The ECB Announcement Returns

Zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften
(Dr. rer. pol.)
von der KIT-Fakultät für
Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte
DISSERTATION

von
M.Sc. Elmar Jakobs

Referent:	Prof. Dr. Maxim Ulrich
Korreferent:	Prof. Dr. Martin E. Ruckes
Tag der mündlichen Prüfung:	18. Februar 2020

Karlsruhe, im Februar 2020

Acknowledgements

I would like to thank my supervisor Prof. Dr. Maxim Ulrich for his excellent guidance and continuous support as well as inspiring discussions throughout my doctoral studies. Further, I am grateful to Prof. Dr. Martin E. Ruckes for serving on my thesis committee. I would like to thank Prof. Dr. York Sure-Vetter and Prof. Dr. Clemens Puppe, who were part of the examination committee.

I am grateful to all my colleagues at the Institute of Finance, Banking, and Insurance. In particular, I thank Stephan Florig, Simon Walther, Lingjie Ni, Lukas Zimmer, Ralph Seehuber and Tobias Kuhlmann for many stimulating conversations and for making the past years such an enjoyable time. I also thank the C-RAM students for their contributions to our research projects.

I gratefully acknowledge the financial support from my PhD scholarship provided by the DZ BANK-Stiftung.

Finally and most importantly, I would like to thank my family and friends. In particular, I would like to thank my parents for their unlimited support throughout my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Recent research shows that 60% to 80% of the annual equity risk premium can be earned on just a few days per year (see Savor and Wilson, 2013; Lucca and Moench, 2015). These trading days are known to investors well in advance since on these days important macroeconomic news about inflation, unemployment, or interest rates is scheduled for announcement. In contrast, holding stocks on all other days does not yield a return that is statistically different from zero.

This pattern is particularly strong for days on which members of the Federal Open Market Committee (FOMC), the Federal Reserve's monetary policy-making body, convene. For these scheduled FOMC announcements, which take place eight times per year, the literature has documented large excess returns: Since 1994, the S&P 500 index has on average increased by about 0.5% in the 24 hours before these meetings, accounting for about 80% of the annual excess returns in U.S. equity markets (see Lucca and Moench, 2015).[1] A simple trading strategy of buying the index one day before the meeting and holding it until the announcement would have yielded an annualized Sharpe ratio of over 1.1. In general, this pre-FOMC announcement drift can be observed across various industries and is present also in major international stock indices like the German DAX index. For other major central banks the authors do not find a similar

[1]Note, their sample period ended in March 2011.

return pattern.

Brusa et al. (2019) argue that these findings suggest ‘a leading role for the Fed among central banks’ in terms of setting monetary policy. But with the beginning of the European sovereign debt crisis in 2010, the relationship between the two major central banks U.S. Federal Reserve and European Central Bank has changed. Whereas the economy in the U.S. started to recover after the global financial crisis of 2007-2009, in Europe a debt crisis evolved as a consequence of the Great Recession. These different economic regimes also resulted in different monetary policies: In the U.S. the Federal Reserve (Fed) started to raise interest rates again whereas in Europe, the ECB first lowered rates until the zero lower bound and then announced further unconventional monetary policy measures, like refinancing operations or asset purchase programs. Thus, the first time in the history of the European Central Bank[2] the course of these two major central banks fundamentally diverged leading to an increased importance of the ECB and an emancipation from the Fed.

In this thesis, we now want to analyze whether this new increased importance of the European Central Bank is reflected in the financial markets, in particular in a return pattern around the ECB’s monetary policy decision days similar to the pre-FOMC announcement drift. And if we can observe such an ECB announcement return then we want to understand what factors are determining this pattern. Particularly, we want to examine how these returns are related to the ECB’s important role as a lender or buyer of last resort (see Acharya et al., 2018) during the high uncertainty periods of the euro crisis, when the concerns about a eurozone break-up were quite severe. Specifically, we want to analyze whether expectations and hopes from market participants about the ECB’s supportive actions, the actual (realized) announcement of unconventional monetary policy measures, or the central bank’s communication in general generate the observed pattern in equity prices around the disclosure of the monetary policy decision.

[2]The ECB was created in June 1998.

The literature on announcement returns agrees upon the notion that the 24h pre-ECB return is zero, whereas the respective pre-FOMC announcement drift is positive (see Lucca and Moench, 2015; Brusa et al., 2019). In this thesis, we contribute to the literature by documenting a statistically significant and sizeable 24 hour pre-ECB return for the period from 2010 to 2015. Opposite to the pre-FOMC drift, which persists after the announcement, we also show that the pre-ECB return exhibits a mean-reversion that starts with the beginning of the ECB's press conference. Using modern textual analysis tools and the captions of the press conference webcasts we develop a novel methodology that automatically creates timestamps for each part of the press conference statement. Based on these timestamps we then study the real-time impact of the information revealed by the ECB Chairman on financial markets and relate the mean-reversion dynamic of the pre-ECB return to fundamental central bank news. With this latter analysis we also contribute to the large literature that studies how central bank communication affects asset prices. Particularly, we complement the analysis of Schmeling and Wagner (2019) who study the impact of the aggregate tone of the ECB's introductory statement on the post-announcement returns.

1.2 Structure of the Thesis

The structure of this thesis is as follows:

Chapter 2 is based on the working paper 'The Euro Crisis and the 24h Pre-ECB Announcement Return' of Ulrich et al. (2019a). We show that the European Central Bank also exhibits a large pre-announcement return of 0.5% (on average) which is similar to the U.S. evidence reported by Lucca and Moench (2015) when studying the time period including the euro crisis from 2010 to 2015. In contrast, for the time span from 2000 to 2009 we do not find an announcement return that is statistically different from zero. We then study the determinants of the pre-ECB return series and find that its time series dynamic is highly driven by uncertainty before the event. Examining different theoretical explanations, we find that the theories can explain several of the stylized facts of the observed ECB announcement return but not the whole pattern. Our analysis also shows that the high uncertainty periods of the euro crisis, which can be characterized as times of severe eurozone break-up fear, were the main driver of the average pre-ECB announcement return.

After presenting evidence for a pre-ECB drift and discussing the determinants of its time series we analyze which information that gets revealed during the ECB's press conference is driving the mean-reversion in the equity market. Therefore we rely on novel textual analysis methods whose foundations are presented in Chapter 3: After a simple introductory example that illustrates the fundamental problems when working with textual data we discuss how to automatically collect data from the Internet. We also present standard pre-processing techniques and how to convert the qualitative strings into a numerical representation. We conclude the chapter by showing how to extract topics and the tone from textual statements and how to gain knowledge from textual data using regression techniques.

Chapter 4 is based on the working paper 'The Real-Time Impact of ECB Press Conferences on Financial Markets' of Ulrich et al. (2019b). In this chapter we present our novel methodology that uses textual analysis tools and the captions of the ECB's press conference webcasts to create timestamps for each part of the introductory statement. Using clustering techniques we extract fine granular topics from the press conference statement. Based on the generated timestamps and topics we study the real-time impact of the different information content on financial markets. We find that the largest impact stems from the announcement of the ECB's quantitative easing measures and also from negative information revealed during the economic analysis, in particular news related to the course of the European sovereign debt crisis.

Chapter 5 summarizes the main findings of this thesis and provides a concise outlook on future research questions.

Chapter 2

The Euro Crisis and the 24h Pre-ECB Announcement Return

2.1 Introduction

As pointed out in Chapter 1, the literature on the 24h pre-monetary policy announcement premium appears to broadly agree upon the notion that the 24h pre-ECB announcement premium is zero, whereas the respective 24h pre-FOMC premium is positive (see e.g. Lucca and Moench, 2015; Brusa et al., 2019). In this chapter we use a previously unexplored rich cross-section of tick-by-tick European industry sector and country equity return data to uncover that the statement of a zero 24h pre-ECB announcement premium is inaccurate; as highlighted by Figure 2.1 and Figure 2.2. For the period from 2010 to 2015, the 24h pre-ECB announcement premium is statistically significant and sizeable for a variety of European industry sectors.[1]

We further document that the magnitude of the pre-ECB announcement return is mainly driven by periods of high uncertainty during the euro crisis which can be characterized as times of severe eurozone break-up fear. The magnitude of the pre-ECB return is 2.5 times higher than for meetings that were held in non-stressed periods. This finding is especially

[1]Current research on announcement drifts in Europe do either use intraday returns until 2011 or rely on daily data and hence could not detect the sizeable and significant upward drift in different sectors of the euro denominated economy; see e.g. Lucca and Moench (2015) or Brusa et al. (2019), which as our analysis shows exists in particular in systemically relevant sectors of the economy and during the height of the euro crisis.

pronounced for the systemically important banking industry, where the magnitude of the average premium almost reached 3% (per announcement) for the ECB's monetary policy decisions that occurred during times of high uncertainty.

In general, for the time period from 2010 to 2015 a simple strategy that buys the Euro Stoxx 50 24 hours prior to the ECB monetary policy announcement and sells it shortly before the announcement, earned an annualized return of 6%, which translates to an annualized Sharpe ratio of 1.5. For selected industries, such as the European banking sector, the respective premium per year was 12% (Sharpe ratio of 1.6); at a time when the annual return of the European banking sector was on average flat. In contrast to the 24h pre-FOMC announcement premium, the 24h pre-ECB announcement premium mean-reverts back to zero within two hours after the announcement, making it undetectable when looking at daily close-to-close prices.

We do not only study the aggregate European equity market, but also 19 European industries and 8 EU country indices for the time span of 2000 to 2015. For the time period from 2010 to 2015 we find that the average 24h pre-ECB announcement return has been positive for all the 19 EU industries; but the magnitude varies a lot among our set of industries and countries. Especially the banking, financial services and insurance (BFSI) sectors and also the cyclical industries (like automobile) exhibit large pre-ECB returns.

The European country indices confirm that the 24h pre-ECB announcement premium is not a general phenomenon, but rather concentrated among the euro-peripheral countries. We find the largest pre-ECB returns for the GIIPS countries of our sample (Italy and Spain). Moreover, we also document that the CAPM explains the cross-section of European equity on ECB announcement days, while the CAPM fails on non-ECB announcement days; very similar to the U.S. evidence documented in Savor and Wilson (2014) and Lucca and Moench (2015).

Looking at data from 2000 to 2015 highlights that 2010 marks a structural break in the euro area. We hence study the characteristics of the 24h pre-ECB announcement return for the euro crisis (2010 – 2015) episode. The results show that the time series of the pre-ECB return strongly covaries with the level of uncertainty prior to the meeting, as measured by the VSTOXX and also its dissolution shortly before the announcement (R^2 of 25 - 30%). The pre-ECB run-up can also predict the subsequent communication shock (as defined in Leombroni et al. (2017)) and the post-announcement return in the equity market, in particular the price change of the Euro Stoxx 50 during the Q&A session (14:50 to 15:30) and the return until the end of the trading day (15:30 to 17:30 CET). We document a positive relationship between the corresponding returns and a R^2 of 10 to 30%.

Our finding that the positive pre-ECB announcement drift is mainly driven by the level of uncertainty and its subsequent dissolution is consistent with the theories of Ai and Bansal (2018) and Wachter and Zhu (2018) who argue that the uncertainty around an information event requires investors to demand a premium for holding risky assets. It is also in line with Hu et al. (2019) who state that the pre-FOMC drift is a 'premium for heightened uncertainty'. Whereas these explanations are plausible for the observed pre-ECB return they cannot explain the mean-reversion that follows the announcement. The behavioral anticipation story of Peterson (2002) ('buy on the rumor, sell on the news') is able to explain both, the positive pre-announcement return and the following mean-reversion for the overall euro crisis period from 2010 to 2015. But when splitting the sample into high and low uncertainty meetings we cannot confirm the finding. Thus, we conclude that existing theories and explanations can rationalize several stylized facts of the return dynamic around ECB meetings but not the whole pattern.

In addition to relating the pre-announcement drift around central bank meetings to general equity market uncertainty (see Hu et al., 2019; Martello and Ribeiro, 2018) we further connect the uncertainty in Europe from 2010 to 2015 to the course of the European sovereign debt crisis.We first use several standard proxies for measuring economic and policy uncertainty in the EU and then create a new ECB Uncertainty Index using textual analysis. Based on the ECB's uncertainty characterization of the economic environment, revealed in the press conference's intro statement, we construct a simple wording indicator. We find that the documented pre-ECB announcement return is mainly driven by the meetings that are classified as 'high uncertainty' by our index. Using the information content of the respective press conference's Q&A session allows us to link the general uncertainty classification to the actual events and developments of the euro crisis. We find that the high uncertainty periods of the euro crisis can be characterized as times of severe concerns and fear about a eurozone break-up. Therefore we conclude that this eurozone break-up fear was a major driver of the 24h pre-ECB announcement return.

The remainder of the chapter is organized as follows. Section 2.2 provides a brief discussion on how the ECB conducts monetary policy. In Section 2.3, we describe the data used in the empirical analysis that follows. We present the main empirical findings in Section 2.4. In Section 2.5, we study the determinants of the time series of the pre-ECB announcement return and Section 2.6 discusses potential explanations and theories. Section 2.7 relates the pre-ECB announcement return to the euro crisis and Section 2.8 concludes.

Related Literature

This chapter contributes to the large literature on announcement returns around central bank meetings. We document a pre-ECB announcement return similar to Lucca and Moench (2015), who report large average excess returns for U.S. equities. We also find a mean-reversion that starts with the begin of the press conference which Schmeling and Wagner (2019) relate to a change in the tone of central bank communication. We do not find evidence for ECB cycle returns as reported in Cieslak et al. (2019). In contrast to Gu et al. (2018) who document positive average returns after FOMC announcements that are accompanied by the SEP release[2] and a press conference we find a mean-reversion of the announcement return with the begin of the ECB's press conference. Brusa et al. (2019) argue that the Fed is generally the leader among all central banks since it is the only one that exhibits pre-announcement returns. Our evidence suggests that this has changed with the start of the euro crisis. Kaul and Watanabe (2015) study potential driver of the pre-FOMC announcement drift and argue that market beta completely explains the drift. As in Savor and Wilson (2014) we also find that the CAPM holds on

[2]SEP release stands for Summary of Economic Projections release.

ECB announcement days. Newer studies argue that the FOMC drift has vanished now or is no longer significant based on empirical evidence covering the last years (see Azar and Lo, 2016; Gilbert et al., 2018; Ben Dor and Rosa, 2019). One potential reason could be that the drift has been arbitraged away since its discovery (see Gilbert et al., 2018).[3]

2.2 The European Central Bank and Governing Council Decisions

The European Central Bank (ECB) is an EU institution that conducts a single monetary policy strategy for 19 EU member countries. The primary goal of the ECB is to ensure price stability in the Euro-zone. As long as there is no conflict with price stability, the ECB shall also support sustainable growth (Article 127 (1) of the AEU contract). The ECB decides upon the monetary policy strategy for its member countries and implements policies to operationalize the strategy (Article 127 (2) of the AEU contract).

The decision-making committee of the ECB is the ECB Governing Council. It consists of the 19 member country central bank presidents, as well as six members that are appointed by the European Council. These six members make up the executive board and consist of the ECB president, the vice-president and four additional members.

The ECB Governing Council meets bi-weekly and decides upon monetary policy measures during every third meeting. Prior to 2015, monetary policy decisions were announced during every second meeting. Similar to monetary policy decisions of the U.S. Federal Reserve (Fed), monetary policy decision meetings of the ECB Governing Council follow a strict protocol. The public announcement about the actual monetary policy decision is released on the day of the meeting at 13:45 a.m. CET. Forty-five minutes later, at 14:30 CET, the president of the executive board reads a public statement in which he explains the reason that led to the monetary policy decision and takes questions afterwards. Four weeks later, the ECB publishes a document that summarizes the Governing Council's discussion that preceded the monetary policy decision.

[3] In a recent post Lucca and Moench (2018) argue that the pre-FOMC drift can still be observed, but only for meetings that are followed by a press conference.

2.3 Data

Our analysis focuses on the 24h pre-ECB announcement return at scheduled ECB meetings from 2000 to 2015. We study the anticipatory price movements across different European equity markets. We accomplish this by studying the 24h return that investors could realize when systematically buying European equity at 13:30 CET one day prior to a scheduled ECB monetary policy meeting and selling on the day of the meeting at 13:30 CET. We also compare these 24h pre-ECB announcement returns to the daily close-to-close return. Dates and exact times of scheduled ECB monetary policy meetings are taken from the webpage of the ECB.

In order to study the 24h pre-ECB announcement return in European equity we use a previously unexplored rich panel of European equity data provided by Deutsche Börse. Our analysis relies on tick-by-tick (15 seconds) return data of the Euro Stoxx 50, 19 Euro Stoxx supersector indices and seven Stoxx country indices. The 19 supersector indices include the following European sectors: banking, automobile, basic resources, insurance, oil and gas, consumer products, chemicals, industry goods, financial services, telecommunication, travel and leisure, real estate, utilities, technology, personal and household goods, food and beverages, media, retail and health care. The seven country indices are comprised of the Stoxx UK 50, Stoxx France 50, Stoxx Spain 20, Stoxx Italy 20, Stoxx Eastern Europe 50, the Stoxx Nordic 30 and the Stoxx Sub-Balkan 30. Each of these country indices tracks the value of the respective regions' blue-chips. All data are available from 06/29/2000 to 06/30/2015, except for three sector indices and the Stoxx country indices which are only available from 2010 onwards.[4] As a proxy for the German market we include in addition to the Stoxx country indices also the DAX index.

To study the determinants of the pre-ECB announcement return time series we rely on data from various sources. First, we obtain data on policy rates, non-standard monetary policy measures and the press conference transcripts from the ECB webpage. Then, we

[4]Data on the Euro Stoxx 50 and the 19 Euro Stoxx supersectors span the period 06/29/2000 to 06/30/2015; except for the Euro Stoxx Personal & Household Goods, Retail, Travel & Leisure whose data starts in 09/20/2004; and for the Euro Stoxx Real Estate, whose data starts in 09/22/2008. Data on Stoxx Eastern Europe 50, Nordic 30, Sub-Balkan 30 are available since 01/04/2010. Last not least, the country indices France 50, Spain 20, UK 50 and Italy 20 are available since 02/21/2011.

use LexisNexis for historical coverage of major newspapers in the English language. We also use intra-day prices of German Schatz government bond futures and VSTOXX futures from the EUREX exchange.[5] Daily end of day VSTOXX values are obtained from the STOXX webpage.[6] In addition, we collect the monthly time series of the Industrial Production Index and Consumer Price Index from the eurostat webpage.[7] To calculate excess returns we use the 6-month yield from the German yield curve as provided by Bundesbank and convert it to a daily horizon.[8] [9] The time series for the Economic Policy Uncertainty Index of Baker et al. (2016) and the categorical index for Greece as created by Hardouvelis et al. (2018) are downloaded from the EPU webpage[10] The Sentix EBI index is obtained from Bloomberg.

2.4 Empirical Findings

This section summarizes the empirical findings of our analysis. We first study the 24h pre-ECB announcement return in the eurozone from 2000 to 2015 and then divide the sample into a pre-crisis period, spanning 2000 to 2009, and a time window covering the euro crisis (2010 to 2015). Then we move on summarizing the cross-sectional findings.

2.4.1 Pre-ECB Announcement Return since 2000

We run two dummy-variable regressions to assess the magnitude of abnormal returns on ECB monetary policy announcement days,

$$rx_t = \beta_0 + \beta_1 \times \mathcal{I}_t(\text{ECB}) + \epsilon_t, \tag{2.1}$$

[5]Note, for the later analysis we always select the most liquid contract.

[6]The historical data can be downloaded here: https://www.stoxx.com/index-details?symbol=V2TX.

[7]The eurostat database can be accessed via https://ec.europa.eu/eurostat/web/short-term-business-statistics/data/database.

[8]The 0.5 years yield is the shortest maturity provided by the Bundesbank.

[9]The German yields are highly correlated with the ECB AAA yields but available for a longer time period (see Schmeling and Wagner, 2019).

[10]See https://www.policyuncertainty.com.

where rx is the percentage value of the log excess return on the Euro Stoxx 50 over the risk-free rate. The constant β_0 measures the unconditional excess return that is earned on days where the ECB does not hold a scheduled monetary policy meeting. The term $\mathcal{I}_t(\text{ECB})$ is an indicator function that takes the value of one if t is an ECB announcement day and zero otherwise. Hence, β_1 captures the unconditional abnormal excess return on days where the ECB announces their decision about future monetary policy. The residuals of this regression are represented by ϵ. The first dummy regression uses excess returns earned during the 24h period prior to the public announcement of the monetary policy decision (13:30 CET on the day before the ECB meeting to 13:30 CET on the announcement day), while the second dummy regression uses daily excess returns (based on the market close on the day before the event to the market-close on the announcement day).

Table 2.1 summarizes the results of both dummy regressions. We document a 27 basis points abnormal 24h pre-ECB announcement excess return. The robust t-statistic of 2.66 confirms significance at the 99% confidence level. During non-ECB days, the excess return has been -4 basis points with an absolute t-statistic of 1.9. Column 3 and 4 of Table 1 show that the daily excess return on a close-to-close basis is statistically speaking zero, indicating that there is a mean-reversion in prices following the announcement. Thus, one cannot detect this pattern using end of day returns. In general, these findings are supplementary to Lucca and Moench (2015) who document a pre-FOMC drift but report that there is no such price pattern before meetings of the European Central Bank. Compared to their analysis we include the recent time period of the euro crisis for our calculations which yields a significant pre-ECB drift.

Table 2.2 quantifies the investment implication of the 24h pre-ECB announcement return. Investing 12 times a year to earn the 24h pre-ECB announcement excess return would have yielded on average an annual risk premium of 2.78%, which translates to an annualized Sharpe ratio of 0.56. The analogous 24h excess return for non-ECB announcement days would have been -10.27%. The last column of Table 2 shows that an investment strategy based on end of day returns would have lost money on average (for both announcement

Table 2.1: **Dummy Regressions - Euro Stoxx 50**
This table reports the results from the pre-ECB dummy regressions based on daily Euro Stoxx 50 returns. The log excess return series is calculated (1) based on the 1:30 p.m. price on date t-1 and the price at 1:30 p.m. on date t and (2) the closing prices of the ES50 on date t-1 and t. 'Pre-ECB Dummy' is a variable that is equal to one when an ECB monetary policy decision is scheduled in the following 24 hour interval and zero otherwise. The sample period is from 06/30/2000 to 06/30/2015. Robust t-statistics shown in parenthesis. *, ** indicate significance at the 5% or 1% level (|t-stat| > 1.96 or 2.58).

	2000 - 2015			
	Euro Stoxx 50 Returns			
	1.30pm-to-1.30pm		**close-to-close**	
Constant	-	-0.043	-	-0.024
		(-1.89)		(-1.07)
Pre-ECB Dummy	0.232*	0.274**	-0.122	-0.098
	(2.29)	(2.66)	(-0.96)	(-0.76)
Obs.	3815	3815	3815	3815
N. of ECB	195	195	195	195

and non-announcement days).

Table 2.2: **Sharpe Ratios - Euro Stoxx 50**
This table reports the cumulative annual excess return and Sharpe-ratio earned on (non-) ECB announcement days. The log excess return series is calculated (1) based on the 1:30 p.m. price on date t-1 and the price at 1:30 p.m. on date t and (2) the closing prices of the ES50 on date t-1 and t. '8x' ('12x') refers to the assumed number of ECB monetary policy decisions within a year. 'ECB Sharpe-Ratio' refers to the annualized Sharpe-Ratio earned on ECB announcement days. It is calculated as $\sqrt{8}$ ($\sqrt{12}$) * per-meeting Sharpe-Ratio (sample mean of pre-ECB returns divided by its sample standard deviation). The sample period is from 06/30/2000 to 06/30/2015.

	2000 - 2015			
	1.30pm-to-1.30pm		**close-to-close**	
	8x	**12x**	**8x**	**12x**
Annual ex-return ECB	1.85 [%]	2.78 [%]	-0.98 [%]	-1.47 [%]
Annual ex-return non-ECB	-10.44 [%]	-10.27 [%]	-5.77 [%]	-5.68 [%]
ECB Sharpe-Ratio (Ann.)	0.45	0.56	-0.2	-0.24

Figure 2.1 plots the cumulative percentage return of the Euro Stoxx 50 from one day prior to a scheduled ECB announcement to one day after the announcement. The average 24h pre-ECB announcement return of 0.27% is visible, together with the substantial amount of noise around the estimate.[11] In contrast to the U.S. experience, as documented in Lucca and Moench (2015), our study for Europe reveals that the ECB announcement return mean-reverts sharply right after the announcement, while it persists in U.S. markets.

2.4.2 Pre-ECB Announcement Return: Structural Break

We next split the sample into two sub-periods: 2000 - 2009 and 2010 - 2015. The former resembles a period without concerns of a eurozone break-up, while the latter covers the European sovereign debt crisis. Figure 2.2 shows that the pre-ECB announcement return is a phenomenon of the euro crisis. Prior to 2010 there is no evidence of an upward

[11]Notice that the cumulative return in the figure starts at 09:00 CET whereas the 24h pre-ECB announcement return starts at 13:30 CET.

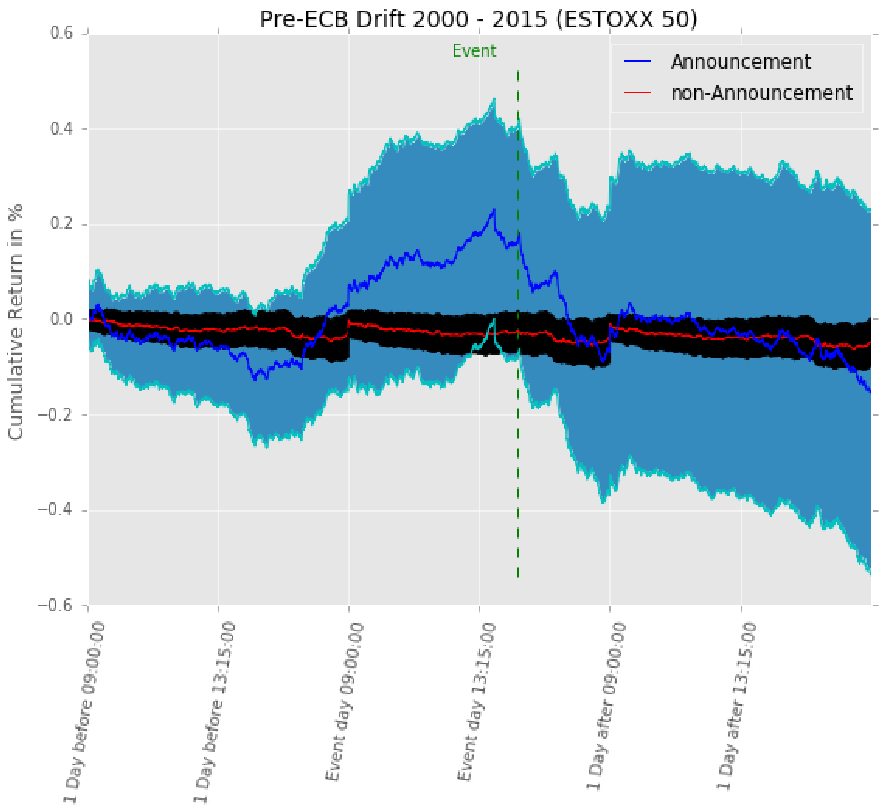


Figure 2.1: **Cumulative Return Series of the Euro Stoxx 50 Index**
This figure plots the average cumulative 15 seconds return of the Euro Stoxx 50 on a three day window around scheduled ECB announcement days (i.e. one day prior to the event, event, one day after the event). The blue line is the average return around ECB announcement days, the blue shaded area is the respective pointwise 95% confidence interval. The red line is the average return around non-ECB announcement days, the black shaded area is the respective pointwise 95% confidence interval. The vertical dashed line marks the start of the ECB press conference, 2:30 p.m. (local time), on the event day. The sample period is June 2000 to June 2015.

drift in the value of the Euro Stoxx 50, while after 2010, the upward drift as well as the fast mean-reversion after the ECB announcement are clearly visible.

Table 2.3 highlights the previous observation using regression tools. It summarizes the regression estimates of the dummy regression model from Equation (2.1), applied to both sub-samples, separately. The table reveals three insights. First, the 24h pre-ECB announcement return exists only from 2010 onwards. For the period 2010 - 2015, the 24h pre-ECB announcement premium adds up to an average of 0.55% per meeting with a robust t-statistic of 3.6. Second, return data at the daily frequency does not detect this pattern as the realized pre-ECB announcement return mean-reverts completely hours after the actual announcement. Third, statistically speaking, the 24h pre-ECB announcement return for the pre-crisis sub-sample is zero.

Table 2.4 compares the economic magnitude of the pre-ECB announcement return in both sub-samples. The realized Sharpe ratio for the 24h pre-ECB announcement return during the non-crisis period (2000 - 2009) is 0.2. In contrast, the same simple buy and hold strategy would have resulted in an annual Sharpe ratio of 1.47 or 1.2 during the crisis period (2010 - 2015), depending on whether one works with 12 or 8 scheduled ECB meetings per year, respectively.

In summary, we document evidence for a 24h pre-ECB announcement excess return during the period of the European sovereign debt crisis. This abnormal pre-ECB return mean-reverts back to zero right after the ECB monetary policy decision is publicly announced.

The next subsection analyzes whether this pattern originates in a few industries or whether it is broad-based.

2.4.3 Cross-sectional Evidence

Table 2.5 reports the results of the dummy regression model from Equation 2.1, estimated separately for the 19 different European industries for the two sub-samples. Prior to the euro crisis (2000 - 2009), only the health-care industry exhibits weak evidence of a 24h pre-ECB announcement return (robust t-statistic of 2.3). During the euro crisis

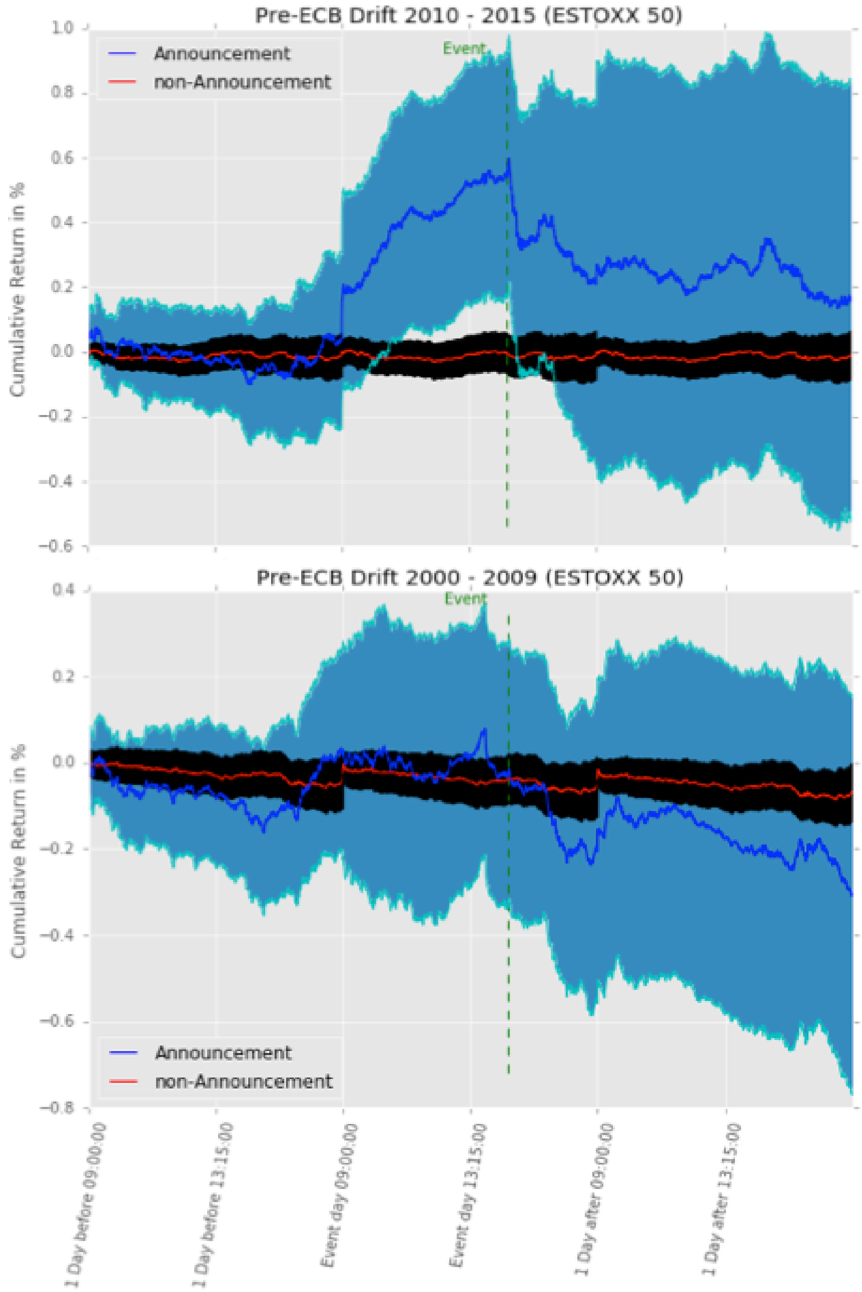


Figure 2.2: **Cumulative Return Series of the Euro Stoxx 50 Index: Sub-Periods** This figure plots the average cumulative 15 seconds return of the Euro Stoxx 50 on a three day window around scheduled ECB announcement days (i.e. one day prior to the event, event, one day after the event). The blue line is the average return around ECB announcement days, the blue shaded area is the respective pointwise 95% confidence interval. The red line is the average return around non-ECB announcement days, the black shaded area is the respective pointwise 95% confidence interval. The vertical dashed line marks the start of the ECB press conference, 2:30 p.m. (local time), on the event day. The upper panel analyzes the sample period of the euro crisis, January 2010 to June 2015. The lower panel studies the pre-Euro crisis period, i.e. June 2000 to December 2009.

Table 2.3: **Dummy Regressions - Euro Stoxx 50: Sub-Periods**
This table reports the results from the pre-ECB dummy regressions based on daily Euro Stoxx 50 returns for two distinct time periods: The upper table shows the dummy regression results for the sample period from 01/01/2010 to 06/30/2015, the lower one shows the results for the period of 06/30/2000 to 12/31/2009. The log excess return series is calculated (1) based on the 1:30 p.m. price on date t-1 and the price at 1:30 p.m. on date t and (2) the closing prices of the ES50 on date t-1 and t. 'Pre-ECB Dummy' is a variable that is equal to one when a ECB monetary policy decision is scheduled in the following 24 hour interval and zero otherwise. Robust t-statistics shown in parenthesis. *, ** indicate significance at the 5% or 1% level (|t-stat| > 1.96 or 2.58).

	2010 - 2015			
	Euro Stoxx 50 Returns			
	1.30pm-to-1.30pm		**close-to-close**	
Constant	-	-0.016	-	0.001
		(-0.50)		(0.05)
Pre-ECB Dummy	0.531**	0.548**	0.148	0.147
	(3.42)	(3.56)	(0.77)	(0.76)
Obs.	1401	1401	1401	1401
N. of ECB	64	64	64	64

	2000 - 2009			
Constant	-	-0.058	-	-0.038
		(-1.91)		(-1.30)
Pre-ECB Dummy	0.085	0.144	-0.254	-0.216
	(0.67)	(1.08)	(-1.56)	(-1.29)
Obs.	2414	2414	2414	2414
N. of ECB	131	131	131	131

Table 2.4: **Sharpe Ratios - Euro Stoxx 50: Sub-Periods**
This table reports the cumulative annual excess return and Sharpe-ratio earned on (non-) ECB announcement days for two distinct time periods: The upper table shows the results for the sample period from 01/01/2010 to 06/30/2015, the lower one for the period from 06/30/2000 to 12/31/2009. The log excess return series is calculated (1) based on the 1:30 p.m. price on date t-1 and the price at 1:30 p.m. on date t and (2) the closing prices of the ES50 on date t-1 and t. '8x' ('12x') refers to the assumed number of ECB monetary policy decisions within a year. 'ECB Sharpe-Ratio' refers to the annualized Sharpe-Ratio earned on ECB announcement days. It is calculated as $\sqrt{8}$ ($\sqrt{12}$) * per-meeting Sharpe-Ratio (sample mean of pre-ECB returns divided by its sample standard deviation).

	2010 - 2015			
	1.30pm-to-1.30pm		**close-to-close**	
	8x	**12x**	**8x**	**12x**
Annual ex-return ECB	4.25 [%]	6.38 [%]	1.18 [%]	1.78 [%]
Annual ex-return non-ECB	-3.94 [%]	-3.88 [%]	0.36 [%]	0.36 [%]
ECB Sharpe-Ratio (Ann.)	1.2	1.47	0.27	0.33

	2000 - 2009			
Annual ex-return ECB	0.68 [%]	1.02 [%]	-2.03 [%]	-3.05 [%]
Annual ex-return non-ECB	-14.25 [%]	-14.02 [%]	-9.37 [%]	-9.21 [%]
ECB Sharpe-Ratio (Ann.)	0.16	0.2	-0.39	-0.48

period of 2010 - 2015, roughly half of all industries exhibit a pre-ECB announcement premium with a robust t-statistic larger than 2. The highest average pre-ECB announcement premium of 1.1% (per announcement) was earned in the banking sector (t-statistic of 4.1). Also the other BFSI[12] sectors exhibit high pre-ECB returns, 0.7% for financial services and 0.5% for the insurance industry, respectively. In addition, cyclical industries like the automotive sector or the oil & gas industry also show large pre-announcement returns, whereas non-cyclical industries like food & beverages or health care exhibit on average much smaller pre-ECB returns.[13] These findings are consistent with Lucca and Moench (2015) who also report high pre-FOMC returns for the banking, insurance and trading industry or the automotive sector.

Next, we run CAPM regressions for each of the 19 European supersectors and for each of the two sub-periods,

$$rx_{it} = \alpha_i + \beta_i \times rx_{m,t} + \epsilon_{it}, \tag{2.2}$$

where rx_i is the excess return of supersector i, rx_m is the excess return of the Euro Stoxx 50, α_i, β_i are constant parameters and ϵ_i captures the residuals. The upper panel of Figure 2.3 plots the average realized 24h pre-ECB announcement excess return (y-axis) on scheduled ECB announcement days (upper left panel) and on all other days (upper right panel) against the respective industry beta for the time of the euro crisis. It is evident that sectors with a higher amount of systematic risk do also pay on average higher 24h pre-ECB announcement returns, whereas on non-announcement days the CAPM does not hold. This finding is consistent with Lucca and Moench (2015) who find the same behavior for pre-FOMC industry returns and also with Savor and Wilson (2014) who extend this observation to other macroeconomic announcements (like inflation or employment). In addition, we document that for the pre-euro crisis period the CAPM does also not hold for European industries.

[12] BFSI stands for banking, financial services and insurance.

[13] The internet appendix contains additional figures on the cumulative returns one day prior and after an ECB monetary policy announcement for different industries and for both sub-samples.

Table 2.5: **Dummy Regressions - Industries: Sub-Periods**
This table reports the results from the 19 pre-ECB dummy regressions based on daily returns for each of the Euro Stoxx sector indices for two distinct time periods: The left panel shows the regression results for the sample period of 01/01/2010 to 06/30/2015, the right panel from 06/30/2000 to 12/31/2009, except for Travel and Leisure, Personal and Household Goods and Retail sector which start on 09/21/2004 and the Real Estate sector which starts on 09/23/2008. The log excess return series is calculated based on the 1:30 p.m. price of the corresponding Euro Stoxx sector index on date t-1 and the price at 1:30 p.m. on date t. 'Industry' refers to the corresponding Euro Stoxx supersector index. 'beta_1' ('beta_0') reports the estimation coefficient for β_1 (β_0) of the linear model stated in Equation (1) run for each sector seperately. Robust absolute t-statistics shown in parenthesis. *, ** indicate significance at the 5% or 1% level (|t-stat| > 1.96 or 2.58).

