REP. OF TURKEY

TED UNIVERSITY

GRADUATE SCHOOL

APPLIED DATA SCIENCE

AN ANALYSIS OF THE CENTRAL BANK OF THE REPUBLIC OF TÜRKİYE (CBRT)'S MONETARY POLICY COMMUNICATION WITH TEXT MINING

ŞEYDA AYAN ÖZBEK

ANKARA, 2022

AN ANALYSIS OF THE CENTRAL BANK OF THE REPUBLIC OF TÜRKİYE (CBRT)'S MONETARY POLICY COMMUNICATION WITH TEXT MINING

A Thesis Submitted To
The Graduate School
of
TED University

by

Şeyda Ayan Özbek

In Partial Fulfillment of The Requirements
For
Master of Science
in
Applied Data Science

ANKARA, 2022

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Şeyda Ayan Özbek

Signature :

ABSTRACT

AN ANALYSIS OF THE CENTRAL BANK OF THE REPUBLIC OF TÜRKİYE (CBRT)'S MONETARY POLICY COMMUNICATION WITH TEXT MINING

Şeyda Ayan Özbek
Master of Science, Applied Data Science
Supervisor: İbrahim Ünalmış
August, 2022

In the last twenty years, most of the central banks have started to use short term interest rates as a main monetary policy tool to control inflation. Moreover, communication strategy has become one of their most important policy tools through time. Especially, after the 2008 Global Financial Crisis, when the nominal interest rates hit the zero lower bound, central banks have used communication policy very effectively to form expectations. As a result of these developments, there is a growing body of literature examining the communication strategies of central banks. Advances in text mining literature also enables to analyze central bank texts by using alternative methods. Under this background, this thesis analyses the Central Bank of the Republic of Türkiye's (CBRT) Monetary Policy Committee decision announcements for the period 2006-2020. In the analysis, different text mining techniques, data cleaning and preprocessing, descriptive text statistics and sentiment analysis of announcements, and machine learning applications are used and contributions to policy making process are discussed.

Keywords: Central Bank of the Republic of Türkiye, central bank communication, monetary policy, text mining, sentiment analysis

ÖZET

TÜRKİYE CUMHURİYET MERKEZ BANKASI (TCMB) PARA POLİTİKASI İLETİŞİMİNİN METİN MADENCİLİĞİ İLE ANALİZİ

Şeyda Ayan Özbek
Master of Science, Uygulamalı Veri Bilimi
Tez Yöneticisi: İbrahim Ünalmış

Ağustos, 2022

Son 20 yıldır merkez bankaları kısa vadeli faiz oranlarını enflasyonu kontrol etmek için temel politika aracı olarak kullanmaktadır. Ayrıca, zaman içerisinde iletişim stratejisi merkez bankalarının en önemli politika araçlarından biri haline gelmiştir. Özellikle, nominal faiz oranlarının sıfır alt sınırına geldiği 2008 Küresel Krizi sonrası, merkez bankaları beklentileri yönetmek için iletişim politikasını oldukça etkin kullanmışlardır. Tüm bu gelişmelerin sonucu olarak giderek büyüyen bir literatür merkez bankalarının iletişim stratejilerini incelemektedir. Öte yandan, metin madenciliği literatüründeki gelişmeler, alternatif yöntemlerle merkez bankası metinlerinin analiz edilmesine yardımcı olmaktadır. Bu doğrultuda, bu tez çalışmasında Türkiye Cumhuriyet Merkez Bankası (TCMB)'nin Para Politikası Kurulu kararları 2006-2020 dönemi için incelenmiştir. Analizlerde; farklı metin madenciliği yöntemleri, veri temizliği ve önişlemesi, tanımlayıcı metin istatistikleri, karar metinlerinin algı analizi ve makine öğrenmesi uygulamaları kullanılmış, politika sürecine katkıları tartışılmıştır.

Anahtar Kelimeler: Türkiye Cumhuriyet Merkez Bankası, merkez bankası iletişimi, para politikası, metin madenciliği, algı analizi

To my mom…

ACKNOWLEDGMENTS

I would like to express my most sincere thanks to my supervisor İbrahim Ünalmış. His guidance, broad knowledge, and support have been very invaluable for both this thesis and my career. I would also like to express my deepest gratitude for his understanding and positive attitude throughout my study.

