latter half of the sentence often dictates the overall sentiment polarity
Preceding Tri-grams	Tri-grams preceding a lexical feature that is highly indicative of sentiment often reverses the polarity	"The food here **isn't really all that great**"

TABLE 3.1: VADER Sentiment Heuristics (Hutto and Gilbert, 2014, p.221)

There are therefore three main advantages to using VADER. Firstly, in practical terms, it is quick and computationally efficient whilst remaining highly accurate. Secondly, the rules governing VADER are completely accessible, rather than being contained within a black-box, and are thus more easily modified. Finally, although VADER is self-contained given the rules it is built upon, it is highly generalisable and performs well without the need for a large training set. Evidence for this can be seen from its wide use in the literature, with tasks ranging from detecting hate speech (Davidson et al., 2017) to conducting text analysis for consumer research (Humphreys and Wang, 2018). Its use in both economic and non-economic contexts therefore gives me confidence in employing it in this research.

In terms of assessing its performance, VADER's accuracy was compared to popular benchmarks including the LIWC, SentiWordNet, and other methods based on Naïve Bayes classifiers or Support Vector Machine algorithms, among others. Overall, VADER generalises better than these benchmarks and actually outperforms even human judges of sentiment, achieving an F1 classification score of 0.96 vs 0.84 for humans.

After assigning sentiment scores to the posts gathered in 3.1.2, in some cases, posts that were actually negative in their sentiment were assigned positive scores. This tended to occur in cases where the poster was using sarcasm, with sarcasm detection being one key limitation of VADER. However, this issue is not exclusive to VADER as the difficulty surrounding sarcasm detection has been noted throughout the literature (Afiyati et al., 2022; Joshi, Bhattacharyya, and Carman, 2017; Rajadesingan, Zafarani, and Liu, 2015).

3.3 Data Manipulation

Firstly, with regards to the BEIR, the data was not available for every day, with some missing values. To account for this, I imputed values where they were missing, using the average value across the five days prior.

The main part of the data manipulation process, however, involved sentiment analysis, the initial steps of which involved text pre-processing. This included (i) removing certain punctuation, (ii) removing stopwords, and (iii) lemmatisation. Those punctuation characters that added little to the sentiment of a post, such as commas and semicolons as opposed to exclamation marks that add to the sentiment of a piece of text, for example, were removed using regular expressions. With regards to the removal of stopwords, these are words that, while critical to the structure of a sentence, add little to no value to the sentimental meaning, such as 'the', 'and', or 'I'. To do this I used the nltk list of stopwords (Bleier, 2010). Lemmatisation, which was conducted using spaCy (spaCy, 2022), aims to remove inflectional endings of words and "to return the base or dictionary form of a word, which is known as the lemma" (Stanford, 2008). For example, if we consider the word *saw*, lemmatisation would attempt to return either *see* or *saw* depending on whether the use of the word was as a verb or as a noun. This is in contrast to stemming, another popular text pre-processing method, where endings of words with the same 'stems' are removed (e.g. runs = running = run), which might return just *s*.

After processing the text and assigning sentiment scores using VADER as described in 3.2, I created a sentiment variable, the Overall Sentiment Score (OSS). All sentiment scores from posts made on the same day were summed to create a daily sentiment score, with each daily score then being collated to create a sentiment time series across the 1-year period. The same logic applied to each post's attention score, which was defined by:

$$\frac{Number\ of\ mentions\ of\ "inflation"\ or\ "inflationary"\ in\ the\ post}{Number\ of\ words\ in\ the\ post}$$

the daily sums of which similarly created the Overall Attention Score (OAS). This was built upon the attention score metric created by Aromi and Llada (2020), which was defined by the total number of mentions divided by the total number of posts, which is over-representative in my view.

3.4 Data Exploration

3.4.1 Overall Sentiment and Attention Scores

Figures 3.1 and 3.2 show the changes in the OSS and OAS. We can see that they exhibit trends we might expect - as time has passed between June 2021 and May 2022 and inflation has become a more pressing and central issue, people have begun to pay more attention to it online, with this attention becoming more and more negative.

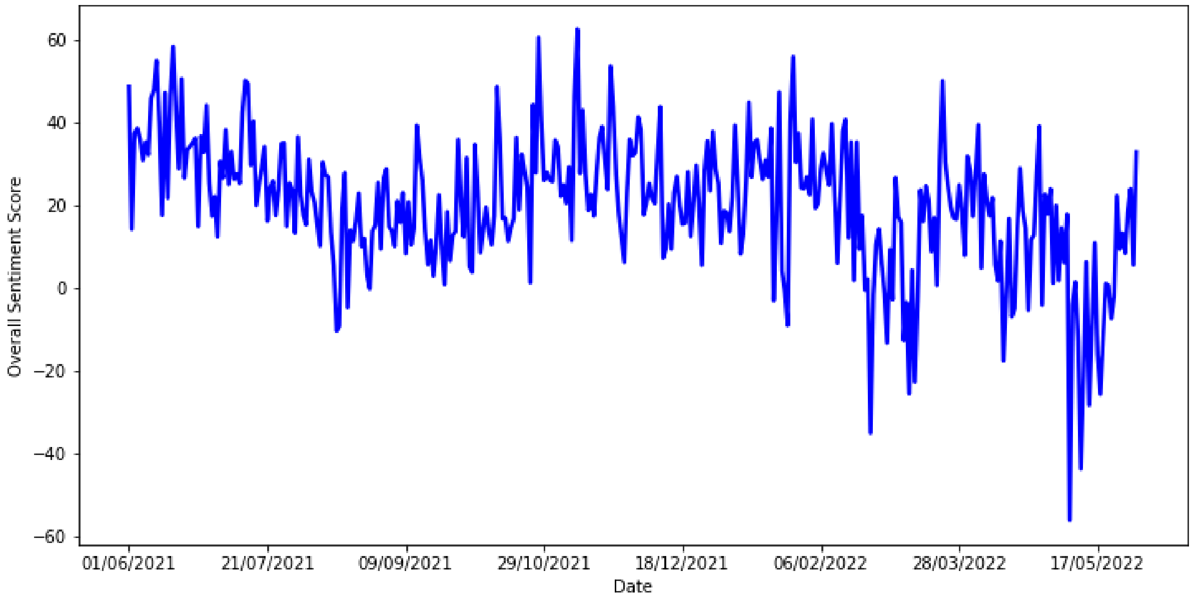

FIGURE 3.1: Overall Sentiment Score

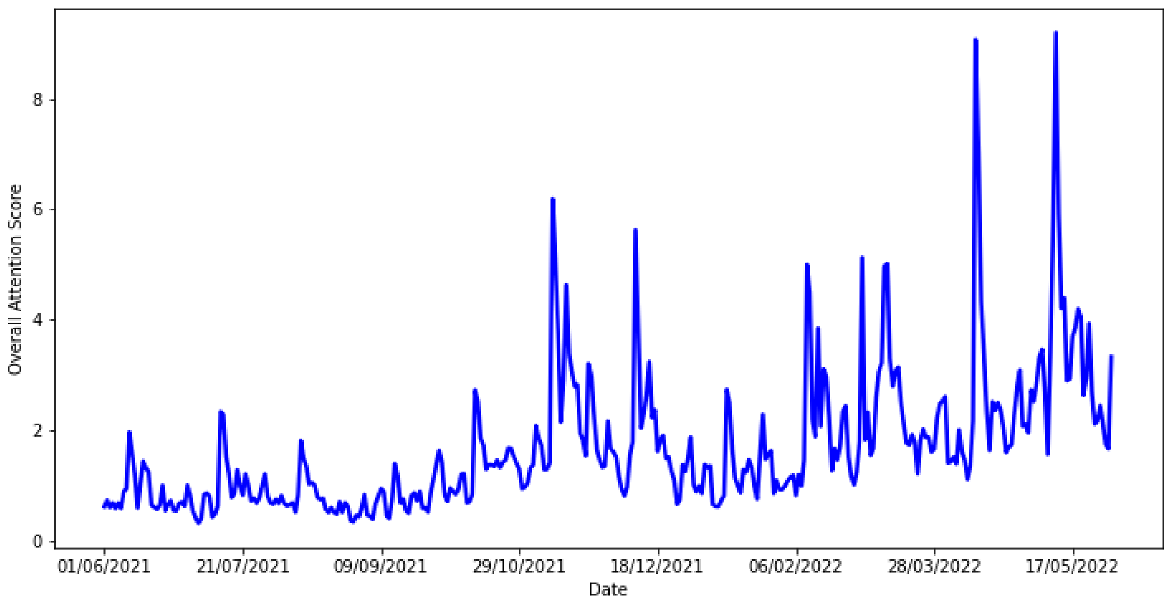

FIGURE 3.2: Overall Attention Score

3.4.2 Sentiment Score - Platform Comparison

Figure 3.3 shows that the sentiment score from both platforms trend downward over time, which is what we would expect as people become more pessimistic about inflation as it has begun to impact them more over the last 6 to 8 months. However, sentiment on Reddit has been much more positive than on Twitter, whilst sentiment on Twitter also appears to trend downward at a greater rate. One reason could be the greater sense of community on Reddit, an idea which is expanded upon in 3.4.4. However, this could also be due to Reddit's stronger meme culture. Hence people might make jokes that appear positive to VADER but in reality are used to express negative sentiment, as per the discussion of its struggles in detecting sarcasm in 3.2.

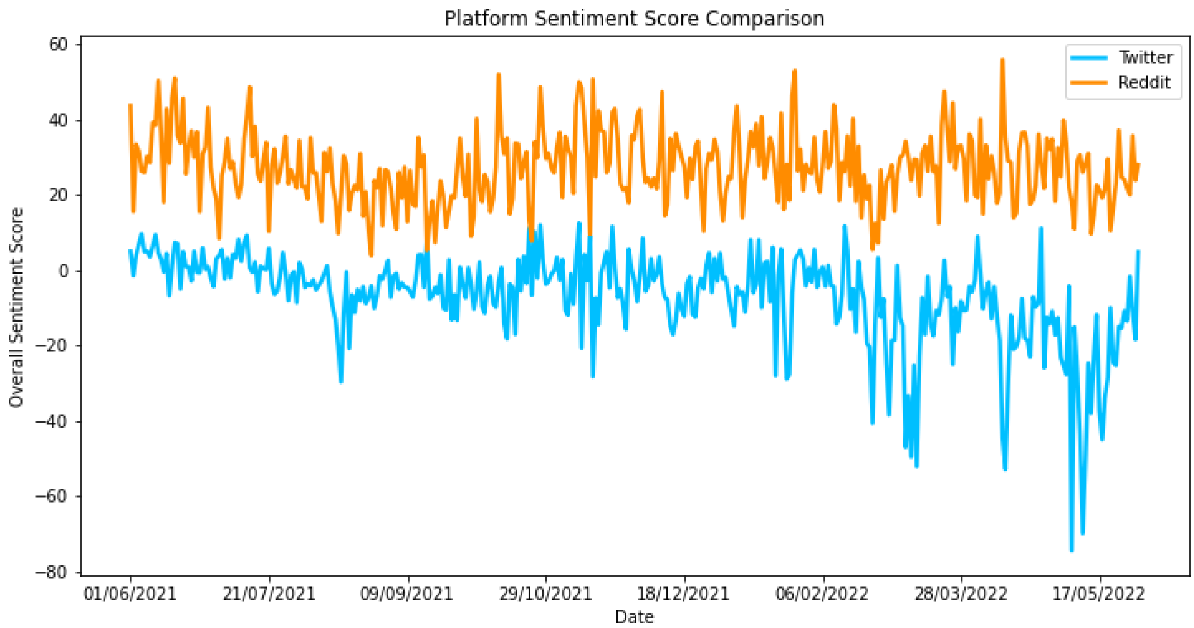

FIGURE 3.3: Twitter vs Reddit Data Sentiment Scores

3.4.3 Attention Score - Platform Comparison

In the case of the attention scores, Figure 3.4 shows that while the metrics increase over time which again we would expect as inflation has gradually become the main focus within many economies over the last 6 to 8 months, the Reddit metric appears to be relatively constant in comparison. However, we again observe a large disparity in the scores from each platform. One reason for this could simply be the fact that more posts were retrieved from Twitter than Reddit. However, while there were around 2.5 times more Twitter than Reddit posts, the attention scores differ by more than 2.5 times at any given time, hence this disparity requires further explanation.

