Erasmus University Rotterdam

Master Thesis Quantitative Finance

Analysing FOMC statements and their predictive power using the contextual natural language processing model BERT

Author

Daniel Vlaardingerbroek

Student number: 579736

Supervisor

dr. Anastasija Tetereva

Second Assessor

dr. Rutger-Jan Lange

Abstract

This paper studies whether the sentiment of the Federal Open Market Committee (FOMC) in their statements can predict the short-term returns of the two-year US bond futures. Each FOMC statement is dissected into five topics, for which the surprise in sentiment is determined separately. This paper finds that the surprise of the FOMC's outlook on future economic conditions and inflation is the only topic with significant predictive power. This is then used in combination with the surprise in the change of the federal funds rate to predict short-term returns. This paper finds that this combination can explain roughly 12% of the variation in returns for an in-sample analysis over the time period of 1999-2020. Furthermore, adding a proxy for current economic conditions and the size of the pre-FOMC announcement drift to the model almost doubles the fit. However, these findings only hold for the time period before 2009, which is shown in both an in-sample and out-of-sample framework.

August 2, 2022

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Contents

1 Introduction

The Federal Open Market Committee (FOMC) consists of twelve members, who hold eight scheduled meetings per year to discuss whether any changes in monetary policy should be made. The term “monetary policy” refers to the means by which government authorities can influence the pace and direction of overall economic activity, both in terms of the level of aggregate output and employment, but also the rate of inflation (Friedman, 2000). To assert this influence, the FOMC has the ability to change the Federal funds rate (FFR) or to employ newer tools such as Quantitative Easing (QE), purchases of private securities, negative interest rates, funding for lending programs, and yield curve control, all of which have been proven to be helpful in some circumstances (Bernanke, 2020).

Before 1994, the FOMC was deliberately opaque regarding their communication on changes in monetary policy, which they would not convey to the market participants until much later after the meeting was held. However, since 1994, the FOMC has significantly increased transparency in their communication (Mishkin, 2004). One of these was to start releasing a statement after each meeting at a scheduled time. This statement contains the main findings of the FOMC and any changes in monetary policy, but also signals any possible changes in monetary policy in future meetings (Hansen and McMahon, 2016). Thus, it is unsurprising that the release of these statements can cause a temporal increase in volatility, which is studied extensively by Farka (2009) and Rosa and Verga (2008), among others. Furthermore, Rosa (2011) shows that the content of these statements has significant predictive power for short-term returns on the major US indices and the VIX. This paper builds on this strand of literature, as it aims to further study the predictive power of the contents of a statement.

This research is relevant for many sectors within Quantitative Finance that deal with trading listed instruments. Knowing in which direction an instrument will likely move on a short timescale can be useful for the risk management of dealing rooms in banks, pension funds and market makers, as they can then hedge or close any open positions. Furthermore, this research also has the possibility to generate alpha for traders and hedge funds.

As mentioned earlier, the predictive power of FOMC statements is already studied by Rosa (2011), who finds that both the unexpected change in FFR and the contents of the FOMC statements provide significant predictive power for short-term returns. However, they cannot cover the out-of-sample performance of their model, since they use data that is not available until after the release of the statement to explain the variation in short-term returns. Thus, they can only show what moves the markets from a retrospective point of view. Furthermore, they manually read and score each statement, which can introduce human biases into the results. This paper investigates whether the novel natural language processing (NLP) model BERT can extract the sentiment of the FOMC statement and whether this too provides significant predictive power for short-term returns. Formally, the main research question is:

- *Does the (surprise in) sentiment of FOMC statements, as determined by the BERT model, contain statistically significant predictive power for short-term bond future returns?*

To support the main research question, this paper aims to answer the following three supporting research questions:

- *Do all sentences in a statement contribute to the market's reaction?*
- *Does the predictive power of the sentiment change over time?*
- *Are there other measures that can influence the market's reaction to the release of an FOMC statement?*

To answer these research questions, a Latent Dirichlet Allocation (LDA) model is used to extract the different topics within an FOMC statement. An LDA model is used since it is widely used for topic modelling and is already successfully applied to FOMC statements by Hansen and McMahon (2016). Then, using the BERT model, the sentiment of each discussed topic is determined. The BERT model is chosen since it is one of the best-performing contextual models for semantic text analysis (Devlin et al., 2019). BERT can detect the context of words within a sentence, which is a significant improvement over the more conventional "bag-of-words" approach that does not look at the position of words within a sentence. In addition, the standard BERT model can easily be further trained on domain-specific corpora, which improves its performance for specific language uses. Araci (2019) further trains the BERT model on finance-related text, which results in the FinBERT Model. This model has a 15% higher accuracy than the standard BERT model on sentiment classification tasks that are related to financial text. As such, the FinBERT model is applied to the FOMC statements instead of the original BERT model. As it has been shown that the market reacts to the surprise of the sentiment, instead of the actual sentiment itself, the markets' expectancy of each topic is determined using an OLS model, from which the news shock can be calculated. The news shock for each topic is then used in an OLS model to determine which topics contain predictive power for short-term returns. Lastly, the news shocks of the topics that show significant predictive power are used in a Macroeconomic Random Forest (MRF) model to investigate whether machine learning (ML) can increase the predictive power of short-term bond future returns. The MRF model is effectively an OLS model, but with time-varying parameters that are determined through a Random Forest (RF) model. Thus, the MRF model can incorporate reaction asymmetries which can influence the model's predictive power.

This research uses three types of data. Text data from FOMC statements is used to extract its topics and their respective sentiment. Furthermore, high-frequency bond futures data is used to measure the market's reaction to each statement's release, which this paper aims to predict. Lastly, a set of macroeconomic variables is used to determine the market's expectancy of the sentiment regarding specific topics.

This paper finds that the surprise of the FOMC's outlook on future economic conditions and inflation contains significant predictive power for short-term bond future returns. It also finds that this is the case for the surprise in the change of the FFR, which is already established by a large strand of literature. The combination of these two surprises can explain 11.9% of the variation in returns for an in-sample analysis over the time period of 1999-2020. Furthermore, when a proxy for current economic conditions and the size of the pre-FOMC announcement drift are added to the model it can explain 21.2% of the variation in returns over the same time period. However, the significance of the two surprises only holds for the time period before 2009. The out-of-sample analysis reflects the in-sample finding that both surprises lose their significant predictive power after 2009, and shows that the MRF model consistently underperforms across all time windows.

The remainder of this paper is structured as follows: In Section 2, the relevant literature to this research is discussed. Then, in Section 3, the data that is used for this research is discussed, in terms of its characteristics, origins and possible manipulations that are applied to the data. Section 4 and Section 5 describe the textual analysis that is used to extract the sentiment from the FOMC statements and the methodology regarding the use of this sentiment for the prediction of short-term returns. To discuss the results, the FOMC statements are firstly analysed in Section 6. Secondly, the in-sample results, two robustness checks and the out-of-sample results are discussed in Section 7. Lastly, the paper is concluded and recommendations for future research are given in Section 8.

2 Literature review

2.1 The effect of monetary policy changes on equity returns

2.1.1 Predicting post-announcement returns

Early studies on the effect of monetary policy on equity returns find that this effect is not insignificant, as Rozeff (1974) finds that an increase in the growth rate of money can positively affect equity returns. Furthermore, Thorbecke (1997) and Patelis (1997) give concrete evidence that monetary policy variables hold significant forecasting power for future equity returns. Specifically, by using an event study over the sample period of 1987-1994, Thorbecke (1997) finds that changes in the target funds rate have a statistically significant negative relationship with the 24-hour returns surrounding the release of the new target funds rate.

However, no distinction is made between expected and unexpected changes in target rates. To address this, Kuttner (2001) uses the futures market for Federal funds to obtain changes in the target funds rate. They separate changes into expected and unexpected components and find that the response of the interest rates to the expected changes is small. In contrast, if this change is unexpected, the response is large and significant. This method is then used by Bernanke and Kuttner (2005), Gürkaynak et al. (2005) and many others for further research. Bernanke and Kuttner (2005) find that an unexpected 25 bps (basis-point) cut in the target funds rate corresponds to an increase of 1% in equity prices. They use daily equity returns, which can cause endogeneity, as other economic data could be released on the same day. To correct for this, they add the values of any other macroeconomic releases on the same day to the regression. Another way to circumvent endogeneity is by using high-frequency data to narrow down the “event window” surrounding the FOMC announcement; this lowers the chances of any other economic releases happening in the same time period. Gürkaynak et al. (2005) use this method to perform the same research as Bernanke and Kuttner (2005), from which they draw the same conclusions.

Furthermore, Gürkaynak et al. (2005) show that the market not only reacts to target funds rate changes, but also to the FOMC’s stand on the future path of monetary policy. This future path is closely related to the announcements made by the FOMC. Rosa and Verga (2008) build onto this strand of literature by creating an indicator that captures the sentiment of the introductory statements for the ECB's monthly press conferences. They do this by manually reading these statements, and mapping words and sentences to a value that indicates how dovish/hawkish they are. By predicting the market's expectancy of the ECB's stance on future policy and subtracting this from ECB's actual stance, they can capture the surprise. They find that this surprise has a significant and sizeable impact on Euribor futures returns. Rosa (2011) follows a very similar methodology, but then for the FOMC's impact on three leading US indices and the VIX. The same conclusion as Rosa and Verga (2008) follows for the three US indices. Specifically, their framework explains about 20% of the variation in US equity returns in the event window.

2.1.2 Reaction asymmetries

Orphanides (1992) is one of the first to provide empirical evidence that, depending on the state of the economy, reactions of financial assets to macroeconomic news announcements may vary. Guo (2004), Andersen et al. (2007), Chuliá et al. (2010) and Law et al. (2018), among others, build onto this idea and show that the reaction of stock prices to changes in target rates is sensitive to the current phase of the business cycle. Specifically, they find that the reaction is larger (smaller) when business conditions are bad (good).

Chuliá et al. (2010) and Farka (2009) also show that negative surprises in the target rate change have a more substantial impact on equity prices than positive surprises. This contradicts Bernanke and Kuttner (2005), who do not find any evidence regarding this asymmetry. However, Chuliá et al. (2010) argue that this is because of the use of daily stock data instead of high-frequency data.

Furthermore, Ehrmann and Talmi (2017) find that similar statements of the Canadian central bank reduce short-term volatility after the release of such a statement. In contrast, changes in statements after a long streak of similar statements cause a much higher short-term volatility.

Next to these three proven reaction asymmetries, this paper investigates whether the pre-FOMC announcement drift also impacts the reaction. In the 24 hours leading up to the statement's release, assets tend to have a lot of upwards momentum. This phenomenon is called the pre-FOMC announcement drift and has two main economic interpretations. The first one is the so-called "announcement premium", where the price gets driven upwards to account for the possible uncertainty that an FOMC statement brings (Hu et al. (2019) and Wachter and Zhu (2018)). Secondly, Vissing-Jorgensen et al. (2015) provide evidence of systematic informal communication of Fed officials with the media and financial sector as the information transmission channel, which could explain the pre-FOMC drift, as investors then start incorporating new information before the statement is released. Furthermore, Lucca and Moench (2015) find that pre-FOMC returns are higher in periods when the slope of the Treasury yield curve is low, implied equity market volatility is high, and when past pre-FOMC returns have been high. These interpretations all imply that the magnitude of the pre-FOMC announcement drift could impact the short-term returns after the release of a statement.

This paper aims to contribute to this strand of literature by combining all four mentioned asymmetries in one single framework for predicting short-term bond future returns around FOMC announcements. Specifically, this paper uses the framework from Rosa (2011). Their framework does not take any asymmetric effects into account, which could increase the predictive power for bond future returns.

2.2 Natural Language Processing models

Natural language processing (NLP) is a theory-motivated range of computational techniques for the automatic analysis and representation of human language (Cambria and White, 2014). This analysis is often split up into two parts: syntactic analysis and semantic analysis. Syntactic analysis is regarding the structure of words within a sentence, whereas semantic analysis looks at the meaning of words within a sentence. In this paper, semantic analysis is applied to FOMC statements, by extracting the sentiment from each sentence within a statement. As such, this section discusses commonly used NLP models for semantic analysis and their respective advantages/disadvantages.

2.2.1 FFN

The first neural network to be used for language modelling is the feed forward neural network (FFN), which is done by Bengio et al. (2003). The model is designed to address the shortcomings of the n-gram model, which is introduced in Brown et al. (1992). The n-gram model focuses on predicting the next word when the n-1 previous words in the sentence are given. This model forms the basis of modern language modelling for speech-to-text analysis (Schwenk, 2004), but has a significant limitation in that it has minimal use of context, as the number of parameters to be estimated grows exponentially when n becomes larger. Bengio et al. (2003) show that the FFN model can achieve a 24% lower perplexity (a performance measure that indicates how good the fit of the model is, where a lower perplexity indicates a better fit) than the n-gram model and show that they can take advantage of more words to provide context, without having an exponentially larger set of parameters.

2.2.2 RNN

A downside of the FFN model is that it uses a fixed number of words to gather the context of a sentence. This needs to be specified before training, which makes the model inflexible. To overcome this problem, Kombrink et al. (2011) use a recurrent neural network (RNN) for language modelling. They show that the RNN model has superior performance over the n-gram model while retaining flexibility for more uses outside the scope of training, as the RNN model does not use a fixed number of words for their context. They do not have an explicit comparison between the FFN and RNN model, but show that the RNN model has half the perplexity of the n-gram model, compared to the 24% lower perplexity of the FFN model. The main problem for the RNN model is that it is difficult to train using backpropagation through time, as the RNN model suffers from the so-called vanishing gradient problem, which is studied in detail by Hochreiter and Schmidhuber (1997). The vanishing gradient problem entails that in an RNN model, the gradient of the hidden layer from the previous word, multiplied by a number, is used as input for the hidden layer for the current word. Depending on the value of this multiplier and the length of the sentence, earlier gradients can then decay (grow exponentially) if the multiplier is smaller (larger) than one.

2.2.3 LSTM

To overcome the vanishing gradient problem of the RNN model, Hochreiter and Schmidhuber (1997) introduce the Long Short-Term Memory (LSTM) model. Sundermeyer et al. (2012) are the first to use the LSTM model for NLP and achieve an 8% lower perplexity compared to the RNN model, which makes it the best performing NLP model at that time.

2.2.4 Transformers

All earlier NLP models have one major constriction: The input must be given sequentially. This significantly increases the time it takes to train the model and reduces the amount of context it can use. To overcome this, Vaswani et al. (2017) introduce the Transformer, a network that allows for parallel computing. This means that the whole sentence can be processed simultaneously, which significantly reduces training time and allows the model to use the whole sentence to determine the context of a specific word. The Transformer network is made to solve language translation tasks and outperforms the at-the-time best model by more than 7.5%.

An important aspect of the Transformer network is the concept of self-attention, which enables the Transformer to capture contextual relationships of words in a sentence. This is done by determining the relevancy of each word in a sentence with respect to all other words in the sentence. The advantage of self-attention is that the computational complexity scales with $O(\mathit{1})$, opposed to the $O(n)$ of most recurrent networks, where n is the length of the sentence. Thus, self-attention looks at all words in a sentence simultaneously, while recurrent networks have to cycle through each word in the sentence to capture contextual relationships. The Transformer is the first network to solely rely on this technique, which causes the training time of the model to be significantly faster than existing NLP models for translation tasks.