Moreover, I wish to thank Taha Eren Sarnıç for his contributions and insights during our classes in Master's program since day one and also for this thesis.

I also would like to express my thanks to my former managers, colleagues and old but gold colleagues from the Central Bank of Republic of Türkiye, Selim and Ahmet for taking the first steps of this study together.

Further, I wish to state my thanks to my current managers and colleagues at TÜBİTAK for their support and understanding.

This thesis have witnessed many important changes and developments in the transition period to adulthood, which made this process more challenging in some times. Thus, the support of my beloved ones has become more precious.

First of all, I wouldn't be where I am now without the endless support and encouragements of my family. I would like to thank to each member of my family; my grandparents, my parents, my Özbek family, and siblings… I thank my brother Berkay Ayan for being not only a brother but a best friend and for turning the light off whenever I need. I thank my father Baki Ayan for giving me strength and encouragement. Especially, I think I am more than lucky and grateful for my mother, Şenay Gündüz. Her endless love and belief in me always enlighten my life.

It is a cliché but I literally believe that friends are our chosen family. I am so, so grateful that I have Büşra in this family. She is more than a friend, a sister for me. Each and every moment we have shared, our talks, laughs, and dreams have given me inspiration, joy, and hope since our highschool years. Without her in my life, I would stumble more, for sure.

I am so lucky that I have Bahar and her friendship, support, and humor in my life. She is always there for me and this is very precious. I am so glad that we have what we have in that “İktisat Kantini” and have collected many memories together since then with “Date” and “Farketmez Buluşmaları”.

I am so happy that we have grown up with Doğukan and bringing our primary school years to our adult lives. Moreover, I am grateful for “Tatiling” and “Asil Beyin Melekleri” for our past and future holidays, for their valuable friendship, and bringing joy to my life.

And I am glad that we had the chance to have Samet Albasar in our lives. His friendship and existence have taught a lot. I hope he is in a better place watching us.

Last but most precious, deserving more than thanks to my better half, team mate, and lovely spouse Ekin for sharing this life with me and always standing by me. Being a team and family with him is one of the best things that ever happened to me. I am so grateful for his presence as we make our dreams and way together. We have come a looong way and even more is waiting. And for anything to come “No, I won't be afraid just as long as you stand stand by me…”

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ARI	Automatic Readability Index
CBRT	Central Bank of the Republic of Türkiye
CDS	Credit Default Swap
DTM	Document Term Matrix
ECB	European Central Bank
EUR	Euro
FED	Federal Reserve
FOMC	Federal Open Market Committee
HC	Hierarchical Clustering
HTML	Hyper Text Markup Language
IDF	Inverse Document Frequency
LDA	Latent Dirichlet Allocation
LM Dictionary	Loughran-McDonald Dictionary
LSA	Latent Semantic Analysis
ML	Machine Learning
MPC	Monetary Policy Committee
SVM	Support Vector Machines
TCMB	Türkiye Cumhuriyet Merkez Bankası
TF	Term Frequency
TRY	Turkish Lira
USD	United States Dollars
VAR	Vector Autoregression
VSM	Vector Space Model
XML	Extensible Markup Language

1. INTRODUCTION

1.1 Motivation

One of the benefits of data science is making the unobservable observable. With the increase in the volume of data generated, as much data as possible is tried to be collected and used in order to create more information and gain insights. Thus, many different approaches have emerged and developed to explore data. The general term used for these approaches is "data mining" which is described as the efforts on collection, cleaning, process, analysis, and getting useful insights from data (Aggarwal, 2015, p.1).

Since data to analyze exist in various forms such as image, audio, text etc., many different data mining techniques have emerged. Text mining is one of those techniques which is used to make use of text data and create more information. The importance of text mining stems from its data mining methods and creating additional value from an unusual source: the text itself. Text mining, also known as text data mining or knowledge discovery from textual databases, generally refers to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents (Tan, 1999). In addition, traditional data mining and text mining differ from each other in terms of data sources: One uses the structured databases and the other extracts patterns from natural language text, which is in an unstructured form (Hearst, 2003). Moreover, text mining is interdisciplinary since it utilizes information retrieval, data mining, machine learning, statistics and computational linguistics (Gupta & Lehal, 2009).

