Predictive Fedspeak: Forecasting US equities with FOMC Minutes tone

Bachelor Thesis

Submitted to

the Chair of Sustainable Asset Management – Prof Dr Alexander Hillert

at the Faculty of Economics and Business Administration

of Goethe Universität Frankfurt am Main

in partial fulfillment of the requirements for the degree of

Bachelor of Science in Economics and Business Administration

by

Markus Brobeil

Matriculation number: [redacted]

[redacted]

July 07, 2021

Abstract

Insight into the effect of policymaker sentiment onto financial markets is valuable to both investors and regulators and has therefore seen an increasing amount of attention in recent years. This thesis aims to amend this literature by examining the impact of FOMC tone onto US equity markets. To this end, I (a) examine the same-day impact of FOMC Minutes tone onto the S&P500 and (b) test its predictive power for returns in the same instrument. I find that a simple measure of overall document tone is associated with changing volatilities on the day of the release, but does not significantly aid prediction. An extended tone metric, which links tone to specific discussion topics computed using Non-Negative Matrix Factorization, is in some cases, however, able to predict equity returns for up to a year after the Minutes' release.

Keywords

Federal Open Market Committee, Non-Negative Matrix Factorization, Return Predictability

Acknowledgements

I would like to thank my advisor Professor Alexander Hillert for all the support and insight provided throughout the process of writing this thesis. His reputation for being highly committed to every one of his advisees precedes him. Simon Heß, PhD provided insight on a few technical details regarding the implementation of my analyses in Stata, for which I'd also like to extend my gratitude. During the writing process, [redacted], [redacted], and [redacted] provided invaluable feedback, without which the following pages would certainly lack a lot of clarity and conciseness. Finally, this thesis would not have been possible without the unwavering support of my friends and family, who I am deeply grateful for.

Contents

List of Tables

List of Figures

List of symbols

V	Intra-day volatility of the S&P500
R	Intra-day return of the S&P500
SPY	S&P500 ETF with ticker symbol SPY
Λ	Naive tone, ie document tone without content topic weighting
λ	Topic tone, ie document tone by content topic
t	Time index for Minutes documents
c	Tone emotion, either net, positive, negative or uncertain
p	Numerical index for paragraphs within Minutes documents
ν	Sentiment word count
ω	Paragraph word count
Ω	Overall word count in Minutes document
θ	Model-generated weight of topic n in paragraph p
∂	Partial derivative
$\sum$	Summation sign
r	Return
μ	Arithmetic mean
σ	Standard deviation

List of abbreviations

FED	Board of Governors of the Federal Reserve System of the United States of America
FOMC	Federal Open Market Committee
GFC	Global Financial Crisis
LM dictionary	Loughran & McDonald Sentiment Word List
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
NMF	Non-Negative Matrix Factorization
TF-IDF	Term Frequency - Inverse Document Frequency

1 Introduction

Views expressed and policy decisions made by the Board of Governors of the US Federal Reserve System draw an exceeding amount of attention both from inside and outside the United States. As part of its communication to markets and society at large, the committee employs a rich qualitative language to report on issues such as inflation, employment, and projections for economic conditions. Yet, although there are a large number of papers that make use of quantitative information like interest rate decisions[1], little attention has been paid in academic research to the soft information transmitted alongside them.[2] This thesis therefore seeks to add to the burgeoning literature in this realm by analyzing the impact of such qualitative information onto US equity markets, as well as by examining the predictive power of the tone measures constructed for future returns. To this end, I make use of the Meeting Minutes of the Federal Open Market Committee, which report on discussions in the meetings of the policy-making body of the Federal Reserve.

My first main result is that using a purpose-built tone measure, the overall sentiment of a Minutes document is neither indicative of equity returns on the day of the release, nor for a large number of days afterwards. However, volatility on the day of the event depends significantly on some of the tone measures, a result that still holds when I control for macroeconomic trends. Motivated by my finding of increased activity on the day of the release conditional on certain emotional utterances, I implement a machine learning method to extract content topics from the textual data. And indeed, when I associate tone with the prevalence of different topics and compute my topic metric thereon, I find highly significant predictability of equity returns over a large number of days into the future for a handful of topic-tone combinations.

To the best of my knowledge, no paper has been published yet that uses the specific method known as ”Non-Negative Matrix Factorization” as a topic modelling approach on the FOMC Meeting Minutes. Further, although a number of publications study the market reaction to policymaker tone, there is no paper that extends the contemporaneous impact analysis to a prediction framework that can be applied to any number of financial variables.[3]

[1] Some of the most prominent papers that examine the impact of monetary policy on equity markets are Bernanke and Kuttner, 2005, Kuttner, 2001, Jensen et al., 1996, Bjørnland and Leitemo, 2009.

[2] Some of the most important articles on sentiment analysis and its relationship with financial markets have been authored by Tetlock, 2007, Antweiler and Frank, 2004, Hanley and Hoberg, 2010, and, specifically concerning the impact of FOMC tone, by Jegadeesh and Wu, 2017.

[3] Boukus and Rosenberg, 2006 study the correlation of five distinct themes as identified by Latent Semantic Analysis with current and future macroeconomic conditions, but not concern-

This thesis is structured as follows. Section 2 presents the choice of textual and financial time series data. Next, section 3 computes a naive tone metric that is then used to analyze the effect on and predictive potential for US equity markets. Section 4 then extends this metric by information on the prevalence of content topics in the documents and revisits the analyses from the previous section. Finally, section 5 concludes.

2 Data

2.1 Data sources

The data set used for further empirical analyses combines time series from three separate sources:

1. Textual data of the FOMC Meeting Minutes is scraped from the website of the Federal Reserve (FED), parsed, and filtered on a word-by-word basis. The outputs of this pipeline are further used to compute metrics capturing the overall tone as well as topic-specific sentiment scores for every Minutes document.
2. Time series of the S&P500 open, high, low, and close captured on a daily basis by the ETF *SPDR S&P 500 ETF Trust* with ticker symbol *SPY*, sourced from Yahoo Finance using the Python library *yfinance*. This ETF differs from the regular S&P500 in that its value is only one tenth of the original vehicle.
3. Time series of several macroeconomic variables employed primarily as regression controls from the publicly accessible database *FRED* of the Federal Reserve Bank of St. Louis.

The following passage discusses in detail why I select the Minutes as a textual basis to analyze policymaker sentiment.

2.2 Textual data: FOMC Meeting Minutes

Although the Meeting Minutes are arguably the most common document employed by the Fed watching literature to gauge policymaker mood (eg Rosa, 2013, Cieslak and Schrimpf, 2019), there are a number of other publications that would at first glance lend themselves to this end. This section clarifies my

ing financial time series.

choice of the Minutes as a textual basis for tonal analyses of policymaker sentiment. The information presented draws loosely on Danker and Luecke, 2005 and the website of the FED[4].

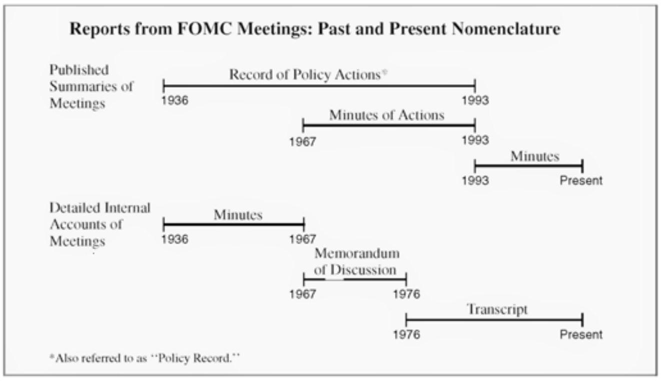
Reports from FOMC Meetings: Past and Present Nomenclature
Published Summaries of Meetings
Record of Policy Actions*
1936
1993
Minutes of Actions
1967
1993
Minutes
1993
Present
Detailed Internal Accounts of Meetings
Minutes
1936
1967
Memorandum of Discussion
1967
1976
Transcript
1976
Present
*Also referred to as "Policy Record."

Figure 1: Overview over the different types of FOMC publications and the respective period of years in which they were published. Although in theory, the second strand of documents would, due to its greater level of detail, be better suited for measuring policymaker tone, the significant lags of several years between committee meetings and publication render them unsuitable for my purposes. The first strand on the contrary offers a good compromise between detail and timeliness. The availability of relevant market data finally induces the decision to select the Minutes since 1993 as textual data. Copied from Danker and Luecke, 2005.

Beginning with the inception of its modern form through the Banking Act of 1935, the FOMC started publishing the ”Record of Policy Actions” that was primarily designed as a declarative record of policy actions, but over time evolved to also include detailed information on the background of the decisions recorded. In 1967, this document was amended by the ”Minutes of Actions”, which served the additional purpose of documenting administrative and organizational details of meetings. Among fundamental shifts in the policy making patterns of the FOMC in 1993 (Pakko, 1995), both publications were merged into the modern ”Minutes” that have been published in the same form since. Because of their timely release schedule relative to other publications, all three of these documents would qualify as a basis for a market impact study. As trading volume in most securities and indices was comparatively low before the turn of the century, focusing on the most recent type of document seems most promising and has the additional benefit of not having to account for differing text structures in the data preparation and analysis stages.

A second strand of documents that report directly on FOMC meetings would by its level of detail be close to ideal for sentiment analyses. However, because the pertaining documents contain large amounts of unedited information, they have historically been kept from the public for several years to not impede implementation of policy actions and are thus not well-suited for a market impact

[4] https://www.federalreserve.gov/monetarypolicy/fomc_historical.htm

study. The earliest Minutes from the first half of the twentieth century, for example, were originally kept as archival records and only released with a 5-year lag in the wake of the ”Freedom of Information Act” in 1967. Similarly, near verbatim transcripts of audio recordings have been published with a 5-year time lag, only starting in 1993 after political pressure on the then FED chair Alan Greenspan (Hansen et al., 2018).

Finally, the ”Statements” - arguably the most prominent publication by the FOMC - are released shortly after the meeting takes place and give the public a brief summary of meeting contents and resulting policy actions. Although the near-immediate release would rule out any indirect information diffusion and thus give results measuring the market impact a greater significance, their brevity poses significant problems to general sentiment analyses and topic modelling techniques. Because both of these methods make use of the composition of each individual text to estimate underlying characteristics of author tone, a lower number of words will increase estimation variance and thereby reduce the reliability of empirical findings. Since the Minutes offer a much more elaborate basis and thereby more granular insights into the proceedings of the meeting, I select them over the much shorter Statements as basis for tonal analyses in the following.

In summary, the Minutes published since 1993 offer the ideal tradeoff between the level of detail and a release date close to the original committee meeting and are thus chosen as a textual basis for the following analyses.

2.3 Text processing

Prior to analyzing the tonal quality and thematical composition of the Minutes, I subject them to preprocessing procedures that are ubiquitous in many applications of computational linguistics. After downloading the raw bodies of text from the FED’s website starting with the meeting on February 2-3, 1993, the text corpora are split into paragraphs based on their HTML structure, with sections enclosed in HTML paragraph tags (”<p>Text</p>”) usually containing one passage.

Next, by tokenizing all words in a paragraph and treating the resulting cluster as a so-called ”bag of words”, I dispose of any information conveyed by the arrangement and order of words within sentences and the structure of sentences within a paragraph. The information content lost this way is surprisingly low relative to overall information (Jurafsky and Martin, 2020). Specifically, every paragraph is run through an off-the-shelf, pre-trained word-classification pipeline provided by the open-source NLP framework *spacy* (Honnibal et al., 2020). Besides tokenization, it also provides an estimate of grammatical function and word class, which allow for the removal of stopwords as well as targeted selection of the most meaningful word classes from the remaining pool, specifically nouns, verbs, and adjectives. When paragraphs are treated as a collection of words with no particular syntactical order or semantical structure, other word classes generally carry little to no meaning and can thus be safely discarded.

A second modification frequently employed in this context is the removal of linguistic flections which carry no relevance if word order is neglected and would in fact impede the efficiency of many advanced algorithms. For example, the topic modelling technique employed in section 4 groups similar words together by discussion topic to produce term clusters. If the algorithm can take different words that convey the same underlying notion as one and the same entity, it can perform its task with much greater efficiency. This technique known as ”lemmatization” is closely related to, but different from another widely-used preprocessing technique known as ”stemming”. While the prior attempts to identify the underlying word root based on syntactic and semantic context, the latter applies a more or less complex set of rules to a single word while disregarding clues in its vicinity about the underlying lemma. This second approach offers the advantage of being fully transparent and generally also requires less computational resources. If however accuracy of results is preferred, the black-box method of lemmatization yields better results. Because the latter applies for this thesis, I use Spacy’s built-in lemmatizer along with the tokenization feature discussed above.[5]

Table 1 provides summary statistics of word and paragraph count on the document level. In total, I process 225 documents over a span of 28.25 years, which amounts to approximately 8 meetings per year. The removal of stopwords shrinks the total number of words per document by about 56%.

