Karlsruhe Institute of Technology

Department of Financial Economics, Banks and Insurances

Chair of Financial Economics and Risk Management

Prof. Dr. Maxim Ulrich
Advisor: M. Sc. Elmar Jakobs

The Information Content of FOMC and ECB Meetings

Seminar Thesis

Author: Lukas Struppek
1954638
Wirtschaftsingenieurwesen (M.Sc.)
lukas.struppek@student.kit.edu

Karlsruhe, December 18, 2018

Assertion

Ich versichere wahrheitsgemäß, die Arbeit selbstständig verfasst, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde sowie die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, December 18, 2018 Lukas Struppek

Contents

List of Figures

List of Tables

1 Introduction

In this seminar thesis a machine learning approach is implemented and applied for analyzing meeting minutes of the Federal Open Market Committee (FOMC), as well as of the Governing Council of the European Central Bank (ECB). Both institutions hold eight regular meetings per year. The related minutes contain detailed summaries of discussed topics in the meeting and are released three to four weeks after the corresponding meeting took place. Minutes are provided publicly by the *Federal Open Market Committee*[1] and the *Governing Council of the European Central Bank*[2].

Analogous to the paper of Jegadeesh and Wu (2015) the minutes of each meeting are divided into individual paragraphs which are assigned to distinct economic topics using a Latent Dirichlet Allocation (LDA) algorithm. Furthermore the net tone and uncertainty level of each topic and meeting are extracted. While the paper of Jegadeesh and Wu (2015) only considers a time frame from June 1991 until December 2012, this thesis extends the observation period to recent days. The Governing Council of the European Central Bank started to publish similar minutes on February 2015. Therefore, the database for the ECB is much smaller than the one of the FOMC.

In analyzing proportion, net tone and uncertainty of different topics in the minutes over time a qualitative view on the work of central banks is taken. The main idea behind this approach is to identify additional qualitative information being released in minutes besides quantitative data like federal funds target rates. Even if the minutes of each meeting are released about a few weeks after, Jegadeesh and Wu (2015) show that it still contains some additional information content based on the superior information of central banks. The analysis of the content of these soft data creates a certain transparency in the intentions and knowledge of central banks.

In section 2 the methodology of textual analysis using a Latent Dirichlet Allocation (LDA) algorithm is introduced, as well as the description of content extraction. Section 3 describes the implementation of the web scraping process and the data processing. Section 4 interprets the results of these approach. Section 5 concludes this thesis.

[1] Available at https://www.federalreserve.gov/monetarypolicy/fomccalendars.htm
[2] Available at https://www.ecb.europa.eu/press/accounts/html/index.en.html

2 Methodology

2.1 Textual analysis

Textual Analysis in its basic principles describes the process of exploring data in the form of written texts with the aim of gaining new insights in their structure and content. The process of discovering new patterns in and relationships between different texts are also called text mining. A huge problem in text mining is the so-called curse of dimensionality which refers to the challenge of analyzing high-dimensional data (Dietrich et al., 2015, p. 256). For this reason, different approaches and algorithms exists to deal with high-dimensional text data. In this thesis a Latent Dirichlet Allocation (LDA) algorithm is used for textual analysis. In another step, the tonality and uncertainty of the texts are examined.

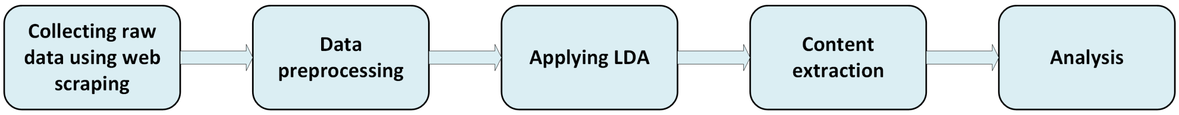


Figure 1: Textual analysis process of this thesis (source: own figure)

The entire textual analysis process applied in this thesis for analyzing the minutes published by the FOMC and ECB is shown in 1. The steps of web scraping and preprocessing are examined in more detail in sections 3.1 and 3.2. These steps produce the basic data set analyzed in this thesis. The following subsections in the current section cover the theoretical fundamentals of the LDA algorithm, and content extraction which refers to the tonality and uncertainty anaylsis. The analysis of the overall results is done in chapter 4.

2.2 Latent dirichlet allocation (LDA)

The Latent Dirichlet allocation (LDA) algorithm has been introduced by Blei, Ng and Jordan in 2003. The algorithm is a generative probabilistic model of a corpus and is used for natural language processing. A corpus is basically a collection of written texts. The components of such a corpus are called documents in the following. Each document itself consists of distinct paragraphs (Krestel et al., 2009, pp. 61-63).

The main idea behind the LDA is that an author of a text has a finite set of topics in mind. While writing about one of these topics the author picks particular words with a certain probability out of a bag of words. For each topic there is a distinct bag of words. When analyzing documents, the LDA groups the content expressed by different words into unobserved sets. Each of these sets represents one topic. Each document can be constituted by combining different topics and their word sets (Krestel et al., 2009, pp. 61-63).

In the LDA, documents are generally regarded as probability distributions over a finite set of topics. Each topic, on the other hand, is represented as a probability distribution over different words. Thus documents could be described as a mixture of the underlying topics (Blei et al., 2003, p. 993).

Both probability distributions are not observable but hidden properties. Under the assumption of Dirichlet distributions for both hidden distributions, the posterior distributions of the documents and topics can be calculated using the standard Bayes theorem. Using a newspaper as an example, a resulting topic from the LDA on *politics* should contain words like *government*, *parliament* and *president* with a high probability or weight, respectively. (Jegadeesh and Wu, 2015, pp. 9-10).

The only manually set parameter for the LDA is the number of different topics K. The number of different terms N, in this case referring to distinct words, is defined by the corpus D itself. The corpus is the only input the algorithm needs to calculate the probability distributions. A corpus contains M different documents $w_m, m \in \{1, ..., M\}$. For each topic $z_k, k \in \{1, ..., K\}$ and each distinct term $t_n, n \in \{1, ..., N\}$ a multinomial random variable $\beta_{n,k}$ is assigned.

The variable $\beta_{n,k} = p(t_n|z_k)$ represents the probability, that one specific term t_n will be used in a document about the topic z_k. Using the previous example of a newspaper article about *politics*, the probability of $\beta_{p,k} = p(government|politics)$ should be quite high while the value of $\beta_{q,k} = p(tournament|politics)$ should be very low, since tournament is not a common word when writing about politics.. This shows that the probability of the word *government* is much more likely in an article about *politics* than the word *tournament* (Blei et al., 2003, pp. 993-997) (Jegadeesh and Wu, 2015, pp. 9-10).

For each document w_m there is a K-dimensional vector $\theta_m \widehat{=} (\theta_{m,1}, \theta_{m,2}, ..., \theta_{m,K})$ defined which represents the topic mixture of this document. The goal of the LDA is to find these mixtures of topics $\theta_{m,k} = p(z_k|d_m)$ for each document (Krestel et al., 2009, pp. 62-63) (Blei et al., 2003, pp. 993-997). Assuming there are only three topics in a newspaper: *politics*, *sports* and *economy*. Analyzing an article (document) w_i about an election, the resulting vector θ could have the following form: $\theta_i = (\theta_{i,politics}, \theta_{i,sports}, \theta_{i,economy}) = (0.73, 0.01, 0.26)$. These values represent the posterior topic mixture of this specific document and show that the article is mainly about *politics*, but also includes some aspects of *economy* (Jegadeesh and Wu, 2015, p. 10).

The LDA model is being built by a generative process. The probability distributions for $\beta_{n,k} = p(t_n|z_k)$ and $\theta_{m,k} = p(z_k|w_d)$ are estimated by the LDA through an iterative sampling process. The resulted model can be used to estimate the proportions of different topics in one document or paragraph (Krestel et al., 2009, pp. 62-63). For a more detailed explanation of the generative process, please refer to the paper of Blei et al. (2003).

2.3 Content extraction

In addition to calculating the topic proportions, for each document a bag-of-words approach is used to determine the tone and uncertainty of each combination of documents and topics. Two dictionaries are used for tonal analysis: the *Harvard IV-4 Psychosociological Dictionary* and a *financial tonal list* developed by Loughran and McDonald[3]. To calculate the uncertainty of topics an *uncertainty words lexicon* also developed by Loughran and McDonald is used. For each paragraph p of a document w_m, the number of words occurring in the dictionaries are counted. Net tone is calculated as the difference between the frequency of positive and negative words, while uncertainty is defined as the frequency of words from the *uncertainty words lexicon.* This results in the variable $F_{p,c}^{w}$ which contains the net tone or number of uncertainty words for paragraph p in document w_m. The parameter $c = \{nettone, uncertainty\}$ specifies the type of content (Jegadeesh and Wu, 2015, pp. 16-17).

