A Deep Learning approach to FOMC Sentiment

M. Osborne R. D. L. Hanes M. V. Cassidy

December 9, 2020

Abstract

Communication is one of the main mechanisms that central bankers use to exert monetary policy. Sentiment expressed in xxx with respect to yyy... in particular has been empirically shown to be informative within the context of modelling policy. We go beyond published work and generalise these findings, moving from simple dictionary approaches to Deep Learning-based methods.

After demonstrating that these more advanced sentiment models improve on predicting the sentiment of individual FOMC sentences, we then consider our sentiment models in the context of monetary policy rules and the modelling of interest rate changes. We find that the best explanation of observed FOMC interest rate policy is obtained when using all monetary policy rules in tandem with all sentiment models. Finally, we qualitatively explore sentiment against a variety of economic indicators and named economic shocks.

TODO

Brain dump of all tasks left:

- Come up with a list of possible target jnl venues.
- Record location of all data sources in a json/txt file
- Consider adding more summary statistics about the data (section 3)

- Integrate ECB examples into the text.
- Label ECB examples and compute performance statistics.
- Emphasise contributions
  - Gazetteer-based vs ML
  - Multiple policy rules
  - Aggregating from sentence to paragraph to section to minutes
  - Taylor rule components
  - all of the above together
- Add equations to §4
- Experiment with varying the 90% in §5 or at least explain it better.
- What is the impact of "Specifically we selected 43 documents from regimes where there was considerable economic uncertainty" on the external validity of the experiments?
- The end of Section 6.2 needs more details, and there is another todo in that section.
- Add Bonferroni correction for multiple comparisons?

1 Introduction

Central Bankers (⇐ capitalized?) exert monetary policy through decisions such as setting interest rates. Orthogonally, Central Banking *communication* can be seen as another means to exert monetary policy [5]. This communication can include discussions of macro-economic climates, possible outlooks, risks and a whole host of other factors. Given how important Central Banks are to open markets, policy communication is closely monitored and reacted to by markets.

Now, communication can be overt (for example, actually discussing topics such as COVID-19) or can be implicit and expressed through sentiment and tone. A line of research aims to quantitatively capture such sentiment. For example, [3, 10, 13, 16] do xyz . . . let's elaborate. Frequently these approaches

rest upon handcrafted dictionaries, with explicit lists of words for positive, neutral and negative sentiment.

We go beyond such approaches and bring to bear recent developments in machine learning when modelling sentiment. Specifically, we show how Deep Learning-based models (⇐ capitalized?), trained on massive quantities of general language can be used to predict the sentiment of complex Federal Open Market Committee (FOMC) minutes without any handcrafting, other than labelling examples of sentences with their sentiment. We compare our Deep Learning-based model against a commonly used dictionary-based baseline, as well as against a simpler machine learning-based approach and show that—at the sentence-level—better modelling carries over to better sentiment prediction over unseen sentences.

We then extend a variety of monetary policy rules and explicitly incorporate sentiment to better capture FOMC interest rate changes. We show how our improved way of gauging sentiment yields improved predictive power and, in particular, that the best explanation of observed rate changes is found when using all of our monetary policy rules and sentiment models together. Finally, we qualitatively explore relationships between our sentiment models and other economic time series, revealing interesting patterns and reinforcing the fact that there is a close relationship between FOMC sentiment and the wider economic world.

The structure of the rest of this paper is as follows. After discussing related work (Section 2) we give a description of FOMC minutes and sentiment (Section 3), followed by a presentation of our sentiment models (Section 4). Next (Section 5), we show how these models can be used to extend monetary policy rules. Having set the scene, we then explore these sentiment models in a series of experiments (Section 7). The first set of experiments consider sentiment prediction at the sentence level. We then move onto seeing how well our models perform at a less direct level, namely within the context of extending Tayloresque monetary policy rules. Afterwards (Section 7.5) we construct a time series of meeting-level sentiment and compare it against a variety of other economic indicators. The paper ends in Section 8 with a discussion.

2 Related Work

Blinder et al. [5] surveyed research on Central Bank communication, sum-

marising the point that it is used as an additional tool for guiding markets. Market expectations of interest rates evolution unfolds gradually and this evolution can be guided by such communications. In the short term, asset prices are known to move on the basis of Central Bank sentiment [17]. Federal Reserve communications (in tandem with Taylor Rules) are known to help model interest rate decisions [11].

Inferring the sentiment of a snippet of text with respect to a certain target based upon dictionaries is popular. For example, Baranowski et al. [4] used this technique for the European Central Bank, whilst Cannon [6] used dictionaries to analyse the FOMC communications. Other dictionary-based studies in this context include [3, 10, 13, 16].

Sentiment has been used within the context of Taylor Rules [12, 13, 18]. Typically this is performed in a regression setting with only a single monetary policy rule is considered. We go beyond such approaches and use multiple policy rules as well as multiple ways to compute sentiment. A key difference is that we work over raw sentiment scores, rather than attempt to isolate sentiment information (i.e., so-called "animal spirits") that is independent of factors such as business cycles, market behaviours and so on [2]. Isolating exogenous shocks poses empirical challenges such as having to specify which factors we should exclude, as well as reasoning about the extent to which these named factors captures these factors.[1] Algaba et al. [1] survey sentiment and econometrics in a broad setting, touching upon both dictionary and machine learning approaches.