Figure 2 shows the progression of both the number of words and paragraphs per document over time. The regular spikes of document length can be traced to an additional section covering the authorization of domestic open market and foreign currency operations, which is generally prepended in the second meeting of the year. Because attempts to programmatically remove this section across documents prove unreliable, I decide to use the entire body of text for this analysis. Considering that these passages are mostly concerned with legal minutiae and therefore are unlikely to contain emotional language of significant amplitude, chances are small that a bias is introduced this way.

[5] A great overview and history of lemmatization can be found here: https://devopedia.org/lemmatization

	Mean	Std	Min	Max
Word count unprocessed	6584.28	2081.46	2962	12590
Word count filtered	2870.12	915.31	1390	5703
Paragraph count	44.29	15.40	16	96
Avg filtered word count per paragraph	65.66	7.93	47.73	126.31

Table 1: Summary statistics of word and paragraph count before and after processing. Overall, 227 documents are processed over a span of 28.25 years. Processing removes about half of the words per document. The fact that the minimum and maximum before and after processing lie at around 1.5 and 3 standard deviations away from the mean respectively indicates that the changes are applied in a relatively uniform fashion.

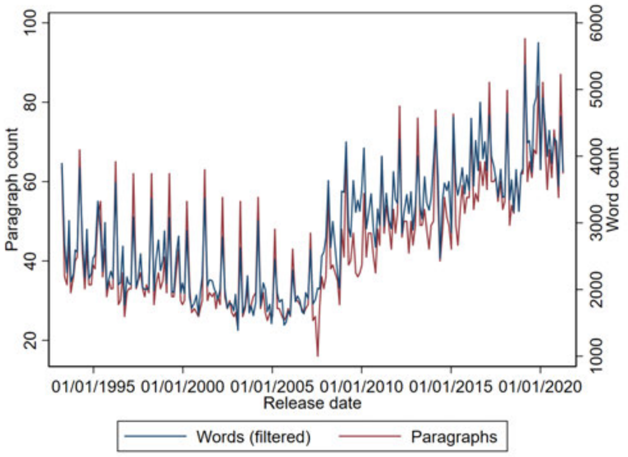

Figure 2: Word and paragraph count over time. The word count captures the number of words in the FOMC Meeting Minutes after applying my custom-built filtering pipeline that aims to remove words with little meaning, when word order is disregarded. The periodical spikes are due to a legal section that is appended only once a year and is by its nature unlikely to introduce any bias to the tone measures.

2.4 Financial variables

Measurement of the same-day market impact relies on two variables, calculated from the daily time series of the S&P500. Returns are computed as the difference between the open and closing price of the S&P500 on day t, relative to the open price on the same day.

$$R_t = \frac{SPY_t^{close} - SPY_t^{open}}{SPY_t^{open}} \tag{1}$$

To construct a volatility measure, I deviate from the convention of squaring intraday or 24-hour returns. Instead, I make use of daily range data by taking the difference between a day's highest and lowest trading price. To obtain a relative measure that is qualitatively comparable to the return definitions above, I also scale the result by the entire-period mean of the intra-day range.

$$V_t = \frac{SPY_t^{high} - SPY_t^{low}}{\frac{1}{N}\sum_{n=1}^{N} SPY_n^{high} - SPY_n^{low}} \tag{2}$$

This definition is superior to traditional squared-returns volatilities as it is more closely aligned with most people's intuitive understanding of financial vicissitude and, more importantly, refrains from applying a non-linear penalty function to deviations from the mean. Despite its wide-spread usage, the second-order moment is in fact not well-suited for descriptive purposes, as its quantitative value, due to its non-linear construction, bears little to no practical meaning. My measure on the contrary can be easily interpreted as the ratio of day t's trading range relative to that of the average day.[6]

Considering the existing literature that measures the impact of central bank policy announcements, I clarify my choice of variable time frame in appendix A.

3 Naive Tone

As mentioned in the introduction, I first explore the association of overall document tone with financial markets and establish a topic-tone link in the following section.

[6] This is not to suggest that the second moment is per se an inferior volatility measure, but simply that in certain situations, other metrics are better suited for the task at hand.

3.1 General observations

To quantify author tone in academic research of all realms, the most ubiquitous technique in the literature is to simply count the number of words belonging to certain emotional dimensions as specified by sentiment dictionaries like the Harvard IV-4 Psychosociological Dictionary. Recent research by Loughran and McDonald, 2011, however, has shown that in financial contexts, the fraction of misclassified words can be as high as 75% when general-purpose word lists are employed, thus necessitating a separate dictionary for financial research. The classification record they suggest as part of the paper has been widely used in the financial research literature over the last decade, which is why I choose to also employ it for this thesis.

To get a first overview over the data, I thus proceed to count the number of word occurrences in the sentiment dimensions ”positive”, ”negative”, and ”uncertain”, as defined by the Loughran & McDonald (LM) sentiment word dictionary,[7] per document. Examining these raw word counts already reveals surprisingly much about the changing language of the committee in regards to macroeconomic events. Figure 3 plots the absolute word counts of the three sentiment dimensions over time.

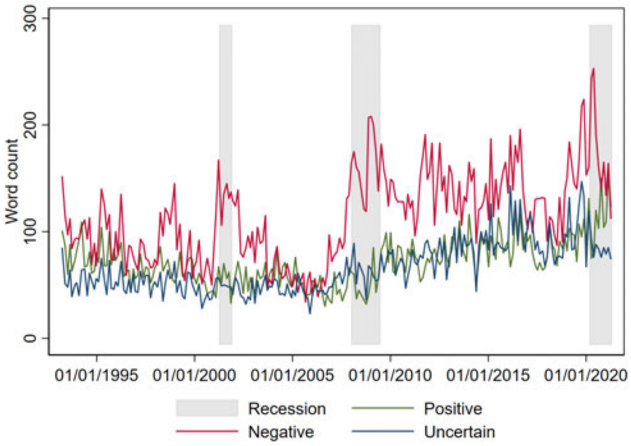

Figure 3: Progression of absolute word counts of the three emotion dimensions. Words are counted as defined by the most recent Loughran&McDonald dictionary, recessions are displayed as indicated by the NBER index. The key takeaway from this graph is that a simple count of emotion words already shows significant patterns relating to economic crises.

While the number of positive and uncertain words closely follows the pattern of the overall word count in figure 2, negativity seems much more sensitive to economic and financial crises, with the number of words spiking sharply during every recession relative to the respective pre-crisis period. The positive word count seems to be largely unaffected by periods of crises and primarily driven by

[7] The most recent word list is available at https://sraf.nd.edu/textual-analysis/resources/

document length. As non-stationarity in the overall number of words impedes an intertemporal comparison of the ”emotion amplitude”, I divide the sentiment word count per Minutes release by the overall number of words in the respective document to obtain the fraction of sentiment words, displayed in figure 4.

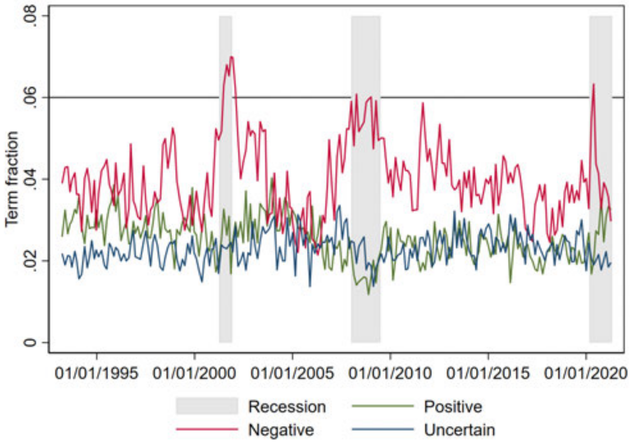

Figure 4: Progression of word counts of the three emotion dimensions relative to the overall number of words in the document. Words are counted as defined by the most recent Loughran&McDonald dictionary, recessions are displayed as indicated by the NBER index. This graph shows that emotional patterns become even more distinct when controlling for non-stationarity in the underlying document length.

The progression of these relative measures reveals that the three major crises in the past two decades did in fact cause the highest (relative) spikes in the usage of negative words. The 6%-level seems to be the relevant threshold that this measure clears in every crisis. Interestingly, pessimism in the Minutes' tone seems to have spiked again close to this particular level only three to four years after the Global Financial Crisis (GFC) before settling down to non-crisis levels until the global pandemic struck in 2020. Although positivity and negativity appear to be mostly white noise, the measure for positive sentiment will prove to be the most valuable in the following.

3.2 Tone computation

For further analyses, I collect the sentiment word count analyzed above on the level of individual paragraphs and weigh the result by the proportion of the overall document body which that respective passage constitutes. Specifically, the number of terms ν of the three select LM sentiment topics $c \in \{positive, negative, uncertain\}$ in paragraph p at time t is multiplied with the proportion of words ω that the same paragraph comprises of total document length Ω. This score is summed over all paragraphs in a Minutes document, such that the resulting metric Λ measures the average score of sentiment type c at time

t over all paragraphs P, weighted by the length of the respective paragraph relative to overall document length:

$$\Lambda_{c,t} = \sum_{p=1}^{P} \nu_{c,p,t} \frac{\omega_{p,t}}{\Omega_t} \tag{3}$$

t - Minutes document

p - paragraph index in document t

c - sentiment topic (c for content)

ν_c - number of words of topic c

$\omega_{p,t}$ - number of words in paragraph p

Ω_t - number of words in document t

In addition to the three basic sentiment topics *positive, negative, uncertain*, I construct a fourth emotion *net tone* as the difference between *positive* and *negative* sentiment. For this measure specifically, $\nu_{net} = \nu_{positive} - \nu_{negative}$. Figure 13 in appendix F shows the progression of $\Lambda_{c,t}$ for all four emotion indices constructed.

The tone computation setup overall was initially inspired by Jegadeesh and Wu, 2017, who also employ a topic modelling scheme to separate talking points as I do in section 4, but only to examine the contemporaneous market impact of their tone measures. Although certain architectural decisions are still the same[8], I have made two significant modifications, which I elaborate on in appendix B.

3.3 Evaluation

To examine the impact of Minutes tone onto US equity markets, I present regression results in three stages. First, a ”placebo test”[9] determines whether equities exhibit non-random behavior on the daily time frame solely because of the Minutes release. Second, I examine whether the amplitudes of my four tone measures are related to market reactions. Finally, before turning to more involved methods for tone computation in the next section, I test the predictive power of tone sentiment measures over a large number of days.

Prior to utilizing tone as a regressor, it is advisable to assess whether the day of the publication itself has an impact on the variables of interest. Specifically in the case of intra-day returns of equities proxied by the S&P500, Lucca and Moench, 2015 provide evidence for a systematic upwards drift prior to

[8] The most significant hereof is the computation of tone scores on the paragraph level and subsequent aggregation to the document level.

[9] ”Placebo test” is in quotation marks here, also because the term is borrowed from medical research, but primarily because this analogy only holds for directional price changes. Particularly, if volatilities were to systematically increase on release days, this result would imply a high information content of publications. Market participants would rationally drive volatility by pricing in the new information instead of behaving irrationally as is implied by the word ”placebo”.

scheduled FOMC Meetings, which, due to the temporal offset, cannot be explained by the content of subsequent policy announcements itself.

	V	R
Release day dummy	-0.070	-0.000
	(-1.32)	(-0.64)
Constant	1.002***	0.000
	(78.71)	(0.45)
Observations	7112	7112
adj. R^2	0.000	0.000
N(days) on which dummy is 1	225	225

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Market reaction to the release of FOMC Meeting Minutes. The table reports regressions of each of the column variables onto the ”Release day dummy” that takes values 1 if a Minutes document is released on that day and 0 if not. Volatility is measured as the difference between the day’s high and low in the S&P500, scaled by the average daily high-low spread. R measures the distance between the day’s close and open price of the S&P500 relative to the day’s open price. The purpose of the table is to show that there is no event day effect, neither in the volatility, nor in the price change of equities.

Table 2 however shows no evidence for systematically higher volatility and returns on the day of the release.

Next, to establish whether the tonal scores have an impact on investor behavior, I regress the volatility of the release day onto the four tone scores *net, positive, negative*, and *uncertain* (2). To avoid multicollinearity, *net* tone is treated as an alternative specification for *positive* and *negative* tone and run as a separate model. Additionally, for both variations, I also present a specification that controls for a few important macroeconomic trends, specifically the Federal Funds Rate, the US unemployment rate, and the NBER recession indicator. Finally, I also include the variable *dissent* as a regressor, which indicates whether one or more of the committee members voted against the final set of policy actions at the end of the meeting. As Madeira and Madeira, 2019 show this dummy variable to be correlated with market reactions, it may help in producing a better explanatory model.

Table 3 attests both positive and net tone a significant impact on volatility on the day of the release, with the prior still holding significance after introducing macroeconomic control variables. Even for positive tone however, the coefficient’s value changes sufficiently much from the uncontrolled to the controlled model for the specification to be likely not free of omitted variable bias.