To perform language translation tasks, the Transformer network uses an encoder-decoder-based architecture, as seen in the Transformer schematic in Figure 1. The encoder takes the to-be-translated sentence as input and transforms it into a set of vectors that encapture the meaning and context of each word in the sentence. The decoder then uses these vectors, in combination with the previously translated word, to translate the following word in the sentence. It does this in sequential order until the end of the sentence is reached. Loosely speaking, by training the model on a set of sentences and their corresponding translations, the encoder learns how to understand language and its context, while the decoder learns how to map the words in one language to the corresponding words in the other language.

The Transformer network forms the basis for more advanced NLP models that can handle various NLP tasks, instead of just one. For example, Radford et al. (2018) remove the encoder from the Transformer network and stack 12 decoders to obtain the OpenAI

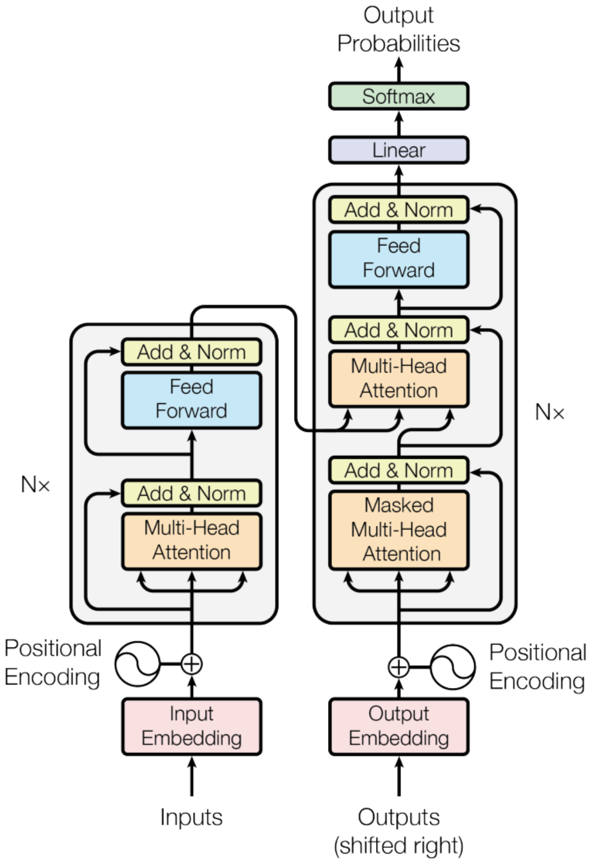


Figure 1: High-level schematic of the Transformer network. Copied from Vaswani et al. (2017).

GPT model, which can handle textual entailment, question answering, semantic similarity assessment, and document classification. Moreover, Devlin et al. (2019) remove the decoder and stack 12 encoders onto each other to obtain the base model of BERT, which sets new high scores for eleven different NLP tasks. From these eleven NLP tasks, two are sentence classification tasks. This makes BERT the best performing NLP model for sentence classification at the time, albeit for non-finance related text. However, the BERT model can easily be further trained on domain-specific corpora, which makes the BERT model highly flexible and ensures that it can be used for very different use cases. For example, Araci (2019) further pre-trains the BERT model on finance-related text. This results in the FinBERT model, which has a 15% higher accuracy than the standard BERT model on sentiment classification tasks that are related to financial text. Since then, numerous improvements are suggested (e.g. by Liu et al. (2019), among others), which results in models that have an even better performance than the original BERT model. However, none of them are further pre-trained on finance-related text. This is the main reason why the FinBERT model is used instead of the improved successors of the standard BERT model.

Even though BERT is designed to be used for differing use cases, no major research is published regarding the use of BERT models on FOMC speech. This paper aims to contribute to this strand of literature by applying the FinBERT model to FOMC statements.

3 Data

For this research, three main types of data are used: text data derived from FOMC statements, high-frequency bond futures data and macroeconomic data. This section covers all the data in terms of its characteristics, origins and possible manipulations that are applied to the data.

3.1 FOMC statements

After each meeting of the FOMC, a statement is released in which they summarise the main findings of the meeting. This started in 1994, when the FOMC decided to explicitly announce monetary policy changes at 14:15 EST after the meeting, but only if they decided to change the monetary policy. From the 18th of May 1999 onwards, the FOMC started releasing a statement after each meeting, regardless of any changes in monetary policy. From this point onwards, the FOMC has consistently released statements after each meeting that also include their opinions on current and future economic conditions (Farka and Fleissig, 2013). Thus, to avoid endogeneity, the sample period for this research starts on the 18th of May 1999.

The FOMC archives all their released statements, which can be found on their website[1]. Using the *BeautifulSoup* package in Python, the statements and their respective release dates are scraped. The time of the statement's release is not stated in their archives, but it can be found on external websites[2]. The time of the release is given in EST (GMT-4) or EDT (GMT-5), depending on the time of year.

Table 1: Descriptive statistics of the FOMC statements over the whole sample period.

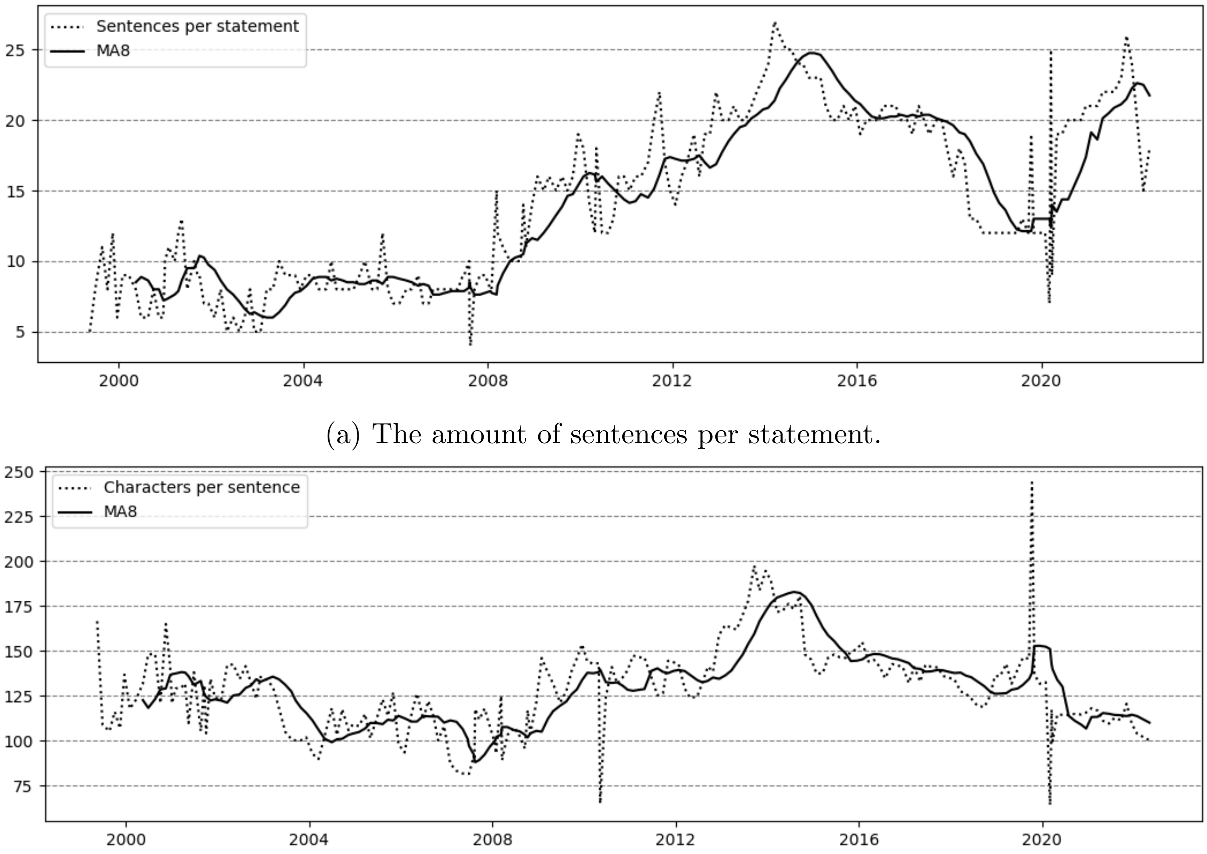
	Full sample
Number of statements	198
Number of sentences	2450
Number of words	73513

Between the 18th of May 1999 and the 4th of May 2022, a total of 198 meetings have been held, of which 15 were unannounced. An interesting note to the statements is that

[1] https://www.federalreserve.gov/monetarypolicy/fomc_historical_year.htm

[2] https://www.investing.com/economic-calendar/fomc-statement-398

they have become substantially more complex since the financial crisis. This phenomena is studied in-depth by Hernández-Murillo et al. (2014) and Coenen et al. (2017), and can be seen in Figure 2. This figure plots the number of sentences per statement and the number of words per sentence. It is clear that after the financial crisis (2009) both ratios start to increase significantly.



(a) The amount of sentences per statement.

(b) The amount of characters per sentence in a statement.

Figure 2: The amount of sentences per statement and characters per sentence in each statement, together with the moving average (MA) over the last eight values. The sample period is 1999-2022.

3.2 Financial data

3.2.1 High-frequency bond futures

To capture the market's response to the release of the statement, high-frequency bond futures data is used. Specifically, the two-year US bond future is used, since short-term bonds have a higher sensitivity to monetary policy surprises (Kuttner, 2001). Thus, the market's response is easier to capture. For the remainder of this paper, the two-year US bond future will be referred to as TU, which is its ticker symbol. The dataset runs from the 1st of April 1993 up to the 6th of January 2021. Thus, the sample period for this paper will end with the last statement of 2020. The data is given in an open, high, low and close (OHLC) format and is updated per minute, which can cause the prices to be contaminated with microstructure noise, such as the bid-ask bounce. Due to this noise, the observed fluctuations in the bond futures prices become less representative of the actual variance in the prices (Zhang et al., 2005). To overcome this noise, it is common to reduce the update frequency of the prices to once per five minutes. This is done by taking the last closing price of five one-minute intervals. Per example, for the closing price of the interval between 8:00 and 8:05, the closing price of the one-minute interval between 8:04 and 8:05 is substituded. Furthermore, the time in this dataset is given in CST (GMT-6) or CDT (GMT-5). Thus, in terms of timezone, the bond futures are an hour behind the releases of the statement. This paper aims to predict the log-returns between t-10 and t+40, where t is the time of release of the statement and is given in minutes. Thus, the log returns are calculated as follows:

$$\Delta P_t = \log(\frac{P_{t+40}}{P_{t-10}}) * 100\%. \tag{1}$$

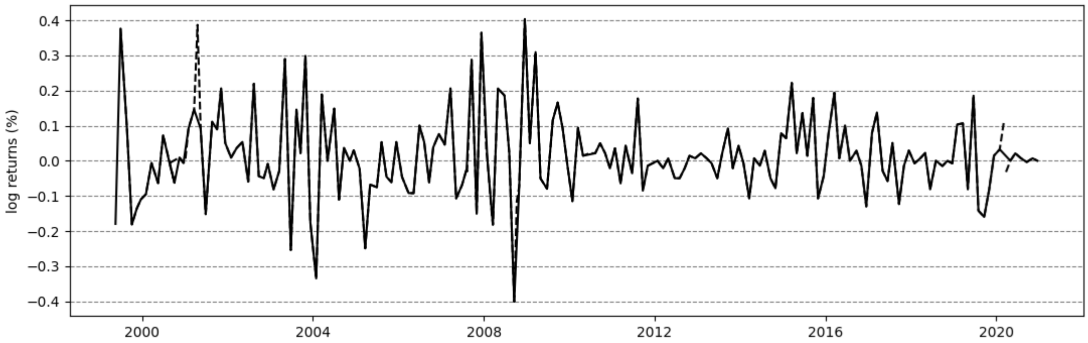
For the remainder of this paper, these returns will be referred to as the market reactions. The descriptive statistics of the market reactions, excluding unannounced meetings, are shown in Table 2.

Table 2: Descriptive statistics

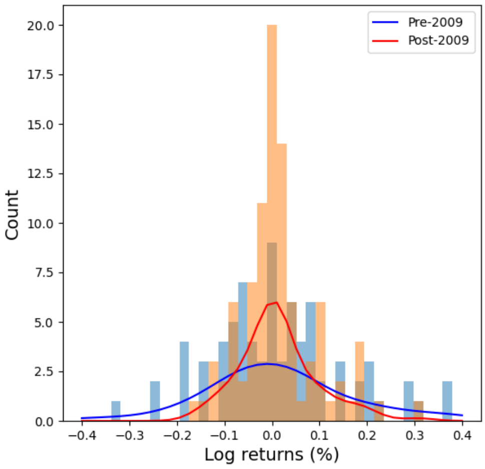
	Mean	Min	Max	St. dev.	Skewness	Kurtosis
ΔP_t	0.013	-0.401	0.403	0.119	0.357	4.844

This table shows the descriptive statistics for the 40-minute market reactions of the US 2-year Treasury note future (TU). The sample period is from 1999-2021.

The market reactions are plotted in Figure 3(a), where the dashed line represents the unannounced meetings. This figure shows that unannounced statements are not often met with a much larger reaction than announced statements. Furthermore, the market reactions are smaller after the financial crisis. To further investigate this, the estimated distribution of the market reactions is plotted for both the sample period of pre-2009 and post-2009 in Figure 3(b). This figure provides further evidence that the size of the market reactions is smaller after 2009, since the distribution of the pre-2009 sample has much fatter tails.



(a) The magnitude of the 40-minute market reactions.



(b) A histogram and its estimated distribution from kernel-density estimation using Gaussian kernels.

Figure 3: In these two figures, the magnitude of the 40-minute market reactions and their estimated distribution is plotted for the sample period of 1999-2021.

3.2.2 Federal funds futures

One of the variables that is used by Rosa (2011) to predict the market's expectancy of the content of the following FOMC statement is the difference between the three-month-ahead federal funds futures contract and the current federal funds futures contract. The federal funds futures contract is a future that tracks the federal funds rate (FFR), according to the following equation:

$$P = 100 - r, \tag{2}$$

where r is the current federal funds rate, and P is the price of the future. Thus, the price of the current federal funds futures contract directly reflects the current FFR, while the price of the three-month-ahead federal funds futures contract reflects the market's expectancy of the FFR in three months. The difference between these two contracts, as shown in Figure 4, then directly reflects the market's expectancy of any future movements in the FFR. A positive difference shows that the market expects a rise in FFR and vice-versa. To obtain the price difference, daily updated futures data is used.

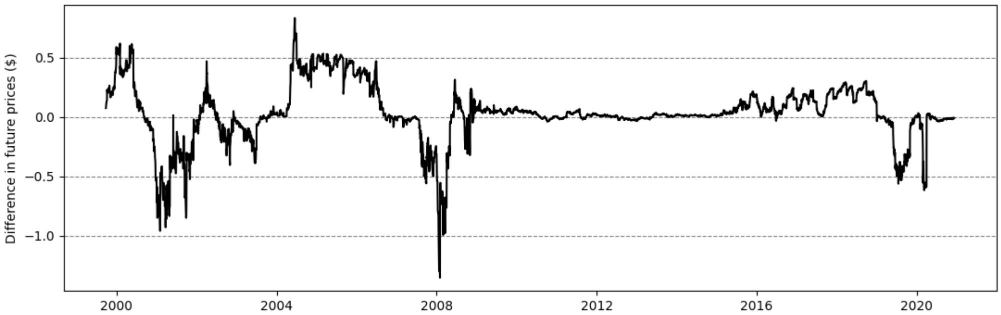


Figure 4: The price difference between the three-month-ahead federal futures contract and the current federal funds futures contract over the whole sample period of 1999-2021.