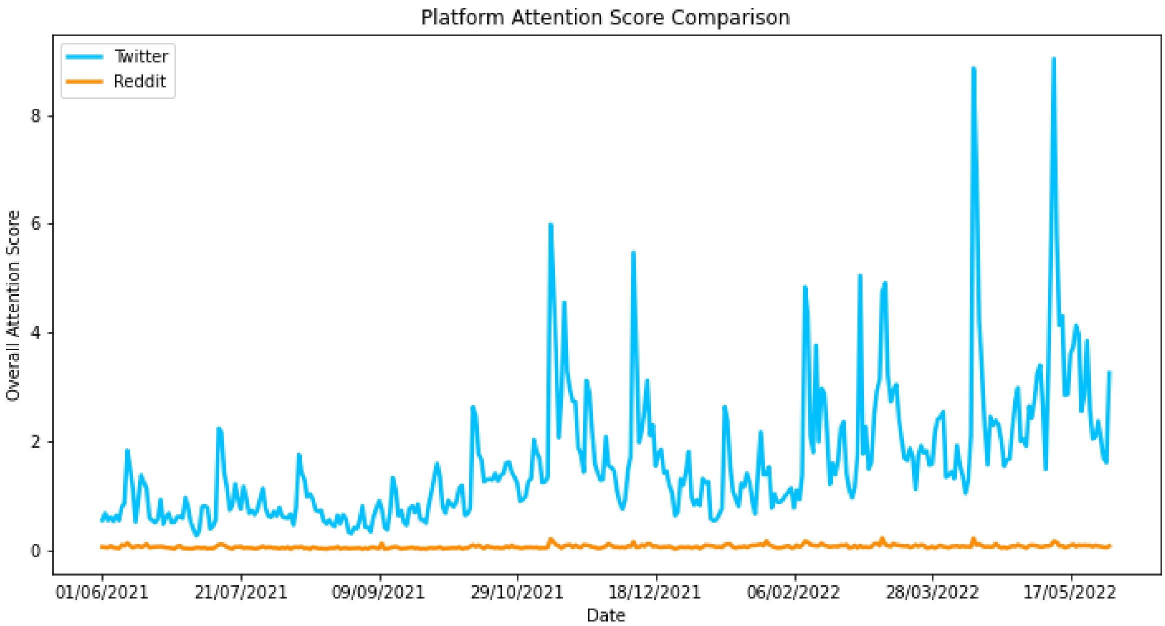

FIGURE 3.4: Twitter vs Reddit Data Attention Scores

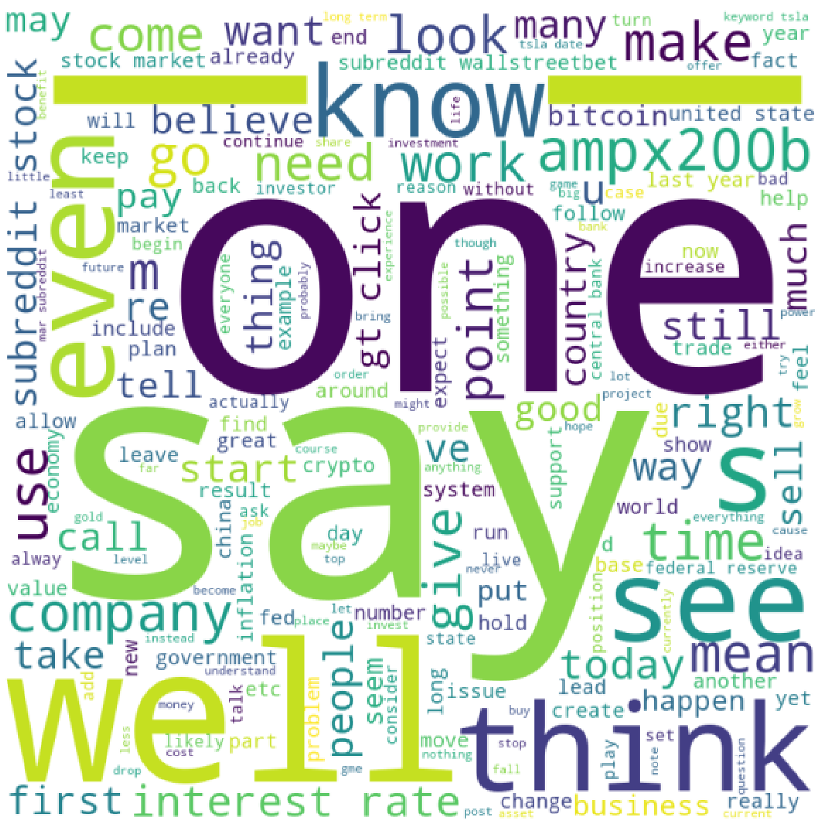

FIGURE 3.5: Reddit Data Wordcloud

FIGURE 3.6: Twitter Data Wordcloud

3.4.4 Wordcloud Comparison

We can use wordclouds to visualise how the discourse surrounding inflation differs between the two platforms and comparing them in Figures 3.5 and 3.6 seems to help us to understand why Reddit sentiment appears to be much more positive than Twitter. The Twitter word cloud shows the presence of more terms that are likely to be associated with other negative terms such as 'war' or 'cost', in addition to many more overtly political terms, such as 'republican', 'democrat', 'trump', and 'biden' which most likely further fuelled negativity as these terms tend to create division online. The Reddit wordcloud shows less focus on these kinds of terms and more focus on terms that people are likely to discuss in a positive light on the platform, such as 'bitcoin', 'stock', 'tsla', and 'crypto'. This is most likely because Reddit places a greater focus on communities (via Subreddits), whereas Twitter users are more isolated. In this context specifically, given that communities like WallStreetBets, who often work towards a common goal or share a common ideology, tried to capitalise on movements in the markets, this collective action means that these terms would be more likely to be surrounded by positive discourse.

Chapter 4

Methodology

4.1 General Formulation

In terms of the general problem formulation (see Equation 4.1), I seek to make predictions for the values of the dependent variable BEIR, based on the variation in the two independent variables of interest OSS and OAS.

$$BEIR = OSS + OAS + X' + \epsilon \tag{4.1}$$

X' represents the vector of control variables: the interest rate (Fed Funds Effective Rate), the Producer Price Index for commodities, the unemployment rate, Michigan Survey sentiment, M2 money supply, GDP, and USD/EUR exchange rate. ϵ represents the error present in the general model. Although this appears simple, I seek to employ a parsimonious approach such that I can introduce complexity via the use of a variety of model architectures.

4.2 Vector Autoregression

VAR is a statistical model used to capture the relationship between multiple time series. It is often the case that these time series influence each other, which VAR allows us to manage. It extends the idea of a univariate autoregression, where the output variable depends on the lagged values of itself, allowing for k time series regressions where all k series are also used as regressors. A benefit of VAR in this case is that, whilst typical autoregressive models are unidirectional, VARs are bidirectional (the dependent and independent variables influence each other). Each variable in the model is a linear combination of past values of itself and the other variables, hence a model including 5 time series would yield 5 equations. VARs have proven to be popular in the realm of dynamic economic and financial time series analysis and forecasting in particular (Zivot and

Wang, 2006) given their success relative to univariate time series models, for example. Since many of the variables included in the model do in fact influence each other, these factors combine to make the use of VAR appropriate for this research.

Before initiating the forecast and building the VAR model, there are both various tests that must be conducted and hyperparameters that must be selected. In terms of lag length selection, the general approach is to fit VAR(p) models (where p is the number of variables included in the model), with orders from 0 to p, and then choose the value of p that minimises some selection criteria, such as the Akaike Information Criteria (AIC) (Konishi and Kitagawa, 2008). In this case, the optimal lag indicated was 1 period according to the AIC. Another requirement for running a VAR is ensuring the stationarity of the variables being used in the model. A stationary time series is one whose characteristics such as mean and variance do not change over time – or in other words, there is no clear trend in the data. To do this, I use the Augmented Dickey-Fuller test (Wei, 2016). Only the OSS was found to be stationary at first, hence I made the remaining variables stationary by taking first differences of the series, by computing the differences between consecutive observations. Doing this helps to stabilise the mean of a time series by removing changes in the level of that time series, thus eliminating (or at least reducing) the trend. The effect of taking first differences to yield stationarity can be seen by comparing Figures 4.1 and 4.2, from which we can see that the trends that were initially present in all variables except OSS disappear after first differencing. Finally, the Granger Causality test allows us to determine whether one time series is useful for forecasting another time series (Wei, 2016), which is clearly important to test in this context. Similarly to other hypothesis tests, if the obtained probability value is less than a given significance level, the null hypothesis that a given time series is not useful in predicting another is rejected. Here it is important to note that rejecting the null in this context does not imply true causality, rather it only implies forecasting ability. In this case, with reference exclusively to the results of the test where BEIR is the dependent variable, we reject the null for both the OSS and OAS at the 5% level, yielding p-values of 0.0227 and 0.0356 respectively. Hence we can infer that both hold the property of causality of BEIR in a forecasting context which is encouraging.