Thanks to its advantages, text mining is used in many different areas such as social media, marketing, scientific literature mining, sociology, security, banking, politics, economics and so on. Central bank communication is one of the areas in which text mining is used extensively. The main reason is that central banks use communication as a policy tool to form expectations. In this thesis, text mining is used to analyze the Central Bank of Türkiye's (CBRT) communication strategy through time.

Central bank communication can be defined as “the provision of information by the central bank to the public regarding the objectives of monetary policy, the monetary policy strategy, the economic outlook, and the outlook for future policy decisions” (Blinder et al., 2008, p.10).

Over the last decades, as the central banks have become more transparent in their monetary policy decisions, central bank communication has become one of the most important components of monetary policy. The transparency of the central bank is defined by Geraats (2002) as “the absence of asymmetric information between monetary policymakers and other economic agents” (p.1). The central bank transparency has become important due to the adoption of inflation-targeting monetary policy (De Haan et al., 2007). This is important for central banks since it contributes to the monetary policy effectiveness by helping to manage market expectations and increase accountability (Montes et al., 2016). De Haan et al. (2007) highlighted that although central bank transparency has been justified by central bank accountability, it is also important from an economic point of view since monetary policy has been becoming the art of expectation management. Therefore, with more transparency, central banks are giving more importance to central bank communication to explain their decisions and possible future actions to avoid information asymmetry and manage expectations for a more effective monetary policy. In addition, central bank communication is also important for financial markets and the general public to understand the role of the central bank and act accordingly.

Thus, the common studies on central bank communications investigate the macroeconomic effects of communication, meanings of the messages from statements, future guidances, dovishness/hawkishness tones of the decisions, etc. Especially, the market reactions to policy rate decisions, the effects of policy rate announcements on interest rates, other macroeconomic indicators and expectations, sentiment in decisions and topics in the communications have been examined. For this aim, different quantitative and qualitative techniques are used (Hughes & Kesting, 2014). In particular, while investigating the macroeconomic effects of central bank communication event analysis and vector autoregression (VAR) methods are adopted widely.

Yet, there is no consensus on the optimal communication strategy. Despite the practicing details and timing differences, the main monetary policy communication tools are policy rate decision announcements, press statements, minutes of monetary policy committee (MPC) meetings, regular reports, speeches, press conferences and macroeconomic forecasts. As it can be inferred, most of those communication tools are in text format. This is where central bank communication meets with text mining. In addition to all of the research about the macroeconomic effects of central bank communication, text mining is a continuation of such research and enables the quantification of the communication.

When it comes to text mining; the potential worth of text mining for central banks is emphasized as there are different text sources such as MPC speeches, policy reports, social media, and news (Bholat et al., 2015). As Bholat et al. (2015) indicate, the use of text mining techniques is still limited; however, there is a growing literature investigating central bank communication and other related documents with different techniques of text mining; such as sentiment analysis, supervised-unsupervised machine learning methods, topic modeling etc.

In this thesis, the main motivation is to investigate the CBRT's monetary policy communication from a different perspective by using the power of data science and focus on text mining techniques to contribute to the related literature.

In the case of the CBRT, according to the official website of the CBRT, the main communication tools of monetary policy are MPC decision announcements with evaluations regarding the decision and Inflation Reports. Moreover, other reports about monthly inflation, financial stability, announcements, presentations, meetings with different shareholders such as investors, analysts, financial and non-financial sector representatives are other components of the CBRT's communication strategy.

For this thesis, among those communication tools, MPC decision announcements in the period 2006-2020 are investigated. This study will be an example of the practical usage and benefits of text mining for central banks. Different text mining techniques, descriptive text statistics and sentiment analysis of announcements, machine learning applications and contributions to policy making process will be discussed.

Since the text mining applications will be conducted via R, the thesis will be an example of text mining applications with R.

1.2 Outline of Thesis

In Chapter 2, the related literature on text mining, central bank communication, the CBRT's monetary policy communication, text mining and central banks will be introduced.

Chapter 3 will cover the methodology used in this thesis by focusing on text preprocessing, text statistics, readability and formality scores, word-based and sentence-based sentiment analysis and supervised-unsupervised machine learning methods.