The negative value of the coefficient however makes intuitive sense, as a greater number of positive words is likely associated with more stable economic conditions and thereby makes erratic investor behavior in equity markets less

	V	V	V	V	R	R	R	R
naive Net Tone			-0.180*	-0.028			0.001	0.001
			(-2.50)	(-0.54)			(0.96)	(0.65)
naive Positive Tone	-0.413***	-0.275*			-0.000	-0.000		
	(-3.53)	(-2.18)			(-0.11)	(-0.30)		
naive Negative Tone	0.107	-0.071			-0.001	-0.001		
	(1.62)	(-1.11)			(-1.39)	(-1.14)		
naive Uncertain Tone	-0.173	-0.012	-0.314*	-0.193*	0.003	0.003	0.002	0.002
	(-1.44)	(-0.12)	(-2.52)	(-2.29)	(1.95)	(1.75)	(1.48)	(1.31)
Dissent Dummy	0.015	-0.009	0.024	-0.002	-0.001	-0.000	-0.001	-0.000
	(0.15)	(-0.11)	(0.24)	(-0.03)	(-0.63)	(-0.34)	(-0.59)	(-0.32)
Federal Funds Rate		-0.089*		-0.106**		-0.000		-0.000
		(-2.33)		(-3.10)		(-0.51)		(-0.78)
Unemployment Rate		-0.025		-0.052		0.000		0.000
		(-0.29)		(-0.65)		(0.51)		(0.29)
Recession		0.969**		0.893**		-0.004		-0.004
		(3.28)		(3.12)		(-1.13)		(-1.28)
Constant	1.810***	2.070***	1.338***	1.773***	-0.001	-0.002	-0.003	-0.003
	(5.71)	(5.06)	(5.44)	(3.58)	(-0.22)	(-0.39)	(-1.15)	(-0.68)
Observations	225	224	225	224	225	224	225	224
Adjusted R^2	0.085	0.247	0.059	0.215	0.009	0.009	0.007	0.009

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Market reaction to the naive tone scores. The first four columns report the coefficients of different specifications for regressions onto volatility as defined in equation 2. The right four columns in turn show the same model variations with the intra-day return of the S&P500 as the dependent variable, as defined in equation 1. Tone scores Λ_c are defined by equation 3. Control variables are provided by the FRED database and described in detail in table 6. The variable ”dissent” takes values 0 and 1 depending on whether the final meeting vote was unanimous or whether at least one committee member voted against the final action. t-statistics are based on White, 1980 standard errors. The estimates use 225 FOMC Minutes released between 1993 and 2021.

likely. This statement of course assumes full exogeneity of forecast errors, ie that a more positive tone is not associated with systematically larger errors in investor expectations about the Minutes' contents. If higher errors due to more positive tone were to occur, the mere adjustment of prices to incorporate new information reveleaved in the Minutes would increase volatility and thereby introduce a positive bias into the coefficient in question. Although there have been attempts to proxy control for this hypothetical problem in the pertaining literature, I decide to not present it as part of my main results. The interested reader, however, may refer to appendix C, which covers the exact mechanism of this potential issue and explains why this particular control may in my case even introduce more distortion than remove it.

The interpretation of the uncertainty coefficient is challenging, both because of its value and because of its inconsistent significance across specifications. Column 4 suggests that one additional uncertain word per paragraph[10] yields a 19.3% lower than usual trading range in the S&P500 on the day of the release. Intuition, however, would suggest that more uncertainty should be associated with a higher volatility, as investors are left to use their own imaginations if policymakers refrain from making unambiguous forecasts.

In summary, although the association of tone and volatility seems to be complicated, it is very likely that a link exists, meaning that policymaker tone does impact investor behavior. The lack of significant coefficients in the right four columns of table 3 in turn suggests that the overall document tone bears little relevance for directional changes in equity prices on the event day though. The combination of these two findings motivates section 4, in which the Minutes are scrutinized more closely by identifying a set of topics in every document. In contrast to the findings above, I demonstrate that an extended version of the tone measure is in some cases capable of predicting equity returns over a large number of days into the future.

Finally, to examine the predictive power of the tone measures, I regress the return of the S&P500 over k days starting with the day of the release onto every one of the four tone scores:

$$\frac{SPY_{t+k} - SPY_t}{SPY_t} = \alpha_{c,k} + \beta_{c,k}\Lambda_{c,t} + \varepsilon_{c,t,k} \tag{4}$$

Results are plotted as a line graph that describes the slope coefficient β of this regression as a function of the number of periods k that the difference is projected into the future.[11] Additionally, a 95% confidence interval is calculated using heteroskedasticity-robust standard errors (White, 1980) and plotted as a dashed line around the coefficient estimate. Graphs are produced for every one of the four sentiment scores and jointly presented in figure 5.

For the evaluation of these graphs, the position of the ”null line” relative to the confidence interval is decisive. If this level lies outside the interval bounds, standard statistical theory provides that the null hypothesis positing no statistically significant effect can be rejected. For the vast majority of predicted values

[10] $\frac{\partial\Lambda}{\partial\nu_p} = \sum_p^P \frac{\omega_p}{\Omega} = 1$. Thus, for every (positive) increment in $\nu_p \; \forall \; p$, Λ increments the same amount.

[11] Specifically, I predict returns in 10-day steps for up to $K = 250$ of the ensuing trading days, comprising roughly the time span of one year after the release. On the New York Stock Exchange for example, equities are currently traded on 252 to 253 days of the year, see https://www.nyse.com/publicdocs/Trading_Days.pdf

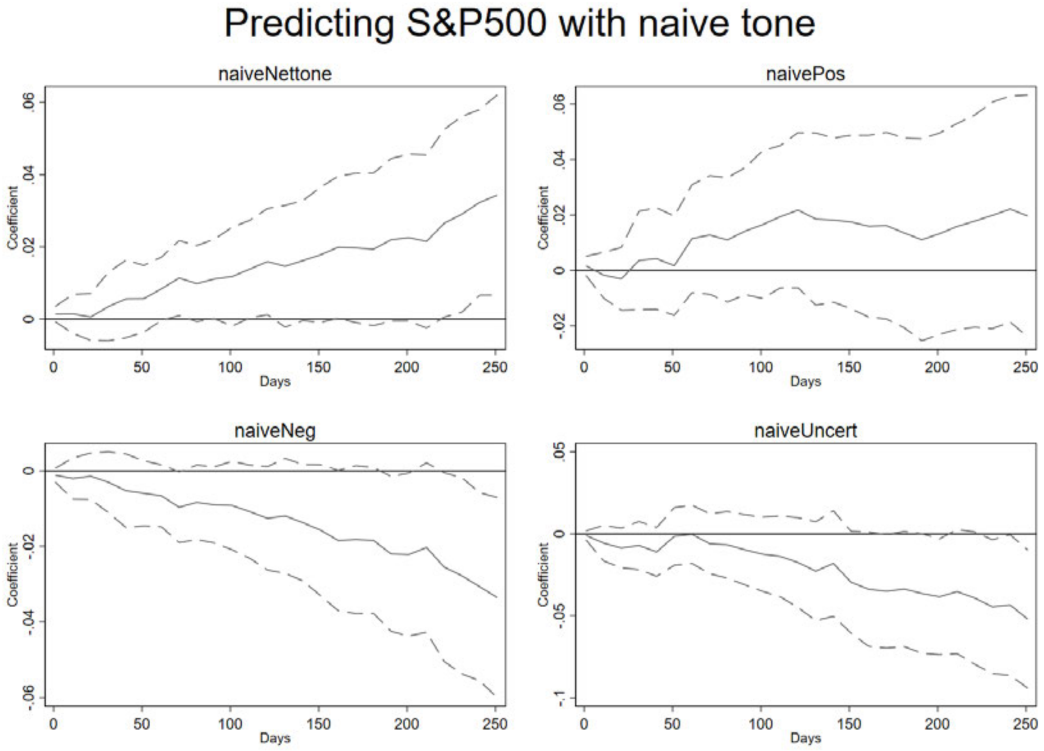

Figure 5: Predicting the S&P500 with naive tone. The purpose of these plots is to visualize when and by what magnitude predictability based on naive tone measures shows significance. The solid line in every graph represents the series of slope coefficients from regressing the change in the S&P500 over k days after the release (with k denoted on the x-axis) onto the respective naive tone metric Λ_c as specified in equation 4. With one year comprising roughly 250 trading days, the graphs cover about one full year of movement in the S&P500. Dashed lines report a 95% confidence interval based on White standard errors (White, 1980) for the corresponding coefficient value on the solid line.

in the four graphs above however, the horizontal line is situated well between the dashed lines, thus providing little ground for claims of significant overall predictability. I thus infer that the measures for overall tone provide little information that would aid in predicting future returns of the S&P500.

If one were to apply a less strict significance niveau, the general trend of the four graphs would, however, run well with an intuitive narrative: While a higher net and positive tone indicates favorable conditions, thereby raising the price of the S&P500, more negativity or uncertainty signals the opposite and thereby depresses prices.

In summary, although altered investor behavior conditional on policymaker tone can be observed via changing volatilities on the release day, I find little evidence for any behavior yielding strong directional price changes in equity markets. My naive tone scores thus do not lend themselves particularly well to predictive modelling.

4 Topic Tone

My results so far suggest that Minutes with an increased amplitude in select emotional dimensions induce higher volatility on release days, indicating that investors do react to tonal information. However, as I was unable to establish a strong directional price change in the S&P500 using the naive tone scores that measure overall document tone, the question arises whether this metric really captures the entire picture or whether a more granular approach might yield more insights. To this end, I employ what is known as a topic modelling technique that identifies the prevalence of six different topics in every Minutes document. Appendix D covers the theoretical background and parametrization of this method, alongside a brief examination of the model constructed.

The topics produced are labelled as follows:

1 - Economic activity
2 - Policy action
3 - Economic outlook
4 - Employment
5 - Financial markets
6 - Inflation

To assess the predictive power of these model-generated themes, I define a topic-specific tone score λ as an extension of naive tone Λ (equation 3). Specifically, the proportion-weighted term count $\nu_{c,p}$ is additionally multiplied with the estimated model-generated weight θ of topic n in paragraph p before summing over all paragraphs by document:

$$\lambda_{n,c,t} = \sum_{p=1}^{P} \theta_{n,p,t} \nu_{c,p,t} \frac{\omega_{p,t}}{\Omega_t} \tag{5}$$

I thereby obtain a topic-tone score that describes the strength of emotion c associated with topic n for every Minutes publication t over the entire sample period. As for overall document tone (equation 3), the composite emotion *net tone* is constructed using $\nu_{net} = \nu_{positive} - \nu_{negative}$.

4.1 Market reaction to topic tone

As for naive tone in section 3.2, I begin by examining market reactions to the tone measures on the day of the release. To this end, I regress both intra-day volatility (equation 2) and intra-day returns (equation 1) of the S&P500 on the release day onto all n topic measures, optionally controlling for macroeconomic trends:

$$V_{c,n,t} = \alpha_{c,n} + \sum_{n=1}^{6} \beta_n \lambda_{n,c,t} + \beta_7 int_t + \beta_8 unemp_t + \beta_9 rec_t + \varepsilon_t \tag{6}$$

$$R_{c,n,t} = \alpha_{c,n} + \sum_{n=1}^{6} \beta_n \lambda_{n,c,t} + \beta_7 int_t + \beta_8 unemp_t + \beta_9 rec_t + \varepsilon_t \tag{7}$$

As controls, int describes the Federal Funds Rate provided by the FED, $unemp$ captures the official unemployment rate published by the US Bureau of Labor Statistics, and rec yields the binary recession indicator for the US economy produced by the National Bureau of Economic Research.

	Net tone		Pos tone		Neg tone		Uncertain tone	
Topic - Economic Activity	1.227*	1.353	**-2.979*****	**-2.945***	**-1.406***	**-1.705***	-2.503*	-2.443
	(2.07)	(1.87)	(-3.58)	(-2.22)	(-2.44)	(-2.39)	(-2.16)	(-1.92)
Topic - Policy Action	0.773	0.866	-0.763	-0.623	-0.124	-0.803	-0.575	-0.433
	(0.98)	(1.33)	(-1.31)	(-1.05)	(-0.22)	(-1.52)	(-0.62)	(-0.53)
Topic - Economic Outlook	-1.521	-1.071	0.364	-0.535	0.577	0.473	1.011	0.273
	(-1.83)	(-1.40)	(0.56)	(-0.82)	(0.92)	(0.98)	(1.00)	(0.33)
Topic - Employment	0.378	0.284	0.365	1.555	-0.247	0.194	-1.125	-0.158
	(0.57)	(0.36)	(0.65)	(1.57)	(-0.61)	(0.53)	(-1.83)	(-0.30)
Topic - Financial Markets	-2.386	-1.863	0.358	-1.947	2.032*	1.101	3.552*	1.895*
	(-1.67)	(-1.76)	(0.48)	(-0.68)	(2.25)	(1.80)	(2.46)	(2.09)
Topic - Inflation	0.048	0.002	-0.850	-0.741	0.192	-0.047	0.136	0.090
	(0.15)	(0.00)	(-1.88)	(-1.49)	(0.77)	(-0.14)	(0.37)	(0.22)
Control - Federal Funds Rate		-0.104***		-0.023		-0.059		-0.060
		(-3.38)		(-0.42)		(-1.35)		(-1.30)
Control - Unemployment		-0.068		0.017		-0.056		-0.047
		(-0.87)		(0.13)		(-0.61)		(-0.53)
Control - Recession		0.921**		0.979**		1.062**		0.933***
		(2.89)		(3.29)		(3.18)		(3.40)
Constant	0.659***	1.391**	1.851***	1.542**	0.818***	1.542**	1.272***	1.542*
	(8.13)	(2.63)	(5.50)	(2.82)	(5.81)	(2.60)	(5.24)	(2.51)
N	225	224	225	224	225	224	225	224
adj. R^2	0.079	0.239	0.146	0.277	0.125	0.270	0.116	0.252

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: NMF topic impact on intra-day volatility in the S&P500. Every column pair reports the coefficients of regressing intra-day volatility onto the 6 topics generated by NMF without and with the 3 control variables capturing macroeconomic trends, as defined in 6. In this sense, columns represent the different values of emotion c, whereas the first 6 rows represent the topics n. Tone scores $\lambda_{c,n}$ are defined by equation 5. Control variables are provided by the FRED database and described in detail in table 6. t-statistics are based on White, 1980 standard errors. The estimates use 225 FOMC Minutes released between 1993 and 2021.