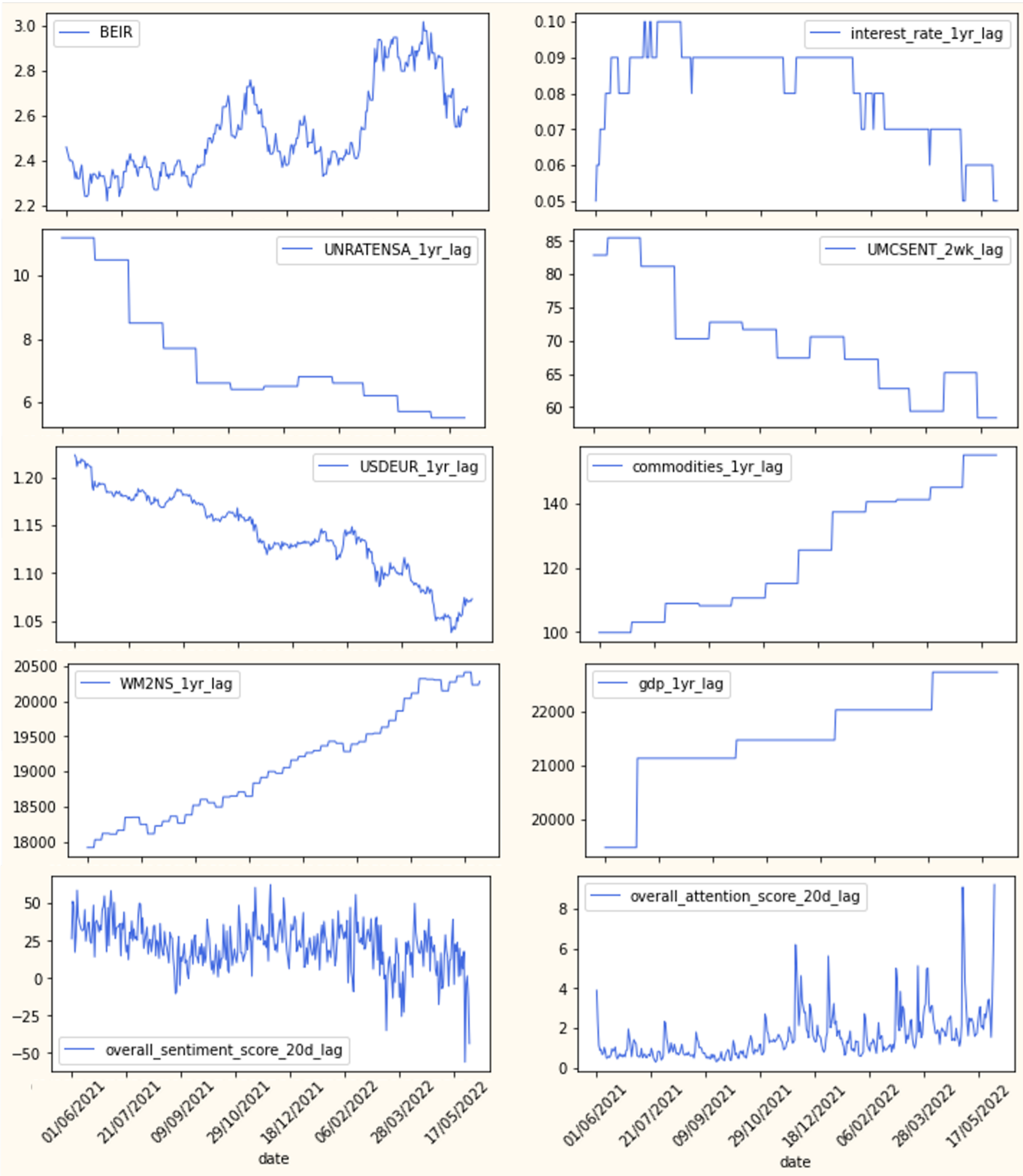

FIGURE 4.1: Time Series - Non-stationary

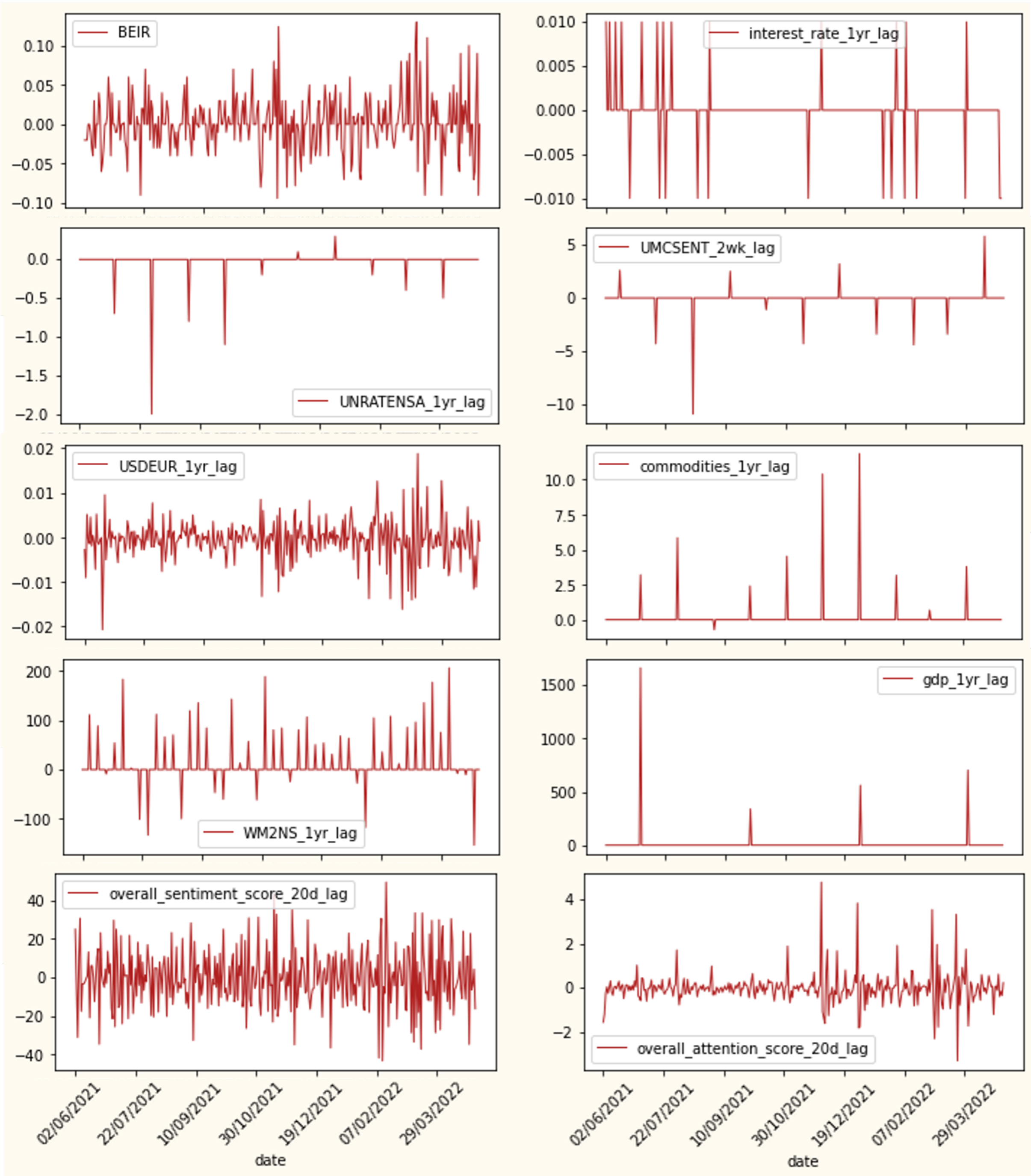

FIGURE 4.2: Time Series - Stationary

4.3 Hyperparameter Optimisation

In both cases, I look to optimise the selection of hyperparameters used in each model, doing this via either grid search or random search. In both cases, we specify the hyperparameters we wish to optimise alongside a list of values for those hyperparameters that we want to test. Models are then built with combinations of the different values of the hyperparameters in order to determine which combination leads to the best performing model. With grid search, each and every combination of the grid of hyperparameters provided is tested. For example, in the case of a tree based model, if we wanted to optimise for the number of trees created and the maximum depth of the model, both with potential values of 1 to 100, there would be 4950 possible combinations ($= \frac{100!}{2!(100-2)!}$).

However, random search, a variation of grid search, involves testing various random combinations of model hyperparameters in order to determine which combination yields the best results. Whilst grid search tests every combination of the grid of hyperparameters provided, random search tests only a random subset of those hyperparameters, the size of which we are able to specify. A key benefit of random search over grid search is that over the same domain random search is able to yield models that produce results that are either equally good or better, whilst also taking less time computationally (Bergstra and Bengio, 2012). Hence, given the same computational budget, random search is preferable to grid search and so I mainly use this to optimise the feature selection models.

4.4 Least Absolute Shrinkage and Selection Operator

LASSO is a regularisation technique, which constrains coefficient estimates by shrinking certain coefficients that are not so useful to 0. This is as opposed to another regularisation technique, Ridge Regression, which simply reduces the value of certain coefficients. The benefit of this is that it reduces variance whilst also minimising bias, which should improve the accuracy of predictions.

For Ordinal Least Squares (OLS) regression, we estimate β by minimising the Residual Sum of Squares (RSS), the difference between our predictions and the true values:

$$RSS = \varepsilon' \varepsilon = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \tag{4.2}$$

However, with regularisation, we seek to minimise Penalised Residual Sum of Squares (PRSS). Thus, the LASSO coefficients $\hat{\beta}^L_\lambda$ minimise the quantity:

$$PRSS(\lambda) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{4.3}$$

where λ is the shrinkage parameter and $\lambda \sum_{j=1}^{p} |\beta_j|$ is the L1 norm of β.

Given that LASSO shrinks some coefficients to exactly 0, we say that it performs variable selection, yielding a sparse model that involves only a subset of all variables used. As a result, selecting a good value of the shrinkage parameter λ is crucial. Thus I select λ via grid search, with values between 8.69315254e-06 to 8.69315254e-03 being searched, in order to determine which value minimises cross validation error. In this case, the optimal value is 0.001, as can be seen in Figure 4.3 below.

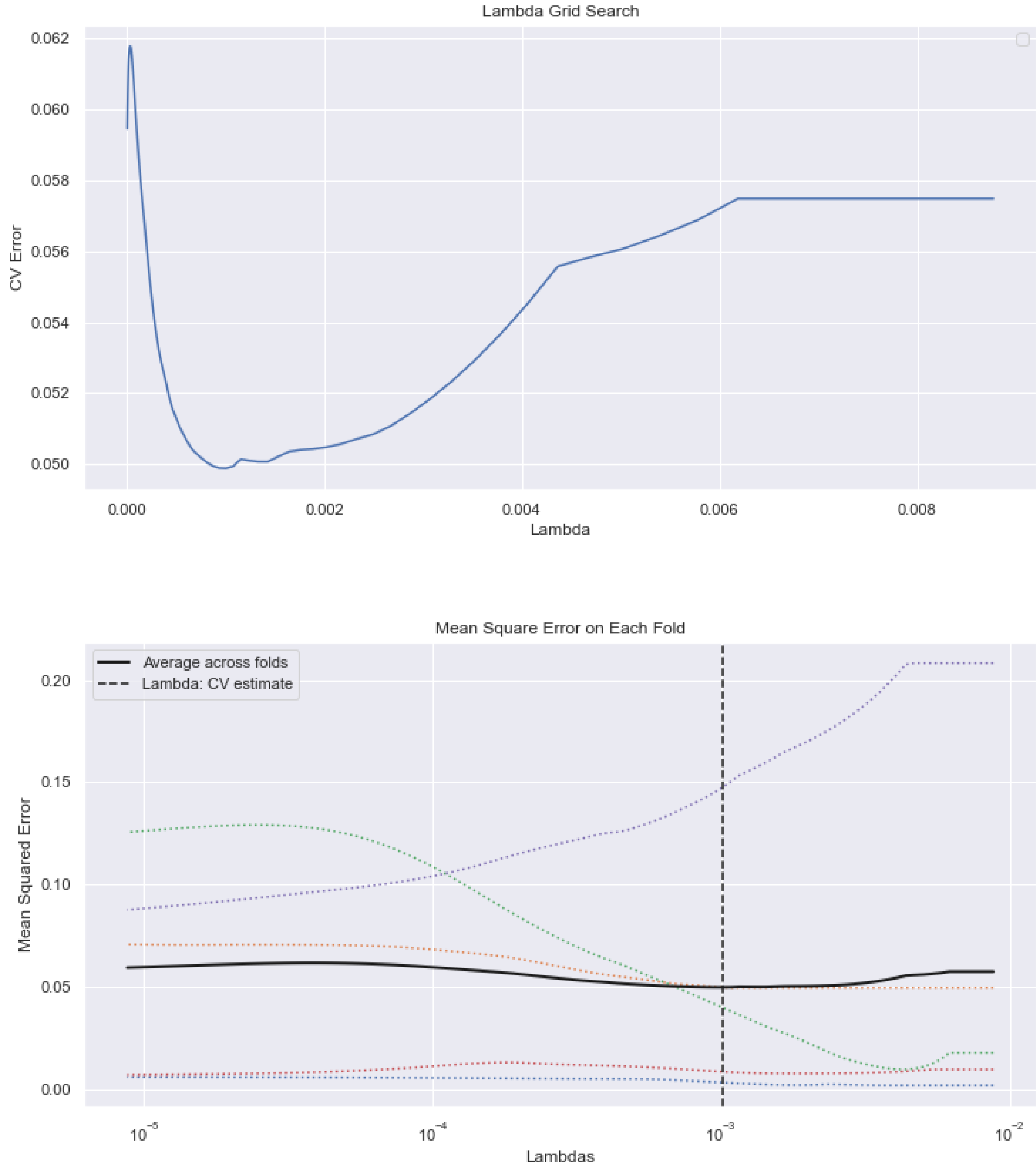

FIGURE 4.3: λ Value Grid Search

4.5 Support Vector Regression

Like LASSO, SVR (Drucker et al., 1996) is a development on Simple Linear Regression. SVR allows us to add a degree of flexibility in terms of how much error we are willing to accept. This flexibility manifests itself in the form of a corridor of acceptable error around the estimated function, of width ε, such that errors of size ε or less (both above and below the estimate) are ignored, whilst those that exceed ε are penalised, as per Vapnik's ε-insensitive approach (Cortes and Vapnik, 1995). This ε-insensitive region is how the move is made from the binary classification problem, solved by Support Vector Machines (Cortes and Vapnik, 1995), to the regression problem in this case. In the classification case, the optimisation problem consists in finding the maximum margin separating the hyperplane, the dividing line between the two classes, whilst maximising the number of training points classified correctly. In this case, support vectors, those points lying on the margin boundary, represent the optimal hyperplane. However, in the regression case, the optimisation problem involves (i) defining a convex ε-insensitive loss function to be minimised and then finding the flattest corridor that contains as many training observations as possible, whilst (ii) minimising prediction error in the form of the distance between the true and predicted values.