In the next step for each topic z_k of every paragraph $p \in \{1, ..., P_w\}$ in document w_m a content score for content c is calculated as $Score_{p,n,c}^{m} = \theta_{p,k}^{m} F_{p,c}^{m}$. After calculating all content scores on a paragraph level, the results are aggregated to document level. For this purpose, a mean value weighted by the inverse length of each paragraph is calculated as $Score_{k,c}^{w} = \sum_{p=1}^{P_w} Score_{p,n,c}^{m}(\frac{1}{T_p^k})$ where T_p^k represents the number of terms in paragraph p of document w_k. For better comparison between different documents the document level scores are standardized to a mean of zero and a standard deviation of one (Jegadeesh and Wu, 2015, pp. 17-18).

[3] Available at https://sraf.nd.edu/textual-analysis/resources

3 Implementation

3.1 Collecting raw data through web scraping

In order to obtain the meeting minutes of the FOMC[4] and the ECB[5], an automatic web scraping approach is used, whereby the data are downloaded in a processable format. Two separate Python applications are written for this task, one for each central bank's web site. With *Beautiful Soup*[6], there already exists a powerful library designed for web scraping and accessing specific content in HTML files. The library creates parse trees from the underlying HTML files, allowing the user to navigate and search the content in a structured and easy way. Prerequisites for this are well-structured HTML documents as an input. Additionally, Beautiful Soup converts HTML documents to Unicode so that the text data is prepared for further processing.

While the ECB minutes are built on a clean and constant structure over the observation period, the FOMC minutes pages are less constant in their structure. Therefore, a more complex solution is implemented for the FOMC protocol, covering some special cases and changes in the page structure. Generally, both applications should be able to also work with upcoming minutes releases, provided that the HTML documents retain their structure. The result of each scraping process is a simple .txt document containing the text data of one meeting.

3.2 Data preprocessing

After scraping the content of the minutes, some pre-processing steps are applied. Analogous to the paper of Jegadeesh and Wu (2015), the administrative details and reviews of previous open market operations are removed from each document. After that, each document is split into individual paragraphs. Since there are no precise information on splitting rules in the paper, paragraphs are separated by blank lines. In the reported period (June 1991 - December 2012), Jegadeesh and Wu produced a total of 4,784 paragraphs with an average sentence length of 28 words for the FOMC minutes. In the same period, the implementation of this study generates 5,219 paragraphs with an average sentence length of 28 words. For the entire period (February 1990 - June 2018) a total of 7,345 paragraphs are produced (Jegadeesh and Wu, 2015, pp 5-6).

[4] Available at https://www.federalreserve.gov/monetarypolicy/fomc_historical_year.htm

[5] Available at https://www.ecb.europa.eu/press/accounts/2018/html/index.en.html

[6] Available at https://www.crummy.com/software/BeautifulSoup/bs4/doc

For the purpose of pre-processing the ECB minutes, each document is split into single paragraphs. As before, the administrative details are removed. This results in a total of 1,869 paragraphs with an average sentence length of about 29. It should be noted that the minutes are only available since February 2015. The remarkably high number of paragraphs for this short period results from the fact that the minutes of the ECB are very detailed, while the FOMC minutes have very little content in early years.

3.3 Application of LDA and content extraction

After extracting individual paragraphs from each document, two separate LDA models are trained on the data sets. The first model is trained on the FOMC minutes. Analogous to the paper of Jegadeesh and Wu (2015), only FOMC data in the period of June 1991 until December 2012 are used for training the first model. A second model is trained with all available ECB data from February 2015 until October 2018. From both data sets the texts are tokenized, which leads to the fact that each sentence is broken down into its individual terms. Using a Porter Stemmer algorithm[7], every word is transformed into its root form. In this step the commoner morphological and inflexional endings of words are removed. Then stop words and the 100 most common words are removed from each data set. (Jegadeesh and Wu, 2015, pp. 10-11).

In the next step, a dictionary is created for each of the two models, which consists of the collection of the remaining stems. These dictionaries are mappings between words and their individual integer ID. From the FOMC dictionary all elements which occur in less than 200 paragraphs or in more than 60% of the paragraphs are removed, as they are not expected to contain much information. For the ECB dictionary, an appearance in at least 30 paragraphs is set as a lower bound threshold while the upper threshold is set at 70%. With these resulting dictionaries, each model is trained on the paragraphs of one central bank. The models are calculated by using the *gensim* Python library[8]. When using the LDA, the number of topics must be specified manually. This parameter is set to eight topics for both models.

[7] Available at https://tartarus.org/martin/PorterStemmer

[8] Available at https://radimrehurek.com/gensim/models/ldamodel.html

Fed Financial Markets Topic

Weight	Word	Weight	Word
0.032	credit	0.022	spread
0.032	loan	0.019	debt
0.025	yield	0.019	dollar
0.023	secur	0.018	equiti
0.023	treasuri	0.017	bond

ECB Financial Markets Topic

Weight	Word	Weight	Word
0.042	bond	0.016	increa
0.028	yield	0.015	declin
0.020	financ	0.015	purcha
0.018	sinc	0.014	equiti
0.016	meet	0.013	curv

Table 1: Distribution of Top LDA Financial Markets Topic Keywords for Fed and ECB (source: own table based on Jegadeesh and Wu (2015))

As a result, the LDA algorithm creates eight topic distributions. For both models, Table 1 shows the top ten keywords for one of their eight topics. This topic has manually been identified as *Financial Markets*. Each weight represents a $\beta_{n,k}$ describing the probability that a keyword t_n characterizes a specific topic z_k (see also section 2.2). The top 20 keywords for each of the eight topics are shown in appendix A (Fed) and D (ECB), respectively (Jegadeesh and Wu, 2015, pp. 10-11, 13-14).

Analogue to the work of Jegadeesh and Wu (2015), the eight topics from the FOMC model are grouped by hand into four mandates: *Growth*, *Inflation*, *Financial Markets* and *Policy*. For the ECB model (appendix D) the same steps are performed. A more detailed description of this assignment process can be found in section 4.1.

After that, for each document the proportion of each mandate is calculated. These data are then plotted over the respective observation period in appendix B (Fed) and E (ECB). In a next step, the net tones and uncertainty scores are calculated and standardized as described in section 2.3. For computing the net tone and uncertainty, the Python library *Pysenment* is used[9]. The results are also plotted over the observation periods applying a 10-period moving average. The plots are shown in appendix C (Fed) and F (ECB).

[9] Available at https://pypi.org/project/pysentiment

4 Interpretation

4.1 LDA topic distribution

As described in section 3.3, the application of the LDA algorithms produces 8 distinct topics for the FOMC and ECB minutes. Tables A (Fed) and D (ECB) in the appendix show the posterior vocabulary distributions for each topic. Unlike Jegadeesh and Wu (2015), the resulting keywords are not replaced in the table by their original form before stemming. This is because in some cases the original form is ambigious. For example, the root form *secur* can result from both, the verb *secure* and the subject *security.* Therefore, in some cases a replacement seems misleading.

By analyzing the components of the individual abstract themes of the LDA, some of them can be clearly identified with real economic themes. Topic 2 in table A, for example, consists of keywords such as *energy*, *core* and *cost.* This suggests that the topic can be assigned to the *Inflation Mandate* of the Fed. In the same way, the *financial market mandate* is unambiguously represented by topic 3 consisting of keywords like *credit*, *loan* and *yield.* The topics are assigned directly to the four Fed mandates *Growth*, *Inflation*, *Financial Markets* and *Policy.* As already mentioned, topic 2 is assigned to *Inflation* and topic 3 to *Financial Markets.* Similarly, topic 5 is assigned to *Policy* and the remaining topics to *Growth* (Jegadeesh and Wu, 2015, pp. 14, 38).

In order to be consistent with the Fed model, an attempt is made to assign the resulting ECB topic keywords into the same four mandates. Again, the mandates *Inflation* and *Financial Markets* are identified very clearly and consist of topic 2 (*Inflation*) and the topics 3 and 4 (*Financial Markets*). Further, topic 5 is assigned to *Policy* and the remaining topics to *Growth.* It is striking that the weights of the top keywords for each distinct topic are generally much higher in the ECB model than the Fed model. This could be an indicator that the ECB minutes are thematically more structured than the FOMC minutes. However, it cannot be ruled out that this phenomenon is due to the much smaller database of the ECB minutes compared with the FOMC dataset.

4.2 Topic proportions over time

The FOMC and ECB topic proportions over time are shown in appendices B and E. Figure 4 shows the 1-period moving average of the FOMC topic time series, while the figures 5 and 6 show the 3-period and 10-period moving average, respectively. In Addition, the dot-com crisis and the financial crisis of 2007/2008 are stated in the figures as shaded areas. The time frames of both recession periods are analogous to the National Bureau of Economic Research[10]. For a better comparison with the study of Jegadeesh and Wu (2015), the beginning and end of the period under review of their studies is marked with black vertical lines.