Table 4 displays the results of the volatility regression (6). Every column pair reports the coefficients β_i for a different tone measure λ_c, while the first 6 rows list the topics n generated by the topic model.

Examining the coefficients, the most striking message seems to be the high relevance of *Economic Activity* for volatility in the S&P500, with every tone score producing a significant coefficient on either the controlled or uncontrolled specification. Particularly positive and negative tone seem to be of high relevance, with the coefficient for the association of both of these emotions with *Economic Activity* being significant in both the uncontrolled and controlled version. Moreover, because the values are even relatively unaffected by the addition of the control vector, indications of high validity for this particular result are strong.

Value-wise, the coefficient suggests that a more negative tone on the current state of the economy is associated with lower volatility on the event day. Specifically, if one assumes a uniform proportion of $\frac{1}{6} = 16.\overline{6}$% of the topic in every paragraph across documents, one additional positive (negative) word in every paragraph would be linked to a 50% (28%) lower volatility on the day of the release as compared to the average day.[12] Since a more negative tone is generally associated with times of crises as per figures 3 and 4 however, one would expect the coefficient for the latter to be positive, as higher levels of pessimism would be grounds for increased trading activity. Refraining from speculations about the source of the negative sign on the coefficient at this point, I conclude that some topic-tone combinations appear to be highly relevant for same-day equity volatilities, although the exact mechanisms remain obscure.

Next, table 5 displays coefficients for regressing intra-day returns onto the set of topic-tone measures (equation 7), structured the same way as the previous table. Results are similar to the naive tone regressions in table 3, with almost none of the coefficients showing statistical significance. The only exception is *net tone* regarding plans for future policy, which on average induces a 0.28 percentage points higher return on the day of the release if one positive word was added to every paragraph (assuming again a uniform content topic proportion of one sixth)[13]. Interpretation again is ambiguous, as the mechanism relating policy tone with current and future economic conditions is unclear. When discussing Quantitative Easing programs for example, it is possible that

[12] Assuming the proportion of topic n fixed across paragraphs and documents, θ can be extracted from the summation term in equation 5, yielding

$$\frac{\partial \beta_{c,n}\lambda_{n,c,t}}{\partial \nu_{c,p,t}} = \beta_n \times \theta_{n,t} \times \underbrace{\sum_{p}^{P} \frac{\omega_{p,t}}{\Omega_t}}_{=1} = \beta_{pos/neg} \times 0.1\overline{6} \times 1 \approx -50\% \text{ and } -28\%$$

[13] $\frac{\partial \beta_{c,n}\lambda_{n,c,t}}{\partial \nu_{c,p,t}} = \beta_n \times \theta_{n,t} \times \sum_p^P \frac{\omega_{p,t}}{\Omega_t} = 0.017 \times 0.1\overline{6} \times 1 = 0.28\overline{3}$%

	Net tone		Pos tone		Neg tone		Uncertain tone	
Topic - Economic Activity	0.009	0.003	0.004	0.003	-0.005	-0.003	-0.004	-0.002
	(1.30)	(0.33)	(0.32)	(0.23)	(-0.98)	(-0.44)	(-0.32)	(-0.13)
Topic - Policy Action	**0.018***	**0.017***	0.009	0.008	-0.003	-0.002	0.006	0.002
	(2.18)	(1.99)	(1.04)	(0.97)	(-0.43)	(-0.35)	(0.68)	(0.26)
Topic - Economic Outlook	-0.008	-0.012	-0.022*	-0.019	-0.006	-0.003	-0.005	-0.001
	(-0.99)	(-1.35)	(-2.32)	(-1.92)	(-1.04)	(-0.58)	(-0.67)	(-0.12)
Topic - Employment	-0.008	-0.002	0.014	0.010	0.008	0.005	0.005	0.002
	(-1.31)	(-0.31)	(1.49)	(1.00)	(1.75)	(1.08)	(0.59)	(0.21)
Topic - Financial Markets	0.011	0.006	-0.005	-0.002	-0.008	-0.006	-0.004	-0.002
	(0.81)	(0.47)	(-0.54)	(-0.12)	(-0.84)	(-0.64)	(-0.27)	(-0.17)
Topic - Inflation	0.004	0.004	0.000	0.001	-0.002	-0.000	0.006	0.007
	(0.79)	(0.76)	(0.03)	(0.08)	(-0.47)	(-0.10)	(1.14)	(1.22)
Control - Federal Funds Rate		-0.000		0.000		-0.000		-0.000
		(-0.85)		(0.06)		(-0.39)		(-0.51)
Control - Unemployment		0.000		0.000		0.000		0.000
		(0.18)		(0.30)		(0.42)		(0.43)
Control - Recession		-0.005		-0.004		-0.004		-0.005
		(-1.37)		(-1.15)		(-1.06)		(-1.21)
Constant	-0.001	-0.000	-0.003	-0.003	0.001	-0.000	-0.004	-0.004
	(-0.59)	(-0.11)	(-0.69)	(-0.57)	(0.41)	(-0.12)	(-1.30)	(-0.80)
N	225	224	225	224	225	224	225	224
adj. R^2	0.003	0.008	0.003	0.005	-0.007	-0.008	-0.012	-0.003

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: NMF topic impact on intra-day returns of the S&P500. Every column pair reports the coefficients of regressing intra-day returns onto the 6 topics generated by NMF without and with the 3 control variables capturing macroeconomic trends, as defined in 7. In this sense, columns represent the different values of emotion c, whereas the first 6 rows represent the topics n. Tone scores $\lambda_{c,n}$ are defined by equation 5. Control variables are provided by the FRED database and described in detail in table 6. t-statistics are based on White, 1980 standard errors. The estimates use 225 FOMC Minutes released between 1993 and 2021.

positive language (”greater”, ”higher”, ”better”) would be associated with adverse conditions in the overall economy, thus necessitating a ”greater” level of policy intervention.

Overall, topic tone appears to provide a similarly low level of insight into same-day market reactions as naive tone. Although I show that certain tone measures are related to contemporaneous market reactions, there is little evidence for my measures capturing the impact exceedingly well.

4.2 Predictive power of topic tone

Thus far, I have shown evidence that naive as well as topic tone are weakly associated with market reactions on the same day, with naive tone also holding a very limited amount of predictive power for equity returns over a longer time horizon. In the following, I demonstrate that certain topic tone measures are able to predict equity returns remarkably well.

The methodology employed is very similar to that used for naive tone predictions (figure 5), only with the appropriate adjustments made to accommodate for the set of 6 tone measures for one emotion at a time. Specifically, I regress the relative difference in the S&P500 ETF starting on the day of the release onto every topic-tone measure separately,

$$\frac{SPY_{t+k} - SPY_t}{SPY_t} = \alpha_{c,n,k} + \beta_{c,n,k}\lambda_{c,n,t} + \varepsilon_{c,n,t,k} \tag{8}$$

plotting the results grouped for the emotional category *positive tone* in figure 6. As before, a 95% confidence interval is calculated based on White, 1980 standard errors and laid around the coefficient line. Positive tone is selected to be shown here as an example, as it exhibits the most remarkable predictive potential for equity returns over a span of two years from the day of the Minutes release.

Particularly topics two and five (top middle and bottom middle) indicate that following a release that discusses planned policy actions or financial markets in a particularly positive light, a significant upward trend ensues in equity markets over the following year. Specifically, assuming again a uniform proportion of topics across paragraphs and documents, an additional positive word in every paragraph associated with the topic ”Financial Markets” would predict a 5% rise in the S&P500 over the ensuing year.[14]

To ensure that these results are not entirely spurious, appendix E assembles, as a practical test, a simple strategy based on the topic tone measures.

5 Conclusion

In this thesis, I study the link of US policymaker tone as proxied by FOMC Meeting Minutes with US equity markets. Quantifying the direct impact of qualitative information is relevant for policymakers themselves to better understand the consequences of policy communication, but may also be of interest to the sophisticated investor who can make better investment decisions by hardening

[14] $\frac{\partial \beta_{c,n}\lambda_{n,c,t}}{\partial \theta_{n,p,t}} = \beta_n \times \theta_{n,t} \times \sum_p^P \frac{\omega_{p,t}}{\Omega_t} = 0.3 \times 0.1\overline{6} \times 1 = 5\%$

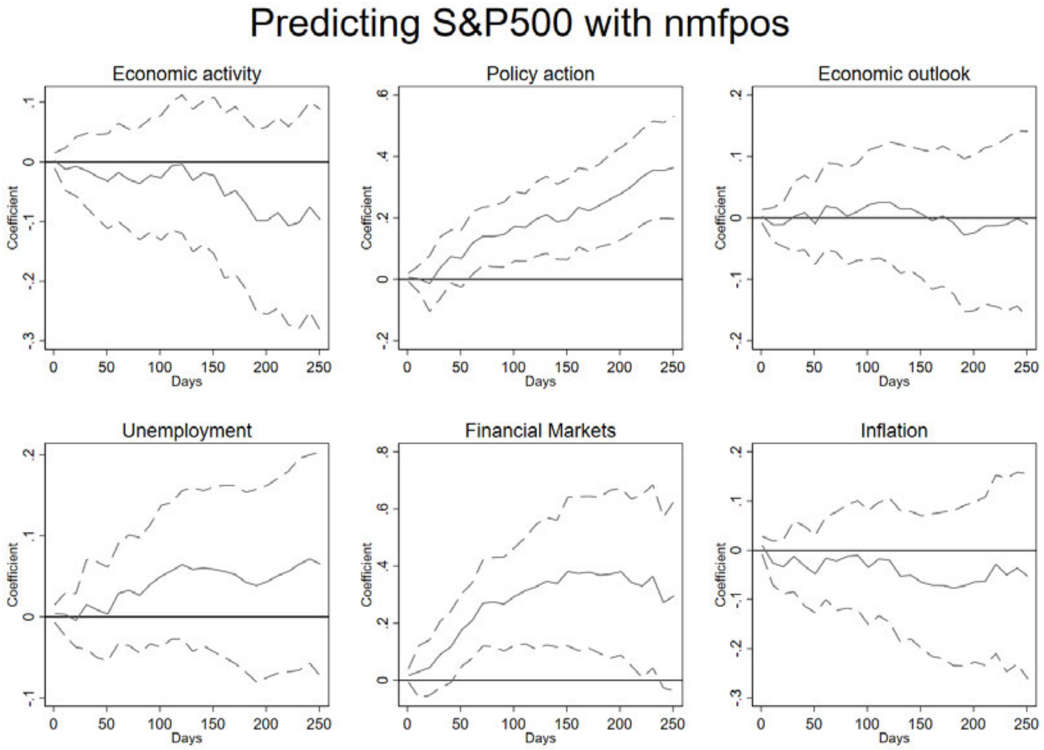

Figure 6: Predicting the S&P500 with positive topic tone. The purpose of these plots is to visualize when and by what magnitude predictability based on topic-tone measures shows significance. The solid line in every graph represents the series of slope coefficients from regressing the change in the S&P500 over k days after the release (with k denoted on the x-axis) onto the respective topic tone metric $\lambda_{pos,n}$ as specified in equation 8. With one year comprising roughly 250 trading days, the graphs cover about one full year of movement in the S&P500. Dashed lines report a 95% confidence interval based on White standard errors (White, 1980) for the corresponding coefficient value on the solid line.

this originally soft information.

I show that tonal scores for different dimensions of sentiment based on simple word counts are neither helpful in explaining returns on the day of the release, nor in predicting returns up to a year past the event. Nevertheless, because volatility on the release day does appear to be associated with some of the tone measures, I decide to scrutinize the information content of the Minutes more closely.

By applying a topic modelling technique, I extend my original tone metrics with information on the type of content that the particular tone is associated with. Using these enhanced measures to revisit the tests for impact and predictability shows that a handful of the extended measures perform very well as predictors for future equity returns over the span of one year. Their comprehensive exploration and implementation into a systematic strategy is at this point left open for future research.

References

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of finance*, *59*(3), 1259–1294.

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

Bjørnland, H. C., & Leitemo, K. (2009). Identifying the interdependence between us monetary policy and the stock market. *Journal of Monetary Economics*, *56*(2), 275–282.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning*, 113–120.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993–1022.

Boukus, E., & Rosenberg, J. V. (2006). The information content of fomc minutes. *Available at SSRN 922312*.