3 FOMC Communications and Sentiment

The Federal Reserve Bank issues a host of communications, ranging from press releases and individual speeches to collective minutes recording decision making processes. These communication items are issued relatively regularly throughout the year and, as mentioned earlier, help markets understand and anticipate their decision making. Minutes in particular are the most detailed type of communication, offering insight into collective decision making, degrees of certainty, economic topics and commentary. Typically these consist of long reports that are split into *sections*, with some sections backwards-looking whilst others look forwards. Sections in turn are com-

[1] In unreported experiments we computed sentiment shock versions of our raw sentiment scores and found that performance degraded.

posed of paragraphs of text, where each paragraph contains fluent and at times quite densely written sentences. Meetings generally have a consistent structure and include sections reviewing developments, economic and financial conditions as well as staff economic outlooks and, most importantly of all, views on the current economy.[2]

Sentiment (or tone) within the context of Central Banking can be thought of as the extent to which members are hawkish, dovish or centerist (neutral) about various aspects of the open economy and especially with respect to interest rates. Historically FOMC communication was famously seen as being opaque ('turgid') but, as is well known, this transitioned to greater transparency and simplicity. Under the current chair, Jerome Powell, press releases are expressed in plain English. This is in contrast to previous releases that were couched in complex economic jargon. Clearly, the extent to which sentiment can be captured will hinge on how complex the language used is. Nevertheless, we note that FOMC meeting minutes (the subject of our work) continue to be densely written.

Given how complex FOMC communication can be, a natural question is whether it is even meaningful to capture sentiment. For example, if trained economists cannot agree on the sentiment of FOMC communication, machines are unlikely to do any better. Table 1 shows three example sentences illustrating the difficulty of this task. As can be seen, even for humans, this can be challenging. [Do we wanna move this paragraph to the "Data" section:?] To answer this question we measure the inter-annotator agreement rate, which considers the extent to which two humans agree on labelling examples and is captured by Cohen's kappa coefficient [7]. A value of 1.0 means the annotators are in complete agreement, whilst a value of 0.0 implies there is no agreement other than what would be expected by chance. Low agreement rates imply the task is highly subjective and puts an upper bound upon how well a machine might do. We asked two domain experts to independently label the same 25 sentences, randomly sampled from the 1821 sentences we already had labels for (see Section 6.2). The inter-annotator agreement value calculated from this sample is 0.60, implying that the task is well-founded, i.e., there is 'good' agreement, but still has room for disagreement.

[2]For older meetings, some section titles have changed and some sections were merged. In order to maintain a coherent view of each section over time, we use a heuristic to map sections in older documents to their counterparts from recent documents.

Table 1: Example FOMC Sentences

Sentence	Source	Comments
Regarding the labor market, many participants commented that the pace of employment gains, which was quite strong in May and June, had likely slowed.	July 28-29, 2020	*Sentiment is governed by the final clause.*
One of them judged that the low level of inflation compensation could reflect increased concern on the part of investors about adverse outcomes in which low inflation was accompanied by weak economic activity, and that it was important **not** to dismiss this possible interpretation.	March 17-18, 2015	*Negation does not invert sentiment.*
Recent data along with anecdotal reports indicated some loss of vigor in the nation's housing markets, though overall activity was still at a high level.	December 21, 1999	*Mixed sentiment.*

4 Sentiment Modelling of FOMC Meetings

We now explain how we compute a sentiment score for FOMC meetings which will be used both directly as well as input to predict interest rate changes, as detailed in the next section. Our approach to predicting the sentiment associated with FOMC meetings is to first predict sentiment at the sentence level. We then aggregate these sentence-level predictions, eventually

Table 2: Top most highly weighted features per class (Linear model)

Neutral		Positive		Negative	
Feature	Weight	Feature	Weight	Feature	Weight
committees	1.02	increased	1.12	weakness	1.23
committee	0.93	gains	0.88	weak	1.14
had remained	0.87	strong	0.88	uncertainty	1.10
mixed	0.86	rise	0.86	below	0.82
information	0.78	up	0.82	short	0.78

producing a sentiment score for entire FOMC meeting minutes.

4.1 Sentence-level Sentiment Models

The first approach (called **Simple**) is a dictionary-based method that uses Financial wordlists compiled by Loughran and McDonald [14]. As such, it is perhaps the most commonly used dictionary for this task. There are two wordlists; the set of positive words totals 354, whilst the negative words totals 2355. As we will see later, this imbalance results in a tendency towards predicting negative sentiment (see Section 7.5.2). Essentially for sentences, it measures the ratio of positive to negative sentiment-bearing words as found within the wordlists. Sentence order is ignored and there is no term-specific importance weighting. The definition of sentiment therefore rests solely on the choice of words. Cannon [6] considered the impact of using other wordlists, including more general consumer-based versions and found that ....

Our second pair of approaches are more data-driven in that they do not rely on pre-defined lists of words, but rather upon sentences taken from FOMC minutes that have been labelled with macroeconomic sentiment (see Section 6.2 for details). The first approach (called **Linear**) is a multiclass logistic regression machine learning classifier which predicts the sentiment label of unseen sentences. It is trained on a set of labelled sentences for which it takes each sentence, tokenises it into words and builds a vocabulary containing all the words and word pairs. It then uses numerical optimisation techniques maximising the conditional log-likelihood of all meetings in order to infer weights associated with these words and word pairs, expressing how

much they contribute towards an overall positive, negative or neutral sentiment. The final classifier has one model per sentiment class and uses a one-versus-all scheme at prediction time.
[Add Equation]

Some examples of words and word pairs with the largest weights are shown in Table 2. To calculate the sentiment of unseen sentences, we perform the same steps as for training and discard words that were not in the model's vocabulary, i.e., that were not seen at training time. One of the advantages of this approach compared to using a predefined set of words is that it learns which words and word pairs are important to the sentiment score from the data itself. One disadvantage is that it largely ignores the order of words within sentences and so only has a weak understanding of syntax.