3.3 Macroeceonomic data

In addition to the difference in federal funds futures contracts, Rosa (2011) uses two macroeconomic variables to predict the market's expectancy. The first variable is the Purchasing Managers Index (PMI)[3], which is an index that shows whether purchasing managers think that market conditions are expanding, contracting or staying the same. Afshar et al. (2007) show that the PMI is (partially) responsible for large variations in GDP, and Koenig et al. (2002) find evidence that the PMI can be associated with rising short-term interest rates. The second variable is a monthly survey of consumers on their view about future movements of inflation (inflation expectation)[4], conducted by the University of Michigan. Both the PMI and the inflation expectation are updated monthly. The reason to use surveys, instead of backwards-looking measures of economic activity and inflation, is that survey data takes the forward-looking nature of the statements better into account, as stated by Rosa (2011).

[3] https://ycharts.com/indicators/us_pmi

[4] https://fred.stlouisfed.org/series/MICH

4 Methodology - Textual analysis

This research extracts two main pieces of information from all FOMC statements. Firstly, the topics that are discussed in each statement and its associated sentiment are extracted, of which its methodology is discussed in Section 4.1 and Section 4.2. Secondly, the similarity between statements is extracted, which is discussed in Section 4.3.

4.1 Topic analysis

To extract the topics and their associated sentences from an FOMC statement, the Latent Dirichlet Allocation (LDA) model, introduced by Blei et al. (2003), is used. This model is widely used in the literature to extract topics from corpora, because of its simplicity and the fact that it is proven to classify text in a similar way to humans (Chang et al., 2009), which makes the results relatively easy to interpret. To support this, Hagen (2018) finds that 87% of all LDA-generated topics make sense to human judges. In addition, the LDA model is successfully used to extract topics from FOMC speech, as demonstrated by Hansen and McMahon (2016) and Jegadeesh and Wu (2017). Due to the high interpretability and its earlier success with FOMC speech, this paper uses this model to extract topics from FOMC statements. A theoretical background of the LDA model can be found in Appendix A.1.

As it is known that FOMC statements consist of more than one topic (Hansen and McMahon, 2016), the aim is not to find the main topic of a statement as a whole, but rather of each sentence individually within a statement. Then, each statement can be dissected into a set of sentences that belong to each of the k topics. If this is done for all the released statements, the distribution of discussed topics over time can be inferred. Furthermore, the sentiment of each discussed topic within a statement is inferred by performing sentiment analysis on each set of sentences that belong to the k topics. Then, by averaging the sentiment of each set of sentences, the sentiment regarding a specific topic in a statement is determined.

As input for the LDA model, a corpus of sentences from all released statements in the time period of 1999-2022 is generated. Thus, unannounced statements are also included. These sentences are then pre-processed by splitting them up into words (1), removing all “stop words” (set of common words within a language, such as “the” and “is”) and special tokens (2), and stemming all remaining words (3). Stemming a word means to remove the inflectional ending of a word, e.g. reflecting, reflected, and reflects can all be reduced to reflect. An example of this process is given in Table 3. Next to these individual words, *n-grams* can be added to the corpus. *n-grams* are n subsequent words that co-occur frequently, and can provide more context. However, adding these n-grams tends to lead to an explosion in the size of the corpus, due to the combinatorial nature of *n-grams* (Denny and Spirling, 2018). Therefore, for simplicity, this paper does not include them in the corpus.

Table 3: Example of pre-processing a sentence for the LDA model.

Sentence	Inflation has risen, largely reflecting transitory factors.
(1)	['inflation', 'has', 'risen', ',', 'largely', 'reflecting', 'transitory', 'factors', '.']
(2)	['inflation','risen', 'largely', 'reflecting', 'transitory', 'factors']
(3)	['inflat', 'risen', 'larg', 'reflect', 'transitori', 'factor']

The LDA model has three parameters that can be optimised: The number of topics k and the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which are used as parameters for two Dirichlet distributions, as discussed in Appendix A.1. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are set to $\frac{1}{k}$, similar to Hoffman et al. (2010). The choice of k depends on the interpretability of the results and the goals of the analysis (Blei and Lafferty, 2009), along with the type of data that is analysed. For example, a corpus containing news articles can be expected to cover many topics, while corpora with more focused text data will cover a smaller number of topics. To further illustrate this, Hansen and McMahon (2016) use an LDA model with 15 topics to analyse the FOMC statements and Jegadeesh and Wu (2017) use 8 topics to analyse FOMC minutes. In comparison, Blei et al. (2003) use up to 100 topics to analyse abstracts from various scientific papers.

Next to this, the perplexity of the model can be used for a formal comparison between models and parameter choices. The perplexity of a model is defined as:

$$\text{perplexity}\left(\boldsymbol{n}, \alpha, \beta\right) \triangleq \exp\left\{-\left(\sum_i \log p\left(n_i \mid \alpha, \beta\right)\right) / \left(\sum_i n_i\right)\right\}, \tag{3}$$

where n_i is the vector of word counts for the ith sentence. However, there are questions regarding the usefulness of perplexity in deciding between models. Chang et al. (2009)

find that the perplexity of a model is inversely correlated to the interpretability of the results, because perplexity looks solely at the fit of the model. Thus, to determine k, the perplexity of the different models are compared with one another, but only if the results of the models are interpretable. Furthermore, the search window is narrowed to $k = 20$, since both Hansen and McMahon (2016) and Jegadeesh and Wu (2017) find optimal results below 20 topics.

The topics and their respective meanings are interpreted through three means: By analysing the words that have a high probability of belonging to a given topic, the distribution of topics in statements over time and the output from the LDAvis system, a web-based visualisation system designed by Sievert and Shirley (2014). The LDAvis system provides a two-dimensional plot that shows the inter-topic distances between all topics and the percentage of the corpus that is labelled as each topic. To compute the inter-topic distances, Jensen-Shannon divergence (JSD) is used to compute the similarity between the probability distributions of words for each topic, after which principal component analysis is used to reduce the dimensionality of the output of JSD (k by k) to k by 2. Topics with similar meanings are then placed close to each other and vice-versa.

4.2 Sentiment analysis

To perform sentiment analysis on all the sentences in a statement, the FinBERT model, introduced by Araci (2019), is used. To obtain the FinBERT model, Araci (2019) pretrains BERT further on a corpus which consists of a subset of Reuters' TRC2 corpus. The TRC2 corpus consists of all news articles that were published by Reuters between 2008 and 2010. However, due to limited availability of computer power, Araci (2019) filters out these news articles on a set of financial keywords. The resulting corpus contains 46163 documents, more than 29M words and almost 400K sentences. Furthermore, Araci (2019) fine-tunes the FinBERT model for sentiment classification tasks using the Financial PhraseBank dataset (created by Malo et al. (2014)). This dataset consists of 4845 randomly selected sentences from financial news articles, which are subsequently labelled by a group of 16 people with backgrounds in finance and business. Araci (2019) shows that the FinBERT model has a 15% higher accuracy than the standard BERT model on classification tasks for finance-related text. Since the FOMC statements are related to finance, the FinBERT model is the preferred choice for this use case. For a theoretical background of the BERT model (and therefore the FinBERT model), see Appendix A.2.

To apply the FinBERT model to the statements released by the FOMC, the model is further fine-tuned on a labelled dataset that consists of all unique sentences that are in the first released statement of each year. This is because the statements have changed significantly throughout the past 23 years, as seen in Section 3. Thus, taking the first released statement of each year ensures that a good representation of the general language used by the FOMC is present in the dataset. Each sentence is manually labelled as either positive, negative or neutral (see Table 4), which is subsequently fed into the BERT model to fine-tune it further.

Table 4: The size of the dataset and the distribution of labels within the dataset. The dataset consists of all sentences in the first released statement of the year in the period of 1999-2021.

Total	229
Positive	77
Negative	35
Neutral	117

To select the hyperparameters for fine-tuning, this paper follows Devlin et al. (2019). They mostly use the same parameters for fine-tuning as they do for pre-training, but with three exceptions: The batch size, learning rate and the number of epochs. When choosing the values for these three hyperparameters, one has to be careful to avoid catastrophic forgetting, which means that the pre-trained knowledge is erased during fine-tuning. Sun et al. (2019) find that especially the learning rate should be kept small to avoid this problem. Concurring, Devlin et al. (2019) find that the following values work well across all NLP tasks:

- **Batch size**: 16, 32
- **Learning rate**: 2e-5, 3e-5, 5e-5
- **Number of epochs**: 2, 3, 4

To find the optimal values for these hyperparameters, a grid search is performed across all the mentioned values for the learning rate and the number of epochs. Due to computational limitations, the batch size is kept at 16. Then, the optimal set of values is determined by comparing the fit of the in-sample framework (Section 5.1).

4.3 Similarity analysis

The cosine similarity measure is used to determine the similarity between two statements. This measure is defined as the cosine angle between two vectors, which is calculated by dividing the dot product between two vectors with the product of their respective lengths. To obtain these vectors, the embedding vector of each statement as a whole is determined using the embedding layer of BERT, as explained in Appendix A.2. The reason for using this embedding vector, instead of other measures such as word-counting vectors, is that the embedding vector contains contextual information about the content in the statement. This additional information yields more accurate results for similarity analysis, as discussed by Taeyoung et al. (2020). The definition for the cosine similarity measure is shown in Eq. (4), where S_i and S_j are the embedding vectors with dimension 768 of statement i and j respectively.

$$\text{Sim(statement i, statement j)} = \text{cosine}(S_i, S_j) = \frac{S_i.S_j}{\|S_i\|\|S_j\|} \tag{4}$$

5 Methodology - Predicting short-term returns

To analyse the predictive power of the contents in a FOMC statement, an in-sample analysis is first performed in Section 5.1 to determine which discussed topics have significant impact on the market. In Section 5.2, the fit of the in-sample model is improved by including reaction asymmetries. Then, to determine the predictive power of the contents, an out-of-sample analysis is performed in Section 5.3. Lastly, in Section 5.4, the findings of this research are applied to a macroeconomic random forest (MRF) model to investigate whether this can improve the predictive power.

5.1 In-sample framework

As a first step in answering the main research question of this paper, the news shocks of the sentiment measures are determined for each of the k topics, which are then used to explain the variation in short-term bond future returns after the release of a statement. The framework of Rosa (2011) is mainly followed to determine the news shocks. Firstly, to predict the sentiment of each topic i, the following AR(1) model is used:

$$\widehat{Index}^*_{i,t} = \gamma_{1i} Index^{OLD}_{i,t^-} + \gamma_{2i} PMI_{t^-} + \gamma_{3i} \pi^e_{t^-} + \gamma_{4i} Slope_{t^-} + \varepsilon_t, \tag{5}$$

where $\widehat{Index}^*_{i,t}$ stands for the predicted sentiment measure of topic i in the FOMC statement that is released at time t, $Index^{OLD}_{i,t^-}$ stands for the sentiment measure from the previously released FOMC statement, PMI_{t^-} stands for the PMI index at time t, $\pi^e_{t^-}$ stands for the inflation expectation at time t, $Slope_{t^-}$ stands for the slope of the federal funds futures at time t and ε_t is the error term of the AR(1) model. t^- indicates that the value is known to the market participants before the statement is released at time t. An AR(1) model is used because monetary policy tends to change slowly over the course of several months (Clarida et al. (2000) and Rudebusch (2002)). Thus, the sentiment too could show signs of persistence. Furthermore, the PMI and inflation expectation both give indications about future levels of economic activity and inflation, which could influence the sentiment of the FOMC. Lastly, the slope of the federal funds futures reveals what the market expects the change in FFR to be. This could indicate the FOMC's future monetary policy decisions and therefore also their sentiment regarding future conditions.

The predicted sentiment from Eq. (5) is used to determine the news shock (NS) by subtracting it from the actual sentiment, which is obtained through the method described in Section 4. Thus, the news shock is determined as follows:

$$NS_{i,t} \equiv Index_{i,t}^{NEW} - \widehat{Index}_{i,t}^{*}, \tag{6}$$

where $Index_{i,t}^{NEW}$ is the actual sentiment of topic i in the statement released at time t. Note that this paper uses the terms news shocks and surprises interchangeably.

The k news shocks are then used in an OLS model to explain the variation in the 40-minute returns after the statement has been released, as given by the following equation:

$$r_{t+40} = 100 \cdot \log\left(P_{[t+40 \cdot m]}/P_{[t-10 \cdot m]}\right) = \alpha + \beta_{i,NS} NS_{i,t} + \varepsilon_t. \tag{7}$$

By looking at the significance of $\beta_{i,NS}$, it can be inferred whether the market responds to a specific topic. To correct for small deviations in the release time, the price difference is taken between 40 minutes after the release and 10 minutes before the release. Furthermore, the 40-minute returns are used since Rosa (2011) shows that the market incorporates FOMC monetary surprises within 40 minutes of the announcement release. As such, this ensures that the whole price reaction after the release of a statement is captured. Moreover, the unannounced statements are disregarded throughout this analysis, since endogeneity issues can arise due to the unexpected nature of these unannounced meetings. This is in line with Rosa (2011).

In addition to the k different topics, the news shock for changes in the FFR are also determined, since Rosa (2011) shows that this news shock has significant predictive power. However, to determine the news shocks, they follow the equation from Kuttner (2001), which is given as:

$$MPS_t \equiv \Delta f_t \frac{D}{D-d}, \tag{8}$$

where Δf_t is the change in the one-month federal funds futures contract in a narrow window (t-10, t+20) around the release of the FOMC statement, d is the day of the month of the meeting and D is the number of days in the month. It is clear that this method suffers from the so-called "look-ahead bias", where information that is unknown at time t is used to predict future returns. Thus, this method can not be used for out-of-sample analysis and is therefore replaced by the following equation:

$$NS_{\Delta FFR,t} \equiv \Delta FFR_t - (FFR_{\text{three-month ahead}} - FFR_{\text{one-month ahead}}), \tag{9}$$

where $\Delta\ FFR$ is the change in FFR that is mentioned in the statement, $FFR_{\text{three-month ahead}}$ is the price of the three-month ahead FFR future and $FFR_{\text{one-month ahead}}$ is the price of the one-month ahead FFR future. This difference reflects the market's expectancy of the FFR change, as discussed in Section 3.2.2.

5.2 Reaction asymmetries

For each topic that has a significant impact on short-term returns, the fit of the regression is improved by including four possible reaction asymmetries (as discussed in Section 2.1.2) to Eq. (7), by including an interaction term between the reaction asymmetry and the news shock. In doing so, this paper follows Gardner et al. (2021), who test the effect of reaction asymmetries on regression coefficients using the following equation:

$$r_{t+40} = \alpha + \beta_{i,NS} NS_{i,t} + \beta_{i,j,X} NS_{i,t} X_{j,t} + \beta_{j,X} X_{j,t} + \epsilon_t, \tag{10}$$

where $X_{j,t}$ is an array that contains the variables that represent each reaction asymmetry j at time t. A statistically significant $\beta_{X,i,j}$ shows that the effect of the news shock for topic i on short-term returns changes depending on the value of the variables in $X_{j,t}$. To make the results easily interpretable, $X_{j,t}$ is standardised by subtracting the mean and dividing it with its respective standard deviation.