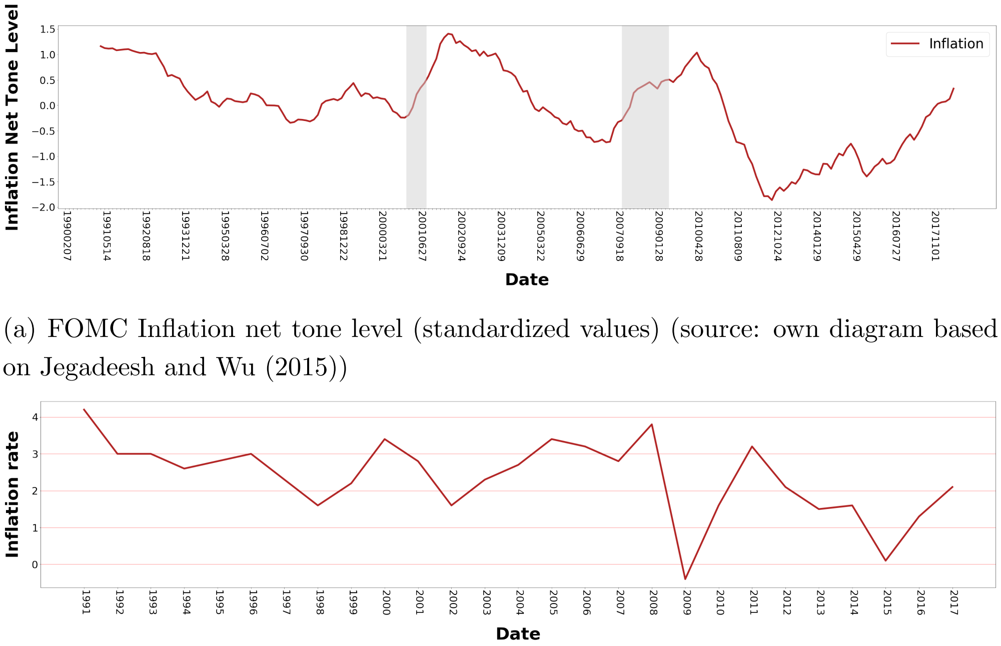
The figures of the FOMC topics clearly show that the proportions of the four main topics are changing over time. While the *Growth* mandate made up the largest part at the end of the 20th century with at around 70%, its proportion declines to about 40% in the current days. During the last financial crisis, the FOMC focused more strongly on the *Financial Markets* mandate. Its proportions has doubled over time from about 10% before the crisis of 2007/2008 to 20% recently and peaked at around 30% during the financial crisis. This change of topic can be explained by the instability of the financial markets during the increasing banking crisis and the necessary intervention of the Fed. Surprisingly, the dot-com crisis did not have any strong impact on the topic proportions. After the end of the crisis, only a slight decline of the *Growth* mandate is discernible (Jegadeesh and Wu, 2015, p. 15).

For the ECB minutes, the topic proportions are shown in figure 9 with a 1-period and in figure 10 with a 3-period moving average. As only data from 30 meetings are available to date, a 10-period moving average is omitted. In contrast to the FOMC minutes, the topic proportions of the ecb minutes show no major changes in the period under review. At least there seems to be a trend to increasing proportions of the topics *Growth* and *Financial Markets*, while the *Inflation* topic is losing importance. This could be a consequence of rising inflation, which reduces the risk of deflation, and falling stock prices in recent months. However, more data is needed to make a reliable statement.

[10] Available at https://www.nber.org/cycles.html

4.3 Topic net tones and uncertainty over time

The 10-period moving average of the standardized document-level net tone of each topic is shown in figure 7 for the FOMC minutes and in figure 11 for the ECB minutes. Evaluating the FOMC results, it stands out that the net tone of the *Financial Markets* topic strongly increases during the last financial crisis. The reason for that could be the Fed using their instruments to stabilize and calm the capital markets with a more positive language. Therefore, the *Financial Markets* mandate appears to be countercyclical (Jegadeesh and Wu, 2015, p. 10). The net tone of the *Growth* mandate, on the other hand, tends to stay neutral in recent years. On the contrary, the net tone of the *Policy* has become more positive in the recent years, while it was fairly neutral in the years before 2012. This is likely because the monetary policy instruments which were used, had a good impact on the markets. Between the net tone of the *Inflation* topic and the U.S. inflation rate[11] a strong correlation is visible. For this purpose both time series are shown in figure 2. Especially in the years 2010 and 2015 it is noticeable that the inflation rate is far away from the target rate of 2% and that the tonality in the minutes is correspondingly negative.



(a) FOMC Inflation net tone level (standardized values) (source: own diagram based on Jegadeesh and Wu (2015))

(b) Unadjusted annual inflation rate (CPI) in the U.S. (source: own diagram based on data from the Bureau of Labor Statistics)

Figure 2: Comparison of the FOMC inflation net tone level and the U.S. inflation rate

[11] Data available at https://www.bls.gov/cpi/

Evaluating the net tones of the ECB minutes shown in figure 11, it appears that the net tone of the topics *Growth*, *Financial Markets* and *Inflation* increased significantly for each of them over the period under review. The main reason for the more positive tone of the *Inflation* topic is probably the increasing inflation rate from a negative rate back in 2015 up to about 2% in recent months[12]. Since the ECB was driven by fears of a deflation in the euro area, it is plausible that the increasing inflation rate reflects itself in the tonality. Economic growth of the euro zone increased in the period under review. Real GDP in the euro zone grew from 1.6% in Q2 2016 up to 2.8 % in Q4 2017 [13]. Similarly, the net tone of *Growth* changed from a neutral to more positive. A graphical comparison between the net tone of these two topics, and the inflation rates and GDP growth in the Euro area is shown in figure 3.

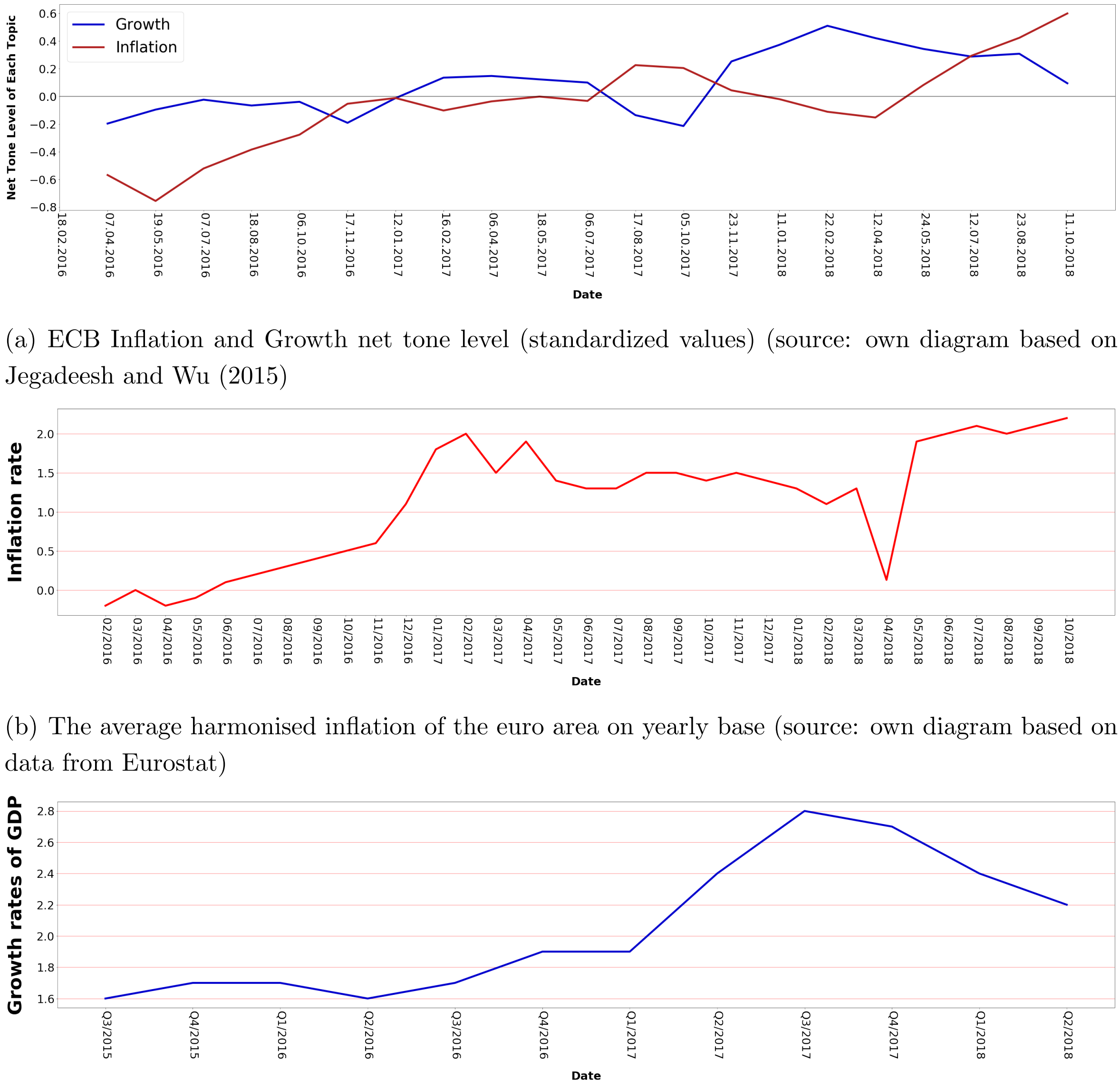
The tonality regarding *Financial Markets* improved sharply in the period under review. Taking the EuroStoxx50[14] as an indicator, its decline below 3,000 points at the beginning of 2016 is reflected in the ECB's tonality. At that point, the tonality of topic was very negative compared to other times. In the following months, the index rose to 3,600 points, while net tone rose as well. In contrast to the other topics, net tone of the *Policy* topic decreased from a very positive view in 2016 down to a negative view in 2018. A reason for this could be a louder criticism of the ECB's quantitative easing program and a stronger internal debate about it.