	2010 - 2015		*2000 - 2009*	
Industry	**beta_1**	**beta_0**	**beta_1**	**beta_0**
Banks	1.11** (4.08)	-0.08 (1.51)	0.22 (1.49)	-0.06 (1.42)
Automobile and Parts	0.82** (3.53)	0.03 (0.61)	-0.16 (0.82)	-0.02 (0.39)
Basic Resources	0.77** (3.08)	-0.06 (1.43)	-0.02 (0.10)	-0.00 (0.07)
Insurance	0.67** (3.11)	-0.00 (0.11)	0.14 (0.73)	-0.08 (1.68)
Oil and Gas	0.59** (3.36)	-0.04 (0.99)	-0.09 (0.61)	-0.03 (0.83)
Construction and Material	0.56** (2.72)	-0.01 (0.34)	-0.02 (0.12)	-0.02 (0.52)
Chemicals	0.46** (2.93)	0.02 (0.70)	-0.08 (0.53)	-0.00 (0.09)
Industrial Goods and Services	0.46** (2.74)	0.01 (0.41)	0.12 (0.88)	-0.05 (1.49)
Financial Services	0.46** (3.09)	0.01 (0.21)	0.11 (0.83)	-0.04 (1.15)
Telecommunications	0.40** (3.16)	-0.03 (0.85)	0.17 (1.13)	-0.08* (2.27)
Utilities	0.33* (1.99)	-0.04 (1.24)	-0.05 (0.49)	-0.02 (0.71)
Technology	0.31* (2.15)	0.02 (0.69)	0.47 (1.89)	-0.12* (2.40)
Food and Beverage	0.23* (2.10)	0.03 (1.39)	-0.01 (0.06)	-0.02 (0.73)
Media	0.19 (1.22)	0.02 (0.82)	0.10 (0.71)	-0.08* (2.47)
Health Care	0.03 (0.25)	0.04 (1.63)	0.27* (2.32)	-0.05 (1.73)
	2010 - 2015		*2004 - 2009*	
Travel and Leisure	0.35* (2.22)	0.03 (0.97)	-0.08 (0.35)	-0.02 (0.51)
Personal and Household Goods	0.31* (2.34)	0.03 (1.13)	-0.15 (0.88)	0.00 (0.09)
Retail	0.14 (1.05)	0.03 (0.91)	-0.05 (0.34)	-0.01 (0.41)
	2010 - 2015		*2008 - 2009*	
Real Estate	0.34* (1.97)	0.01 (0.18)	0.55 (0.81)	-0.08 (0.62)

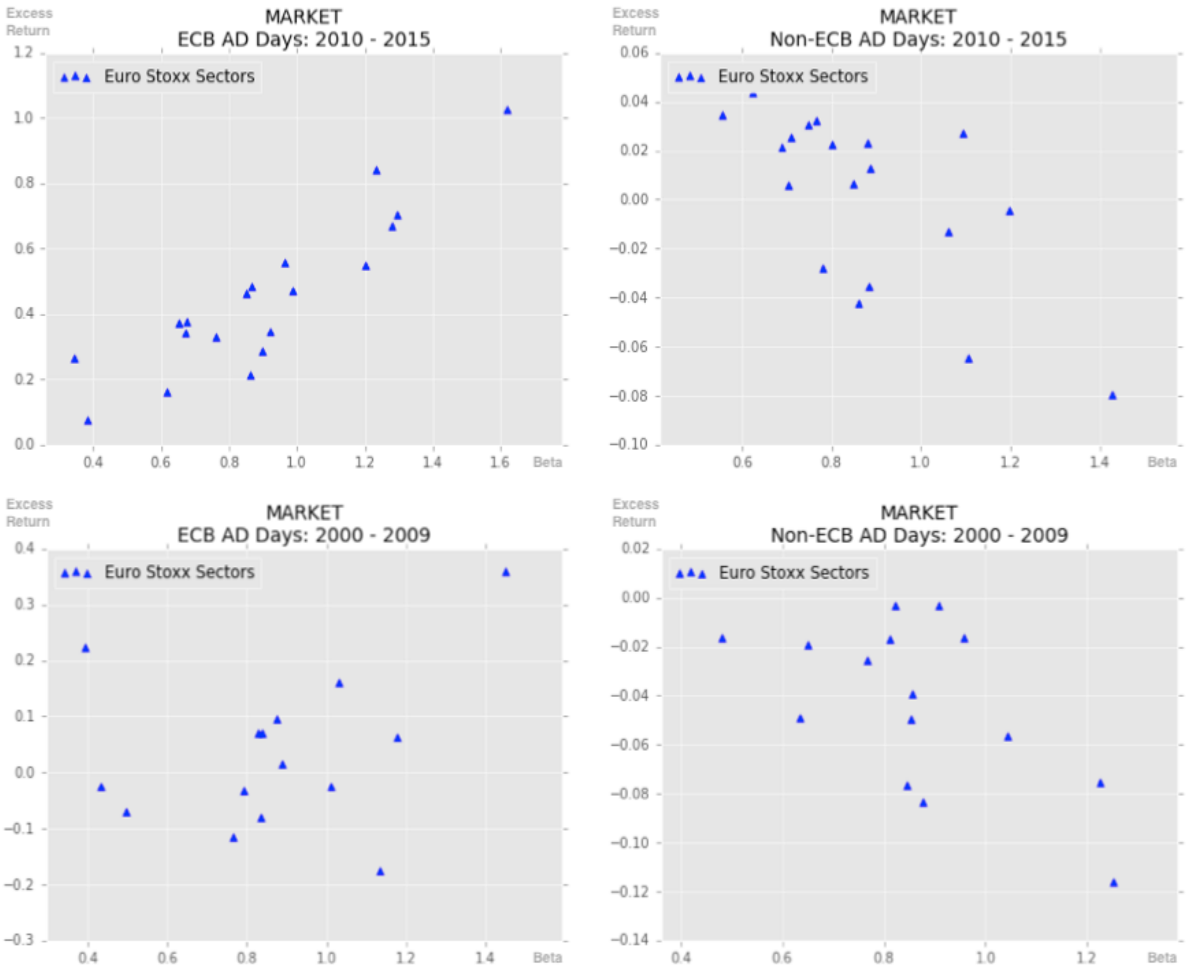


Figure 2.3: **CAPM on ECB Announcement Days - Industries: Sub-Periods** This figure plots the average 24h excess return of the EU supersectors against their respective CAPM beta. The log return is calculated based on the index value at 1.30 pm on the day before the ECB meeting and the index value at 1.30 pm on the announcement day. The upper panels summarize the results for ECB announcement days (upper left) and all other days (upper right) for the period of the Euro crisis (2010 - 2015). The two lower panels summarize the results for ECB announcement days (lower left) and all other days (lower right) for the pre-euro crisis episode (2000 - 2009).

Table 2.6 summarizes regression results of the dummy regression model from Equation (2.1) for the Stoxx country indices. One can see that the GIIPS countries of our sample (Italy and Spain) exhibit the highest pre-announcement returns with an estimate of 0.6% and 0.5%. France and Germany also have high estimates with a β_1 coefficient of around 0.4%, whereas the other countries in our sample have much lower estimates.[14]

Table 2.6: **Dummy Regressions - Countries**
This table reports the results from the 8 pre-ECB dummy regressions based on daily returns for each of the Stoxx country indices and the DAX. The log excess return series is calculated based on the 1:30 p.m. price of the corresponding Stoxx index on date t-1 and the price at 1:30 p.m. on date t. 'Country' refers to the corresponding Stoxx country index or the DAX. 'beta_1' ('beta_0') reports the estimation coefficient for β_1 (β_0) of the linear model stated in Equation (1) run for each country seperately. The sample period starts for 'Nordic', 'Eastern Europe', 'Sub Balkan' and 'Germany' on 01/05/2010 and for all of the other Stoxx country indices on 02/22/2011. The sample period ends for all sector indices on 06/30/2015. Robust absolute t-statistics shown in parenthesis. *, ** indicate significance at the 5% or 1% level (|t-stat| > 1.96 or 2.58).

	2010 - 2015			
Country	**beta_1**	(| t-stat |)	**beta_0**	(| t-stat |)
Germany	0.45**	(3.02)	0.02	(0.52)
Nordic	0.29	(1.92)	0.02	(0.84)
Eastern Europe	0.23	(1.47)	-0.03	(0.84)
Sub Balkan	0.14*	(2.07)	-0.04	(1.86)
	2011 - 2015			
Italy	0.61*	(2.42)	-0.03	(0.65)
Spain	0.54**	(3.38)	-0.01	(0.27)
France	0.42*	(2.54)	-0.00	(0.05)
UK	0.24*	(2.00)	0.01	(0.30)

To sum-up, opposite to Lucca and Moench (2015) we report a significant pre-ECB return that mean-reverts after the announcement. When splitting the sample one can see that the pre-ECB drift is a recent phenomenon, driven by the 2010 to 2015 sample which

[14]Running a CAPM like regression for the country indices yields a similar result as when using the different industry returns. On announcement days the CAPM holds, whereas on non-announcement days it fails.

covers the European sovereign debt crisis. For the cross-section we document the largest pre-announcement returns for the BFSI and cyclical industries and also for the GIIPS countries of our sample.

2.5 Determinants of Pre-ECB Announcement Returns

We next study the determinants of the pre-ECB announcement return series. We draw inspiration from recent studies on what drives markets on central bank announcement days, and analyze how realized monetary policy decisions, market expectations and monetary policy surprises drive our results. In addition, we examine the impact of market uncertainty and economic activity on the magnitude of the pre-ECB return.

2.5.1 Monetary Policy Actions

First, we study the relationship between the ECB's realized monetary policy actions, like the change in the policy rate or the announcement of non-standard monetary policy measures, and the pre-ECB return. Therefore, we regress the 1.30pm-to-1.30pm Euro Stoxx 50 log excess return[15] on the change in the policy rate announced at the press conference (ΔMRO) and a dummy variable (UMP) that takes the value of one for meetings at which unconventional monetary policy measures are announced and zero otherwise (as used in Schmeling and Wagner, 2019).
Table 2.7 shows the regression results. The positive constant with a magnitude of 0.4 reflects the positive pre-announcement drift. We can also observe a negative estimate of -2.3 for the standard monetary policy decision variable which is consistent with the intuition that a lower interest rate should have a positive impact on the equity market. This finding is similar to Lucca and Moench (2015) who find larger pre-FOMC

[15]Notice this return is calculated based on the price at 13:30 CET on the event day and 13:30 CET on the day before the announcement.

returns for the easing cycle but with an insignificant estimate.[16] Also the announcement of supportive measures by the ECB should be positive news for the stock market as reflected in the positive coefficient for the UMP variable. Nevertheless, both estimates are not statistically different from zero suggesting that they are not a major driver of the pre-ECB announcement return. In general, these results do not provide support for a channel where the leakage of (on average) good information, e.g. a lower interest rate or non-standard monetary policy measures, are responsible for the pre-ECB announcement return (see e.g. Bernile et al., 2016; Cieslak et al., 2019).

2.5.2 Market Expectations

A positive pre-announcement return can also be the result of market expectations of supportive central bank actions. To test for this hypothesis we first construct expectations measures for monetary policy decisions using textual analysis tools, following the methodology of Lucca and Trebbi (2011).[17] We use all articles containing the term 'European Central Bank' and that were published in the week before the start of the pre-ECB drift.[18] Like in Lucca and Trebbi (2011) we first split the articles into the respective sentences and then filter out the ones relevant for the ECB's monetary policy decision. We define a sentence to be relevant if it contains the term 'interest rate'. Then we count the occurrence of words associated with positive rate changes based on the following list $\mathbf{P} = \{$'increase', 'rise', 'high'$\}$ and negative ones $\mathbf{N} = \{$'fall', 'decrease', 'low'$\}$. The final expectation measure for the interest rate ($Exp\ MRO$) is then calculated as the log ratio of the number of terms associated with a positive and negative rate change. Correspondingly we define an expectation measure for the non-standard monetary policy measures ($Exp\ UMP$) based on the number of sentences that contain any word out of the following list $\mathbf{M} = \{$'QE', 'easing', 'stimulus', 'liquidity', 'purchases', 'support', 'help', 'rescue',

[16] Note, during our sample period the ECB mostly lowered the interest rates because of the adverse economic developments during the euro crisis. Therefore, the period from 2010 to 2015 can be interpreted as a period of monetary policy easing.

[17] We use this textual analysis approach since it allows us to extract expectations for both standard and non-standard monetary policy measures. Using information from the derivatives markets or surveys would only provide us with an expectation on the interest rate decision.

[18] Note, since our pre-ECB announcement return starts already the day before the event day t we take all articles from day t-2 to t-9.

Table 2.7: **Time Series Regressions - Euro Stoxx 50: Monetary Policy**
The dependent variable of the regression model is the time series of log excess returns calculated from the 13:30 CET price on date t-1 and the price at 13:30 CET on date t on scheduled ECB monetary policy decisions. 'ΔMRO' denotes the change in the policy rate announced at the ECB meeting. '*UMP*' is a dummy varaible that takes the value one for ECB meetings at which unconventional monetary policy actions are announced and zero otherwise. '*Exp MRO*' ('*Exp UMP*') captures the market's expectations of the interest rate decision (announcement of UMP measures) based on the newspaper coverage prior to ECB meetings. '*Monetary Policy Shock*' denotes the yield changes of the German Schatz futures from 13:40 to 15:30 CET on the announcement day, respectively from 13:40 to 14:25 (14:25 to 15:30) for the '*Target Shock*' ('*Communication Shock*'). '*ES50*' is the return of the Euro Stoxx 50 for the respective time window on the announcement day. The sample period is from 01/01/2010 to 06/30/2015. Robust t-statistics shown in parenthesis. *, ** indicate significance at the 5% or 1% level.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
Constant	0.413**	0.089	0.572**	0.550**	0.534**	0.473**	0.542**	0.647**	0.661**
	(2.76)	(0.24)	(4.49)	(4.41)	(3.85)	(3.51)	(3.98)	(5.26)	(4.32)
ΔMRO	-2.277								
	(-0.94)								
UMP	0.363								
	(1.26)								
Exp MRO		-0.189							
		(-0.22)							
Exp UMP		5.257							
		(1.63)							
Monetary Policy Shock			-5.698**						
			(-3.31)						
Target Shock				0.940					
				(0.31)					
Communication Shock				-7.008**					
				(-3.65)					
ES50 13:30-14:30					-0.049				
					(-0.09)				
ES50 14:30-14:50						-0.304			
						(-1.78)			
ES50 14:50-15:30							1.089**		
							(4.70)		
ES50 15:30-17:30								0.815**	
								(5.05)	
ES50 13:30-17:30									0.423**
									(3.38)
Adj. R^2 [%]	3.5	1.4	13.0	17.3	0.0	1.9	11.2	29.0	16.6

'bailout'}.[19] Finally, we normalize the number of occurrences by the total number of articles.

In Table 2.7 we can see that expectations about a potential interest rate change do not correlate with the pre-announcement return series. The estimate is negative with a magnitude of -0.2 which is not statistically significant. In contrast the estimate for the expected unconventional monetary policy measures is positive with an estimate of 5.3 and has a p-value slightly higher than 10% suggesting that our parsimonious measure is able to capture expectations about non-standard monetary policy actions. In general, this finding provides weak evidence for the hypothesis that speculations about supportive central bank actions drive the pre-ECB run-up.

2.5.3 Monetary Policy Surprises

In a next step we want to understand how information published in the ECB's announcement relates to the pre-announcement return. As a proxy for the news revealed we are going to use the overall monetary policy shock as well as its decomposition into target and communication shock as defined in Leombroni et al. (2017) and also applied in Schmeling and Wagner (2019).[20] We calculate these shocks using intra-day changes of the German Schatz futures prices for the respective time window.[21] The target shock covers the time period of the ECB's press release (13:40 to 14:25 CET) and the communication shock the press conference from 14:25 to 15:30 CET. The overall monetary policy shock is then defined as the change from 13:40 to 15:30 CET. We calculate the surprise components for both the fixed income and equity market. For the latter we use the tick-by-tick values of the Euro Stoxx 50.

Column (3) of Table 2.7 reports a negative loading for the monetary policy shock with an estimate of -5.7 which is significant to the 1% level. The negative relationship for the yield

[19] The choice of this parsimonious wordlist is motivated by a close monitoring of the news coverage of ECB events around the announcement of non-standard monetary policy measures. For example, the headline of a news item on the Bloomberg newswire in March 2015 was 'European Stocks Advance as Investors Seek ECB *Stimulus* Details'.

[20] Leombroni et al. (2017) build upon the methodology of Gürkaynak et al. (2005) which is also used in Lucca and Moench (2015) to calculate the surprise measures.

[21] We first calculate log returns based on the futures trade prices and then convert these returns into approximative bond yield changes following Cieslak and Schrimpf (2018).

changes of the Schatz futures implies a positive correlation for the futures returns. Thus a higher pre-announcement drift predicts higher returns for the short-end of the German yield curve after the announcement. Column (4) shows that this finding is driven by the communication part of the monetary policy shock with a remarkable high R^2 of almost 20%. These findings are different to the U.S. where Lucca and Moench (2015) report for the pre-FOMC announcement drift that the ex-post policy surprise is not associated with the magnitude of the ex-ante returns. A reason for this difference could be that the FOMC meetings analyzed in Lucca and Moench (2015) are not accompanied by a press conference[22] whereas our ECB meetings were always followed by a press conferences.
When looking at the ex-post surprise in the equity market we see that the results depend on the specific time window that is used for the calculation of the shock. The overall post-announcement return from 13:30 to 17:30 CET exhibits a positive relationship with the pre-ECB run-up. The different sub-windows reveal that there is first a negative coefficient for the surprise from 14:30 to 14:50 (p-value $<$ 10%) which then then turns positive for the shock from 14:50 to 15:30 and also for the return until the end of the trading day (with a notable R^2 of almost 30% for the latter). The first time window corresponds to the time of the press conference where the intro statement is presented by the ECB chairman and is then followed by the Q&A session (until 15:30). Thus this result suggests that investors were first disappointed from news revealed during the intro statement, whereas the questions and answers session reaffirmed the previous run-up in equity prices. For the U.S. Lucca and Moench (2015) does not find a significant correlation between pre- and post-announcement returns in the equity market.[23]

2.5.4 Uncertainty

We now assess whether the pre-ECB announcement returns are related to market uncertainty as measured by the EURO STOXX 50 Volatility Index (VSTOXX).[24] The

[22]Only since 2019 every FOMC meeting is followed by a press conference.

[23]Notice in unreported results we also regress the pre-announcement return series on the ECB tone level and tone change as defined in Schmeling and Wagner (2019). We do not find a significant relationship between these variables for our time period.

[24]The VSTOXX is the European counterpart to the VIX index and is based on Euro Stoxx 50 realtime options prices.

option-implied volatility index is used by various authors to study uncertainty around FOMC announcements.[25]

We first want to examine how the level of uncertainty is related to the pre-ECB returns. Table 2.8 shows highly significant estimates for the VSTOXX index value two days and one day prior to the announcement and on the event day itself. The estimate is always positive with the highest loading (around 0.08) for the day before the start of the pre-ECB drift. The impact is large with a one standard deviation shock translating into an increase of 0.5% in the pre-ECB return. The R^2 ranges from almost 25% to roughly 6% on the event day. Thus, a large part of the variation in the pre-announcement return series can be explained by the level of uncertainty prior to the ECB meeting. This finding is similar to Lucca and Moench (2015) who find a highly significant relationship between pre-FOMC returns and the two-day lagged level of the VIX, though the R^2 is roughly half the size with 10%.

In a next step we want to understand how the change in uncertainty affects the pre-ECB drift. Columns (6) and (7) of Table 2.8 show that a drop in uncertainty on the day prior to the ECB announcement and on the event day significantly increases the pre-ECB return. E.g. for the one-day lagged change in VSTOXX a one standard deviation shock implies a large 0.7% increase in the pre-ECB drift. Also this change in uncertainty can explain a considerable amount of the time series variation in the pre-ECB announcement return, with a R^2 of around 30%. Looking at the intra-day uncertainty changes on the event day we can see that the change in the VSTOXX from the opening of the trading day to shortly before the ECB announcement explains over 30% of the pre-ECB return variation, whereas the changes for the second half of the day only account for around 10%. In general, such an uncertainty resolution can reflect new information entering the market, e.g. through information leakage or arising speculations in the market place about the central bank's actions.

To sum-up, we observe that heightened uncertainty, especially before the announcement, and uncertainty resolution shortly prior to the event lead to significantly larger pre-ECB

[25] See, for example Lucca and Moench 2015; Boguth et al. 2018; Gu et al. 2018; Hu et al. 2019.

returns. This finding is similar to Hu et al. (2019) who report a pattern of large pre-FOMC returns after heightened VIX values with the uncertainty dissolving mostly prior to the announcement.

2.5.5 Economic Activity

We next want to analyze whether the pre-ECB announcement returns depend on fundamental economic activity. We therefore consider, as in Lucca and Moench (2015), annual growth rates of industrial production (IP) and annual inflation as measured by the consumer price index (CPI). Similar to the results in Lucca and Moench (2015) we find both estimates not to be statistically different from zero and with no explanatory power. Thus we conclude that the pre-ECB return series does not strongly co-vary with the business cycle or inflation dynamics.

To sum-up, based on our regression analysis we discovered two strong relationships for the pre-ECB announcement return series. On the one hand it is highly driven by the level of uncertainty prior to the event and its dissolution, with a R^2 ranging from around 25% to over 30%. On the other hand we find a strong relationship to the post-announcement returns in the fixed income and equity market, that starts with the begin of the press conference (R^2 from 17% to 29%). The sign depends on the specific time window, in particular whether the shock covers the intro statement or the Q&A session. This suggests that the pre-announcement run-up is related to information revealed during the communication part of the announcement.

2.6 Potential Explanations

In this section we discuss potential explanations for the 24h pre-ECB announcement return. Therefore we first briefly present the intuition of the respective theory and then test its implications for the pre-ECB drift. In general, a valid theory has to explain the following six characteristics: (1) positive price drift around an event, (2) pre-announcement

Table 2.8: **Time Series Regressions - Euro Stoxx 50: Economic Activity and Uncertainty**
The dependent variable of the regression model is the time series of log excess returns calculated from the 1:30 p.m. price on date t-1 and the price at 1:30 p.m. on date t on scheduled ECB monetary policy decisions. 'Δ^{12}*Log(IP)*' ('Δ^{12}*Log(CPI)*') denotes the 12-month log change of the Industrial Production Index (Consumer Price Index). '*VSTOXX*' is the level of the VSTOXX index at the market close. '*VSTOXX lag1*' ('*VSTOXX lag2*') denotes the level of the VSTOXX one (two) day(s) prior to the ECB meeting. 'Δ*VSTOXX*' captures the change in the VSTOXX from the current market close to the previous one. 'Δ*VSTOXX lag1*' ('Δ*VSTOXX lag2*') is the change in the VSTOXX from the market close two (three) days prior to the ECB meeting to the market close one (two) days before the event. 'Δ*VSTOXX 09:00-13:30*' ('Δ*VSTOXX 13:30-17:30*') captures the change in the VSTOXX on the event day between 09:00 (13:30) and 13:30 (17:30) CET. The sample period is from 01/01/2010 to 06/30/2015. Robust t-statistics shown in parenthesis. *, ** indicate significance at the 5% or 1% level.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
Constant	0.334	-1.424**	-1.368**	-0.595	0.519**	0.446**	0.372**	0.225*	0.559**
	(1.84)	(-4.24)	(-3.77)	(-1.84)	(4.13)	(3.89)	(2.89)	(1.99)	(3.99)
Δ^{12}*Log(IP)*	1.732								
	(0.35)								
Δ^{12}*Log(CPI)*	0.092								
	(0.81)								
VSTOXX lag2		0.081**							
		(6.20)							
VSTOXX lag1			0.080**						
			(5.60)						
VSTOXX				0.048**					
				(3.16)					
Δ*VSTOXX lag2*					0.028				
					(0.95)				
Δ*VSTOXX lag1*						-0.149**			
						(-5.23)			
Δ*VSTOXX*							-0.090**		
							(-3.01)		
Δ*VSTOXX 09:00-13:30*								-0.370**	
								(-4.40)	
Δ*VSTOXX 13:30-17:30*									-0.078*
									(-2.27)
Adj. R^2 [%]	0.8	23.2	17.7	6.2	2.0	30.3	28.2	32.0	11.2

timing of price run-up, (3) variation in the time series of pre-ECB returns, (4) variation in the cross-section (industries and countries), (5) mean-reversion following the meeting and (6) correlation between pre- and post-announcement return.

2.6.1 Ai and Bansal (2018)

Ai and Bansal (2018) develop a theoretical model that allows macroeconomic announcements 'to carry information about the prospect of future economic growth' and find that generalized risk sensitivity in investor's preferences is crucial to generate a non-negative announcement premium. For example under the robust control preferences of Hansen and Sargent (2001), the uncertainty-averse investor overweighs bad states before the announcement which results in an extra discounting of the payoffs. After the release of the macroeconomic information the uncertainty is resolved and asset prices are discounted using marginal utilities which then gives rise to a positive announcement premium.

Hypotheses Testing

Next we want to test more formally the implications of the here discussed theory for our observed pre-ECB drift. Therefore, we are going to discuss each of the six characteristics as stated above: (1) According to the theory of Ai and Bansal (2018) investors prefer the resolution of uncertainty and thus require a risk premium to hold risky assets around an important announcement. This implies a positive price drift as documented in the upper panel of Figure 2.2 and Table 2.3. (2) In general, the premium is realized when the information gets revealed and thus the uncertainty is resolved. This implies that the price increase occurs at the time of the announcement. If information would enter the market already prior to the event, e.g. via an information leakage channel, then the uncertainty could be resolved before the announcement and thus give rise to a pre-ECB announcement return. Table 2.8 shows that the resolution of uncertainty one day before and on the morning of the announcement can explain around 30% of the pre-ECB return series, indicating that new information is entering the market place already prior to the announcement. (3) Ai and Bansal (2018) explain the time series variation of announcement returns due to 'differences in their informativeness and the significance of their welfare implications'. In Table 2.8 we can observe that the magnitude of the pre-ECB drift strongly comoves with the level of uncertainty prior to the event. In times of high financial market uncertainty or an economic crisis news from the central bank e.g. about the economic outlook is of high importance to market participants and thus more informative than in a low volatility state. (4) A similar argument holds for explaining the cross-sectional results. The information revealed by the central bank has a different information content for certain industries (e.g. banking sector) as well as the different EU countries (e.g. GIIPS) which gives rise to heterogeneous pre-announcement returns as documented in Table 2.5 and 2.6. (5) The theory of Ai and Bansal (2018) does not provide an explanation for the mean-reversion of the pre-ECB drift and thus (6) also contains no implications for the correlation between pre- and post-announcement returns. To sum-up, the theory of Ai and Bansal (2018) provides a plausible explanation for the pre-ECB announcement return assuming the leakage of information before the event. But it cannot explain why the return mean-reverts with the start of the press conference.

2.6.2 Wachter and Zhu (2018)

Wachter and Zhu (2018) explain the macroeconomic announcement premium using a model with rare events in which scheduled announcements inform agents about a latent disaster probability. For example during the recent euro crisis such a disaster would have been the break-up of the eurozone and the subsequent major economic downturn. To prevent these adverse economic developments the European Central Bank decided on non-standard monetary policy measures which often got revealed on announcement days.[26] Therefore, the meetings of the ECB were highly informative for market participants about the future course of the European sovereign debt crisis and thus the latent disaster probability. In the model of Wachter and Zhu (2018) this 'big news' characteristic of (certain) announcements is reflected in the assumption that they convey full information,

[26]Notice beside the announcement of the unconventional monetary policy measures also news about the economic outlook were of high interest for investors.

that is, they perfectly reveal whether the economy is in a good or bad state. Especially the probability of a rare negative event to the economy (e.g. a eurozone break-up) represents risk that is realized on announcement days and thus investors require a premium to hold assets over the risky announcement period.[27]

Hypotheses Testing

In general, this mechanism is similar to Ai and Bansal (2018) since the revelation of information on an announcement day is associated with uncertainty and to bear this risk investors require a risk premium. Therefore, the arguments from the previous discussion in Chapter 5.1 also hold for Wachter and Zhu (2018). The main difference is that Wachter and Zhu (2018) specifically relate the risk to (negative) information about a disaster probability whereas the model of Ai and Bansal (2018) is more general in relating uncertainty to the release of unknown information. The former specification is particularly appealing in the context of the euro crisis where the fear and probability of a eurozone break-up and a subsequent recession was quite severe and the ECB's announcements revealed important information about the future course of the European sovereign debt crisis. Table 2.9 confirms that several variables that proxy for the economic and policy uncertainty of the euro crisis and eurozone break-up fear show a positive significant relationship to the pre-announcement return series. Overall, the theory of Wachter and Zhu (2018) can rationalize a positive pre-announcement return assuming information entering the market prior to the event. But it also cannot provide an explanation for the subsequent mean-reversion.

2.6.3 Heightened Uncertainty

Hu et al. (2019) argue that the pre-FOMC announcement returns[28] can be interpreted as a premium for heightened uncertainty. In their analysis they do not find conventional measures of risk (like volatility, skewness or kurtosis) to be elevated around FOMC meet-

[27] This preference for early resolution of uncertainty is consistent with the concept of risk-sensitivity as defined by Ai and Bansal (2018), which is a necessary condition for a nonzero announcement premium.

[28] Notice the authors also document pre-announcement returns for other macro releases in the U.S., like Nonfarm Payroll, GDP, and ISM.

ings but instead document heightened uncertainty in the market as measured by the CBOE VIX. They relate this uncertainty to market-moving information that is released on announcement days. They show that prior to the announcement uncertainty slowly builds up which results in initial price depression, the resolution of uncertainty shortly before the announcement then gives rise to the significant increase in price, as documented for the pre-FOMC announcement return. This pattern of heightened uncertainty is particularly pronounced for FOMC meetings that exhibit a high pre-FOMC drift.

Hypotheses Testing

In the following we test whether Hu et al. (2019) can rationalize the 24h pre-ECB announcement return: (1) Similar to Ai and Bansal (2018) and Wachter and Zhu (2018) the authors argue that uncertainty related to market-moving information revealed during macro and central bank announcements can generate a positive premium as can be observed in the upper panel of Figure 2.2 and Table 2.3. First, this (heightened) uncertainty prior to the event results in price depression which then gives rise to increasing prices when the uncertainty is resolved. Table 2.8 documents a strong positive relationship between the two-day lagged level of uncertainty as measured by the VSTOXX and the magnitude of the pre-ECB drift. Also it shows that a resolution in uncertainty prior to the ECB meeting leads to increasing stock price, explaining nearly a third of the variation in the return time series. (2) Compared to the previous two theories in this chapter Hu et al. (2019) state that the price drift should be observed prior to the meeting, as for the pre-ECB return, due to the observed resolution of uncertainty. (3) The authors also argue that the magnitue of the pre-announcement return depends on the uncertainty prior to the event: E.g. high uncertainty before the announcement gives rise to large pre-FOMC returns. For the pre-ECB drift we also find evidence for such a relationship (see Table 2.8). (4) The uncertainty related to market-moving information does not need to affect every industry or country the same way. E.g. cyclical industries (like the automotive sector) are more sensitive to economic outlook information revealed during a central bank's meeting than for example the food and beverages sector. Thus, we can observe heterogeneous pre-announcement returns for different industry sectors and country's blue chip indices as reported in Table 2.5 and 2.6. (5) Like the previous two explanations from this chapter Hu et al. (2019) cannot rationalize the mean-reversion of the pre-ECB return and therefore (6) do not provide any (testable) implications for the relationship between pre- and post-announcement returns.

To sum-up, Hu et al. (2019) provide a plausible explanation for the pre-ECB announcement return based on heightened uncertainty but cannot rationalize the mean-reversion that follows.

2.6.4 Buy on the Rumor, Sell on the News (BRSN)

The previous explanations associate uncertainty around an information event with being negative since investors either apply an extra discounting to the cash flows due to the uncertainty or face actual risk of bad market-moving news which lets them require a positive risk premium. In contrast, Peterson (2002) demonstrate a relationship between investor psychology and asset prices around scheduled events explaining that prices first rise prior to and then mean-revert after the positively anticipated announcement.

According to Peterson (2002) prior to a positively anticipated event investors are in a positive affect sate (e.g. euphoria) due to the anticipated reward. This affect leads to an increasing risk-taking and purchasing behavior and thus increasing prices. After the delivery of the expected reward, the (positive) news, investors' affect regresses to neutral which results in a net decrease in positive affect, leading to decreased risk-aversion and protective investing behavior like selling, and thus a mean-reversion in asset prices.

Hypotheses Testing

We next test the implications of this rationale for the 24h pre-ECB announcement return: (1) To generate a positive premium, as documented in the upper panel of Figure 2.2 and Table 2.3, the mechanism as described in Peterson (2002) requires a positively anticipated announcement. We argue that the monetary policy decisions of the European Central Bank represent such an event since these meetings are scheduled in advance and, especially in times of financial stress or a crisis (e.g. the European sovereign debt crisis), market participants expect the ECB to support the European economy which represents positive news for financial markets.[29] In addition, Column (2) of Table 2.7 provides (weak) evidence for such speculations or rumors prior to ECB meetings with a p-value of roughly 10% for the $Exp\,UMP$ variable. This variable proxies in general for the market's expectations of unconventional monetary policy measures but can also be interpreted as speculations or rumors in the market place regarding supportive measures.[30] (2) According to Peterson (2002) the price drift should occur prior to the event due to the anticipation of reward which leads to purchasing behavior and thus rising stock prices. This is consistent with the in this study documented pre-ECB drift (see e.g. top panel of Figure 2.2). (3) Larger pre-announcement returns could be explained in this framework by stronger purchasing behavior caused by a more positive affect state due to a larger anticipated reward. Such a large anticipated reward could have been the prospect of increasing stock prices after the ECB announcing supportive measures for the economy particularly in times of high uncertainty. Thus, we would expect higher pre-ECB returns in times of high uncertainty as documented in Table 2.8. (4) The anticipated reward, the rising stock prices, can be different for industries and countries since the central bank's information and actions have a different impact on the various sectors or blue-chip indices of the European countries. This can explain the heterogeneous pre-ECB returns as shown in Table 2.5 and 2.6. (5) As the first theory discussed in this chapter Peterson (2002) predicts a mean-reversion after the announcement as observed for the ECB meetings (see upper panel of Figure 2.2). Following the rationale of the author a reason for this behavior could be that the positive news revealed at the ECB meeting, e.g. the announcement of non-standard monetary policy measures, rarely exceeded the high market expectations of a potential stimulus or bailout. Thus, on average investors were disappointed which lead to a decrease in positive affect and thus protective investing

[29]In general, monetary policy tend to have an asymmetric impact on stock prices as financial conditions are eased in times of a crisis but not tightened accordingly in good times. This is called a 'government put' in the literature (see e.g., Diamond and Rajan, 2012).

[30]In their analysis of the ECB's foreign exchange interventions in the beginning of the 21th century Fatum and Hutchison (2002) similarly extract rumors of such interventions from related newswire reports.

behavior like the selling of stocks. (6) If market participants were indeed disappointed after their high expectations we would expect a negative correlation between pre- and post-announcement returns. Table 2.7 shows a weakly significant negative relationship (only to 10% level) between the pre-ECB return and the post-announcement return for the time window from 14:30 to 14:50 CET. This is the time of the press conference where the president of the ECB presents the introductory statement. Whereas the press release (for our sample period) only comprises the interest rate decision the intro statement contains also the announcement of non-standard monetary policy measures. This finding documents a disappointment in the stock market after a corresponding pre-announcement increase.

To sum-up, all of the here discussed theories and explanations can rationalize our main finding from the previous chapters, namely an in general positive price drift around announcement days which covaries strongly with uncertainty prior to the event and also the resolution of this uncertainty. The latter also documents that information are entering the market already before the announcement which could indicate the possibility of information leaks or could be a sign for speculations and rumors that are building-up before the event. Whereas the first three theories (Chapter 5.1 to 5.3) interpret the uncertainty on these event days due to the announcement of unknown information as negative, the theory of Peterson (2002) argues based on investor psychology that a positively anticipated event can also give rise to a price run-up. In addition, it is the only explanation that predicts a mean-reversion after the announcement due to market participants being disappointed from the news revealed. We find weak empirical evidence for such a BRSN pattern. Overall, we conclude that there are two different fundamental mechanisms, uncertainty resolution and positive anticipation, that can rationalize the positive pre-ECB drift but only one that can, in addition, explain the observed mean-reversion.

2.7 ECB Announcement Returns and Euro Crisis Uncertainty

In the previous chapters we have learned that the pre-ECB drift is a rather recent phenomenon, occurring only for the time period of the euro crisis. In addition, we have seen that it strongly co-moves with uncertainty prior to the respective ECB meeting, as measured by the VSTOXX. Figure 2.4 visualizes this main finding. The plot shows that the uncertainty peaks of the VSTOXX correspond to major events of the European sovereign debt crisis. Giving this observation we are now going to study the relationship between the pre-ECB announcement return and the uncertainty of the euro crisis in more detail.

2.7.1 Risk Aversion, Uncertainty and Eurozone Break-up Concerns

The VSTOXX index is a general measure of equity market uncertainty. Therefore in a first step we apply the methodology of Bekaert et al. (2013) to decompose the VSTOXX index into proxies for uncertainty and risk aversion.[31] The former is estimated as the conditional variance obtained from regressing the realized variance on the lagged realized variance and the lagged squared VSTOXX. The difference between $VSTOXX^2$ and the conditional variance measure serves then as a proxy for risk aversion.

Table 2.9 shows that the results are similar to when using the original VSTOXX index. We observe a larger pre-ECB announcement return in times when risk aversion or uncertainty in the market is higher. Both estimates are highly significant to the 1% level with a t-statistic larger than 5. Also the R^2 with around 22% is considerable large, comparable to when using the pure VSTOXX variable. Therefore, we conclude that the result is not just driven by the expected volatility component of the VSTOXX but also by fundamental risk aversion.

[31] Schmeling and Wagner (2019) also use the methodology of Bekaert et al. (2013) to decompose the VSTOXX.

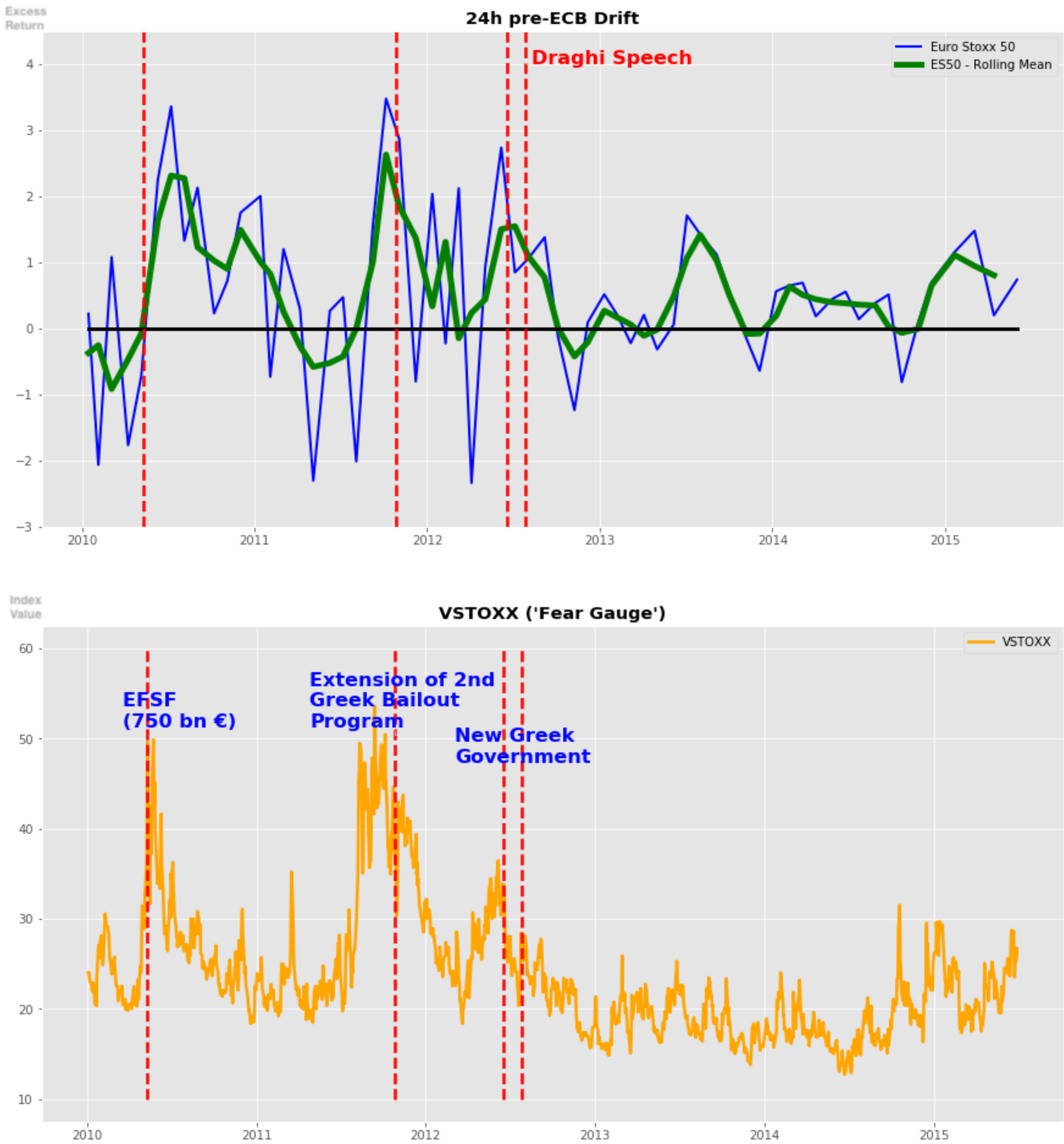


Figure 2.4: **Daily Pre-ECB Drift Returns and Uncertainty During the Euro Crisis**
This figure plots the 24h pre-ECB announcement excess return together with the VSTOXX index. The upper panel depicts the realized 24h excess return (blue line) and a quarterly moving average (green line). The log excess return series is calculated based on the 1:30 p.m. price on date t-1 and the price at 1:30 p.m. on date t, for t being an ECB announcement day. The lower panel plots the time-series of the two-day lagged VSTOXX index with a couple of major political events. The time period is 01/01/2010 to 06/30/2015.

Next we want to understand what is driving the risk aversion observed prior to scheduled ECB monetary policy decisions. Therefore, we regress the risk aversion variable on several proxies for economic and policy uncertainty in Europe. We first use the Economic Policy Uncertainty Index (EPU) for Europe as created by Baker et al. (2016). The index relies on frequency counts of articles containing specific combinations of terms associated with economy, uncertainty and policy. Table 2.10 shows a significant positive relationship between the economic and policy uncertainty index and our risk aversion proxy. The EPU Index can explain roughyl 1/4 of the variation in risk aversion.[32] Hardouvelis et al. (2018) build upon the methodology developed in Baker et al. (2016) and construct an index to specifically measure the economic and policy uncertainty related to the Greek economic crisis. In column (2) of Table 2.10 we can see again a significant positive relationship between the Greek specific EPU and the pre-ECB return series with a stronger t-statistics indicating that this measure better reflects the fundamental uncertainty underlying our risk aversion proxy. In addition to the general EPU index the authors also construct several sub-indices covering different categories, e.g. monetary or pension policy.[33] In our analysis we also include the EPUC Index which measures uncertainty related to currency or Grexit possibility. An important keyword for this index is the phrase 'grexit' which directly relates to the fear of Greece leaving the EU and the possibility of a subsequent eurozone break-up. Regressing our risk aversion variable on this sub-index shows again stronger t-statistics than for the previous two EPU indices suggesting that uncertainty related to a potential Grexit of Greece was an important determinant of risk aversion during the euro crisis period. Table 2.9 also shows that this Grexit related uncertainty measure is directly related to the magnitude of the pre-announcement return.

To test more directly for this eurozone break-up risk we use the Sentix Euro Break-up Index (EBI) which is based on a survey among individual and institutional investors and captures euro break-up news. Table 2.10 documents again a significant positive

[32]Note as a robustness exercise we also regress our economic and policy uncertainty measures directly on the pre-announcement return series. The results can be found in Table 2.9.

[33]The construction of these indices follows always the same methodology, the only difference is the selection of the corresponding keywords.

Table 2.9: **Time Series Regressions - Euro Stoxx 50: Economic and Political Uncertainty**
The dependent variable of the regression model is the time series of log excess returns calculated from the 1:30 p.m. price on date t-1 and the price at 1:30 p.m. on date t on scheduled ECB monetary policy decisions. '*VSTOXX lag2*' denotes the level of the VSTOXX index at the market close two days prior to the ECB meeting. '*Uncertainty*' ('*Risk Aversion*') proxies the uncertainty (risk aversion) component of the two-day lagged VSTOXX following the methodology of Bekaert et al. (2013). '*EPU Europe*' is the economic policy index of Baker et al. (2016) for Europe. '*EPU Greece*' denotes the respective index for Greece as created by Hardouvelis et al. (2018) and '*EPUC Greece* ' a sub-index of the same authors measuring currency or Grexit possibility. '*Sentix EBI*' is the Sentix Euro Break-up Index and '*ECB Uncertainty Index*' our own created uncertainty index using the press conference transcripts of the European Central Bank. The sample period is from 01/01/2010 to 06/30/2015. Robust t-statistics shown in parenthesis. *, ** indicate significance at the 5% or 1% level.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Constant	-1.432**	-0.078	-0.425*	-0.654	-0.534	-0.226	0.106	0.231
	(-4.26)	(-0.55)	(-2.34)	(-1.04)	(-1.45)	(-0.80)	(0.81)	(1.49)
VSTOXX lag2	0.081**							
	(6.26)							
Risk Aversion		0.003**						
		(7.53)						
Uncertainty			0.002**					
			(5.40)					
EPU Europe				0.007				
				(1.80)				
EPU Greece					0.009**			
					(2.83)			
EPUC Greece						0.006**		
						(3.53)		
Sentix EBI							0.015**	
							(2.95)	
ECB Uncertainty Index								0.241**
								(2.64)
Adj. R^2 [%]	23.3	21.6	22.6	4.9	4.2	6.8	11.3	6.9

relationship to our risk aversion proxy suggesting that the fear of a eurozone break-up was a major driver of risk aversion. The regression also shows a remarkable R^2 of 37% which underlines the strong relationship between these two variables. But one drawback of this measure is that it is only available since Summer 2012 and thus misses the first two stress periods of the euro crisis. Therefore, we are going to construct now our own European uncertainty index using textual analysis of the ECB's press conference transcripts.