The data used in this thesis, the application process and the evaluation of the results will be the topics of Chapter 4.

Finally, in Chapter 5, the conclusion and future works will be discussed.

2. LITERATURE REVIEW

This chapter aims to examine the related studies in the literature. In this thesis, the related literature is separated into three parts: Text Mining, Central Bank Monetary Policy Communication, Text Mining and Central Banks.

2.1 Text Mining

Text mining is originated from library science, information science and natural language processing with the applications of information retrieval and summarization such as document grouping, indexes and abstracts. However, since the 1990s, with the emerging techniques in the field of data mining, statistical analysis, and statistical learning, text mining has changed significantly (Miner et al., 2012).

Consequently, text mining is currently considered an interdisciplinary method that uses information retrieval, machine learning, statistics, computational linguistics, and data mining (Hotho et al., 2005). While Hearst (2003) defined text mining as "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources" (p.1); Hotho et al. (2005) emphasized that to define text mining, one can refer to the related research area and there could be three different definitions of text mining in different perspectives. According to their study, the three definitions of text mining are as follows: text mining as information extraction, text mining as text data mining, and text mining as knowledge discovery in databases.

As it is deduced, there are different approaches and definitions of text mining in the literature. Besides, the importance of text mining is also discussed with the increase in the amount of text data in recent years. Allahyari et al. (2017) emphasized the importance of text mining by highlighting the tremendous amount of text data available, which is an invaluable source of information and knowledge.

The importance of text mining is also discussed along with the increase of unstructured data in the literature. According to Fan et al. (2006), with the increase in unstructured or semi-structured data beyond the reach of data mining, text mining is used for further analysis since it is designed to handle structured data from databases or XML files and also works with unstructured or semi-structured data sets (such as email, full-text documents, and HTML files). In addition, according to IBM (2020), 80% of data in the world is unstructured, and since the volume of text data increases every day, text mining is highly valuable.

The methods of text mining draw on different areas; natural language processing, statistics, machine learning, reasoning, information extraction, and information retrieval knowledge management (Gupta & Lehal, 2009). Since text mining is an interdisciplinary field, there are slight differences between its usage areas, methods and approaches. Miner et al. (2012), summarizes the main five areas of usage of text mining as follows: getting meaning from an unstructured text (sentiment analysis, themes in a text etc.), automatic text categorization (automatic email routing, spam filtering, fraud detection, automatic translation, etc.), improving accuracy in predictive modeling or unsupervised learning (clustering, etc.), identifying specific or similar/relevant documents, and extracting specific information from the text i.e. entity extraction. Atan (2020) listed the usage areas of text mining as follows: information extraction, topic recognition and classification, similarity and plagiarism detection, topic summary, concept link detection, and sentiment analysis.

Moreover, the steps of text mining may differ according to its usage areas, analysis units and methods. However, the general steps are outlined by Sumathy and Chidambaram (2013) as follows: text preprocessing (tokenization, stopword removal, stemming, etc.), text transformation (bag of words, vector spaces format), feature selection, text mining methods (clustering, classification, machine learning algorithms, information retrieval, etc.) and interpretation. According to their study, text mining applications are often used in telecommunication, media, banking, financial markets, political analysis, energy, healthcare, bio-informatics, national security, and service sector.

As an example of a different usage of text mining, Hong and Park (2019) examined airline review data with text mining methods. By using keyword extraction and clustering methods, they classified keywords into three sections and they found that there are different relationships between keywords and corporate performance. Wegrzyn-Wolska & Bougueroua (2012) conducted text mining on Twitter to analyze French presidential election trends. Tweets are examined using sentiment and content analysis. Kolini and Janczewski (2017) collected national cybersecurity strategies of different regions and applied clustering and topic modeling to examine the differences and similarities between them. A text categorization application by using the machine learning methods Naïve Bayes, Support Vector Machines (SVM) and neural networks are conducted by Shahi and Pant (2018) to be able to classify Nepali news. Fung et al. (2005), used text mining and data mining techniques to predict the movements of stock prices by analyzing the impacts of the news stories on the stocks. More than 350.000 documents from Reuters news stories and intraday stock prices of all the Hong Kong stocks were analyzed.