Cao, L., & Fei-Fei, L. (2007). Spatially coherent latent topic model for concurrent object segmentation and classification.

Chen, Y., & Filliat, D. (2015). Cross-situational noun and adjective learning in an interactive scenario. *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 129–134.

Cieslak, A., & Schrimpf, A. (2019). Non-monetary news in central bank communication. *Journal of International Economics*, *118*, 293–315.

Danker, D. J., & Luecke, M. M. (2005). Background on fomc meeting minutes. *Fed. Res. Bull.*, *91*, 175.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Devarajan, K. (2008). Nonnegative matrix factorization: An analytical and interpretive tool in computational biology. *PLoS Comput Biol*, *4*(7), e1000029.

Fei-Fei, L., & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *2*, 524–531.

Garla, A. (2021). Nmf — a visual explainer and python implementation. https://towardsdatascience.com/nmf-a-visual-explainer-and-python-implementation-7ecdd73491f8

Gürkaynak, R. S., Sack, B. P., & Swanson, E. T. (2004). Do actions speak louder than words? the response of asset prices to monetary policy actions and statements. *The Response of Asset Prices to Monetary Policy Actions and Statements (November 2004)*.

Hanley, K. W., & Hoberg, G. (2010). The information content of ipo prospectuses. *The Review of Financial Studies*, *23*(7), 2821–2864.

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the fomc: A computational linguistics approach. *The Quarterly Journal of Economics*, *133*(2), 801–870.

Hillert, A., & Schäfer, T. (2021). *The filing of amendments and investors' behavior* [unpublished], Goethe-Universität Frankfurt.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303

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

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent dirichlet allocation (lda) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169–15211.

Jensen, G. R., Mercer, J. M., & Johnson, R. R. (1996). Business conditions, monetary policy, and expected security returns. *Journal of Financial Economics*, *40*(2), 213–237.

Jiang, S., Qian, X., Shen, J., Fu, Y., & Mei, T. (2015). Author topic model-based collaborative filtering for personalized poi recommendations. *IEEE transactions on multimedia*, *17*(6), 907–918.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Jurafsky, D., & Martin, J. H. (2020). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. *published online*. https://web.stanford.edu/~jurafsky/slp3/

Kuang, D., Choo, J., & Park, H. (2015). Nonnegative matrix factorization for interactive topic modeling and document clustering. *Partitional clustering algorithms* (pp. 215–243). Springer.

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.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*(1), 1–22.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, *66*(1), 35–65.

Lucca, D. O., & Moench, E. (2015). The pre-fomc announcement drift. *The Journal of finance*, *70*(1), 329–371.

Madeira, C., & Madeira, J. (2019). The effect of fomc votes on financial markets. *Review of Economics and Statistics*, *101*(5), 921–932.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the 2011 conference on empirical methods in natural language processing*, 262–272.

M'sik, B., & Casablanca, B. M. (2020). Topic modeling coherence: A comparative study between lda and nmf models using covid'19 corpus. *International Journal*, *9*(4).

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, 100–108.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine learning*, *39*(2), 103–134.

O'callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, *42*(13), 5645–5657.

Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, *5*(2), 111–126.

Pakko, M. R. (1995). The fomc in 1993 and 1994: Monetary policy in transition. *Federal Reserve Bank of St. Louis Review*, *77*(2), 3.

Pascual, F. (2019). *Introduction to topic modelling (monkeylearn blog)*. https://monkeylearn.com/blog/introduction-to-topic-modeling/ Accessed: 2021-06-05

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Rigobon, R., & Sack, B. (2004). The impact of monetary policy on asset prices. *Journal of Monetary Economics*, *51*(8), 1553–1575.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the eighth ACM international conference on Web search and data mining*, 399–408.

Rosa, C. (2013). The financial market effect of fomc minutes. *Economic Policy Review*, *19*(2).

Russell, B. C., Freeman, W. T., Efros, A. A., Sivic, J., & Zisserman, A. (2006). Using multiple segmentations to discover objects and their extent in image collections. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, *2*, 1605–1614.

Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., & Freeman, W. T. (2005). Discovering object categories in image collections.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, 952–961.

Suri, P., & Roy, N. R. (2017). Comparison between lda & nmf for event-detection from large text stream data. *2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT)*, 1–5.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association*, *101*(476), 1566–1581.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, *62*(3), 1139–1168.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582.

Wang, X., & McCallum, A. (2006). Topics over time: A non-markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 424–433.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, 817–838.

Appendices

A Connection with event study literature

To quantify the impact of central bank communication, many papers employ what has become known as the ”event study” approach (eg Rosa, 2013, Cieslak and Schrimpf, 2019, Gürkaynak et al., 2004)[15]. The key problem that this literature attempts to tackle is a potential simultaneous causation bias in the estimator, as monetary policy decisions are not entirely disjunct from asset price reactions. Rigobon and Sack, 2004 provide mathematical proof that the resulting endogeneity bias depends negatively on the policy shock magnitude relative to unexpected changes in asset prices.

The event study literature attempts to mitigate this problem by analyzing the effect of monetary policy shocks in a very tight window around the policy announcement, thereby reducing the chance of co-occurring news that are not directly linked to the policy release. This approach warrants the argument that the change in asset prices during this window can in its entirety be *approximately* attributed to the policy announcement. This result holds only *in approximation*, as true unbiasedness would only be given if the event window were to be of size null, which in turn would obviously obstruct measurement of asset price changes, as markets take time to fully incorporate information into prices. To trade off unbiasedness and relevance, most papers construct a window of a few minutes to a few hours around the news event[16] but generally do not work with data on the daily timeframe.

Even though I am aware of this problem, I choose to work with financial data on the daily level, as the contribution of this thesis is primarily to show the predictive power of FOMC sentiment concerning certain topics in the Meeting Minutes and not the precise impact quantification of the releases.

[15] According to Rigobon and Sack, 2004, this method stands in the tradition of Cook and Hahn (1989).

[16] Rigobon and Sack, 2004 show that this approach still yields biased estimators and suggest a truly exogenous estimator by instrumenting changes in volatilities around the event to measure the impact on asset prices.

B Connection with Jegadeesh and Wu, 2017

The construction of tone scores is inspired by Jegadeesh and Wu, 2017. This appendix serves to clarify my choices for and against certain features of their setup.

Before computing individual tone scores, the authors choose to merge the Harvard IV-4 Psychosociological Dictionary with the Loughran&McDonald Financial Dictionary to count the occurrences of positive, negative, and uncertain words. Although this choice ultimately provides them with richer tone metrics, it warrants concern for bias in the calculation. Because the LM dictionary was precisely constructed to reduce misclassifications for analyses of financial language that occured all too often with the general Harvard dictionary, a combination of both word lists would defeat the purpose of using the LM dictionary in the first place. I thus do not merge the dictionaries and compute tone solely based on the LM dictionary.

A second modification concerns the tone weighting scheme employed in the final computation of tone scores. Jegadeesh and Wu weigh the paragraph-level tone score by the inverse paragraph length, arguing that the relevance of sentiment-specific term occurence declines in the length of the paragraph. This notion is based on the assumption that longer sections are generally more difficult to comprehend and thus transmit a lower impact per emotion-specific word. Although I would see this specification fit for purposes with a more popular audience, I believe that because publications of this type are targeted specifically at a professional audience, section length does not compromise the perceived weight of sentiment-specific words. Hence positing a (positive) linear relationship between the number of words assigned to sentiment category c and the paragraph sentiment score, I weigh paragraphs proportionally to their relative length instead of inversely to their absolute length.

Dividing by the document's overall length Ω as part of the weighting scheme has the additional advantage of not assuming stationarity of the overall level of detail of the Minutes' writing style, which could be expressed as a higher document length despite a constant (intended) sentiment amplitude. A weighting scheme that does not account for this type of variability would - assuming uniform sentiment-specific word distribution across documents - assign higher sentiment scores to longer publications, even if differing document length was entirely caused by variations in the level of detail. Evidence for non-stationarity is apparent in figure 2, showing the total word count of the entire document steadily declining by about one third until 2005 and since doubling until today.

C Accounting for lagged volatility

A concern noted by Jegadeesh and Wu, 2017 about the underlying cause for increased volatility on the day of the release relates to how expectations about the Minutes' contents are treated by investors. As the document is released a considerable amount of time after the meeting takes place, market opinion on the meeting's substance could become biased, eg by political speeches, the information content of which would be overestimated. The resulting price changes to incorporate the pseudo-information into prices would increase volatility in the days prior to the release. As soon as the documents were released and the falsehood of the initial information set recognized, the market would then adjust prices, thereby leading to another bout of increased volatility.

It is therefore clear that a higher degree of bias in the run-up to the release is associated with a stronger reaction on the day of, thus necessitating some form of control for the strength in biased expectations. If this phenomenon were to go unchecked, estimates of the market impact of the Minutes could be both over- and underestimated. To this end, Jegadeesh and Wu, 2017 employ volatilities over several days prior to the event but find no meaningful change in the significance of the ”day effect”, thus indicating little to no bias of the sort proposed above. Stating that little change in the significance of the day effect implies little bias in expectations overall is of course contingent on the assumption that the bias is not conditional on the tone of the subsequent announcement. If it were, then estimates of the tone impact coefficients could still be distorted.

As their findings indicate that this potential phenomenon is likely of minor concern, I decide to not discuss it as part of my main results. Moreover, I need to stress at this point that my technique is inferior to that of Jegadeesh and Wu, 2017, since my ”event window” (see appendix A) is at least 16 times larger than the one they construct. It is therefore not unlikely that my results are confounded by higher volatility in the days running up to the release. It would, however, be imprecise to attribute this effect in its entirety to the bias-in-expectations phenomenon described above by using lagged volatilities as an unmediated control, as my daily-timeframe approach does not distinguish between volatility within a short event window around the release and volatility over any other part of the day.

D Topic modelling technique

D.1 Overview

As shown in section 3, Minutes releases with a higher net tone or uncertainty score induce higher volatility on release days, indicating that investors do react to emotional information. However, as I was unable to establish a unidirectional price change in the S&P500, the question arises whether the naive tone score captures the entire picture or whether a more granular approach might yield different results. To this end, I replace the assumption of the Minutes documents being homogeneous content pieces in favor of viewing them as a heterogeneous collection of discussion points, which calls for a technique that can reliably separate these topics. Computing individual tone scores for every single one of these issues would then allow me to test whether a more granular approach offers more insight into the behavior of investors and whether the Minutes might even be useful for forecasting various financial variables.

The family of methods referred to above are known as ”topic models” in computational linguistics and have been the subject of a growing amount of research in the past years, not least because of their wide array of applications in a number of fields. Customer support centers can for example categorize and quantify the content of complaints with much greater objectivity and efficiency than a team of human professionals could (Pascual, 2019). Apart from text mining, topic models have also been used to enhance computer vision (Sivic et al., 2005, Russell et al., 2006, Cao and Fei-Fei, 2007, Fei-Fei and Perona, 2005), recommendations in social networks (Jiang et al., 2015) and have even been adopted in bioinformatics (Liu et al., 2016, Devarajan, 2008) where they are used for genome sequencing.

Of all methods that have been developed over the years, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is arguably the most prominent one, in part due to its general applicability (Jelodar et al., 2019). Other techniques focus on scenarios where the inter-temporal evolution of topics is of importance[17] or where the length of input documents is very limited[18]. Without going into detail on the history and confines of other methods at this point,[19] I apply an alternative known as Non-Negative Matrix Factorization (NMF) that produces

[17] Dynamic Topic Model (Blei and Lafferty, 2006) and Topics Over Time (Wang and McCallum, 2006)

[18] Mixture of unigrams (Nigam et al., 2000)

[19] The interested reader may want to refer to Vayansky and Kumar, 2020 for an extensive albeit not exhaustive review of the most widely-used topic modelling methods. Liu et al., 2016 also outline the construction and application of a topic model in the context of bioinformatics in a relatively easy-to-understand manner.

qualitatively similar results to LDA and can thus be seen as a direct substitute. This choice is in part motivated by the lack of research that applies this method specifically to the FOMC Meeting Minutes, in part also by the vastly shorter runtimes to train the model, but primarily by the fact that NMF seems to produce topics with a higher degree of robustness to changes in the model parameters. The higher quality of topics produced has also been observed by O'callaghan et al., 2015, by Garla, 2021 and by Hillert and Schäfer, 2021, who use NMF to examine the reaction of investors to 10-K amendments on a granular level, as market participants do not seem to respond to overall document tone, conceptually similar to my findings in section 3.[20]

NMF was initially developed as an extension to the dimensionality-reduction method Latent Semantic Analysis (LSA) (Deerwester et al., 1990) and published by Paatero and Tapper, 1994. Later popularized by Lee and Seung, 1999 in the journal *Nature*, the method has attracted a high amount of attention in the two decades since (Kuang et al., 2015).

D.2 Technical setup

In the following, I discuss the mechanics of the topic modelling process, drawing primarily on Kuang et al., 2015 for the explanation of the NMF method.