Table 2.10: **Time Series Regressions - Risk Aversion: Economic and Political Uncertainty**
The dependent variable of the regression model is the value of the risk aversion proxy as defined in Chapter 6.1 two days prior to scheduled ECB monetary policy decisions. '*EPU Europe*' is the economic policy index of Baker et al. (2016) for Europe. '*EPU Greece*' denotes the respective index for Greece as created by Hardouvelis et al. (2018) and '*EPUC Greece* ' a sub-index of the same authors measuring currency or Grexit possibility. '*Sentix EBI*' is the Sentix Euro Break-up Index and '*ECB Uncertainty Index*' our own created uncertainty index using the press conference transcripts of the European Central Bank. The sample period is from 01/01/2010 to 06/30/2015. Robust t-statistics shown in parenthesis. *, ** indicate significance at the 5% or 1% level.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**
Constant	-224.427	-203.245	-76.255	57.495**	84.086**
	(-1.15)	(-1.8)	(-1.05)	(4.18)	(-3.55)
EPU Europe	2.520*				
	(1.98)				
EPU Greece		3.760**			
		(2.90)			
EPUC Greece			2.216**		
			(3.12)		
Sentix EBI				3.233**	
				(4.88)	
ECB Uncertainty Index					112.443**
					(4.40)
Adj. R^2 [%]	23.7	23.1	36.2	36.6	51.4

2.7.2 ECB Uncertainty Index

In a next step we create our own uncertainty index based on the ECB's economic analysis as revealed in the press conference's intro statement. Based on a textual analysis of these statements we extract the uncertainty keywords that characterize the economic environment. For example, in January 2010 the 'Governing Council continues to view the risks to this outlook as broadly balanced',[34] whereas in May 2010 an additional passage was added: 'The Governing Council continues to view the risks to this outlook as broadly balanced, in an environment of *high uncertainty*'.[35] This phrase reflected the increasing tensions in the financial markets around the debt problems of Greece and the danger of a potential Grexit. In summer 2012, before the famous 'Whatever it takes' speech of Draghi on the 26th of July the statement observes 'renewed weakening of economic growth and heightened uncertainty' for the euro area economy.[36] At the end of the year after Draghi's speech the economic environment was then only characterized as times of 'high uncertainty' (November 2012).[37] In December 2013 the statement uses the phrase 'related uncertainties'[38] and in October 2014 the terms 'uncertainty' or 'uncertainties' were completely dropped.[39] We convert this uncertainty characterization of the economic environment into a simple wording indicator variable using a four-value scale[40]: '0' indicates that the term 'uncertainty' is not used at all and '4' characterizes an environment of 'particularly high uncertainty'.[41]

Figure 2.5 plots the 24h pre-ECB announcement return together with our ECB Uncertainty Index. It highlights, that the 24h pre-ECB announcement return peaks at times when the ECB's uncertainty classification takes the highest values. To assess this relationship more formally we next regress the pre-announcement return series on this newly

[34] https://www.ecb.europa.eu/press/pressconf/2010/html/is100114.en.html.
[35] https://www.ecb.europa.eu/press/pressconf/2010/html/is100506.en.html.
[36] https://www.ecb.europa.eu/press/pressconf/2012/html/is120705.en.html.
[37] https://www.ecb.europa.eu/press/pressconf/2012/html/is121206.en.html.
[38] https://www.ecb.europa.eu/press/pressconf/2013/html/is131205.en.html.
[39] https://www.ecb.europa.eu/press/pressconf/2014/html/is141002.en.html.
[40] Note, such an approach is often used in the former literature that analyzed the impact of central bank communication on financial markets, see e.g. Rosa (2011).
[41] An example phrase for the value '1' would be 'environment of uncertainty', for the value '2' 'uncertainty is elevated' and for value '3' 'heightened uncertainty'.

created index. Table 2.9 shows a positive relationship which is significant to the 1% level between the pre-announcement return series and our newly created ECB Uncertainty Index confirming the previous visual impression.

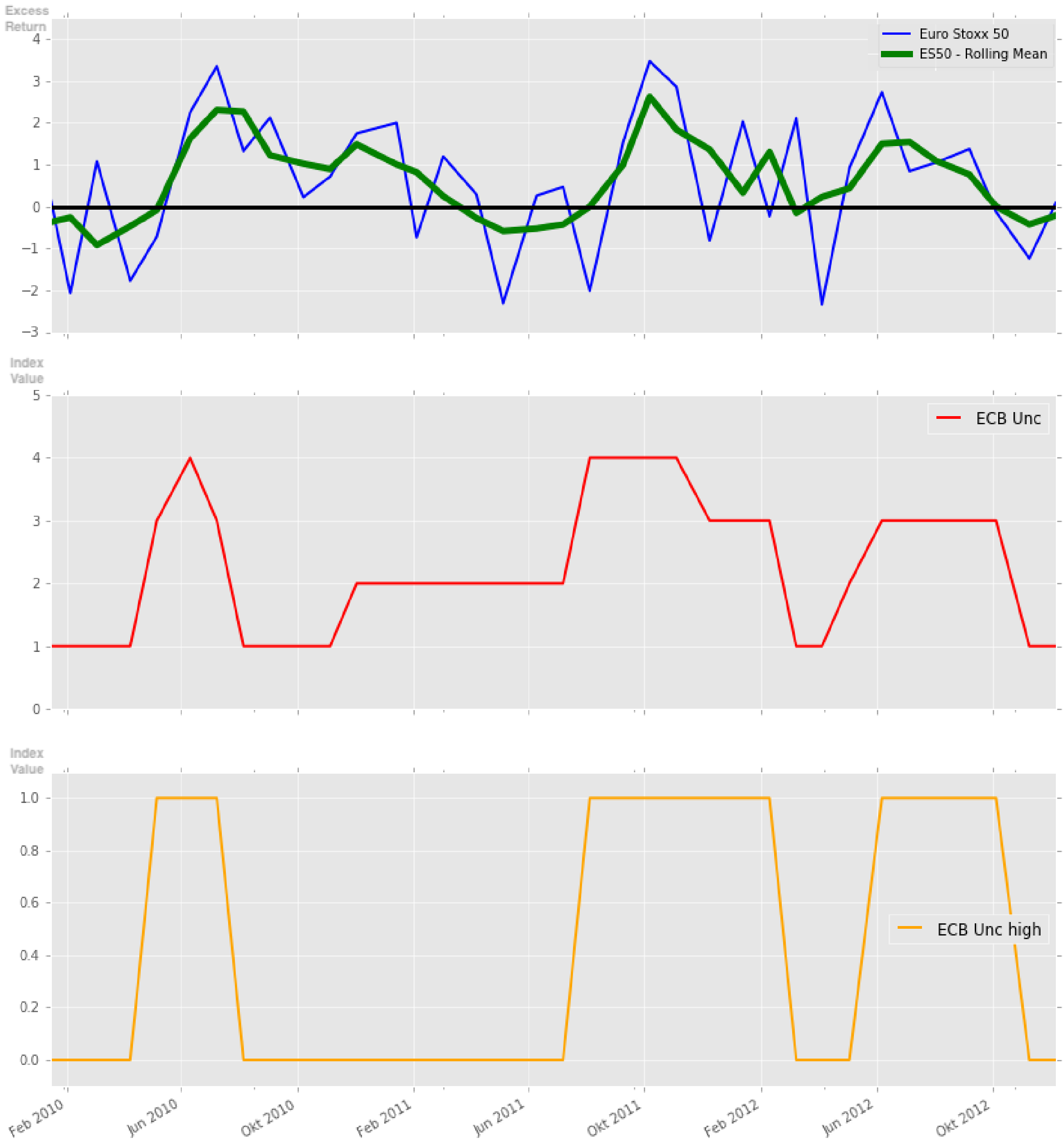


Figure 2.5: **Pre-ECB Announcement Returns and the ECB Uncertainty Classification**
This figure plots the 24h pre-ECB announcement return together with the ECB's uncertainty evaluation, for the period January 2010 to December 2012. The upper panel plots the realized 24h pre-ECB announcement return (blue line) and a one-quarter moving average (green line). The middle panel depicts the ECB's uncertainty characterization of the Euro economy. The lower panel maps the ECB uncertainty into states of low and high Euro area uncertainty.

In a further analysis, similar to Hu et al. (2019) and Martello and Ribeiro (2018), we split our sample of ECB meetings and then re-calculate the average cumulative announcement return series for each class. Using our newly created ECB Uncertainty Index we create a high uncertainty group (index values '3' and '4') and a low uncertainty category which comprises all the remaining meetings. The upper panel of Figure 2.6 plots the cumulative announcement returns for the high uncertainty group. One can see that the pre-ECB drift first rises to a level above 1% and then drops sharply with the start of the ECB's press conference. The lower panel shows the same plot for the low uncertainty meetings.[42] One can clearly see that the pre-announcement drift for this group is much less pronounced than for the high uncertainty periods. It only rises to around 0.4% which is over 2.5 times lower than in the upper panel. Also the mean-reversion following the announcement is much smaller. This observation is consistent with Martello and Ribeiro (2018) who argue that the finding of Lucca and Moench (2015) actually only occurs for FOMC meetings during stressed periods. A similar observation is also reported in Hu et al. (2019) who show a strong correlation between large pre-FOMC returns and high uncertainty prior to the FOMC announcement. In general we can conclude that the pre-ECB announcement drift is mainly driven by rather a few ECB meetings which were held in periods of high uncertainty.[43][44]

In a last step, we are going to connect these 'high uncertainty' meetings to the fundamental economic developments in the eurozone using the respective ECB transcripts. In particular, we are going to exploit the information content of the ECB's Q&A session since the reporters' questions are often about the current economic situation that caused the intro statement's uncertainty classification in the first place.

Figure 2.5 shows that there are three distinct high uncertainty periods in the time frame from 2010 to 2012; another one can also be observed around September and October

[42] To better compare the magnitude of both pre-announcement drifts we deployed the same scale in the upper and lower panel.

[43] Only 18 out of 86 meetings are classified as 'high uncertainty' which is around 20% of all announcements.

[44] This finding is also in line with Ben Dor and Rosa (2019) who report that 'one single FOMC meeting contributes roughly 0.1% to the average pre-FOMC effect'.

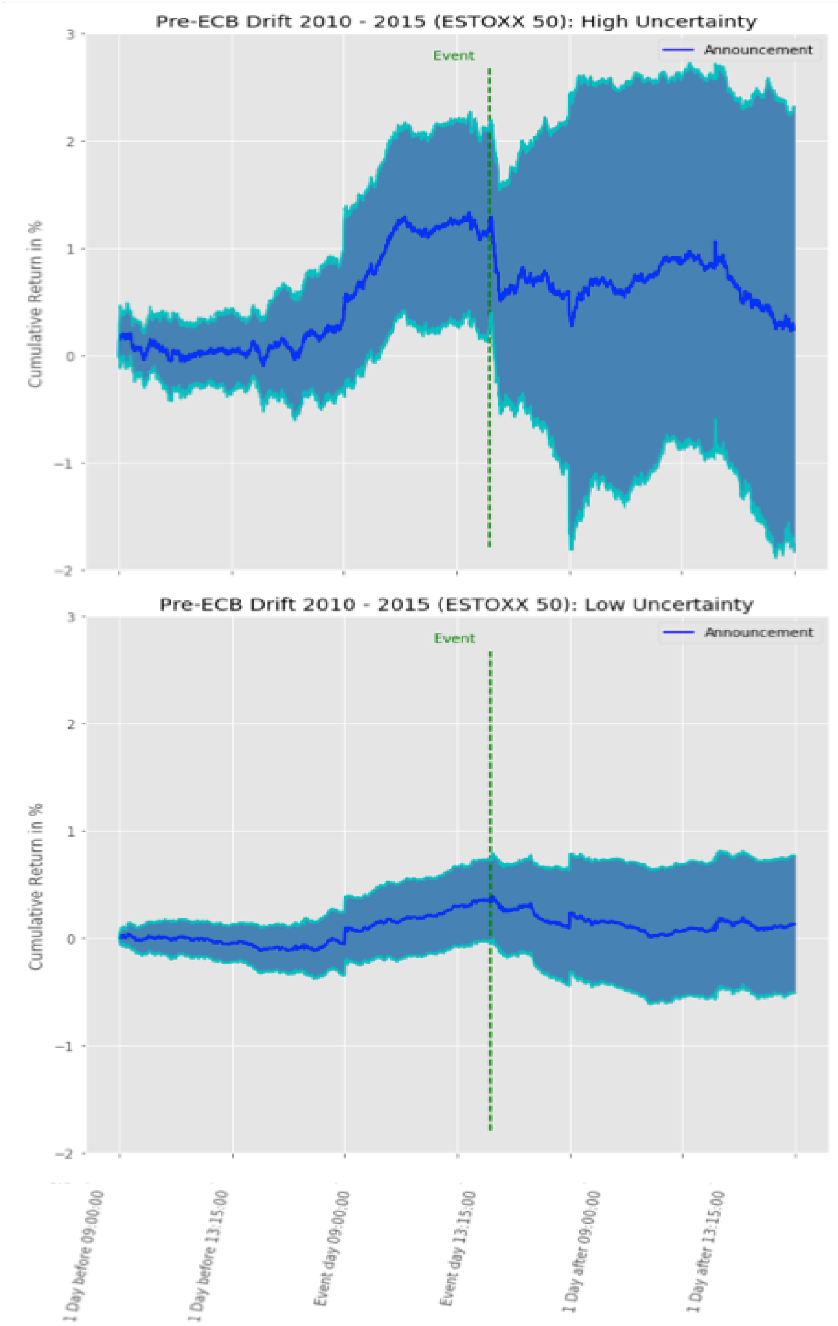


Figure 2.6: **Cumulative Return Series of the Euro Stoxx 50 Index: High and Low Uncertainty**

This figure plots the average cumulative 15 seconds return of the Euro Stoxx 50 on a three day window around scheduled ECB announcement days (i.e. one day prior to the event, event, one day after the event). The blue line is the average return around ECB announcement days, the blue shaded area is the respective pointwise 95% confidence interval. The vertical dashed line marks the start of the ECB press conference, 2:30 p.m. (local time), on the event day. The upper panel comprises all ECB meetings from January 2010 to June 2015 that occur within a period of high uncertainty. The lower panel studies all remaining meetings in the same time period.

2015 (not shown in this plot). The first high uncertainty period started in May 2010 and lasted until July. During the Q&A session of the May 2010 meeting the first reporter asked all of his three questions about the 'consequences of Greece's fiscal crisis on financial markets', supportive ECB actions and whether the European Union needs an 'orderly default procedure for its member countries'.[45] In addition, most of the questions that followed were about the same topic reflecting the first euro crisis peak with the danger of a potential Grexit.[46]

The second high uncertainty period then occurred in summer 2011 and lasted until the beginning of the next year. Again, the dominating topic was the European sovereign debt crisis, in particular its contagion from Greece to other member states like Spain and Italy and the supportive measures of the ECB. For example, in the August meeting one reporter asked all of his three questions about the 'Securities Markets Programme', 'contagion in the eurozone' and about the economic outlook that the ECB characterized as 'particularly high'.[47]

The third stress period occurred in summer 2012. During the Q&A session of the July press conference the first questions of the reporters were again about the ongoing euro crisis, in particular about the European banking sector and that 'some countries [...] are still in kind of credit crunch territory'. Also other journalists asked (and speculated) about further non-standard monetary policy measures, e.g. another LTRO program. Other topics targeted the EFSF and ESM[48] which were created to support European countries which had problems raising funds in the financial markets.[49]

The last high uncertainty period occurred around late summer and autumn 2015. On 27th of June 2015 the Greek Prime Minister Alexis Tsipras announced a referendum

---

[45] See https://www.ecb.europa.eu/press/pressconf/2010/html/is100506.en.html.

[46] At the end of the year 2009 Standard and Poor's downgraded Greek government bonds leading to speculations about a bankruptcy of Greece. After rising tensions and risk premia in the financial markets the EU member states agreed on a first austerity program for Greece finally leading to the set-up of the 750 € bn European Financial Stability Facility (EFSF) in June 2010 (see Illing, 2013). For a detailed description of the Greek crisis also see the timeline of the 'tagesschau' (German television news service): https://www.tagesschau.de/wirtschaft/griechenland640.html.

[47] See https://www.ecb.europa.eu/press/pressconf/2011/html/is110804.en.html.

[48] EFSF stands for the European Financial Stability Facility and ESM for the European Stability Mechanism.

[49] See https://www.ecb.europa.eu/press/pressconf/2012/html/is120705.en.html.

about the bailout conditions in the country's government-debt crisis proposed by the Troika. It took place on 5th of July 2015 and a majority of over 61% voted with 'No' resulting in a new Greek legislative election in September 2015 and re-emerging Grexit and eurozone break-up fears. These adverse developments were also reflected in the ECB's assessment of the economic situation as can be seen in their July statement: '[...] June 2015, again associated with increased political uncertainty related to the breakdown of the negotiations and the announcement of the referendum'. In addition, the Grexit fear was reflected in the reporters' questions in the corresponding Q&A session.[50]

To sum-up, we can observe a close relationship between periods of high uncertainty in Europe and the peaks of the European sovereign debt crisis which can be characterized as times of severe eurozone break-up concerns. Given the finding that the pre-ECB announcement return is especially pronounced for these high uncertainty periods we conclude that the Grexit fear and eurozone break-up concerns have been a major driver of the pre-ECB drift as documented in this study.

2.7.3 Grexit Fear, the European Central Bank and BRSN Pattern

Especially during the high uncertainty times of the euro crisis where the danger of a potential Grexit was quite severe, market participants expected or hoped for supportive actions from the ECB as the lender or buyer of last resort (see Acharya et al., 2018), rendering the ECB's monetary policy decisions a very positively anticipated event in the sense of Peterson (2002) ('buy on the rumor'). Also Figure 2.6 and 2.7 show that the mean-reversion of the pre-ECB announcement return happens entirely during the hour of the communication event suggesting investors being disappointment from the news revealed ('sell on the news'). In particular, we can observe a sharp drop that starts with the begin of the press conference. Taking together these arguments with the preliminary

[50]One of the questions asked by the first journalist was: '[...] question regarding terminology: time out on the eurozone, a potential temporary Grexit that's been talked about in the last week and a half, and at 4am on Monday morning it seemed a potential Grexit was a possibility. I just wondered whether using that terminology has opened Pandora's box on the perception of eurozone membership, and by inference the euro?' See https://www.ecb.europa.eu/press/pressconf/2015/html/is150716.en.html.

evidence for a ‘buy on the rumor - sell on the news’ story in the sense of Peterson (2002) from Chapter 5.4 we now test whether this behavioral anticipation explanation also holds for the high uncertainty periods or even shows stronger explanatory power than for the overall euro crisis period. Therefore, we re-run the regressions of columns (2) and (6) of Table 2.7 for the high uncertainty meetings only. We find neither the estimates for the ‘rumors’ variable nor for the ‘disappointment’ proxy to be significant and. Therefore we cannot conclude that the BRSN explanation following Peterson (2002) fully explains the pattern of the ECB announcement returns. Instead the regression results suggest that the sharp decrease during the press conference is decoupled from the pre-announcement run-up, at least during the high uncertainty periods of the euro crisis.

2.8 Conclusion

With this chapter’s analysis we contribute to the large literature on announcement returns, in particular around central bank decision days. Supplementary to Lucca and Moench (2015) we provide empirical evidence for a pre-ECB announcement return that occurs prior to the scheduled meetings of the European Central Bank and which as opposed to the pre-FOMC drift mean-reverts with the start of the ECB’s press conference. We argue that the pre-ECB announcement return in Europe is a recent phenomenon that started when the course of the Federal Reserve and the European Central Bank diverged with the begin of the European sovereign debt crisis in 2010. We also document that the high uncertainty periods of the euro crisis, when the fear of a eurozone break-up was severe, were a major driver of the average pre-announcement premium. Due to the bank-sovereign nexus (see Acharya et al., 2018) the financial sector was particularly affected by the euro crisis and thus benefitted greatly from the non-standard monetary policy measures which is reflected e.g. in the huge pre-ECB drift for the banking sector (almost 3% for the high uncertainty periods). In addition, find larger pre-ECB returns for cyclical industries (like the automotive sector) and the GIIPS countries of our sample (Italy and Spain). We also document that the CAPM holds

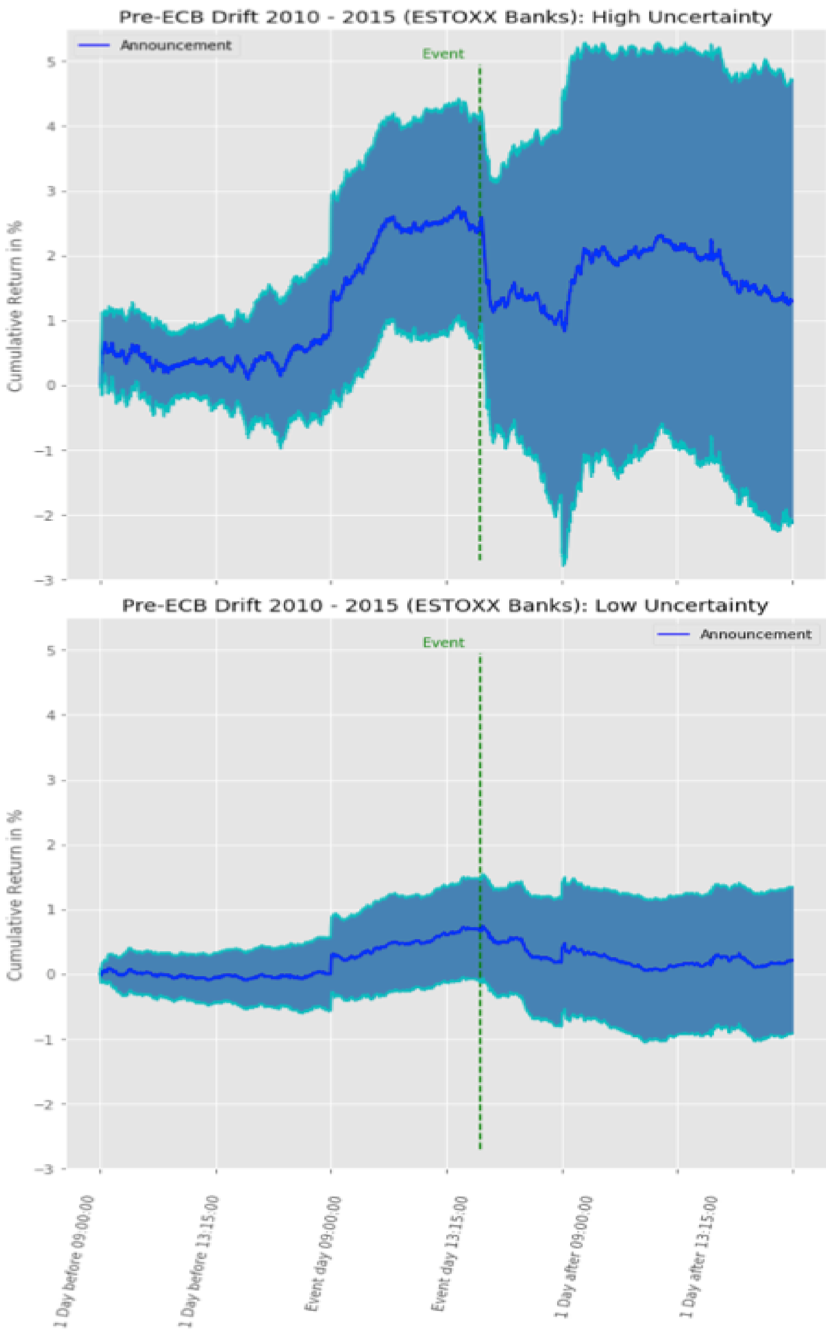


Figure 2.7: **Cumulative Return Series of the Euro Stoxx Banks Index: High and Low Uncertainty**
This figure plots the average cumulative 15 seconds return of the Euro Stoxx Banks on a three day window around scheduled ECB announcement days (i.e. one day prior to the event, event, one day after the event). The blue line is the average return around ECB announcement days, the blue shaded area is the respective pointwise 95% confidence interval. The vertical dashed line marks the start of the ECB press conference, 2:30 p.m. (local time), on the event day. The upper panel comprises all ECB meetings from January 2010 to June 2015 that occur within a period of high uncertainty. The lower panel studies all remaining meetings in the same time period.

(only) on announcement days for European industry and country's blue-chip indices.

The magnitude of the observed pre-announcement return is strongly driven by the level of uncertainty prior to the meeting and also highly co-moves with the resolution of uncertainty shortly before the event (with an R^2 of around 25% to over 30%). The pre-ECB return strongly predicts post-announcement returns in the fixed income and equity market (R^2 ranging from roughly 20% to 30%), especially for the time period of the press conference and its aftermath. The former findings can be reconciled with the theories of Ai and Bansal (2018) and Wachter and Zhu (2018) who argue that the uncertainty around an information event requires investors to demand a risk premium. They are also in line with Hu et al. (2019) who argue that the pre-FOMC drift is a 'premium for heightened uncertainty'. The announcements first bring heightened uncertainty to the market, its subsequent dissolution then gives rise to the pre-event price appreciation. Whereas these explanations are plausible for the pre-ECB return they do not explain the mean-reversion that follows the announcement. In this study we also discuss the behavioral anticipation theory of Peterson (2002) that matches all of the stylized facts of the ECB announcement return for the overall euro crisis period; but when testing it for the high uncertainty periods we cannot confirm these results.

Compared to recent studies that relate the pre-FOMC drift also to heightened uncertainty (see Hu et al., 2019; Martello and Ribeiro, 2018) we connect the general equity market uncertainty as measured by the VSTOXX to the course of the European sovereign debt crisis using textual analsis of the press conference statements. We find that the pre-ECB drift is particularly pronounced for the high uncertainty periods of the euro crisis, where the concerns about a potential Grexit were quite severe, with the magnitude being 2.5 times higher than for the low uncertainty meetings. We therefore argue that the fear and risk about a eurozone break-up was a major driver of the pre-ECB announcement return.

Chapter 3

Textual Analysis in Finance: A Primer

3.1 Introduction

In this chapter we are going to introduce fundamental techniques used in textual analysis and discuss how they are applied in the field of financial economics. Particularly, we focus on the concepts that are necessary to understand Chapter 4 of this thesis. In general, the goal of this primer is to give the reader a sense of how textual analysis is applied in financial studies instead of giving a comprehensive review. For a more thorough and extensive discussion of the natural language processing concepts we refer to Jurafsky and Martin (2017) and Manning and Schütze (2003). For methodological surveys that cover contemporary statistics and machine learning in general we recommend Hull (2019), Bishop (2006) or Murphy (2012).[1]

3.1.1 Introductory Example

In this section we are going to present and discuss a simple introductory example to illustrate the key challenges and problems inherent in the analysis of textual statements.

[1]Note, sometimes we omit a formal introduction of a statistical concept and focus only on the main idea underlying the concept and how it is applied in the economic analysis. In these cases we refer the interested reader to the recommended literature for further details and explanations.

In general, the main goal of textual analysis is the automatic content extraction and knowledge discovery from textual information. For example, every six weeks the European Central Bank announces its monetary policy decision where the Governing Council reveals among others things whether the key ECB interest rates have been changed.[2] These information are of high interest for all market participants since they directly affect the valuation of stocks and bonds. Therefore, the announcement is highly followed and the information is incorporated into asset prices immediately. Processing this newly published information in an automated and thus very fast way can be beneficial for different trading strategies or risk and asset management decisions in general. Also it allows ex-post to analyze which kind of information was driving asset prices.

Usually, the interest rate decision is announced within a short statement. See for example the press conference from 25th of October 2018:

> *"Based on our regular economic and monetary analyses, we decided to keep the key ECB interest rates unchanged."*

The key economic information here is that the central bank did not change the short-term interest rates. A human investor can easily understand the statement and extract the valuable investment insights but how can a machine automatically extract the relevant piece of information? In particular, the following questions and challenges arise:

1. How can a computer work with the qualitative raw text data?
2. How can the text mining algorithm separate the important words (e.g. 'rates', 'unchanged') from the uninformative ones, the so-called 'noise' (e.g. 'Based on our ... we decided')?

[2]For an introduction on monetary policy see https://www.ecb.europa.eu/mopo/intro/html/index.en.html.

3. How does the machine know that some terms form an expression or entity (e.g. 'key ECB interest rates')?

4. How can it understand that some expressions relate to each other and together constitute the valuable information?

In our example the important information about the monetary policy decision is contained in the following two expressions and their relationship to each other: 'key + ECB + interest + rates' and 'keep + unchanged'.

> *"~~Based on our regular economic and monetary analyses, we decided to~~ **keep** ~~the~~ **key ECB interest rates unchanged.**"*

In the following we will present a first naive approach of how a machine could process this raw text data automatically and generate valuable insights.

In general, for a computer a textual document is internally represented as a string data type (sequence of characters) and thus cannot serve directly as an input for quantitative methods. Therefore, one has to transform it first to a quantitative representation. A simple approach is to create a count vector where every entry of the vector corresponds to one term and comprises the number of occurrences of the respective word in the overall text document.

Figure 3.1 illustrates the idea for our example. The first row corresponds to the first word in our document ('Based') which occurs exactly one time in this sentence. Note, the order does not matter in the construction of the count vector (often it is constructed based on the alphabetical order). In our simple example this vector representation already exhibits a dimensionality of 18×1. In a more realistic setting the number of unique terms can easily exceed 10,000. So a lot of techniques in textual analysis relate to the task of reducing this high dimensionality of the vector representation.

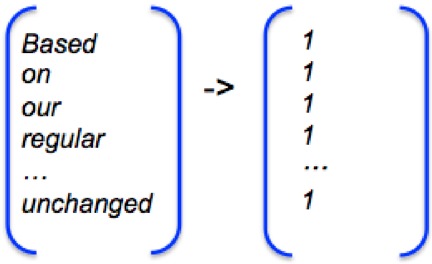


Figure 3.1: Count Vector Representation of the Interest Rate Decision Statement

Now imagine that we also observe the intra-day asset price response, for example in the equity market, to this news and that we collect a time series of the statements with their corresponding returns for all monetary policy announcements of the last 10 years. To analyze which words are responsible for the variation in the returns we can now, as a simple approach, employ a standard regression model and check the estimates for statistical significance:

$$Y_t = \beta_0 + \beta_1 \times X_t^{(Term\ 1)} + ... + \beta_V \times X_t^{(Term\ V)} + \epsilon_t, \tag{3.1}$$

with Y_t being the intra-day return for the interest rate decision statement published at time t, $X_t^{(Term\ v)}$ being the number of occurrences of term v in the respective monetary policy decision statement and V the number of unique terms across all announcements. The error term of this regression is represented by ϵ_t, with $\epsilon_t \sim i.i.d.(0, \sigma^2)$. For example, in our use case one can imagine a positive and statistically significant relationship between the word 'decrease' or 'lower' and the corresponding announcement return since a lower discount rate leads to higher stock prices.[3]

A problem within that naive approach is that the resulting X-matrix is sparse and high dimensional and in comparison we usually only have few Y observations. This is a typical example of the multi co-linearity problem which leads to non-trustworthy estimates in the previous specified regression model (see Equation 3.1). Thus, the key

[3]In the following sub-section we show an implementation of this example using an artificially created dataset.

challenge is to reduce the dimensionality of this machine learning problem. This can be done for example via an efficient pre-processing (e.g. filtering out stopwords, see Chapter 3.3), a sophisticated choice of the representation (e.g. tf-idf, see Chapter 3.4), an aggregation of the information to topic (see Chapter 3.5) and tone level (see Chapter 3.6) or via the statistical method (e.g. using a penalized regression, see Chapter 3.7).

Another challenge is how to incorporate the context of a given word into the machine learning model. In our example the important information is a combination of 'interest rates' and 'unchanged'. Within our specified regression model every term is an independent input variable. One possibility would be to also include interaction terms for the different words in the regression. Here the question arises which interaction terms to include? Adding all possible combinations would yield the curse of dimensionality problem again as discussed before. One popular technique to include the word order and thus the context in the analysis is the n-gram representation (see Chapter 3.4).

Simulation Exercise

In this sub-section we are going to look at an implementation of the simple regression problem from the previous chapter using simulated data. In particular, we are going to discuss the following steps:

1. First, we are going to create an artificial dataset which comprises short textual statements about the interest rate decision (increase, decrease, unchanged) and a quantitative number for the corresponding stock market response. In total, we are simulating 60 statement-return pairs.

2. Next, we are going to transform the raw textual data into a quantitative representation using count vectors.

3. In a last step, we are going to deploy a standard regression model to learn about the relationships between the independent words on the one side and the respective intra-day return series on the other side.

Figure 3.2 shows our three different interest rate statements which announce an increase or decrease of the key ECB interest rates or no change at all. If there is a decrease (increase) in the interest rate we assume a positive (negative) stock market reaction of 20 bp. If there is no change at all then we associate a zero response for the equity returns. The only variation in the three statements that can explain the variance in the returns, lies in the keywords 'increase', 'lower', 'keep' and 'unchanged'. In total, we create a balanced dataset of 60 observations which comprises of the three different statement-return pairs to an equal share.

```
SIMULATED DATA
--------------

-> Simulation of 60 interest rate decision statement-return pairs.

20x  Based on our regular economic and monetary analyses, we decided to increase the key ECB interest rates.
 Return: -0.2 [%]

20x  Based on our regular economic and monetary analyses, we decided to keep the key ECB interest rates unchanged.
 Return:  0.0 [%]

20x  Based on our regular economic and monetary analyses, we decided to lower the key ECB interest rates.
 Return: +0.2 [%]
```

Figure 3.2: Simulated Interest Rate Decision Dataset

In a next step, we are going to transform the textual data into a quantitative representation constructing count vectors as discussed before. Then we put them together into one matrix with the count vectors as the rows. Figure 3.3 shows the first 5 rows of this matrix containing the observations for the interest rate announcements and their corresponding stock returns. The columns represent the unique terms across all statements and the number indicates the occurrence of the respective word for each statement (beside the return). In total, we observe 20 different terms, 18 as discussed in the previous chapter and two more words characterizing the statements for the interest rate 'increase' or decrease ('lower'). Since the count for the term 'unchanged' is 1 and the return is 0.0 the first rows only show statements where the interest rates were not changed.

	analyses	and_	based	decided	ecb	economic	increase	interest	keep	key	...	monetary	on	our	rates	regular	the	to	unchanged	we	Return
0	1	1	1	1	1	1	0	1	1	1	...	1	1	1	1	1	1	1	1	1	0.0
1	1	1	1	1	1	1	0	1	1	1	...	1	1	1	1	1	1	1	1	1	0.0
2	1	1	1	1	1	1	0	1	1	1	...	1	1	1	1	1	1	1	1	1	0.0
3	1	1	1	1	1	1	0	1	1	1	...	1	1	1	1	1	1	1	1	1	0.0
4	1	1	1	1	1	1	0	1	1	1	...	1	1	1	1	1	1	1	1	1	0.0

Figure 3.3: Quantitative Representation of the Simulated Dataset

After this preparation we can now use a standard regression model to analyze which information or words are driving the variation in the returns. Figure 3.4 shows the estimated coefficients for the regression problem as defined in Equation 3.1.

	coef	std err	t
analyses	-5.421e-19	3.01e-19	-1.802
and_	3.903e-18	3.01e-19	12.974
based	4.77e-18	3.01e-19	15.857
decided	-3.003e-17	3.01e-19	-99.829
ecb	2.168e-19	3.01e-19	0.721
economic	2.168e-19	3.01e-19	0.721
increase	-0.2000	6.17e-18	-3.24e+16
interest	2.168e-19	3.01e-19	0.721
keep	5.378e-17	4.01e-18	13.402
key	2.168e-19	3.01e-19	0.721
lower	0.2000	6.17e-18	3.24e+16
monetary	2.168e-19	3.01e-19	0.721
on	2.168e-19	3.01e-19	0.721
our	2.168e-19	3.01e-19	0.721
rates	2.168e-19	3.01e-19	0.721
regular	2.168e-19	3.01e-19	0.721
the	2.168e-19	3.01e-19	0.721
to	2.168e-19	3.01e-19	0.721
unchanged	-3.296e-17	4.01e-18	-8.214
we	2.168e-19	3.01e-19	0.721

Figure 3.4: Estimated Coefficients of the Regression Model

One can clearly see the high t-statistics ($> |8|$) for our four keywords that indicate how the interest rate was adjusted ('increase', 'lower', 'keep' and 'unchanged'). Also the magnitudes of the estimates correspond to the simulated returns from our dataset (-0.2, +0.2 and 0.0) which suggests a reasonable fit for this naive approach. But one has to keep in mind that within our simulation study we looked at an easy toy example with only 20 dimensions. In a realistic setting we can easily end up with over 10,000 unique terms and thus independent explanatory variables. In addition, Figure 3.4 shows another problem: Some of the other words that occur in all of the three statements and thus do not contain explanatory power for the variation in the stock returns also show up significantly (e.g. 'and', 'based', 'decided').[4] The problem here are the relatively few y observations compared to the high dimensionality of the X-matrix which makes it very difficult to clearly identify the different effects. To get rid of this multicollinearity problem one needs to significantly reduce the number of independent variables in the regression model. This is one of the key challenges in textual analysis and popular methods to achieve that goal are discussed in the next sections.

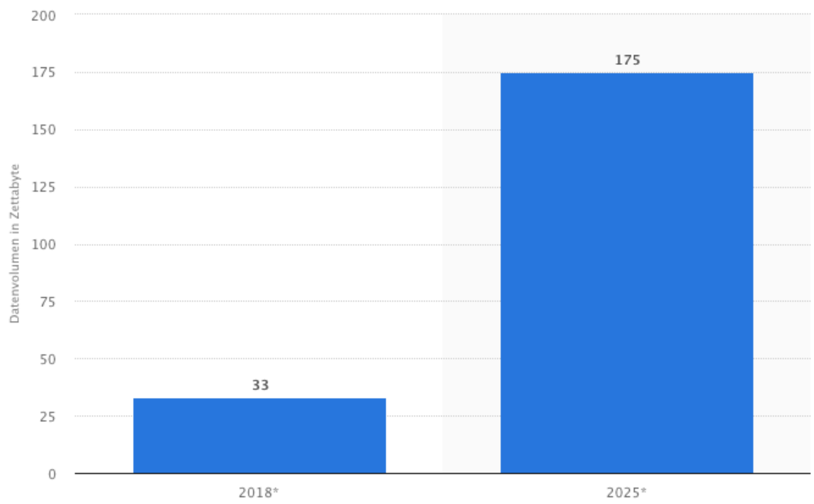
3.1.2 Motivation

Through the digital revolution, the digitalization, there has been an exponential growth in the amount of data created. Figure 3.5 shows that the volume of the digital data generated worldwide in 2018 was around 33 zettabytes (33 trillion gigabyte) and the projection for 2025 is over five times as high: 175 zettabytes. This data is generated not only by computers but all kinds of devices like smartphones, or the internet of things and industry 4.0. In our daily lifes we also generate lots of data e.g. through Google search queries, the usage of social media or messaging services or online shopping. Nowadays, a lot of companies, especially the big tech firms like Google, Facebook or Amazon, exploit this data and monetarize its valueable information.

Clearly one cannot manually screen the vast amount of data and then extract the valuable information by hand. Instead one needs a technical system that processes the different data automatically (see Figure 3.6). Given that a lot of this raw data is text-based and unstructured[5] one needs to deploy text mining techniques to extract the information. In addition, one often needs to first collect the textual documents in an automatic way using Web crawling algorithms.

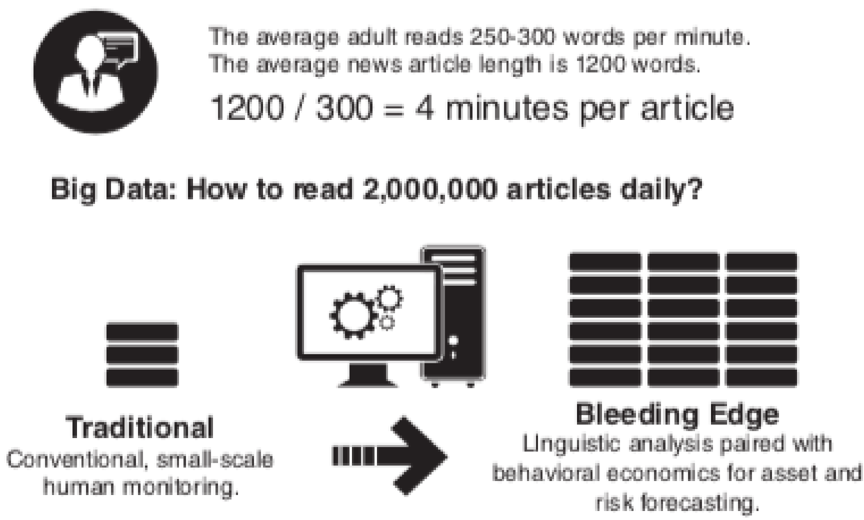
[4]The other terms exhibit very small estimates that are not statistically different from zero (to the 5% level).

[5]Unstructured data usually does not have a pre-defined data model compared to a series of stock prices which typically consists of the information 'date' and 'price' for each time point.