As a side note, to conduct these analyses with text mining, there are both paid software, such as the products of IBM and SAS, and free and open source programming languages, such as R and Python, which have certain packages for text mining.

2.2 Central Bank Communication

Central banks use many different tools to achieve targets of monetary policy. One of these important tools has been central bank communication in recent years. Blinder et al. (2008) emphasized that before the 1990s, central banks preferred to be shrouded in mystery and believed they should be. According to Woodford (2005), there has been a notable change in central banking over the last decades and there has been a worldwide movement towards an increase in central bank communication about policy decisions, targets behind those decisions, and views on the economy.

In one of her speeches in 2012, the chairperson of the Federal Open Market Committee (FOMC) Jennet Yellen focused on revolution and evolution in central bank communication and stated that revolution in central bank communication is not a result of technological advances, but it is a result of improvements in the understanding of effective monetary policy. Yellen pointed out that growing research and experience show that clear communication is a vital tool for raising the efficiency and reliability of the monetary policy.

In the literature, mainly, inflation targeting, transparency, accountability, and effectiveness of monetary policy are related to this change in central bank communication. "Inflation targeting" is explained by Bernanke and Mishkin (1997) as following: "This approach is characterized by the announcement of official target ranges for the inflation as at one or more horizons, and by an explicit acknowledgment that low and stable inflation is the overriding goal of monetary policy." (p.97). Also in their study, they indicated that the important features of this strategy are increased transparency, communication with the public, and in many cases, increased accountability.

After 1990s, with the adoption of inflation targeting monetary policy by New Zealand, Canada, United Kingdom, and Sweden central banks, transparency has increased substantially for both these central banks and also for the other central banks having different monetary policy strategies (De Haan et al., 2007). With the increase in the adoption of inflation targeting strategies, central bank independence, accountability, and transparency have started to come to the fore. According to Mishkin (2004), with the advent of inflation targeting in the early 1990s, central banks started to follow a different route to solve the time-inconsistency problem, and they recognized the importance of transparency and improved communication with the public and the markets to conduct a successful monetary policy. Mishkin also pointed out that the emphasis on transparency and communication has many benefits for central banks such as increasing credibility and anchoring inflation expectations.

Blinder et al. (2001) defended that central banks' becoming more open is a consequence of the new approaches to monetary policy conduct. According to their study, there are two reasons for the openness of central banks: one is the effectiveness of monetary policy and the other is democratic accountability. They highlighted that with more transparency, market expectations will respond to policy changes better and faster. Moreover, openness is a need for accountability and transparency which can be seen as a result of central bank independence according to their study.

In the literature, central bank communication and its effects on monetary policy transmission mechanism and expectation management are also discussed widely. In his speech in 2018, Jens Weidmann, the President of the Deutsche Bundesbank, highlighted the importance of central bank communication as a monetary policy instrument by saying that "Communication is vital not only to make a central bank more transparent but also, and above all, to steer expectations.".

Ehrmann and Fratzscher (2005) stated that "Central bank communication is a key determinant of the market's ability to anticipate monetary policy decisions and the future path of interest rates." (p. 4). They also emphasized that the success of central banks is related to the ability to influence asset prices and interest rates in all maturities while the only direct control they have is over a single interest rate.

Eusepi and Preston (2010) highlighted that one of the potential benefits of inflation targeting is anchoring expectations by stabilizing the effect on macroeconomic activity; thus, with this role of expectations, central bank communication is a crucial part of this regime.

Feldkircher et al. (2021) indicated that although central bank communication is not placed in monetary policy transmission mechanism diagrams by most central banks, it is seen as a separate role and function which is essential. They pointed out that there are at least three reasons why central bank communication can play a distinct role in monetary policy transmission mechanisms.

Those reasons can be listed as follows: “First, inflation expectations are not the only expectations that matter, expectations of future interest rates, that is, forward guidance, are also important for investors, households and firms. Second, central banks provide information not only about inflation but also about different topics. Third, since the global financial crisis and the euro area sovereign debt crisis, the emphasis of central bank communication has shifted toward topics related to ensuing implications for financial stability.” (p.62).

De Haan et al. (2007) emphasized that the ability of central banks to affect the economy depends on the ability to affect market expectations about the future overnight interest rates, and not only the current level; thus, monetary policy is increasingly becoming the “art of managing expectations”.