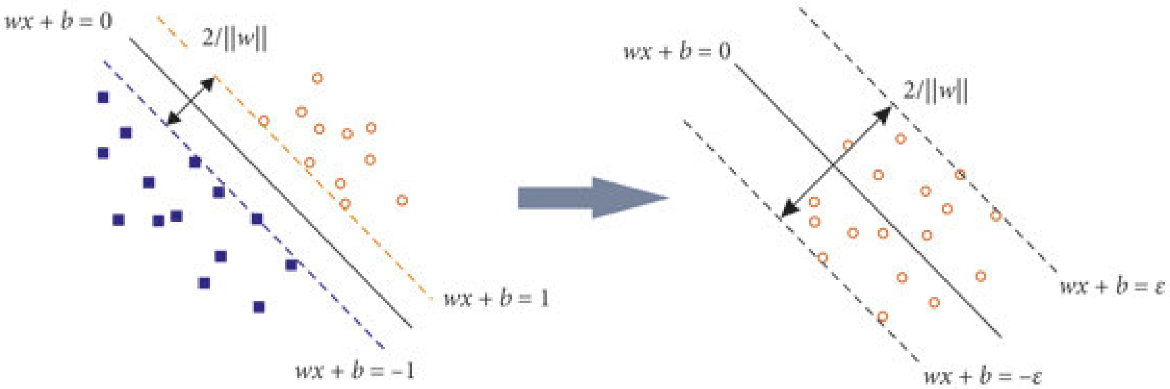

FIGURE 4.4: Support Vector Machine vs Support Vector Regression (Wang et al., 2020)

The multivariate regression problem takes the form of:

$$f(x) = \begin{bmatrix} 2 \\ b \end{bmatrix}^T \begin{bmatrix} x \\ 1 \end{bmatrix} = w^T x + b \tag{4.4}$$

$$x, w \in \mathbb{R}^{M+1} \tag{4.5}$$

where M is the order of the polynomial used to approximate the function.

Condition (i) produces the objective function $MIN\frac{1}{2}\|W\|^2$, where $\|W\|$ is the magnitude of the normal vector to the surface that is being approximated.

Incorporating slack variables (ξ_i) adds further flexibility to the model, where ξ_i denotes the deviation of observations that fall outside of ε from the margin. A further hyperparameter, C, acts a regularisation, which corresponds to the total amount of acceptable error, the sum of ξ_i. Increasing C increases the toleration of these errors discussed above.

This combination of features of the model leads us to the objective function to minimise:

$$MIN\frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}|\xi_i| \tag{4.6}$$

Subject to the constraint:

$$|y_i - w_i x_i| \leq \varepsilon + |\xi_i| \tag{4.7}$$

Some of the advantages of SVR are that its computational complexity does not depend on the dimensionality of the input space, whilst it also has strong generalisability and high prediction accuracy (Awad and Khanna, 2015).

With regards to hyperparameter tuning, the grid of values searched can be seen in Table 4.1, with particular interest in the value of C. The optimal hyperparameters used according to random search are gamma = 'scale' (Kernel coefficient - in this case using $\frac{1}{(n\_features * var(X))}$) and a C value of 8.8, in addition to the use of a Radial Basis Function (RBF) kernel.

Hyperparameter	**Values (Start, Stop, Step)**	**Optimal Value**
C	1, 10, 0.1	8.8
gamma	auto, scale	scale

TABLE 4.1: Support Vector Regression Hyperparameter Search Grid

4.6 Tree Based Models

4.6.1 Decision Trees in General

In this section, I explore the use of Random Forests and XGBoost, both of which are comprised of a combination of decision tree predictors. Simple decision trees aim to recursively split a feature space **x** into j smaller regions $(R_1, ..., R_j)$. For each observation that eventually falls into R_j, we predict its value to be the mean value of all training observations in that region. The goal of this process is to create the appropriate regions such that RSS error is minimised. These regions are created by making binary splits at the internal nodes, at a given value of some input variable. We select the predictor X_j and some cutpoint s such that splitting the predictor space $< s$ and $>= s$ leads to the greatest possible reduction of error. This process repeats, looking for the best predictor and cutpoint to split one of the two previously identified regions, and stops upon reach a given criterion, for example, a maximum specified depth. Once the final regions (also known as the terminal nodes) are created, each one corresponds to a prediction for the value of some output, which in the case of regression is the average value of the observations present in a given region. The idea behind decision trees is outlined visually in Figure 4.5. Although this example describes the use of trees for classification, the same idea is applied in the case of regression - the outcomes would be predicted values, constituted by the average of the observations in each region, as opposed to predicted classes.

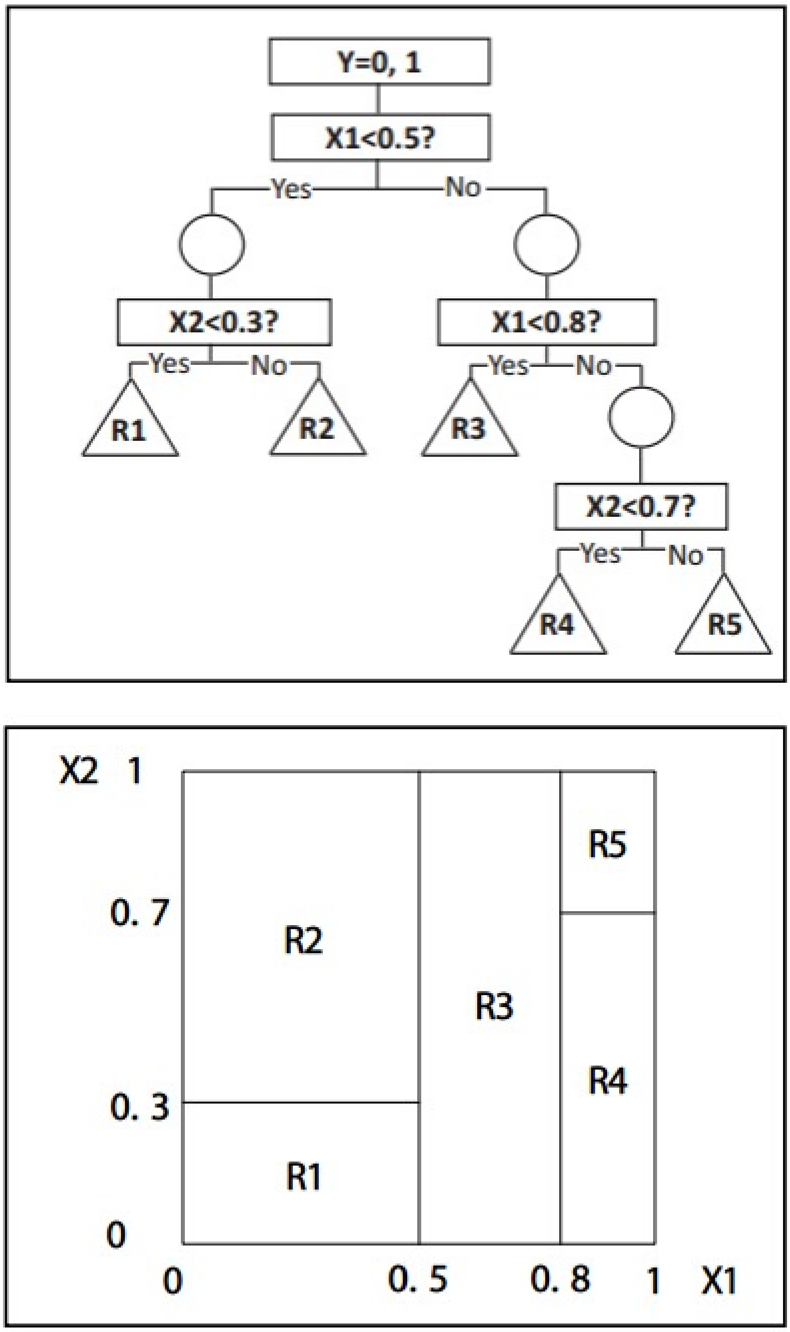

FIGURE 4.5: Decision Trees (Song and Lu, 2015)

4.6.2 Random Forest Regression

Random Forests are comprised of a number of decision trees and are thus an ensemble method of prediction, where multiple models are combined in order to yield a prediction. In the case of regression, we make predictions by taking the mean of each of the predictions made by the constituent models.

Random Forests are built on the idea of Bootstrap Aggregation (also known as Bagging), a procedure for reducing variance. The idea behind Bagging is as follows. Given a set of n independent observations $Z_1, ..., Z_n$, each will have a variance of σ^2. However, the variance of the mean of these observations falls to σ^2/n. Bagging thus reduces the variance of statistical methods which are high variance, as decision trees are.

The bootstrapping component of Bagging is required since we do not have access to multiple training sets. Therefore, we create these training sets artificially by taking repeated samples from our single training set, generating B boostrapped training sets. The individual decision tree models are then trained on each of the B training sets in order to get a predicted value for our output at given input values, which we yield by averaging the predictions across each model.

Random Forests build on this idea, but involve a slight tweak. In the same way, we create B bootstrapped training sets and grow a decision tree on each training set. However, rather than choosing one variable to split on from the entire set of variables (p), we instead consider a random subset of variables (m) at each split. This randomisation helps to control for overfitting. Random Forests therefore improve upon the benefits of bagged trees by further decorrelating the trees, which reduces variance more when we finally average the predictions. Here, the optimal value of m is given by $\sqrt{p}$, as can be seen in Table 4.2, along with the optimal values of the other hyperparameters according to random search.

Hyperparameter	**Values (Start, Stop, Step)**	**Optimal Value**
Number of Trees	1, 20, 1	2
Maximum Features	Auto, Sqrt	Sqrt
Maximum Depth	10, 120, 10	110
Minimum Samples Split	2, 10, 2	2
Minimum Samples Leaf	1, 4, 1	1
Bootstrap	True, False	True

TABLE 4.2: Random Forest Hyperparameter Search Grid

Algorithm 1 - Bagging Algorithm (For Regression) (Geicke, 2022)

1. Sample with replacement from training data to obtain bootstrap sample
2. Construct new tree T_1
3. Repeat B times to obtain B trees
4. Given a new point x, predict the average value
5. Make predictions on out-of-bag observations (test set)

4.6.3 XGBoost

XGBoost is a decision tree boosting method, where trees are grown sequentially – each tree is grown using information from the previously grown trees. This is in contrast with the trees being grown simultaneously, as is the case with Bagging and Random Forests. Unlike fitting a large decision tree to the data, potentially leading to overfitting, boosting learns slowly. Given the current model, we fit a decision tree to the residuals from the model. We then add this new decision tree into the fitted function to update the residuals. Each of these trees can be rather small with only a few terminal nodes, determined by the parameter d in the algorithm. By fitting small trees to the residuals, we slowly improve the fit of the function in areas where it does not perform well. The shrinkage parameter λ slows the process further, which allows for more and different shaped trees to attack the residuals. Thus, boosting converts many 'weak' learners into a more complex and effective predictor. A summary of the algorithm is outlined below (Geicke, 2022). In terms of hyperparameter tuning, I used random search to determine the optimal values. The grid of values searched and the optimal values yielded can be seen in Table 4.3.