Source: Statista 2019

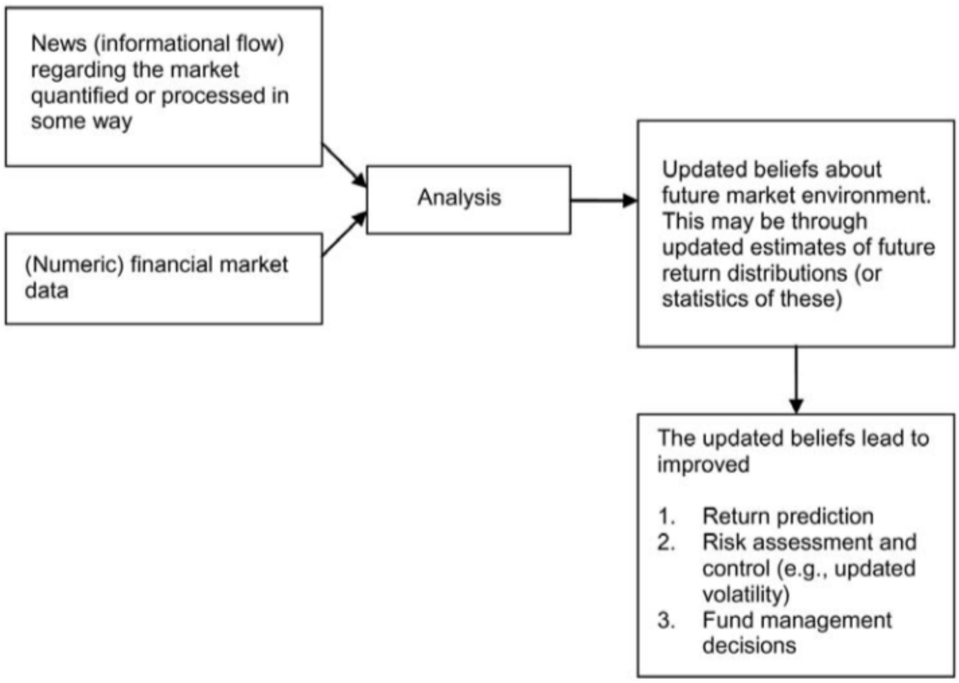
Figure 3.5: Projection of Annually Generated Volume of Digital Data Worldwide in 2018 and 2025 (in Zettabyte)



Source: Peterson 2016, p. 57

Figure 3.6: Large-Scale Text Analytics

These systems and methods are also popular in the field of news analytics in finance. 'It is widely recognized news plays a key role in financial markets. [...] There is a strong yet complex relationship between market sentiment and news. The arrival of news continually updates an investor's understanding and knowledge of the market and influences investor sentiment' (see Mitra and Mitra 2011a, p.1) and hence, affects their risk and asset management decisions. Figure 3.7 illustrates this idea.



Source: Mitra and Mitra 2011a, p. 3

Figure 3.7: A Simple Representation of News Analytics in Financial Decision Making

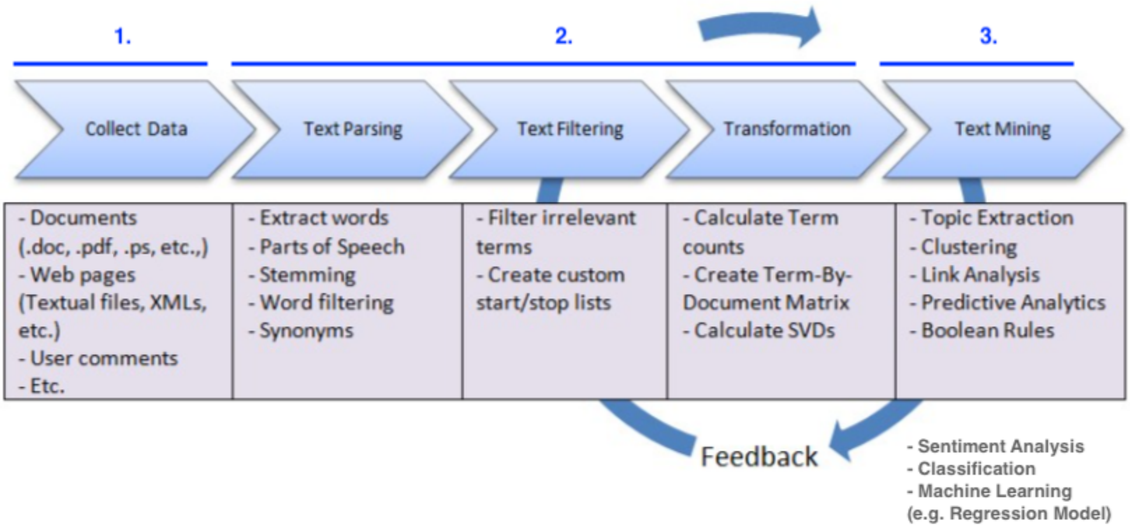
Using this alternative data risk and asset management companies try to make better investment decisions (generate an 'alpha') or to improve their risk monitoring. Important news and data sources are e.g. SEC filings, social media or central bank communication.

3.1.3 Definitions and Text Mining Process

In this chapter we are going to provide definitions for 'text mining' in general, technical terms used in textual analysis and we are also going to present the classical text mining process.

Compared to entries in a database, which follow a pre-defined pattern or schema, textual data is not structured. There is no emphasis on the key words or fundamental concepts in the text, their relationships to each other and how certain parts of the text relate to the key terms. A human reader can intuitively grasp these characteristics but a computer needs a set of tools to extract that structure. Heyer, Quasthoff and Wittig define text mining as a group of methodological approaches to structure texts and thereby extract new and relevant information (see Heyer et al. 2012, p. 3f).

Figure 3.8 shows the classical text mining process. It consists of the three logical steps: 'Data Collection', 'Pre-Processing' and 'Text Mining'.



Source: Chakraborty et al. 2013, p. 7

Figure 3.8: Text Mining Process Flow

First, the basis for every text mining process is a collection of documents. This collection can comprise of hundreds to millions (and more) of documents. In general, one can use a static collection of documents like PubMed[6] or one can apply Web crawling techniques to collect articles from the Internet (e.g. Wikipedia articles or central bank

[6]https://www.ncbi.nlm.nih.gov/pubmed.

statements).

Second, after collecting the potentially large amount of documents one usually has to apply some filtering techniques to obtain the most relevant ones. Also one has to consider and check the quality of the articles. In a next step the documents need to be transformed into a quantitative representation (e.g. tf-idf, see Chapter 3.4).

Third, based on the created representations one can extract the topic of the documents or perform a sentiment analysis (e.g. calculate the tone). These measures can then be used in a machine learning model (e.g. regression) to further analyze the information content of the respective documents.

3.1.4 Applications

We now turn to applications of textual analysis in finance. In the following two subchapters we are going to present a selection of important papers that illustrate the typical research study design and methodologies used, rather than providing a comprehensive literature survey. The interested reader we refer to more extensive survey papers.

Equity Market

A popular application of news text analytics is the prediction of stock prices. Already in the early 20th century Cowles (1933) examined whether financial publications could actually forecast the stock market. For the analysis he first selected, during the time period from January 1928 to June 1932, 24 publications from professional financial services, financial weeklies, bank letters, or investment house letters, manually read them and then subjectively classified them into ‘bullish’, ‘bearish’ or ‘in doubt’. These forecasts have then been tested with actual fluctuations of the stock market as reflected by the Standard Statistics Company index of 90 representative stocks (a precursor of the S&P 500). On a week-to-week basis the average forecasting performance fell approximately 4% per year below a record created by pure change.

A similar idea lies behind the study of Antweiler and Frank (2004) who, roughly 70 years later, use modern computational linguistic and machine learning techniques to automatically and objectively classify more than 1.5 million messages posted on stock message boards. Based on the classes ‘bullish’, ‘bearish’, or ‘neither’ they construct a bullishness index and use it to predict the returns of the 45 companies in the Dow Jones Industrial Average and the Dow Jones Internet Index. They find that stock messages help to predict market volatility but their effect on stock returns is economically small. A doubling of the number of messages posted leads to a -0.2% decrease in the stock price.

In general, there are two ways to determine which words are most useful to discriminate between a positive and negative sentiment of a document. Either one can use a machine learning algorithm (like in Antweiler and Frank 2004) which, based on a training set, automatically learns the most important words or one can specify a wordlist up-front (like a dictionary) and then for example count the number of negative words in the respective document.
An example for the latter method is the seminal work of Tetlock (2007) who used the Harvard Psychosociological Dictionary to extract the sentiment from the ‘Abreast of the Market’ column of the Wall Street Journal and study its relationship to stock returns. He finds that high media pessimism predicts negative returns on the next day. A one standard deviation change in pessimism leads to a -8.1 bp change in Dow Jones returns which (in absolute terms) is higher than the unconditional mean of these returns (5.4 bp).
A few years later Loughran and McDonald (2011) emphasize the importance of the choice of the dictionary. In their study they show that a wordlist developed for other disciplines misclassifies common words in financial documents (e.g. almost three-fourths for the Harvard negative wordlist). Based on SEC 10-K company filings they propose their own lexicon (LM dictionary) which is nowadays the most commonly used dictionary

in financial textual analysis studies.

Jegadeesh and Wu (2013) point out that when using the dictionary approach to calculate a sentiment score the implicit assumption is that all words in the list have the same weights and thus are of same importance. In their study they present a new approach to determine the weights using the market's reactions to 10-K filings within a simple regression model (the words from the dictionary serve as the independent variables). They find that their quantitative scores have very low correlation with the ones assigned by the LM dictionary. In addition, their methodology objectively determines whether words carry a negative or positive sentiment based on the sign of the estimated coefficient.
This popular text regression method is also applied in Manela and Moreira (2017) who construct a text-based measure of uncertainty. In particular, they run a support vector regression (SVR)[7] using front-page articles of the Wall Street Journal (as the explanatory variables) on the VIX index. Based on the long coverage of this newspaper they create a time series of the news implied volatility index (NVIX) starting 1890. Another advantage of their methodology is that one can relate uncertainty peaks to the specific news events. They find that news related to wars and government policy explains most of the variation in risk premia over time.

For a more thorough literature review we refer the interested reader to the following survey papers: Loughran and McDonald (2016), Tetlock (2014) and Kearney and Liu (2014).

Central Bank Communication

The Bank of England's discussion paper (Bank of England, 2015) emphasizes the important role that central bank communication played 'in the crisis response,

[7]Compared to the standard linear regression model a SVR selects only a relatively small number of observations called support vectors and ignores the rest. See Manela and Moreira (2017) for a more detailed introduction of this concept and further references.

from forward guidance to disclosure of stress-test results.' But 'how frequently, in what form and about what, should central banks communicate?' And how does the communication affect financial markets? Is there a 'risk of communication' which adds volatility to financial markets or encourages the public to overrely on public signals (see Bank of England, 2015)? To find an answer to these questions a new strand of literature emerged which uses textual analysis tools to extract the information content from different central bank statements and to quantify the respective market impact.

Focusing on the U.S. central bank Hansen and McMahon (2016) examine the information released by the Federal Reserve. Applying computational linguistic tools to FOMC statements they first measure the news on the state of economic conditions, as well as the forward guidance. Within a factor augmented VAR framework they find that in the last 18 years shocks to forward guidance are more important than the economic information. In particular, a more expansionary forward guidance about future rates tends to decrease longer maturity bonds significantly whereas there is almost no significant reaction of yields when the FOMC statement is talking more about an economic contraction.

Jegadeesh and Wu (2017) analyze the FOMC minutes which are released about three weeks after the monetary policy decision and contain additional details and comments about the rationale of the decision. From the raw statements they first extract eight distinct topics and then calculate for each topic the corresponding tone and uncertainty.[8] In a last step, they regress these measures on the market reaction around the release of the documents and find that the FOMC minutes contain significant incremental informational value, despite that they are released a few weeks after each FOMC meeting. Specifically, the market finds the Fed's discussion of its policy stance, inflation and employment to be the most informative, compared to topics like trade or consumption which seem to be not informative for market participants.

[8] In particular, to extract the topics they deploy the Latent Dirichlet Allocation (LDA) which was first introduced in Blei et al. (2003). An intuitive video explaining the basic idea behind LDA can be found here https://www.youtube.com/watch?v=3mHy4OSyRf0&t=1058s.

For the European Central Bank Schmeling and Wagner (2019) show that the tone of the introductory statement of the ECB's press conference moves asset prices. The authors use standard textual analysis tools to first pre-process the raw text data and then calculate the tone as a simple word count of negative terms obtained from the LM dictionary. They find that a more positive tone leads to higher stock prices and interest rates whereas credit spreads decrease. Thus, central bank communication seems to provide a channel through which monetary policy can affect risk premia in financial markets.

Analyzing the Bank of England's inflation report Hansen et al. (2018) show that information about future expected economic conditions can affect short- and long-run market rates. While studying both qualitative and quantitative information in the inflation report they find that the narrative about risks and uncertainty affects long-run yields, in particular their term premium component.

Figure 3.9 illustrates the common research design of these studies:

- First, these studies focus on official communication releases published by the major central banks (ECB, Fed or Bank of England). To automatically collect the statements they deploy standard Web crawling techniques.

- In a next step, they use standard textual analysis tools to pre-process the raw text data and transform it into a quantitative representation.

- Then they calculate the tone either of the aggregate statement, the individual topics or on a sentence level, often using a dictionary approach. To extract the topics a clustering algorithm is deployed.

- In a final step, the extracted topic and tone measures or the individual terms are used as the explanatory variables in an often standard regression model. For the independent variable the authors use stock returns or bond yields, either end of day values or for a narrow time window around the release of the statement.

These analyses allow to answer the question which information content is driving financial markets.

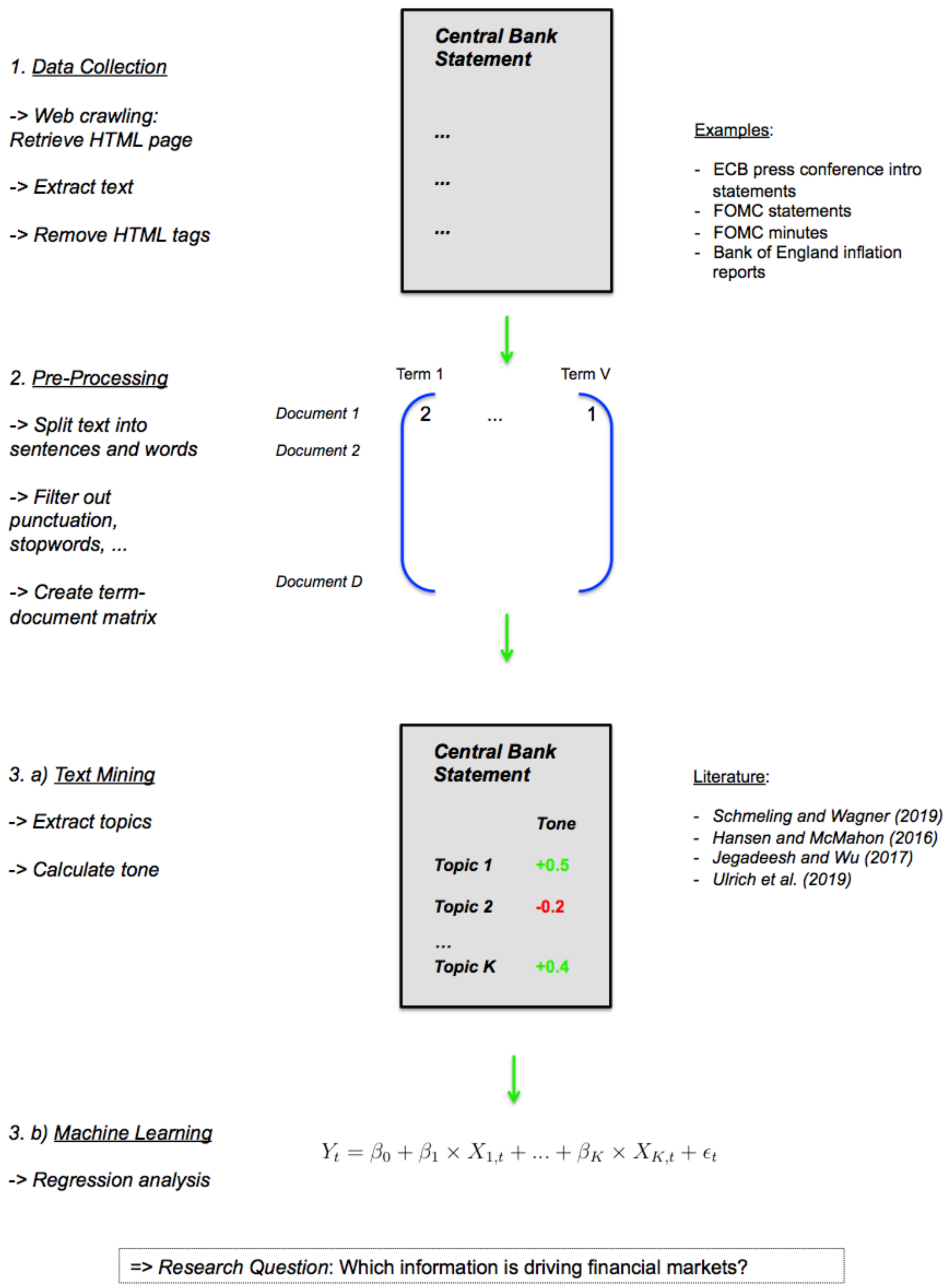


Figure 3.9: Typical Research Study Design

Overall, this common research design follows closely the three major steps from the classical text mining process as defined earlier in this chapter (see Figure 3.8).

The following chapters will explain the fundamental techniques used in text mining (in finance). Since there are a vast amount of textual analysis topics we focus on a selection that is necessary to understand this thesis. For further concepts or a more detailed explanation we refer the interested reader to fundamental books or papers.

3.2 Data Collection: Web Crawling

Some texts (e.g. newspaper articles) can be retrieved via news databases like Nexis[9] or Factiva[10]. But a lot of textual data, like social media feeds or press releases from central banks, has to be collected from the World Wide Web ('the Web'). To automatize this process one needs to understand the basic architecture underlying the Web and the Internet. Therefore, in this chapter we will briefly introduce the most important Web technologies and present the technique of Web crawling.

3.2.1 Web Technologies

The Web was invented in 1989 by Tim Berners-Lee when he was working at CERN (Halsall, 2005, p. 568) and is the most popular service of the Internet (short for interconnected network). The Internet is the underlying network of networks whereas the Web is the actual application which uses the Internet to exchange messages and documents between different computers or devices. Other applications are for example e-mail, Voice-over-IP or file sharing.

In our daily lifes the Web is often used for information browsing. For example, if you like to inform yourself about the latest monetary policy decision you can go to the ECB's web page and look up the information you need. Figure 3.10 illustrates this use case and

[9]See https://www.nexis.com.
[10]See https://www.dowjones.com/products/factiva/.

highlights the main technologies providing the service.

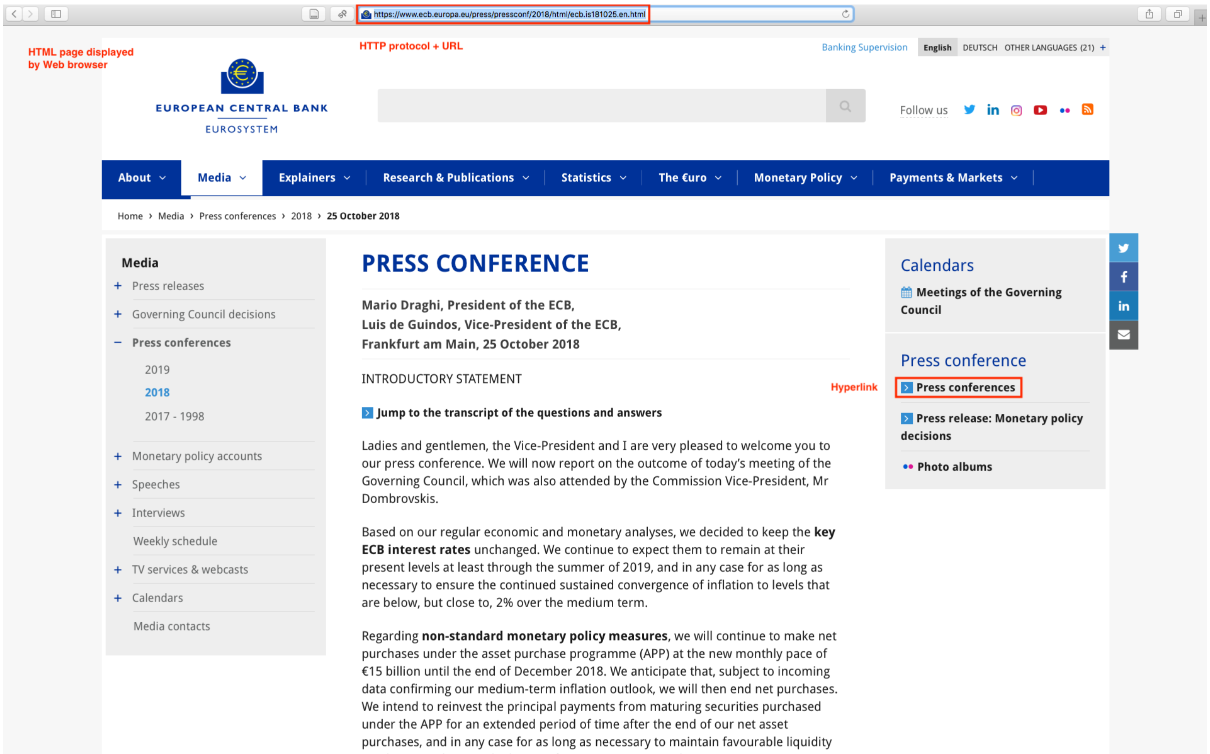


Figure 3.10: Web Page of the European Central Bank

Usually, one starts with opening the **web browser** that is installed on your computer and then types in a **URL**, which is the address for the web page of interest. If the request is submitted (by hitting 'enter') one can see that the full URL starts with 'http(s)' which stands for the **HTTP protocol**[11] and 'www' which is the acronym for the World Wide Web. An address for a FTP file sharing service usually starts with 'ftp://' instead of 'http://www'. The requested resource is then provided by the respective server as a **HTML** page which gets displayed by the web browser. If one wants to navigate to the next page or resource one usually uses a **hyperlink** which contains the new address. When clicking on the respective link the new page is requested by using its URL and the HTTP protocol (see Halsall, 2005, p.569 ff).

In the following sub-chapters we are going to present the most important technologies in more detail. For a comprehensive discussion we refer the interested reader to funda-

[11] If one uses the secure version of the HTTP protocol then this is indicated by 'https'.

mental text books (like Halsall, 2005; Tanenbaum, 2003; Kurose and Ross, 2013) or the respective technology specification of the World Wide Web Consortium (W3C).

HTML

'To publish information for global distribution, one needs a universally understood language, a kind of publishing mother tongue that all computers may potentially understand. The publishing language used by the World Wide Web is HTML' (Faulkner et al., 2017, Chapter 2.2).

The key challenge for the Web lies in the heterogeneity of computers or devices (e.g. mobile phones) that want to use it. These clients have different operating systems, web browsers (also with different versions) and large or small screens with various resolutions but everybody wants to have a good user experience. Therefore, one needs a specific set of rules about how a page is structured (e.g. headline, paragraphs) and how it should get displayed on the different client systems. This is specified by the HyperText Markup Language (HTML), specifically by annotating the raw text with **HTML tags**.

This is illustrated in Figure 3.11 which shows the official ECB web page in the Safari web browser and its underlying HTML file. One can see that the document starts with the $< html >$ tag which defines that it is a HTML file. Then the meta information follow within the $< head >$ tag.[12] The $< body >$ of this file contains the content of the web page. Our press conference text is then located within the $< article >$ tag and the different paragraphs of the introductory statement are in a list of $< p >$ elements. Each of these $< p >$ tags contains the actual text that is shown on the web page (e.g. 'Ladies and gentlemen ...'). Note, one can exploit this explicit structure in the later analysis, e.g. for studying the information content of different paragraphs instead of the aggregate document.

[12]The $< /head >$ command closes the previous $< head >$ tag and thus indicates that the next element starts.

For a more detailed description we refer the interested reader either to (Halsall, 2005, Chapter 9.4) or the official HTML specification (Faulkner et al., 2017).

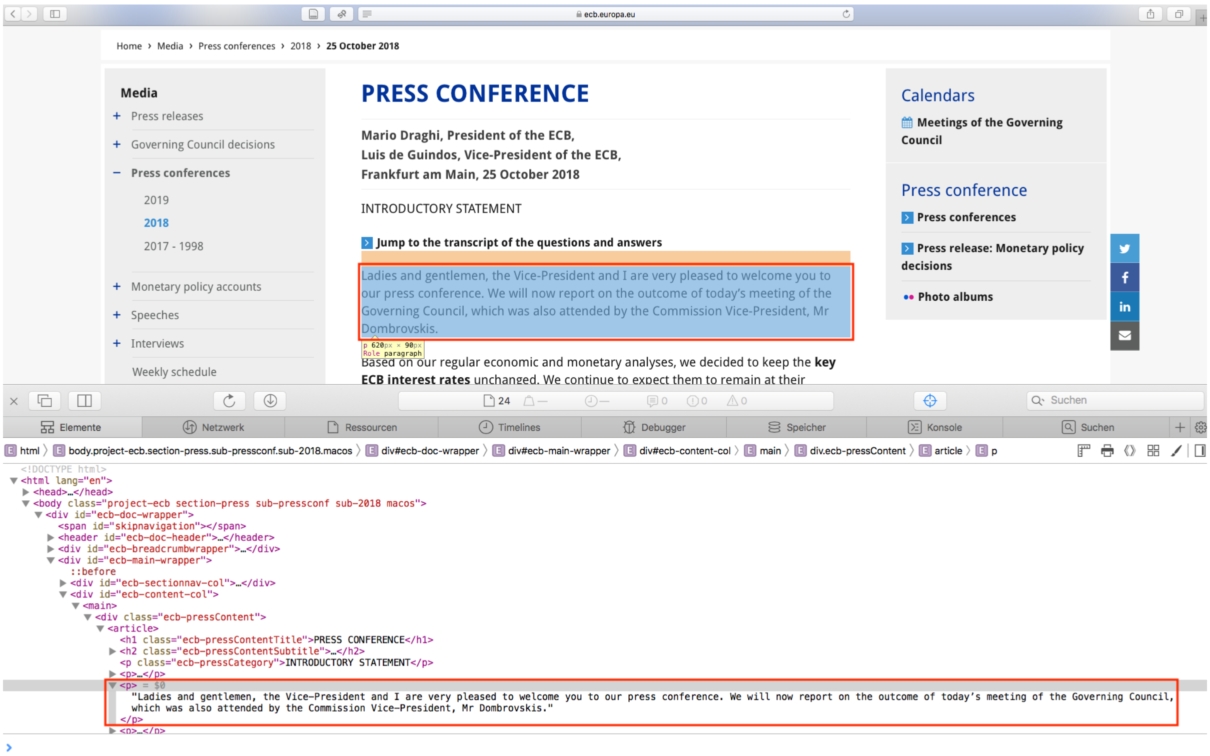


Figure 3.11: Source Code of the ECB Web Page

URLs and HTTP

The Web comprises of a vast collection of documents that are stored on millions of (server) computers distributed all over the world. Two questions or challenges arise. First, how can we identify all these files? And second, how can we access them on all the different machines?

The first challenge is addressed by the Uniform Resource Locator (URL). The URL is a uniform naming scheme for locating resources on the Web.[13] The standard format of an URL comprises of (see Halsall, 2005, p. 577):

- the application protocol which is used to retrieve the page,

[13]This concept is similar to a post address in the real world.

- the domain name of the server,
- the pathname or the file,
- the file name.

Figure 3.12 shows the URL for the ECB's web page as used before. The application protocol in this example is the secure version of 'http', the domain name is 'www.ecb.europe.eu', '/press/pressconf/2018/html/' represents the pathname and 'ecb.is.181025.en.html' is the file name. The ending of the file indicates that the requested resource is a html page. Note, one can also see that the date of the press conference is incorporated in the name of the file (2018/10/25). This suggests a systematic structure of the URLs and allows for an automatic retrieval of all press conference transcripts.

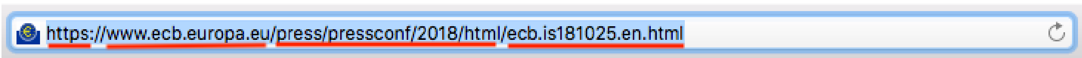


Figure 3.12: URL for the ECB Web Page

To establish a communication between all the heterogeneous computers in the Internet (e.g. Linux vs. Windows based servers) we need a standardized set of rules, a protocol. For the Web HTTP is the standard application protocol. It follows a simple request-response pattern: the browser side sends a request message and the server side returns a response message (Halsall, 2005, p. 581).[14]

Figure 3.13 shows a schematic diagram for a HTTP communication. It usually starts with the user typing in the URL of the web page of interest in his web browser: 'https://www.ecb.europe.eu/press/pressconf/2018/html/ecb.is.181025.en.html' (1). After hitting 'enter' the browser creates an HTTP request message and sends it to the server computer as identified by the first part of the URL ('www.ecb.europe.eu') (2).[15]

[14]For a more detailed and technical explanation see the W3C specification of the HTTP https://www.w3.org/Protocols/.

[15]Note, for the sake of brevity we skip the host name resolution by the Domain Name System (DNS), that maps the logical names on the corresponding IP address (like a phone book in the real world).

Due to the standardized application protocol (HTTP) computers with a different technology stack (e.g. Windows or Linux operating system) can communicate with each other. The web server program that runs on the server machine (and listens to port 80) receives that message and routes it to the corresponding implementation or class (using the relative path information) and method (via the HTTP method).[16] Within the GET method the html file can then be retrieved, e.g. from a database or file server (3). It is then attached to the HTTP response message and sent back to the client (4). The web browser of the user receives that message, extracts the html file and displays it accordingly to the structure as defined by the html tags (5). Note, a HTTP GET request does not necessarily need to be issued by a web browser but can also be executed by any other program, e.g. a web crawler (see also Halsall, 2005, p. 577 f).

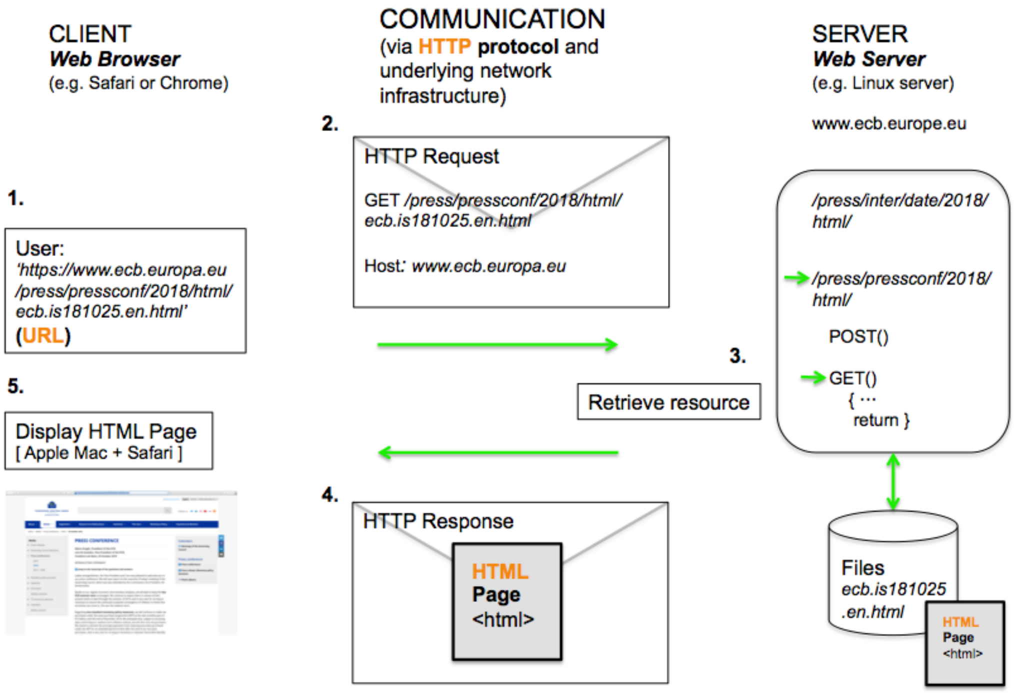


Figure 3.13: Example for a HTTP Communication

A more detailed discussion can be found in Chapter 9.3 of Halsall (2005) and in the specifications of the HyperText Transfer Protocol.[17]

[16]Note, the standard HTTP method that is used when requesting a Web page is 'GET'.
[17]See https://www.w3.org/Protocols/.

3.2.2 Web Crawler

Knowing the basic web architecture and technologies one can now write a program that automatically issues the HTTP GET requests and then extracts and saves the html files from the response messages. Such a program is called a web crawler.

Figure 3.14 shows the structure of a web crawler that automatically collects all the press conference transcripts from the ECB web page. The program performs for every press conference date the following steps:

1. Construct the specific press conference URL by combining the base URL and the dynamic part, here the press conference date (e.g. ‘181025’): ‘https://www.ecb.europa.eu/press/pressconf/2018/html/ecb.is’ + DATE + ‘.en.html’.

2. Send a HTTP GET request to the specified URL (like a normal web browser).

3. Extract the HTML page from the HTTP response message.

4. Instead of displaying the HTML page to a user we need to parse and pre-process it. In particular, one needs to extract the main text (that is located between the $< article >$ tag)

5. and remove the HTML tags.

6. In a last step we save the processed text files for further analyses.

3.3 Pre-Processing

When working with text files one usually starts with preparing the raw textual data. In this chapter we will present the standard pre-processing steps that are used in textual analysis.

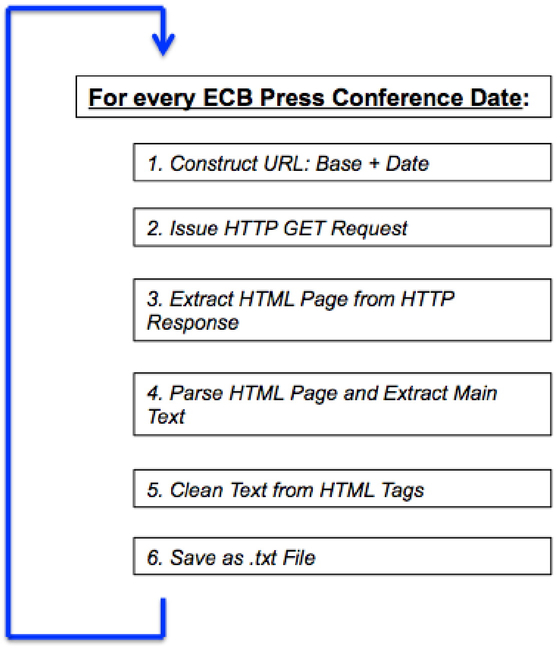


Figure 3.14: Web Crawler for ECB Press Conference Transcripts

3.3.1 Tokenization

In a first step, we need to split the raw text document into its constituents, initially into the sentences and then the words. Manning and Schütze (2003, Chapter 4.2.2) define tokenization as the process 'to divide the input text into units called *tokens* where each is either a *word* or something else like a number or a punctuation mark.'

A simple rule-based approach is to split the document on white spaces and punctuation. But this will also split hyphenated phrases like 'vice-president' or numbers like '1.8%'. Therefore, in practice, one should use a specialized library for this task.

For example, if we tokenize our previous interest rate decision statement

> *"Based on our regular economic and monetary analyses, we decided to keep the key ECB interest rates unchanged."*

we would obtain the following list of individual terms:

['Based', 'on', 'our', 'regular', 'economic', 'and', 'monetary', 'analyses', ',', 'we', 'decided', 'to', 'keep', 'the', 'key', 'ECB', 'interest', 'rates', 'unchanged', '.'].

Multi - Word Expressions

As one can see in the above example the tokenization process splits the raw text into its individual tokens regardless whether certain terms belong together. E.g. the words 'key', 'ECB', 'interest' and 'rates' refer to the ECB's official interest rates for the euro area.

To mitigate this problem one can use a multi-word expression (MWE) tokenizer which groups these tokens together. In our example the previous four interest rate terms would become the following unified token 'key ECB interest rates'. In general, the information on which expressions should be treated as one must be specified by the user.

3.3.2 Punctuation, Numbers or Non-Alphabetic Characters

To reduce the number of items in the list and thus the dimensionality of the later representation one should filter out tokens that are not informative ('noise'). A first method is to remove non latin alphabet tokens, like punctuation (',' or '.'), numbers or non-alphabetic characters (e.g. '@' or ':)').

Note, in some applications it might be important to keep the numbers (e.g. 'GDP increased by 0.5%') since they contain valuable information. Also when working with social media content, e.g. tweets, non-alphabetic characters like smileys ':)' can represent an informative feature.

Filtering out the non latin alphabet tokens from our interest rate statement yields the following list of tokens (reduced by two items):

[‘Based’, ‘on’, ‘our’, ‘regular’, ‘economic’, ‘and’, ‘monetary’, ‘analyses’, ‘we’, ‘decided’, ‘to’, ‘keep’, ‘the’, ‘key’, ‘ECB’, ‘interest’, ‘rates’, ‘unchanged’].

3.3.3 Stopwords

A second popular method to reduce the dimensionality is filtering out stopwords. According to Das (2014, Chapter 3.3) ‘stopwords are non-contextual words, i.e., not germane to interpretation of the text that are removed from the data before conducting textual analysis.’ Typical examples are ‘a’, ‘are’, ‘for’, ‘the’, ‘to’, etc. which appear often in the English language but can be seen as being redundant. Since there is no definitive list of stopwords one should always be careful when using a predefined one.[18] A good practice is to use the list of stopwords incorporated in standard NLP packages.[19]

Applying this technique to our interest rate decision sentence we obtain the following wordlist:

[‘Based’, ‘regular’, ‘economic’, ‘monetary’, ‘analyses’, ‘decided’, ‘keep’, ‘key’, ‘ECB’, ‘interest’, ‘rates’, ‘unchanged’].

After these basic pre-processing steps we already decreased the dimension of our list of tokens from 20 to 12.

3.3.4 Linguistic Roots

One can further decrease the dimensionality by collapsing similar word forms to a common root word. For many applications one does not loose information when using the linguistic root instead of the grammatical form of a word. For example, one could treat ‘economical’ and ‘economic’ as equivalent tokens.

[18] Also sometimes it is necessary to keep all the words of a text, e.g. when calculating a readability measure which is based on the length of a sentence.

[19] Also common is to drop rare words, for example those that appear less than a given threshold.

In general, there are two options to create the root forms: stemming and lemmatization. In Chapter 4.2.3 Manning and Schütze (2003) define stemming as 'a process that strips off affixes and leaves you with a stem.' A popular stemming algorithm (for the English language) that is used in the textual analysis literature is the Porter stemmer. This algorithm maps for example the two above mentioned terms to 'econom'.

Alternatively, the process of lemmatization first tags each token with its part of speech (POS) and then looks up each word-POS pair in a dictionary to find the root. For example, in the sentence 'we saw a recovery in the economy' 'saw' would be tagged as a verb and thus be converted to its root form 'see'. In the case of the phrase 'this is a big saw' it would be tagged as a noun and left unchanged.

3.4 Representations

Based on the list of terms obtained from the pre-processing one can create a quantitative representation for the collection of documents. In the following we will present the most commonly used methods.

3.4.1 Count Vector - Term-Document Matrix

A popular approach is the bag of words (BoW) model. Within this model 'all the structure and (linear) ordering of words within the context is ignored' (see Manning and Schütze, 2003, p. 237).

Given the tokenized, filtered and stemmed list of words from the pre-processing part the BoW model counts the occurrences of unique tokens in each document. These word counts $x_{d,v}$ are then collected in the *term-document matrix* $\mathbf{X}$ which is defined as follows:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \dots & x_{1,V} \\ \dots & \dots & \dots \\ x_{D,1} & \dots & x_{D,V} \end{bmatrix}.$$

Every row represents one document d and each columns an unique term v from the overall corpus. The dimensions of this matrix are $D \times V$.

The key characteristics of this matrix are its high dimensionality and its sparsity. The former is a result of the large number of different words used in a collection of text documents. The latter stems from the fact that not all the words appear in every document and thus exhibit a count of zero.

Example

We now extend the previous interest rate statement example by two more 'decisions': A rate increase and one lowering of the key ECB interest rates:

> *"Based on our regular economic and monetary analyses, we decided to keep the key ECB interest rates unchanged."*

> *"Based on our regular economic and monetary analyses, we decided to decrease the key ECB interest rates."*

> *"Based on our regular economic and monetary analyses, we decided to increase the key ECB interest rates."*

We can now create a quantitative representation of these three decisions. The respective term-document matrix is shown in Figure 3.15. Note, in our example the terms only occur at most once in every document. In practice, the counts are often larger than one.

	analyses	based	decided	decrease	ecb	economic	increase	interest	keep	key	monetary	rates	regular	unchanged
0	1	1	1	0	1	1	0	1	1	1	1	1	1	1
1	1	1	1	1	1	1	0	1	0	1	1	1	1	0
2	1	1	1	0	1	1	1	1	0	1	1	1	1	0

Figure 3.15: Example Term-Document Matrix

Compared to our introductory example from Chapter 3.1 we reduced the number of columns of this matrix from 20 to 14 due to the applied pre-processing steps.

3.4.2 TF-IDF Weighting

'If we count up how often each word (type) of a language occurs in a large corpus, and then list the words in order of their frequency of occurrence, we can explore the relationship between the frequency of a word f and its position in the list, known as its *rank* r' (see Manning and Schütze, 2003, Chapter 1.4.3). According to Zipf's law there is an inversely proportional relationship between the frequency of a particular term and its rank. This means that 'in human languages [...] there are a few very common words, a middling number of medium frequency words, and many low frequency words' (see Manning and Schütze, 2003, Chapter 1.4.3). This observation can distort raw word counts as discussed previously.

Also usually the most commonly used words carry very little meaningful information about the actual contents of the document (see also stopwords). In the above example we can see that certain words, like 'analyses' or 'decided', appear in every of our documents and thus have no discriminative power if one likes to differentiate between the different interest rate decisions. We know that the three statements only differ with respect to the terms 'keep', 'unchanged', 'decrease' and 'increase' which appear only once in the overall corpus. So the question arises how can we incorporate this characteristic into our quantitative representation?

Definition

A popular approach is to use the TF-IDF (*term frequency - inverse document frequency*) weighting scheme which is defined as follows (see Jurafsky and Martin, 2017, p. 805):

$$tf - idf_{d,v} = tf_{d,v} \times idf_v.$$

The tf-idf score consists of two components: First the term frequency $tf_{d,v}$ of word v in document d as used before which is then scaled by the second term, the inverse document frequency:

$$idf_v = log(\frac{D}{df_v}),$$

where df_v is the number of documents containing term v and D is the overall number of documents. This weighting scheme assigns high scores to words that occur many times in few documents.

Example

In our example the term 'analyses' appears once in all three interest rate statements and thus is not very characteristic for a single document. The idf score is therefore very low (around 1).[20] To correct for the document length the value is also normalized using the L2 norm. Since the L2 value for the first count vector is roughly 4, the first matrix entry yields 0.25 (1 x 1/4).

	analyses	based	decided	decrease	ecb	economic	increase	interest	keep	key	monetary	rates	regular	unchanged
0	0.252	0.252	0.252	0.000	0.252	0.252	0.000	0.252	0.427	0.252	0.252	0.252	0.252	0.427
1	0.279	0.279	0.279	0.472	0.279	0.279	0.000	0.279	0.000	0.279	0.279	0.279	0.279	0.000
2	0.279	0.279	0.279	0.000	0.279	0.279	0.472	0.279	0.000	0.279	0.279	0.279	0.279	0.000

Figure 3.16: Example TF-IDF Matrix

The four most important terms, according to the tf-idf score, are 'unchanged', 'keep', 'increase' and 'decrease' and match our keywords from the simulated dataset. Their

[20]Note, to prevent a division by zero the implementation of the used NLP package differs slightly from the standard textbook notation. For the implementation details, see Chapter 4.2.3.4. of https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction.

weighted score is factor 1.5 to 2 times higher than for the uninformative features.