As a result, communication has developed into a key instrument in the central bankers’ toolbox in recent years. Neuenkirch (2012) studied transparency and informal central bank communication effects on money market expectations by investigating nine major central banks from January 1999 to July 2007. He found out that transparency reduces the bias in money market expectations and informal communications help to manage financial market expectations thanks to reducing the variation.

The macroeconomic effects of central bank communication are also discussed in the literature widely. Gürkaynak et al. (2004) investigated the FOMC announcements’ effects on financial markets with an event-study analysis. According to their study, both monetary policy actions and statements affect asset prices. Also, statements have more effects on long-term Treasury yields. Neuenkirch (2013) analyzed the role of communication in monetary policy transmission with VAR analysis. In this study, European Central Bank (ECB)’s communication was on the spot and the role of communication in the ECB’s monetary policy transmission to inflation, inflation expectations and output was investigated. The results of the study pointed out that communication has important effects on inflation, inflation expectations, and output.

Another discussion in the literature is about what and how central banks communicate. There is no consensus on the optimal communication strategy of central bank communication.

According to Blinder et al. (2008) and Şen-Taşbaşı (2011), central bank communication includes "objectives, strategies, reasons behind a policy decision, economic outlook, assessments of future indicators and economic activity, projections regarding future monetary policy decisions", the content and communication channels can differ across central banks, and the messages can be sent by committee or a committee member. Also, Şen-Taşbaşı (2011) indicated that there are two targetted audiences of central bank communication; broad public and financial markets.

Brouwer and De Haan (2021) pointed out that central banks primarily focused on audiences of specialists such as financial market specialists, academics, and journalists, but recently, general public is also targeted because of the increased accountability and its help to achieve price stability by anchoring inflation expectations and stabilizing economic conditions thanks to managing inflation expectations.

Additionally, the communication tools of central banks are listed as decision statements, minutes, inflation reports, speeches, press conferences, etc., in the literature.

2.2.1 Communication of the CBRT

One of the institutional publications of the CBRT is focused on central banks' communication, the CBRT's monetary policy communication evaluation, and current strategy (2011). In this publication, it is emphasized that the communication strategy of the CBRT has evolved in time similar to other central banks and 2001 is an important year for both institutional structure and policies. Thus, the communication of the CBRT is investigated in two periods: before and after 2001.

Before 2001, the CBRT's communication strategy was more cautious, conservative and in line with the general trend. Also, in this period, the CBRT was not independent and policy priorities were not defined which had effects on transparency and communication.

Until 2001, there were several important advancements in communication, but the most important year was 2001. In 2001, with the amendment of the Central Bank Law, the CBRT's target was defined as "price stability", the CBRT gained instrument independence, MPC was founded, and the design and implementation of policies were determined to be based on transparency and communication. In 2002, the CBRT announced that the ultimate objective of the monetary policy was inflation targeting regime. Between 2002 and 2005, there was a transition period named as Implicit Inflation Targeting Regime and necessary preconditions were ensured.

In 2006, after the disinflationary process in the transition period and attaining the necessary preconditions, Inflation Targeting Regime was adopted. In 2010, after the global financial crisis in 2008-2009, the CBRT added new policy instruments to its toolbox and financial stability became important.

In the CBRT publication named "Central Banks and Communication" in 2011, the communication tools of the CBRT are listed as Inflation Reports, MPC decision statements and minutes, speeches, press releases, open letters to the government, monthly price developments, Financial Stability Reports, press relations and briefings, annual monetary and exchange rate policy texts, strategic plan, the CBRT general web page, booklets – newsletters, publications – statistics and other tools.

Currently, according to the CBRT's general website[1], the main communication tools of the CBRT are listed as MPC announcements and Inflation Reports; other reports and the CBRT presentations are also listed as communication tools. Also, on the website, it is indicated that in 2016 there were innovations in communication policy which were having regular meetings with real sector representatives and financial institutions, increasing the number and capacity of technical meetings with investors and analysts, actively using social media and opening up the CBRT blog.

In the literature, the communication of the CBRT is investigated in different manners.