Algorithm 2 - Boosting Algorithm (Geicke, 2022)

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set
2. For $b = 1, 2, ..., B$ repeat:
- Fit a tree $\hat{f}(x)_b$ with d splits $(d + 1)$ terminal nodes to the training data (X, r)
- Update $\hat{f}$ by adding in shrunken version of the new tree: $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}_b(x)$
- Update the residuals: $r_i \leftarrow r_i - \lambda \hat{f}_b(x)$
3. Repeat B times to obtain B trees
4. Given a new point x, predict the average value
5. Make predictions on out of bag observations (test set)

Hyperparameter	Values (Start, Stop, Step)	Optimal Value
Number of Trees	1, 20, 1	17
Maximum Depth	10, 120, 10	60
Learning Rate	0.001, 0.1, 0.001	0.098

TABLE 4.3: XGBoost Hyperparameter Search Grid

Chapter 5

Results

In terms of the data used to conduct the analysis outlined in 4, in each case I split the 365-day data set described in 3 into a training set containing observations from 1st June 2021 to 30th April 2022 and a test set consisting in data points from 1st May 2022 to 31st May 2022. Each model is trained on the former data set, with forecasts being made on and compared to the latter. Throughout this section I use the Root Mean Squared Error (RMSE) and the percentage of the mean value of the BEIR true values (the test data) that this represents as performance metrics, making comparisons of the performance both between the different model architectures and also within each model by comparing forecasts that include the OSS and OAS with those that do not. All figures are reported to 3 significant figures unless stated otherwise.

5.1 Vector Autoregression

The results of the VAR model can be seen in Table 5.1. Here, we consider the two variables of interest, OSS and OAS. In terms of OSS, its effect aligns with our expectations in terms of its direction, implying that the BEIR increases as online sentiment becomes more negative, also achieving statistical significance at the 10% level. However, its corresponding coefficient implies that it only has a very small effect in this way. In terms of OAS, curiously the model finds that BEIR increases as online attention falls, however this effect is not statistically significant. For the full regression results, please see Table 5.7.

Variable	Coefficient	Standard Error	p-value
Overall Sentiment Score	-0.000236	0.000128	0.065
Overall Attention Score	-0.00400	0.00286	0.889
BEIR	-0.0462	0.0557	0.407
Interest Rate	-0.560	0.707	0.428
Commodities Index	0.000818	0.00216	0.705
Unemployment Rate	0.00729	0.0145	0.614
Michigan Sentiment	0.00105	0.00220	0.633
M2 Money Supply	-0.000049	0.000054	0.368
GDP	-0.000001	0.000021	0.626
USD/EUR Exchange Rate	-0.420	0.436	0.335

TABLE 5.1: VAR Regression Results

After generating the model, I also conduct the Durbin Watson test to check for serial correlation of residuals (errors) as a robustness test. If serial correlation is present, this implies that there is some pattern in the time series that requires explanation by the model. In that case, we might increase the order of the model or include more predictors. Thus, the Durbin Watson test allows us to ensure that the model is sufficiently able to explain the variances and patterns in the time series. If the Durbin Watson test statistic is close to 2, this implies no significant serial correlation. As it tends to 0 this indicates positive serial correlation, while a value close to 4 shows negative serial correlation. The results of the test in Table 5.2 show that the test statistic for all variables is close to 2, hence we can be confident that there is no serial correlation.

Variable	Test Statistic
Overall Sentiment Score	2.23
Overall Attention Score	2.08
BEIR	2.00
Interest Rate	1.98
Commodities Index	2.06
Unemployment Rate	2.01
Michigan Sentiment	2.01
M2 Money Supply	2.03
GDP	2.02
USD/EUR Exchange Rate	2.10

TABLE 5.2: Durbin Watson Test Results

I then look to use the trained model to generate a forecast of the BEIR for the period 1st May 2022 to 31st May 2022. Prior to doing this, I reverse the first differencing transformation made in 4.2 to revert the data to the original scale. As can be seen in Figure 5.1, the model forecasts BEIR to start at 2.64% and slightly increase across the 31-day period, thus lacking the nuanced movements of BEIR in reality. This forecast achieves a RMSE of 0.121, representing 4.49% of the mean value of the BEIR test data. In this case, the model excluding OSS and OAS is slightly outperformed, only achieving a RMSE of 0.123.

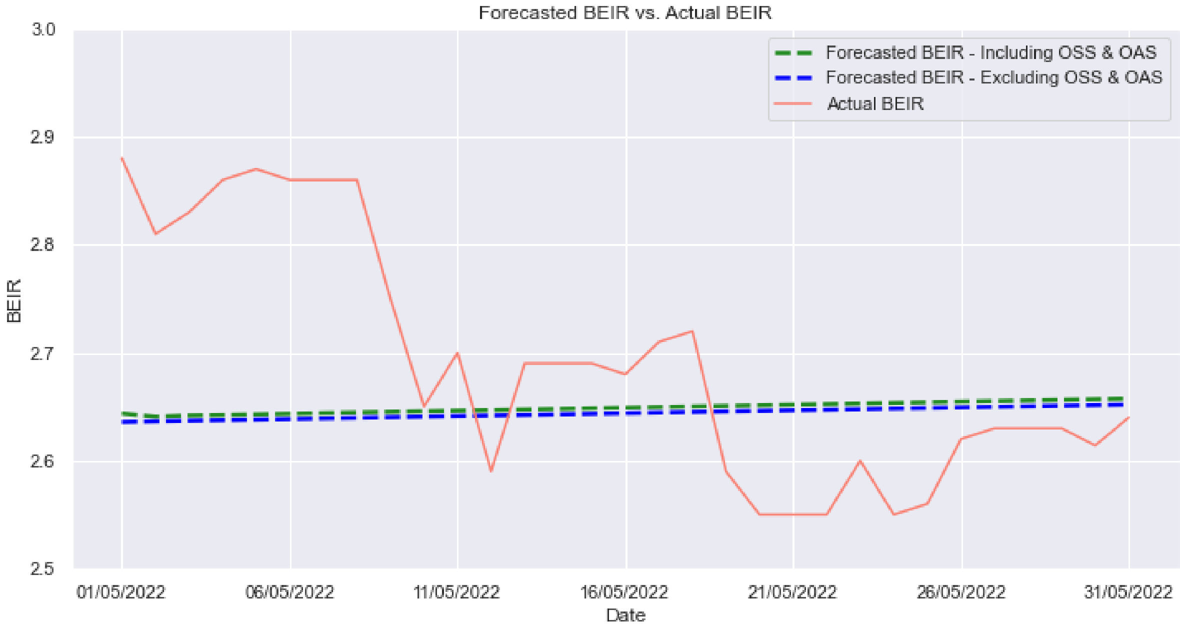

FIGURE 5.1: VAR Forecast

5.2 Least Absolute Shrinkage and Selection Operator

Considering the features selected, as can be seen in Table 5.3, it is encouraging that the LASSO model includes both the OSS and OAS. The corresponding coefficients also align more with our intuition. The OSS coefficient entails that a 1 unit increase in online sentiment is associated with a 0.000242 fall in the BEIR, whilst the OAS coefficient implies that a 1 unit increase in online attention is associated with a 0.00106 increase in the BEIR. Please note that the p-values are not reported, since the variables that are selected by the model will tend to be those that are indeed statistically significant - hence, by making interpretations based off of the p-values, as per Lee et al. (2016), we would be 'overfitting' to a particular realisation of the data. For the full regression results, please see Table 5.7.

The LASSO model forecasts BEIR to vary between values of 2.84 and 2.95 across the 31-day period. This forecast achieves a RMSE of 0.231, representing 8.59% of the mean value of the BEIR test data, however, is outperformed by the model excluding OSS and OAS which achieved a RMSE of 0.228.

Feature Selected	**Coefficient**
Overall Attention Score	0.00106
M2 Money Supply	0.0000033
Overall Sentiment Score	-0.000242
Unemployment Rate	-0.000358
USD/EUR Exchange Rate	-2.78
Interest Rate	-3.36

TABLE 5.3: Features Selected by LASSO

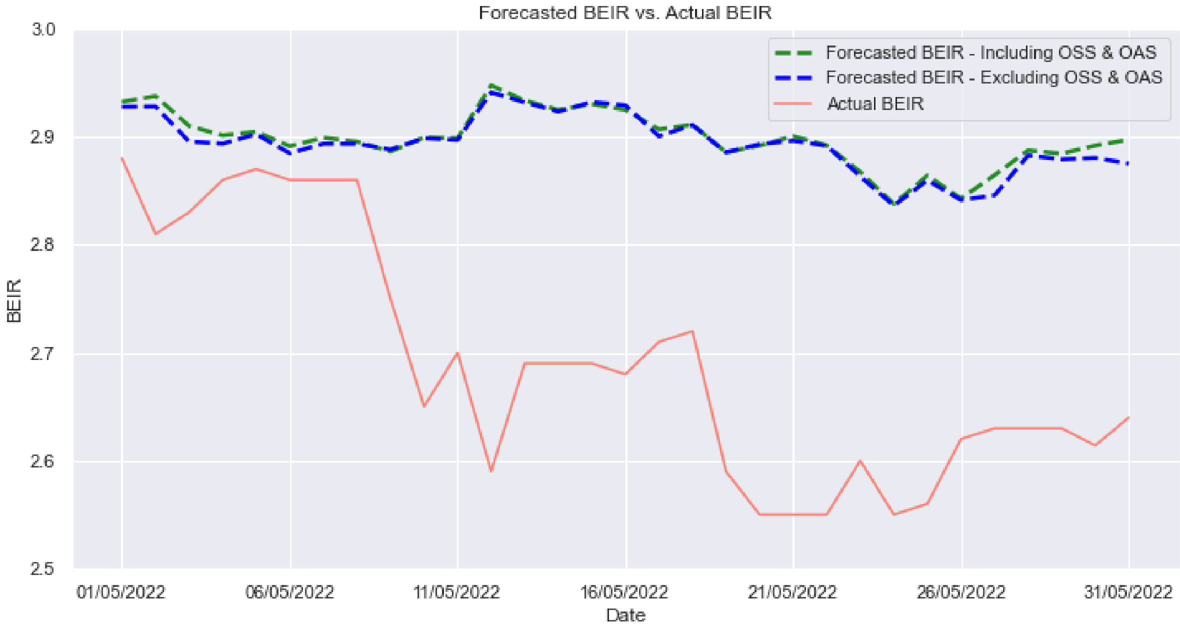

FIGURE 5.2: LASSO Forecast

5.3 Support Vector Regression

The regression coefficients of the SVR model are only available with the use of a linear kernel, however, I use a RBF kernel and so cannot compare them with those produced by VAR and LASSO. In terms of a forecast, the SVR makes a linear forecast of BEIR of 2.58% across the 31-day period. This forecast achieves a RMSE of 0.157, representing 5.83% of the mean value of the BEIR test data. However, despite this relatively good performance in terms of RMSE, its linear nature implies a lack of generalisability and so the model in its current form does not seem to be particularly useful. In addition, in contrast to some of the other model architectures, the SVR excluding OSS and OAS actually performs better, with a RMSE of 0.153.