Our final approach (called **Deep Learning**) is also machine learning-based, utilising a neural network architecture called a Transformer [19]. This model has a number of extensions over our simple linear classifier. Firstly, it models sentences as sequences rather than as sets of words and word pairs, thereby capturing syntax. This, in turn, gives it the capacity to learn more complex relationships between words. Secondly, it decomposes words into subword units, thereby better handling newly seen words. It does this by breaking unknown words apart into sub-word units (stopping when it reaches a unit it does know) the smallest of which are individual characters. Finally, it exploits the idea that many natural language processing tasks are actually fairly similar to each other and can be trained ahead of time using massive quantities of found and possibly unrelated text (such as Wikipedia or a crawl of the World Wide Web). These models are therefore task-neutral and then *fine tuned* for the task in question, which here is sentiment prediction.
[Add Equation]

We use a very large model that has been pretrained by others on a huge corpus of text[3] [8]. The pretraining step primes the model with a general understanding of how words relate to one another. We then fine tune the model on training data in order to train a model that can classify sentiment. The number of parameters in the model is around 100M, compared to the 66k parameters of the Linear model. This means it takes a lot more computational power to train and perform inference, but it achieves better results.

[3] 2.5B words from the English language Wikipedia and 800M words from BooksCorpus

4.2 Meeting-level Sentiment Model

Our sentiment models each assign a sentiment label to sentences: positive, neutral or negative. Because meetings consist of paragraphs, organised into *sections*, we need to compute sentiment at the sentence and section level, and then ultimately at the entire meeting level.

For each sentence we predict the sentiment using a sentence-based model. Negative sentiment is assigned a score of -1; neutral sentiment a score of 0 and positive sentiment a score of 1. As discussed earlier (Section 3), meeting minutes are generally structured into sections, each of which consists of sentences. Empirically we find that the bulk of useful sentiment information is ⇐ if I were a reviewer I would want to know how/why present within the "Participants View" (PV) section(⇐ Above we said that another section ("views on the current economy") was more important. Let's be consistent.). We therefore assign a weight of 90% to this section and assign the remaining 10% equally to all other sections. ⇐ magic number? The score assigned to a section is then the weighted average sentiment score of all sentences within it. The overall sentiment score of a meeting is then the average section scores. Paragraph boundaries are ignored.

5 Modelling Interest Rate Changes

We now show how we incorporate our previously mentioned sentiment models into monetary policy rules. Following [12], we assume interest rate changes are smooth with respect to previous interest rates:

$$i_t^* = \rho i_{t-1} + (1-\rho) i^F, \tag{1}$$

where i_t^* represents the current interest rate at meeting time t, $i_{t-1}$the previous meeting's interest rate and i^F is the optimal (target) interest rate guiding decision making. ρ is a mixing parameter governing the relative contributions of these two components. Note that time here is measured in terms of when meetings happen, so the notation $t-1$ means with respect to the previous meeting. Meetings can be traditional events with associated minutes or—as in recent times—video calls and so on. We use the term "meeting" to stand for any event where a rate change might occur and where there is some kind of record concerning FOMC decision making.

Within a monetary policy rule-based approach, the target interest rate can be modelled taking FOMC sentiment into account as follows:

$$i_t^F = \eta i_t^P + (1-\eta) i_t^S, \tag{2}$$

where i_t^P refers to a *family* of monetary policy rules (including a standard Taylor Rule) explaining the interest rate, whilst i_t^S is a component that captures the *communicative intent* of the committee. This communicative intent is represented through committee sentiment towards the economy. As before, we have mixing parameters controlling the relative contributions.

Table 3: Monetary Policy Rules

Name	Specification
Basic Taylor Rule	$i_t^T = r^* + \pi^* + 0.5(\pi^* - i) + 0.5(y_t - y_t^P)$
Inertia Rule	$i_t^{Ti} = 0.85 \cdot r_{t-1} + 0.15(r^* + \pi^* + 0.5(\pi^* - i) + (y_t - y_t^P))$
First Differences Rule	$i_t^{Tf} = r_{t-1} + 0.1(\pi^* - i) + 0.1(y_t - y_{t-4})$
Taylor Rule with Okun's Law	$i_t^{To} = r^* + \pi^* + (A \cdot (\pi^* - i)) + (B * \Sigma * (U^* - U))$

We consider the following set of policy rules i_t^P shown in Table 3. The first three rules are named on the Federal Reserve Website.[4] The final rule is in the spirit of rules that encode Okun's law [15]. Table 4 outlines the variables and the values they take here. We do not expect that changing these settings will dramatically alter any of the results we report upon.

The communicative intent can be captured as follows:

$$i_t^S = \alpha_s S(t). \tag{3}$$

Here, $S(t)$ is the predicted sentiment of the FOMC at meeting time t and α_s is a scaling constant should we wish to directly regress interest rates from our sentiment models.[5]

We use a standard logistic regression machine learning classification model to capture when the FOMC raises, decreases or maintains interest rates on a

[4] https://www.federalreserve.gov/monetarypolicy/policy-rules-and-how-policymakers-use-them.htm

[5] Our sentiment models are not on the same scale as interest rates.

Table 4: Policy Rule Components

Symbol	Indicator	Value
r^*	Neutral Real Rate	2%
A	Weight given to the Inflation Gap	0.5
i	Target inflation rate	2.0%
π^*	Current rate of inflation	US Personal Consumption Expenditure Core Price Index[a]
B	Weight given to the Unemployment Gap	0.5
Σ	Factor to convert from the Unemployment to the Output gap	2.0
U^*	Non-accelerating Inflation Rate of Unemployment (NAIRU)	5.0
r_t	Actual upper bound on FOMC interest rates.	FDTR Index.
y_t	The logarithm of Real Gross Domestic Product (GDP) in the quarter containing time t[b]	GDPC1 Index
y^P	The logarithm of Real Potential Gross Domestic Product in the quarter containing time t[c]	GDPPOT Index
U	Current unemployment rate	US Unemployment Rate Total in Labor Force Seasonally Adjusted[d]

per meeting basis. Free parameters are set as experimental conditions. For example, should we decide to ignore interest rates, we can set the associated weight to zero. Turning this into a regression rather than a classification setting means solving for interest rates after each meeting.