Application

In their seminal study Loughran and McDonald (2011) also compare their newly constructed dictionary (based on the SEC 10-K filings) with the Harvard negative wordlist using the tf-idf weighting scheme. They find in their return regression analysis that when using tf-idf weights both wordlists are 'essentially identical in their impact'. This finding emphasizes the importance of considering a weighting scheme.

3.4.3 Ngram Models

The simple bag of words model considers only individual terms and destroys all information on the word order. This might be sufficient for certain applications but in some contexts we need the order information. Consider the following two sentences which have the same representation according to the BoW model, but share a different meaning:

"The interest rate was not increased."

"Not the interest rate was increased."

Here, the different semantics stems from the context of the term 'not'. If it is located next to the verb 'be (was)' then the sentence says that the action was not executed. Placed in the beginning it negates the noun 'interest rate'.

To include the context of a given word one can alternatively consider adjacent phrases, called n-grams. Typically, people use n = 2, 3 or 4 and these models are referred to as *bigram*, *trigram* or *four-gram* (see Manning and Schütze, 2003, p. 193).

In the example a bigram would be 'the_interest', 'interest_rate', 'rate_was' and so on.[21]

[21]Note, one can construct the n-grams also after the pre-processing steps when stopwords are already filtered out.

A trigram would include three adjacent terms, e.g. 'interest_rate_was', 'rate_was_not' and a four-gram would be 'interest_rate_was_not' or 'rate_was_not_increased'.

One can see that the last four-gram accurately captures the context of the term 'not' and 'increased', saying that the interest rate was not increased which is the relevant piece of information here. In general, working with the n-gram model is similar to using individual terms (v now indexes the unique bigrams/ trigrams/ four-grams rather than the unigrams). The problem is how to choose n so that the context is accurately captured? If n is too high then the curse of dimensionality problem arises (again).

The n-gram model is used for example in Manela and Moreira (2017) who construct a news-implied volatility index based on an uni- and bigram representation of Wall Street Journal abstracts.

3.5 Similarity and Clustering

In this section we are going to introduce several measures that can be used to quantify how similar two sentences or documents are to each other. Given the similarity scores one can then apply a clustering algorithm which groups related documents together.

3.5.1 Document Similarity

In general, one can interpret the individual rows of the term-document matrix as vectors in a V-dimensional space. Thus, to determine how alike two documents are we can now calculate the similarity between the corresponding vectors.

As a first approach one could calculate the Euclidean distance which is defined as (see Manning and Schütze, 2003, p. 301):

$$|\vec{x}_i - \vec{x}_j| = \sqrt{\sum_v (x_{i,v} - x_{j,v})^2},$$

with vector $\vec{x}_i$ ($\vec{x}_j$) representing the row of the term-document matrix that contains document i (j) and $x_{i,v}$ ($x_{j,v}$) being the respective row entry for term v.

The problem with this measure is that vectors which comprise of an identical set of words but have a different word count (e.g. $\vec{x}_1 = [3, 6]^\top$ and $\vec{x}_2 = [1, 2]^\top$) exhibit a large Euclidean distance whereas our intuition would tell as that they cover a similar topic. In Figure 3.17 this is illustrated by the vectors $\vec{x}_1$ and $\vec{x}_2$ which point in the same direction but have a different document length. A better approach would be to consider whether the vectors point in the same direction or not.

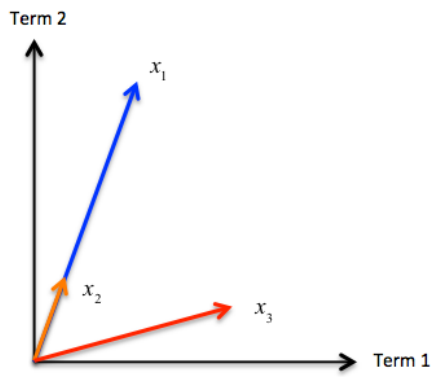


Figure 3.17: Vector Space Example

This idea is implemented by the popular cosine similarity. It measures 'the distance between two documents by the cosine of the angle between their vectors. When two documents are identical they will receive a cosine of 1; when they are orthogonal (share no common terms) they will receive a cosine of 0' (see Jurafsky and Martin, 2017, p. 803). The cosine similarity between two documents is calculated as:

$$cos\_sim(i, j) = \frac{\vec{x}_i \cdot \vec{x}_j}{||\vec{x}_i||\ ||\vec{x}_j||},$$

with $\vec{x}_i$ and $\vec{x}_j$ as defined above.

Example

Prior to the announcement of the interest rate decision the ECB President welcomes the audience to the respective press conference. We now add this ‘welcome’ sentence to our previous example. One can immediately see that this sentence is very different to the other three, since it uses a completely different vocabulary:

> *“Ladies and gentlemen, the Vice-President and I are very pleased to welcome you to our press conference.”*

In a next step, we first calculate the term-document matrix for the four sentences and then the cosine similarity for each document with respect to all the others. The results are shown in Figure 3.18:

```
matrix([[1.   , 0.87 , 0.87 , 0.   ],
        [0.87 , 1.   , 0.909, 0.   ],
        [0.87 , 0.909, 1.   , 0.   ],
        [0.   , 0.   , 0.   , 1.   ]])
```

Figure 3.18: Example Cosine Similarity

The first three columns correspond to the three interest rate decision statements as defined previously. The fourth column represents the similarity scores for the ‘welcome’ sentence. One can see that the first three documents share a high similarity score (around 0.9), whereas the newly added one is not similar at all (the score is zero).

For the following illustration we first transform the similarity matrix into a distance matrix

$$dist = 1 - cos\_sim$$

and then reduce the dimensionality of this matrix (to two) which allows us to visualize the distances between the different documents.[22]

[22] Specifically, we use the Multi-Dimensional Scaling (MDS) algorithm which is similar to the principal component analysis but uses a distance matrix as an input instead of the covariance matrix. For a detailed introduction of this algorithm we refer the interested reader to a basic statistics textbook.

The left panel of Figure 3.19 plots only the distances between the three interest rate decision statements. One can see that they are all a bit different but when adding the new ‘welcome’ sentence (as shown in the right panel) one can clearly see that they share a commonality compared to the introductory sentence (represented by point far to the right). Interestingly, the three interest rate decisions mostly differ across the second axis, the value for the first factor is roughly the same. Thus, one could conclude that factor one separates the documents across different topics and factor two differentiates within the topic, e.g. whether there was an increase or decrease in the interest rates.

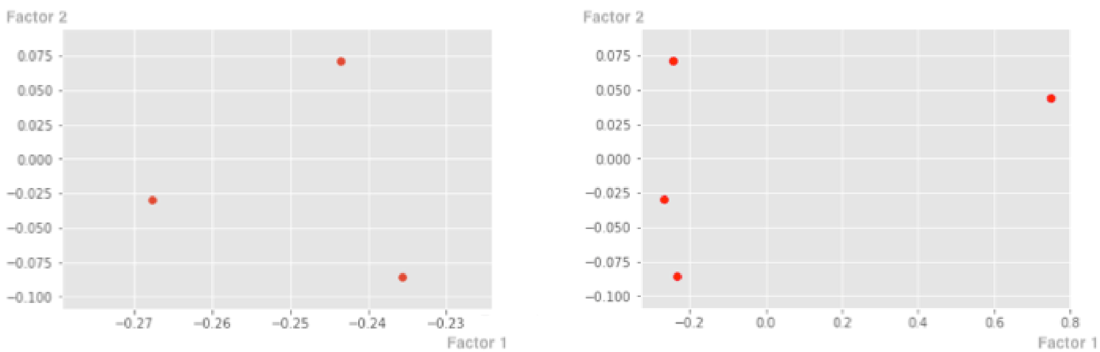


Figure 3.19: Example Visualization of Document Similarity

Application

Hoberg and Phillips (2010) apply the cosine similarity measure to product descriptions of 10-K filings to extract a continuous representation of the pairwise similarity of any two firms. They use it as an extension to the normal industry classification methodologies like the Standard Industry Classification (SIC) or the North American Industry Classification System (NAICS).

Also Merkley (2014) uses the cosine similarity and a vector space model to measure the amount of repetitive information within R&D narrative disclosures.

In addition to the cosine similarity measure also other scores are used. For example, Tetlock (2011) uses the Jaccard coefficient to determine the staleness of a news story, extracted from the textual similarity of a new information compared to the older ones. For a comprehensive list of similarity measures and a more detailed discussion of their properties we refer the interested reader to Manning and Schütze (2003, p. 299) or Jurafsky and Martin (2017, p. 697 ff).

Sequence Matching

Instead of calculating a similarity score using the respective document vectors one can compare directly two words or sentences with each other based on their string representation. One distance measure for strings is the minimum edit distance which 'is the minimum number of editing operations (insertion, deletion, substitution) needed to transform one string into another' (see Jurafsky and Martin, 2017, p. 107 f). If each operation has the cost of one then it is called the *Levenshtein* distance.

The Levenshtein distance for the example in Figure 3.20 is 2 since two characters have to be substituted to convert the one expression into the other.

Figure 3.20: Example Edit Distance

This minimum edit distance can now be used to compare words/ strings/ sequences with each other and match them based on the lowest distance value.

3.5.2 Clustering

Based on the obtained similarity scores we can now group similar observations (documents) together. Therefore, we deploy a clustering algorithm. These algorithms belong to the class of unsupervised learning methods, which means that they uncover previously unknown patterns in the data set without pre-existing labels. The hidden structure in textual documents can be thought of as topics, the underlying theme of the document.

K-Means

A popular method to extract topics from text is the K-means clustering algorithm. In general, it 'partitions a set of vectors $\{\vec{x}_1, \vec{x}_2, ... \vec{x}_n\}$ into a set of clusters $\{D_1, D_2, ... D_K\}$' (see Feldman and Sanger, 2007, Chapter V.3.1).

The algorithm proceeds as follows (Feldman and Sanger, 2007, Chapter V.3.1):

Initialization:

K seeds, either given or selected randomly, form the core of K clusters. Every other vector is assigned to the cluster of the closest seed.

Iteration:

The centroids $\vec{u}_k$ of the current clusters are computed (based on the documents d that are assigned to the respective cluster):

$$\vec{u}_k = \frac{1}{|D_k|} \sum_{d \in D_k} \vec{x}_d. \tag{3.2}$$

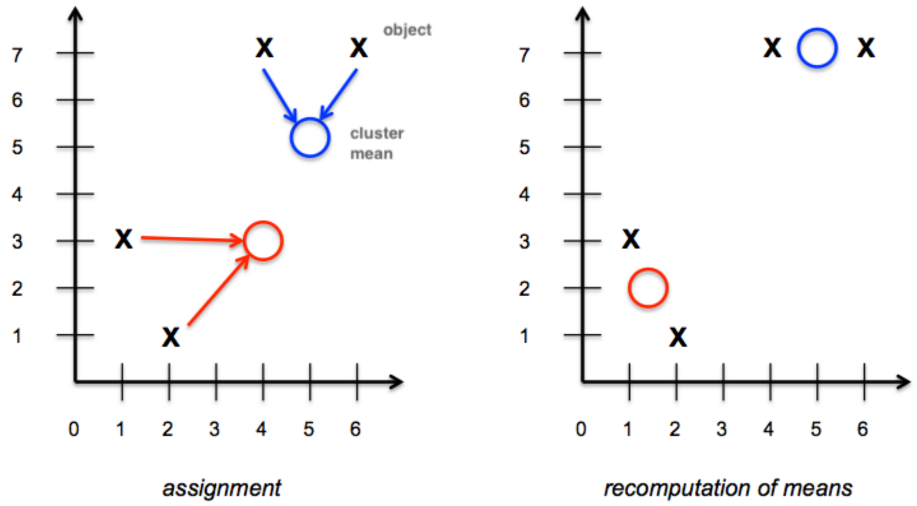
Each vector is reassigned to the cluster with the closest centroid.

Stopping Condition:

At convergence – when no more changes occur.

To determine the closest centroid one usually uses the Euclidean distance as the standard distance function (see Manning and Schütze, 2003, p. 515 f).

Figure 3.21 illustrates one iteration of the K-means algorithm. The first step assigns objects (documents) to the closest cluster mean (shown as circles). In the second step cluster means are recomputed as the center of mass of the set of objects that are members of the cluster (see Manning and Schütze, 2003, p. 517).



Source: Manning and Schütze 2003, p. 517

Figure 3.21: One Iteration of the K-Means Algorithm

Example

We now apply the K-means clustering procedure to our previous example. The algorithm takes two input values: First, the term-document matrix (e.g. the tf-idf matrix) and second the number of clusters (e.g. n = 2). The output is a list of topic labels [0, 0, 0, 1] which correspond to our four documents.

Figure 3.22 illustrates the result. The algorithm groups the three interest rate decision statements together, the introductory sentence forms a separate topic.

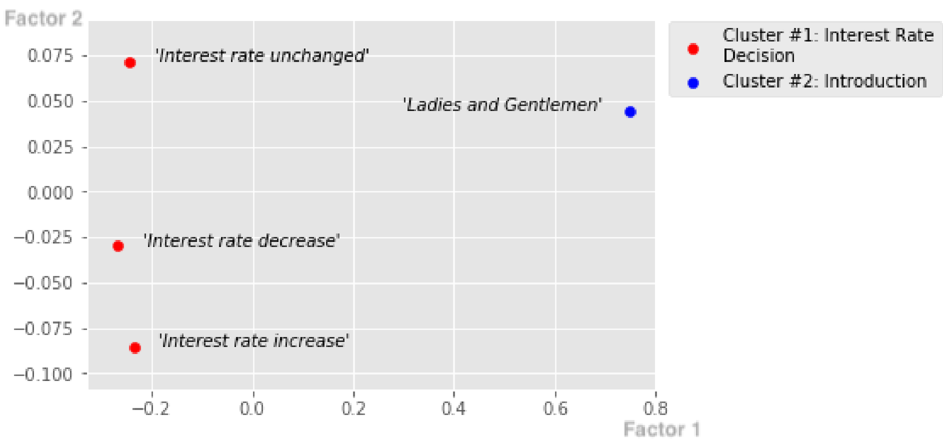


Figure 3.22: Example K-Means Clustering

In addition, Figure 3.23 shows the top words per cluster. One can see that cluster 0 clearly represents the interest rate decision (e.g. 'interest', 'rates', 'decided') whereas cluster 1 captures the introduction (e.g. 'welcome', 'gentlemen').

```
Top terms per cluster:
---------------------

Cluster 0 words: analyses, interest, based, regular, rates, decided

Cluster 1 words: welcome, gentlemen, conference, press, president, pleased
```

Figure 3.23: Top Words of K-Means Cluster

In general, one can choose the number of clusters either based on expert knowledge or via the within-cluster sum-of-squares criterion (inertia), which is calculated based on the sum of squares between each document and its cluster centroid:

$$\sum_{k} \sum_{d \in D_k} ||\vec{x}_d - \vec{u}_k||^2. \tag{3.3}$$

Table 3.1 shows the calculated inertias for our example.[23]

# Clusters	1	2	3	4
Inertia	16.5	2.67	1.0	0.0

Table 3.1: K-Means Inertias

Figure 3.24 illustrates how the error diminishes with the increasing number of clusters. The error decreases a lot when using two instead of one cluster, but only slightly when adding a third one. This 'kink' in the function can serve as a rule of thumb when deciding on how many clusters to incorporate in the clustering procedure. In our example, both the rule of thumb and our expert knowledge yield the same number of clusters (n = 2).

[23]Note, when using four clusters for four documents then the error is zero.

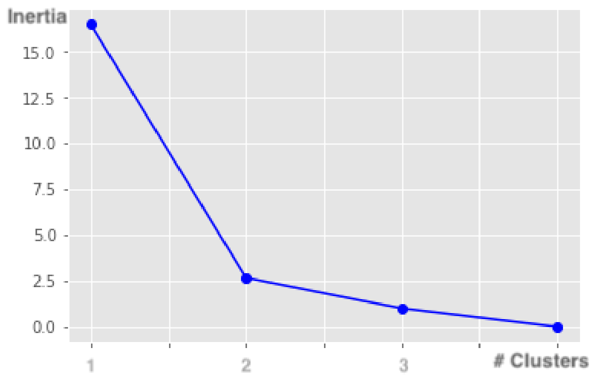


Figure 3.24: Inertia for K-Means Clustering

Other Algorithms

In general, the K-means algorithm produces a one hot encoding for each document, which means that it belongs to exactly one topic (see the topic label list in the previous example). This works well in certain applications but can be too restrictive for documents that contain several topics, e.g. information about inflation **and** economic growth. In these cases one should use mixed-membership models. A popular algorithm in this field is for example the Latent Dirichlet Allocation (LDA) (see Blei et al., 2003).

3.6 Tone and Sentiment

In this chapter we are going to discuss two main techniques to extract the sentiment or tone from a given document: First, based on a lexicon or wordlist which classifies specific words ex-ante into positive or negative. The second approach builds upon classification algorithms which learn from a labeled training set the words that are characteristic for a negative or positive sentiment.

3.6.1 Dictionary

A lot of textual analysis studies in finance use a simple wordlist to assess the tone of a certain document. This method works surprisingly well since people, often unconsciously, tend to use certain words (e.g. 'risk', 'uncertain', 'afraid') more often when they are in a specific mood (e.g. anxiety). If one then just counts the number of occurrences of these terms then this can yield already a meaningful sentiment score.

In general, this wordlist has to be specified prior to the analysis and thus depends on a subjective choice made by the researcher. Also such a pre-specified sentiment or affective lexicon focuses 'only on certain words, ones that carry particularly strong cues to sentiment or affect' (see Jurafsky and Martin, 2017, p. 327) rather than considering every term as a feature. This is advantageous in the sense that it reduces the dimensionality of the term-document matrix significantly by focusing only on a subset of words.

A popular lexicon that was used for example in Tetlock (2007) is the General Inquirer[24] which stems from psychology research and was developed by Philip Stone and his colleagues in the 1960s in Harvard (Stone et al., 1966). It comprises of 77 categories, covering e.g. negative, positive, strong, weak, active or pleasure related words.

A more recent dictionary which was developed especially for financial applications is the LM dictionary.[25] It was constructed in Loughran and McDonald (2011) based on the most frequent words in SEC 10-K company filings.

After selecting the wordlist one needs to define the actual sentiment measure. In the literature one often considers the fraction of the number of positive (negative) words P (N) to the total number of words in the document T. E.g. Schmeling and Wagner (2019) define their tone measure as:

$$1 - \frac{N}{T}.$$

Also one can derive a net tone measure, e.g. as used in Ardia et al. (2019):

$$\frac{P - N}{T}.$$

Dong et al. (2018) also use an extreme tone measure, defined as the maximum of the

[24] http://www.wjh.harvard.edu/ inquirer.

[25] https://sraf.nd.edu/textual-analysis/resources/.

positive or negative tone:

$$\max\left(\frac{P}{T}, \frac{N}{T}\right).$$

In general, dictionaries are based on the raw word count. If one applies e.g. the tf-idf weighting scheme (as defined previously) to the LM dictionary then it yields similar results as the Harvard dictionary in a stock return regression analysis (see Loughran and McDonald, 2011). Also Jegadeesh and Wu (2013) emphasize the importance of term weighting when using a dictionary. They develop a methodology which uses a simple regression approach and the stock market reaction to determine the corresponding term weights (direction and magnitude).

Das (2014, p. 52) points out that 'word count classifiers may be enhanced by focusing on "emphasis words" such as adjectives and adverbs, especially when classifying emotive content. One approach used in Das and Chen (2007) is to identify all adjectives and adverbs in the text and then only consider words that are within +/- 3 words before and after the adjective or adverb. This extracts the most emphatic parts of the text only, and then mood scores it.'

Example

We now illustrate the usage of a lexicon within our interest rate decision example. Therefore, we are going to define a simple central bank dictionary ourselves since other popular wordlist tend to misclassify terms in the monetary policy context.[26]

Based on our expert domain knowledge we specify the following positive and negative wordlists:

$$Pos = [decrease,\ lower,\ cut],$$

$$Neg = [increase,\ raise,\ hike].$$

[26] For example, the Harvard dictionary classifies the term 'decrease' as negative whereas a lower interest rate should be positive news to financial markets.

We now use the following definition of the tone:

$$\frac{P - N}{T},$$

with P, N and T as defined previously.

> *"Based ~~on our~~ regular economic ~~and~~ monetary analyses, ~~we~~ decided ~~to~~ keep ~~the~~ key ECB interest rates unchanged."*
>
> **Tone:** (0 - 0) / 12 = 0.00[27]

> *"Based ~~on our~~ regular economic ~~and~~ monetary analyses, ~~we~~ decided ~~to~~ **decrease** ~~the~~ key ECB interest rates."*
>
> **Tone:** (1 - 0) / 11 = 0.09

> *"Based ~~on our~~ regular economic ~~and~~ monetary analyses, ~~we~~ decided ~~to~~ **increase** ~~the~~ key ECB interest rates."*
>
> **Tone:** (0 - 1) / 11 = -0.09

> *"Ladies ~~and~~ gentlemen, ~~the~~ Vice-President ~~and I are~~ very pleased ~~to~~ welcome ~~you to our~~ press conference."*
>
> **Tone:** (0 - 0) / 8 = 0.00

Overall, one can see that in our example the defined tone measure (which uses our pre-specified wordlists) nicely captures the different semantics of the ECB statements. Nevertheless, in a more realistic setting this task is quite challenging since the choice of the appropriate dictionary is difficult (one cannot always create tailored wordlists) and also the meaning of the specific terms can be ambiguous and context-dependent.

In a next step, one can either calculate the aggregate tone across all statements or

[27]Note, for calculating the tone we do not consider the stopwords.

group it with respect to the different topics (e.g. interest rate decision and introduction). Then, one can use the tone measure within a regression model to examine for example the relationship between the sentiment of the statement and the corresponding return.

Boolean Search - Regular Expression

Another technique that uses keywords to reduce the dimensionality is the Boolean search. Whereas dictionary methods consider the intensity (frequency) of word use to be informative the Boolean search just looks for the occurrence of the specified keyword. The result of a boolean query is either 1 (word is in text) or 0 (expression is not included). To check whether a specific keyword occurs in a text one often uses regular expression, 'a language for specifying text search strings' (Jurafsky and Martin, 2011, p. 51 ff).

Within our previous example one could imagine that we are only interested in whether a statement is about the interest rate decision or not. To check this one can define the regular expression pattern

$$\textbf{pattern} := `\textit{interest rates}'$$

and search for it in all the texts (strings). If the expression occurs (at least once) in the respective statement the query responds with 'True' or '1' and thus we know that the document is, at least partly, about the interest rate decision.

In general, a pattern can be defined in a very flexible way. One can use combinations of alphabetical and numerical characters as well as other symbols (e.g. HTML tags). For example, when one needs to split central bank statements based on (bold) keywords, like $< strong > key\ ECB\ interest\ rates < /strong >$ or $< strong > economic\ analysis < /strong >$[28] one can define these expressions as patterns and search for them in the textual document. Based on their position in the

[28]See https://www.ecb.europa.eu/press/pressconf/2018/html/ecb.is181025.en.html.

text one can then split the document.

To construct the Economic Policy Uncertainty (EPU) index[29] (see Baker et al., 2016) or the Equity Market Volatility (EMV) Tracker (see Baker et al., 2019) the authors use Boolean searches. In particular, they specify three groups with certain keywords (for the EPU e.g. 'uncertainty' in group 1, 'economic' in group 2 and 'congress' or 'deficit' in group 3)[30] and retrieve all newspaper articles that contain at least one keyword from each category. Then, based on the raw counts of the retrieved articles they construct their respective index.

Lucca and Trebbi (2011) also rely on raw counts of Boolean type queries. They use the Google search engine to extract a semantic orientation score with respect to the expected interested rate decision of the Federal Reserve. Note, in this application the authors do not have control over the corpus (the Internet), they just receive the number of hit counts for their query (e.g. statement + 'hawkish').

3.6.2 Supervised Learning - Classification

As an alternative to using an ex-ante specified or selected dictionary one can 'let the data speak' and apply supervised machine learning techniques to determine the document sentiment. According to Murphy (2012, p. 2), in the 'supervised learning approach, the goal is to learn a mapping from inputs x to outputs y, given a labeled set[31] of input-output pairs $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$. Here $\mathcal{D}$ is called the training set, and N is the number of training examples.' As a result these machine learning algorithms determine ex-post which input values (e.g. words) are most important for a specific category or class.

This methodology works especially well with shorter documents where hardly any terms from a wordlist are used, like social media tweets. Antweiler and Frank (2004)

[29] See http://www.policyuncertainty.com.

[30] The EMV keywords are for example: 'economic' in group 1, 'stock market' in group 2 and 'volatility' in group 3.

[31] Class labels could be for example 'positive', 'neutral' or 'negative'.

apply a supervised machine learning algorithm to classify more than 1.5 million messages about U.S. companies posted on Yahoo! Finance and Raging Bull. Based on the different number of positive and negative messages they construct a simple bullishness index:[32]

$$B_t \equiv \frac{M_t^{BUY} - M_t^{SELL}}{M_t^{BUY} + M_t^{SELL}},$$

where M_t is the respective number of buy (sell) messages.

To classify the messages Antweiler and Frank (2004) apply the simple Naive Bayes algorithm.[33] As a robustness exercise they also use a Support Vector Machine[34] for their classification task which yields similar results in their study.

A critical issue when applying a supervised machine learning technique is to obtain the class or category labels. One possibility is to use the Web which 'contains an enormous number of on-line reviews for restaurants, movies, books, or other products, each of which have the text of the review along with an associated review score: a value that may range from 1 star to 5 stars, or scoring 1 to 10' (see Jurafsky and Martin, 2017, p. 333 ff). In a financial application one can often use stock returns or other asset prices to obtain the class label (e.g. return ≥ 0 and < 0). Alternatively, one can either manually classify the text or crowdsource a training set (e.g. via Amazon Mechanical Turk).[35]

An overview of financial studies using dictionary or supervised machine learning methods for sentiment analysis is presented in Kearney and Liu (2014). Further details and explanations with respect to sentiment analysis in finance can also be found in Mitra and Yu (2016).

---

[32]This idea is similar to counting positive or negative words when using the lexicon approach.

[33]See Jurafsky and Martin (2017, Ch. 6) for a detailed explanation on how to use this algorithm for sentiment classification.

[34]See Murphy (2012, Ch. 14.5).

[35]See https://www.mturk.com.

3.7 Text Regression

After transforming the qualitative textual data into a quantitative representation and after extracting the topic and/ or tone we can now use standard statistical and machine learning methods to analyze the relationship between the information revealed and e.g. the financial markets. When performing these analyses one always has to keep in mind the high-dimensionality of the input data which can easily lead to multicollinearity problems due to the curse of dimensionality (see Murphy, 2012, p. 18 f).

In the following we will briefly discuss the most popular methods used in the (financial) text analysis literature to avoid this problem.

3.7.1 Linear Regression Model

When using a standard linear regression model one usually cannot plug-in all the unique words as an input variable due to the multicollinearity problem (see also the introductory example from Chapter 3.1). Instead one needs to reduce the dimensionality up-front, e.g. through working with the topics and the tone of the document. The comparatively low number of topic dummy variables or tone measures serve then as the input for the regression model avoiding the curse of dimensionality problem.

Topic

Equation 3.4 shows the specification of the linear regression model when using the topics as the explanatory variables:

$$Y_t = \beta_0 + \beta_1 \times \mathcal{I}_t^{(Topic\ 1)} + ... + \beta_K \times \mathcal{I}_t^{(Topic\ K)} + \epsilon_t, \tag{3.4}$$

with Y_t being for example the time series of stock returns and $\mathcal{I}_t^{(Topic\ k)}$ an indicator function which is one when the document covers topic k.

In our ECB statement example from the previous chapters we identified two clusters (topics): interest rate decision and introduction. For our analysis we would use the time series of both dummy variables over our press conference sample to assess the impact of the two topics on the stock market (the respective price changes would serve as the dependent variable Y_t). As a result we would get the individual average asset price response to each information revealed. This approach is used in Chapter 4 of this thesis to quantify the real-time response of financial markets to the different content of ECB press conferences.

Tone

Compared to the general topic it is also important to quantify the actual semantics of the respective document. For example, whether the information revealed carries a more positive or negative sentiment for investors. Schmeling and Wagner (2019) construct an aggregate tone measure for the intro statement of ECB press conferences and study the impact of the tone change on European stock returns and bond yields using the following specification:

$$Y_t = \beta_0 + \beta_1 \times \triangle Tone_t^{(Agg)} + \epsilon_t, \tag{3.5}$$

with Y_t being the time series of asset price changes and $\triangle Tone_t^{(Agg)}$ the shift in aggregate sentiment for the whole ECB statement.

Topic and Tone

A drawback of the previous aggregate tone measure is that one cannot disentangle the different effects that the various information have on asset prices. Therefore, Hansen and McMahon (2016) first extract several financial and economical topics from the Federal Reserve's FOMC statements and then analyze the impact of each topic's tone on financial markets. In our model specification we can account for this by using an interaction term between the topic and tone measures:

$$Y_t = \beta_0 + \beta_1 \times \mathcal{I}_t^{(Topic\ 1)} \times Tone_t^{(Topic\ 1)} + ... + \beta_K \times \mathcal{I}_t^{(Topic\ K)} \times Tone_t^{(Topic\ K)} + \epsilon_t, \quad (3.6)$$

with the variables as defined previously.

A similar methodology is also used in Jegadeesh and Wu (2017) who first analyze the information content of the overall FOMC minutes using a tone measure and then study the individual topics.

3.7.2 Penalized Regression

When using the full term-document matrix as an input one needs a mechanism that discriminates within the high-dimensional input vector. One popular technique from statistics is the ridge regression (see Hoerl and Kennard (1970) or Murphy (2012, Ch. 7.5)).

A ridge regression is similar to a standard linear regression but incorporates a penalty term that punishes the complexity of the model[36] (hence it is also called penalized least squares). Specifically, it adds a $\ell 2$ regularization term to the normal objective function (see Murphy, 2012, p. 226 f):

$$J(w) = \frac{1}{T}\sum_{t=1}^{T}(Y_t - (\beta_0 + \beta_1 \times X_t^{(Term\ 1)} + ... + \beta_V \times X_t^{(Term\ V)}))^2 + \lambda \sum_v \beta_v^2,$$

with the first term being the mean squared error as usual and the second term the complexity penalty. λ is a free so-called hyperparameter that determines the 'strength' of the penalty term.[37]

[36] Given the penalty term one cannot estimate the regression model via Ordinary Least Squares (OLS) but instead has to use, for example, a Maximum Likelihood Estimator (MLE).

[37] Note, to determine this hyperparameter one can use cross validation techniques (see Murphy, 2012, p. 207f).

In general, adding the regularization term encourages the parameters to be small and thus can prevent multicollinearity. A related technique is the lasso regression which uses the $\ell 1$ regularization instead of the $\ell 2$ norm.

This regression technique is for example applied in the study of Ardia et al. (2019) who use the penalized least squares method to discriminate between a high-dimensional input vector of sentiment scores obtained from different dictionaries and weighting schemes to forecast economic growth in the United States. Generally, in such a regression setup one can use the full term-document matrix as an input and let the algorithm decide (ex-post) which words are most informative.[38]

3.7.3 Other Regression Techniques

The standard linear regression model and also the penalized least squares assume a linear relationship between the dependent and the independent variables which could be too restrictive in some settings (especially when working with textual data). Therefore, Manela and Moreira (2017) use a (non-linear) support vector regression (see Murphy, 2012, Ch. 14.5.1) to analyze the correlation between uni- and bigrams of the Wall Street Journal abstracts and the VIX index. Based on the found relationship (trained model) they construct the news-implied volatility index (short NVIX) which can be calculated back until July 1889.

Kelly et al. (2019) develop a high-dimensional selection model, the hurdle distributed multiple regression (HDMR), which improves machine learning from text. Their framework can be used to backcast, nowcast, and forecast financial variables using newspaper articles and outperforms alternative state-of-the-art approaches in an out-of-sample exercise.

[38] This is different to the dictionary approach where one decides ex-ante which terms are most important.

Another methodology is called the ‘Principal Component Regression’ (PCR) which basically first applies a PCA to the term-document matrix and then includes the latent factors in a then standard regression model. Another option is to use regression trees to account for the non-linearity in the analysis.

Classification Exercise

Instead of running a regression analysis to examine the relationship between financial variables and text one can formulate the problem as a classification task: Class 1 corresponds to e.g. positive returns (≥ 0) and Class 2 to negative ones (returns < 0). But in the financial text analysis literature this approach is seldom used.[39]

[39] In general, one can use the same algorithms as presented in Chapter 3.6.2 of this work.

Chapter 4

The Real-Time Impact of ECB Press Conferences on Financial Markets

4.1 Introduction

'The arrival of news continually updates an investor's understanding and knowledge of the market and influences investor sentiment, [...] hence asset prices, asset price volatility and risk' (Mitra and Mitra, 2011b).[1] In todays information society all kinds of news stream in 24 hours a day from a variety of sources thus it is challenging to understand the overall information flow and in particular to 'identify the piece of information that initially triggers the change' (Moniz et al., 2011).

According to financial theory, stock prices rapidly incorporate all relevant information about a company (Fama et al., 1969) leading to the notion of 'efficient markets' (see Fama, 1970). Given todays speed in the market place with which news are incorporated into prices[2] we can use the real-time response of asset prices to gauge the information content of the different news elements and are able to identify 'what type of information causes [e.g.] analysts to revise their earnings expectations' (Moniz et al., 2011) or whether information about the economic outlook revealed by a central bank matters to

[1] According to Barber and Odean (2008) 'significant news will often affect investors' beliefs and portfolio goals heterogeneously, resulting in more investors trading than is usual'.

[2] Recent developments in automatic and high frequency trading lead to trades executed within milliseconds (see Mitra et al., 2011).

market participants.

In this chapter we develop a new methodology that explicitly 'allows [us] to trace the information flow, and thus to investigate to what type of statements financial markets react predominantly' (Ehrmann and Fratzscher, 2009) using Webcasts of the press conferences accompanying the communication event.[3] Using the captions of the video and modern textual analysis tools we fully automatically create timestamps for the different news items which allows us to align them with the corresponding asset price respones in the financial markets and thus study their information content. To the best of our knowledge we are the first study that uses the information implicit in the subtitles of videos.

Our approach is comparable to exploiting the news flow of a financial news service provider as in Ehrmann and Fratzscher (2009) who use Reuters news snaps for 13 consecutive month (2004 - 2005) to study the impact of ECB statements on financial markets. But working directly with the press conference content avoids several caveats as outlined in Ehrmann and Fratzscher (2009): For example, statements can be wrongly reported or misinterpreted. Also different newswire services can report different news items based on the same original source. Our methodology overcomes these problems since we work directly with the press conference statement. In addition, we can use textual analysis tools to exploit the structure inherent in these statements and thus can examine the impact of fine granular topics. Another advantage of our methodology is that we extract the exact time the ECB president needs for reading the respective part of the statement while news snaps from a newswire service always come with a delay (since journalists have to prepare and publish them). In general, our approach allows to study every communication event with a press conference, even if it is not covered in detail by the news service provider (or the coverage has only started lately).

[3]In a press conference the information gets revealed piece by piece which enables us to trace the information flow. In contrast, a press release reveals all its information at once and thus it is difficult to study the impact of the individual information on financial markets.

Also closely related to our analysis is the work of Busse and Green (2002) who use taped videos of the Morning Call or Midday Call TV show from CNBC. While the authors manually record the videos and the starting times our methodology makes use of the Webcast videos (which are available for the last years) and of financial news services to extract the exact start time of the communication event. Instead of manually labeling the information content we identify the topics using modern textual analysis tools. Thus, our methodology is fully automatic and can be easily applied to a long history of press conferences.

We then apply our newly developed methodology to the press conferences of the European Central Bank[4] whose broadcasts are highly followed by market participants. By applying our methodology to the ECB we can directly relate the mean-reversion of the pre-ECB announcement drift (as documented in Schmeling and Wagner, 2019 and Chapter 2 of this thesis) to the information revealed in the introductory statement. Using fixed income futures we can also shed light on the question which topic contributes the most to the communication shock (as defined in e.g. Leombroni et al., 2017) or has the largest impact across the yield curve.

In detail, applying our methodology to the ECB press conference statements from 2011 to 2017[5] yield the following results:
First, we document that the pre-ECB announcement return as shown in Schmeling and Wagner (2019) and in Chapter 2 of this thesis further increases by 20% (+10 bps) with the start of the press conference. We find that this appreciation in equity prices is entirely driven by the non-standard monetary policy measures as announced by the ECB, in particular the refinancing operations, the asset purchase programs and by further implementation details and comments.

[4]The ECB was one of the first central banks that streamed their press conferences and thus provides the longest history of videos.

[5]The sample period is defined by the starting point of the ECB webcasts and our data availability.

Second, we observe that after this first initial increase equity returns drop by about half the size of the pre-ECB run-up (-24 bps). Information from the ECB's economic analysis account for over 50% of this fall, with a roughly equal contribution of real economy and inflation news. We can relate the former to information about the course of the European sovereign debt crisis and its subsequent recovery.

Third, information about the monetary analysis (e.g. money growth) seem to be of second-order importance to investors with a small estimate of -4 bps.

Fourth, we find that comments by the ECB Chairman about the situation of the European banking sector and the implementation of structural reforms and fiscal policy within the EU member states further decreased prices in the equity market by -7 bps.

Fifth, using the rich cross-section of Euro Stoxx supersector and Stoxx country indices we find that the positive effect of the ECB's non-standard monetary policy measures and the negative impact of the economic information is particular pronounced for the banking and insurance sector as well as the GIIPS countries of our sample. These industries and countries have been at the core of the euro crisis (also connected via the bank-sovereign nexus) and thus very sensitive to the announced actions of the central bank as well as its superior economic (outlook) information.

Sixth, using fixed income futures on German, French and Italian government bonds we document that the non-standard monetary policy announcements and the information on the real economy have the biggest impact on yield changes. The effect is particularly pronounced for the long-end of the yield curve and here specifically for the Italian government bond.

The remainder of the chapter is organized as follows. Section 4.2 provides an overview over the related literature. Section 4.3 presents our new method to create the content timestamps. Section 4.4 explains our clustering approach and Section 4.5 describes the financial data we use in our empirical analysis. Section 4.6 shows our empirical results and Section 4.7 concludes this study.

4.2 Literature

Our analysis contributes to several strands of literature. The first strand of literature quantifies monetary policy and growth shocks through studying asset price responses around central bank's announcement days. The seminal paper of Kuttner (2001) proposes to measure the unexpected change in the target funds rate with changes in the price of Federal funds futures settling in the month of the meeting. Bernanke and Kuttner (2005) also use the change in Federal funds to examine the impact of (unanticipated) changes in monetary policy on equity prices. Gürkaynak et al. (2005) improve on this single-factor analysis by extracting two latent factors, to proxy for the information content of FOMC statements, using a standard principal component analysis from the asset price response (money market rates) in a narrow window around the release of the information. They label these factors 'current federal funds rate target' and 'future path of policy'. Brand et al. (2010) and Leombroni et al. (2017) build upon this methodology and exploit the institutional feature of the ECB with a separate announcement for the monetary policy decision and its explanation to extract separate target and communication shocks from high-frequency asset prices. Recently, Altavilla et al. (2019) extend this literature by extracting both conventional and unconventional monetary policy communication surprises from the press release and press conference windows.

Other studies exploit the comovement of bonds and stocks to distinguish between monetary policy and economic growth shocks. Jarocinski and Karadi (2018) disentangles the two components using a structural vector regression and Cieslak and Schrimpf (2018) exploit the comovement combined with monotonicity restrictions across the yield curve to separate the monetary policy, economic growth shocks and shock to risk premia. They find that the non-monetary news accounts especially during a economic crisis period.

A problem with these 'indirect' asset price based approaches is that it is unclear what kind of information is revealed to investors (Woodford, 2012) and thus driving the actual shock(s).[6] Instead, our approach aligns the different information revealed by the central bank with the corresponding returns and yield changes and thus allows to study directly the different information content. In addition, our methodology quantifies information shocks on a fine granular level which, first, allows us to separate the effect of standard versus non-standard monetary policy measures (see target vs. QE factor in Altavilla et al., 2019) and to distinguish between monetary and non-monetary news (see Cieslak and Schrimpf, 2018) in one framework and second, to even further decompose the different news shocks.[7]

The second strand of literature studies directly the impact of information revealed by the central bank on financial markets (instead of infering the news component from asset prices). In earlier work narrative approaches and subjective communication indicators have been used to quantify the qualitative textual information. Nowadays, modern textual analysis tools gain popularity for analyzing central bank statements in an automatic and objective manner.

In their important contribution Romer and Romer (2004) use a narrative approach to subjectively identify monetary policy shocks from central bank statements. Subjective indicators (glossaries) are also used by other authors to measure the tone of ECB communication and to analyze its impact on financial markets (see e.g. Rosa and Verga, 2007, 2008; Heinemann and Ullrich, 2007; Gerlach, 2007 or Berger et al., 2011 amongst others). Rosa (2011) first uses a narrative approach in the sense of Romer and Romer (2004) to measure the tone of FOMC statements and then estimates the market expectations of the Fed's announcement using a forecasting regression for extracting the surprise component.

Compared to the subjective indicators Lucca and Trebbi (2011) use tools from computational linguistics to develop an automated and objective procedure to measure central

[6]See also the criticism of Lucca and Trebbi (2011) who state that 'such an immediate approach is, however, limited by its indirectness. Econometric models measuring the impact of communication on interest rates based on these measures, in essence explain rates with rates'.

[7]For example, we can decompose the economic growth shock into information about inflation and the real economy and the latter into news about the economic outlook compared to realized GDP information.

bank communication and apply it to FOMC statements. Hansen and McMahon (2016) also deploy tools from computational linguistic to analyze in addition to the monetary policy stance the economic information revealed in the FOMC statements based on a topic and tone approach. Jegadeesh and Wu (2017) extract the general topics of FOMC minutes using a LDA clustering approach. Schmeling and Wagner (2019) analyze the ECB's introductory statement and document that a change in central bank tone has a significant effect on asset prices. Hansen et al. (2018) use the Bank of England's Inflation Report to study how quantitative and qualitative information about future economic conditions affect the short and long end of the yield curve.

With this chapter we contribute to this strand of literature with respect to two dimensions. First, we use modern textual analysis tools to (objectively) exploit the template structure of the ECB's introductory statements which allows us to study the revealed information on a fine granular level which is new to the literature. Second, we develop a fully automatic methodology to create timestamps for the different information pieces using the videos of the corresponding press conferences. This allows us to study the real-time impact of the information revealed by the ECB Chairman on financial markets and thus to quantify the different information shocks (news component).