The result of the pre-processing methods detailed in section 2.3 can be understood as a matrix that associates each document with an absolute count of every unique word (lemma). With documents as rows and words as columns, the integers in the document-word matrix can now be modified to yield higher values for terms with a greater level of meaning to the underlying set of documents. This step provides the NMF algorithm with more information, thus producing better results in the quality of topics, as it reduces irrelevant noise in the matrix (as shown for example by Chen and Filliat, 2015). To this end, I employ the TF-IDF weighting scheme (Jones, 1972) that is commonly used in applications of natural language processing that aim to extract the most relevant words in a text.

TF-IDF stands for ”Term Frequency - Inverse Document Frequency” and multiplies, as the name suggests, the frequency of every unique word (lemma) per document with the inverse of the term frequency over all documents. In its

[20] Opinions on the superiority of LDA vs NMF in terms of topic coherence diverge in the literature. For example, Stevens et al., 2012, M'sik and Casablanca, 2020, Suri and Roy, 2017 find that LDA outperforms in this regard for their respective use cases. These papers, however, also agree on the markedly higher computational efficiency of NMF.

most common form, TF-IDF is specified as

$$\text{tf-idf}_t = \text{tf} \times \text{idf} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \times (-\log \frac{n_t}{N})$$

t - term identifier n_t - term t count over all documents
d - document identifier N - total word count

If therefore a particular term t occurs often in document d relative to the sum of frequencies of all other words in that document, the *tf* factor will bear a high value as the word presumably describes the document better than other ones. Because there are a large number of words that appear often in every document, the *idf* factor then downweights words such as ”the”, ”to”, ”of”, that add little meaning to the text.[21] The highest weight is therefore assigned to words that appear often in a given document, but rarely in the entire set of documents, since this type of word uniquely identifies the particular document in which it occurs often. Words with a moderate frequency that is relatively stable across all documents would conversely receive a low weight, as they do not aid in separating documents by topic in the following.[22]

The modified document-term matrix is now used by the NMF algorithm to produce two separate factor matrices **W** and **H** that, when multiplied, approximate the original document-term matrix **A**. While **W** contains relative topic weights for each document (documents as rows, topics as columns), matrix **H** describes the importance of a particular word to a given topic (topics as rows, words as columns).

$$\mathbf{W}_{d\times t} \times \mathbf{H}_{t\times w} \approx \mathbf{A}_{d\times w}$$

$$\begin{bmatrix} w_{11} & \ldots \\ \ldots & \\ & \\ & w_{dt} \end{bmatrix} \times \begin{bmatrix} h_{11} & \ldots & \\ \ldots & & h_{tw} \end{bmatrix} \approx \begin{bmatrix} a_{11} & \ldots & \\ \ldots & & \\ & & \\ & & a_{dw} \end{bmatrix}$$

As NMF by its nature and origin is a dimensionality reduction technique, the number of topics that connects both factor matrices is required to be set by the researcher to a value lower than both the number of documents and words in matrix **A** ($t < \min\{d, w\}$). My choice for this parameter is discussed

[21] Those three examples would of course have already been removed in section 3.2 as I used *spacy*’s built-in word classifier to remove all word types but nouns, verbs and adjectives.

[22] Since my weighting scheme relies on term counts instead of frequencies, the f would need to be replaced by a ν for my implementation. The tf-idf mechanism is however unaffected by this adjustment.

in the next subsection. For lack of a closed form solution, the two matrices are found by minimizing the difference of the approximated and the actual matrix, as measured by the Frobenius norm:

$$\min_{W \geq 0, H \geq 0} f(W, H) = ||\mathbf{A} - \mathbf{WH}||_F^2$$

With the Frobenius norm defined as $\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}$, this objective function sums the squared deviations of corresponding elements in the estimated and real document-term matrix. Like many machine-learning algorithms, a local minimum is then found through an iterative process commonly known as gradient descent in the machine learning literature.

To ensure interpretability of the estimated document-topic and topic-term values, every element of the two factor matrices is constrained to be non-negative during the optimization, since negative weights would have little real-life meaning (hence the name). Also, because it is often convenient to work with the topic weights by document as absolute percentages or probabilities, a widely-used transformation is to row-normalize matrix **W** as $\hat{\mathbf{W}}$

$$\mathbf{WH} = (\mathbf{WD}_W)(\mathbf{D}_W^{-1}\mathbf{H}) = \hat{\mathbf{W}}\tilde{\mathbf{H}}$$

with the diagonal matrix $\mathbf{D}_W$ holding the column sums of **W** on its diagonal. $\hat{\mathbf{W}}$ can now be interpreted analogously to the output of LDA, which differs from NMF primarily in using a probabilistic approach and thus generates a fraction matrix directly. The same is however not applicable to matrix $\tilde{\mathbf{H}}$, whose values still have to be handled as ”relative weights” as before.

To avoid the estimation being dominated by a few extreme TF-IDF values, particularly common as well as rare words were removed from the document-term matrix before applying TF-IDF. Specifically, words (columns) that appear fewer than 3 times across all documents (column sum < 3) and those that are found in more than 85% of documents (fraction of non-zero elements in column > 85%) are discarded, lowering the matrix dimensions from 10037×4943 to 10037×3301. Accordingly, dimensions of the two factor matrices produced by NMF are of dimensions 10037×6 for **W** and 6×3301 for **H**. These transformation steps along with the subsequent topic modelling were implemented with the open-source Python library *gensim* provided by Řehůřek and Sojka, 2010.

D.3 Parameter selection

As is the case for most topic modelling algorithms, NMF unfortunately does not determine the optimal number of topics endogenously.[23] Because most researchers usually desire a certain degree of objectivity in setting this parameter however, a number of metrics have been developed that quantify the quality of topics produced by topic models. With the most commonly used measure being the *coherence score* as proposed by Newman et al., 2010, I proceed to compute the average topic coherence as a function of the number of topics, shown in figure 7.

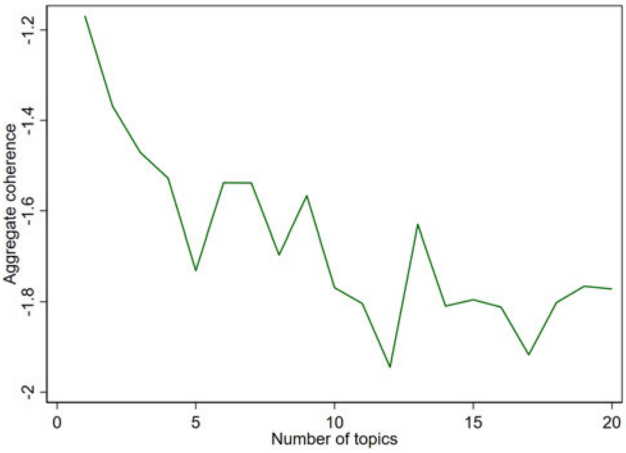

Figure 7: Coherence measure, proposed by Newman et al., 2010, as a function of the number of topics. Because coherence is computed for every topic individually, the aggregated measure in the graph above is the arithmetic mean of all k topic coherence values corresponding to the k topics used as a parameter for the model. The specific type of coherence measure used is *UMASS* (Mimno et al., 2011). Although it is the least accurate of all currently known measures, it is also the computationally least demanding one. (For an overview over and comparison of the different measures, see Röder et al., 2015). As computation of more accurate alternatives would not be feasible inside a 24-hour time window with the computing power at my disposal, I decide to employ this specific variant.

At first glance, the graph seems to suggest that the overall quality of the model deteriorates in the number of topics employed. Figure 8 however provides evidence that the coherence score of an individual topic remains relatively stable under a changing total number of topics, with the behavior of the aggregate score being primarily driven by a lower score *on the margin*. To still make use of this metric, one would need to define an arbitrary cutoff value, below which no new topics would be added.

Since I want to avoid making unfounded assumptions, I decide to proceed like Hillert and Schäfer, 2021 and select the number of topics based on my

[23] The Hierarchical Dirichlet Process (HDP), as presented by Teh et al., 2006, is one example of a method that learns the appropriate number of topics automatically (Vayansky and Kumar, 2020).

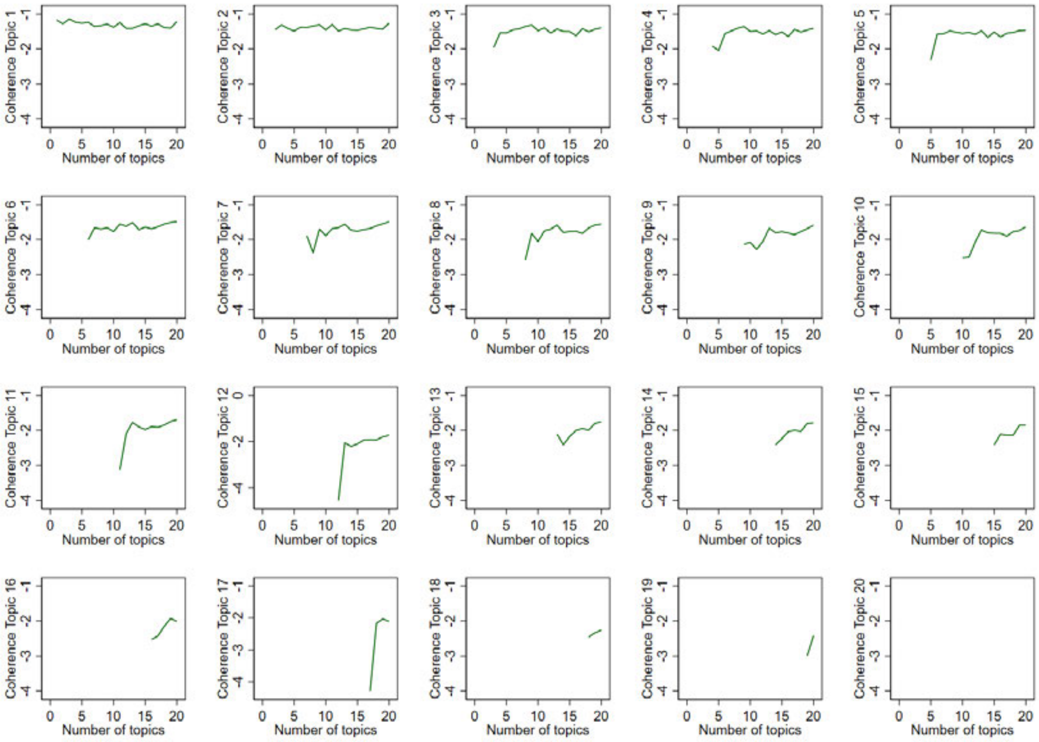

Figure 8: Progression of coherence for individual topics as a function of the overall number of topics. As in figure 7, *UMASS* was employed as a coherence specification. Since a coherence score for topic k can only be computed once the number of topics $K \geq k$, values for $K < k$ do not exist. The key takeaway from this graph is that while coherence for individual topics is relatively stable, the average value of the topic *on the margin* keeps decreasing.

subjective judgement of topic coherence. After a careful revision of the most relevant terms for all 20 models generated, I decide to continue with the model that produces 6 topics. A smaller number yields ambiguous clusters, while a greater number leads to redundancy, as the same underlying topic seems to be captured by more than one cluster. Notably, the aggregate coherence score in figure 7 also has a maximum at 6 topics if values below 5 are disregarded, which is likely a reasonable assumption. Incidentally, another paper in the field that applies Latent Semantic Analysis (LSA), a similar technique, has identified five topics in the Minutes independently from me (Boukus and Rosenberg, 2006). The fact that I learned about their selection only after having implemented the technique is another pointer towards the validity of my choice.

Before assessing the predictive power of the tone by topic in the next section, I briefly describe the 6-topic model by examining the two factor matrices in the following passage.

D.4 Model evaluation

Figure 9 visualizes the topic-word matrix $\tilde{\mathbf{H}}$ by showing the most important words for each topic, with the specific topic-term weights determining the size of the word in the cluster. Additionally, table 8 in appendix F showcases one representative paragraph per topic with the respective link to the corresponding Minutes document.

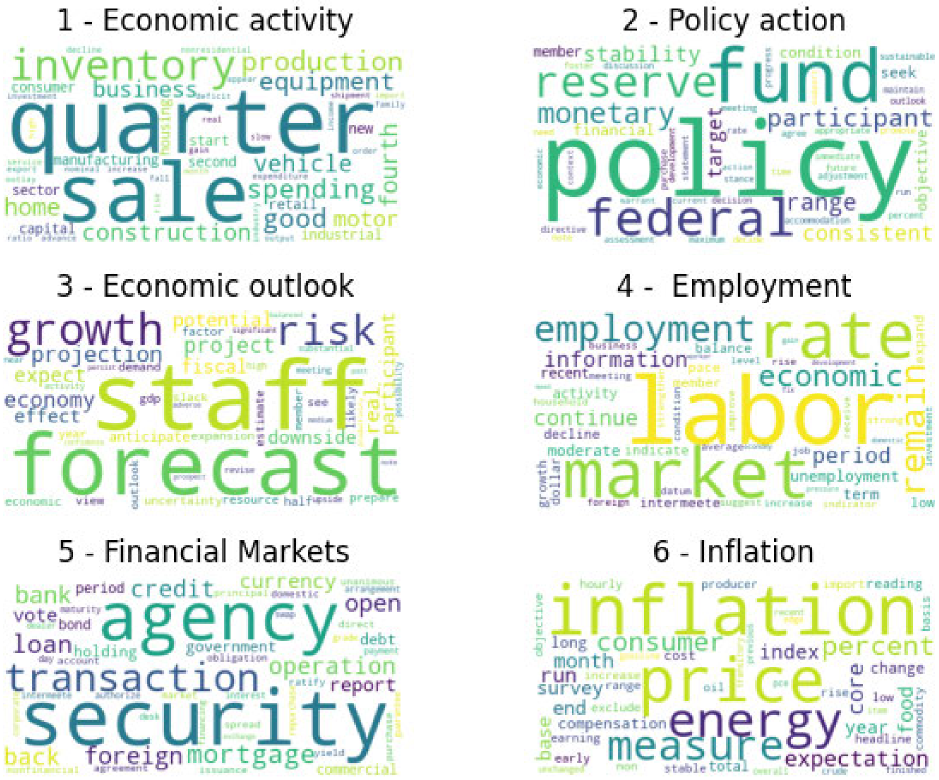

Figure 9: World clusters or so-called "word clouds" for every topic produced by the NMF algorithm. The size of the individual terms is determined by the relevance that the specific word bears for the respective topic, as specified in the modified factor matrix $\tilde{\mathbf{H}}$

.