Taking Equation 2, the family of equations in Table 3 and Equation 3 together, we have the following target for optimisation:

$$\Delta it = \alpha_1 i_{t-1} + \alpha_2 i_t^T + \alpha_3 i_t^S, \tag{4}$$

[Miles: shouldn't that i_t^T be i_t^F ?] where $\Delta i(t)$ represents the decision at meeting time t; α_1 to α_3 are free parameters. Equation 4 explains FOMC decision making as a combination of previous interest rates, a monetary policy interest rate and our sentiment modelling component. Later on, we might add

multiple sentiment models and policy rules. This simply means adding corresponding extra terms to Equation 4. Note the model has three parameters per term, i.e., one per sentiment class, and so has a linear complexity. We have deliberately designed it so that the model cannot trivially account for all interest rate decisions by memorisation, allowing it to support interpretation at the same time.

6 Data

In this section we detail how we obtained the minutes we use and how we label our data.

6.1 From Minutes to Sentences

We used FOMC meeting minutes from January 2000 until August 2020, a total of 177 documents. The FOMC provides access to recent and historical meetings via their website, which we used to download the HTML version of each document[6]. We manually mapped minute release dates to actual meeting dates or, more specifically, the date of the press release when the rate change became public. We then used the rate change date to identify the minutes documents throughout the remainder of this study. We extracted text from each section of the document, using the HTML markup to identify the section boundaries and titles. Whilst processing the meeting data we noticed that during times of significant economic turmoil, the FOMC makes more frequent rate decisions. These unplanned decisions are based on telephone or video conferences, with the press release the same day. The transcript of these conferences is then appended to the prior (if the conference is with the 3 week lag time) or next official meeting document. We extracted 17 such events and treat them as full meetings with a single section. Having identified named sections, we tsplit them into sentences, strip punctuation and lower case the terms. We did not stem or remove stop words.[7]

[6]Meetings prior to 2007-09-19: https://www.federalreserve.gov/fomc/minutes/, Meetings from 2007-09-19 until the present: https://www.federalreserve.gov/monetarypolicy/fomccalendars.htm

[7]Our Deep Learning model in Section 4.1 works directly with raw sentences and so we do not need to process sentences when using it.

Table 5: Regimes where we annotated documents.

Date range	Regime
Dec 2000 - Dec 2001	Strong 'cut' after the .com crash
June 2004 - June 2005	Rise in rates before the Great Recession
June 2007 - June 2008	Cut in rates leading into the Great Recession
Jan 2011 - Dec 2011	Lengthy trough associated with the Great Recession
Jan 2018 - Dec 2018	Bulk of the latest 'hiking' language

Table 6: Sentiment distribution over labelled sentences

Sentiment	Percentage
Negative	38.2
Neutral	27.4
Positive	34.4

6.2 Labelling Sentences for Sentiment

The full range of 177 documents contained approximately $100,000$ sentences in total. Clearly we cannot manually label all of them, so we decided upon focussing labelling attention on interesting economic intervals. Specifically we selected 43 documents from regimes where there was considerable economic uncertainty (Table 5) and the three most recent documents, for a total of 46 meetings.Check these numbers in the raw data From these documents, we labelled the sentences in the *Participants' Views* section, giving a total of 1821 labelled sentences.

Table 6 shows the distribution over sentiment labels for these sentences. As can be seen, this is approximately equal (with an overall preference for negative polarity). Compare this with our dictionary, which has an imbalance towards negative terms (Section 4.1). So, ...

7 Experiments

We now explore how our sentiment models behave both intrinsically, i.e., within the context of monetary policy rules, and in a time series setting.

7.1 Evaluation Details

For our classification results—both at the sentence-level and when modelling interest rate changes—we measure how well we can predict labels. In particular, we use F-scores which consider both recall and precision together. This, rather than accuracy, is appropriate when dealing with imbalanced labels. For example, most rate decisions maintain the status quo. Simply always predicting that rate decisions are maintained will yield misleadingly high accuracy levels, even though rarer—and arguably more important—rate changes are never predicted.

Our sentence-based experiments are all *out-of-sample* as they operate over unseen sentences. Our subsequent monetary policy experiments are all *in-sample* in that they operate over seen meetings, albeit only partially observed by the underlying sentence-based approaches how so? because some of the training set sentences are in there?.

When reporting on statistical significance for classification tasks, it is common to use McNemar's Test [9]. The null hypothesis is that errors made by two classifiers are statistically indistinguishable from each other. For our sentence-level models it is feasible to compare all models against each other. For our rate change experiments we take as our yardstick a simple Taylor Rule model.

7.2 Sentential Model Training

We divided our labelled data into train-development-test splits. Sentences in the training set were used for our Linear and Deep Learning models. Development sentences were used for tuning the hyperparameters of the models and test sentences were used only for evaluation purposes. Figure 7 summarises our datasets.

These all consisted of sentences drawn from non-overlapping meetings and with training sentences chronologically earlier than development sentences, which in turn were from meetings earlier than test sentences.