Third, our analysis relates to the recent literature documenting and explaining announcement returns around central bank decision days. For example, Lucca and Moench (2015) documents a pre-FOMC announcement drift for the U.S. equity market that persists after the FOMC meeting. The price pattern exists across various industries and countries and for the U.S. market also over the full FOMC cycle (see Cieslak et al., 2019). Chapter 2 of this thesis and Schmeling and Wagner (2019) document similar pre-ECB announcement returns for the European equity market, especially for the euro crisis period. Compared to the U.S. returns the EU drift mean-reverts with the begin of the ECB's press conference resulting in large negative post-announcement returns.

We contribute to this literature by identifying the information revealed in the press conference that leads to the decline in European equity returns.

Fourth, with this chapter we contribute to the literature that evaluates the ECB's unconventional monetary policy actions during the European sovereign debt crisis. Acharya et al. (2018) for example find that the ECB's long-term refinancing operations (LTRO) temporarily reduced the funding pressure for banks, but did not help to contain sovereign risk. In contrast, the ECB's announcement of the Outright Monetary Transaction (OMT) program was able to reduce the bank-sovereign nexus. Krishnamurthy et al. (2018) also compare the effects of the Securities Market Programme (SMP), the OMT and the LTROs on government bond yields and find that on average across Italy, Spain and Portugal the SMP and OMT program lead to significant lower yields. Decomposing the fall they find that reduced market segmentation accounts for 50% of this reduction in yields. Other studies (amongst others) also find a beneficial impact of the ECB's non-standard monetary policy measures for asset prices in the euro area (see Fratzscher et al., 2016) and that the announcements of asset purchase programmes have lowered market uncertainty (see Coenen et al., 2017).

We contribute to this strand of literature by quantifying the real-time response of the financial markets to the announcement of the ECB's non-standard monetary policy measures.

4.3 Methodology

In this section we introduce a novel methodology to the literature which uses the captions of press conference webcasts and their start time to automatically create timestamps for the different information content. These timestamps allow us to explicitly study the information flow and also align it with the price responses in the financial markets which can then be used to directly assess its information content. To the best of our knowledge this study is the first that makes use of the information contained in the subtitle files. In the following we are going to apply our methodology to the press conferences of the European Central Bank.

4.3.1 Start Times

As in Fleming and Piazzesi (2005) and Lucca and Moench (2015) we rely on timestamps of Bloomberg newswires.[8] Around ECB monetary policy meetings there are several news stories published regarding the progress of the press conference. First, usually a few minutes before the actual press conference a reminder on the event is published. As shown in Figure 4.1 for the press conference held on 25th of January 2018 at 14:28 CET the news says 'Mario Draghi Speaks at ECB News Conference: LIVE <GO>'. Second, when the ECB Chairman and Vice-President enter the press conference room a further news element is published saying e.g. 'MARIO DRAGHI PRESIDES AT ECB NEWS CONFERENCE IN FRANKFURT: LIVE' (14:31 CET). Third, another news element gets published when the ECB Chairman or the host of the press conference starts to address the audience with some personal remarks or with reading the introductory statement, e.g. '*ECB PRESS CONFERENCE STARTS IN FRANKFURT: LIVE' (14:32 CET). This latter news element indicates the actual time when the speech of the ECB chairman starts and is explicitly marked with a '*' by the Bloomberg service. Fourth, for every information that gets revealed in the press conference another news item is published. The first of these news items is usually about the monetary policy decision which follows after the introduction. For our further analysis we use the issuance timestamp (which is available on a second basis, e.g. 14:32:13) of the third news item which indicates the actual start of the press conference and which aligns with the start of our caption.[9][10] A full list of the start times is reported in Table A.2.

[8] In general we consider German and English news items and take the element that is issued first.

[9] For the earlier part of our sample Bloomberg only issued one news item regarding the start of the press conference compared to the two elements (2) and (3) as shown in Figure 4.1. We then use the issuance timestamp of this item. Whenever there is no news item before the start of the press conference we exclude the respective meeting from our sample. This is true for the first four monetary policy decisions in our sample.

[10] As a robustness check we manually compare every timestamp of the first content news item (e.g. 14:33:42 in Figure 4.1) with the timestamp obtained from our allocation procedure. Overall, we observe a very close fit which gives us confidence that our reported start times are close to the true starting points. We exclude the meetings from January 2017 and April 2018 since the time difference between the last item before the start of the meeting and the first content news is too large. We keep the press conference from October 2011 because it starts with introductory remarks by Bundesbank President Weidmann which naturally extends the time difference between the start and the first content news item.

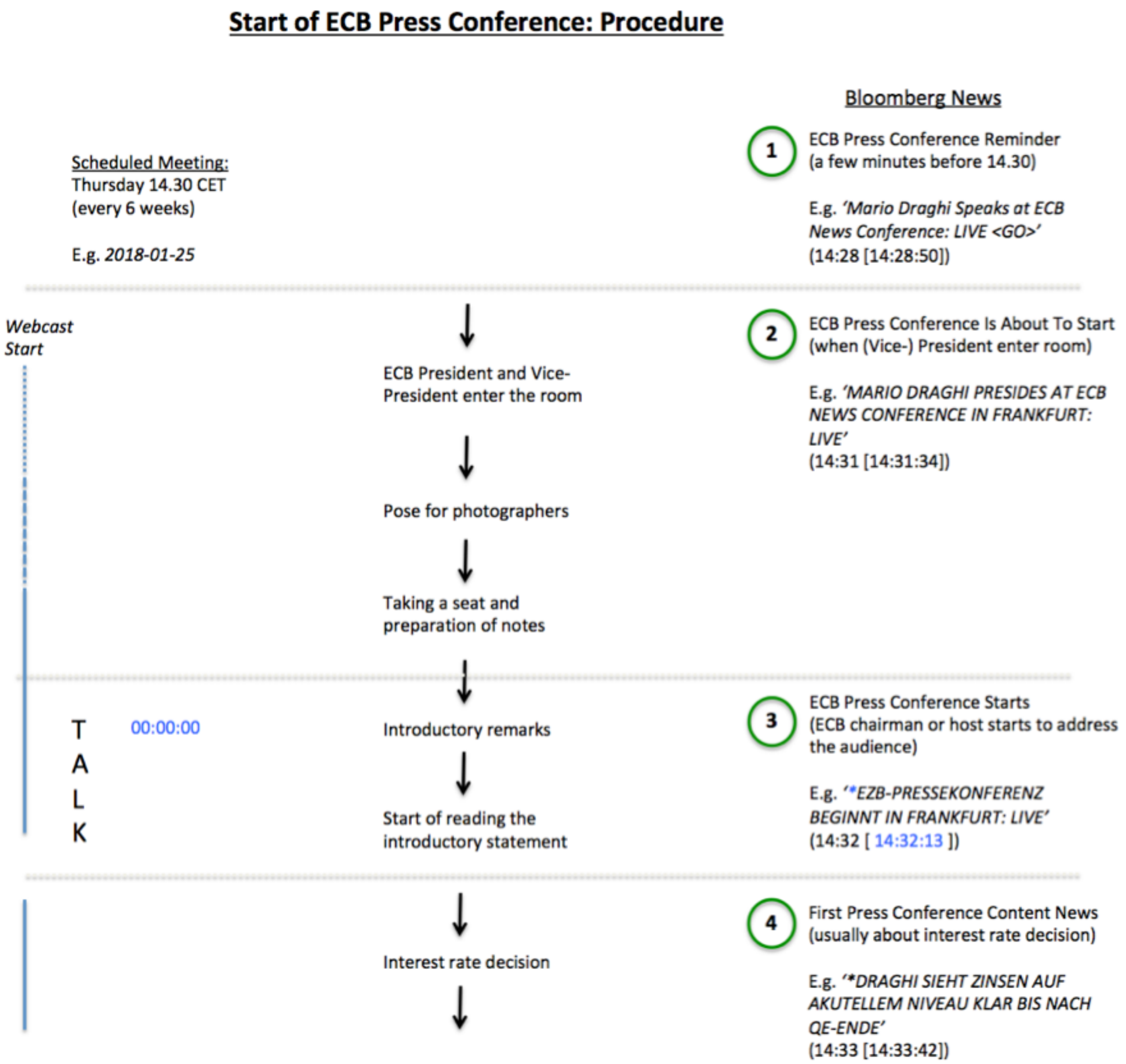


Figure 4.1: **Starting Procedure of the ECB Press Conference**
This figure illustrates the usual starting procedure of an ECB press conference. The dotted, dashed and solid vertical line indicates the different start times of the ECB webcast.

4.3.2 Subtitles

For our further analysis we use automatically generated subtitles obtained from ECB press conferences webcasts. These videos are provided by the ECB via its Web page since

January 2011[11] Figure 4.2 shows a subtitle example for the press conference which was held on 4th of August 2011. The SubRip text (.srt) files comprise formatted lines of plain text which are grouped in blocks separated by a blank line. Every block is numbered sequentially, starting at 1. The next line contains the start and end timestamps of the text that follows. The timecode format is hours:minutes:seconds,milliseconds. After the timestamps the actual subtitle text is reported. Due to the defined structure of these .srt files we can use a parsing algorithm to extract the timestamps and text information for each block and press conference.[12]

```
1
00:00:00,000     00:00:04,410
00:00:00,310 --> 00:00:04,720
ladies and gentleman the vice president
and I are very pleased to welcome you to

2
00:00:04,410     00:00:05,580
00:00:04,720 --> 00:00:05,890
our press conference

3
00:00:05,580     00:00:10,130
00:00:05,890 --> 00:00:10,440
we will report on the outcome of today's
meeting of the Governing Council which
```

Figure 4.2: **Caption of ECB Press Conference Webcast**
This figure shows the beginning of the .srt file containing the subtitle for the ECB press conference which was held on 4th of August 2011. In red are shown the normalized timestamps which are obtained by substracting the start time of the first cell.

The timestamps of the subtitle blocks are reported relative to the starting point of the webcast. Depending on whether the video already includes the scene when the ECB staff is entering the press conference room or whether the video starts right before the

[11] Link: https://www.ecb.europa.eu/press/tvservices/webcast/html/index.en.html.

[12] Generally, the idea of exploiting the timestamp information inherent in subtitles is similar to using the videocassette-recorder counter as in Busse and Green (2002).

ECB Chairman reads the statement the value of the first timestamp varies significantly.[13] Usually, the first words that are recorded in the caption are when the microphone gets turned on and the ECB Chairman addresses the audience which we define as the start of the press conference. We then substract the start time of the first block from all other timestamps to have a normalized starting time of 00:00:00 for each press conference.[14] All other timestamps are then relative to the moment when Draghi starts to talk. Since this time point also corresponds to the reported '*' Bloomberg timestamp we can then add up this normalized relative timestamp and the start time of the press conference. For example, in Figure 4.2 we first substract 310 milliseconds from all other timestamps to set the starting time of the first block to 00:00:00,000 and then add the start time 14:30:43 (from Table A.2). The next subtitle block 'our press conference' does then coincide to the start (end) timestamp 14:30:47,410 (14:30:48,580) and so on.[15][16]

4.3.3 Matching Algorithm

To obtain the start and end time of every sentence of the ECB introductory statement we first split the text using standard textual analysis techniques (sentence tokenizer) and then match each sentence of the transcript with the subtitle of the respective press conference. For this matching task we employ an algorithm that compares pairs of sequences and returns the best match for a given string.

For example, if we want to find the start and end time for the first sentence of

[13]During our sample period from January 2011 to April 2018 the webcasts mostly start right before the audience gets addressed.

[14]Only for the press conference on 3rd of December 2015 the microphones already record a 'Thank you' during shooting the photos. We adjust the timestamp so that the beginning coincides with the real start of the press conference as defined above.

[15]For the earlier part of our sample the ECB press conference webcast does not include the Q&A session. Also most of these videos stop right before the 'Disposal for Questions' expression. Thus, for the further analysis we select the part of the introductory statement that starts with 'Ladies and Gentlemen' and ends right before the 'Disposal for Questions' (or similar) expression (like in Schmeling and Wagner, 2019).

[16]We exclude three press conferences from the further analysis. First, the press conference from 5th of May 2011 since the video does not include the first paragraph of the introductory statement and thus misses the start of the speech. Additionally, we remove the press conferences that were held on the 8th of May 2014 and 22nd of October 2015 since for both meetings the ECB Chairman started reading the transcript while the microphone was not yet turned on. Also we exclude the introduction by Bundesbank president Weidmann from 6th of October 2011 which preceded Trichet's introductory statement.

the introduction from the ECB press conference held on 4th of August 2011 ('Ladies and gentlemen, the Vice-President and I are very pleased to welcome you to our press conference.') we compare this sentence (sequence one) with every block from the caption (sequence two) and take the timestamps of the best match as indicated by the highest score. As can be seen in Figure 4.2 the best match in our example comprises of block one and two which yields the following start (end) timestamp 00:00:00,000 (00:00:05,580). The respective score is 0.95 which is close to the maximum score of 1.0 (which describes two identical sequences).

Since the cells of the caption do not ncessarily cover a full transcript sentence we also use a combination of blocks in the matching procedure. For example, in Figure 4.2 we look for a match not only in the first block on its own ('ladies and gentleman the vice president and I are very pleased to welcome you to') but we also use the combination with block number two ('ladies and gentleman the vice president and I are very pleased to welcome you to (+) our press conference'). The start timestamp is then taken from the first block, the end timestamp from the last block. Overall, we use up to fifteen combined blocks in the matching procedure.[17] We create these combinations on a rolling basis, shifting it always by one block (e.g. Block 1+2, 2+3, 3+4, ...).

Table 4.1 shows descriptive statistics for the matching results. On average we observe a very high score of roughly 0.9 with a very low standard deviation of 0.07 for the matching of the official transcript sentences with the corresponding captions which yields a very good matching result. This can also be seen in the relatively large 5% quintile which shows that 95% of the matches have a score higher than 0.76. Overall, we only observe eight sentences (out of 3641) with a score lower than 0.60 which is around 0.2% of the total sentences. Such a low score can be observed for example if a sentence from the transcript is not said in the press conference and thus the algorithm tries to match it with some other sequence. Also correct matches can have a lower score when e.g. the

[17]The maximum number of combinations used in the matching procedure is twelve.

automatically created caption exhibits spelling errors or uses written out numbers instead of a numerical representation.[18] We manually check all of these low score sentences and exclude bad matches.[19]

The average length of a press conference sentence is around 14 seconds. The shortest sentence spans only two seconds and is said on the press conference held on the 3rd of February 2011 ('We therefore decided to leave them unchanged.'). The maximum length of a sentence is roughly 50 seconds, 90% of the sentences have a duration between 6 and 25 seconds.

4.3.4 Paragraph and Category Time Stamps

Based on the timestamps for the individual sentences we also derive the start and end times for coarser parts of the statement: Paragraphs and categories. A paragraph comprises of several sentences and is separated from the next one by a blank line.[20][21] A category comprises of several paragraphs and is directly related to the ECB's communication approach as explained in the following.

Since 2003 the ECB follows a clear defined structure in their introductory statement as presented in a press seminar on 8th of May 2003. First, they announce the monetary policy decision which 'aim[s] to maintain inflation rates close to 2% over the medium term'. Then, they use the two-pillar structure which provides the support for the policy decision. The economic analysis 'focus[es] on shorter-term price movements' and is followed by the monetary analysis with the 'focus on longer-term price trends'. These two different type of information are then cross-checked so that 'based on all information [the] Council comes to a single assessment' (see Issing, 2003). Based on that structure we derive four categories which make up every press conference statement: (1) 'Monetary

[18] Due to these issues the minimum score of 0.39 still belongs to a correct match.

[19] As an additional robustness check we plot the time difference of the start (end) timestamp of every sentence to the previous start (end) timestamp. A wrong match would show up as an outlier. We do not observe such a data point after removing the wrong low score matches.

[20] It can be easily extracted from the statement's HTML page via the $< p >$ elements.

[21] Jegadeesh and Wu (2017) also use the paragraph level in their analysis of FOMC minutes.

Table 4.1: **Matched ECB Press Conference Transcripts**
This table reports descriptive statistics for the matching procedure of press conference sentences with the corresponding captions of the ECB webcasts. The column 'Score' refers to the matching score reported by the algorithm and lies between 0 and 1 (1 indicates a perfect match). The length of a sentence, paragraph or category is reported in minutes and seconds. 'Paragraphs' ('Categories') refers to the paragraphs (categories) of the ECB introductory statement comprising several sentences (paragraphs). The sentences are obtained from official ECB press conference transcripts starting June 2011 until March 2018.

		Score	**Length [min:s]**
Sentences	Mean	0.89	00:14
	Std dev	0.07	00:06
	Min	0.39	00:02
	Q5	0.76	00:06
	Q25	0.86	00:10
	Median	0.91	00:13
	Q75	0.94	00:18
	Q95	0.97	00:25
	Max	0.99	00:47
	Obs.: 3641		
Paragraphs	Mean		00:56
	Std dev		00:33
	Obs.: 906		
Categories	Mean		03:11
	Std dev		01:20
	Obs.: 264		
Number of ECB PCs:	66		

Policy Decision', (2) 'Economic Analysis', (3) 'Monetary Analysis' and (4) 'Summary' which contains the 'Cross-Check', 'Structural Reforms' and 'Fiscal Policy'. In the statement the keywords of these categories are marked bold and thus can be extracted in a consistent and transparent way using textual analysis tools.

For the start time of a category we use the starting time of the first sentence of the respective category. For the end time we use the starting time of the next category or the end time of the last sentence of the current one (if it is the last category of the statement). This approach produces non-overlapping category timestamps.

Figure 4.3 shows the category length of each press conference within our sample period. First, one can see that the aggregate length matches the one obtained from the original videos.[22] Also our time series of the duration of the press conferences is similar to the plotted number of sentences shown in Cieslak and Schrimpf (2018). Second, the duration of the monetary policy decision category always shows a peak when the ECB announces their non-standard monetary policy measures. Third, the economic analysis takes always longer when the two macro projection paragraphs are added in the statement. The macro projections are included every quarter and thus every third press conference until 2015. Starting 2015 they are incorporated in every second press conference (due to the 6 week cycle). Forth, for the monetary analysis we can see a longer category duration in the first half of our sample which is due to the euro crisis in which the ECB added new paragraphs to the statement regarding the special situation of the European banking sector. After the height of the crisis these parts were dropped again which can be seen in a lower category length. The last part contains the summary paragraph and the remarks on structural reforms and fiscal policy. At first, the summary was rather extensive repeating most of the information from the monetary policy decision. After the first press conferences in our sample this was reduced to a couple of sentences which also lowered the category duration significantly.

[22]One can see notable time difference for the press conferences from October 2011 and May 2012 since in our analysis we exclude the introductions by Bundesbank president Weidmann (2011) and by Governor Ordóñez (2012). Also Draghi's introductory remarks in December 2014 about the new ECB premise are excluded.

Overall, the economic analysis category is the longest with 4 minutes and 18 seconds, before the monetary policy decision with 3 minutes and 42 seconds. The monetary analysis takes on average 2 minutes and 23 seconds and the last part 2 minutes and 20 seconds. The full press conference takes on average 12 minutes and 43 seconds.

Analogously to the categories we use the starting time of the first sentence of the respective paragraph as the paragraph's start time. For the end time we use the next starting time or the end time of the last sentence of the paragraph (if it is the last paragraph for a given press conference).
Unlike the four categories not every paragraph is included in each press conference statement and thus we cannot label them according to a hard-wired rule. Instead we need a more flexible procedure that groups similar paragraphs together. Our clustering approach is explained in the next section.

4.4 Clustering

The introductory statements of the ECB's press conferences are based on a template which gets adjusted from meeting to meeting to reflect the central bank's current assessment and its intended communication to the public.[23][24] Thus, in addition to the category structure as described in Issing (2003) the statements also exhibit a strong similarity on the paragraph level. To exploit this structure and extract the topics we use a standard clustering algorithm as it is popular in the literature.[25] In the following we will outline our approach.

[23]This is a quite popular approach among central banks which 'often [...] start from the previous statement, and update the earlier text at the margin' (see Ehrmann and Talmi, 2017).

[24]Amaya and Filbien (2015) calculate the Jaccard similarity (Jaccard, 1901, as used in Tetlock, 2011) for bigrams of the introductory statement and report a high similarity of about 60% for successive conferences.

[25]See for example Jegadeesh and Wu (2017) who apply a clustering algorithm to FOMC minute paragraphs to extract topics or Hansen and McMahon (2016) who work with clustered sentences from FOMC statements.

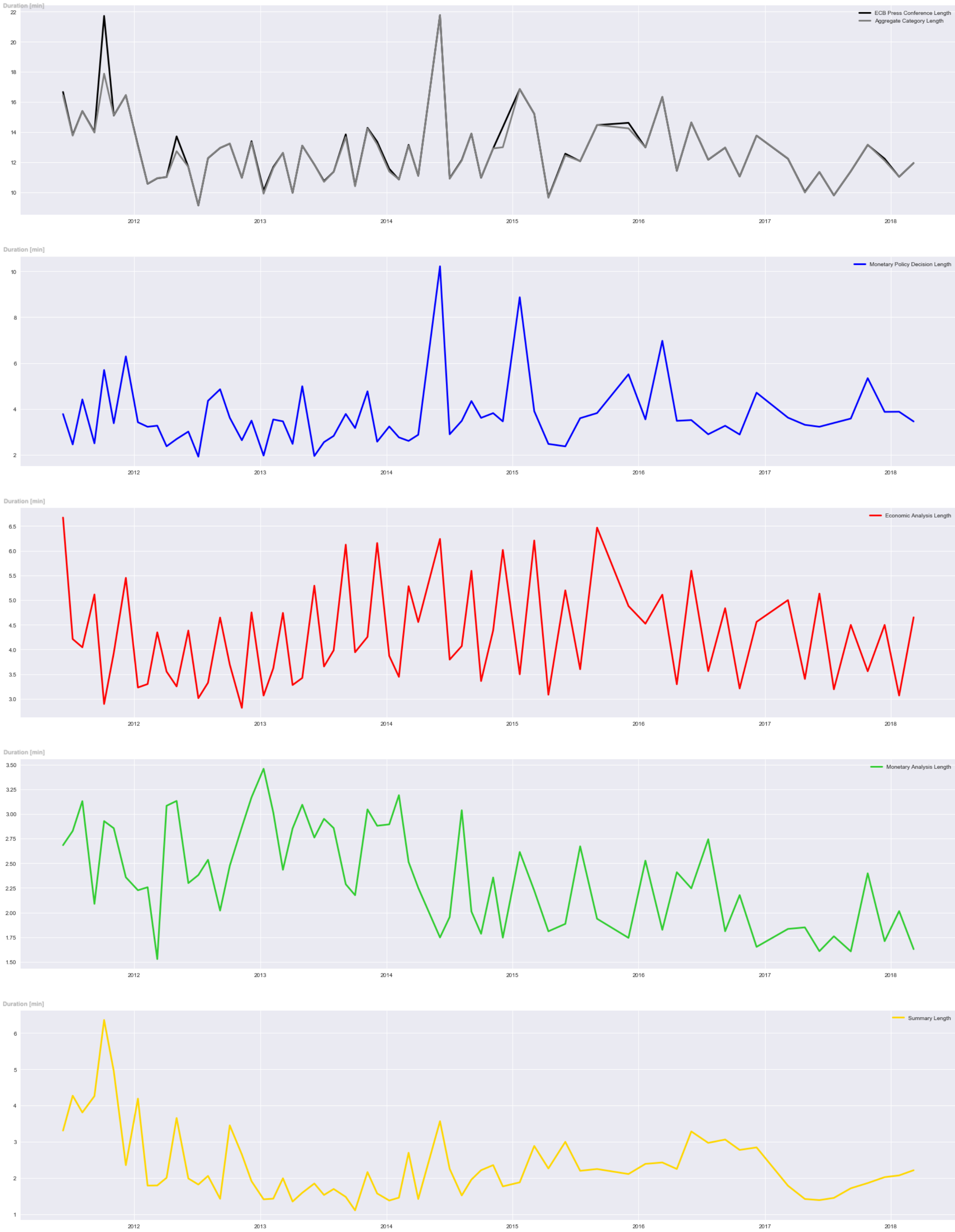


Figure 4.3: **ECB Press Conference Category Length**
This figure shows the time series for the duration of the four categories of the introductory statement as well as the aggregated category length (sum of individual category lengths) and the one obtained from the original ECB webcast videos. The duration is measured in minutes. The sample comprises 66 ECB press conferences from June 2011 to March 2018.

4.4.1 Approach

To facilitate the clustering procedure we incorporate our domain knowledge about the structure of the press conference statements into the algorithm. By studying the statements we can see that each of the different paragraphs of the ECB introductory statement comprises highly characteristic words and phrases that are always used in this section and hardly in any other part of the statement. For example, in the press conference from the 25th of January 2018, as part of the economic analysis the ECB revealed the following information on the economic outlook risks:

> *"The **risks surrounding the euro area growth outlook** are assessed as broadly balanced. On the one hand, the prevailing strong cyclical momentum could lead to further positive growth surprises in the near term. On the other hand, downside risks continue to relate primarily to global factors, including developments in foreign exchange markets."*

The expression 'risks surrounding the euro area growth outlook' is only used in this paragraph and thus highly characteristic for this part.[26] In addition, this expression reflects the overall topic of the paragraph. To incorporate this knowledge into the clustering procedure the algorithm has first to understand that the different words form one expression and then that this phrase is highly informative for the topic of the paragraph.

To group the different words of the expressions together we use the multi-word tokenization (as e.g. in Hansen and McMahon, 2016 or Lopez-Lira, 2019) in the pre-processing of our documents. As a result the seven tokens from our expression above ('risks', 'surrounding', 'the', 'euro', 'area', 'growth', 'outlook') are treated as

[26] Other examples are 'Ladies and gentlemen', 'key ECB interest rates', 'non-standard monetary policy measures' or 'euro area annual HICP inflation'.

one token in the subsequent analysis: 'risks_surrounding_the_euro_area_growth_outlook'.[27]

The entries of the standard term-document matrix contain pure frequency counts of the respective terms and thus all words are of same importance. To increase the weight for the highly informative phrases we are going to use a weighting scheme. The tf-idf weighting scheme, as already used in Loughran and McDonald (2011), puts more weight on terms that occur in only a few documents (paragraphs) compared to words that appear in every part. This works perfectly for our ECB statements since the characteristic phrases only occur in certain paragraphs and thus have a high idf score.[28]

The ECB dedicates in her intro statement each paragraph to a different topic. In the previous example we can see that this part only talks about the risks to the economic outlook. There is no other information revealed in this paragraph. Therefore, we need an algorithm that assigns a 'one-hot encoding' to each of our paragraphs.[29] A very popular clustering algorithm producing such a label is the k-means which is perhaps the most widely used fully automated clustering method (MacQueen, 1967).

To sum-up, to group our different intro statement paragraphs using our domain knowledge we apply the multi-word tokenization in combination with the tf-idf weighting scheme and the k-means algorithm. In the following we will discuss the implementation details.

[27]Note, the single words of this expression are not necessarily characteristic for the paragraph. For example, 'risks' also appears often in the context of inflation risks which is a different topic than risks regarding the economic outlook. Thus, only the full expression is highly informative for the topic of the corresponding paragraph.

[28]Jegadeesh and Wu (2013) also emphasize that 'the appropriate choice of term weighting in content analysis is at least as important as, and perhaps more important than, a complete and accurate compilation of the word list.' In our use case the idf weighting scheme fits perfectly.

[29]One-hot encoding means that the respective document only gets assigned one topic label. Compared to a mixed-topic approach like the LDA algorithm (as used in Hansen and McMahon, 2016) which assigns different topic probabilities to every single word of the document. This would not suit us here.

4.4.2 Pre-Processing and Feature Extraction

In a first step we have to transform the qualitative textual representation of the paragraphs into a quantitative number. Therefore, we apply standard textual analysis methods to split the paragraphs into their respective sentences and words. Then, we remove any tokens not containing letters and apply a stemming algorithm. In a last step we remove stopwords.

In addition to the standard pre-processing steps we use multi-word expressions to group tokens that occur often together and are highly characteristic for the corresponding paragraph. Out of a collection of n-grams with the highest frequency, expressions with the highest likelihood ratio and pointwise mutual information (PMI)[30] we select a parsimonious list of expressions that are very informative about the respective paragraphs and also standard economic phrases like 'inflation_rate' or 'monetary_policy_stance'. A full list can be found in the appendix (Table A.3).

Based on the resulting list of words we create a term document matrix that describes the frequency of terms v that occur in a collection of documents (corpus), D, indexed by d. The documents are our extracted paragraphs from the ECB statements. As discussed before we use 'the standard weighting scheme for term-document matrices in information retrieval, called tf-idf' (Jurafsky and Martin, 2017, p. 278), defined as:

$$tf - idf_{d,v} = tf_{d,v} \times idf_v, \tag{4.1}$$

where the first part is the term frequency (tf), simply the frequency of the word v in the document d. The second inverse document frequency is defined as:

$$idf_v = log(\frac{N}{df_v}), \tag{4.2}$$

where N is the number of documents in the corpus ($|D|$) and df_v the number of documents

[30]As discussed in Manning and Schütze (2003), Chapter 5.

containing the term v. Overall, the tf-idf feature prefers words that appear often in the current document d but are rare in the overall collection.[31]

4.4.3 Clustering of Paragraphs

Using the term document matrix we can now apply the k-means clustering algorithm to assign a single cluster label to every paragraph. The k-means clustering aims to partition N observations into K clusters in which each observation belongs to the cluster with the nearest mean. It also returns the centroids of the K clusters which can be used to obtain topic labels for the training set and also to label new data points (e.g. a new press conference).

Let D_k be the set of all documents that belong to cluster k. The centroid of this cluster k is $\vec{u}_k = \frac{1}{|D_k|} \sum_{d \in D_k} \vec{x}_d$, with $\vec{x}_d \in R_+^V$ being the vector that represents document d. The k-means algorithm chooses the cluster assignments $\{D_1, ..., D_k\}$ to minimize the sum of squares between each document and its cluster centroid:

$$\sum_k \sum_{d \in D_k} ||\vec{x}_d - \vec{u}_k||^2. \tag{4.3}$$

The optimization procedure uses two inputs: First, the $tf-idf$ matrix and second the number of clusters. Since the categories of the intro statement discuss different topics we can apply the algorithm separately to each of the four parts which significantly reduces the dimensionality for the clustering algorithm. Therefore, we first divide the full corpus D into four distinct sets of documents $D_{mon\_pol\_dec}$, D_{econ}, D_{mon} and D_{sum} and then calculate the respective term document matrices $tf-idf_{mon\_pol\_dec}$, $tf-idf_{econ}$, $tf-idf_{mon}$ and $tf-idf_{sum}$ which serve as an input for the k-means optimizations.

To find the second input parameter, the number of clusters, we perform a manual assessment of the ECB introductory statement. As a result we identify the structure as

[31] For example, the highly characteristic expressions 'fixed_rate_tender_procedures' and 'asset_purchase_programme' have one of the highest idf scores with roughly 3.8 and 3.4.

shown in Figure 4.4: After the (1) introduction by the ECB Chairman the monetary policy decision follows. Specifically, the (2) change in the key ECB interest rate gets disclosed which is accompanied by further explanations. Then, especially during the height and aftermath of the euro crisis non-standard monetary policy measures got announced. Here we can distinguish between the ECB acting as the (3) 'lender' or (4) 'buyer of last resort' (see Acharya et al., 2018). These decisions can also be accompanied by (5) further explanations and details, e.g. about the practical implementation of these measures. Then, the information about the real economy is revealed, particularly about (6) realized GDP and the economic environment, the (7) quarterly macroeconomic staff projections and also the (8) risks to the economic outlook. Following a similar structure the information about (9) realized inflation, (10) the inflation projections and the (11) inflation outlook are published. After the economic analysis the monetary assessment follows. Specifically, information about the (12) money supply, loans and the (13) banking sector. The press conference concludes with a (14) cross-check of the economic and monetary analysis and with (15) final remarks or comments about fiscal policy and structural reforms in the euro area.[32] Based on this structure we want to identify 15 (5+6+2+2) topics which serves as our starting value for the number of clusters in the optimization procedure.[33]

In general, our derived topics follow previous studies: For example, Berger et al. (2011) divide the ECB's communicated monetary policy stance into 'price stability', 'real economy' and 'monetary' related indicators. Jansen and De Haan (2005) use the categories 'M3', 'Economic growth', 'Inflation' and 'Interest rates' to analyze statements made by ECB officials and Ehrmann and Fratzscher (2009) differentiate between statements on 'economic outlook', 'inflation', 'money growth' and 'interest rates' when analyzing news

[32]Note, the information about the risks to the price developments were dropped during our sample period and are currently not included in the introductory statement. Also the paragraph about the banking sector was only included temporarily during the height of the European sovereign debt crisis.

[33]Since not every of our defined groups is identified immediately we increase the number of clusters sequentially until we have obtained the structure as shown in Figure 4.4. If one group consists of several (by the algorithm identified) clusters we subsume them within the corresponding topic. For example, separating the macro projections GDP and inflation section is quite difficult since the wording of both paragraphs is very similar so that the algorithm separates other groups first which are based on two or three different templates (e.g. the risks to the economic outlook). Therefore, in total we need 24 clusters to identify the 15 topics.

CATEGORY	META-TOPIC	TOPIC	CLUSTER	TOP WORDS
Monetary Policy Decision	Standard Monetary Policy	Introduction	0 6	‘vice-president’, ‘welcome’, ‘ladies_and_gentlemen’, ‘pleased’, ‘press_conference’ ‘his’, ‘meeting’, ‘excellent’, ‘special’, ‘gratitude’
		Interest Rate Decision and Explanation	1 3	‘unchanged’, ‘key_ecb_interest_rates’, ‘guidance’, ‘forward’, ‘decided’ ‘inflation’, ‘economic’, ‘remains’, ‘euro_area’, ‘prices’
	Non-Standard Monetary Policy	Refinancing Operations	2	‘rate’, ‘refinancing_operations’, ‘conducting’, ‘mros’, ‘fixed_rate_tender_procedures’, ‘three-month’, ‘operations’, ‘ltros’
		Asset Purchase Programs	5	‘purchases’, ‘billion’, ‘inflation’, ‘intended’, ‘case’, ‘any’, ‘end’, ‘monthly’
		Further Details and Comments	4	‘market’, ‘purchases’, ‘measures’, ‘liquidity’, ‘monetary_policy’, ‘banks’, ‘euro_area’, ‘start’
Economic Analysis	Real Economy	Euro Area GDP	8 15	‘quarter’, ‘euro_area’, ‘demand’, ‘sector’, ‘economic’ ‘quarter’, ‘economic’, ‘supported’, ‘reforms’, ‘continued’
		GDP Projections	7	‘staff_macroeconomic_projections’, ‘real_gdp’, ‘revised’, ‘compared’, ‘eurosystem’, ‘ecb’, ‘foresee’, ‘range’
		Economic Outlook Risks	12 14 16	‘risks’, ‘downside’, ‘surrounding’, ‘hand’, ‘growth_outlook’ ‘risks’, ‘relate’, ‘downside’, ‘euro_area’, ‘tensions’ ‘risks’, ‘downside’, ‘insufficient’, ‘countries’, ‘including’
	Inflation	Euro Area Inflation	9	‘prices’, ‘inflation_rates’, ‘inflation’, ‘months’, ‘euro_area_annual_hicp_inflation’, ‘was’, ‘currently’, ‘energy’
		Inflation Projections	11	‘staff_macroeconomic_projections’, ‘hicp’, ‘inflation’, ‘projection’, ‘comparison’, ‘ecb’, ‘eurosystem’, ‘foresee’
		Price Development Risks	10 13	‘risks’, ‘upside’, ‘taxes’, ‘indirect’, ‘outlook_for_price_developments’ ‘monitor’, ‘context’, ‘developments’, ‘close’, ‘exchange’
Monetary Analysis	Money, Credit and Banking Sector	Money and Loans	18 19 20	‘loans’, ‘adjusted’, ‘non-financial’, ‘sales’, ‘securitisation’ ‘m3’, ‘growth’, ‘annual’, ‘monetary’, ‘rate’ ‘loans’, ‘since’, ‘borrowers’, ‘place’, ‘measures’
		Banking Sector	17 21	‘bank’, ‘countries’, ‘strengthened’, ‘euro_area’, ‘transmission’ ‘bank’, ‘key’, ‘provision’, ‘banks_balance_sheets’, ‘soundness’
Summary	Summary and Final Remarks	Cross-Check	23	‘analysis’, ‘monetary’, ‘prices’, ‘confirms’, ‘cross-check’, ‘to_sum_up’, ‘signals’, ‘picture’
		Fiscal Policy and Structural Reforms	22	‘growth’, ‘euro_area’, ‘fiscal’, ‘countries’, ‘implementation’, ‘structural_reforms’, ‘market’, ‘competiveness’

Figure 4.4: **ECB Press Conference Structure and Cluster Allocation**
This figure shows the structure and our cluster allocation for the ECB press conference introductory statements. For determining the clusters we use a k-means algorithm.

snaps from ECB press conferences. In addition to these broader categories we also explicitly exploit the paragraph structure of the ECB intro statement to derive very fine granular topics. To the best of our knowledge our study is the first who makes use of that implicit structure.

4.4.4 Visualization

For a graphical illustration and inspection of the paragraphs and our clustering results we first determine the similarity of each paragraph d_i with all the other paragraphs $d_j,\ j \in [1, ...N]$. Based on the $tf - idf$ matrix for the full corpus D we calculate the cosine similarity:

$$cos\_sim_{i,j} = cos(\theta) = \frac{\vec{d_i} \cdot \vec{d_j}}{||d_i||\ ||d_j||}. \tag{4.4}$$

We then define the respective paragraph distance as:

$$dist_{i,j} = 1 - cos\_sim_{i,j}. \tag{4.5}$$

The result is a N x N matrix containing in each row the distances $dist_{i,j}$ of each paragraph i to all of the other ones $j \in [1, ...N]$. We use the multidimensional scaling algorithm (MDS)[34] to reduce the dimension of this matrix to two which allows us to visually inspect the results. The plot is shown in Figure 4.5. Similar paragraphs are located closer to each other on the two dimensional scale compared to very different sections. For example, the welcome section with the phrase ‘Ladies and gentlemen, ...’ (depicted in purple) has usually a quite similar wording and thus the points lie close together.[35]

[34]The general aim of the MDS algorithm is ‘to find a configuration of points in a space [...] where each point represents one of the objects or individuals, and the distances between pairs of points in the configuration match as well as possible the original dissimilarities between the pairs of objects or individuals’ (Cox and Cox, 2008). In general, the MDS algorithm is similar to the PCA analysis but uses a distance matrix as an input instead of the covariance matrix.

[35]In general, we also used the graphical visualization to verify the clustering results. Figure 4.5 clearly shows the similarity of paragraphs within the same group. To further validate the cluster and topics, we read through the individual cluster elements and compared the documents with each other and also with the top words of the cluster (see Grimmer and Stewart, 2013). Overall, we find a very good fit for our clustering results. An extensive discussion of the clustering results can be found in the appendix.

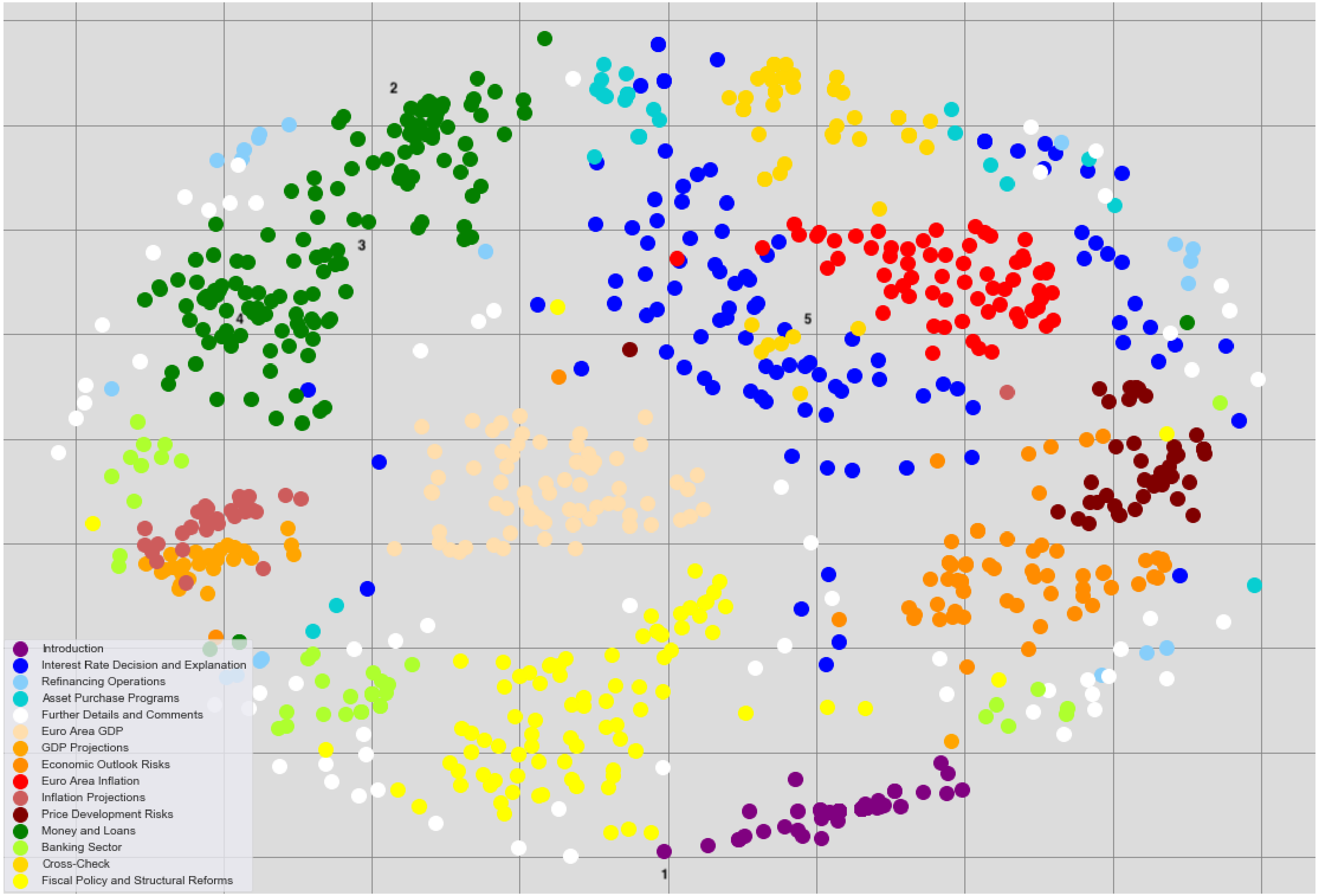


Figure 4.5: **Clustered ECB Introductory Statement Paragraphs**
This figure shows the clustered paragraphs of the ECB's introductory statements. The sample period comprises 66 press conferences from June 2011 to March 2018. (1) marks the introduction of Draghi's first press conference. (2)/ (4) indicates the 'pure' money supply/ loan topic and (3) the paragraphs comprising information of both topics, money supply and loan growth. (5) indicates a 'Cross-Check' paragraph which repeats most the information of the interest rate decision part.