5.3 Out-of-sample framework

After establishing which topics significantly impact short-term returns and which reaction asymmetries significantly affect the reaction, the main research question is answered by means of an out-of-sample analysis. To do this, all significant news shocks from Eq. (7) and all significant reaction asymmetries from Eq. (10) are combined into one OLS regression model as follows:

$$\hat{r}_{t+40} = \alpha + \sum_{i=1}^{k} (\beta_{i,NS} NS_{i,t} + \sum_{j=1}^{r} \beta_{X,j} NS_{i,t} X_{j,t}) + \epsilon_t, \tag{11}$$

where k is the set of significant topics and r is the set of significant reaction asymmetries. Eq. (11) is used in combination with an expanding window to make predictions. Thus, all the available data up to time t is used to determine the coefficients in Eq. (11), which is subsequently used to predict the reaction at time t. This is done for all meetings M, after which the performance of the model is determined using the out-of-sample R^2, which is introduced by Campbell and Thompson (2008). The out-of-sample R^2 is defined as

$$R^2_{\text{oos}} = 1 - \frac{\sum_{m=1}^{M} \left(r_{m,t+40} - \hat{r}_{m,t+40}\right)^2}{\sum_{m=1}^{M} \left(r_{m,t+40} - \overline{r}_{m,t+40}\right)^2}, \tag{12}$$

where $r_{m,t+40}$ is the actual 40-minute return after the release of the statement for meeting m, $\hat{r}_{m,t+40}$ is the predicted return from Eq. (11) and $\overline{r}_{m,t+40}$ is the average over the 40-minute returns from all previous releases. A positive R^2_{oos} indicates that the model performs better than the benchmark, which in this case is assuming that the return equals the historic average, while a negative R^2_{oos} indicates that the model performs worse.

Then, it is tested whether the predicted values are significantly different from the benchmark. As such, this test is equivalent to testing whether the R^2_{oos} significantly differs from zero. The method described in Clark and West (2007) is used to perform this test. First, the test statistic in Eq. (13) is calculated, which is subsequently regressed on a constant. Then, the t-statistic of the constant in this regression reflects the significance level of the test statistic.

$$\left(r_{m,t+40} - \overline{r}_{m,t+40}\right)^2 - \left(r_{m,t+40} - \hat{r}_{m,t+40}\right)^2 + \left(\hat{r}_{m,t+40} - \overline{r}_{m,t+40}\right)^2. \tag{13}$$

5.4 Macroeconomic random forest model

Next to the OLS model in Eq. (10), this paper investigates whether the use of machine learning (ML) can aid the predictive power of the news shocks. To this end, the macroeconomic random forest (MRF) model, which is created by Goulet Coulombe (2020), is employed. The MRF model is effectively an OLS model with time varying coefficients that are determined through a random forest (RF) model. As such, the model is given as

$$\hat{r}_{t+40} = \alpha_t + \sum_{i=1}^{k} \beta_{i,NS,t} NS_{i,t} + \epsilon_t \text{ with } \alpha_t, \beta_{i,NS,t} = \mathcal{F}\left(S_t\right), \tag{14}$$

where $\mathcal{F}$ is an RF model with as input S_t, a set of variables. The OLS model in Eq. (14) determines the linear model that should be time-varying, and the RF is used to generate a set of Generalized Time-Varying Parameters (GTVPs) that are used as coefficients in the OLS model. S_t consists of the variables that are used for all mentioned reaction asymmetries in Section 2.1.2, the k news shocks and a vector that contains the numbers between 1 and the length of S_t, which allows the MRF model to recognise and correct for structural breaks. For a general MRF model with the form

$$\begin{aligned} y_t &= X_t \beta_t + \epsilon_t \\ \beta_t &= \mathcal{F}\left(S_t\right), \end{aligned} \tag{15}$$

the tree fitting procedure can be displayed as follows:

$$\begin{aligned} \min_{j \in \mathcal{J}^-, c \in \mathbb{R}} \Bigg[& \min_{\beta_1} \sum_{\{t \in l | S_{j,t} \leq c\}} \left(y_t - X_t \beta_1\right)^2 + \lambda \left\|\beta_1\right\|_2 \\ & + \min_{\beta_2} \sum_{\{t \in l | S_{j,t} > c\}} \left(y_t - X_t \beta_2\right)^2 + \lambda \left\|\beta_2\right\|_2 \Bigg]. \end{aligned} \tag{16}$$

Similar to a standard RF model, this procedure starts by selecting a random subset of variables from S_t, from which the optimal variable to split the sample with is determined. Then, the threshold c that optimally splits the sample into two children nodes is calculated. However, contrary to standard RF models, the loss function is given as the squared error of the OLS equation in Eq. (15), which is summed with a ridge regularisation for the GVTPs. The same optimisation procedure is performed for each children node until a stopping criterion is met. This procedure generates a single decision tree and is performed a fixed amount of times. Through this procedure, the MRF model can model reaction asymmetries, as the GVTPs are functions of the set of variables that are used in the reaction asymmetries.

The MRF model has a similar set of hyperparameters as a standard RF model, which can be fine-tuned using a search algorithm. However, as Goulet Coulombe (2020) notes, minuscule performance gains after fine-tuning are the norm rather than the exception. Thus, to decrease the computations necessary, this paper primarily uses the default hyperparameters that are determined in Goulet Coulombe (2020). However, there are two changes: Firstly, Goulet Coulombe (2020) advises that the minimal node size is set to 10 for quarterly data and 15 for monthly data. Since there are eight released statements each year, the minimal node size is set to 13. Secondly, Goulet Coulombe (2020) finds that 200-300 trees are often needed to obtain credible regions for the parameters. Thus, 300 decision trees are used in the forest.

To measure the performance of the MRF model and compare it with the OLS model in Eq. (11), the same out-of-sample framework of Section 5.3 is used.

6 Results - Textual analysis

This section discusses the results of applying textual analysis to the FOMC statements. Firstly, the topic analysis is discussed in Section 6.1, after which the similarity analysis is discussed in Section 6.2.

6.1 Topic analysis

6.1.1 Parameter estimation

As a first step in dissecting the FOMC statements into topics, the number of topics k that the LDA model should extract from each statement is determined. As discussed in Section 4.1, both the perplexity and interpretability of the topics are taken into account when determining this. In Figure 5, the perplexity for each of the k topics is plotted. This figure shows that the perplexity drops quickly and levels off when k is set to five. Therefore, k is not set to a value smaller than five. In terms of interpretability, testing shows that large values for k can provide topics that are difficult to interpret. Or, there are multiple topics that are very similar to each other, which can negatively impact the interpretability of these topics. To avoid these issues, a lower value for k is preferred. When $k = 5$, all topics are clearly interpretable and distinguishable. Therefore, it is chosen to have $k = 5$ topics, of which its interpretation is discussed in the sections below. Furthermore, both the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are set to $\frac{1}{5}$, as discussed in Section 4.1.

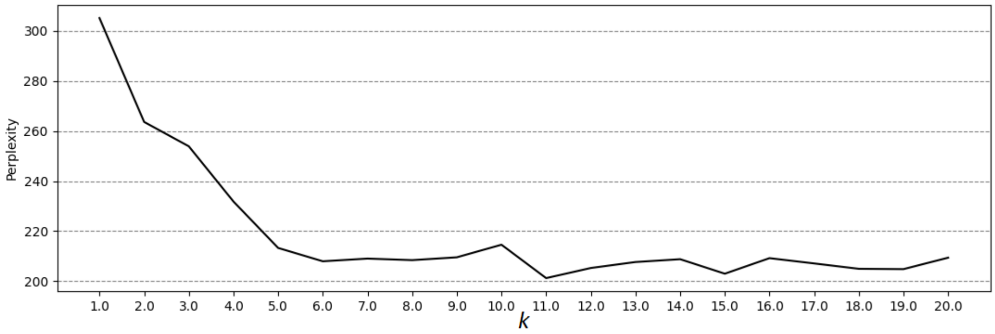


Figure 5: The perplexity for a varying number of topics k.

6.1.2 Interpretation of the topics

With these parameters, the results of the LDA model are shown in Table 5. The table shows the top ten stemmed words for each topic that have the highest probability of belonging to that specific topic. To help with the economic interpretation of the results in Section 7, each topic is interpreted using this set of ten words. To this end, an example of a sentence that belongs to each topic is shown in Table 6.

Table 5: The ten words that have the highest probability of corresponding to each topic.

	Topic 1:	Topic 2:	Topic 3:	Topic 4:	Topic 5:
	Changes in FFR	Monetary policy	Market conditions	Inflation	FOMC's expectation
1	feder	secur	market	inflat	committe
2	rate	agenc	econom	expect	econom
3	fund	back	activ	longer	employ
4	percent	mortgag	labor	committe	stabil
5	target	purchas	condit	remain	maximum
6	committe	hold	growth	term	price
7	rang	committe	remain	run	outlook
8	decid	treasuri	continu	percent	polici
9	market	billion	busi	market	risk
10	open	maintain	committe	pressur	continu

The top ten words that correspond to each topic. The sample consists of all statements released between the 15th of May 1999 and the 4th of May 2022.

Table 6: Examples of sentences that correspond to a specific topic.

Topic	Sentence
1	The Federal Open Market Committee decided today to lower its target for the federal funds rate by 25 basis points to 1 percent.
2	In addition, the Federal Reserve will buy up to $ 300 billion of Treasury securities by autumn.
3	Labor market conditions, however, apparently continue to improve gradually.
4	Inflation recently picked up somewhat, reflecting higher energy prices.
5	.... the Committee believes the risks continue to be weighted mainly toward conditions that may generate heightened inflation pressures in the future.

The table shows examples of sentences that correspond to a specific topic and the date of the released statement from which these sentences are taken. All sentences have a probability of more than 90% of belonging to this topic.

For topic 1, the top three words are regarding the FFR, while the latter seven are regarding decisions made by the committee. Thus, topic 1 most likely covers the single sentence within each statement that states whether changes in FFR are made. Topic 2 talks about securities, purchases and treasuries. It also again contains the word committee, which indicates that topic 2 is regarding any further changes in monetary policy, such as quantitative easing. Topic 3 contains words that are about labour/economic market conditions, which indicates that topic 3 is regarding the current state of these markets. Topic 4 has the main word "inflat", which shows that topic 4 is mainly about inflation, for both the current conditions and the future expected path. Topic 5 has committee as its highest probability word, which is paired with words such as economic, employment and inflation. As such, topic 5 is most likely regarding the committee's expectations of future conditions.

6.1.3 LDAvis

As mentioned in Section 4.1, the LDAvis system is used to visually analyse the similarities between topics, of which the results are displayed in Figure 6.

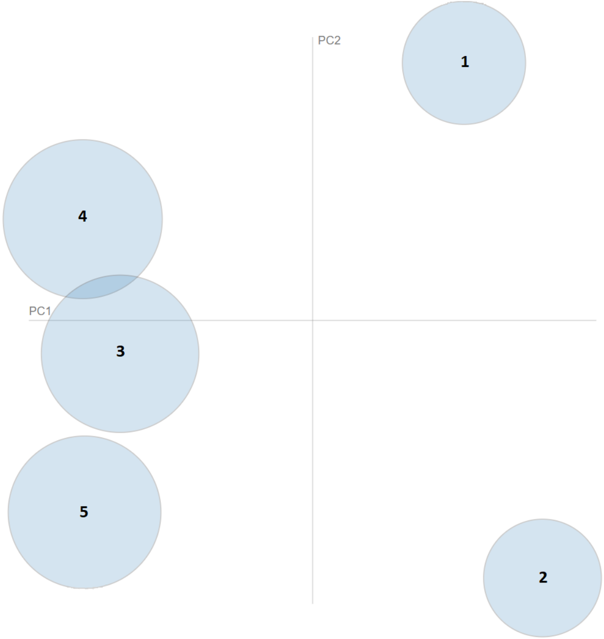


Figure 6: The inter-topic distances for the five topics. Topic 1 is about changes in FFR, topic 2 is about additional monetary policy measures, topic 3 is regarding market/economic conditions, topic 4 is about inflation and topic 5 is regarding the FOMC's expectation on future conditions.

The figure shows that topics 3 through 5 all span an equal part of the corpus, while topics 1 and 2 span a smaller section. Furthermore, it shows that topics 3, 4 and 5 are similar to each other and that topics 1 and 2 are largely isolated from the rest. Intuitively, these results are plausible, since topics 3, 4 and 5 all contain components regarding both current and expected economic conditions. This further confirms the earlier mentioned interpretations of the five topics.

6.1.4 distribution of topics over time

Lastly, to also support the earlier mentioned interpretation for each topic, the distribution of discussed topics over time is plotted in Figure 7. This plot shows that the distribution for topic 1 is inversely related to the number of sentences per statement, as given in Figure 2(a). This further supports the hypothesis that topic 1 covers the one sentence within a statement that mentions the possible change in FFR, since this topic will cover a large proportion of a short statement and vice-versa. Furthermore, it shows that topic 2 is mainly present during times of crisis (both the financial crisis and COVID-19), which further indicates that topic 2 is regarding extra monetary measures to support the economy. In addition to this, Figure 7 also shows that the share of topic 2 within a statement shrank between the financial crisis and the COVID-19 crisis in 2020. An interpretation for this is that after the financial crisis, the need for additional measures slowly dropped as the economy expanded. Especially at the end of 2014, a significant drop is present, which coincides with the ending of the third quantitative easing program in October 2014. The distribution of the other three topics is relatively consistent throughout time. A further interesting note to Figure 7 is that the distribution of topics over time is more stable after the financial crisis in 2009.

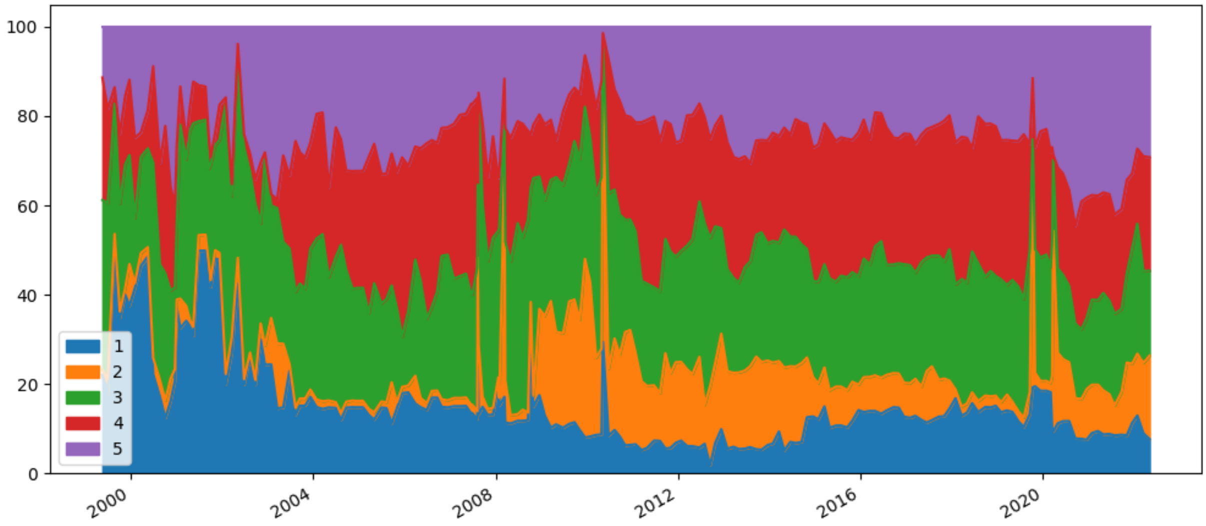


Figure 7: The distribution of topics over the time period between the 15th of May, 1999 and the 4th of May, 2022.