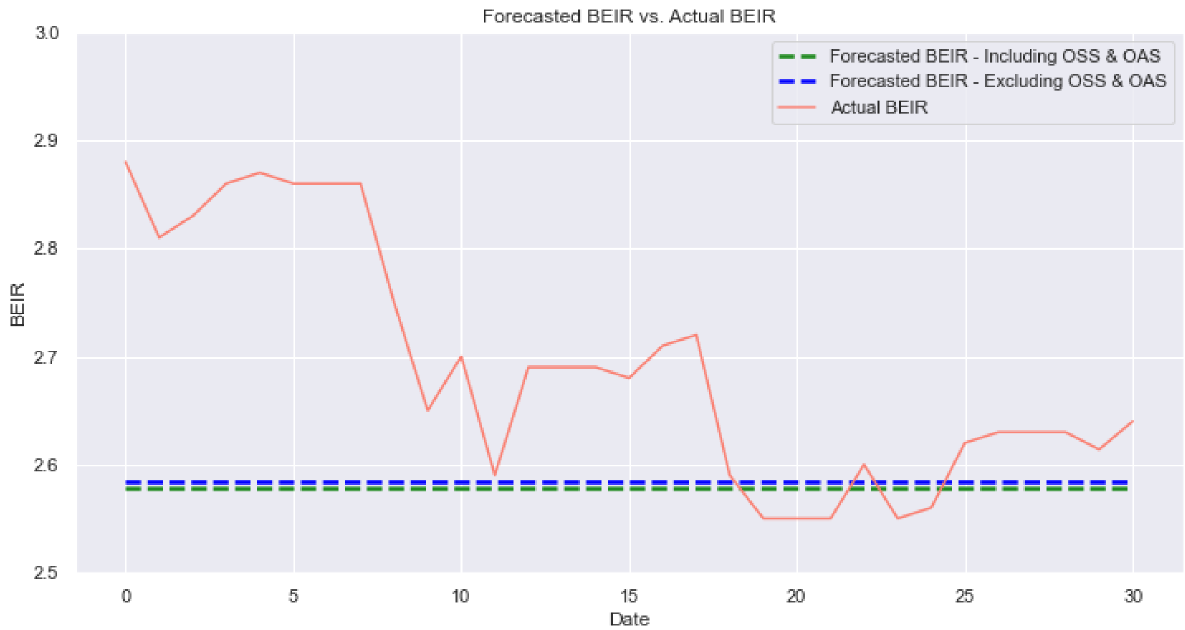

FIGURE 5.3: SVR Forecast

5.4 Tree Based Models

As a quick note, regression coefficients are not produced when using tree based models, and so are not referred to here.

5.4.1 Random Forest Regression

With regards to feature importance, the Random Forest Regression model finds that the Unemployment Rate and M2 Money Supply are the two most important determinants of BEIR, in terms of the total reduction of squared error. In contrast, it finds OSS and OAS to be relatively unimportant.

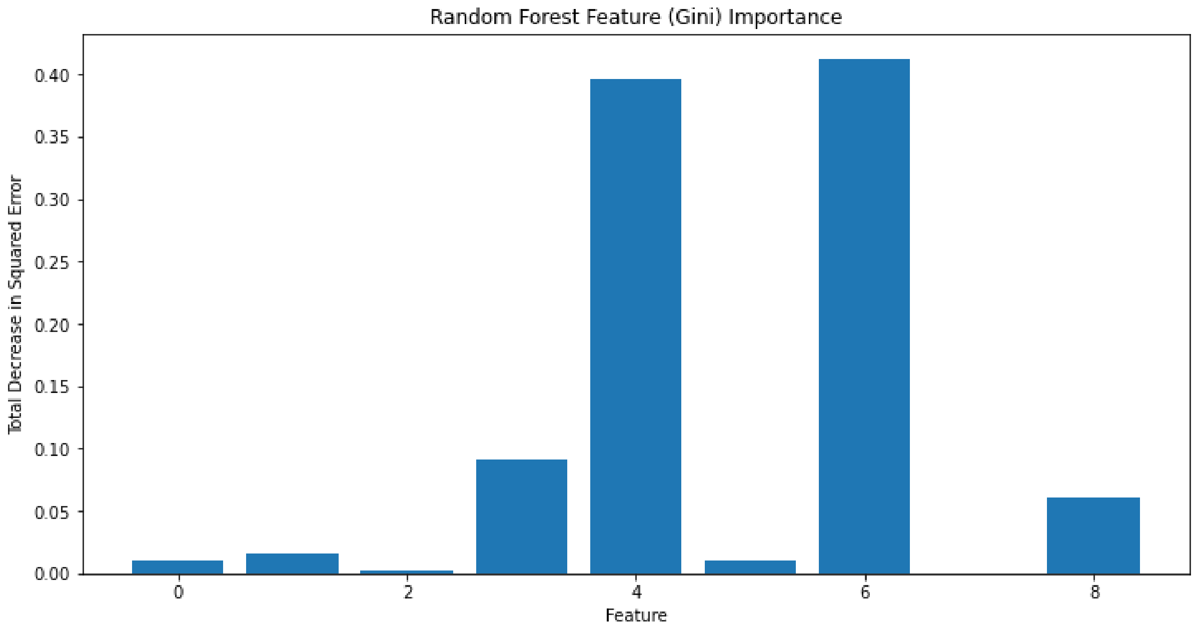

FIGURE 5.4: Random Forest Feature (Gini) Importance

Feature	Feature Selected	Total Decrease in Squared Error
0	Overall Sentiment Score	0.01084
1	Overall Attention Score	0.01547
2	Interest Rate	0.00302
3	Commodities	0.09125
4	Unemployment Rate	0.39631
5	Michigan Sentiment Score	0.00987
6	M2 Money Supply	0.41196
7	GDP	0.00039
8	USD/EUR Exchange Rate	0.06089

TABLE 5.4: Random Forest Feature Importance Scores

The Random Forest Regression model forecasts BEIR to vary between values of 2.86 and 2.93 across the 31-day period. This forecast achieves a RMSE of 0.195, representing 7.25% of the mean value of the BEIR test data. This is in comparison to the model that does not include OSS and OAS, which yields a RMSE of 0.234.

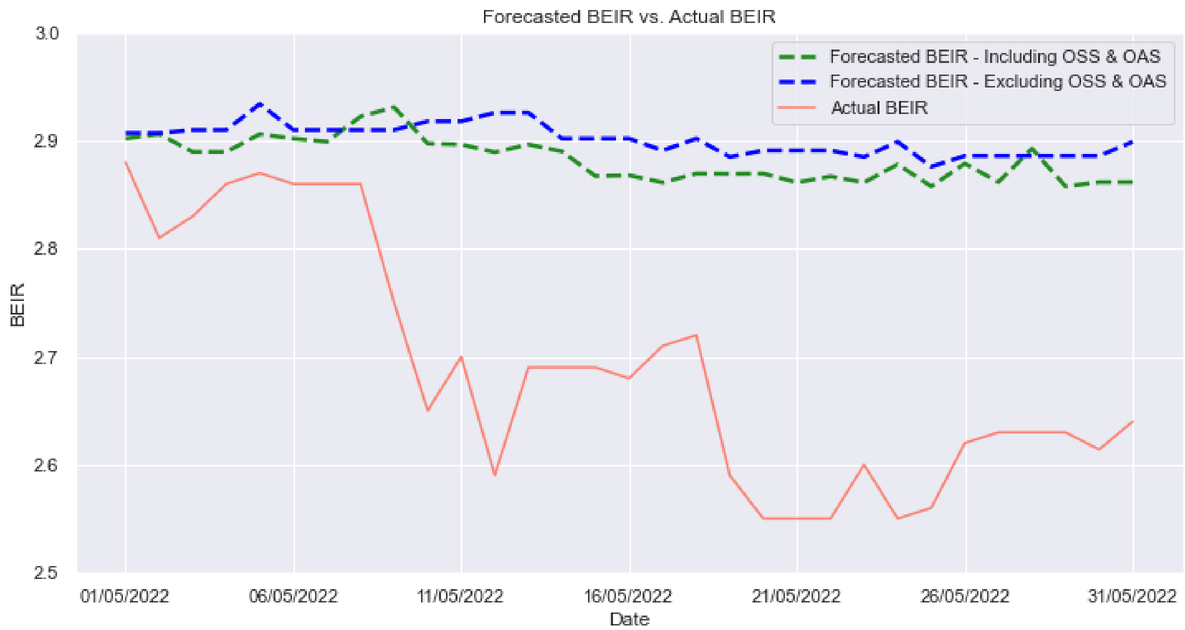

FIGURE 5.5: Random Forest Forecast

5.4.2 XGBoost

In terms of feature selection, as can be seen in Table 5.5 and Figure 5.6, XGBoost indicates that the five most important features, in terms of how many times each feature appears in a tree, are the Michigan Sentiment Score, M2 Money supply, USD/EUR Exchange Rate, OAS, and OSS, in ascending order of importance, with the final two variables being shown to be significantly more important than the rest. The fact that the model highlights the importance of OSS and OAS in conjunction with the fact that it also performs so strongly, as seen in Figure 5.7, strengthens the argument to employ them in inflation forecasting in future.

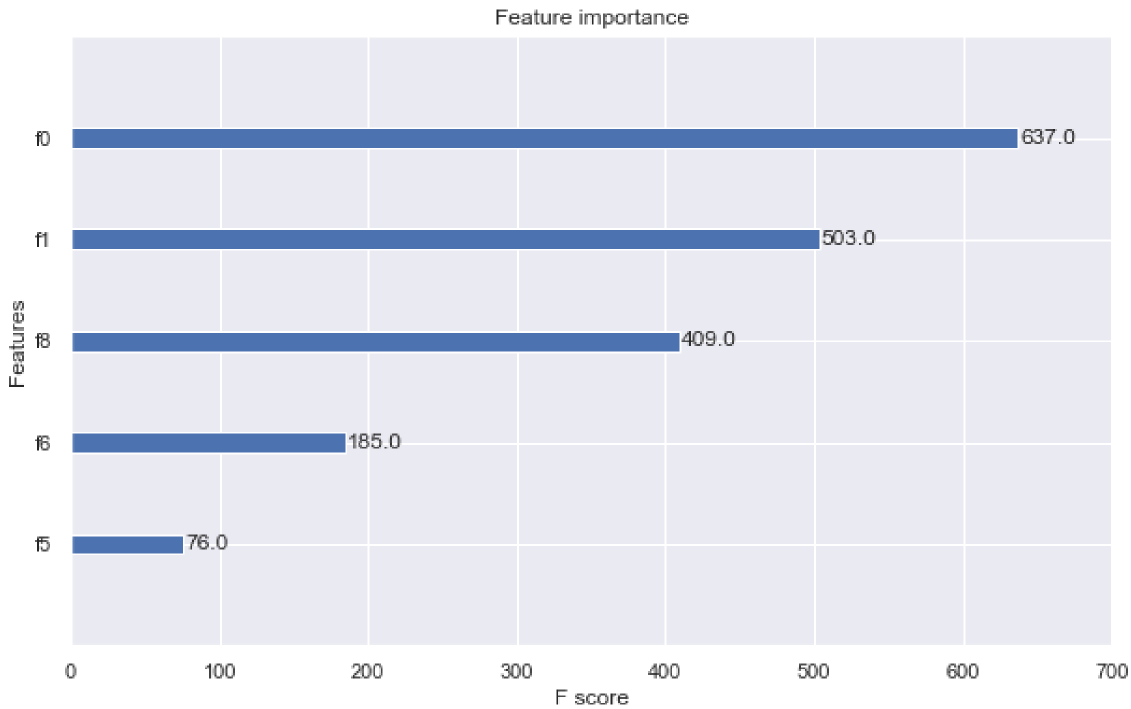

FIGURE 5.6: XGBoost Feature Importance

Feature	Feature Selected	F Score
0	Overall Sentiment Score	637
1	Overall Attention Score	503
8	USD/EUR Exchange Rate	409
6	M2 Money Supply	185
5	Michigan Sentiment Score	76

TABLE 5.5: Features Selected by XGBoost

To clarify with regards to hyperparameter tuning, using random search over grid search led to no improvement in model performance in this case. In terms of the forecast made by XGBoost, we can see below that it provides a more accurate prediction than other models, forecasting BEIR to vary between values of 2.65 and 2.68 across the 31-day period. This forecast achieves a RMSE of 0.113, representing 4.21% of the mean value of the BEIR test data, making a marked improvement upon the model that excludes the OSS and OAS, which achieves a RMSE of 0.227.