Looking at the uncertainty scores of the topics shown in figures 8 (FOMC) and 12 (ECB), both central banks became more certain about most of their topics in recent years. But in recent periods, both institutions have seen a decline in certainty. They became both more uncertain about the topic *Inflation* in the last periods. The FOMC plot shows a decline in certainty about the *Financial Markets* since 2012. A possible reason could be the increasing risks of speculative bubbles and the difficult exit from the low-interest phase.

[12] Data available at https://ec.europa.eu/eurostat/de/web/hicp/publications/news-releases

[13] Data available at https://ec.europa.eu/eurostat/de/web/national-accounts/publications/news-releases

[14] Data available at https://www.stoxx.com/indices



(a) ECB Inflation and Growth net tone level (standardized values) (source: own diagram based on Jegadeesh and Wu (2015)

(b) The average harmonised inflation of the euro area on yearly base (source: own diagram based on data from Eurostat)

(c) Growth in real gross domestic product in the euro area compared with the prior-year quarter in percent (source: own diagram based on data from Eurostat)

Figure 3: Comparison of the ECB *Inflation* and *Growth* net tone level, the inflation rate and the real GDP of the euro area

5 Conclusion

This thesis reproduces the results obtained by Jegadeesh and Wu (2015) regarding their work on analyzing the topic proportions of FOMC meeting minutes. In addition, this thesis takes also ECB meeting minutes into account and extends the period under review until current meetings. As in the original paper, an automated approach is used for scrapping the minutes from the FOMC's and ECB's web sites. For both data sets, a Latent Dirichlet Allocation (LDA) algorithm is used to identify different topics covered in these meetings. For each meeting, the proportions of the topics are calculated as well as the tonality and uncertainty of these topics. When these results are compared with other data such as stock markets or GDP growth rates, correlations are clearly visible. Although the results of this thesis are not entirely identical to the results of Jegadeesh and Wu (2015), this work is able to reproduce the key points of their work and validate its results in a practical way.

The written algorithms and programs for these tasks can not only used to rebuild the presented results, but also provide the ability to set user-defined time frames for review periods and will also work with future publications of meeting minutes[15]. Therefore, it will be interesting to see how current global political trends such as looming trade wars, declining stock markets and growing concerns about highly indebted countries will influence the content of the minutes. Another interesting part to be analyzed would be the changes in central bank boards and possible reflections in the tonality of minutes.

In addition to FOMC and ECB meeting minutes, this approach could be extended to other central banks or important institutions such as the International Monetary Fund or the Bank of England. An important prerequisite for the proper functioning of this approach is a clear structure of the documents being processed.

Another interesting modification of this approach is to use algorithms other than the LDA for clustering the topics. For example, there are popular algorithms such as *K-Means* or *Latent semantic analysis* to analyze and identify different topics. At the same time, consideration could also be given to the definition of other overarching topics like employment or more specific topics like cryptocurrencies.

[15] As long as there are no major changes in the structure of the HTML documents.

A FOMC: Distribution of Top LDA Topic Keywords

Topic 1		Topic 2		Topic 3		Topic 4	
Weight	Word	Weight	Word	Weight	Word	Weight	Word
0.063	project	0.035	energi	0.032	credit	0.077	system
0.057	staff	0.035	employ	0.032	loan	0.054	open
0.050	forecast	0.030	core	0.025	yield	0.049	transact
0.024	gdp	0.029	unemploy	0.023	secur	0.048	secur
0.023	anticip	0.017	averag	0.023	treasuri	0.045	account
0.021	direct	0.016	cost	0.022	spread	0.041	domest
0.019	long-run	0.015	food	0.019	debt	0.040	oper
0.018	dure	0.014	rose	0.019	dollar	0.040	manag
0.018	slightli	0.014	payrol	0.018	equiti	0.039	direct
0.017	consist	0.014	compen	0.017	bond	0.038	vote

Topic 5		Topic 6		Topic 7		Topic 8	
Weight	Word	Weight	Word	Weight	Word	Weight	Word
0.020	agr	0.017	expan	0.040	hou	0.028	equip
0.019	view	0.016	invest	0.039	export	0.027	vehicl
0.017	discuss	0.015	household	0.037	home	0.026	motor
0.016	pressur	0.015	effect	0.032	import	0.026	inventori
0.016	possibl	0.015	mani	0.028	u.s	0.025	manufactur
0.015	current	0.013	factor	0.022	mortgag	0.023	industri
0.015	might	0.012	incom	0.021	trade	0.016	output
0.013	accommod	0.01	improv	0.020	good	0.015	good
0.013	point	0.011	firm	0.019	start	0.014	first
0.013	need	0.011	hou	0.018	new	0.014	gain

Table 2: Distribution of Top LDA Topic Keywords of the Fed Federal Open Market Committee minutes (source: own table based on Jegadeesh and Wu (2015))

B FOMC: Topic Proportions Over Time

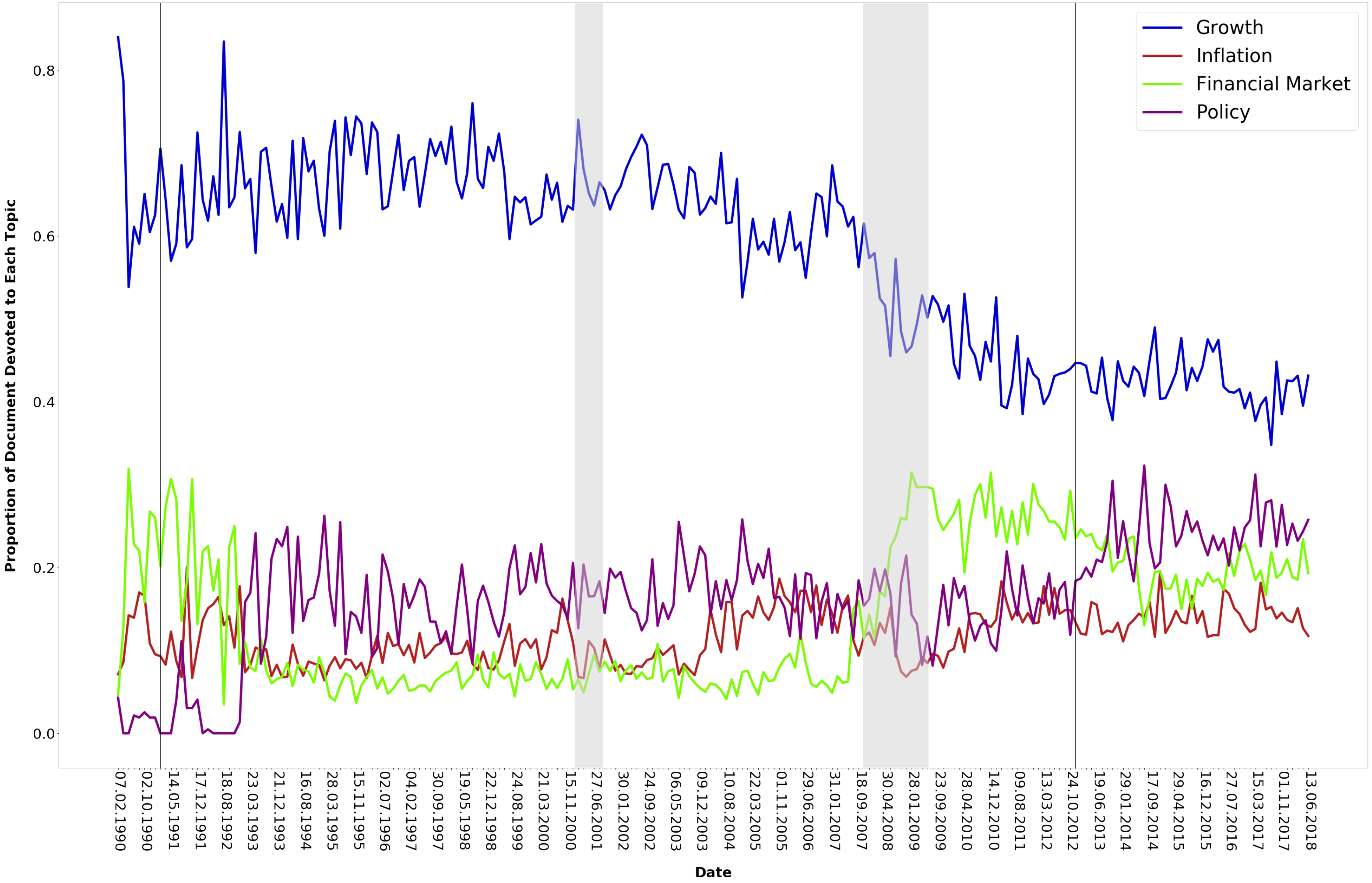


Figure 4: Time series of the 1-period moving average of the proportion of each FOMC topic (source: own diagram based on Jegadeesh and Wu (2015))