[1] www.tcmb.gov.tr

Yetkin (2005), evaluated the communication policy of the CBRT within the framework of both institutional and policy implementations for periods before and after 2001, investigated press releases and communication tools evaluation and also measured transparency and expectation management statistically. She found out that transparency was substantially increased after 2001. The effects of the increase in transparency of expectation management were also investigated and the results showed that predictability has increased from the beginning of 2005 and also the long term expectations were acting in harmony with the realizations. Finally, after all detailed investigations, she highlighted that the CBRT had a substantial improvement in the effective communication policy, but there was still more to be done in the future.

Soylu et al. (2014) investigated the effects of the CBRT interest rate releases on financial markets for the period February 2005-April 2013. In the study, Borsa Istanbul 30 Index, USD/TRY, EUR/TRY exchange rates, and both spot and futures daily return series were variables. The results of the study showed that markets respond differently to rate increases and decreases.

Küçükkocaoğlu et al. (2013), explored the response of the banks' stock returns to the CBRT's monetary policy announcements. According to this study, policy rate increases in the MPC days led to important decreases in the stock returns of all individual banks before May 2010, the traditional inflation targeting period before global financial crisis. After the global financial crisis, since the CBRT adopted different policy tools, the authors named this period as the new monetary policy period and found out that, for this period, the aggregate and individual banks stopped responding significantly to the surprises in MPC days due to flexible timing and decisions, actions made in other days beside MPC days.

Duran et al. (2010) examined the CBRT monetary policy decisions' effects on the Turkish stock market. The results of this study were coherent with the literature: increases in the policy rates lead to decreases in stock prices at varying rates according to the sectors. The authors emphasized that this study showed strong evidence of the transmission from monetary policy to capital markets.

In her study, where she examined the transparency in monetary policy and management of inflationary expectations, Kansu (2007) represented the important developments in directing inflationary expectations and remarked that in the fight against inflation, the importance of transparency in the communication of the CBRT could be understood by comparing before and after 2001. She emphasized that there was an undeniable improvement.

Başkaya et al. (2008) analyzed the inflation expectations' behavioral aspects by using the data from survey respondents and drew implications about the monetary policy communication strategy of the CBRT. In this study, they found out that there are sectoral heterogeneities in the formation of expectations. The study produces evidence of central bank communication's effects on the expectation formation mechanism. It is found out that, after the publication of monthly inflation developments, financial sector becomes inattentive to the short-term inflation surprises. Also, they suggested that while designing an effective communication policy sectoral heterogeneities in the inflation expectation can be important.

Demiralp et al. (2012) investigated the effectiveness of monetary policy communication of the CBRT by quantifying the MPC announcements right after the MPC decision. In their study, they quantified the signal for the next interest rate decision and examined the communication's effect on predictability. Also, by identifying the surprise component of communication, they explored its effects on the term structure of interest rates. According to this study, the role of statements has strengthened after the full adoption of Inflation Targeting Regime. Moreover, they found out that the relative effect of communication over the yield curve has increased in time and the yield curve responded surprising policy statements significantly. They underlined that the written statements on interest rate decisions are key tools under an inflation targeting regime.

2.3 Text Mining and Central Banks

From the literature, it can be observed that central bank communication, as well as other analyzes regarding central banks, are generally studied with quantitative methods such as econometrics and statistical methods. Also, generally, central banks themselves use those more conventional methods.

However, there is a growing literature on text mining usage in both macroeconomic analyses and, in particular, central bank analyses by both other analysts and central banks.

Krukovets (2020) investigated data science opportunities in central banks in his study. For text analysis, he defined text usage as an alternative information source and listed the possible usages of text mining as follows: monitoring news to predict macroeconomic, financial series and shocks, credibility, transparency, social interactions index based on news, estimation of expectations, hot topics in the economy, usage of social network data, internal communication analysis at central banks to increase transparency, topic dynamics, etc.

In another similar study, Ghirelli et al. (2021) studied new data sources for central banks. In the paper, it is emphasized that besides central banks' usage of structured micro and macro data, new data sources have emerged with recent technological developments. Those new sources are generally highly frequent and they could be structured or unstructured (news, social media, etc.). They illustrated successful studies which are mainly the examples of text data usage.