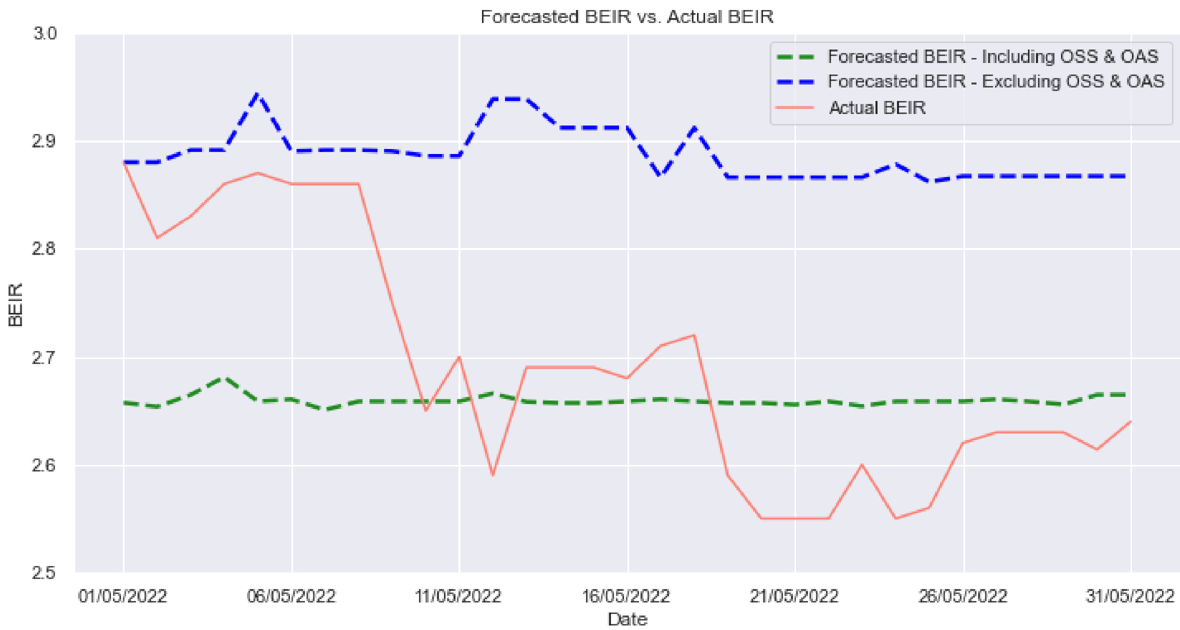

FIGURE 5.7: XGBoost Forecast

5.5 Summary

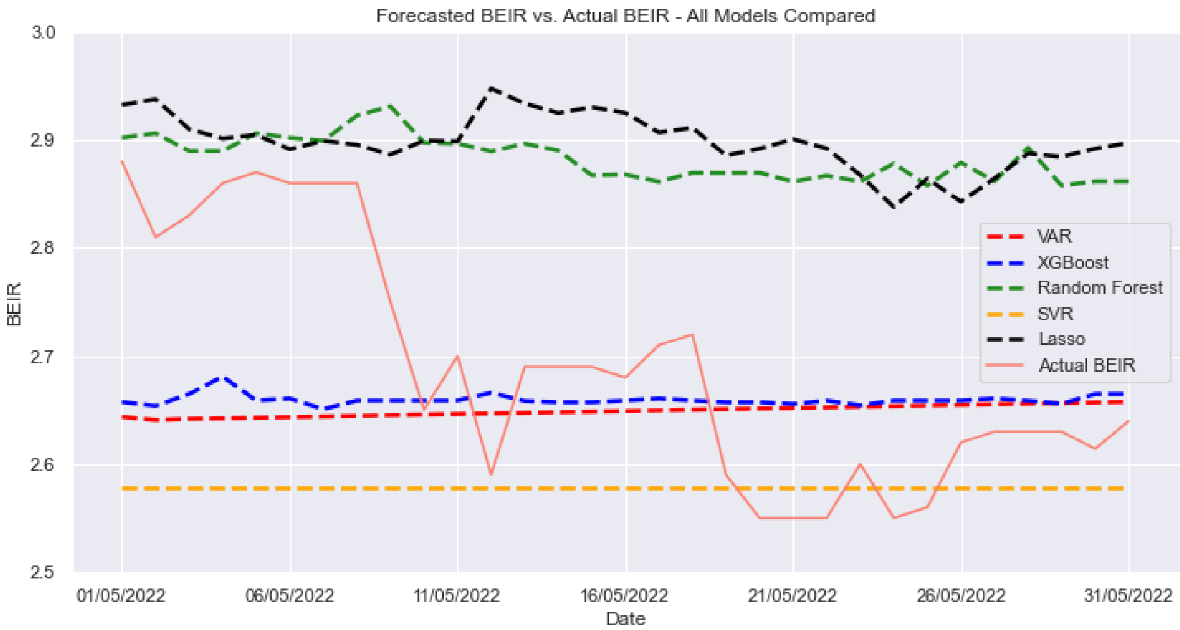

FIGURE 5.8: All Forecasts Compared

Overall, the best performing model was XGBoost, producing a forecast with the lowest RMSE of 0.113. A comparison of the performance of all models can be seen in Figure 5.8. Furthermore, we can take more encouragement from the fact that, as can be seen from Table 5.6, this model, which placed most importance onto both the OSS and OAS, outperformed all other model architectures. Also, in 3 out of the 5 cases, the models including OSS and OAS outperformed those that excluded them. Although the XGBoost forecast does not necessarily exactly track the actual values across the test period, the fact that it provides a forecast that lies around the intermediate values of the highs and lows of BEIR in May 2022 is encouraging - making forecasts with XGBoost whilst allowing for a margin of error either side of the prediction could therefore be extremely useful to policymakers.

OSS & OAS	**Included**		**Excluded**	
Model	**RMSE**	**% of Mean BEIR Test Value**	**RMSE**	**% of Mean BEIR Test Value**
XGBoost	0.113	4.21	0.227	8.46
VAR	0.121	4.49	0.123	4.56
SVR	0.157	5.84	0.153	5.67
Random Forest	0.195	7.25	0.234	8.70
LASSO	0.231	8.59	0.228	8.46

TABLE 5.6: Results Summary

TABLE 5.7: Regression Results (of Those Available)

	LASSO	VAR
overall_sentiment_score	-0.000242	-0.000236*
		(0.000128)
overall_attention_score	0.00106	-0.004
		(0.00286)
W2MNS	0.0000033	-0.000049
		(0.000054)
UNRATENSA	-0.00358	0.00729
		(0.0145)
USDEUR	-2.78	-0.42
		(0.436)
interest_rate	-3.36	-0.56
		(0.707)
commodities		0.000818
		(0.00216)
UMCSENT		0.00105
		(0.0022)
gdp		-0.000001
		(0.000021)
N	334	332
AIC		-12.6442

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Chapter 6

Conclusion

6.1 Limitations

Earlier I alluded to my desire to carry techniques and analyses from studies of how social media data can influence financial markets over to a study of how it might influence traditional economic metrics. However, it could be argued that the effects of the measures of interest OSS and OAS are not analogous. Whereas it is conceivable to think that an online conversation might directly impact an individual company's share price, for example, economic metrics are much wider, more nebulous concepts, with various constituent parts. Hence, changes in inflation at large cannot be effected as directly as changes in share prices in the stock market. Therefore, any causal inferences made from online activity to offline economic activity may be more difficult to justify. A possible development on this for future research could include the use of an instrumental variable to combat this.

6.2 Extensions

Extensions to this research could be both methodological and substantive in nature. In methodological terms, I believe deep learning techniques, such as Convolutional Neural Networks (You et al., 2015), could be applied to analyse the influence of text and video, in the form of memes for example, in this context, given the impact that meme culture has had more broadly. Including these forms of media, as opposed to text data alone, would allow for greater granularity in terms of analysing the sentiment of discussion occurring online. Furthermore, conducting this analysis over a longer time period, which could not be done in this case due to both time and computational constraints, would add further variation, thus allowing for a wider test of generalisability. In terms of content, the methods could be applied in other domains to see, for instance, how online discussion may impact the implementation of government policy, including U-turns. Finally, a further, more practical extension could be the creation of an inflation-sentiment index, which could be used to indicate the path inflation is likely to take in the near future, similar to those that already exist in the context of Fear vs Greed in the financial markets to see which emotions are driving the market at a given point in time (CNN, 2022).

6.3 Research Questions

RQ1 - Do social sentiment towards and attention to US inflation impact inflation expectations?

RQ2 - How can public expectations be operationalised?

With regards to RQ1, this paper finds that social sentiment and attention surrounding inflation do impact US inflation expectations, indicated by both the significance of the OSS variable in the VAR model and the highlighting of the importance of both OSS and OAS in the feature selection models. The hypothesis that both more negative sentiment and increased attention are likely to contribute to increased inflation expectations also seems justified to an extent. In terms of RQ2, we can see that we can indeed operationalise public expectations, thanks to the power of social media and sentiment analysis.

6.4 Final Remarks

Overall, this paper finds that models that include social media indicators, in the form of OSS and OAS, produce more accurate forecasts than those that omit them, with XGBoost providing the best prediction out of the models tested. Thus, the main contributions of this paper to the literature are as follows. Firstly, in terms of more practical contributions, it demonstrates how we can effectively operationalise public expectations in the form of online conversation, thanks to the power of social media and sentiment analysis, for use in economic research in conjunction with traditional economic methods. Secondly, in terms of the data used in this research, I combine the use of posts made on both Twitter and Reddit, something that is uncommon in the existing literature. Combining posts from these platforms in my analysis helps to alleviate the effects of the opposing biases that exist on the individual platforms, thus arguably providing more balanced and valid results. Finally, and arguably most significantly, this study pushes a new source of data, in the form of OSS and OAS, that indicates the preferences and behaviour of individual economic agents into the policymaking spotlight. As mentioned earlier, the main disadvantage of the typical surveys used is that, given their relative infrequency, they are likely to be unable to account for the rapid changes in economic activity that we see today and so relying on these more slow-moving sources of information means that policymakers will be less agile in responding to these changes.

In summary, as the internet has matured, it is no longer the escape from real life that it was in the early years after its conception. Rather, our digital identity has become a key part of our lives and has augmented them, as Jurgenson (2012) argues, rather than being something independent of them. Our real-life decisions are hugely influenced by the opinions and actions of others, and social media has massively increased our ability to spread those opinions to the masses in a mere instant. Therefore, information on social media platforms that provides financial or economic signals, of a much higher frequency than those available in the typical surveys, is something which policymakers can certainly leverage and use to their advantage. Just as there is a clear benefit to the use of monthly economic indicators in conjunction with the quarterly or semi-annual measures (Guzmán, 2011), I believe we can further increase the richness of our predictive models by incorporating daily data in the form of social media sentiment and attention. Given their successful use in financial markets (Nguyen, Shirai, and Velcin, 2015; Luo, Zhang, and Duan, 2013), it therefore seems almost negligent to ignore their potential impact in more traditional economic contexts. This seems especially true given policymakers' misdiagnosis of the inflation emerging in 2021 as merely 'transitory'. Therefore, we should embrace this new form of data, in combination with the historically used sources of information that have brought policymakers demonstrable success in the past. As such, I feel that this research will help to bring economic policymaking into the 21st century.