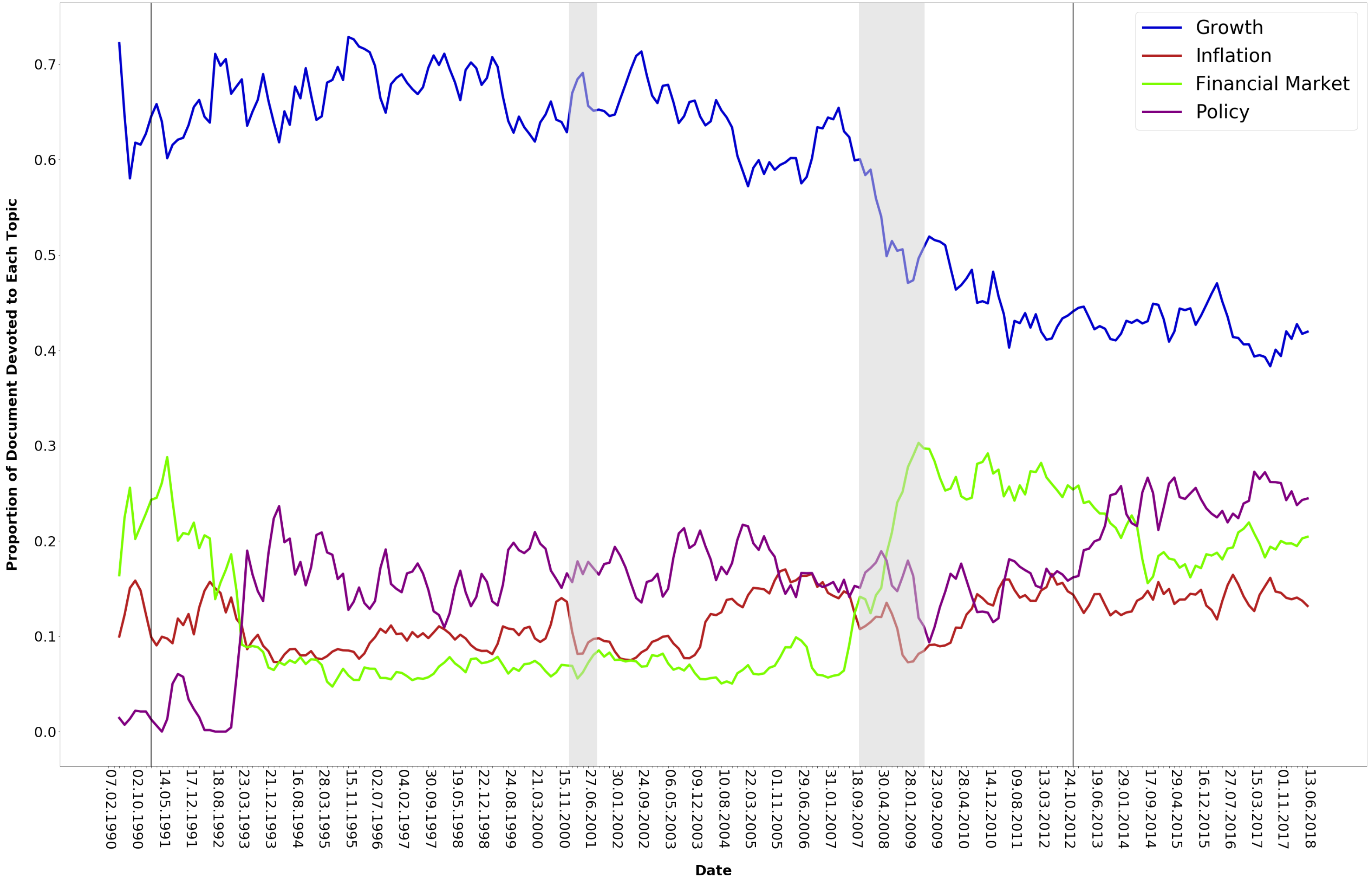


Figure 5: Time series of the 3-period moving average of the proportion of each FOMC topic (source: own diagram based on Jegadeesh and Wu (2015))

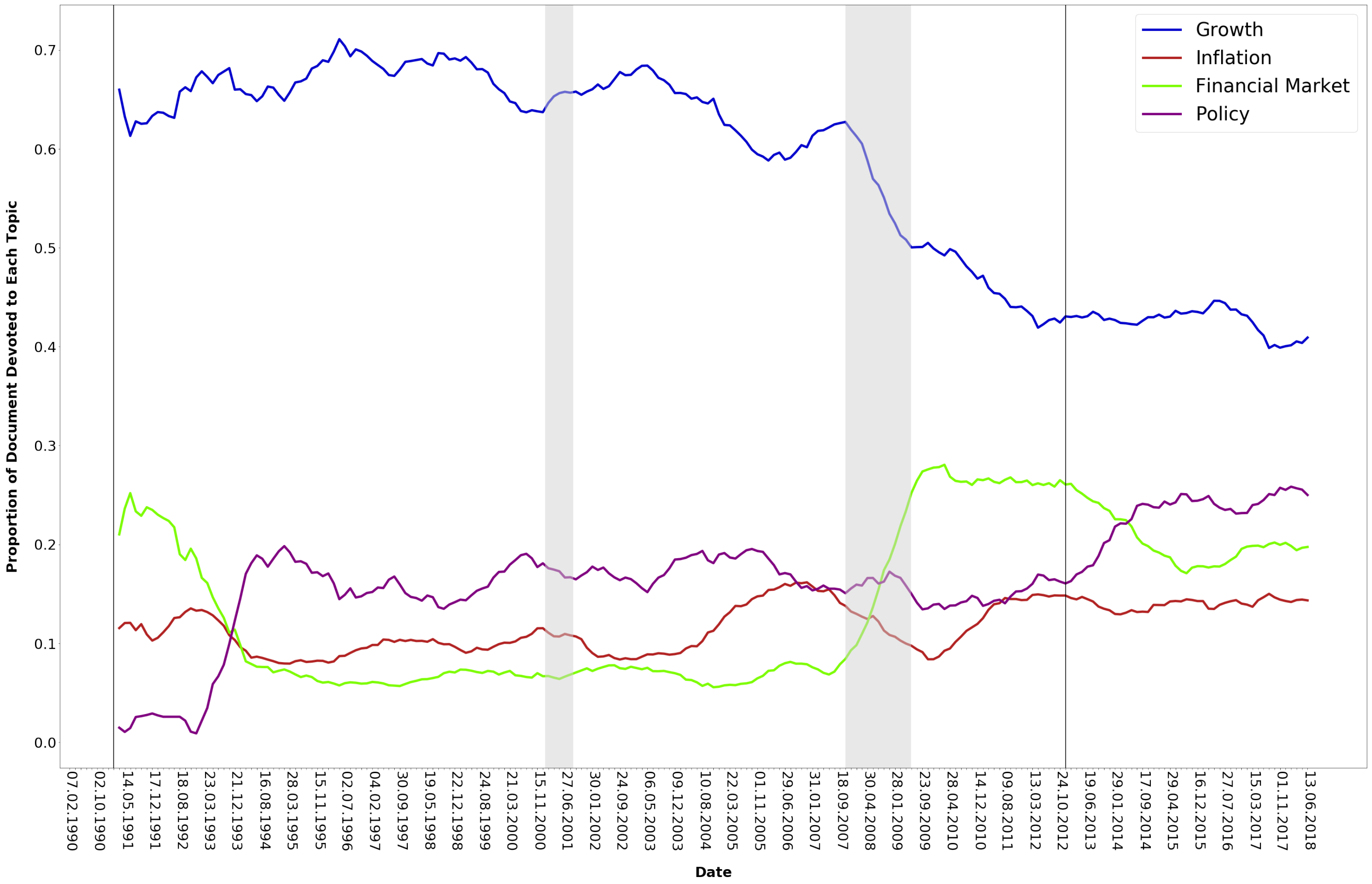


Figure 6: Time series of the 10-period moving average of the proportion of each FOMC topic (source: own diagram based on Jegadeesh and Wu (2015))