4.5 Financial Market Data

To analyze the impact of the different ECB topics on the stock market we first use actual Euro Stoxx 50 index futures trades obtained from the EUREX exchange. For the analysis we use the most liquid futures which usually coincides with the contract with the nearest maturity.[36] As a robustness exercise we also use tick-by-tick (15 seconds) index values of the Euro Stoxx 50. The data is available for all of our press conference days from June 2011 until the end of September 2017.

To study the cross-section we use the 15 seconds index values for 16 Euro Stoxx supersector indices: banking, automobile, basic resources, insurance, oil and gas, consumer products, industry goods, financial services, telecommunication, travel and leisure, utilities, personal and household goods, food and beverages, media, retail and health care.[37]

For our cross-country analysis we include the following Stoxx country indices: Stoxx UK 50, Stoxx France 50, Stoxx Spain 20, Stoxx Italy 20, Stoxx Eastern Europe 50, the Stoxx Nordic 30 and the Stoxx Sub-Balkan 30. Each of these country indices tracks the value of the respective regions' blue-chips companies. The data is available since June 2011 until the end of December 2014. All of the tick-by-tick index values are provided by Deutsche Börse.

For our fixed income analysis we use the German Schatz, Bobl and Bund futures contracts. To study the impact on European government yields we also include the Italian BTP Long and the French OAT futures. The data is available from June 2011 until the end of September 2011 for the German and the Italian futures and since 16th of April 2012 for the French contract. The trade prices are obtained from the EUREX

[36] Only a few days before the expiration date the new contract has a higher trading volume than the old one when investors roll over to the new futures.

[37] Note, we do not have data for the three remaining Euro Stoxx supersector indices: chemicals, real estate and technology.

exchange and we select the most liquid contracts as stated above.

We then join the futures trades with the created timestamps for the press conference content based on the first trade after a given timestamp. The tick-by-tick values are joined with the rounded timestamps (on 15 second basis). Based on the prices/ index values for the start and end times we then calculate log returns for the respective categories and paragraphs.

4.6 Empirical Analysis

Using our previously created timestamps and the clustered topics we can now study which information revealed by the European Central Bank during its press conference has the largest impact on financial markets and thus seems most important to market participants.[38]

4.6.1 ECB Press Conference Returns

Compared to the pre-FOMC announcement drift as documented by Lucca and Moench (2015) the pre-ECB announcement return mean-reverts after the start of the press conference (see Schmeling and Wagner, 2019 and Chapter 2 of this thesis).

Table 4.2 shows that until the end of the trading day the magnitude of the average mean reversion return is -37 basis points which accounts for 70% of the 53 bps pre-announcement increase as reported in Chapter 2. By splitting the time window into three distinct parts we can see that the biggest drop falls within the time window from 2:30 p.m. to 2:45 p.m. (with a p-value of 9%) when the introductory statement is presented by the ECB Chairman. This return on its own accounts for roughly 40% of the

[38] In general, our analysis is similar to Ehrmann and Fratzscher (2009) who studied how the ECB's explanations of their monetary policy decision affect the 3-month Euribor futures market. Based on taped versions of the press conferences they quantify the aggregate impact of the introductory statement on absolute Euribor futures returns and using Reuters news snaps they also provide a minute-by-minute analysis. Compared to their approach our methodology allows us to study the real-time impact of very fine granular topics based on the automatically created timestamps and the textual analysis of the ECB statement.

overall mean-reversion. The point estimates for the Q&A session and the remaining hours after the press conference are over 20% lower in magnitude and exhibit higher p-values.

Table 4.2: **Post-ECB Announcement Returns - Euro Stoxx 50**
This table reports the results from the post-ECB announcement dummy regressions using intra-day Euro Stoxx 50 returns. We use the following model specification: $r_t = \beta_0 + \beta_1 \times \mathcal{I}_t(\text{ECB}) + \epsilon_t$, where r is the percentage value of the derived log return using the index values of the respective start and end time. The constant β_0 measures the unconditional return for the given time window that is earned on days where the ECB does not hold a scheduled monetary policy meeting. The term $\mathcal{I}_t(\text{ECB})$ is an indicator function with a value of one if time t is a day where the ECB scheduled a monetary policy meeting and zero otherwise ('ECB Dummy'). We split the trading hours into (i) the time window of the ECB introductory statement (14:30 -14:45 CET), (ii) the time of the ECB press conference Q&A session from 14:45 to 15:30 CET, (iii) the remaining trading hours after the press conference from 15:30 to 17:30 CET and (iv) the full time window: 14:30 - 17:30 CET. The sample period is from June 2011 to September 2017. Robust t-statistics are shown in parenthesis. *, ** indicate significance at the 5% or 1% level.

	2011 - 2017			
	Euro Stoxx 50 Returns			
	14:30-14:45	**14:45-15:30**	**15:30-17:30**	**14:30-17:30**
Constant	-0.004	0.003	-0.01	-0.011
	(-0.91)	(0.57)	(-0.78)	(-0.73)
ECB Dummy	-0.144	-0.112	-0.111	-0.367*
	(-1.7)	(-1.57)	(-1.0)	(-2.32)
Obs.	1721	1721	1721	1721
N. of ECB	62	62	62	62

The upper part of Figure 4.6 illustrates this empirical finding showing that the steepest drop in the cumulative ECB announcement return series falls within the presentation of the introductory statement. This finding is in line with Ehrmann and Fratzscher (2009) who emphasize the importance of the press conferences as a tool for explaining monetary policy decisions.

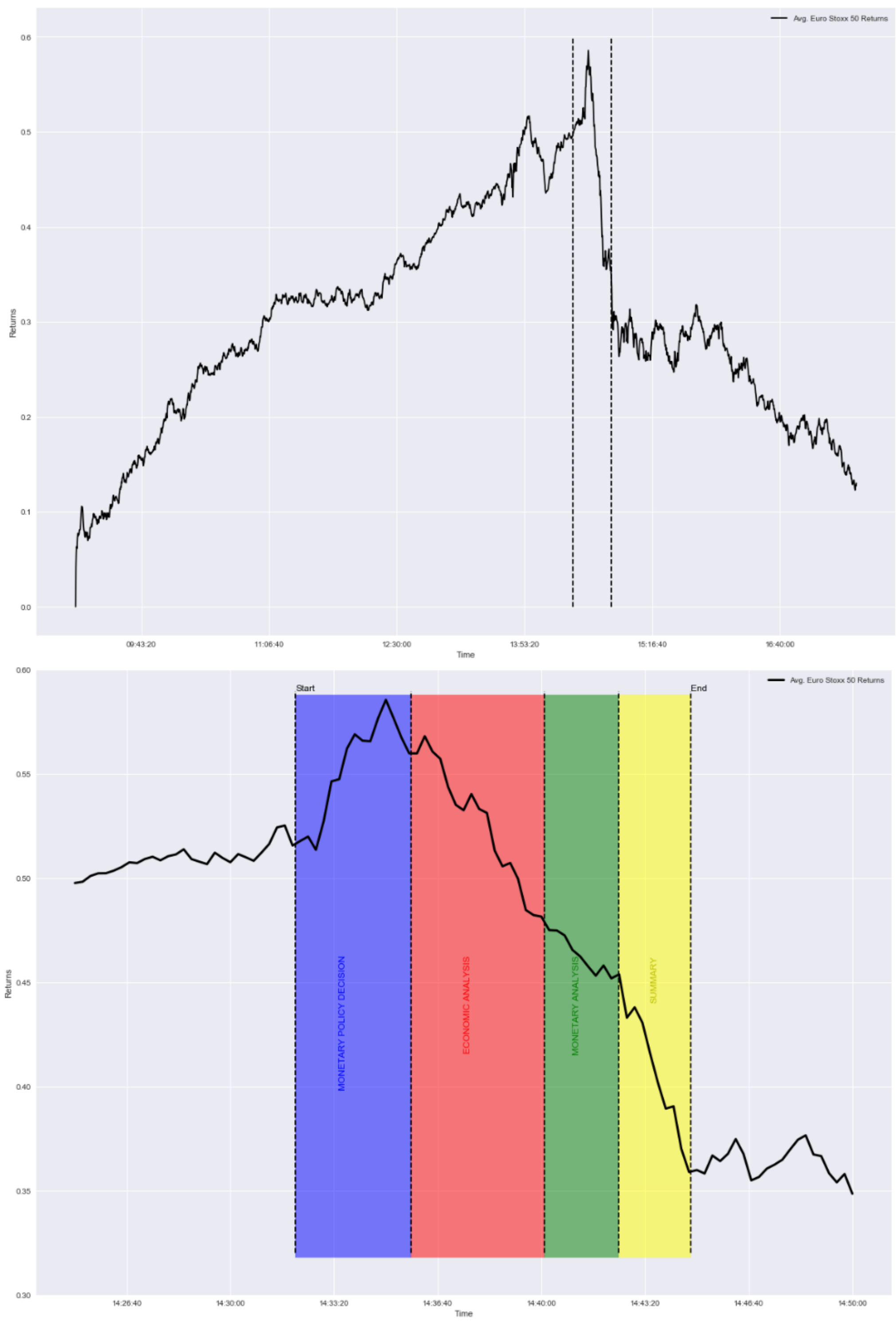


Figure 4.6: **ECB Announcement Returns**
The upper plot shows the average returns on full ECB announcement days. The black line shows the mean of the cumulative return series from 09:00 to 17:30 CET. The dashed vertical lines mark the time window of the lower plot (14:25 - 14:50 CET) spanning the introductory statement of the press conference. The lower plot shows the average cumulative return series around the introductory statement from 14:25 - 14:50 CET. The average starting (end) time of the introductory statement is marked with 'Start' ('End'). The blue/ red/ green/ yellow shaded areas mark the average duration of the respective category ('Monetary Policy Decision', 'Economic Analysis', 'Monetary Analysis' and 'Summary'). The sample comprises 62 press conferences from June 2011 to September 2017.

4.6.2 ECB Press Conference Category Returns

After the observation that the largest part of the mean-reversion occurs in the 14:30 to 14:45 CET time window we now want to analyze which information is responsible for this drop in the stock market. Therefore, we are going to decompose the overall introductory statement return into the contribution of each ECB press conference category using our previously obtained timestamps. Table 4.3 shows the respective ECB press conference category returns. One can see that the pre-ECB announcement return as reported in Chapter 2 first further increases by around 20% (10 basis points) before the mean-reversion starts. The beta estimate on the monetary policy decision dummy variable is positive and significant to the 5% level. After this first initial increase the Euro Stoxx 50 return experiences a large drop of in total 24 basis points where over 50% (-13 bps) is realized during the economic analysis. This point estimate shows significance to the 1% level. The drop continues during the monetary analysis but only to a moderate extent (-4 bps). At the end of the introductory statement the return decreases by another 7 bps making the intra press conference price changes the largest contribution to the overall mean reversion dynamic.[39] The lower plot of Figure 4.6 summarizes the findings by showing the time window of the cumulative ECB announcement return series which covers the introductory statement.

Overall, the monetary policy decision and the economic analysis account for the largest movements in the financial market. This finding is in line with Cieslak and Schrimpf (2018) who find that for the ECB growth and monetary shocks explain most of the variance of stock returns. In contrast, the monetary analysis exhibits the lowest returns of all categories suggesting that its information are of less importance to the financial markets. This is consistent with Berger et al. (2011) who conclude that 'developments in the monetary sector [...] only played a minor role most of the time'. Given our sample period from 2011 to 2017 the on average negative return during the economic analysis might be driven by negative news related to the European sovereign debt crisis. The positive effect of the monetary policy decisions might reflect the supporting actions of

[39] Adding up the individual category returns closely resembles the 14 basis point estimate for the 14:30 to 14:45 time window of Table 4.2.

the European Central Bank. To find evidence for these hypotheses we are now going to look at the time series of the category returns and relate it to important events of the euro crisis and decisions of the ECB.

Table 4.3: **ECB Introductory Statement Returns - Euro Stoxx 50**
This table reports the results from three dummy regressions assessing the impact of the different content of the ECB's introductory statement on financial markets. The regression model is defined as follows: $r_t = \beta_0 + \sum_{j=1}^{J} \beta_j \times \mathcal{I}_t(\text{ECB Content}^{(j)}) + \epsilon_t$, with β_0 being set to zero and j $\in \{1, ..., J\}$ with J being the number of different categories, meta-topics and topics as defined previously. $\mathcal{I}_t(\text{ECB Content}^{(j)})$ is an indicator function with a value of one if the section of the introductory statement belongs to a certain category, meta-topic or topic as described in the table and zero otherwise. r_t represent the log returns of the different sections calculated using futures prices around the respective start and end time. Specification (1) reports the β estimates for the four categories, (2) for the six meta-topics and (3) for the 15 individual topics. The sample period comprises 62 ECB press conferences from June 2011 to September 2017. Robust t-statistics are shown in parenthesis. *, ** indicate significance at the 5% or 1% level.

	Euro Stoxx 50 Index Future		
ECB Press Conference - Introductory Statement	**(1)**	**(2)**	**(3)**
Category: Monetary Policy Decision	0.101* (2.20)		
Standard Monetary Policy		0.005 (0.15)	
Introduction			0.009 (0.73)
Interest Rate Decision and Explanation			-0.003 (-0.11)
Non-Standard Monetary Policy		0.122** (3.05)	
Refinancing Operations			0.087* (2.05)
Asset Purchase Programs			0.044 (1.51)
Further Details and Comments			0.054* (2.24)

Category: Economic Analysis	-0.126**		
	(-3.41)		
Real Economy		-0.067*	
		(-2.21)	
Euro Area GDP			-0.049
			(-1.86)
GDP Projections			-0.033
			(-1.63)
Economic Outlook Risks			-0.002
			(-0.13)
Inflation		-0.059**	
		(-3.69)	
Euro Area Inflation			-0.049**
			(-3.53)
Inflation Projections			-0.008
			(-0.41)
Price Development Risks			-0.009
			(-1.01)
Category: Monetary Analysis	-0.039		
	(-1.66)		
Money, Credit and Banking Sector		-0.039	
		(-1.66)	
Money and Loans			-0.012
			(-1.29)
Banking Sector			-0.021*
			(-2.06)
Category: Summary	-0.074**		
	(-2.83)		
Summary and Final Remarks		-0.074**	
		(-3.08)	
Cross-Check			-0.018
			(-1.73)
Fiscal Policy and Structural Reforms			-0.043**
			(-2.76)

Figure 4.7 shows the decomposition of the overall introductory statement return series into the respective categories. As a robustness check we also plot the sum of the individ-

ual category returns and the 14:30 to 14:45 and 14:30 to 15:00 CET return series.[40] One can see that the aggregate return tracks the other two time series very closely, especially the 14:30 to 14:45 returns with a correlation of 95%.

In general, the time series of the full introductory statement returns shows a lot of variation with large negative returns for example during summer 2012 where the chances of a eurozone break-up were quite severe. Interestingly, there is also a large negative drop in December 2015 where the ECB Chairman announced further accommodating non-standard monetary policy measures which should have had a positive impact on financial markets. The negative reaction indicates that the market expected the ECB to announce more supportive measures and thus was disappointed by the announcement. Support for this interpretation can also be drawn from the first question of the Q&A session where a reporter shared his observation: '...it seems like what you've done is a little bit on the low end of the range of what the financial markets had expected, in terms of your stimulus package today. It seems like the initial reaction in the financial markets bears this point...'. This observation is the only large negative return in the first category.

Overall, the return series for the monetary policy decision category shows a lot of positive returns which correspond to the non-standard monetary policy measures announced by the European Central Bank. If there are no supportive measures announced (or no decrease in the key ECB interest rate) then returns are roughly zero. This can be seen particularly for the last year of our sample period where no further supporting measures were announced due to the ongoing recovery of the euro area economy.

The negative point estimate for the economic analysis stems from a lot returns being below zero as illustrated in Figure 4.7. Especially, in the first half of our sample we can observe large negative category returns reflecting the adverse economic conditions during the height of the euro crisis. For example, the large drop in August 2011 corresponds to a period which the ECB characterized as a time when 'uncertainty [was] particularly high'. The last year of our sample period shows mostly zero or slightly positive returns for this category reflecting the ongoing economic recovery in the eurozone.

[40] Note, some press conferences take a bit longer than 15 minutes.

The monetary analysis returns do not show a lot of variation in their time series with most of the returns being around zero. This suggests that information for example about money growth are not as important for market participants as news about the economic conditions or the monetary policy decision.

The returns of the last category are mostly close to zero with some larger negative observations. These often correspond to press conferences where the ECB Chairman urged the EU countries to implement more labour market reforms and to consolidate their fiscal policies which emphasized the structural problems underlying the European sovereign debt crisis.

In a next step we are going to analyze the impact of the respective category topics on the financial markets to establish a closer link between the press conference content and the accompanying stock market reaction.

4.6.3 ECB Press Conference Topic Returns

Table 4.3 shows the contribution of each of the press conference topics to the respective category return. In addition to the individual topics we also aggregate similar paragraphs into meta-topics which share a common theme.[41]

Column two of Table 4.3 reveals that the large positive monetary policy decision return is entirely driven by the non-standard monetary policy measures. The average response in the equity market is 12 basis points with large t-statistics (above 3). In contrast, the return for the standard monetary policy is close to zero and not statistically significant. Looking at the respective topics we see that both estimates for the introduction and the interest rate decision are not statistically different from zero.[42] Assuming efficient markets this observation comes not as a surprise since the interest rate decision is already revealed at 13:45 CET in a press release and thus markets should already have

[41] For example the three paragraphs about 'Euro Area Inflation', 'Inflation Projections' and 'Price Development Risks' are all about inflation. The meta-topic timestamps are then derived based on the start (end) time of the first (last) paragraph and the returns are calculated based on these timestamps.

[42] Note, since the introduction and the interest rate decision are included in every introductory statement the aggregate estimate of 0.005 for the standard monetary policy can be obtained by adding up the two individual returns (the slight difference is due to rounding errors).

Figure 4.7: **Time Series of ECB Category Returns**
This plot shows the return time series for the four categories of the ECB's introductory statement as well as the time series based on the sum of the individual category returns and the 14:30 to 14:45 (15:00) CET returns. The sample comprises 62 ECB press conferences from June 2011 to September 2017.

incorporated this information with the start of the press conference.[43]

The provision of excess liquidity through the ECB's refinancing operations (MROs, LTROs and TLTROs) was perceived very positively by the financial markets with a point estimate of 0.09 which is significant to the 5% level. Especially, during the high uncertainty period in Summer and November 2011 the announcement of these operations yielded a very high return (e.g. roughly 0.8% in August 2011 or 0.4% in October 2011) which is consistent with the finding of Acharya et al. (2018) that the LTRO programs calmed down the markets temporarily. Also the announcement of the targeted longer-term refinancing operations (TLTROs) in January 2015 had a large positive impact on Euro Stoxx 50 returns (about 0.6%).

The estimate for the asset purchase programs (APP) is positive (4.4 bps) but surprisingly not significant to the 1% or 5% level. This is puzzling at first since the announcement of the extensive purchases by the ECB should have a positive effect on the economy and expected dividends and thus on stock prices. Looking into the cluster composition we can see that this group comprises mainly the asset purchase program as announced in January 2015. After large positive returns following the announcement in January and the expansion to €80 billion in March 2016 the returns fluctuate around zero.[44] We attribute this to the fact that starting March 2016 the APP information are already included in the press release which gets published at 13:45 CET. Therefore, we cannot report a dominant QE-related factor as in Altavilla et al. (2019) but rather a medium-sized response.

In contrast, additional information about the implementation details are not included in the press release and thus represent actual news. Therefore, the 'Further Details and Comments' about the non-standard monetary policy measures exhibit a large positive (above 5 bps) and significant estimate. This is consistent with Coenen et al. (2017) who state that the contextual release of implementation details for the ECB's non-standard

[43]In July 2013 the ECB newly introduced its forward guidance on the future monetary policy intentions. Therefore, after mid of 2013 one can observe a few larger positive returns (e.g. 0.39% in July 2013) when the ECB expected the key interest rates 'to remain at present or lower levels for an extended period of time'.

[44]The mean return of this cluster is 0.07% until March 2016, from April 2016 to September 2017 it is only -0.02%.

monetary policy measures has further lowered market uncertainty and thus was perceived as good news by investors and market participants.

The large negative return for the economic analysis is driven to a similar extent by information about the real economy (-7 bps) and inflation (-6 bps). In particular, news about the 'Euro Area GDP' yield a negative estimate (-5 bps) reflecting the adverse economic conditions during the European sovereign debt crisis. The estimate is slightly not significant to the 5% level which might be due to the positive economic recovery in the later part of our sample period. The estimate for the macro staff GDP projections is about one third smaller and has a large p-value indicating that the ECB's forecasts do not contain new information for market participants. The information about the economic outlook is also not statistically different from zero which comes as a surprise since one would have expected an impact of the ECB's superior outlook information on financial markets, especially during the euro crisis. To shed light on this finding we re-run our regression model with dummy variables for each of the individual clusters which compose the different topics (as stated in Figure 4.4). The results are shown in Table 4.4 and one can see that there are two opposite effects that cancel out. First, a positive impact of cluster 12 (5 bps) and then a negative one from cluster 14 and 16 with a magnitude of around -4 basis points. When looking into the respective cluster elements we observe that cluster 14 and 16 contain only economic outlook paragraphs from the first half of our sample. This period covers the height of the euro crisis and thus one would expect the negative estimate. This is also reflected in the keywords as can be seen in Figure 4.4: 'euro_area', 'tensions' and 'insufficient'. On the other side, cluster 12 contains the paragraphs after the height of the crisis when the European economy started to recover. The top words do not include the expressions 'euro_area' or 'tensions' anymore but the more optimistic phrase 'growth_outlook'. A similar observation can be made for the 'Euro Area GDP' topic which comprises of two groups: Cluster 8 has a negative significant impact of -8 bps and cluster 15 a point estimate close to zero. The former one contains the paragraphs starting June 2011 (until the end of 2014) which includes the high uncertainty periods of the euro crisis. The other cluster starts in January 2015 and thus spans the recovery period of the euro economy. This asymmetric response to positive and negative news within the same topic is also found in previous studies (see e.g. Jansen and De Haan, 2005) and is also in line with Cieslak and Schrimpf (2018) who find that economic news mattered especially in the period until 2013. Overall, these observations are very interesting since they show that the wording and phrases used by the central bank differ significantly in periods of economic crisis and recovery which can be detected by our text-based clustering approach.

Table 4.4: **ECB Economic Analysis Returns - Euro Stoxx 50**
This table reports the results from two dummy regressions assessing the impact of the different topics and clusters of the ECB's introductory statement on financial markets. The regression model is defined as follows: $r_t = \beta_0 + \sum_{j=1}^{J} \beta_j \times \mathcal{I}_t(\text{ECB Content}^{(j)}) + \epsilon_t$, with β_0 being set to zero and j $\in \{1, ..., J\}$ with J being the number of different topics and clusters as defined previously. $\mathcal{I}_t(\text{ECB Content}^{(j)})$ is an indicator function with a value of one if the section of the introductory statement belongs to a certain topic or cluster as described in the table and zero otherwise. r_t represents the log return of the different sections calculated using future prices around the respective start and end time. Specification (1) reports the β estimates for the three real economy topics and (2) for the five clusters composing these topics. For the sake of brevity we only report here the estimates for the real economy section. The sample period comprises 62 ECB press conferences from June 2011 to September 2017. Robust t-statistics are shown in parenthesis. *, ** indicate significance at the 5% or 1% level.

	Euro Stoxx 50 Index Future	
ECB Press Conference - Economic Analysis	**(1)**	**(2)**
Real Economy		
Euro Area GDP	-0.049 (-1.86)	
CLUSTER 8 ('Crisis')		-0.076* (-2.19)
CLUSTER 15 ('Post-Crisis')		0.009 (0.25)
GDP Projections	-0.033 (-1.63)	

CLUSTER 7		-0.033 (-1.63)
Economic Outlook Risks	-0.002 (-0.13)	
CLUSTER 12 ('Post-Crisis')		0.050 (1.56)
CLUSTER 14+16 ('Crisis')		-0.035* (-2.26)

The inflation news about the euro area has a large negative effect on Euro Stoxx 50 returns. The dummy regression coefficent is around -0.05 [%] and significant to the 1% level which is consistent with the finding of Ehrmann and Fratzscher (2009) that inflation statements have a large impact on financial markets during the press conference. Especially in the beginning of our sample in 2011 one could observe high inflation rates (e.g. in August 2011 the reported HICP inflation rate for July was 2.5%) which were accompanied by large negative returns (-0.4%). The estimates for the other two inflation related topics are close to zero and not significant on a stand alone basis.

Looking at the constituents of the monetary analysis we can see that the point estimate for the money and loans section (which covers e.g. the broad monetary aggregate M3) is negative but with a low magnitude and not statistically significant to the 5% or 1% level suggesting a second-order importance of these information to investors. This finding is in line with Berger et al. (2011) who state that 'money has become less important in the ECB monetary strategy'. The estimate of the banking related information is around -2 bps and slightly significant to the 5% level reflecting the importance of news about the banking sector especially during the European debt crisis.

The last category comprises the cross-check of the economic information with the monetary signals and comments from the ECB Chairman on the implementation of structural reforms and fiscal policies in the EU member states. The estimate for the cross-check is negative but not statistically different from zero reflecting the low information content of this section.[45] The concluding comments in the last paragraph often stress the need

[45] It mostly repeats information from the earlier part of the statement.

for structural reforms and fiscal consolidation in the eurozone and have on average a negative effect on the stock market (-4.3 bps). The estimate is also significant to the 1% level suggesting that information emphasizing the structural problems of the euro area and its economy lead to downward pressure on equity prices.

Overall, we can see that not every piece of information that is revealed during the press conference is of same importance to investors. For our sample period we find the largest asset price responses to news concerning the European sovereign debt crisis and also stronger inflation dynamics. In general, our derived real-time response measure provides a natural way to determine the information content of an information piece, in particular its news component.[46] Other studies explicitly measure the ex-ante market expectations to derive a news shock based on the difference between the new information and the expectation (see e.g. Rosa, 2011). Our study presents an alternative methodology to measure directly (ex-post) the corresponding news or information shock.

U.S. Jobless Claims

Every Thursday at 14:30 CET the United States Department of Labor publishes the new figures on the U.S. jobless claims. Since this news can affect our measured ECB communication impact we supplement our regression specification to control for this news. We follow Ehrmann and Fratzscher (2009) who use the (absolute) value of the surprise component of the release, calculated as the difference between the announced figures and the market expectations as measured by the median response to a Bloomberg survey. We can report that neither the actual nor the absolute value of the surprise component affect our topic estimates as stated in Table 4.3. Both estimated coefficients are very close to zero and not statistically significant (t-values < 0.85).

4.6.4 Volatility and Trading Activity

In the following analysis we are going to study the effect of the different information revealed during the press conference on market volatility and trading activity. Therefore,

[46] Assuming efficient markets asset prices should only react to the news component of an information.

we follow Ehrmann and Fratzscher (2009) and use absolute returns to study the volatility and as a measure of trading activity the number of ticks recorded within a given minute and the traded volume of the Euro Stoxx 50 futures contracts.[47]

Table 4.5 shows the estimates for the different topics ordered by the average absolute returns. One can observe that the top three topics with the largest average absolute returns are all directly related to the European sovereign debt crisis. Each of the topics has an average market impact of about 14 basis points. The two non-standard monetary policy paragraphs also exhibit the by far highest trading activity which can be seen in the high number of ticks per minute (200 and 300) and also the large number of contracts traded (10,000 and 16,000). These findings underline the importance of these information as already reported in our previous analysis.

The lower end of Table 4.5 comprises topics with little or no information content. For example the introduction does not contain any investment relevant information and thus has the lowest absolute return on average. In addition, the number of ticks is by far the lowest with only a fifth compared to the refinancing operations. The information content of the cross-check is also small compared to the other topics. As mentioned previously this section mostly repeats sentences from the earlier part of the statement and thus does not contain new information. The risks to price developments topic is also located in the lower part of this table indicating a minor importance of this section. Maybe that is the reason why it is dropped from the ECB's intro statement in the later half of our sample period.

[47]As an alternative measure we also use squared returns as a proxy for market volatility.

Table 4.5: **ECB Statement Topic Volatility and Trading Activity - Euro Stoxx 50**
This table reports the average absolute and squared returns of the 15 topics of the ECB introductory statement. The paragraph returns are calculated using the Euro Stoxx 50 futures trades around the respective paragraph start and end times. The number of ticks and the futures trading volume are first calculated for the full topic length and then scaled to a minute basis. The sample period comprises 62 ECB press conferences from June 2011 to Sept 2017.

	Euro Stoxx 50 Index Futures			
ECB Press Conference - Introductory Statement	**Absolute Return**	**Squared Return**	**Number of Ticks** (per minute)	**Trading Volume** (per minute)
Refinancing Operations	0.143	0.054	304	15941
Further Details and Comments	0.139	0.046	218	9558
Euro Area GDP	0.137	0.047	150	5979
Interest Rate Decision and Explanation	0.125	0.039	138	6239
Fiscal Policy and Structural Reforms	0.091	0.022	104	4013
Asset Purchase Programs	0.090	0.019	90	2363
Euro Area Inflation	0.087	0.014	131	5273
Inflation Projections	0.081	0.009	123	4516
Economic Outlook Risks	0.075	0.015	135	5673
GDP Projections	0.074	0.012	140	5525
Money and Loans	0.066	0.010	118	4314
Banking Sector	0.051	0.005	88	3842
Cross-Check	0.050	0.007	112	4273
Price Development Risks	0.045	0.004	116	5137
Introduction	0.043	0.009	60	1916

4.6.5 Cross-Section: Industries and Countries

Euro Stoxx Industries

To assess the impact of the different press conference content on the cross-section of industries we estimate our previous dummy regression model for each of our Euro Stoxx supersector indices. The results are shown in Table 4.6.[48][49] We can see that the banking and insurance sector have the highest loadings on the non-standard monetary policy topics. Especially the equity prices of the banking industry appreciate when the ECB announces the refinancing operations. The 15 bps increase in returns is by far the largest of all industries. The same holds for the additional details and comments section (+10 bps). These findings are in line with Krishnamurthy et al. (2018) who find that financial sector stocks increase around OMT events. In addition, the two financial sectors show a more negative response to information about the real economy than the overall market index (see Table 4.4). We relate these findings to the fact that the European banks and insurance companies were closely linked to the course of the European sovereign debt crisis through the sovereign-bank nexus (see Acharya et al., 2018). A possible default of one the EU member states would have caused severe losses in the government bond portfolio of these companies. The findings are also in line with Chan-Lau et al. (2015) who disentangle the effect of different factors on equity prices and find that the 'deterioration of the growth outlook' had a major impact on the bank returns. In addition, the EU banking sector also exhibits the strongest loading on the euro area inflation topic.

The lowest average response to the non-standard monetary policy topics can be observed for sectors like 'Travel and Leisure', 'Health Care' or 'Food and Beverages'. They also exhibit the lowest sensitivity to the real economy information. These industries can be characterized as non-cyclical and thus were not as closely connected to the course of the euro crisis as the cyclical or the financial sectors.

[48]For the brave of clarity we only report estimates for the most interesting paragraphs. The results are ordered by the loading on the refinancing operations topic.

[49]Note, since our dataset on the Euro Stoxx supersector indices ends in December 2014 the estimate for the asset purchase program (which starts in January 2015) is zero.

Table 4.6: **ECB Introductory Statement Returns - Industries**
This table reports the results from 16 separate dummy regressions assessing the impact of the different content of the ECB's introductory statement on the cross-section of industries using the Euro Stoxx supersector indices. $r_t^{ind} = \beta_0 + \sum_{j=1}^{J} \beta_j \times \mathcal{I}_t(\text{ECB Content}^{(j)}) + \epsilon_t$, with β_0 being set to zero and j $\in \{1, ..., J\}$ with J being the number of different topics as defined previously. $\mathcal{I}_t(\text{ECB Content}^{(j)})$ is an indicator function with a value of one if the section of the introductory statement belongs to a certain topic as described in the table and zero otherwise. r_t^{ind} represent the log returns of the different sections for the various industries calculated using tick-by-tick index values of the respective start and end time. For the sake of brevity we only report the estimates for a subste of topics. 'Ref Ops' refers to the 'Refinancing 'Operations', 'APP' to the 'Asset Purchase Programs', 'Details & Comm' to the 'Further Details and Comments' section, 'Euro GDP' to 'Euro Area GDP', 'Econ Outl' to the 'Economic Outlook Risks' and 'Euro Infl' to the 'Euro Area Inflation' topic. The sample period comprises 42 ECB press conferences from June 2011 to December 2014. Robust t-statistics shown in parenthesis. *, ** indicate significance at the 5% or 1% level.

	Monetary Policy Decision			*Economic Analysis*		
Industries	**Ref Ops**	**APP**	**Details & Comm**	**Euro GDP** (Cluster 8)	**Econ Outl** (Cluster 14+16)	**Euro Infl**
Banks	0.151**	0.00	0.102**	-0.106*	-0.042	-0.099**
	(2.81)	(0.00)	(2.98)	(-2.15)	(-1.85)	(-4.27)
Insurance	0.106**	0.00	0.057*	-0.080*	-0.043**	-0.058**
	(2.60)	(0.00)	(2.35)	(-2.57)	(-2.60)	(-3.01)
Telecommunications	0.098*	0.00	0.050*	-0.060*	-0.027	-0.055**
	(2.49)	(0.00)	(2.34)	(-2.34)	(-1.89)	(-3.38)
Industrial Goods & Services	0.076*	0.00	0.039*	-0.067**	-0.024*	-0.038**
	(2.30)	(0.00)	(2.07)	(-2.62)	(-2.24)	(-2.69)
Construction & Material	0.073*	0.00	0.053*	-0.074*	-0.028*	-0.056**
	(1.98)	(0.00)	(2.39)	(-2.38)	(-2.00)	(-3.15)
Oil & Gas	0.072*	0.00	0.058*	-0.066*	-0.021	-0.044*
	(2.20)	(0.00)	(2.17)	(-2.21)	(-1.59)	(-2.41)
Utilities	0.070*	0.00	0.059*	-0.073*	-0.039**	-0.049**
	(2.01)	(0.00)	(2.35)	(-2.57)	(-2.59)	(-3.30)
Automobile & Parts	0.066*	0.00	0.054*	-0.087**	-0.033	-0.055**
	(2.17)	(0.00)	(2.04)	(-2.68)	(-1.92)	(-2.87)
Financial Services	0.063**	0.00	0.026	-0.050*	-0.031**	-0.032**

	(2.59)	(0.00)	(1.53)	(-2.24)	(-2.73)	(-2.59)
Basic Resources	0.054	0.00	0.054*	-0.083**	-0.025	-0.043*
	(1.48)	(0.00)	(2.57)	(-3.18)	(-1.79)	(-2.41)
Personal & Household Goods	0.053	0.00	0.032	-0.052**	-0.026**	-0.038**
	(1.72)	(0.00)	(1.82)	(-2.69)	(-2.80)	(-3.00)
Media	0.047	0.00	0.031*	-0.046*	-0.021**	-0.031**
	(1.89)	(0.00)	(2.10)	(-2.37)	(-2.60)	(-2.99)
Travel & Leisure	0.045**	0.00	0.018	-0.041**	-0.009	-0.020*
	(2.90)	(0.00)	(1.62)	(-2.67)	(-1.09)	(-1.93)
Health Care	0.045*	0.00	0.041	-0.048*	-0.017	-0.038**
	(2.17)	(0.00)	(2.06)	(-2.28)	(-1.79)	(-3.55)
Retail	0.040	0.00	0.044*	-0.059**	-0.026**	-0.030*
	(1.35)	(0.00)	(2.45)	(-2.74)	(-2.62)	(-2.53)
Food & Beverages	0.037	0.00	0.036	-0.038*	-0.008	-0.032**
	(1.69)	(0.00)	(2.44)	(-1.96)	(-1.10)	(-2.91)

Euro Stoxx Countries

Table 4.7 shows the results for the STOXX blue-chip indices of different European countries. The GIIPS countries in this sample (Italy and Spain) exhibit the by far highest estimates on the non-standard monetary policy content. The average response of Italian stocks alone to the announcement of the ECB's refinancing operations is 11 basis points which is 25% higher than for the aggregate Euro Stoxx 50 index. Also their loading on the additional details and comments section is significantly higher than for the other countries in our sample. These findings are in line with Krishnamurthy et al. (2018) who report rising stock returns for the announcement of the SMP and OMT programs for all GIIPS countries and also positive reactions to the LTRO programs by Italy and Spain. We can also observe that Italy and Spain show similar point estimates for the economic variables: A negative coefficient of -0.09 [%] for the 'Euro Area GDP' section and around -0.03 [%] for the economic outlook. Similar to our previous observations we interpret these findings as reflecting the course of the European sovereign debt crisis which particularly affected the GIIPS countries.[50]

The estimates for France share the same signs as for Italy and Spain but are lower in magnitude. The remaining countries (UK, Nordic, Eastern Europe and Sub Balkan) show significantly lower estimates compared to the first three reflecting that they have not been at the core of the euro crisis.

Table 4.7: **ECB Introductory Statement Returns - Countries**
This table reports the results from 7 separate dummy regressions assessing the impact of the different content of the ECB's introductory statement on the cross-section of countries using the Stoxx country indices. $r_t^{country} = \beta_0 + \sum_{j=1}^{J} \beta_j \times \mathcal{I}_t(\text{ECB Content}^{(j)}) + \epsilon_t$, with β_0 being set to zero and j $\in \{1, ..., J\}$ with J being the number of different topics as defined previously. $\mathcal{I}_t(\text{ECB Content}^{(j)})$ is an indicator function with a value of one if the section of the introductory statement belongs to a certain topic as described in the table and zero otherwise. $r_t^{country}$ represent the log returns of the different sections for the various countries calculated using tick-by-tick index values of the respective start and end time. For the sake of brevity we only report the estimates for a subste of topics. 'Ref Ops' refers to the 'Refinancing 'Operations', 'APP' to the 'Asset Purchase Programs', 'Details & Comm' to the 'Further Details and Comments' section, 'Euro GDP' to 'Euro Area GDP', 'Econ Outl' to the 'Economic Outlook Risks' and 'Euro Infl' to the 'Euro Area Inflation' topic. The sample period comprises 42 ECB press conferences from June 2011 to December 2014. Robust t-statistics shown in parenthesis. *, ** indicate significance at the 5% or 1% level.

	Monetary Policy Decision			*Economic Analysis*		
Industries	**Ref Ops**	**APP**	**Details & Comm**	**Euro GDP** (Cluster 8)	**Econ Outl** (Cluster 14+16)	**Euro Infl**
Italy	0.108*	0.00	0.081**	-0.092*	-0.024	-0.065**
	(2.42)	(0.00)	(2.81)	(-2.32)	(-1.33)	(-3.61)
Spain	0.088	0.00	0.076**	-0.090*	-0.035*	-0.067**
	(1.95)	(0.00)	(2.83)	(-2.46)	(-2.12)	(-3.29)
France	0.075*	0.00	0.054*	-0.068*	-0.028*	-0.054**
	(2.18)	(0.00)	(2.29)	(-2.32)	(-2.20)	(-3.31)
UK	0.044*	0.00	0.032	-0.048*	-0.000	-0.015
	(2.13)	(0.00)	(1.83)	(-1.96)	(0.03)	(-1.60)

[50]Chouliaras and Grammatikos (2017) report that the stock market indices of the GIIPS countries suffered the most during the euro crisis, showing a lot of extreme negative returns.

Nordic	0.041	0.00	0.042*	-0.047*	-0.018	-0.025*
	(1.26)	(0.00)	(2.50)	(-2.28)	(-1.93)	(-2.13)
Eastern Europe	0.036*	0.00	0.018	-0.017	-0.012	-0.008
	(2.19)	(0.00)	(1.23)	(-0.88)	(-1.29)	(-0.87)
Sub Balkan	-0.009	0.00	0.006	-0.004	0.003	-0.012*
	(-0.28)	(0.00)	(0.80)	(-0.59)	(0.52)	(-2.21)

4.6.6 Other Assets: Fixed Income

European Government Bonds

In Table 4.8 we report the average absolute yield changes for the Italian, French and German 10 year government bond for the respective intro statement topics.[51] One can see that the by far largest impact on the bond market stems from topics directly related to the European debt crisis. Especially information related to the ECB's non-standard monetary policy actions seem to be of high importance with an average yield change of 64 bps for the LTRO programs and 60 bps for the additional details and comments for the Italian bond. This finding is in line with Krishnamurthy et al. (2018) who report falling Italian yields due to the announcement of the SMP, OMT[52] and LTRO programs.[53] The estimates for the German Bund futures are over 10 to 20 bps lower for the first topics.

In addition to the central bank's supporting actions also fundamental information about the European economy seems important, especially for the Italian bond yields (over 50 bps). Again the loadings for the French and German bond are much lower. We relate these findings to the different sensitivity of the European countries to the course of the euro crisis.[54] These observations are in line with Leombroni et al. (2017) who

[51]Since we do not find evidence for a pre- or post-announcement drift in the fixed income market and not observe significant estimates for the raw bond yield changes we omit the respective table here.

[52]Note, the SMP and OMT programs are included in our 'Further Details and Comments' topic.

[53]The average Italian yield change for the non-standard monetary policy measures using our real-time methodology is negative and large in magnitude (-0.3%).

[54]Note, our data on the French government bonds starts only in April 2012 and thus does not include the first uncertainty peaks of the euro crisis. This could also be a reason for the lower estimates.

report that around the European sovereign debt crisis the ECB's communication drove a wedge between yields of core and peripheral European countries. Overall, these results are comparable to the country's equity indices as reported earlier where Italy showed the highest loadings on the refinancing operations and also the economic topics. The lowest impact seem to have the introduction and also the paragraph about the price development risks. As explained earlier these topics do have little or no information content and thus should not be a major driver of financial markets.

Table 4.8: **ECB Introductory Statement and Government Bond Yields**
This table reports the average absolute yield changes of European government bond futures for the 15 topics of the ECB introductory statement. The paragraph returns are calculated using actual trades of the German Schatz, Bobl, Bund, the Italian BTP Long and the French OAT futures contracts around the respective paragraph start and end times. To convert the futures returns into yield changes we divide them by the negative of duration as in Cieslak and Schrimpf (2018). The sample period comprises 62 ECB press conferences from June 2011 to Sept 2017. Note, for the French government bond futures the data is only available since April 2012.