6.2 Similarity analysis

The cosine similarity between two subsequent statements is shown in Figure 8. This is plotted for the whole set of FOMC statements (marked by the dashed line) and the set of FOMC statements where unannounced meetings are excluded (marked by the solid line). The figure shows that FOMC statements are typically speaking very similar to each other, especially after the financial crisis in 2009. Furthermore, it shows that unannounced statements are often very different from announced statements. A comparison between this figure and Figure 7 shows that there is an overlap in the similarity of unannounced statements and the spikes of topic 2. This is due to the fact that unannounced meetings are often held in times of economic distress, during which new monetary policy measures are discussed to boost the economy.

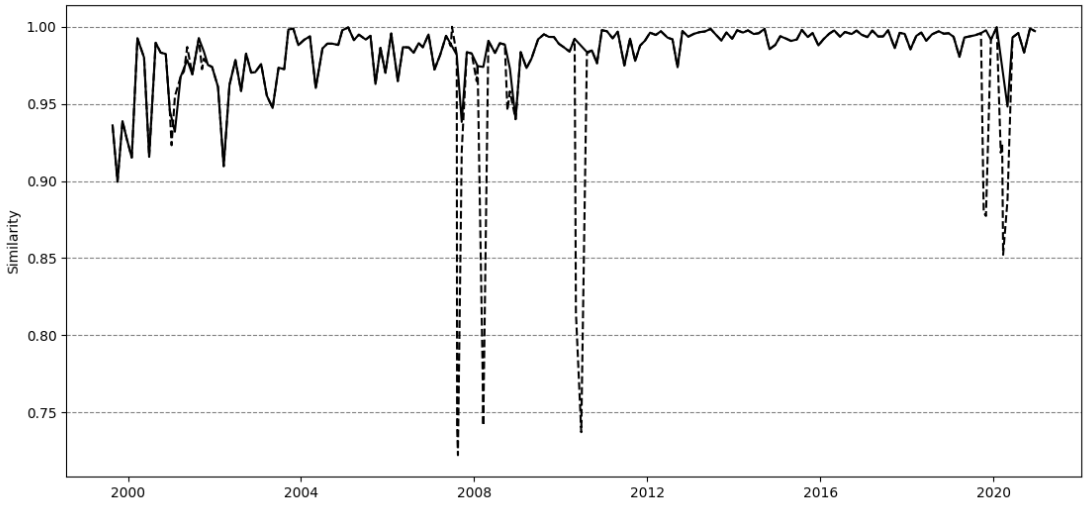


Figure 8: The cosine similarity between all FOMC statements that have been released between the 15th of May, 1999 and the 16th of December, 2020. The solid line represents the cosine similarity for the set of FOMC statements that excludes the unannounced statements, while the dashed line represents the whole sample.

In Figure 9, the similarity between a statement and all other statements is plotted in the form of a heat map. This is plotted for the set of FOMC statements that also includes unannounced meetings. This heat map shows that the unannounced statements differ from all other announced and unannounced statements. An interesting finding in this heat map is that the announced statements after 2009 are very similar to each other, and not just similar to the statement before, which is shown in Figure 7. This indicates that the FOMC has generalised their communication through their released statements, even more so than before the financial crisis. Concurringly, this finding coincides with the finding in Section 6.1 that states that the distribution of topics over time is more stable after the financial crisis in 2009.

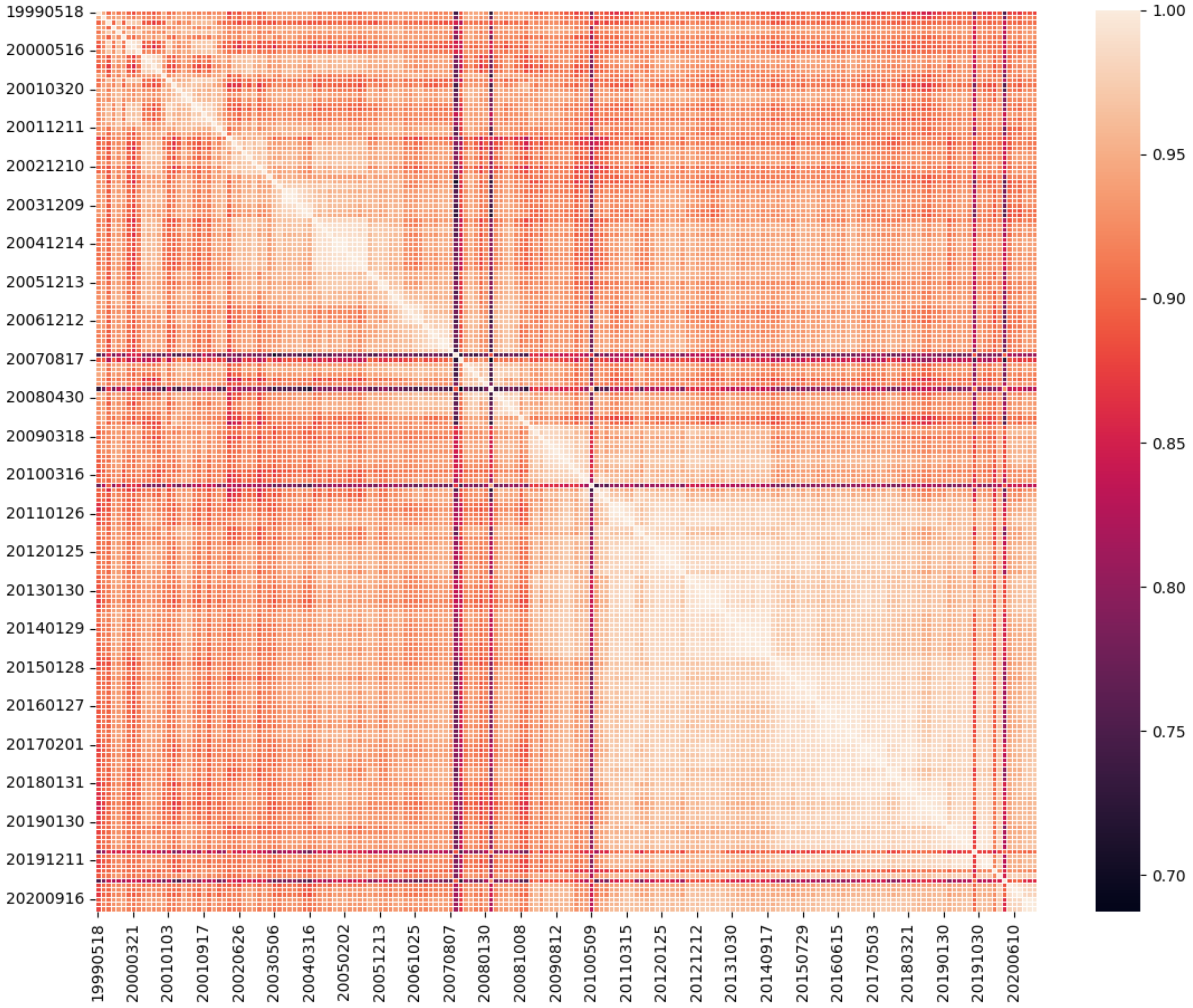


Figure 9: The cosine similarity between a statement and all other statements that have been released between the 15th of May, 1999 and the 16th of December, 2020.

7 Results - Predicting short-term returns

This section discusses the results regarding the predictive power of the FOMC statements are discussed. First, in Section 7.1, the results from the in-sample framework are discussed. This entails determining the news shocks and measuring to what extent the news shocks can explain the variation in returns and possible reaction asymmetries that can affect these results. Then, in Section 7.4, the out-of-sample predictive power of the FOMC statements is discussed using an OLS and MRF model. Lastly, the two robustness checks are discussed in Section 7.5.

7.1 In-sample framework

7.1.1 Determining the market's expectancy

The market's expectancy of the FOMC's sentiment regarding a topic is determined using the OLS model as given in Eq. (6). The results of this regression are given in Table 7, where T_i stands for the sentiment of each of the five topics. To capture this sentiment, an arbitrary number of 2 epochs and a learning rate of 2e-5 are used for the FinBERT model. These hyperparameters are further optimised in Section 7.1.3.

Table 7: Summary of the regressions that determine the market's expectancy.

	intercept	y_{t-1}	*PMI*	*FFR Slope*	*Inflation Exp.*	R^2_{adj}
T_1	-0.028	0.586***	0.000	0.197***	-0.014	0.449
T_2	-0.097	0.154**	-0.005*	0.037	0.035	0.036
T_3	0.322***	0.371***	0.024***	0.157	-0.080***	0.366
T_4	-0.126*	0.371**	-0.003	-0.076*	0.037*	0.148
T_5	-0.067	0.405***	0.007**	0.116*	-0.004	0.357

This table summarises the regressions that determine the market's expectancy on each topic that is discussed in a FOMC statement. The sample period of this regression runs between the 15th of May 1999 and the 16th of December 2020 and contains 172 observations. The 10%, 5% and 1% significance levels are denoted by *, ** and ***, respectively. Furthermore, throughout these regressions, all unannounced FOMC statements are excluded.

As hypothesized in Section 5.1, all sentiments follow an AR(1) process, since their first lag is highly significant. In addition, y_{t-1} explains a large portion of the variation in the sentiment for each topic, since the R^2_{adj} is highly correlated with y_{t-1}. A larger value for y_{t-1} indicates that the sentiment is more persistent, which causes the R^2_{adj} to increase.

Furthermore, each sentiment has at least one macroeconomic variable that is significant. *PMI* is significant for T_2, T_3 and T_5, where topics T_3 and T_5 both discuss economic conditions and how it could develop in the future, and T_2 discusses any changes in monetary policy. Whether any changes in monetary policy are made is dependent on the current state of the economy, since poor economic conditions might trigger new monetary policy measures, as discussed in Section 6.1.4. The *FFR Slope* variable is significant for T_1, T_4 and T_5, where T_1 discusses any changes in FFR, T_4 talks about inflation and its expected course and T_5 discusses the committee's expectations on future conditions. Lastly, the *Inflation Expectation* is significant for T_3 and T_4, two topics that discuss inflation and economic conditions. Thus, the significant variables are in line with the interpretation of the topics in Section 6.1.

7.1.2 Measuring the effect of news shocks on short-term returns

To measure the effect of news shocks on the 40-minute returns of TU, the model as given in Eq. (7) is used, from which the results are given in Table 8.

Table 8: Summary of the in-sample regressions on 40-minute returns.

intercept	NS_1	NS_2	NS_3	NS_4	NS_5	$NS_{\Delta FFR}$	R^2_{adj}
0.013*	0.103						0.013
0.013*		0.041					-0.003
0.013*			-0.047*				0.011
0.013*				-0.006			-0.006
0.013*					0.106**		0.010
0.012*						-0.493***	0.108
0.012*			-0.036		0.110***	-0.476***	**0.123**
0.012*					0.106***	-0.493***	0.119

This table summarises the in-sample regressions of Eq. (7) that determine what effect the news shocks for each discussed topic and the news shocks for ΔFFR have on the 40-minute returns of TU. The sample period runs between the 15th of May 1999 and the 16th of December 2020 and contains 172 observations. The 10%, 5% and 1% significance levels are denoted by *, ** and ***, respectively. The value in bold denotes the highest R^2_{adj}. Furthermore, throughout these regressions, all unannounced FOMC statements are excluded.

This table shows that the news shock for topic 3 (NS_3), topic 5 (NS_5) and the news shock for ΔFFR ($NS_{\Delta FFR}$) have a significant effect on 40 minute returns. A positive news shock for topic 5 has a positive effect on the price of TU, whereas a positive news shock for ΔFFR and NS_3 have a negative impact on the price of TU. When all non-significant variables are excluded from the regression, the model with the best fit is obtained. However, this causes NS_3 to lose its significance, which signals that this news shock does not add any additional information next to NS_5 and $NS_{\Delta FFR}$. For this reason, NS_3 is excluded from the regression, which results in the model with the second-best fit. From here on forth, this model will be referred to as the "standard OLS model".

A positive news shock for topic 5 indicates that the FOMC is more positive about future conditions than expected. This can lead investors to change their beliefs about future conditions, which drives the price upwards. Next to this, an unexpected increase in FFR indicates that the interest rates for 2Y bonds will rise too, which drives the price of TU down.

To further interpret these results, the five topics are divided into two sets: A backwards-looking set and a forward-looking one. The backwards-looking set consists of topics that discuss current economic conditions, or how conditions have evolved since the last meeting. The forward-looking set contains topics that discuss how the FOMC expects future conditions to evolve or topics that discuss changes in monetary policy. Using the topic interpretations from Section 6.1, the five topics are divided as follows: $Backwards : \{T_3, T_4\}$ and $Forward : \{T_1, T_2, T_5\}$. Table 8 shows that none of the backwards-looking topics are significant, which is expected, since those topics discuss information that is already known to the market participants.

Interestingly, topic 1 and topic 2 are not significant, even though they are part of the forward-looking set. This can be caused by the fact that both topics discuss changes in monetary policy, which makes the extraction of the true sentiment more difficult. For example, topic 1 mainly contains a single sentence in a statement that gives the ΔFFR. As such, the FinBERT model should be able to tell whether a rise in interest rates is positive or negative. If this information is not sufficiently present in the fine-tuning dataset, the model might not be able to fully extract the true sentiment of that specific sentence. Similarly for topic 2, the model should know whether the additional monetary policy measures are positive or negative for the economy, which can be challenging to achieve due to the complex nature of the sentences that belong to this topic.

7.1.3 Hyperparameters selection

The FinBERT model has two hyperparameters that need to be optimised, namely the learning rate and the number of epochs, as discussed in Section 4.2. These hyperparameters are optimised by using a grid search across all possible sets of values for the hyperparameters. Specifically, the sentiment of topic 5 is determined for each set of values, since this is the only topic that has a significant impact on returns, as shown in Table 8. This sentiment is used in Eq. (7) to determine the fit of the model, which is then compared across all possible sets of values. The results of this process are given in Table 9, in which the R^2_{adj} and the significance level of NS_5 are given for each set of values. They are also given for the FinBERT model that is not further fine-tuned, as indicated by the value for a learning rate of zero and zero number of epochs.

Table 9: The R^2_{adj} and the significance of NS_5 for various values of the two hyperparameters of the FinBERT model.

	Learning rate				
Epochs	0	2e-5	3e-5	4e-5	5e-5
0	-0.004	-	-	-	-
2	-	0.010**	0.009	0.013	0.019
3	-	0.000	0.021*	0.022	0.011
4	-	0.005	0.008	0.016	-0.001

This table displays the R^2_{adj} for the regression model given in Eq. (7) for NS_5. The sample period runs between the 15th of May 1999 and the 16th of December 2020 and contains 172 observations. The 10%, 5% and 1% significance levels of NS_5 are denoted by *, ** and ***, respectively. Furthermore, throughout these regressions, all unannounced FOMC statements are excluded.

The table shows that NS_5 is only significant for two sets of hyperparameters, which shows that NS_5 is highly sensitive to the paramaters of the FinBERT model. To choose the optimal set of hyperparameters, a higher significance level is preferred over a higher R^2_{adj}. Therefore, two epochs and a learning rate of 2e-5 are used to obtain the sentiment from the statements for the rest of this paper. Furthermore, this table provides evidence that fine-tuning is necessary for training the FinBERT model, since the non-fine-tuned model has the weakest fit.

7.2 Reaction asymmetries

As NS_5 and $NS_{\Delta FFR}$ are the only two significant variables for short-term returns, it is determined whether reaction asymmetries significantly affect these two variables. The results from Eq. (10) are given in Table 10. First, in regression (1), it is determined whether a negative surprise in ΔFFR causes a larger reaction than a positive surprise, by multiplying NS with the two indicator variables I_1 and I_2. I_1 is set to one if the surprise has a positive sign and I_2 is set to one if the surprise has a negative sign.