C FOMC: Topic Content Scores Over Time

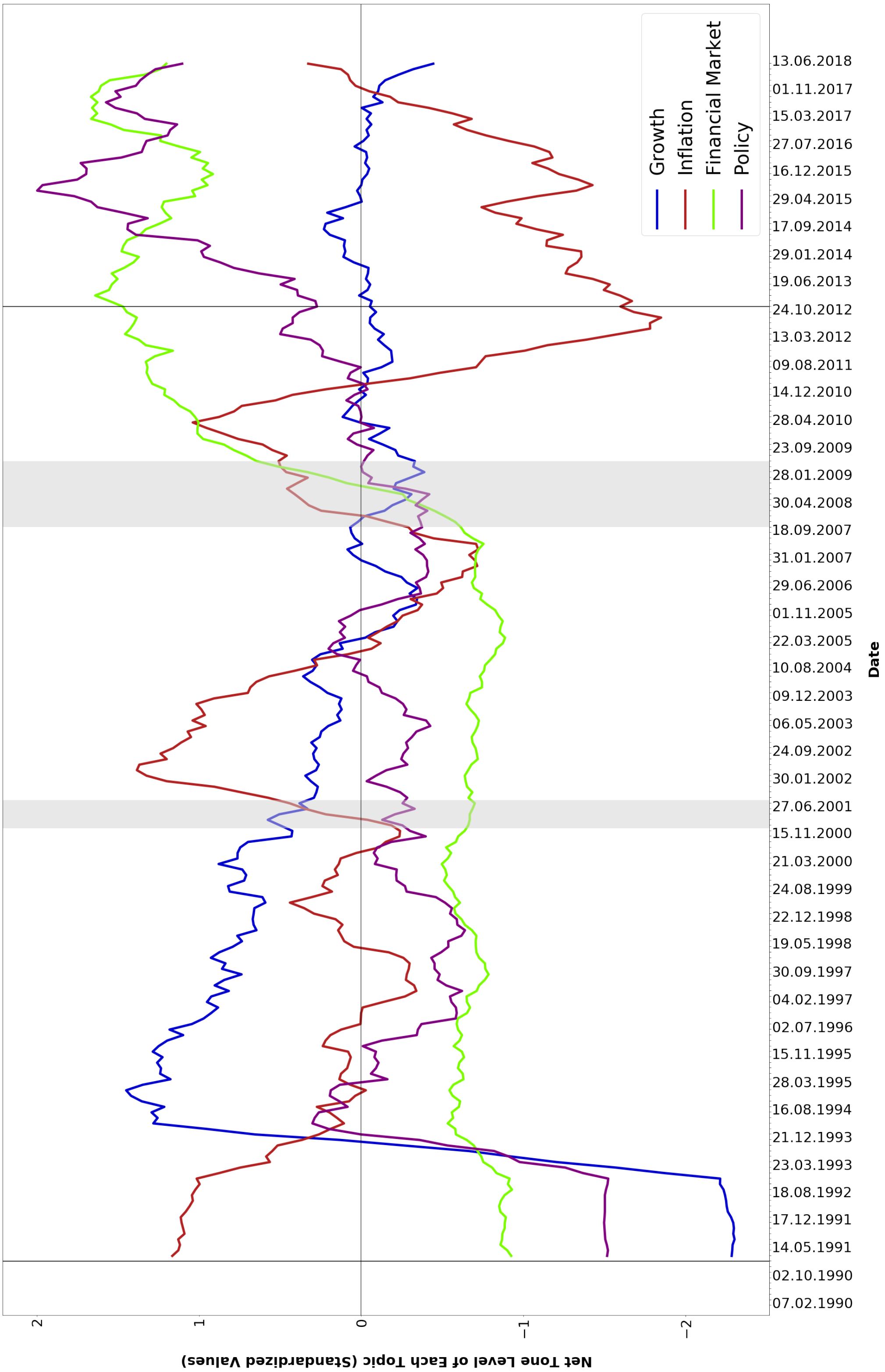


Figure 7: Time series of the 10-period moving average of the document-level Net Tone Scores of each of the four FOMC topics (source: own diagram based on Jegadeesh and Wu (2015))

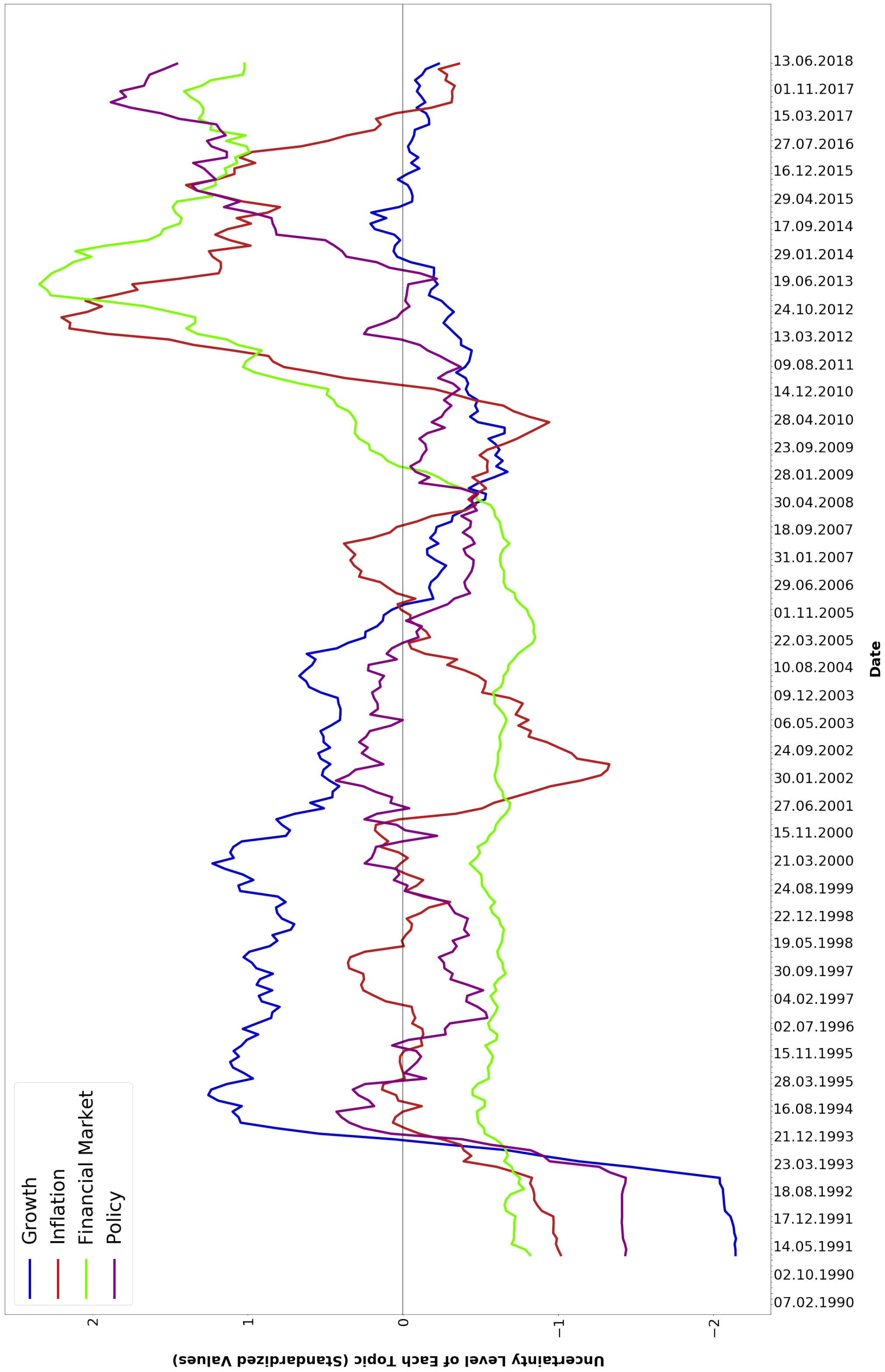


Figure 8: Time series of the 10-period moving average of the document-level Uncertainty Scores of each of the four FOMC topics (source: own diagram based on Jegadeesh and Wu (2015))

D ECB: Distribution of Top LDA Topic Keywords

Topic 1		Topic 2		Topic 3		Topic 4	
Weight	Word	Weight	Word	Weight	Word	Weight	Word
0.058	purcha	0.084	price	0.034	bank	0.048	market
0.030	would	0.030	energi	0.032	loan	0.028	bond
0.030	asset	0.028	oil	0.028	credit	0.027	expect
0.029	govern	0.021	hicp	0.027	financ	0.026	rate
0.025	app	0.019	remain	0.026	condit	0.019	euro
0.024	rate	0.018	annual	0.023	growth	0.018	yield
0.020	council	0.018	food	0.023	continu	0.017	sinc
0.017	net	0.017	rate	0.021	lend	0.015	govern
0.016	end	0.017	increa	0.019	rate	0.015	meet
0.015	polici	0.015	develop	0.019	remain	0.013	area