Bibliography

Afiyati, Afiyati et al. (Nov. 2022). "Challenges of Sarcasm Detection for Social Network : A Literature Review | Afiyati | JUITA: Jurnal Informatika". In: *JUITA* 8. URL: http://www.jurnalnasional.ump.ac.id/index.php/JUITA/article/view/8709.

Ante, Lennart (2022). *How Elon Musk's Twitter Activity Moves Cryptocurrency Markets*. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3778844.

Aromi, J Daniel and Martin Llada (2020). "Forecasting Inflation with Twitter". In: *Asociación Argentina de Economía Política*. URL: https://ideas.repec.org/p/aep/anales/4308.html.

Awad, Mariette and Rahul Khanna (2015). *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. First. Apress. ISBN: 978-1-4302-5990-9.

Batini, Nicoletta and Edward Nelson (Nov. 2001). "The Lag from Monetary Policy Actions to Inflation: Friedman Revisited". In: *International Finance* 4.3, pp. 381–400. ISSN: 1367-0271, 1468-2362. DOI: 10.1111/1468-2362.00079.

Baumgartner, Jason Michael (July 2022). *Pushshift Reddit API Documentation*. URL: https://github.com/pushshift/api.

Bergstra, James and Yoshua Bengio (2012). "Random Search for Hyper-Parameter Optimization". In: *Journal of Machine Learning Research*, p. 25. URL: https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf.

Bleier, Sean (2010). *NLTK's List of English Stopwords*. URL: https://gist.github.com/sebleier/554280.

Brown, Stephen J. and Jerold B. Warner (1985). "Using Daily Stock Returns: The Case of Event Studies". In: *Journal of Financial Economics* 14.1, pp. 3–31. URL: https://leeds-faculty.colorado.edu/bhagat/brownwarner1985.pdf.

Cabral, Inês da Cunha, Pedro Pires Ribeiro, and João Nicolau (Feb. 2022). "Changes in Inflation Compensation and Oil Prices: Short-Term and Long-Term Dynamics". In: *Empirical Economics* 62.2, pp. 581–603. ISSN: 0377-7332, 1435-8921. DOI: 10.1007/s00181-021-02032-4.

CNN (2022). *Fear and Greed Index - Investor Sentiment*. URL: https://www.cnn.com/markets/fear-and-greed.

Copypasta (2022). URL: https://copypastatext.com/.

Cortes, Corinna and Vladimir Vapnik (Sept. 1995). "Support-Vector Networks". In: *Machine Learning* 20.3, pp. 273–297. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/BF00994018.

Davidson, Thomas et al. (May 2017). "Automated Hate Speech Detection and the Problem of Offensive Language". In: *Proceedings of the International AAAI Conference on Web and Social Media* 11.1, pp. 512–515. ISSN: 2334-0770. URL: https://arxiv.org/pdf/1703.04009.pdf.

Drucker, Harris et al. (1996). "Support Vector Regression Machines". In: *Advances in Neural Information Processing Systems*. Vol. 9. MIT Press. URL: https://papers.nips.cc/paper/1996/file/d38901788c533e8286cb6400b4 Paper.pdf.

Fed, Philadelphia (2022a). *Livingston Survey - Summarizes Forecasts of Economists from Industry, Government, Banking, Academia - Philadelphia Fed*. URL: https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/livingston-survey.

— (2022b). *Survey of Professional Forecasters*. URL: https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/survey-of-professional-forecasters.

Francis-Devine, Brigid et al. (Apr. 2022). "Rising Cost of Living in the UK". In: *Background to Spring Statement 2022*. URL: https://researchbriefings.files.parliament.uk/documents/CBP-9500/CBP-9500.pdf.

FRED (2022). *Federal Reserve Economic Data | FRED | St. Louis Fed*. URL: https://fred.stlouisfed.org/.

Friedman, Milton (1972). "Have Monetary Policies Failed?" In: *American Economic Review* 62, pp. 11–18. URL: https://www.jstor.org/stable/1821518.

Geicke, Friedrich (2022). *MY474 Lecture 8: Ensemble Methods, Bagging, Random Forests, Boosting*. London School of Economics and Political Science.

Giles, Chris (June 2021). "Bank of England Says Inflation Likely to Exceed 3% in Coming Months". In: *Financial Times*. URL: https://www.ft.com/content/a2df9e5c-2dd4-4295-9938-890329c4e61f.

Girola, James A (2019). *The Treasury Breakeven Inflation Curve*. URL: https://home.treasury.gov/system/files/226/bri_mar2019_acc.pdf.

Greenberg, Stan (Nov. 2016). "Why Did Pollsters like Me Fail to Predict Trump's Victory?" In: *The Guardian*. ISSN: 0261-3077. URL: https://www.theguardian.com/commentisfree/2016/nov/15/pollsters-fail-predict-trump-victory-fbi-clinton.

Guzmán, Giselle (2011). "Internet Search Behavior as an Economic Forecasting Tool: The Case of Inflation Expectations". In: URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2004598.

Humphreys, Ashlee and Rebecca Jen-Hui Wang (Apr. 2018). "Automated Text Analysis for Consumer Research". In: *Journal of Consumer Research* 44.6, pp. 1274–1306. ISSN: 0093-5301. DOI: 10.1093/jcr/ucx104.

Hutto, C. and Eric Gilbert (May 2014). "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text". In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1, pp. 216–225. ISSN: 2334-0770. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550/14399.

Joshi, Aditya, Pushpak Bhattacharyya, and Mark J. Carman (Sept. 2017). "Automatic Sarcasm Detection: A Survey". In: *ACM Computing Surveys* 50.5, 73:1–73:22. ISSN: 0360-0300. DOI: 10.1145/3124420.

Jurgenson, Nathan (Mar. 2012). "When Atoms Meet Bits: Social Media, the Mobile Web and Augmented Revolution". In: *Future Internet* 4.1, pp. 83–91. ISSN: 1999-5903. DOI: 10.3390/fi4010083.

Kalamara, Eleni et al. (May 2020). *Making Text Count: Economic Forecasting Using Newspaper Text by Eleni Kalamara, Arthur Turrell, Chris Redl, George Kapetanios, Sujit Kapadia :: SSRN*. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3610770.

Karami, Amir et al. (June 2021). "Analysis of Geotagging Behavior: Do Geotagged Users Represent the Twitter Population?" In: *ISPRS International Journal of Geo-Information* 10.6, p. 373. ISSN: 2220-9964. DOI: 10.3390/ijgi10060373.

Kellner, Douglas (2020). "Jean Baudrillard". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2020. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/entries/baudrillard.

Konishi, Sadanori and Genshiro Kitagawa (2008). *Information Criteria and Statistical Modeling*. Springer. ISBN: 978-0-387-71886-6. URL: https://link.springer.com/book/10.1007/978-0-387-71887-3.

Lee, Jason D. et al. (June 2016). "Exact Post-Selection Inference, with Application to the Lasso". In: *The Annals of Statistics* 44.3. ISSN: 0090-5364. DOI: 10.1214/15-AOS1371. arXiv: 1311.6238 [math, stat].

Luo, Xueming, Jie Zhang, and Wenjing Duan (Mar. 2013). "Social Media and Firm Equity Value". In: *Information Systems Research* 24.1, pp. 146–163. ISSN: 1047-7047. DOI: 10.1287/isre.1120.0462.

Nguyen, Thien Hai, Kiyoaki Shirai, and Julien Velcin (Dec. 2015). "Sentiment Analysis on Social Media for Stock Movement Prediction". In: *Expert Systems with Applications* 42.24, pp. 9603–9611. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2015.07.052.

Politi, James and Colby Smith (May 2021). "Fed Governor Plays down Inflation Risks as 'Transitory Surge'". In: *Financial Times*. URL: https://www.ft.com/content/98c211f4-f097-4ee6-b9da-593422b6a539.

Rajadesingan, Ashwin, Reza Zafarani, and Huan Liu (Feb. 2015). "Sarcasm Detection on Twitter: A Behavioral Modeling Approach". In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM '15. New York, NY, USA: Association for Computing Machinery, pp. 97–106. ISBN: 978-1-4503-3317-7. DOI: 10.1145/2684822.2685316.

Saiidi, Uptin (July 2016). *Here's Why the Majority of Brexit Polls Were Wrong*. URL: https://www.cnbc.com/2016/07/04/why-the-majority-of-brexit-polls-were-wrong.html.

Schlosser, Stephan, Daniele Toninelli, and M. Cameletti (2021). "Comparing Methods to Collect and Geolocate Tweets in Great Britain". In: DOI: 10.3390/JOITMC7010044.

Shiller, Robert J. (Apr. 2017). "Narrative Economics". In: *American Economic Review* 107.4, pp. 967–1004. ISSN: 0002-8282. DOI: 10.1257/aer.107.4.967.

Song, Yan-yan and Ying Lu (Apr. 2015). "Decision Tree Methods: Applications for Classification and Prediction". In: *Shanghai Archives of Psychiatry* 27.2, pp. 130–135. ISSN: 1002-0829. DOI: 10.11919/j.issn.1002-0829.215044.

spaCy (2022). *spaCy · Industrial-strength Natural Language Processing in Python*. URL: https://spacy.io/.

Stanford (2008). *Stemming and Lemmatization*. URL: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.

Thomas, Lloyd B. (Dec. 1999). "Survey Measures of Expected U.S. Inflation". In: *Journal of Economic Perspectives* 13.4, pp. 125–144. ISSN: 0895-3309. DOI: 10.1257/jep.13.4.125.

Tilly, Sonja and Giacomo Livan (July 2021). "Predicting Market Inflation Expectations with News Topics and Sentiment". In: *arXiv:2107.07155 [cs]*. arXiv: 2107.07155 [cs].

Varian, Hal and Hyunyoung Choi (Apr. 2009). *Predicting the Present with Google Trends by Hal R. Varian, Hyunyoung Choi :: SSRN*. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1659302.

Vogels, Emily a, Brooke Auxier, and Monica Anderson (2021). *Partisan Differences in Social Media Use Show up for Some Platforms, but Not Facebook*. URL: https://www.pewresearch.org/fact-tank/2021/04/07/partisan-differences-in-social-media-use-show-up-for-some-platforms-but-not-facebook/.

Wang, Xuancang et al. (Feb. 2020). "A Hybrid Model for Prediction in Asphalt Pavement Performance Based on Support Vector Machine and Grey Relation Analysis". In: *Journal of Advanced Transportation* 2020, pp. 1–14. DOI: 10.1155/2020/7534970.

Wei, Wang (2016). *Achieving Inclusive Growth in China Through Vertical Specialization*. Chandos Publishing. ISBN: 978-0-08-100627-6.

Wojcik, Stefan and Adam Hughes (Apr. 2019). *Sizing Up Twitter Users*. URL: https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.

You, Quanzeng et al. (Sept. 2015). *Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks*. DOI: 10.48550/arXiv.1509.06041. arXiv: 1509.06041 [cs].

"Vector Autoregressive Models for Multivariate Time Series" (2006). In: *Modeling Financial Time Series with S-PLUS®*. Ed. by Eric Zivot and Jiahui Wang. New York, NY: Springer, pp. 385–429. ISBN: 978-0-387-32348-0. DOI: 10.1007/978-0-387-32348-0_11.