Table 10: The effect of four reaction asymmetries on the two news shocks NS_5 and $NS_{\Delta FFR}$.

	NS_5				$NS_{\Delta FFR}$			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
$Intercept$	0.058	0.014*	0.013**	0.013*	0.053	0.013**	0.011**	0.011**
NS	-0.092	0.114*	0.088***	0.098	-0.350**	-0.489***	-0.377**	-0.426***
$NS * I_1$	0.189				-0.264			
$NS * I_2$	0.169				-0.086			
$NS * cos$		0.015				0.023		
$NS * PMI$			-0.026				0.073**	
$NS * PD$				0.002				0.077
I_1	-0.068*				-0.039			
I_2	-0.019				-0.039			
cos		0.003				0.005		
PMI			-0.022***				-0.012*	
PD				-0.038***				-0.035***
R^2_{adj}	0.034	0.012	0.035	0.104	0.100	0.117	0.121	0.192

This table summarises the in-sample regressions of Eq. (10) that determine which reaction asymmetries affect the reaction of the market to NS_5 and $NS_{\Delta FFR}$. The sample period runs between the 15th of May 1999 and the 16th of December 2020 and contains 172 observations. The 10%, 5% and 1% significance levels are denoted by *, ** and ***, respectively. Furthermore, throughout these regressions, all unannounced FOMC statements are excluded.

Table 10 shows that this reaction asymmetry does not hold for both NS_5 and $NS_{\Delta FFR}$, as $NS * I_1$ has the larger coefficient and none of the interaction terms are significant. For $NS_{\Delta FFR}$, this contradicts Chuliá et al. (2010) and Farka (2009). However, this is caused by the fact that they use Kuttner (2001)'s measure to determine $NS_{\Delta FFR}$, whereas this paper uses Eq. (9), as discussed in Section 5.1. To show this, Eq. (9) is replaced with Kuttner (2001)'s measure in Appendix B.1. The results in Appendix B.1 show that the reaction asymmetry does hold true when Kuttner (2001)'s measure is used.

Regression (2) shows that the second reaction asymmetry, namely whether changes in the statement cause a larger reaction, also does not hold for both news shocks, as the interaction terms are not significant.

Regression (3) shows that the current economic conditions, as proxied by the *PMI* index, only affect the reaction of the market to $NS_{\Delta FFR}$. The results show that poor (good) economic conditions cause a larger (smaller) reaction, as poor economic conditions cause the sign for $NS_{\Delta FFR} * PMI$ to be negative, which is the same as the sign for $NS_{\Delta FFR}$. Furthermore, current economic conditions also directly impact the short-term returns, since the coefficients for *PMI* are negative and significant for both news shocks. As the signs are negative, the short-term returns tend to be higher when economic conditions are poor and vice-versa.

Lastly, regression (4) shows that the size of the pre-drift (PD) does not significantly impact the market's reaction to the news shocks, but that the pre-drift does have a highly significant impact on the short-term returns. The negative sign of the coefficient indicates that there is a mean-reverting process, where a large pre-drift has a negative impact on short-term returns and vice-versa. An economic interpretation for this finding is that investors may tend to overestimate the uncertainty and risk a statement brings, or that they overreact to the informal communication of FOMC members, as discussed in Section 2.1.2.

7.3 Final OLS model

To finalise the in-sample regression model, all significant variables of Table 10 are combined into one regression model, as given in Eq. (11). This results in regression (1) in Table 11. The table shows that all variables, except for $NS_{\Delta FFR} * PMI$, are still significant and that the coefficients hold the same signs. When $NS_{\Delta FFR} * PMI$ is omitted from the regression, the model with the best fit as given in regression (2) is obtained. From here on forth, regression (2) will be referred to as the "full OLS model". Interestingly, none of the reaction asymmetries end up being significant for this model. However, adding PMI and PD do cause a 9.3 percent point increase in R^2_{adj}.

Table 11: The full OLS model that incorporates all significant reaction asymmetries.

	(1)	(2)
intercept	0.011**	0.012***
NS_5	0.097**	0.099**
$NS_{\Delta FFR}$	-0.374**	-0.423***
$NS_{\Delta FFR} * PMI$	0.046	
PMI	-0.013**	-0.014***
PD	-0.034***	-0.035***
R^2_{adj}	0.211	**0.212**

This table summarises the in-sample regression of Eq. (11), which includes all the significant reaction asymmetries as shown in Table 10. The sample period runs between the 15th of May 1999 and the 16th of December 2020 and contains 172 observations. The 10%, 5% and 1% significance levels are denoted by *, ** and ***, respectively. The value in bold denotes the highest R^2_{adj}. Furthermore, throughout these regressions, all unannounced FOMC statements are excluded.

7.3.1 Two regimes

Throughout this paper, evidence is found that indicates that there has been a change in communication from the FOMC to the market participants through their released statements. Firstly, as discussed in Section 3.1 and Section 6.2, this paper finds that the statements have become more complex and similar to each other after the financial crisis. Next to this, the responses of the market to the release of the statements have also decreased, as discussed in Section 3.2.1. Therefore, to investigate whether this shift also affects the predictive power of NS_5 and $NS_{\Delta FFR}$, the following regression is run:

$$r_{t+40} = \alpha + \beta_1 NS_{5,t} I_1 + \beta_2 NS_{\Delta FFR,t} I_1 + \beta_3 NS_{5,t} I_2 + \beta_4 NS_{\Delta FFR,t} I_2 + \varepsilon_t, \qquad (17)$$

where I_1 is an indicator variable that is set to one if $t < 2009$ and zero otherwise and I_2 is an indicator variable that is set to one if $t > 2009$ and zero otherwise. The results of this regression are given in Table 12 and show that both NS_5 and $NS_{\Delta FFR}$ are highly significant for the time period of 1999 till 2009, but lose their significance in the time period of 2009 till 2021. Thus, both news shocks lose their predictive power for short-term returns after the financial crisis in 2009.

Table 12: The time-varying effect of news shocks on 40-minute returns.

$NS_5 * I_1$	$NS_{\Delta FFR} * I_1$	$NS_5 * I_2$	$NS_{\Delta FFR} * I_2$	R^2_{adj}
0.110***	-0.608***	0.072	-0.068	0.131

This table summarises the in-sample regression of Eq. (17). The sample period runs between the 15th of May 1999 and the 16th of December 2020 and contains 172 observations. The 10%, 5% and 1% significance levels are denoted by *, ** and ***, respectively. Furthermore, all unannounced FOMC statements are excluded.

7.3.2 Varying time windows

Up till now, all results are obtained using the 40-minute returns of TU, as Rosa (2011) finds that it takes 40 minutes for the market to incorporate new information from a statement into the price. However, different time windows could impact the fit of the in-sample models. Therefore, following Taeyoung et al. (2020), the standard and the full OLS model are run for seven additional time windows that vary between 10 minutes and 120 minutes. In Table 13, the columns corresponding to regression (1) show the results for the standard OLS model, and regression (2) gives the R^2_{adj} for the full OLS model. For the full OLS model, the coefficients and their respective significance levels have been omitted for the sake of redundancy. The table shows that $NS_{\Delta FFR}$ is significant for all time windows, that NS_5 is significant for time windows larger than 20 minutes and that the fit for the standard OLS model is the highest at a time window of 30 minutes, whereas the best fit for the full OLS model is achieved with a time window of 20 minutes. In addition, the table also shows that the largest reaction is reached at 40 minutes after the release of the statement, as both NS_5 and $NS_{\Delta FFR}$ have their highest value for this time window. At a time window of 50 minutes, a reversal seems to occur, after which the price stabilises. Similar to Rosa (2011), this indicates that the market takes roughly 40 minutes to incorporate new information regarding monetary policy.

Table 13: The effect of varying time windows on the fit of the standard and full OLS models.

		(1)		(2)
Time window	NS_5	$NS_{\Delta FFR}$	R^2_{adj}	R^2_{adj}
10	0.045	-0.381***	0.104	0.199
20	0.041	-0.427***	0.120	**0.237**
30	0.066***	-0.464***	**0.126**	0.233
40	0.106***	-0.493***	0.119	0.212
50	0.084**	-0.458***	0.115	0.180
60	0.090**	-0.455***	0.113	0.173
90	0.095**	-0.455**	0.111	0.164
120	0.106**	-0.484**	0.121	0.165

This table summarises the in-sample regressions of Eq. (7) (regression (1)) and Eq. (11) (regression (2)) for eight varying time windows. Regression (1) shows the coefficients and their significance for all variables, while regression (2) only shows the R^2_{adj}. To obtain NS_5, two epochs a learning rate of 2e-5 are used. The sample period runs between the 15th of May 1999 and the 16th of December 2020 and contains 172 observations. The 10%, 5% and 1% significance levels are denoted by *, ** and ***, respectively. The value in bold denotes the highest R^2_{adj}. Furthermore, throughout these regressions, all unannounced FOMC statements are excluded.

7.4 Out-of-sample framework

This section discusses the out-of-sample results for the standard OLS model, the full OLS model and the MRF model. For all three models, an expanding window is used to predict the short-term returns after the release of a statement. Similar to Section 7.3.2, the out-of-sample performance is determined for eight time windows that span 10 and 120 minutes. The out-of-sample prediction starts with the first released statement of 2004 on the 28th of January 2004, and continues till the last statement of 2020 on the 16th of December 2020.

Table 14: R^2_{oos} for the standard OLS, full OLS and MRF model for various time windows.

	10	20	30	40	50	60	90	120
Standard OLS	0.085*	0.122	**0.131**	0.130	0.110	0.107	0.106	0.120
Full OLS	0.118**	**0.174****	0.172*	0.148*	0.141*	0.130	0.126	0.122
MRF	0.037	0.103*	0.127	**0.132***	0.088	0.077	0.073	0.065

This table shows the R^2_{oos} for the standard OLS, full OLS and the MRF model. These values are obtained using an expanding window, where the sample period runs from the 28th of January 2004 till the 16th of December 2020. The sample contains 134 predictions. The 10%, 5% and 1% significance levels are denoted by *, ** and ***, respectively. The value in bold denotes the highest R^2_{oos} for each model. Furthermore, throughout these regressions, all unannounced FOMC statements are excluded.

Table 14 displays the R^2_{oos} and its significance level for each model and each of the eight time windows. The table shows that for each time window all three models outperform the benchmark and that for the standard and full OLS models the best performance is achieved with a time window of 30 and 20 minutes respectively, whereas the MRF model has the highest R^2_{oos} for a time window of 40 minutes.

A noticeable finding is that the full OLS model consistently outperforms the standard OLS model across all time windows. This indicates that adding *PMI* and *PD*, as discussed in Section 7.2 and Section 7.3, also increases the predictive power in an out-of-sample framework. Next to this, Table 14 shows that both OLS models consistently outperform the MRF model. This is met with one exception, where the MRF outperforms the standard OLS model for a time window of 40 minutes. A potential reason for this underperformance is that none of the in-sample reaction asymmetries in Section 7.2 are significant. This, in combination with the expanding window, could cause the MRF model to have a tendency to overfit; Due to the expanding window, the MRF model does not have the whole sample size to determine its trees, which may cause the model to find spurious reaction asymmetries that are not significant over the whole sample.

Furthermore, the table shows that only a few of the R^2_{oos} are significant. This is because there is a single significant outlier in the set of predictions on the 16th of December 2008, when the FOMC unexpectedly cut the FFR by 75 basis points. This single outlier can skew the results of the test statistic in Eq. (13), since this effectively tests the consistency of the out-of-sample performance. In Appendix B.2, this outlier is removed from the dataset to test this finding. If this outlier is removed, most values of R^2_{oos} have an increased significance and are still positive, albeit a lot smaller than displayed in Table 14. This indicates that most of the gain in R^2_{oos} comes from this single correct prediction.

In Figure 10, the evolution of the R^2_{oos} for a time window of 40 minutes is plotted over time. The figure shows that most of the gain in R^2_{oos} is achieved in the time period of 2006 till 2009, after which the R^2_{oos} drops to its displayed level in Table 14. This reflects and further strengthens the finding of Section 7.3.2, which states that after 2009 both NS_5 and $NS_{\triangle FFR}$ lose their significant predictive power for short-term returns.

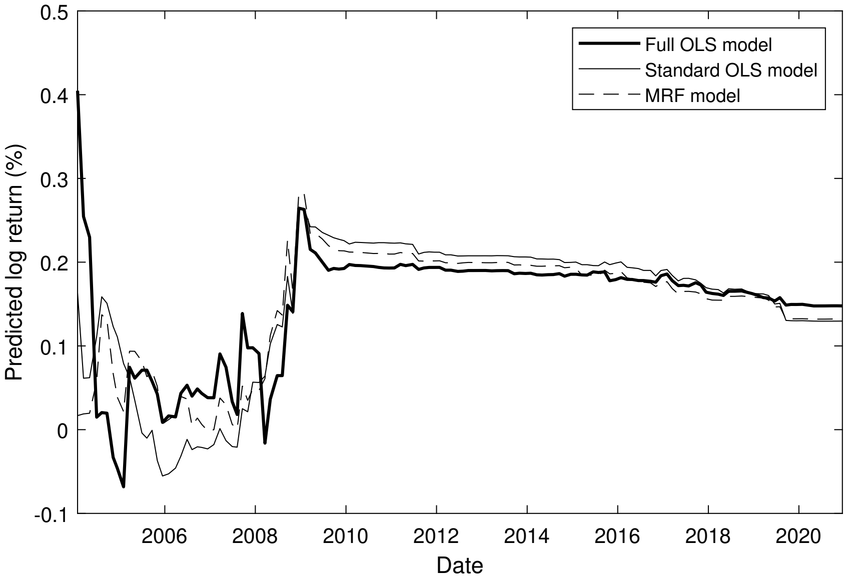


Figure 10: The R^2_{oos} over time for the full OLS model (thick solid line), standard OLS model (thin solid line) and the MRF model (dashed line). The time period runs from the 28th of January, 2004 till the 16th of December, 2020.

7.5 Robustness checks

7.5.1 Including unannounced meetings

In all previous results, unannounced statements were excluded from the dataset. These unannounced statements can cause issues with endogeneity, since it is not known whether the market reacts to the content of these statements, or purely to the fact that a statement is unannounced. However, these unannounced meetings can still have an impact on short-term returns. Thus, for the first robustness check, these unannounced meetings are included in the dataset and the in-sample results for both the standard and the full OLS model are replicated. Table 15 gives the results of using the complete set of statements on both the standard OLS model (1) and the full OLS model (2).

Table 15: Summary of the regressions of the standard and full OLS model when the unannounced statements are included in the dataset.

	(1)	(2)
intercept	0.009	0.009*
NS_5	0.086	0.092*
$NS_{\Delta FFR}$	-0.329***	-0.262***
PMI		-0.011***
PD		-0.036***
R^2_{adj}	0.120	0.208

This table summarises the in-sample regressions for both the standard OLS model (1) and the full OLS model (2) when the unannounced statements are included in the dataset. The sample period runs between the 15th of May 1999 and the 16th of December 2020 and contains 187 observations. The 10%, 5% and 1% significance levels are denoted by *, ** and ***, respectively.