Topic 5		Topic 6		Topic 7		Topic 8	
Weight	Word	Weight	Word	Weight	Word	Weight	Word
0.051	polici	0.034	risk	0.056	project	0.049	quarter
0.041	monetari	0.027	euro	0.034	staff	0.041	growth
0.024	govern	0.022	area	0.028	expect	0.024	area
0.018	need	0.020	could	0.026	outlook	0.022	euro
0.017	measur	0.019	also	0.021	econom	0.021	invest
0.016	council	0.018	market	0.020	growth	0.017	continu
0.012	would	0.018	outlook	0.019	euro	0.014	indic
0.011	stanc	0.017	growth	0.018	area	0.013	global
0.011	thi	0.016	global	0.016	revi	0.013	first
0.010	econom	0.015	economi	0.015	assess	0.013	remain

Table 3: Distribution of Top LDA Topic Keywords of the ECB Governing Council minutes (source: own table based on Jegadeesh and Wu (2015))

E ECB: Topic Proportions Over Time

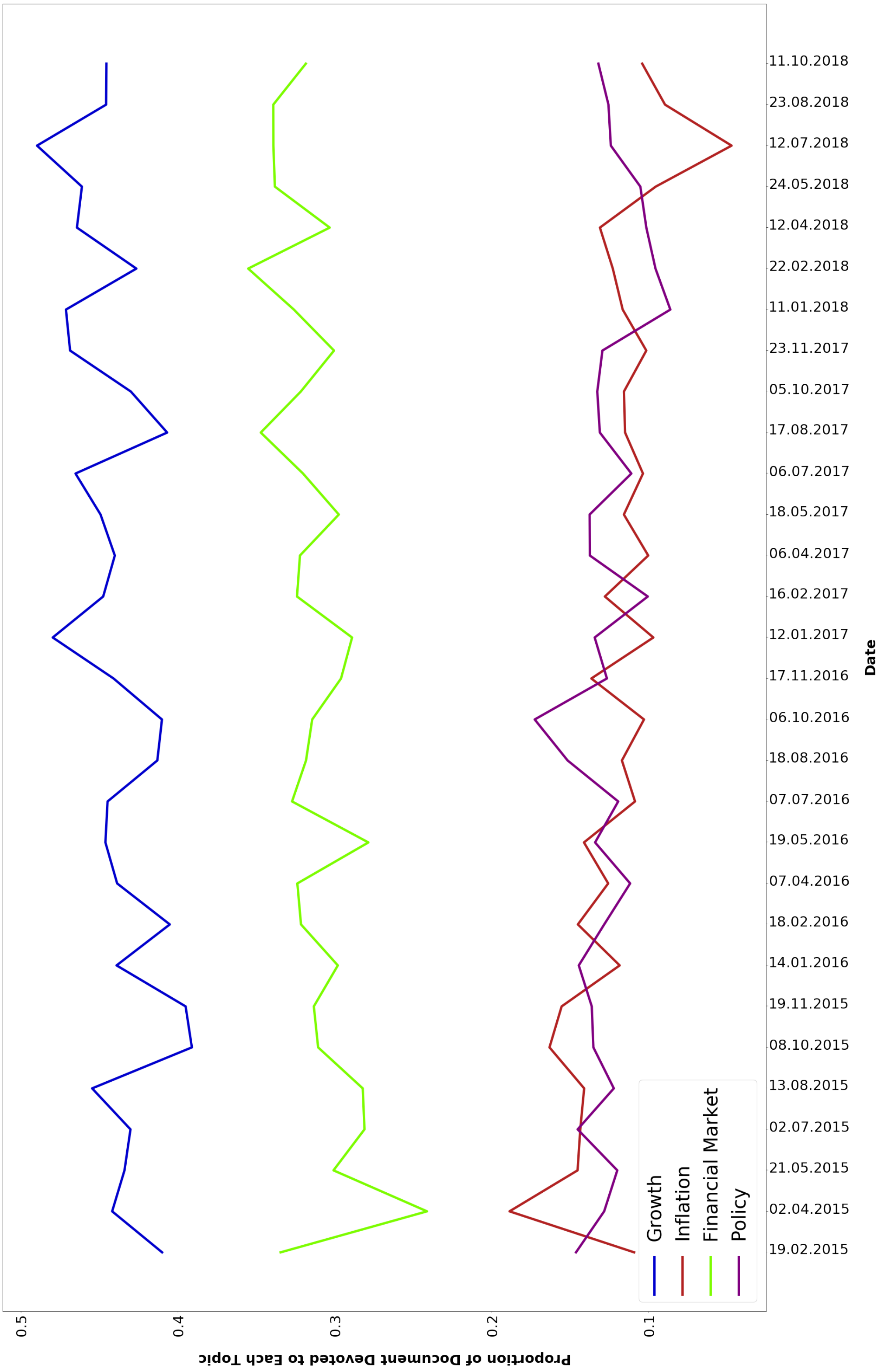


Figure 9: Time series of the 1-period moving average of the proportion of each ECB Governing Council topic (source: own diagram based on Jegadeesh and Wu (2015))

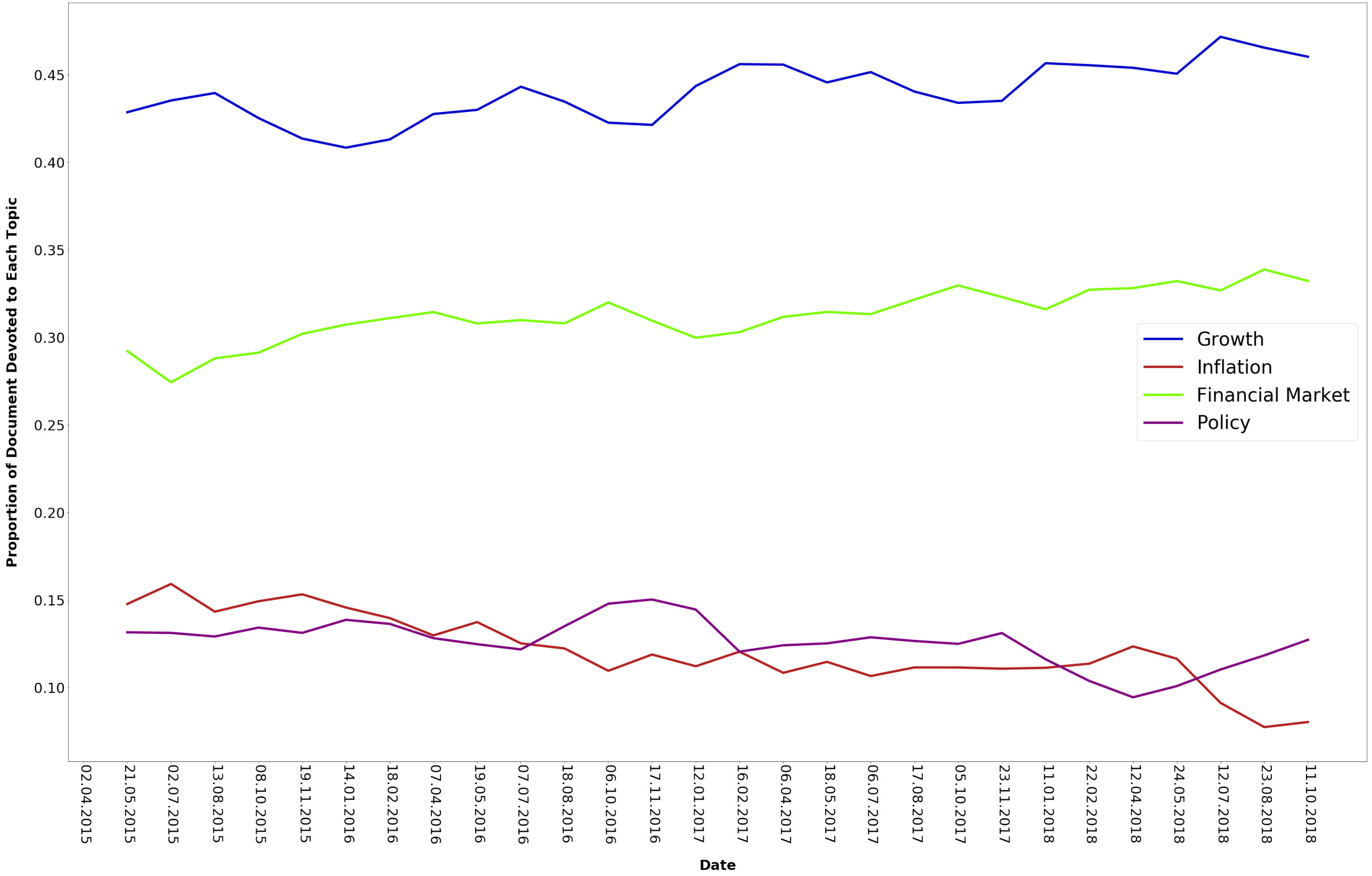


Figure 10: Time series of the 3-period moving average of the proportion of each ECB Governing Council topic (source: own diagram based on Jegadeesh and Wu (2015))