Table 7: Statistics on FOMC Meetings that were labelled

Split	Date Range	Number of Sentences Labelled
Training	2001-02-01 – 2018-10-17	1518
Development	2019-01-09 and 2019-07-10	136
Testing	2019-08-21 and 2019-10-09	166

7.3 Sentence-level Experiments

Before diving into modelling the sentiment of meetings it is worthwhile evaluating how well we can model sentiment at the sentence level. This is an intrinsic evaluation and,to the best of our knowledge, related work does not directly evaluate how well sentence-level sentiment models perform as any form of evaluation tends to be indirect, i.e., on another task, instead.

Table 8: Sentence-level evaluation (F1 scores)

Model	Negative	Neutral	Positive
Simple	0.62	0.28	0.43
Linear	0.60	0.41	0.66
Deep Learning	0.72	0.55	0.80

Table 8 shows how well our sentence models perform on unseen, out-of-sample sentences. As can be seen, the two machine learning approaches (Linear and Deep Learning) both outperform the simple dictionary approach. Additionally we see the Deep Learning approach (which is capable of modelling sentence order) outperforms the simpler linear model (which ignores sentence order). Neutral sentiment appears to be the most challenging to predict. Pairwise differences between the models Simple, Deep Learning and Linear are statisically significant at the $\rho > 99.9\%$ level. The pair of models Simple and Linear are statistically significant at the $\rho > 96.8\%$ level. Table 9 shows examples from the unseen test set where the different models produced different results.

It is interesting to see whether our Deep Learning model finds hard sentences that humans also find difficult to assign sentiment to. We took 25 unseen sentences that the Deep Learning model classified incorrectly and

Table 9: Example sentences with their true and calculated sentiment labels, with ⊕ indicating positive, ⊙ neutral, and ⊖ indicating negative sentiment.

Sentence	True Label	Predicted Sentiment		
		Simple	Linear	Deep Learning
In their discussion of the outlook for inflation, participants generally anticipated that with appropriate policy, inflation would move up to the Committee's 2 percent objective over the medium term.	⊕	⊙	⊕	⊕
Some participants observed that trade uncertainties had receded somewhat, especially with the easing of trade tensions with Mexico and China.	⊕	⊖	⊖	⊕
Furthermore, inflation pressures continued to be muted, notwithstanding the firming in the overall and core PCE price indexes in the three months ending in June relative to earlier in the year.	⊖	⊙	⊕	⊖

collected labels from a second expert annotator. We then computed agreement rates (and hence kappa) between the original human annotator and this second expert. The calculated value of kappa for these sentences was 0.07, meaning that the agreement between the human annotators on these challenging sentences was just above chance. Humans and our Deep Learning model both find the same examples hard to label, which is encouraging and suggests our automated approach is mimicking the behaviour of our domain expert. Table 10 shows some examples of sentences where the Deep Learning model and human annotators disagreed.

Table 10: Sample sentences with disagreement, with ⊕ indicating positive, ⊙ neutral, and ⊖ indicating negative sentiment.

Sentence	Assigned Labels		
	Human 1	Human 2	Deep Learning
They also argued that it was desirable for the Committee to seek and maintain a level of accommodation sufficient to deliver inflation at 2 percent on a sustained basis and that such a policy would be consistent with inflation exceeding 2 percent for a time.	⊕	⊖	⊙
Several participants noted that statistical models designed to gauge the probability of recession, including those based on information from the yield curve, suggested that the likelihood of a recession occurring over the medium term had increased notably in recent months.	⊙	⊖	⊕
Several also noted that, because monetary policy actions affected economic activity with a lag, it was appropriate to provide the requisite policy accommodation now to support economic activity over coming quarters.	⊕	⊖	⊙

7.4 Policy Rule Explanation Experiments

We now consider using our sentiment models within the context of Policy Rules. Here we indirectly evaluate the contribution sentiment makes to modelling.

Table 11 shows our explanation results (Equation 4) using just Monetary Policy rules, then these rules with lagging and finally three sets of experiments where we additionally add sentiment models to lagged policy rules. Policy rules alone largely only capture maintaining interest rates. Adding lagging information does not make any significant differences to these results. When we add our sentiment models, we start to see relatively large improvements. In particular, we begin to be able to explain interest rate decreases as well as increases. Interestingly the Taylor Rule with Okun, when augmented with lagging and our Deep Learning model best explains the data. For some of the rules (especially the basic Taylor Rule), we see no gains over using a simple dictionary approach. Significance results are added in parentheses, taking the original Taylor model as our point of reference.

Now, given that we have multiple policy rules and sentiment models, it is worthwhile investigating whether these rules and models can combine together in a useful way. If this is true, it would suggest that rate change decisions are better modelled in terms of collective (aggregated) decision making, rather than focussing upon simple rules and models in isolation. Table 12 shows modelling results when we use all policy rules and all sentiment rules. We also show a kitchen-sink approach where we use everything apart from our Deep Learning model and finally where we use everything. All settings use lagging information. These results improve over our previous findings. Additionally, we see that sentiment models alone do worse than when using all policy rules, but that the best results are obtained when using everything together. Finally we note that not using our Deep Learning approach reduces performance. Together, this suggests that rate decision modelling is best tackled as a consensus over different policy rules and sentiment models. This however brings greater complexity and makes interpretability harder.