The table shows that NS_5 loses its significance, or has a reduced significance level, when the unannounced statements are included in the dataset. This is due to the fact that most of these unannounced statements are very different in content from the announced statements, as discussed in Section 6.2. These unannounced statements mostly contain sentences that are regarding topic 2, which is about other monetary policy measures. This can be inferred by looking at the similarity between Figure 7 and Figure 8, which shows that the main spikes of the unannounced statements in Figure 8 are also present in Figure 7. These spikes in Figure 7 show that they mostly contain sentences from topic

2. As Table 8 shows, news shocks in this topic do not contain any predictive power. It is therefore expected that the content of the unannounced statements also does not contain predictive power (except for any changes in FFR), which can cause NS_5 to lose its significance.

7.5.2 Excluding possible bias

To gather the training data for the FinBERT model, all of the unique sentences from the first announced statement of each year are taken. However, this could skew the sentiment score for these statements since I effectively "tell" the model whether those statements are positive or negative. Since this sentiment is used in an AR(1) model, this could skew the results for the first two statements of the year. Thus, to test whether any bias is introduced into the FinBERT model, the first two released statements of the year are excluded from the dataset and the in-sample results are replicated for both the standard (1) and the full OLS model (2), which results in Table 16.

Table 16: Summary of the regressions of the standard and full OLS model when the first two statements of the year are excluded from the dataset.

	(1)	(2)
intercept	0.009	0.011^{*}
NS_5	0.082^{**}	0.059^{**}
$NS_{\Delta FFR}$	-0.502^{***}	-0.449^{***}
PMI		-0.006
PD		-0.041^{***}
R^2_{adj}	0.161	0.284

This table summarises the in-sample regressions for both the standard OLS model (1) and the full OLS model (2) when the first two statements of the year are excluded from the dataset. The sample period runs between the 15th of May 1999 and the 16th of December 2020 and contains 128 observations. The 10%, 5% and 1% significance levels are denoted by *, ** and ***, respectively.

The table shows that excluding the first two statements of each year causes *PMI* to lose its significance. However, more importantly, NS_5 is still significant. This shows that the chosen fine-tuning method does not introduce a lot of human bias into the results.

8 Conclusion

In this paper, each FOMC statement is dissected into five topics, for which the sentiment is determined separately. This paper then mostly follows the method described in Rosa (2011) to determine the market's expectancy of the sentiment for each topic. The difference between the market's expectancy and the actual sentiment is defined as the news shock. For each topic, an in-sample OLS regression is performed to investigate whether the news shock significantly impacts the 40-minute return of the two-year bond futures market directly after the release of the statement. Furthermore, this paper investigates whether a set of four reaction asymmetries significantly influences the market's reaction. Then, the in-sample OLS regressions are combined with the reaction asymmetries to improve the fit of the model. Lastly, the findings are applied to the Macroeconomic Random Forest (MRF) model, whose out-of-sample performance is compared with the out-of-sample performance of the OLS models. The results of this paper can be summarised as outlined below.

This paper finds that the market only reacts to the sentences within an FOMC statement that display the FOMC's outlook on future economic conditions and inflation. The surprise of this sentiment is significantly positive, which indicates that a better-than-expected outlook causes a positive reaction to the release of the statement. Next to this, the surprise in changes in the Federal Funds Rate (FFR) is significantly negative, which shows that an unexpected increase (decrease) in FFR has a significant negative (positive) effect on bond future returns. Together, they can explain 11.9% of the variation in returns for an in-sample analysis over the time period of 1999-2020. For the reaction asymmetries, this paper finds that differing FOMC statements do not cause a significantly larger reaction to news shocks. Furthermore, negative surprises in changes in the FFR also do not have a significant effect, but that can be explained by the type of measure that is used to determine this surprise. In addition to this, this paper finds that good (bad) economic conditions cause a significantly decreased (increased) reaction of the market to surprises in the change in FFR, and that good (bad) economic conditions have a significantly negative (positive) effect on short-term returns altogether. Furthermore, the size of the pre-FOMC announcement drift does not impact the market's reaction to news shocks, but a large (small) pre-FOMC announcement drift does have a significantly negative (positive) effect on short-term returns. When a proxy for current economic conditions, the size of the pre-FOMC announcement drift, the surprise in sentiment and surprise in changes in the FFR are combined in a single OLS model, it can explain 21.2% of the variation in returns over the same time period of 1999-2020. Furthermore, this paper provides evidence that both the news shocks for the sentiment and the surprises in changes in the FFR lose their significant predictive power after the financial crisis in 2009.

Performing out-of-sample analysis on the OLS model (both with and without the reaction asymmetries) and the MRF model shows that all three models outperform the benchmark, which in this case is assuming that the reaction equals the historical average. This is tested for eight time windows, ranging from 10 to 120 minutes. The results show that the OLS model that incorporates the earlier mentioned findings consistently outperforms both other models across all tested time windows. Furthermore, the MRF model consistently underperforms, which could be due to a tendency to overfit. In addition, this paper finds that for most time windows the out-of-sample performance is not significant. This is due to a single outlier in the set of predictions. If this outlier is removed from the dataset, most out-of-sample results become significant, but also have a decreased performance. This demonstrates that most of the out-of-sample performance comes from a single outlier. Lastly, this paper shows that the out-of-sample performance peaks in 2009, but then halves over the next 12 years. This is in line with the findings from the in-sample analysis which state that the news shocks lose their predictive power after the financial crisis in 2009.

Thus, to answer the four research questions, this paper finds that the surprise in the sentiment of the FOMC statements, as extracted by the FinBERT model, has significant predictive power for short-term bond future returns. However, this is only the case for one specific topic within an FOMC statement and only holds for the time period before 2009. Furthermore, poor economic conditions cause a significantly larger reaction of the market to news shocks and have a positive impact on the reaction altogether. Lastly, a large pre-FOMC announcement drift has a direct negative impact on the reaction.

For future research, improvements could be made regarding the fine-tuning process of the FinBERT model. This research uses a self-labelled dataset that consists of 229 sentences. However, a lot more text data is available in the form of minutes, speeches and transcripts of meetings. With a better fine-tuned FinBERT model, one might be able to extract a more accurate sentiment from the FOMC statements. This could especially be useful for the period after 2009, as statements are significantly more complex and similar to each other, making it more difficult to obtain the true sentiment. Next to this, *n-grams* are currently excluded from the corpus with which the LDA model is trained. However, adding *n-grams* can provide more detailed topics and an increased insight in the content of a FOMC statement and the topics to which the market responds to. Lastly, to determine the market's expectancy on the FOMC's sentiment, a linear model is used with three variables, following Rosa (2011). However, this relation does not need to be linear, or there could be other variables that have an impact. As this model directly impacts the size and direction of the surprises in the sentiment, a better-performing model could yield more information that could better predict the market's reaction to the release of an FOMC statement.

References

Afshar, T., Arabian, G., Zomorrodian, R., et al. (2007). Stock return, consumer confidence, purchasing managers index and economic fluctuations. *Journal of Business & Economics Research (JBER)*, 5(8).

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Vega, C. (2007). Real-time price discovery in global stock, bond and foreign exchange markets. *Journal of International Economics*, 73(2):251–277.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3.

Bernanke, B. S. (2020). The new tools of monetary policy. *American Economic Review*, 110(4):943–83.

Bernanke, B. S. and Kuttner, K. N. (2005). What explains the stock market's reaction to federal reserve policy? *The Journal of Finance*, 60(3):1221–1257.

Blei, D. M. and Lafferty, J. (2009). Text mining: theory and applications, chapter topic models.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Brown, P. F., Della Pietra, V. J., Desouza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–480.

Cambria, E. and White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2):48–57.

Campbell, J. Y. and Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4):1509–1531.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., and Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 32:288–296.

Chuliá, H., Martens, M., and van Dijk, D. (2010). Asymmetric effects of federal funds target rate changes on s&p100 stock returns, volatilities and correlations. *Journal of Banking & Finance*, 34(4):834–839.

Clarida, R., Gali, J., and Gertler, M. (2000). Monetary policy rules and macroeconomic stability: evidence and some theory. *The Quarterly Journal of Economics*, 115(1):147–180.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291–311.

Coenen, G., Ehrmann, M., Gaballo, G., Hoffmann, P., Nakov, A., Nardelli, S., Persson, E., and Strasser, G. (2017). Communication of monetary policy in unconventional times. *ECB Working Paper*.

Denny, M. J. and Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2):168–189.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *North American Association for Computational Linguistics (NAACL*, 1.

Ehrmann, M. and Talmi, J. (2017). Starting from a blank page? semantic similarity in central bank communication and market volatility. *ECB Working Paper*.

Farka, M. (2009). The effect of monetary policy shocks on stock prices accounting for endogeneity and omitted variable biases. *Review of Financial Economics*, 18(1):47–55.

Farka, M. and Fleissig, A. R. (2013). The impact of FOMC statements on the volatility of asset prices. *Applied Economics*, 45(10):1287–1301.

Friedman, B. M. (2000). Monetary policy. *NBER Working Paper No. 8057*.

Gardner, B., Scotti, C., and Vega, C. (2021). Words speak as loudly as actions: Central bank communication and the response of equity prices to macroeconomic announcements. *Finance and Economics Discussion Series*, 74.

Goulet Coulombe, P. (2020). The macroeconomy as a random forest. *Available at SSRN 3633110.*

Guo, H. (2004). Stock prices, firm size, and changes in the federal funds rate target. *The Quarterly Review of Economics and Finance*, 44(4):487–507.

Gürkaynak, R. S., Sack, B. P., and Swanson, E. T. (2005). Do actions speak louder than words? the response of asset prices to monetary policy actions and statements. *International Journal of Central Banking*, 1.

Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Information Processing & Management*, 54(6):1292–1307.

Hansen, S. and McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99:S114–S133.

Hernández-Murillo, R., Shell, H., et al. (2014). The rising complexity of the FOMC statement. *Economic Synopses*, 23(2).

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9:1735–1780.

Hoffman, M., Bach, F., and Blei, D. (2010). Online learning for latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, 23.

Hu, X., Pan, J., Wang, J., and Zhu, H. (2019). Premium for heightened uncertainty: Solving the FOMC puzzle. *NBER Working Paper No. 25817.*

Jegadeesh, N. and Wu, D. (2017). Deciphering fedspeak: The information content of FOMC meetings. *Available at SSRN 2939937.*

Koenig, E. F. et al. (2002). Using the purchasing managers' index to assess the economy's strength and the likely direction of monetary policy. *Federal Reserve Bank of Dallas Economic and Financial Policy Review*, 1(6):1–14.

Kombrink, S., Mikolov, T., Karafiát, M., and Burget, L. (2011). Recurrent neural network based language modeling in meeting recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2877–2880.

Kuttner, K. N. (2001). Monetary policy surprises and interest rates: Evidence from the fed funds futures market. *Journal of Monetary Economics*, 47(3):523–544.

Law, T. H., Song, D., and Yaron, A. (2018). Fearing the fed: How wall street reads main street. *Available at SSRN 3092629.*

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692.*

Lucca, D. O. and Moench, E. (2015). The pre-FOMC announcement drift. *The Journal of Finance*, 70(1):329–371.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., and Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Mishkin, F. S. (2004). Can central bank transparency go too far? *NBER Working Paper No 10829.*

Orphanides, A. (1992). When good news is bad news: Macroeconomic news and the stock market. *Working Paper, Board of Governors of the Federal Reserve System.*

Patelis, A. D. (1997). Stock return predictability and the role of monetary policy. *The Journal of Finance*, 52(5):1951–1972.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openaiassets/research-covers/languageunsupervised/language understanding paper. pdf.*

Rosa, C. (2011). Words that shake traders: The stock market's reaction to central bank communication in real time. *Journal of Empirical Finance*, 18(5):915–934.

Rosa, C. and Verga, G. (2008). The impact of central bank announcements on asset prices in real time. *International Journal of Central Banking*, 13.

Rozeff, M. S. (1974). Money and stock prices: Market efficiency and the lag in effect of monetary policy. *Journal of Financial Economics*, 1(3):245–302.

Rudebusch, G. D. (2002). Term structure evidence on interest rate smoothing and monetary policy inertia. *Journal of Monetary Economics*, 49(6):1161–1187.

Schwenk, H. (2004). Efficient training of large neural networks for language modeling. In *International Joint Conference on Neural Networks*, volume 4, pages 3059–3064. IEEE.

Sievert, C. and Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pages 194–206. Springer.

Sundermeyer, M., Schlüter, R., and Ney, H. (2012). LSTM neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

Taeyoung, D., Dongho, S., and Shu-Kuei, Y. (2020). Deciphering federal reserve communication via text analysis of alternative fomc statements. *Federal Reserve Bank of Kansas City, Research Working Paper no. 20-14*.

Thorbecke, W. (1997). On stock market returns and monetary policy. *The Journal of Finance*, 52(2):635–654.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Vissing-Jorgensen, A., Morse, A., Cieslak, A., et al. (2015). Stock returns over the FOMC cycle. In *2015 Meeting Papers*, number 1197. Society for Economic Dynamics.

Wachter, J. A. and Zhu, Y. (2018). The macroeconomic announcement premium. *NBER Working Paper No 24432*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100(472):1394–1411.

Appendix A Theoretical background

In this section, a theoretical background is given for both the LDA and BERT model, as they are the two main NLP models that are used in this paper.

A.1 The LDA model

A.1.1 Model description

The LDA model is a generative probabilistic model for text data. The main idea of the LDA model is that documents are mixtures over k latent topics, where each topic is characterised by a probability distribution over the whole vocabulary of the corpus. The LDA model assumes that documents consist of a bag-of-words, where the order of words within a document can be neglected. It further assumes that each document d is generated according to the following generative processs (Blei et al. (2003)):

1. Choose the amount of words (N), according to a Poisson(ξ) process.
2. Choose a topic mixture $\boldsymbol{\theta}_d \sim$ Dirichlet($\boldsymbol{\alpha}$).
3. For each topic, choose a word mixture $\boldsymbol{\varphi}_k \sim$ Dirichlet($\boldsymbol{\beta}$).
4. Determine each of the N words $w_{d,n}$ in document d as follows:
    (a) Draw a topic $z_{d,n} \sim$ Multinomial($\boldsymbol{\theta}_d$).
    (b) Draw a word $w_{d,n} \sim$ Multinomial($\boldsymbol{\varphi}_{z_{d,n}}$).

This generative process can be modelled as a joint posterior distribution of the topic mixture $\boldsymbol{\theta}$, word mixture $\boldsymbol{\varphi}$, the set of topics $\mathbf{z}$ and all words $\mathbf{w}$ in the corpus, which is given as follows:

$$p(\mathbf{z}, \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\varphi} | \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{d=1}^{D} p\left(\boldsymbol{\theta_d} | \boldsymbol{\alpha}\right) \prod_{k=1}^{K} p\left(\boldsymbol{\varphi_k} | \boldsymbol{\beta}\right)) \prod_{n=1}^{N_d} p\left(z_{d,n} | \boldsymbol{\theta_d}\right) p\left(w_{d,n} | z_{d,n}, \boldsymbol{\varphi}\right). \tag{18}$$

By computing this distribution, the topic mixture of a document and the words mixture for each of the k topics is obtained. The topic mixture shows what topic a document likely belongs to, while the word mixture shows which words relate to each topic. In this generative process and the corresponding joint posterior distribution, only the corpus's documents and words $\mathbf{w}$ are observed. The variables $\mathbf{z}$, $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ are all latent. The amount of topics k and the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are chosen in advance. Thus, to use the LDA model, the posterior distribution of the hidden variables, as shown in Eq. (19), should be computed.

$$p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\varphi}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\mathbf{z}, \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\varphi}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})} \tag{19}$$

This posterior distribution cannot be computed directly, but can be approximated by a variety of approximate inference algorithms. One of such algorithms is the online variational Bayes (OVB) algorithm, which is introduced by Hoffman et al. (2010). The main advantage of the OVB algorithm is that it converges much faster than other algorithms, such as the standard variational Bayes algorithm or Markov Chain Monte Carlo sampling, while retaining the same accuracy. Therefore, this algorithm is used to approximate the posterior distribution of Eq. (19).