F ECB: Topic Content Scores Over Time

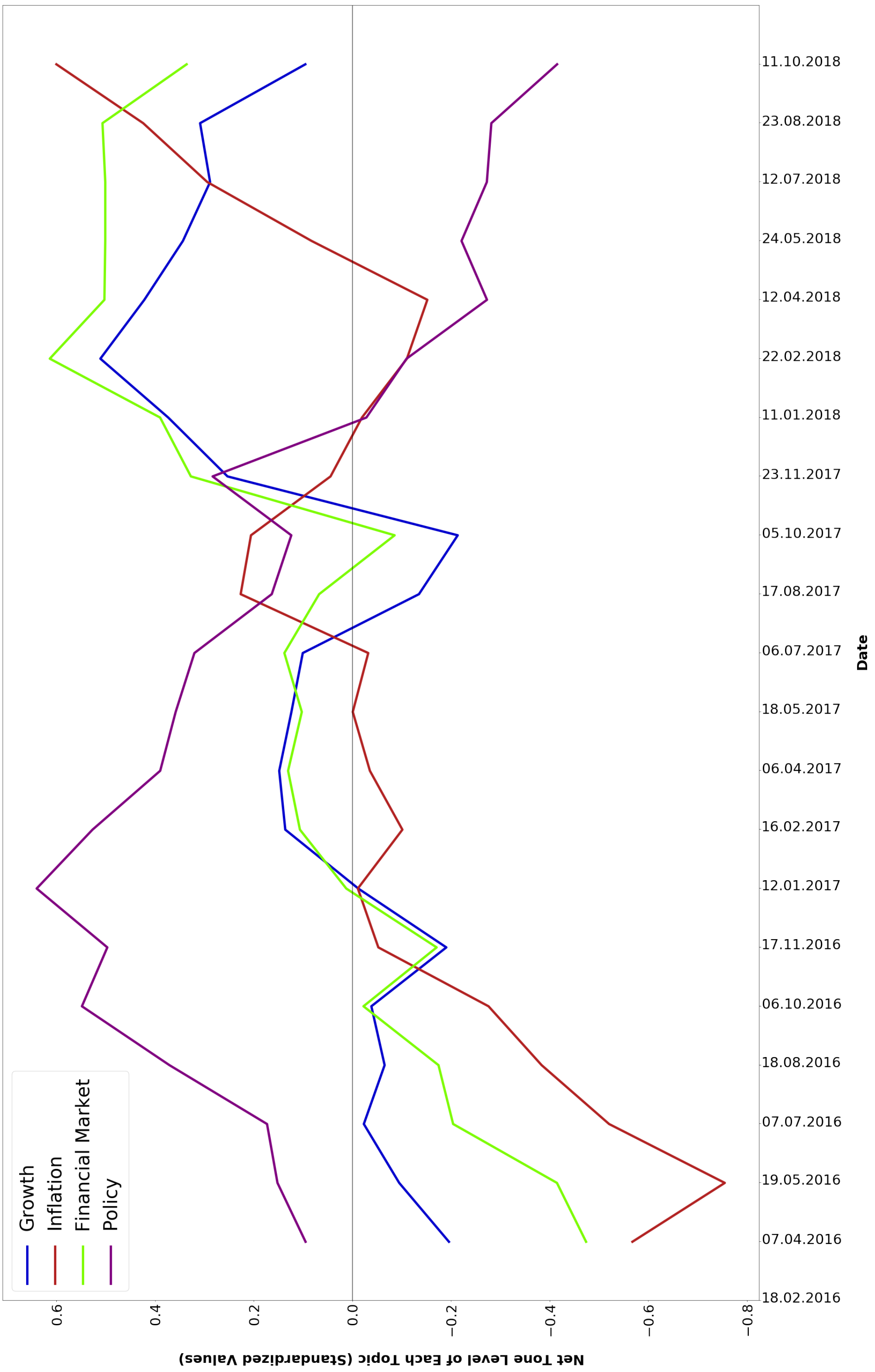


Figure 11: Time series of the 10-period moving average of the document-level Net Tone Scores of each of the four ECB Governing Council topics (source: own diagram based on Jegadeesh and Wu (2015))

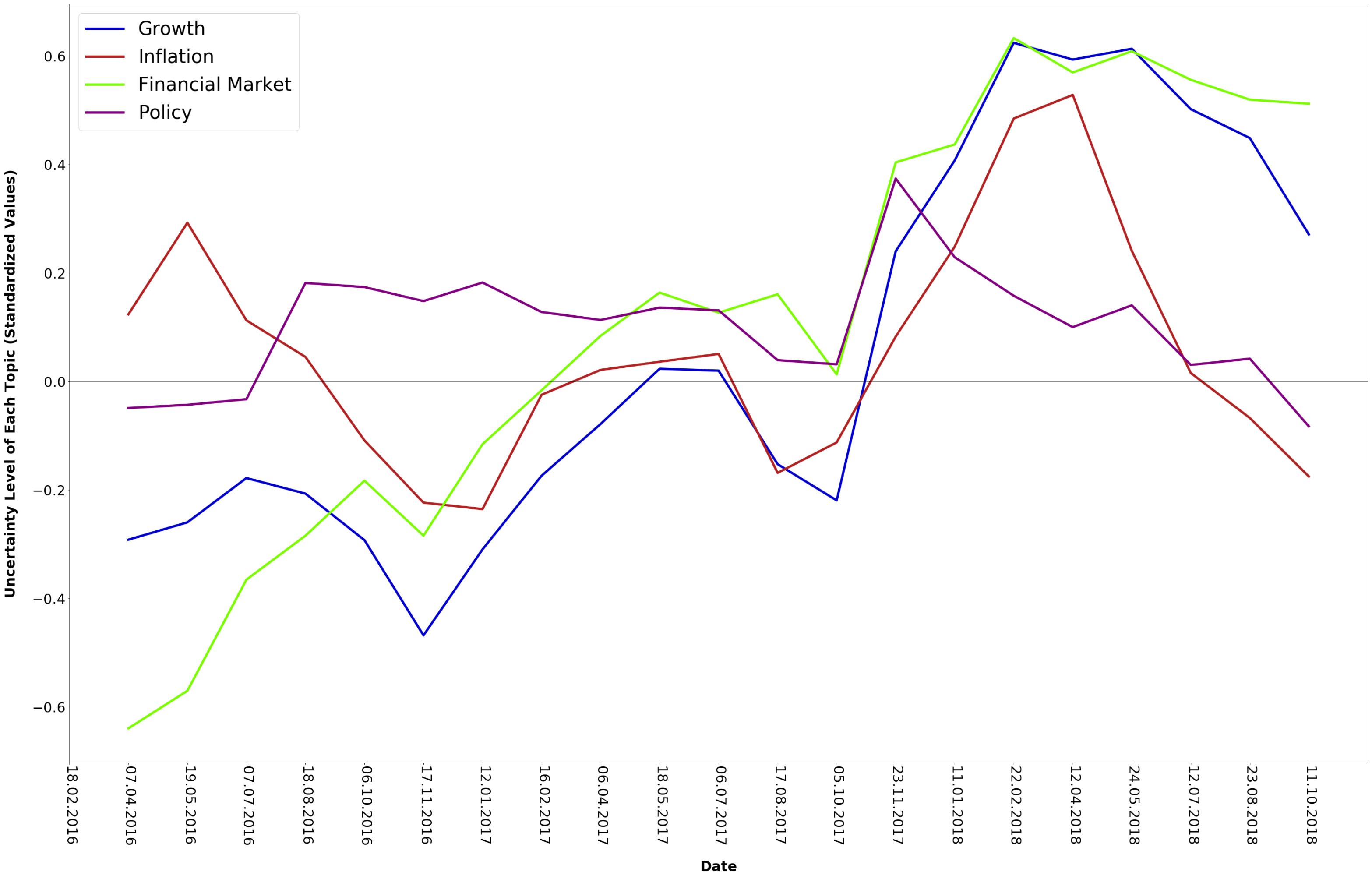


Figure 12: Time series of the 10-period moving average of the document-level Uncertainty Scores of each of the four ECB Governing Council topics (source: own diagram based on Jegadeesh and Wu (2015))

References

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

Dietrich, D., Heller, B., and Yang, B. (2015). *Data science & big data analytics: Discovering, analyzing, visualizing and presenting data.* John Wiley & Sons, Indianapolis, IN.

Jegadeesh, N. and Wu, D. (2015). Deciphering fedspeak: The information content of fomc meetings.

Krestel, R., Frankhauser, P., and Nejdl, W. (2009). Latent dirichlet allocation for tag recommendation.