Based on these word clouds, topics are labelled as follows:

1 - Economic activity
2 - Policy action
3 - Economic outlook

4 - Employment
5 - Financial markets
6 - Inflation

To visualize the document-topic matrix $\hat{\mathbf{W}}$, I plot the normalized topic proportions for each Minutes document over time in figure 10, as the matrix rows simultaneously capture a temporal evolution.

Overall, the topic share seems to be relatively stable over time, which in itself is already a positive sign for the validity of the model, since the documents are by design supposed to be structured in a similar fashion. The most notable

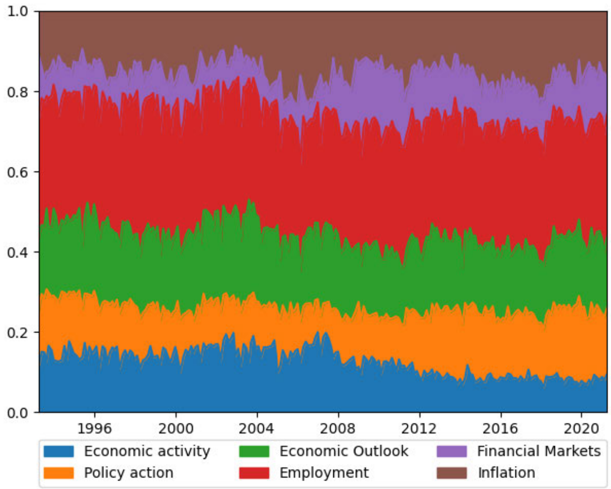

Figure 10: Progression of topic proportions over documents and time. Values are derived from the factor matrix **W** by normalizing so that every row sums to unity. Topics appear to be stable over time, indicating model validity. A notable increase in topic share is visible in ”Financial Markets” around 2008, indicating that financial matters became of increasing concern to the FOMC.

change in the share of any topic seems to be that of topic 5, capturing discussions about financial markets. While taking up less than about 10% before the GFC in 2008, there seems to have been a marked increase in the discussion share of this topic, enduring up until today. A plausible explanation for the shift at this particular time would be the preparation of the Dodd-Frank act that gave the FED greater powers regarding the supervision of financial firms. The lack of an equivalent change in the wake of the Sarbanes-Oxley act of 2002 could be explained by the fact that this earlier package of regulations was primarily designed to remedy accounting errors and fraudulent financial practices, thus not affecting the FED as directly as the Dodd-Frank act.

E Strategy test for topic tone predictions

Based on the results presented in sections 3 and 4, I have deduced that, although naive tone is not overly conducive for predicting future equity returns, certain topic-tone combinations do bear significant potential. To corroborate these findings, I construct a simple strategy in the following that holds a position in the S&P500 over 250 trading days depending on the standardized value of the respective topic tone score:

$$r_{t,t+250}^{Tone} = \frac{\lambda_{n,c,t} - \mu_{\lambda_{n,c}}}{\sigma_{\lambda_{n,c}}} \times r_{t,t+250}^{SPY} \tag{9}$$

Standardization is necessary, as tone scores with the exception of *net tone* are by their definitions (3, 5) strictly positive, which, if unchecked, would imply taking systematic long positions in the equity markets. I choose 250 days as a holding period, as predictability over shorter time horizons (such as 30 days) is very limited and in significance comparable to the case of naive tone. The specific number is selected because it approximately corresponds to the average number of trading days per year.

Again, positive topic tone is presented here in figure 11 as an example; similar predictive power is however also exhibited by other topic-tone combinations. Panel B shows significant variation between the performance of different $\lambda_{n,pos}$ as input, thereby suggesting a high degree of uniqueness and little correlation between the signals. It is important to note here that the strategy does not incorporate information regarding the significance and direction of the prediction from panel A. The negative performance of topics 1, 3, 4 and 6 are thus entirely unconcerning, as based on figure panel A, these strategies would have been discarded for a lack of significance in the prediction.

To further buttress my argument, I also showcase prediction and strategy graphs for *uncertain* topic tone in figure 12. As is clearly visible from panel A, topics 1 and 6 show a strong negative trend following a marginal increase in the respective uncertainty scores. Any rational investor would in both cases of course invert the strategy specified above and according to panel B, earn marked positive returns over the entire sample period from 1993 to 2021.

To avoid overloading this appendix with technicalities, I do at this point not analyze the exact performance metrics of these strategies in greater detail, but leave the comprehensive exploration and exact performance quantification open for future research.

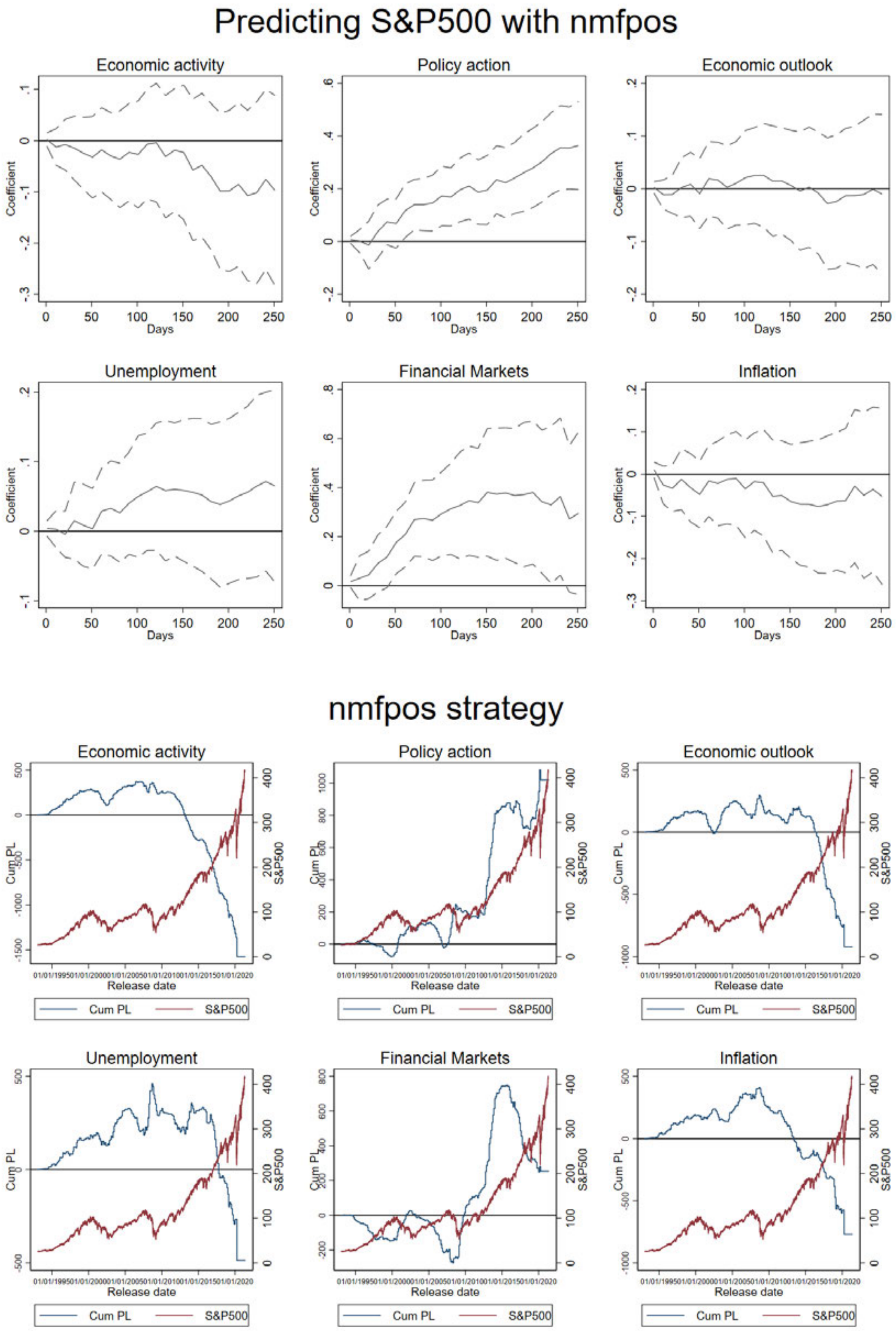

Figure 11: Evaluation of a positive topic tone strategy. The purpose of this graph is to showcase the practical applicability of the significant predictive potential of some topic-tone measures. Panel A shows the predictive potential of positive tone associated with every one of the 6 topics, as discussed in detail in section 4. Panel B shows the performance of a strategy constructed using the topic tone measures (see equation 9) since the start of the sample period in 1993. Importantly, this strategy does not incorporate information on whether panel A predicts an upwards or downwards trend. If panel A therefore shows a predicted downward trend for a particular topic-tone combination, a profit-seeking investor would invert the strategy, ie multiply the right-hand side of equation 9 with -1.

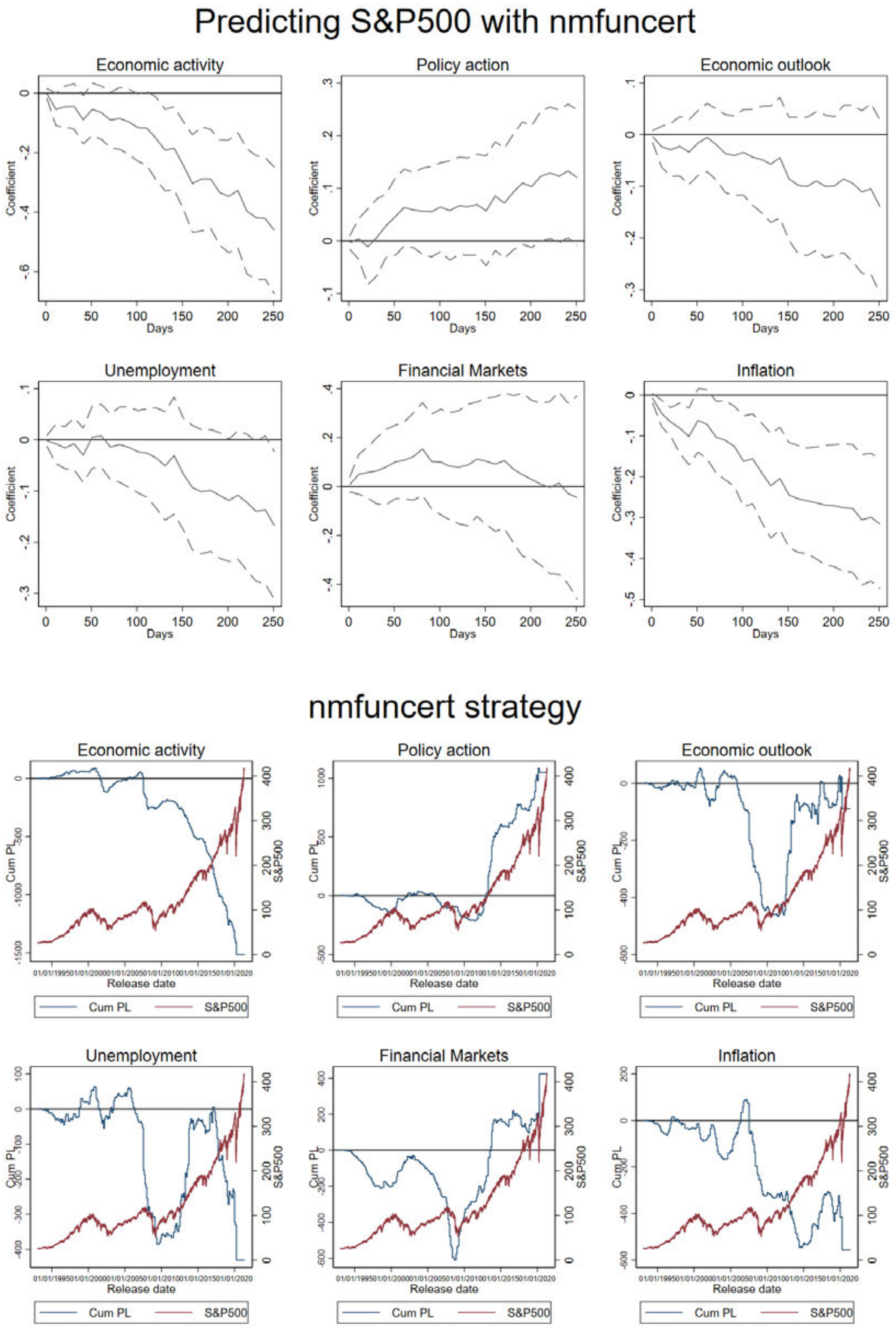

Figure 12: Evaluation of an uncertain topic tone strategy. The purpose of this graph is to showcase the practical applicability of the significant predictive potential of some topic-tone measures. Panel A shows the predictive potential of uncertain tone associated with every one of the 6 topics, as discussed in detail in section 4. Panel B shows the performance of a strategy constructed using the topic tone measures (see equation 9) since the start of the sample period in 1993. Importantly, this strategy does not incorporate information on whether panel A predicts an upwards or downwards trend. If panel A therefore shows a predicted downward trend for a particular topic-tone combination, a profit-seeking investor would invert the strategy, ie multiply the right-hand side of 9 with -1.