	ITALY	*FRANCE*	*GERMANY*		
ECB Press Conference - Introductory Statement	**BTP Long** (10 years)	**OAT** (10 years)	**Bund** (10 years)	**Bobl** (5 years)	**Schatz** (2 years)
Refinancing Operations	0.64	0.39	0.52	0.14	0.03
Further Details and Comments	0.60	0.36	0.36	0.11	0.01
Euro Area GDP	0.53	0.27	0.30	0.09	0.01
Interest Rate Decision and Explanation	0.38	0.30	0.34	0.12	0.02
GDP Projections	0.35	0.14	0.23	0.12	0.01
Fiscal Policy and Structural Reforms	0.30	0.23	0.21	0.07	0.01
Euro Area Inflation	0.26	0.20	0.22	0.08	0.01
Economic Outlook Risks	0.24	0.12	0.15	0.06	0.01

Money and Loans	0.23	0.18	0.25	0.06	0.01
Inflation Projections	0.20	0.18	0.27	0.11	0.01
Asset Purchase Programs	0.16	0.14	0.13	0.03	0.00
Cross-Check	0.16	0.10	0.11	0.04	0.01
Price Development Risks	0.15	0.10	0.13	0.04	0.01
Banking Sector	0.13	0.11	0.12	0.04	0.01
Introduction	0.13	0.06	0.09	0.03	0.00

German Yield Curve

Comparing the responses across the German yield curve one can see that the short-term Schatz futures contract only marginally reacts to the new information revealed in the introductory statement. This finding is consistent with Leombroni et al. (2017) who report that their communication shock hardly affects the short-end of the yield curve. Compared to the 2 year bond the 5 year Bobl futures exhibit significantly higher loadings. The highest average absolute yield change of 14 bps can be observed for the refinancing operations topic. In general, the loadings of the 5 year bond are by factor 2 to 3 lower compared to the 10 year Bund futures. This observation is in line with Brand et al. (2010) who report that the ECB's press conferences have a sizeable impact especially on longer-term yields.

4.7 Conclusion

In this chapter we present a novel methodology to study the information flow of communication events. We use modern textual analysis tools and the information contained in the subtitles of press conference webcasts to fully automatically create timestamps for the different information revealed. These timestamps can then be used to study the real-time impact on financial markets and thus the information content of the respective news element.

We then apply this methodology to the press conferences of the European Central Bank and analyze which information of the introductory statement ist most important to market participants. Exploiting the template structure of the statements using clustering techniques we derive topics on the category and paragraph level which allow for a very fine granular analysis on which information is driving asset prices.

Studying the impact on the equity market we can relate the mean-reversion of the pre-ECB announcement drift as documented in Schmeling and Wagner (2019) and Chapter 2 of this thesis to the information revealed in the press conference. We find that first the positive pre-announcement return increases by another 20% driven entirely by the non-standard monetary policy measures of the ECB before equity prices start to drop by roughly half the size of the pre-ECB run-up. The economic analysis alone accounts for over 50% of this fall, with a roughly equal contribution of information about the real economy and inflation news. We relate the former to news about the European sovereign debt crisis and its subsequent recovery. Comments about the European banking sector and the structural and fiscal problems of some EU member states further significantly decrease the Euro Stoxx 50 returns.

Studying the cross-section of European industries and countries we find that the positive impact of the ECB's unconventional monetary policy measures and the negative loading on the economic information is particular pronounced for the banking and insurance sector as well as the GIIPS countries of our sample. These industries and countries have been at the core of the euro crisis (also connected via the bank-sovereign nexus) and thus very sensitive to the announced actions of the central bank as well as its superior economic (outlook) information. Quantifying the impact on the fixed income markets further underlines the importance of the announcement of the non-standard monetary policy measures as well as the disclosure of the economic information. These effects are particularly pronounced for the long-end of the yield curve and here specifically for the Italian government bond.

Chapter 5

Summary and Outlook

In this thesis, we study announcement returns around ECB monetary policy decision days. In particular, we focus on the time period from 2010 onwards which represents a new era of increased importance for the European Central Bank due to the outbreak of the European sovereign debt crisis. In general, since the start of the euro crisis one can observe a decoupling of the ECB's monetary policy decisions from the course of the U.S. Federal Reserve, the global leader among all central banks.

In Chapter 2 we document a statistically significant and economically large pre-ECB announcement return for this new era of increased importance (2010 to 2015), whereas the respective premium for the time period from 2000 to 2009 is statistically speaking not different from zero. This finding updates the current notion of the literature that high returns on days surrounding monetary policy decisions are only present for meetings of the Federal Reserve. We further show that heightened uncertainty is a main driver of the 24h pre-ECB return which is consistent with the literature that analyzes the determinants of the pre-FOMC announcement drift. Compared to these studies we connect the general equity market uncertainty to the course of the European sovereign debt crisis, in particular we show that the pre-ECB drift is mainly driven by the periods of high uncertainty of the euro crisis which can be characterized as times of severe eurozone break-up risks (e.g. Grexit). During these times the importance of the ECB was even more exaggerated due to its role as the lender or buyer of last resort. Last, we document that the pre-ECB announcement return, as opposed to the pre-FOMC drift, starts to mean-revert after the announcement with the drop in equity prices occurring predominantly during the ECB's press conference.

To study the impact of the news that gets revealed during the press conference on financial markets we use modern textual analysis tools. We lay the foundations for how these techniques are applied in finance in Chapter 3. In particular, we illustrate the main challenges inherent in this kind of analysis, present typical research designs and discuss the most important textual analysis concepts.

Using these text mining methods we then analyze in Chapter 4 which information disclosed by the ECB Chairman is driving the mean-reversion of the pre-ECB return. We develop a novel methodology that uses the captions of the press conference webcasts and modern textual analysis tools to automatically create timestamps for each part of the ECB's introductory statement. In addition, we deploy a clustering algorithm to exploit the template-like structure of the central bank statement and to derive fine granular central bank topics. Using the created topic timestamps we then quantify the real-time impact of the different information on financial markets. We find that the largest impact stems from (positive) news about the ECB's unconventional monetary policy measures and also (negative) information revealed during the economic analysis, in particular about the course of the European sovereign debt crisis and inflation dynamics.

Our findings are important for risk and asset management processes since the here documented pre-ECB announcement return represents an easy and practicable investment strategy with a high Sharpe ratio of 1.5. To realize this premium an investor only needs to buy an Euro Stoxx 50 futures contract (in a very liquid market with low transaction costs) 24 hours before the ECB's monetary policy meeting and sell it right before the announcement. Additionally, one can also trade on the following mean-reversion of the announcement return by establishing a short position (short selling the ES 50 futures).

The insights on how (new) information are perceived by the financial markets are of high importance to central banks who nowadays use communication as an additional tool to their classical monetary policy action ('setting the interest rates'), which is particularly important in times of interest rates being at the zero lower bound or during a financial crisis. Therefore, understanding the effect of the different content and its wording on financial markets is crucial for central bank members. The here developed methodology provides an easy and intuitive measure for quantifying the real-time market reaction!

In general, our novel methodology can be applied to all kind of communication events that are accompanied by a press conference, like other central bank meetings or also annual shareholders' meetings. In addition to analyzing a statement's paragraphs, one could use our methodology on a sentence level, which would allow to find out the exact (key-)words that drive the asset price response (using for example the regression approach of Jegadeesh and Wu 2013). Based on the regression estimates, one could then also create a specific central bank dictionary that comprises of all the significant terms. The sign of the estimate would then indicate whether it is a positive or negative word, also the magnitude could be used as a weight for the corresponding term. Such a central bank dictionary would improve the sentiment analysis of central bank statements, since 'general' dictionaries tend to misclassify terms in the monetary policy context. Also it would be of interest to examine how our real-time asset price responses for the equity and bond market relate to measurements of the monetary policy and information shock (and their respective decomposition) using principal components (see Altavilla et al., 2019; Leombroni et al., 2017) or obtained through exploiting the comovement of stocks and bonds (see Cieslak and Schrimpf, 2018; Jarocinski and Karadi, 2018). Last, it would be interesting to observe whether the ECB announcement returns will also be present in the future or whether they will dissapear like the pre-FOMC announcement drift (see Gilbert et al., 2018).

Appendix A

Additional Information on ECB Real-Time Analysis

A.1 Discussion of Clustering Results

The press conference usually starts with the introduction where the ECB Chairman welcomes the press audience with the phrase ‘Ladies and gentlemen, the Vice-President and I...’. In Figure 4.5 these paragraphs are shown in purple. All of the dots lie close to each other indicating a high similarity between the respective introductions. Paragraphs that are located at the boundary of this topic include for example the presentation of a new commissioner or Draghi’s personal remarks in his first ECB press conference on 3rd of November 2011 which is marked with a ‘1’ in the figure. The top words (e.g. ‘vice-president’, ‘welcome’ and ‘ladies_and_gentlemen’) accurately reflect the specific words and phrases used in the introductory paragraph. The second cluster groups the introductions where the ECB Chairman expresses his ‘special gratitude’ to the organizers of the meeting (e.g. another central bank).

After the introduction by the ECB Chairman the standard monetary policy decision gets revealed, in particular whether there was a change in the key ECB interest rates.[1]

[1]Since July 2013 the ECB also includes a forward guidance on the expected path of future interest rates in their statement which is usually in the same paragraph as the announcement of the interest rate

In addition to the interest rate decision the statement also provides the rationale for the decision, in particular how it aligns with the ECB's monetary policy mandate of 'maintaining inflation rates below, but close, to 2% over the medium-term'. In the more recent press conferences this discussion is presented in a separate section but especially in the beginning of our sample period the statement provides a joint paragraph on the interest rate decision and the accompanying explanations. Therefore, we use a combined topic which represents the interest rate decision and its rationale.

The first cluster with the keywords 'key_ecb_interest_rates', 'forward', 'guidance' and 'decided' includes the paragraphs from the later part of our sample about the pure interest rate decision and the forward guidance. The term 'unchanged' reflects the situation of interest rates being at the zero lower bound which could not be lowered any further within the course of the euro crisis. The second cluster comprises the rationale for the monetary policy action as well as the paragraphs from the earlier part of our sample that cover both the interest rate decision and its explanation. These remarks reconcile the monetary policy actions with the current economic development and its impact on the inflation aim. The main theme of this section centers around 'inflation' which is the most important term of this cluster centroid. Also the terms 'prices' and 'euro_area' underline the focus on price developments within the EU countries. In general, these explanations comprise of highly standardized sentences which allows the clustering algorithm to easily group them together. Overall, the respective top words nicely summarize the content of this topic which contains the interest rate decision and the accompanying explanation focusing on the inflation objective of the ECB.

To support the European economy during the euro crisis the ECB conducted main and longer-term refinancing operations (MROs and LTROs) which provided additional liquidity to the banking sector in the spirit of a 'lender of last resort' (see Acharya et al., 2018). The top words of this cluster reflect on the one hand the words and phrases announcing the decision like 'conducting', 'operations' or 'refinancing_operations'. On

decision.

the other hand the specific measures themselves like 'mros' or 'ltros' are represented in the list of words characterizing the cluster centroid as well as the implementation details like 'three-month' or 'fixed_rate_tender_procedures'.

In addition to the liquidity injection measures the ECB decided in January 2015 to launch an expanded asset purchase program (APP) in the range of €60 billion per month buying public and private sector securities 'encompassing the existing purchase programs for asset-backed securities and covered bonds'.[2,3] The key words of this cluster centroid reflect the wording used in this paragraph. Like for the refinancing operations, the name of the measure ('purchases') is highly characteristic for this topics as well as the implementation details like the frequency ('monthly'), time horizon ('end') or the volume ('billion') of the program. Also the expression 'in any case' contributes to the list of keywords which is often used in this section to underline the determination of the ECB to improve economic conditions.

The last cluster of the first category collects all the remaining paragraphs of the monetary policy decision that are not yet assigned to the previous topics.This includes mainly comments related to the European sovereign debt crisis, like the provision of liquidity to the banking sector, high risk premia, collateral availability, reserve ratio, or the EFSF/ ESM.[4] Compared to the highly standardized inflation rationale for the monetary policy actions these comments are a direct response to the course of the euro crisis and thus are quite diverse in nature. We label this group 'Further Details and Comments' about the non-standard monetary policy of the ECB and the euro crisis.
This can also be seen in the cluster's keywords which on the one hand include phrases like 'monetary_policy' and 'measures' reflecting the non-standard monetary policy measures

[2]See the introductory statement of the press conference held on 22nd of January 2015.

[3]In March 2016 the ECB first extended the volume of the APP to €80 billion and then lowered it again to €60 billion per month in March 2017. In October '17 the monthly volume got adjusted to €30 billion and then further reduced on 14th of June 2018 when the ECB decided to first lower the asset purchases to €15 billion until the end of December 2018 and then end the net purchases.

[4]EFSF stands for the European Financial Stability Facility, ESM for the European Stability Mechanism.

and on the other hand comprise of terms characterizing the problems underlying the euro crisis, like the 'liquidity situation of euro area banks'. Overall, due to the disperse nature of the different paragraphs (e.g. provision of liquidity vs. collateral vs. reserve ratio) the cluster exhibits the highest variance which can be seen in Figure 4.5 with the white dots distributed across the two dimensional plane.

The three paragraphs about the real economy are clustered very accurately as can be seen in the respective top words in Figure 4.4. For example, the most important words for the 'Euro Area GDP' topic are 'quarter', 'euro_area' or 'demand' which reflect the reported quarterly real GDP growth and also the characterization of the economic environment, e.g. with respect to the the domestic demand. The top words for the 'GDP Projections' are 'staff_macroeconomic_projections' and 'real_gdp' which are highly characteristic for this paragraph. The paragraph about the risks to the economic outlook can be characterized by words like 'risk', 'growth_outlook' and also 'downside' or 'tensions' reflecting the negative sentiment during the euro crisis.
Also the three inflation related topics are clustered very accurately as indicated by the top words. E.g. for the 'Euro Area Inflation' paragraph the expression 'euro_area_annual_hicp_inflation' is highly characteristic since it is only used in this section (other parts of the statement just use the word 'inflation'). As for the 'GDP Projections' the phrase 'staff_macroeconomic_projections' is also important for the inflation forecast topic. The two other keywords 'hicp' and 'inflation' show that the cluster actually relates to the inflation and not the GDP forecasts. Keywords for the risks to the price development section are e.g. 'risks' or the phrase 'outlook_for_price_developments'. Overall, the clustering procedure works perfectly for the economic analysis topics since the paragraphs get only slightly adjusted for the different press conferences and exhibit a very similar wording across the statements. This is also illustrated in Figure 4.5 which shows the six economic analysis cluster well centered around their mean with a low variance. The smallest variation exhibit the two macroeconomic staff projection groups where only the forecast target (GDP or inflation) and the corresponding predictions

(e.g. 2%) are changed from statement to statement.

The two monetary analysis topics are also identified accurately by the k-means algorithm due to the characteristic keywords and expressions used in the respective paragraphs. The money and credit section first publishes the broad (M3) and narrow (M1) money growth figures and then discusses the growth of loans to non-financial corporations and households and the borrowing conditions in general.[5] The most important keywords for the three sub-clusters are 'loans' or 'm3'. Also the expression 'turning_to_the_monetary_analysis' is highly characteristic for the cluster composition. Figure 4.5 depicts the 'pure' money supply (loan) topic in the upper (lower) part of the darkgreen cluster (indicated by number two and four). The paragraphs comprising both sub topics are located in between (number three).
The section about the banking sector was added during the height of the euro crisis which is reflected in its content. The paragraphs cover the 'soundness of banks balance sheets' or talk about strengthening the resilience of banks and the funding situation in general. These topics are also reflected in the top words of the cluster: 'bank', 'strengthened', 'soundness' or 'banks_balance_sheets'. Depending on the course of the euro crisis the wording of the ECB changed from requesting banks to 'retain earnings' to 'substantial progress has been made' to 'finalising the assessment' which can be seen in the distinct sub clusters in Figure 4.5. After the height and aftermath of the crisis this section got dropped.

The last category also comprises of two topics. First, the 'Cross-Check' of the economic analysis with the signals from the monetary analysis.[6][7] Often this part repeats information from the interest rate explanations which then yields in a high similarity between these paragraphs (see number five in Figure 4.5). The keywords accurately de-

[5]Since the information about money and loan growth is often presented in a joined paragraph, especially in the earlier part of the sample, we group them into one topic.

[6]Generally, the focus of the economic analysis is on shorter-term price movements. The monetary analysis focuses on longer-term price trends (see Issing, 2003).

[7]In the first year this section was rather large, in the latter part of the sample this paragraph only comprises of one or two sentences.

scribe this cluster (e.g. ‘cross-check’, ‘to_sum_up’).

The second topic contains comments from the ECB Chairman with respect to the fiscal policy and structural reforms of the European countries. These comments also originated from the European debt crisis where the fiscal discipline of some EU member states and the implementation of structural reforms were heavily discussed. The characteristic terms of this cluster like ‘fiscal’, ‘structural reforms’ or ‘implementation’ support this label.

A.2 Sensitivity Analysis

To test the sensitivity of our results to a change in the created timestamps we use the tick-by-tick (15 seconds) Euro Stoxx 50 index values instead of the acutal futures trades. We join the index values with the rounded timestamps (on 15 second basis). Table A.1 shows the results.

Overall, the average category returns are nearly identical. For example, the estimate for the monetary policy decision is now 0.101 compared to 0.101 when using the actual trades. The point estimate for the economic analysis is now -0.124 (vs. -0.126), for the monetary analysis -0.039 (vs. -0.039) and for the last category -0.091 (vs. 0.074). In addition, the significance levels do not change.

The coefficients for the (meta-) topics are also very similar when using futures trades and the 15 seconds tick-by-tick index values. The estimate for the standard monetary policy decision slightly increases from 0.005 (based on futures trades) to 0.014 using the 15 seconds index values. When looking at the topics one can see that the increase stems mainly from a slightly higher coefficient for the interest rate decision and explanation part (+0.005). In contrast, we observe slightly lower estimates for the refinancing operations and APP topic. Since the standard monetary policy explanations typically precede the non-standard measures we can see here a small shift of the asset price response from the former to the later topic due to the use of the rounded timestamps. A similar small shift can also be observed for the economic analysis. Here, the real economy topic slightly increases by 0.007 whereas the inflation section decreases by a similar amount. In the last part of the statement one can see also a shift from the banking related topic to the cross-check. Here the former increases by 0.008 and the later decreases by the exactly the same magnitude. This illustrates that a change in the timestamps affects especially short paragraphs which only consist of one or two sentences, like the cross-check. But in general one can see that the results for the longer paragraphs and also the meta-topics and category level are very robust to a change in timestamps. Our main results do not change when using the 15 seconds index values instead of the actual futures trades!

Table A.1: **ECB Introductory Statement Returns - Euro Stoxx 50 Index Values** This table reports the results from three dummy regressions assessing the impact of the different content of the ECB's introductory statement on financial markets. The regression model is defined as follows: $r_t = \beta_0 + \sum_{j=1}^{J} \beta_j \times \mathcal{I}_t(\text{ECB Content}^{(j)}) + \epsilon_t$, with β_0 being set to zero and j $\in \{1, ..., J\}$ with J being the number of different categories, meta-topics and topics as defined previously. $\mathcal{I}_t(\text{ECB Content}^{(j)})$ is an indicator function with a value of one if the section of the introductory statement belongs to a certain category, meta-topic or topic as described in the table and zero otherwise. r_t represent the log returns of the different sections calculated using tick-by-tick (15 seconds) index values for the respective rounded start and end time (to 15 seconds). Specification (1) reports the β estimates for the four categories, (2) for the six meta-topics and (3) for the 15 individual topics. The sample period comprises 62 ECB press conferences from June 2011 to September 2017. Robust t-statistics are shown in parenthesis. *, ** indicate significance at the 5% or 1% level.

	Euro Stoxx 50 Index Values		
ECB Press Conference - Introductory Statement	**(1)**	**(2)**	**(3)**
Category: Monetary Policy Decision	0.101* (2.26)		
Standard Monetary Policy		0.014 (0.47)	
Introduction			0.011 (1.64)
Interest Rate Decision and Explanation			0.002 (0.11)
Non-Standard Monetary Policy		0.109** (2.82)	
Refinancing Operations			0.066

			(1.70)
Asset Purchase Programs			0.027 (0.89)
Further Details and Comments			0.057* (2.54)
Category: Economic Analysis	-0.124** (-3.35)		
Real Economy		-0.074* (-2.49)	
Euro Area GDP			-0.051* (-2.03)
GDP Projections			-0.037* (-2.04)
Economic Outlook Risks			-0.006 (-0.54)
Inflation		-0.050** (-3.11)	
Euro Area Inflation			-0.038** (-3.10)
Inflation Projections			-0.002 (-0.09)
Price Development Risks			-0.016 (-1.80)
Category: Monetary Analysis	-0.039 (-1.68)		
Money, Credit and Banking Sector		-0.039 (-1.70)	
Money and Loans			-0.015 (-1.51)
Banking Sector			-0.013 (-1.24)
Category: Summary	-0.091** (-2.96)		
Summary and Final Remarks		-0.091** (-3.10)	
Cross-Check			-0.026* (-2.53)
Fiscal Policy and Structural Reforms			-0.049** (-2.58)

A.3 Additional Figures and Tables

Table A.2: **Scheduled ECB Press Conferences: Dates and Start Times**
This table reports the start times for the press conferences following the ECB's monetary policy decisions. The term 'excluded' denotes meetings that are discarded from our sample. The time of the announcements are obtained from Bloomberg news timestamps which indicate the start of the speech. We report the dates and start times from January 2011 to April 2018.

2011-01-13	*excluded*	2013-01-10	14:32:05	2015-01-22	14:37:14	2018-01-25	14:31:34
2011-02-03	*excluded*	2013-02-07	14:38:30	2015-03-05	14:29:59	2018-03-08	14:30:12
2011-03-03	*excluded*	2013-03-07	14:30:36	2015-04-15	14:35:12	2018-04-26	*excluded*
2011-04-07	*excluded*	2013-04-04	14:34:32	2015-06-03	14:30:10		
2011-05-05	*excluded*	2013-05-02	14:33:12	2015-07-16	14:32:23		
2011-06-09	14:33:40	2013-06-06	14:32:07	2015-09-03	14:31:36		
2011-07-07	14:33:50	2013-07-04	14:34:00	2015-10-22	*excluded*		
2011-08-04	14:30:43	2013-08-01	14:29:57	2015-12-03	14:30:38		
2011-09-08	14:30:51	2013-09-05	14:31:44				
2011-10-06	14:37:42	2013-10-02	14:34:11	2016-01-21	14:30:42		
2011-11-03	14:33:11	2013-11-07	14:31:29	2016-03-10	14:34:44		
2011-12-08	14:31:14	2013-12-05	14:35:41	2016-04-21	14:33:48		
				2016-06-02	14:25:17		
2012-01-12	14:30:22	2014-01-09	14:31:14	2016-07-21	14:29:25		
2012-02-09	14:31:23	2014-02-06	14:31:21	2016-09-08	14:30:30		
2012-03-08	14:35:00	2014-03-06	14:32:53	2016-10-20	14:30:10		
2012-04-04	14:32:44	2014-04-03	14:32:21	2016-12-08	14:28:58		
2012-05-03	14:31:43	2014-05-08	*excluded*				
2012-06-06	14:30:35	2014-06-05	14:30:08	2017-01-19	*excluded*		
2012-07-05	14:29:53	2014-07-03	14:30:53	2017-03-09	14:33:25		
2012-08-02	14:31:16	2014-08-07	14:30:06	2017-04-27	14:32:38		
2012-09-06	14:28:45	2014-09-04	14:29:48	2017-06-08	14:30:05		
2012-10-04	14:32:05	2014-10-02	14:36:21	2017-07-20	14:33:55		
2012-11-08	14:32:35	2014-11-06	14:31:59	2017-09-07	14:31:55		
2012-12-06	14:32:55	2014-12-04	14:33:09	2017-10-26	14:32:34		
				2017-12-14	14:32:24		

Table A.3: **Multi-Word Expressions for ECB Press Conference Transcripts**
This table reports a list of the multi-word expressions used in the pre-processing of the press conference statements ordered by category.

Category	**Multi-Word Expressions**
Monetary Policy	’ladies_and_gentlemen’, ’press_conference’, ’report_on_the_outcome’
Decision	’regular_economic_and_monetary_analyses’, ’key_ecb_interest_rates’
	’refinancing_operations’, ’fixed_rate_tender_procedures’, ’full_allotment’
	‘net_asset_purchases’, ‘asset_purchase_programme’, ’purchase_programme’
	’non-standard_monetary_policy_measures’
Economic Analysis	’explain_our_assessment’, ’economic_activity’, ’real_gdp’
	’economic_outlook’, ’growth_outlook’
	’euro_area_annual_hicp_inflation’
	’outlook_for_price_developments’
	’staff_macroeconomic_projections, staff_projections’
Monetary Analysis	’turning_to_the_monetary_analysis’, ’bank_lending_survey’
	’banks_balance_sheets’
Summary	’to_sum_up’
	’fiscal_policy’, ’structural_reforms’
General Expressions:	‘governing_council’, ’interest_rates’, ’euro_area’,
	’monetary_policy_stance’, ‘monetary_policy’, ’inflation_rate’,
	’inflationary_pressure’, ’inflation_expectations’

Bibliography

Acharya, V., D. Pierret, and S. Steffen. 2018. Lender of Last Resort, Buyer of Last Resort, and a Fear of Fire Sales in the Sovereign Bond Market. *Swiss Finance Institute Research Paper Series No. 18-35.*

Ai, H., and R. Bansal. 2018. Risk Preferences and the Macroeconomic Announcement Premium. *Econometrica 86*, pp. 1383–1430.

Altavilla, C., L. Brugnolini, R. S. Gürkaynak, R. Motto, and G. Ragusa. 2019. Measuring euro area monetary policy. *ECB Working Paper Series.*

Amaya, D., and J.-Y. Filbien. 2015. The similarity of ECB's communication. *Finance Research Letters 13*, pp. 234–242.

Antweiler, W., and M. Z. Frank. 2004. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance 59*, pp. 1259–1294.

Ardia, D., K. Bluteau, and K. Boudt. 2019. Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. *International Journal of Forecasting 35*, pp. 1370–1386.

Azar, P. D., and A. W. Lo. 2016. The Wisdom of Twitter Crowds: Predicting Stock Market Reactions to FOMC Meetings via Twitter Feeds. *Journal of Portfolio Management 42*, pp. 123–134.

Baker, S. R., N. Bloom, and S. J. Davis. 2016. Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics 131*, pp. 1593–1636.

Baker, S. R., N. Bloom, S. J. Davis, and K. J. Kost. 2019. Policy News and Stock Market Volatility. *NBER Working Paper.*

Bank of England. 2015. One Bank Research Agenda. *Discussion Paper.*

Barber, B. M., and T. Odean. 2008. All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. *The Review of Financial Studies 21*, pp. 785–818.

Bekaert, G., M. Hoerova, and M. L. Duca. 2013. Risk, uncertainty and monetary policy. *Journal of Monetary Economics 60*, pp. 771–788.

Ben Dor, A., and C. Rosa. 2019. The Pre-FOMC Announcement Drift: An Empirical Analysis. *The Journal of Fixed Income 28*, pp. 60–72.

Berger, H., J. De Haan, and J.-E. Sturm. 2011. Does Money Matter in the ECB Strategy? New Evidence based on ECB Communication. *International Journal of Finance and Economics 16*, pp. 16–31.

Bernanke, B. S., and K. N. Kuttner. 2005. What Explains the Stock Market's Reaction to Federal Reserve Policy? *The Journal of Finance 60*, pp. 1221–1257.

Bernile, G., J. Hu, and Y. Tang. 2016. Can information be locked up? Informed trading ahead of macro-news announcements. *Journal of Financial Economics 121*, pp. 496–520.

Bholat, D., S. Hansen, P. Santos, and C. Schonhardt-Bailey. 2015. Text mining for central banks. *Centre for Central Banking Studies Handbook - No. 33*, pp. 1–19.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning.* Springer.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, pp. 993–1022.

Boguth, O., V. Grégoire, and C. Martineau. 2018. Shaping Expectations and Coordinating Attention: The Unintended Consequences of FOMC Press Conferences. *Journal of Financial and Quantitative Analysis (forthcoming).*

Brand, C., D. Buncic, and J. Turunen. 2010. The Impact of ECB Monetary Policy Decisions and Communication on the Yield Curve. *Journal of the European Economic Association 8*, pp. 1266–1298.

Brusa, F., P. Savor, and M. Wilson. 2019. One Central Bank to Rule Them All. *Review of Finance (forthcoming).*

Busse, J. A., and T. C. Green. 2002. Market efficiency in real time. *Journal of Financial Economics 65*, pp. 415–437.

Chakraborty, G., M. Pagolu, and S. Garla. 2013. *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS.* SAS Institute.

Chan-Lau, J. A., E. X. Liu, and J. M. Schmittmann. 2015. Equity returns in the banking sector in the wake of the Great Recession and the European sovereign debt crisis. *Journal of Financial Stability 16*, pp. 164–172.

Chouliaras, A., and T. Grammatikos. 2017. Extreme Returns in the European financial crisis. *European Financial Management 23*, pp. 728–760.

Cieslak, A., A. Morse, and A. Vissing-Jorgensen. 2019. Stock Returns over the FOMC Cycle. *The Journal of Finance 74*, pp. 2201–2248.

Cieslak, A., and A. Schrimpf. 2018. Non-Monetary News in Central Bank Communication. *Working Paper.*

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

Cowles, A. 1933. Can Stock Market Forecasters Forecast? *Econometrica 1*, pp. 309–324.

Cox, M. A., and T. F. Cox. 2008. *Multidimensional Scaling*, vol. Handbook of Data Visualization, pp. 315–347. Springer Handbooks Comp.Statistics.

Das, S. R. 2014. Text and Context: Language Analytics in Finance. *Foundations and Trends in Finance 8*, pp. 145–261.

Das, S. R., and M. Y. Chen. 2007. Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science 53*, pp. 1375–1388.

Diamond, D. W., and R. G. Rajan. 2012. Illiquid Banks, Financial Stability, and Interest Rate Policy. *Journal of Political Economy 120*, pp. 552–591.

Dong, Y., M. Massa, and H. Zhang. 2018. Guru Dreams and Competition: An Anatomy of the Economics of Blogs. *Working Paper*.

Ehrmann, M., and M. Fratzscher. 2009. Explaining Monetary Policy in Press Conferences. *International Journal of Central Banking 5*, pp. 41–84.

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

Fama, E. F. 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance 25*, pp. 383–417.

Fama, E. F., L. Fisher, M. C. Jensen, and R. Roll. 1969. The Adjustment of Stock Prices to New Information. *International Economic Review 10*, pp. 1–21.

Fatum, R., and M. M. Hutchison. 2002. ECB Foreign Exchange Intervention and the EURO: Institutional Framework, News, and Intervention. *Open economies review 13*, pp. 413–425.

Faulkner, S., A. Eicholz, T. Leithead, A. Danilo, and S. Moon. 2017. HTML 5.2 Specification. W3C recommendation. URL `https://www.w3.org/TR/2017/REC-html52-20171214/`.

Feldman, R., and J. Sanger. 2007. *The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Fleming, M. J., and M. Piazzesi. 2005. Monetary Policy Tick-by-Tick. *Federal Reserve Bank of New York Working Paper.*

Fratzscher, M., M. Lo Duca, and R. Straub. 2016. ECB Unconventional Monetary Policy: Market Impact and International Spillovers. *IMF Economic Review 64*, pp. 36–74.

Gerlach, S. 2007. Interest Rate Setting by the ECB, 1999-2006: Words and Deeds. *International Journal of Central Banking 3*, pp. 1–46.

Gilbert, T., A. Kurov, and M. H. Wolfe. 2018. The Disappearing Pre-FOMC Announcement Drift. *Working Paper.*

Grimmer, J., and B. M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis 21*, pp. 267–297.

Gu, C., A. Kurov, and M. H. Wolfe. 2018. Relief Rallies after FOMC Announcements as a Resolution of Uncertainty. *Journal of Empirical Finance 49*, pp. 1–18.

Gürkaynak, R. S., B. Sack, and E. T. Swanson. 2005. Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements. *International Journal of Central Banking 1*, pp. 55–93.

Halsall, F. 2005. *Computer Networking and the Internet.* Pearson Education Limited.

Hansen, L. P., and T. J. Sargent. 2001. Robust Control and Model Uncertainty. *The American Economic Review 91*, pp. 60–66.

Hansen, S., and M. McMahon. 2016. Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics 99*, pp. 114–133.

Hansen, S., M. McMahon, and M. Tong. 2018. The Long-Run Information Effect of Central Bank Narrative. *Working Paper.*

Hardouvelis, G. A., G. I. Karalas, D. I. Karanastasis, and P. K. Samartzis. 2018. Economic Policy Uncertainty, Political Uncertainty and the Greek Economic Crisis. *Working Paper.*

Heinemann, F., and K. Ullrich. 2007. Does it Pay to Watch Central Bankers' Lips? The Information Content of ECB Wording. *Swiss Journal of Ecomomics and Statistics 143*, pp. 155–185.

Heyer, G., U. Quasthoff, and T. Wittig. 2012. *Text Mining: Wissensrohstoff Text.* W3L Verlag.

Hoberg, G., and G. Phillips. 2010. Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis. *The Review of Financial Studies 23*, pp. 3773–3811.

Hoerl, A. E., and R. W. Kennard. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics 12*, pp. 55–67.

Hu, G. X., P. J., J. Wang, and Z. H. 2019. Premium for Heightened Uncertainty: Solving the FOMC Puzzle. *NBER Working Paper.*

Hull, J. C. 2019. *Machine Learning in Business: An Introduction to the World of Data Science.* Independently published.

Illing, F. 2013. *Die Euro-Krise - Analyse der europäischen Strukturkrise.* SpringerVS.

Issing, O. 2003. Evaluation of the ECB's monetary policy strategy. *Press Seminar.*

Jaccard, P. 1901. Distribution de la Flore Alpine dans le Bassin des Dranses et dans Quelques Régions Voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles.*

Jansen, D.-J., and J. De Haan. 2005. Talking heads: the effects of ECB statements on the euro–dollar exchange rate. *Journal of International Money and Finance 24*, pp. 343–361.

Jarocinski, M., and P. Karadi. 2018. Deconstructing monetary policy surprises: the role of information shocks. *ECB Working Paper Series*.

Jegadeesh, N., and D. Wu. 2013. Word Power: A New Approach for Content Analysis. *Journal of Financial Economics 110*, pp. 712–729.

Jegadeesh, N., and D. Wu. 2017. Deciphering Fedspeak: The Information Content of FOMC Meetings. *Working Paper*.

Jurafsky, D., and J. H. Martin. 2011. *Speech and Language Processing*. 2nd ed. Pearson International Edition.

Jurafsky, D., and J. H. Martin. 2017. Speech and Language Processing. Third Edition draft.

Kaul, A., and M. Watanabe. 2015. Dissecting the Pre-FOMC Announcement Drift. *Working Paper*.

Kearney, C., and S. Liu. 2014. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis 33*, pp. 171–185.

Kelly, B., A. Manela, and A. Moreira. 2019. Text Selection. *Working Paper*.

Krishnamurthy, A., S. Nagel, and A. Vissing-Jorgensen. 2018. ECB Policies Involving Government Bond Purchases: Impact and Channels. *Review of Finance 22*, pp. 1–44.

Kurose, J. F., and K. W. Ross. 2013. *Computer Networking - A Top-Down Approach*. Pearson Education Limited.

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

Leombroni, M., A. Vedolin, C. Venter, and P. Whelan. 2017. Central Bank Communication and the Yield Curve. *Working Paper*.

Lopez-Lira, A. 2019. Risk Factors That Matter: Textual Analysis of Risk Disclosures for the Cross-Section of Returns. *Working Paper*.

Loughran, T., and B. McDonald. 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance 66*, pp. 35–65.

Loughran, T., and B. McDonald. 2016. Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research 54*, pp. 1187–1230.

Lucca, D. O., and E. Moench. 2015. The Pre-FOMC Announcement Drift. *The Journal of Finance 70*, pp. 329–371.

Lucca, D. O., and E. Moench. 2018. The Pre-FOMC Announcement Drift: More Recent Evidence. URL `https://libertystreeteconomics.newyorkfed.org`.

Lucca, D. O., and F. Trebbi. 2011. Measuring Central Bank Communication: An Automated Approach with Application to FOMC Statements. *Working Paper*.

MacQueen, J. 1967. Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1*, pp. 281–297.

Manela, A., and A. Moreira. 2017. News implied volatility and disaster concerns. *Journal of Financial Economics 123*, pp. 137–162.

Manning, C. D., and H. Schütze. 2003. *Foundations of Statistical Natural Language Processing*. The MIT Press Cambridge.

Martello, V. G. R., and R. M. Ribeiro. 2018. Pre-FOMC Announcement Relief. *Working Paper*.

Merkley, K. J. 2014. Narrative Disclosure and Earnings Performance: Evidence from R&D Disclosures. *The Accounting Review 89*, pp. 725–757.

Mitra, G., D. diBartolomeo, A. Banerjee, and X. Yu. 2011. Automated Analysis of News to Compute Market Sentiment: Its Impact on Liquidity and Trading. Tech. rep., Government Office for Science.

Mitra, G., and L. Mitra (eds.). 2011a. *The Handbook of News Analytics in Finance.* Wiley.

Mitra, G., and X. Yu (eds.). 2016. *Handbook of Sentiment Analysis in Finance.* Optirisk Systems.

Mitra, L., and G. Mitra. 2011b. *Applications of News Analytics in Finance: A Review*, vol. The Handbook of News Analytics in Finance, chap. 1. John Wiley and Sons.

Moniz, A., G. Brar, C. Davies, and A. Strudwick. 2011. *The Impact of News Flow on Asset Returns: An Empirical Study*, vol. The Handbook of News Analytics in Finance, chap. 8. John Wiley and Sons.

Murphy, K. P. 2012. *Machine Learning: A Probabilistic Perspective.* The MIT Press Cambridge.

Peterson, R. L. 2002. "Buy on the Rumor:" Anticipatory Affect and Investor Behavior. *Journal of Psychology and Financial Markets 3*, pp. 218–226.

Peterson, R. L. 2016. *Trading on Sentiment - The Power of Minds Over Markets.* John Wiley and Sons.

Romer, C. D., and D. H. Romer. 2004. A New Measure of Monetary Shocks: Derivation and Implications. *The American Economic Review 94*, pp. 1055–1084.

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

Rosa, C., and G. Verga. 2007. On the consistency and effectiveness of central bank communication: Evidence from the ECB. *European Journal of Political Economy 23*, pp. 146–175.

Rosa, C., and G. Verga. 2008. The Impact of Central Bank Announcements on Asset Prices in Real Time. *International Journal of Central Banking 4*, pp. 175–217.

Savor, P., and M. Wilson. 2013. How Much Do Investors Care About Macroeconomic Risk? Evidence from Scheduled Economic Announcements. *Journal of Financial and Quantitative Analysis 48*, pp. 343–375.

Savor, P., and M. Wilson. 2014. Asset pricing: A tale of two days. *Journal of Financial Economics 113*, pp. 171–201.

Schmeling, M., and C. Wagner. 2019. Does Central Bank Tone Move Asset Prices? *Working Paper.*

Stone, P. J., D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis.* MIT Press, Cambridge, MA.

Tanenbaum, A. S. 2003. *Computer Networks.* Pearson Education International.

Tetlock, P. C. 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance 62*, pp. 1139–1168.

Tetlock, P. C. 2011. All the News That's Fit to Reprint: Do Investors React to Stale Information? *The Review of Financial Studies 24*, pp. 1481–1512.

Tetlock, P. C. 2014. Information Transmission in Finance. *Annual Review of Financial Economics 6*, pp. 365–384.

Ulrich, M., E. Jakobs, L. May, and J. Landwehr. 2019a. The Euro Crisis and the 24h Pre-ECB Announcement Return. *Karlsruhe Institute of Technology Working Paper.*

Ulrich, M., E. Jakobs, and R. Tran. 2019b. The Real-Time Impact of ECB Press Conferences on Financial Markets. *Karlsruhe Institute of Technology Working Paper.*

Wachter, J. A., and Y. Zhu. 2018. The Macroeconomic Announcement Premium. *Working Paper.*

Woodford, M. 2012. Discussion of 'Macroeconomic Effects of Federal Reserve Forward Guidance'. *The Brookings Papers on Economic Activity, Spring 2012*, pp. 64–73.

Glossary

Asset purchase programs (APP) Non-standard monetary policy measures implemented by the ECB under which private and public sector securities are purchased to increase inflation over the medium term.

Client Computer hardware or software that uses a service hosted by a server.

Corpus Collection of documents.

Document Basic element in text mining that can be very informally defined as a unit of discrete textual data within a collection that usually, but not necessarily, correlates with some real-world document (Feldman and Sanger, 2007, p. 3).

Document collection A collection of documents can be any grouping of text-based documents (Feldman and Sanger, 2007, p. 2).

European Central Bank (ECB) Central bank of the 19 European Union countries which have adopted the euro.

Feature In machine learning and pattern recognition, a feature is an individual measurable property or characteristic of a phenomenon being observed (Bishop, 2006). In textual analysis popular document features are characters, words, terms or concepts (Feldman and Sanger, 2007, p. 4f).

Federal Open Market Committee (FOMC) Federal Reserve's monetary policy-making body.

Federal Reserve (Fed) Central banking system of the United States of America.

Governing Council Main decision-making body of the ECB.

Longer-term refinancing operations (LTROs) Liquidity-providing operations in euro that offer additional longer-term refinancing to the financial sector.

Markup language System (such as HTML) for annotating a document that describes its logical structure (like headlines) and defines its layout on the page.

Natural language processing (or **computational linguistics**) Set of techniques for computational processing and analysis of naturally occurring human languages (Bholat et al., 2015).

Protocol A protocol defines the format and the order of messages exchanged between two or more communicating entities, as well as the actions taken on the transmission and/or receipt of a message or other event (Kurose and Ross, 2013).

Server Computer program or device that offers functionality to other devices or programs.

Structured data Data stored in fields in a traditional database (Bholat et al., 2015).

Supervised machine learning Set of techniques that learn from classified observations and then assign classes on unseen data, based on the prior observed distribution (Bholat et al., 2015).

Term frequency-inverse document frequency (tf-idf) Common weighting scheme used in the text mining literature that consists of giving lesser weight to words that appear more frequently and greater weight to words that appear less frequently (Bholat et al., 2015).

Term-document matrix Matrix capturing the frequency of each term in the set of documents (Bholat et al., 2015).

Text mining Text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools (Feldman and Sanger, 2007, p. 1).

Tokens Words, phrases, or symbols in text (Bholat et al., 2015).

Training data In text mining, set of classified data used to 'train' a supervised algorithm to 'learn' the co-occurrence of classes and words (Bholat et al., 2015).

Unconventional (or non-standard) monetary policy measures Forms of monetary policy other than setting the interest rates.

Uniform Resource Locator (URL) Reference to a Web resource (e.g. Web page) that defines its location on a computer network and a mechanism to retrieve it.

Unstructured data Data not residing in a traditional row/column database format (Bholat et al., 2015).

Unsupervised machine learning Set of techniques that involve taking unclassified observations (text or otherwise) and uncovering hidden patterns (Bholat et al., 2015).

VSTOXX index Measure of market expectations of the volatility of option prices on the Euro Stoxx 50 index.