7.5 Sentiment as a Time Series

We can treat our sentiment models as time series and examine how they relate to other economic indicators. Finally, it is interesting to examine specific economic shocks such as the Sub Prime Crisis or the more recent

Table 11: Explaining FOMC Decisions using Sentiment

Model	F Score	Rate Change Decreased	Maintained	Increased
Just policy rule				
TR (N/A)	0.27	0.00	0.80	0.00
TR with Okun (94.2)	0.31	0.00	0.82	0.12
Inertia (< 80)	0.33	0.17	0.81	0.00
First Differences (< 80)	0.33	0.17	0.81	0.00
Policy rule and lagging				
TR (< 80)	0.32	0.12	0.79	0.05
TR with Okun (< 80)	0.35	0.06	0.80	0.17
Inertia (< 80)	0.32	0.12	0.79	0.05
First Differences (< 80)	0.32	0.12	0.79	0.05
As above, but with sentiment				
Simple				
TR (83.8)	0.53	0.43	0.81	0.34
TR with Okun (< 80)	0.49	0.37	0.78	0.31
Inertia (89.7)	0.53	0.43	0.82	0.35
First Differences (88.3)	0.54	0.47	0.81	0.34
Linear				
TR (88.3)	0.51	0.21	0.81	0.52
TR with Okun (95.0)	0.52	0.23	0.82	0.50
Inertia (88.3)	0.51	0.21	0.81	0.52
First Differences (< 80)	0.50	0.21	0.80	0.51
Deep Learning				
TR (91.7)	0.53	0.31	0.81	0.48
TR with Okun (90.4)	0.56	0.38	0.81	0.50
Inertia (91.7)	0.53	0.31	0.81	0.48
First Differences (80.2)	0.52	0.31	0.80	0.43

Table 12: Explaining FOMC Decisions using Collective Modelling

Model	F Score	Rate Change Decreased	Maintained	Increased
All Sentiment Models	0.57 (93.7)	0.47	0.82	0.43
All Policy Rules	0.61 (99.6)	0.38	0.83	0.60
Everything (no Deep Learning)	0.68 (99.7)	0.53	0.85	0.67
Everything	0.70 (99.8)	0.57	0.86	0.67

COVID-19 pandemic.

7.5.1 Comparisons between Economic Indicators and Sentiment

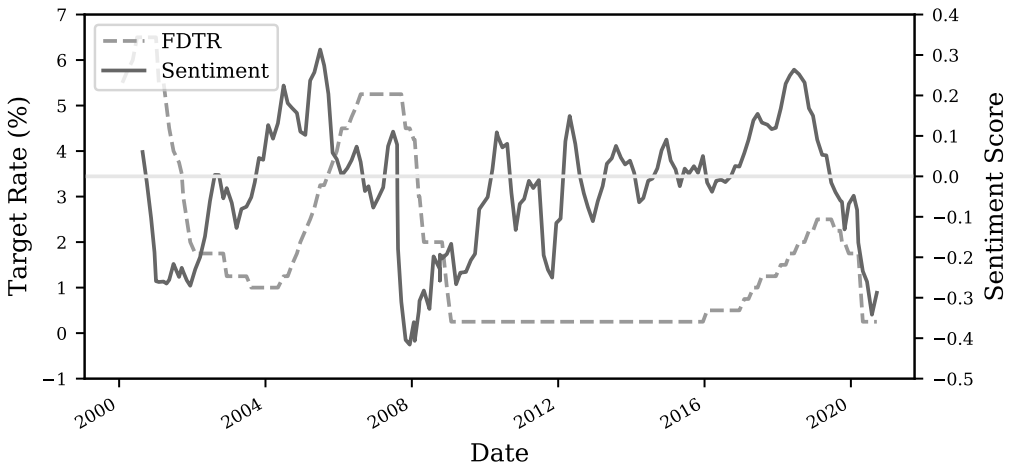

Figure 1: Federal Funds Target Rate (Upper Bound) on the left axis and the smoothed sentiment score calculated using the Deep Learning model on the right axis. The sentiment was smoothed using a size simple moving average over the previous 5 meetings.

Figure 1 shows the Federal Funds Target Rate, Upper Bound (FDTR Index) and a smoothed version of the Deep Learning model produced sentiment score. We averaged the sentiment score using a sliding window containing the previous 5 meetings to highlight the salient features and reduce the noise. The changes that the FOMC make to FDTR coincide and in some cases are

led by changes in the sentiment score. Even with the averaging, the sentiment appears to lead the rate change in a few cases: late 2003 the sentiment rises above 0.0 just before the increase, mid 2007 the sentiment drops below zero before the sharp drop in rates. This implies that the intention of the FOMC to modify rates is already coming through in meetings before the rate change decision. This does not appear to be the case for the COVID-19 pandemic, where the sentiment and rate changes occour contemporaneously. It is unlikey that the FOMC knew how seriously the economy would be affected in the meetings prior to the discovery of the virus.

Figure 2 shows the relationship between the spread between two and ten year treasuries and the same smoothed sentiment score as before. As can be seen in the Spread timeseries, there are broad regions: 2000 to 2004, 2004 to 2008, 2008 to 2011, 2011 to 2014 and finally 2014 until 2020. From 2004 until 2008 we see that sentiment inverts the corresponding spread, perhaps reflecting an increasingly positive outlook corresponding with increased market volatility. This inversion can be seen again from 2012 until 2020 (before the Pandemic), presumably due to markets recovering after the Sub Prime Crisis, where sentiment becomes increasing positive and the spread is seen to decrease.