F Additional figures and tables

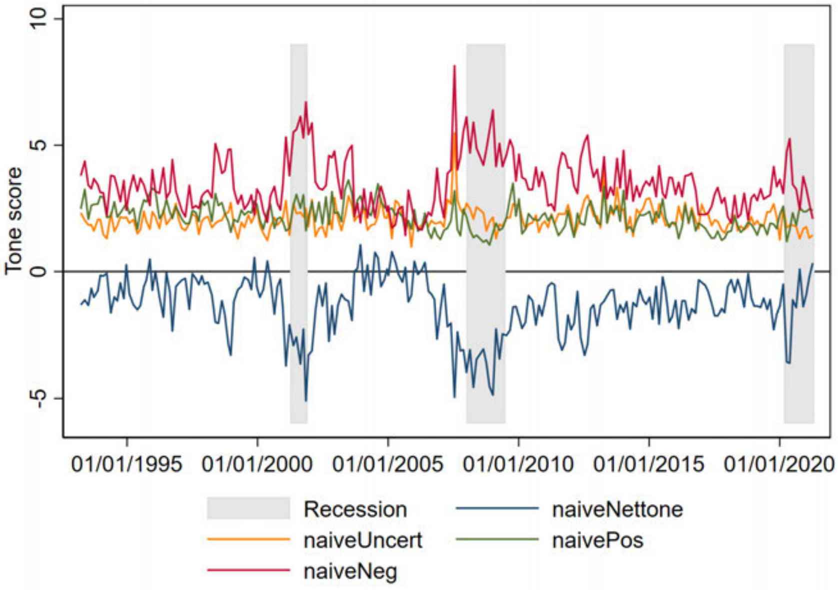

Figure 13: Progression of naive tone scores over documents and time, as defined in equation 3. The number of sentiment-specific words, as defined by the LM dictionary, is counted per paragraph and weighted by the relative length of the paragraph in relation to the overall document length. These weighted paragraph scores are then summed to arrive at an emotion-specific tone score per Minutes document. In the case of net tone, the negative word count is subtracted from the positive word count before weighting by the relative paragraph length. With positivity and uncertainty exhibiting characteristics of white noise, the net tone measure is mostly driven by the negative number of negative words.

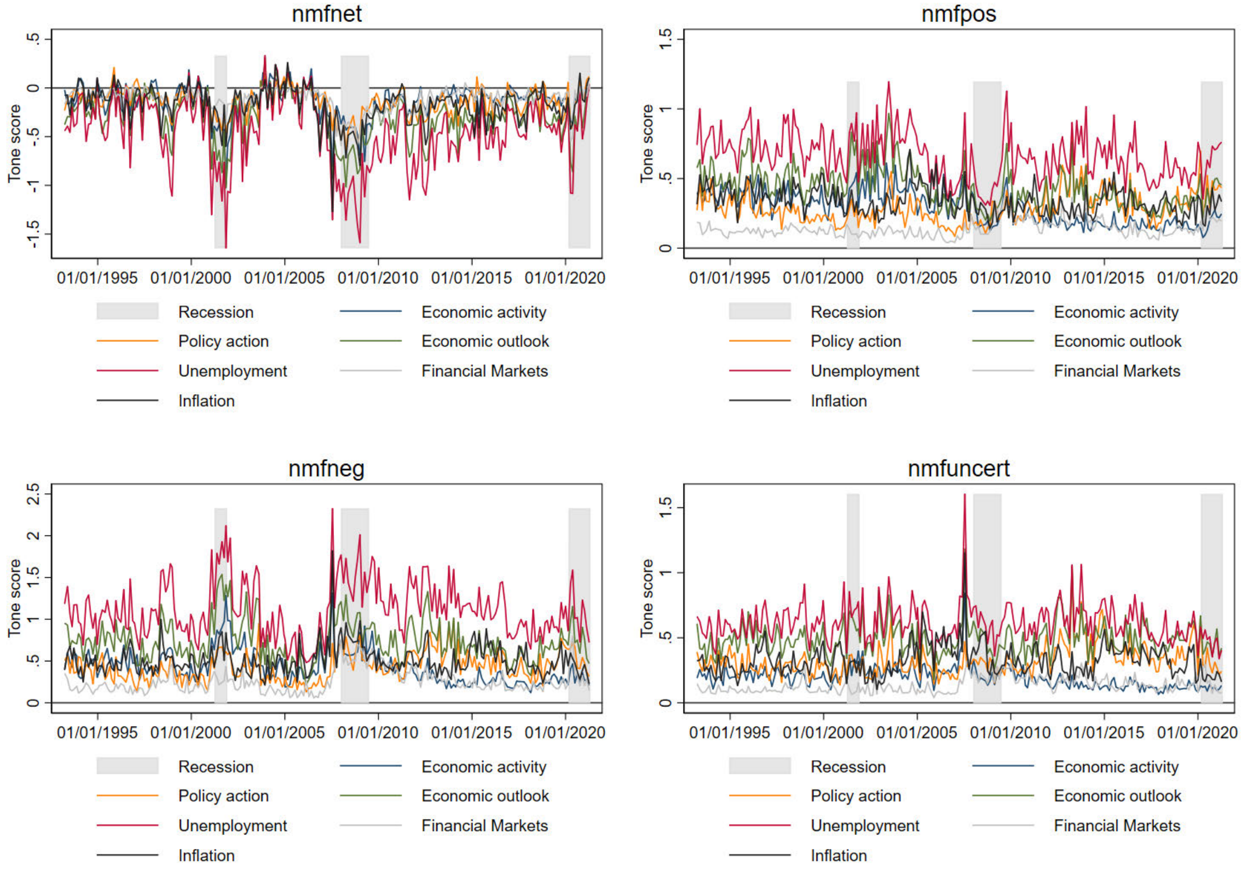

Figure 14: Progression of topic tone scores over documents and time, as defined in equation 5. The number of sentiment-specific words, as defined by the LM dictionary, is counted per paragraph and weighted by the relative length of the paragraph in relation to the overall document length. These weighted paragraph scores are then summed to arrive at an emotion-specific tone score per Minutes document. In the case of net tone, the negative word count is subtracted from the positive word count before weighting by the relative paragraph length.

Variable name	Description	Source	Link
S&P500	SPDR S&P 500 ETF Trust (ticker symbol: SPY)	Yahoo Finance	https://finance.yahoo.com/quote/SPY?p=SPY&.tsrc=fin-srch
ν_c	Word count of sentiment type c in respective passage of text, as defined in Loughran & McDonald dictionary.	Loughran & McDonald Sentiment Word List	https://sraf.nd.edu/textual-analysis/resources/
Federal Funds Rate	The federal funds rate is the interest rate at which depository institutions trade federal funds (balances held at Federal Reserve Banks) with each other overnight.	FRED	https://fred.stlouisfed.org/series/FEDFUNDS
Unemployment Rate	The unemployment rate represents the number of unemployed as a percentage of the labor force.	FRED	https://fred.stlouisfed.org/series/UNRATE
Recession dummy	NBER binary recession indicator	FRED	https://fred.stlouisfed.org/series/USREC

Table 6: Description and source for all variables used in this paper that are of external origin. FRED indicates the ”Federal Reserve Economic Data” database provided by the Federal Reserve Bank of St. Louis.

	V	V	R	R
Economic activity	-11.192***	-10.458*	-0.005	-0.006
	(-3.97)	(-2.43)	(-0.17)	(-0.16)
Policy action	-4.513*	-3.858	0.021	0.017
	(-2.03)	(-1.77)	(0.74)	(0.62)
Economic outlook	7.382**	4.962*	-0.036	-0.029
	(2.87)	(2.18)	(-1.12)	(-0.81)
Unemployment	2.550	4.021*	0.006	-0.005
	(1.65)	(2.54)	(0.24)	(-0.17)
Financial Markets	3.550**	1.485	-0.015	-0.015
	(2.71)	(0.84)	(-0.83)	(-0.76)
Inflation	3.047**	3.133**	0.024	0.026
	(2.97)	(2.73)	(1.42)	(1.52)
Federal Funds Rate		-0.021		-0.000
		(-0.37)		(-0.02)
Unemployment Rate		-0.008		0.000
		(-0.08)		(0.79)
Recession		0.969***		-0.005
		(3.50)		(-1.21)
Observations	225	224	225	224
Adjusted R^2	0.645	0.700	-0.005	0.006

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: Market reaction to topic proportions. The constant is omitted to avoid multicollinearity. The overall picture is very similar to the naive tone score regression: Volatility is affected with the strongest effect observable in the first topic ”Economic activity”, whereas daily returns appear unaffected by FOMC tone.

Topic	Label	Sample paragraph	Link
1	Economic activity	Real residential investment appeared to be weakening significantly in the second quarter. Starts and building permit issuance for single-family homes, along with starts of multifamily units, dropped sharply in April. Sales of existing homes contracted markedly in April, although new home sales edged up.	https://www.federalreserve.gov/monetarypolicy/fomcminutes20200610.htm
2	Policy action	As it noted in its statement of principles regarding longer-run goals and monetary policy strategy released in January, the Committee seeks to explain its monetary policy decisions to the public as clearly as poss ble. With that goal in mind, participants discussed a range of additional steps that the Committee might take to help the public better understand the linkages between the evolving economic outlook and the Federal Reserve's monetary policy decisions, and thus the conditionality in the Committee's forward guidance. The purpose of the discussion was to explore potentially promising approaches for further enhancing FOMC communications; no decisions on this topic were planned for this meeting and none were taken.	https://www.federalreserve.gov/monetarypolicy/fomcminutes20120313.htm
3	Economic outlook	The staff viewed the uncertainty around the forecast for economic activity as normal relative to the experience of the past 20 years. However, the risks were still viewed as skewed to the downside, in part because of concerns about the situation in Europe and the ability of the U.S. economy to weather potential adverse shocks. Although the staff saw the outlook for inflation as uncertain, the risks were viewed as balanced and not unusually high.	https://www.federalreserve.gov/monetarypolicy/fomcminutes20130619.htm
4	Employment	Participants agreed that labor market conditions had strengthened further over the intermeeting period. Payrolls had grown strongly in June, and labor market tightness was reflected in recent readings on rates of private-sector job openings and quits and on job-to-job switching by workers. Although the unemployment rate increased slightly in June, this increase was accompanied by an uptick in the labor force participation rate.	https://www.federalreserve.gov/monetarypolicy/fomcminutes20160921.htm
5	Financial markets	The Committee directs the Desk to continue rolling over maturing Treasury securities at auction and to continue reinvesting principal payments on all agency debt and agency mortgage-backed securities in agency mortgage-backed securities. The Committee also directs the Desk to engage in dollar roll and coupon swap transactions as necessary to facilitate settlement of the Federal Reserve's agency mortgage-backed securities transactions.	https://www.federalreserve.gov/monetarypolicy/fomcminutes20160316.htm
6	Inflation	Both total U.S. consumer price inflation, as measured by the PCE price index, and core inflation, as measured by PCE prices excluding food and energy, were about 1-1/2 percent over the 12 months ending in October; consumer energy prices declined, while consumer food prices rose more than overall prices. Over the 12 months ending in November, total inflation as measured by the consumer price index (CPI) was 1 1/4 percent, partly reflecting the further decline in energy prices, while core CPI inflation was 1-3/4 percent. Measures of expected long-run inflation from a variety of surveys, including the Michigan survey, the Blue Chip Economic Indicators, the Survey of Professional Forecasters, and the Desk's Survey of Primary Dealers, remained stable. In contrast, market-based measures of inflation compensation moved lower.	https://www.federalreserve.gov/monetarypolicy/fomcminutes20141217.htm

Table 8: Sample paragraph for every topic. The table shows a representative paragraph and its corresponding link for every one of the six topics generated by NMF, explained in appendix D. The passages shown are each within the top 5 of the paragraphs with the highest weight for the respective topic. The most relevant words for each topic are visualized in figure 9, where the size of each word is determined by the significance that it bears for the respective topic.