A.2 The BERT model

The BERT model, created by Devlin et al. (2019), stands for Bidirectional Encoder Representations from Transformers. This model is designed so that it can be used for various NLP tasks, for which it takes relatively little time to train. Furthermore, it can easily be further trained on domain-specific corpora, which improves its performance for specific language uses. Since this paper uses BERT for sentiment analysis; this section focuses on this specific task. For a global overview of the BERT model and its other use cases, see Devlin et al. (2019).

A.2.1 Architecture

Figure 11 shows a high-level overview of the architecture of the BERT model for an input sentence of three words, with sentiment analysis as an NLP task. The figure shows that the BERT model contains four main layers: The embedding layer, encoder stack, classification layer and softmax layer. In the remainder of this subsection, each layer of the BERT model is explained in more detail.

Input

Before BERT can interpret a sentence, it needs to be pre-processed, the first step of which is that the sentence needs to be tokenised. Tokenising a sentence means that the sentence is split up into words or subwords and mapped to their respective IDs, such

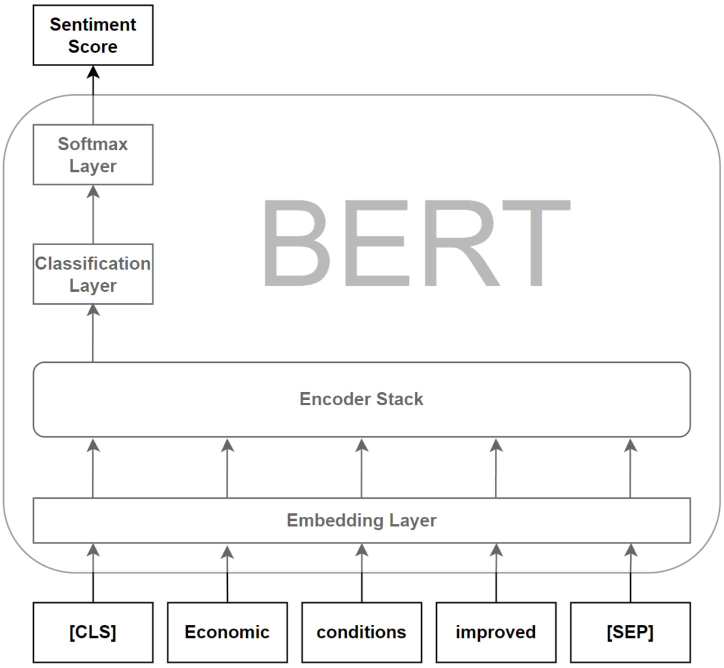


Figure 11: High-level overview of the BERT model.

that each sentence can be conveyed by a vector of numbers, which are called tokens. These vectors can then be used as input for BERT. To perform the tokenisation, BERT uses the WordPiece tokeniser, which is created by Wu et al. (2016). It is a subword tokenisation algorithm which maps common words directly to numbers, but splits more rare and complex words into smaller sub-words before mapping these sub-words to their respective IDs.

Next to this, two extra tokens are added to the sentence: The classification ([CLS]) and sentence separator ([SEP]) tokens. The classification token is placed at the beginning of the sentence and is used by BERT as an output for any classification tasks, such as sentiment analysis. For most other NLP tasks, the encoder stack has an output for each input token. However, in Figure 11 it can be seen that for sentiment analysis the encoder stack has one output which corresponds to the classification token, from which the sentiment of the sentence is deducted.

The sentence separator token is placed at the end of the sentence. This token is typically used when the input consists of more than one sentence, for example when BERT is used to perform next sentence prediction tasks. The input then consists of a classification token, the tokens of sentence 1, a sentence separator token and the tokens of sentence 2. Furthermore, the sentence separator is also used to signal that the end of the input is reached. The base BERT model can take up to 512 tokens as input simultaneously.

Embedding Layer

The first layer of BERT has as its purpose to embed the input tokens; this means that each token is converted into a vector of numbers which hold information about the word itself (token embeddings), to which sentence it belongs (segment embeddings) and its position within the sentence (position embeddings). A representation of this embedding process can be seen in Figure 12. Firstly, the three different embeddings are determined and then summed to form the final embedding for that specific token.

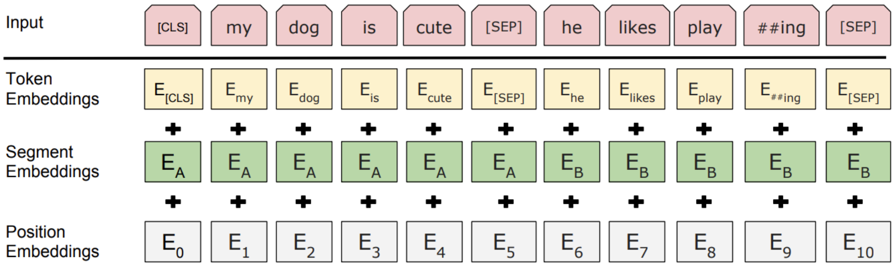

Input
[CLS]
my
dog
is
cute
[SEP]
he
likes
play
##ing
[SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} E_{dog} E_{is} E_{cute} $E_{[SEP]}$ E_{he} E_{likes} E_{play} $E_{\#\#ing}$ $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

Figure 12: Embedding

To obtain token embeddings, the token is mapped onto an embedding space with dimension d. For the base BERT model, this dimension is set to 768. This embedding space places tokens with a similar meaning close together, so that these embeddings hold information about the meaning of the word. The segment embeddings show to which sentence a token belongs, thus whether the token comes before the sentence separator token or afterwards. For sentiment analysis, this embedding is the same for each token, since only one sentence is used at a time. Position embeddings show the position of the token within the sentence. This is determined using a lookup table with dimension 512x768, where each row corresponds to a specific position. The final embedding vector then consists of a summation of the token, segment and position embeddings.

Encoder stack

As mentioned earlier, the base model of BERT consists of $L = 12$ stacked encoders taken from the Transformer network. The input and output of each encoder consists of n vectors with dimension $d = 768$, where n stands for the amount of tokens. The input of the first encoder is the embedding vector for each token, which then gets manipulated and sent to the next encoder. During this process, the dimension does not change. Thus, the output of the encoder stack is still a vector with the same dimension d for each token.

Each of these encoders contains two main layers: A Multi-Head Attention layer and an FFN layer, as seen in Figure 1. In the multi-head attention layer, self-attention is used to calculate an attention vector, which conveys how much each token relates to the others. In every encoder, this is done 12 times for each token, after which a weighted average is taken of these 12 attention vectors. This allows the encoder to capture more information about the relationship between the tokens, since different relationships are discovered in each attention vector. Thus, a total of 144 attention vectors are calculated for a single token. After each multi-head attention layer, the output is normalised and used as input for an FFN network. This network processes the output from the attention layer in such a way that it better fits the input for the next encoder. The parameters in these FFN networks are determined during the pre-training phase of BERT.

Classification Layer

If the BERT model is used for a classification task, such as sentiment analysis, a classification layer is connected to the output of the encoder stack that corresponds to the classification token. This layer consists of an FFN network that converts the vector with dimension d to a vector with dimension l, where l is the number of labels that is used for the classification task. For sentiment analysis, this is set to three, since the output is a vector with values that correspond to how positive, negative or neutral a sentence is. The weights of the FFN network in the classification layer are determined during fine-tuning.

Softmax Layer

Lastly, the softmax layer normalises the output of the classification layer using the softmax function, which is given as follows:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{l} e^{z_j}}, \tag{20}$$

where z is the output vector of the FFN network in the classification layer. The output of the softmax layer is a vector with dimension l which contains the probability of the specific sentence being classified as a certain label. In the case of sentiment analysis, it would contain the probability of a sentence being positive, negative or neutral.

A.2.2 Pre-training

One of the main advantages of the BERT model is that it can be used for various NLP tasks. Furthermore, it takes relatively little time to further train the BERT model for these NLP tasks. This is because the BERT model has already been pre-trained by Devlin et al. (2019), who applied the following two unsupervised learning tasks simultaneously to the model: Masked language modelling and next sentence prediction.

During masked language modelling, Devlin et al. (2019) mask 15% of all tokens in the sentence. The BERT model then has to predict these masked tokens, given all other tokens in the input sequence. This causes the model to understand the bi-directional context within a sentence. Next sentence prediction is a classification task with two sentences as input, which are separated by the sentence separator token. The model then has to predict whether sentence B naturally follows sentence A. Figure 12 shows an example for this task, sentence A: “My dog is cute”, sentence B is: “He likes playing”. In this example, it is clear that sentence B is a natural follow-up to sentence A. This unsupervised task teaches the model how to understand the relationship between different sentences. Using these tasks, the weights in the encoder stack are determined in such a way that the BERT model can understand context and language.

The corpus that is used for pre-training consists of the BooksCorpus (800M words) and English Wikipedia (2,500M words).

A.2.3 Fine-tuning

After pre-training, the model is fine-tuned for a specific NLP task using supervised learning. This is often a relatively straightforward and quick process, since the weights of the encoder stack are already determined during pre-training. When the BERT model is fine-tuned for sentiment analysis, only the weights of the FFN network in the classification layer are determined.

Appendix B Extra checks

B.1 Using Kuttner (2001)'s measure to determine $\boldsymbol{NS}_{\Delta FFR}$

Throughout this paper, the following equation is used to determine $NS_{\Delta FFR}$:

$$NS_{\Delta FFR,t} \equiv \Delta FFR_t - (FFR_{\text{three-month ahead}} - FFR_{\text{one-month ahead}}). \tag{21}$$

For this measure, the sign of $NS_{\Delta FFR}$ does not influence the size of the market reactions, which contradicts Chuliá et al. (2010) and Farka (2009). However, they use Eq. (22) to determine the news shock for the change in FFR. This measure is the standard in the literature, but is not used in this paper since it suffers from look-ahead bias, which this paper aims to avoid. Therefore, this section aims to show that this reaction asymmetry does hold for the measure in Eq. (22) when looking at the short-term returns of bond futures.

$$MPS_t \equiv \Delta f_t \frac{D}{D-d} \tag{22}$$

To this end, the exact same method is used as described in Section 5.2, except for the fact that Eq. (22) is used instead of Eq. (21).

Table 17: Comparison between $NS_{\Delta FFR}$ and MPS.

	$NS_{\Delta FFR}$	MPS
$Intercept$	0.053	0.003
NS	-0.350**	-0.399***
$NS * I_1$	-0.264	-0.108
$NS * I_2$	-0.086	-0.291***
I_1	-0.039	0.011
I_2	-0.039	0.014
R^2_{adj}	0.100	0.127

This table shows whether the sign of $NS_{\Delta FFR}$ and MPS have a significant impact on the size of the market reactions to the mentioned news shocks. The sample period runs between the 15th of May 1999 and the 16th of December 2020 and contains 172 observations. The 10%, 5% and 1% significance levels are denoted by *, ** and ***, respectively. Furthermore, throughout these regressions, all unannounced FOMC statements are excluded.

The result is shown in Table 17, where I_1 is set to one if the surprise has a positive sign and I_1 is set to one if the surprise has a negative sign. The table shows that in this case, $NS * I_2$ is highly significant. Thus, the reaction asymmetry holds for the measure given in Eq. (22).

B.2 Removing an outlier in the set of out-of-sample predictions

In Figure 13, the out-of-sample predictions for a time window of 40 minutes are displayed for each of the three models. As noted in Section 5.3, there is a significant outlier in this figure on the 16th of December, 2008, where the full OLS model predicts a log return of 0.477%. During this meeting, the FOMC unexpectedly cut the FFR by 75 basis points. Next to this, the other two models predict similar values. As discussed in Section 5.3, the test-statistic in Eq. (13) is regressed on a constant, whose significance level indicates the significance level of the R^2_{oos}. Thus, the higher the significance level, the lower the variation of the test statistic. Therefore, a single significant outlier can skew the results of the regression.

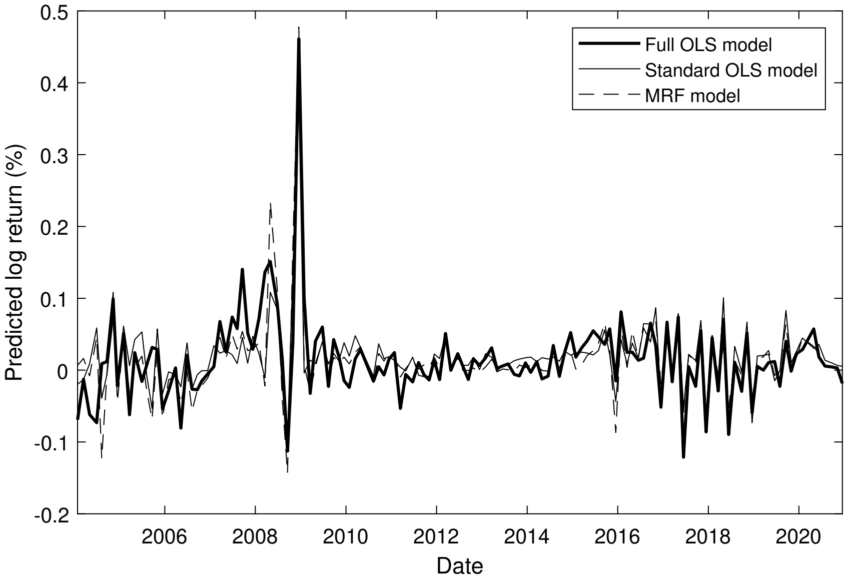


Figure 13: The out-of-sample predictions of the full OLS model (thick solid line), standard OLS model (thin solid line) and the MRF model (dashed line). The time period runs from the 28th of January, 2004 till the 16th of December, 2020.

To measure the effect of this outlier on the significance level of R^2_{oos}, this outlier is removed from the dataset and the results of Section 7.4 are replicated. The results are given in Table 18 and show that most R^2_{oos} become more significant, which indicates that the single outlier skews the results in Table 14. However, an extra note is that most R^2_{oos} are much smaller than their respective values in Table 14. This shows that most of the gain in R^2_{oos} comes from this single correct prediction, since removing this large outlier causes the out-of-sample performance to roughly halve across all time windows and all models.

Table 18: R^2_{oos} for the standard OLS, full OLS and MRF model for various time windows.

	10	20	30	40	50	60	90	120
Standard OLS	0.020*	0.028*	0.040*	**0.041***	0.018	0.013	0.019	0.038*
Full OLS	0.071**	**0.085*****	**0.085****	0.063**	0.050*	0.040*	0.037*	0.039*
MRF	-0.030*	0.004*	0.036*	**0.047***	-0.007	-0.016	-0.015	0.004*

This table shows the R^2_{oos} for the standard OLS, full OLS and the MRF model. These values are obtained using an expanding window, where the sample period runs from the 28th of January 2004 till the 16th of December 2020. This dataset excludes the FOMC statement on the 16th of December 2008. The sample contains 133 predictions. The 10%, 5% and 1% significance levels are denoted by *, ** and ***, respectively. The value in bold denotes the highest R^2_{oos} for each model. Furthermore, throughout these regressions, all unannounced FOMC statements are excluded.