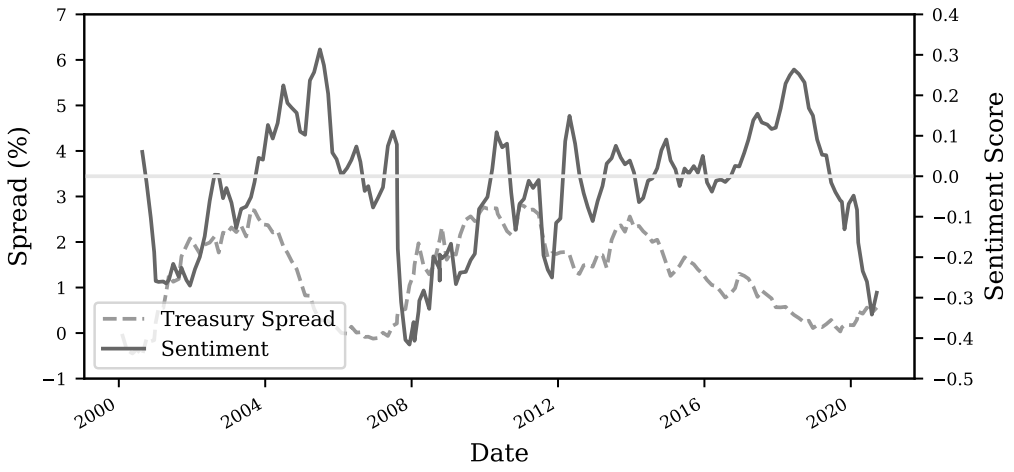

Figure 2: Treasury Spread on the left axis and the smoothed sentiment score calculated using the Deep Learning model on the right axis.

Figure 3 shows the relationship between the Taylor Rule (with Okun)

and the same smoothed sentiment score as before. Visually we can clearly see that there is a close relationship between these two time series and they appear to be correlated (albeit in a complex, non-linear manner).

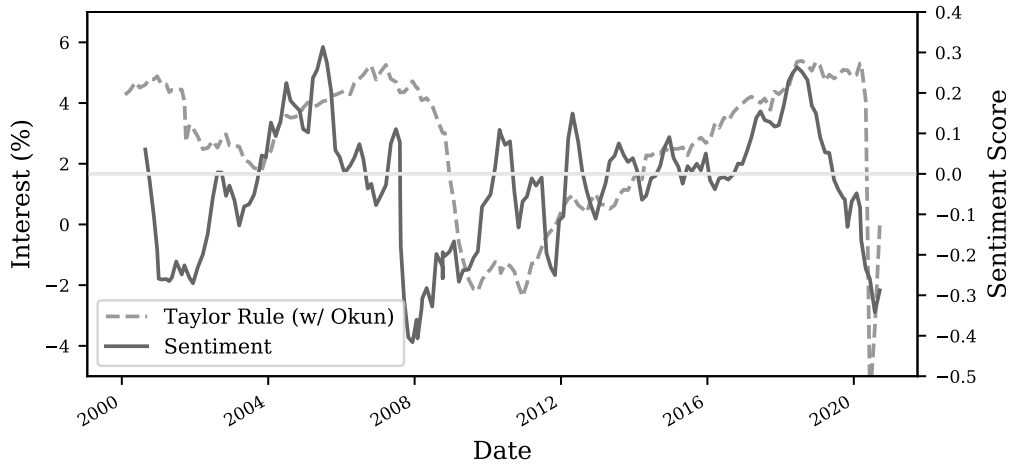

Figure 3: Taylor Rule (with Okun) on the left axis and the smoothed sentiment score calculated using the Deep Learning model on the right axis.

7.5.2 Named Events

Finally, we return to our sentiment models and consider how they treat named time intervals (Figure 4). Broadly speaking we see similar trends between our models, but with some differences. For example, during Yellen's tenure, we note that the dictionary approach consistently assigns a negative sentiment to meetings, whereas the two machine learning approachs are tending towards a positive sentiment over time. We suspect this behaviour arises from the fact that our dictionary approach performs poorly when predicting positive sentiment (recall Table 8). This negative sentiment tendency is also seen for all other intervals, to a varying extent.

8 Conclusions

Sentiment modelling is clearly useful when working with the complex and nuanced language of central bankers. Although dictionary-based approaches

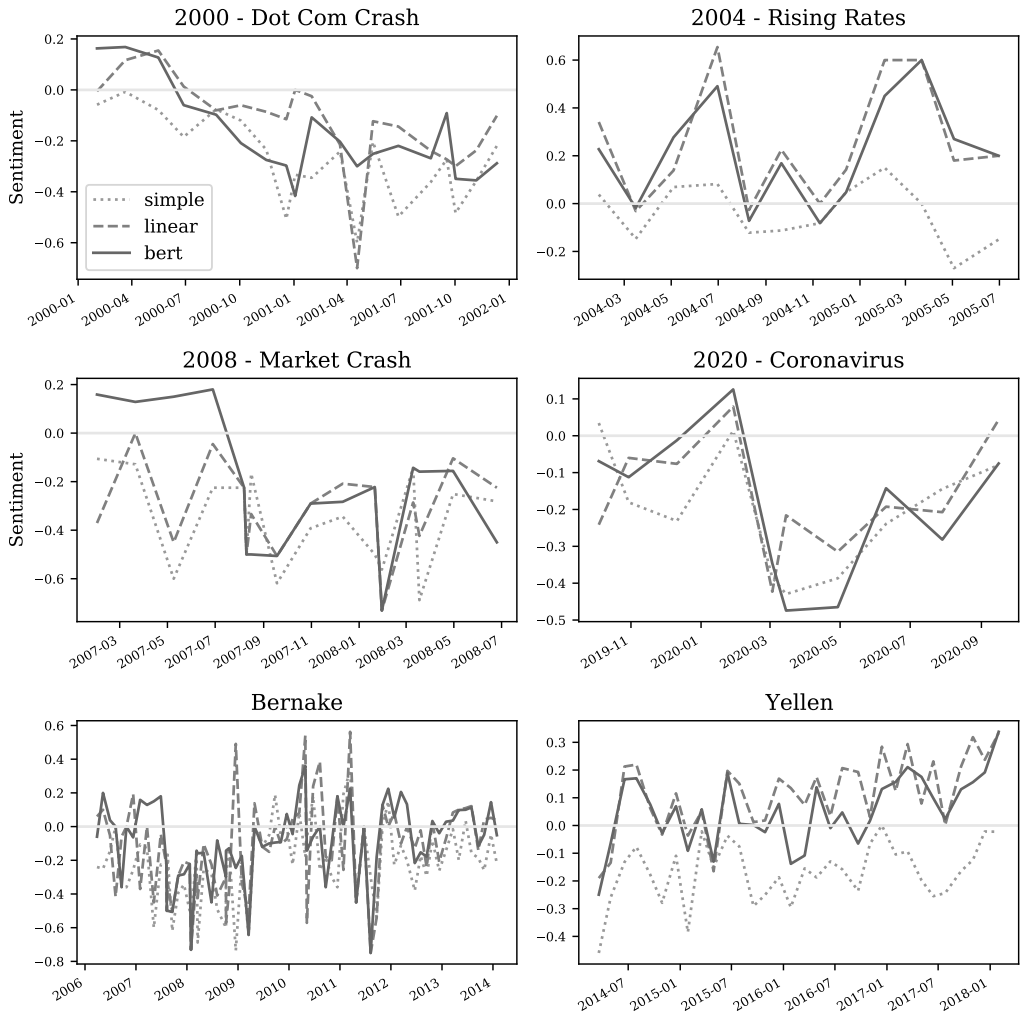

Figure 4: The sentiment timeseries calculated by the three different models for specific time intervals of interest.

are appealing in their simplicity they are limited and, arguably, some of the delicate sentiment information present within minutes is lost. More modern deep learning approaches—although initially more daunting—result in simpler systems that perform well. This simplicity stems from the fact that humans need to make fewer design decisions, with complexity moved over to the machine. This being said, there is still a place for simpler methods

and they can act as a complement. For example, our monetary policy experiments suggest that when aiming for the best explanation of decisions, it is best to use all available policy rules and sentiment models.

We primarily worked with FOMC meetings and it is interesting to consider whether our approach will work with other documents and in particular, with communications from other central banks (for example the European Central Bank).

References

[1] Andres Algaba, David Ardia, Keven Bluteau, Samuel Borms, and Kris Boudt. Econometrics meets sentiment: An overview of methodology and applications. *Journal of Economic Surveys*, 34(3):512–547, 2020. doi: https://doi.org/10.1111/joes.12370. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/joes.12370.

[2] George-Marios Angeletos and Jennifer La'O. Sentiments. *Econometrica*, 81(2):739–779, 2013.

[3] Mikael Apel and Marianna Blix Grimaldi. The Information Content of Central Bank Minutes. *Riksbank Research Paper Series*, (261), 2012.

[4] Paweł Baranowski, Hamza Bennani, and Wirginia Doryń. Do the ECB's Introductory Statements Help Predict Monetary Policy? Evidence from a Tone Analysis. *European Journal of Political Economy*, 10 2020. doi: 10.1016/j.ejpoleco.2020.101964.

[5] Alan S Blinder, Michael Ehrmann, Marcel Fratzscher, Jakob De Haan, and David-Jan Jansen. Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence. *Journal of Economic Literature*, 46(4):910–45, 2008.

[6] San Cannon. Sentiment of the FOMC: Unscripted. *Economic Review-Federal Reserve Bank of Kansas City*, 5, 2015. URL "https://www.kansascityfed.org/~/media/files/publicat/econrev/econrevarchive/2015/4q15cannon.pdf".

[7] Jean Carletta. Assessing Agreement on Classification Tasks: The Kappa Statistic. *arXiv preprint cmp-lg/9602004*, 1996.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Thomas G Dietterich. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural computation*, 10 (7):1895–1923, 1998.

[10] Stephen Hansen and Michael McMahon. Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communication. *Journal of International Economics*, 99:S114–S133, 2016.

[11] Bernd Hayo and Matthias Neuenkirch. Do Federal Reserve Communications Help Predict Federal Funds Target Rate Decisions? *Journal of Macroeconomics*, 32(4):1014–1024, 2010.

[12] Friedrich Heinemann and Katrin Ullrich. Does it Pay to Watch Central Bankers' Lips? The Information Content of ECB Wording. *Swiss Journal of Economics and Statistics*, 143:155–185, 04 2007. doi: 10.1007/BF03399237. URL http://ftp.zew.de/pub/zew-docs/dp/dp0570.pdf.

[13] Paul Hubert and Labondance Fabien. Central Bank Sentiment and Policy Expectations. Technical report, Bank of England Working Paper, 2017. URL https://hal.archives-ouvertes.fr/hal-01374710/document.

[14] Tim Loughran and Bill McDonald. When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66 (1):35–65, 2011.

[15] Athanasios Orphanides and John C. Williams. Monetary Policy with Imperfect Knowledge. *Journal of the European Economic Association*, 4(2-3):366–375, 05 2006. ISSN 1542-4766. doi: 10.1162/jeea.2006.4.2-3.366. URL https://doi.org/10.1162/jeea.2006.4.2-3.366.

[16] Matthieu Picault and Thomas Renault. Words are Not all Created Equal: A New Measure of ECB Communication. *Journal of International Money and Finance*, 79:136 – 156, 2017. ISSN 0261-5606. doi: https://doi.org/10.1016/j.jimonfin.2017.09.005. URL http://www.sciencedirect.com/science/article/pii/S0261560617301808.

[17] Maik Schmeling and Christian Wagner. Does Central Bank Tone Move Asset Prices? *Available at SSRN 2629978*, 2019. URL "https://openaccess.city.ac.uk/id/eprint/16865/1/".

[18] Jan-Egbert Sturm and Jakob De Haan. Does Central Bank Communication Really Lead to Better Forecasts of Policy Decisions? New Evidence Based on a Taylor Rule Model for the ECB. *Review of World Economics*, 147(1):41–58, 2011.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.