While reviewing the literature, it could be observed that descriptive statistics such as words, phrases, sentence frequencies of text sources, readability-formality scores of texts, sentiment analysis, unsupervised machine learning methods such as clustering, topic modeling with Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) and supervised machine learning methods Naïve Bayes classification, SVM are used widely by central banks for text mining to analyze central bank texts.

Bholat et al. (2015) remarked that although the wide usage of text mining in other fields, it is historically less used in economics. However, they defended that text mining could be worthy since it makes many important data sources tractable. This way, monetary and financial stability can be assessed from these data sources which cannot be analyzed by quantitative methods. According to their study, key text data sources are news, articles, social media, reports, etc. In this study, Bholat et al. focused on text preparation process and usage of various text mining methods which are Boolean and dictionary techniques, unsupervised machine learning techniques LSA, LDA, Descending Hierarchical Classification, and also supervised machine learning technique which is Naïve Bayesian. By giving examples about the usage of those techniques by central banks, they indicated that for central banks, text mining can be used for measuring risk and uncertainty, the consistency of its communication, sentiment, topics, effects of transparency on markets, etc.

From the literature, especially the studies using text mining for central bank communication analysis are illustrated in the remaining of this section.

One of the first studies of applying LSA to central bank communication was conducted by Boukus and Rosenberg (2006) to the Federal Reserve (FED) communication for the period of 1987-2005 in order to be able to analyze the relationship between the themes in the FOMC minutes and current-future economic conditions. Also, they measured the FOMC communication effects on markets' expectations about future policy rates. The results of the study showed that the volatilities of different maturities of Treasury yields react significantly to the communication on current and future financial markets. Economic conditions and market participants can receive complex, multidimensional signals from the FOMC minutes.

Communication statements of the Bank of Canada are investigated by Hendry and Madeley (2010) by using LSA to extract information and investigate the information type affecting short-term return, volatility, and long-term interest rate for 2002-2008. This study remarked that discussions of the major shocks affecting Canada, the balance of risk to economic projection, and the forward-looking statements are the focus of markets to respond.

In another study, Bank of Canada's monetary policy reports are investigated by Binette and Tchebotarev (2019) by looking at the language (most frequent words), readability level, length, and sentiment.

Apel and Grimaldi (2012) measured the sentiment and tone of the Swedish Central Bank by using an automated search and word counting approach. With this approach, hawkishness and dovishness of the Riksbank's minutes are tried to be captured. Moreover, it is found out that this measure is useful for predicting future policy decisions.

Hansen and McMahon (2014) analyzed the internal deliberation of the FOMC and the effect of transparency on policy makers' deliberations by using character counting method and LDA for text qualification. In another study, Hansen and McMahon (2016) quantified the FOMC communication about monetary policy by using computational linguistic approaches. By applying LDA and dictionary methods, they extracted the content of the FED interest rate statements. The results of their study showed that over the period 1998-2014, future guidance on interest rates has been more important than communication about economic conditions.

Moniz and Jong (2014) designed an automated system that predicts the impacts of the Bank of England's communication on investors' interest rate expectations. For this system; TextRank for capturing an outstanding aspect of minutes, Naïve Bayesian model for classification, LDA for topic modeling, and ensemble tree for predicting the impact of communications were used.

In his study, Bruno (2016) reviewed some of the main methodologies in text mining and applied those techniques to the Bank of Italy Governor's concluding remarks for the period of 1996-2015 with R. In that study, with text mining techniques, features of the word frequency distributions, sentiment orientation and polarity value, readability and formality scores, and interest/memorability from documents were analyzed. He stated that the main tools for text mining are Vector Space Model (VSM), LSA, and LDA. By focusing on mainly the applications of VSM, it is concluded that the last 20 editions of Concluding Remarks stay neutral. Furthermore, readability index is in the range of 12-15, which is college degree, and formality score is quite high as expected.

Finally, it is found that English versions of the Concluding Remarks hit similar with of the ECB and the FED. Bruno emphasized that, considering the literature, his paper is the first application of these measures to institutional official reports and the objectivity of those figures is a strong ground for increasing accountability and transparency of central bank communication. In another paper, Bruno analyzed the Bank of Italy Financial Stability reports with similar methods (2017).

Tobback et al. (2017) presented Hawkish-Dovish indicator measuring the hawkishness-dovishness degree of the media's perception of the ECB's press conference tone. To construct the indicator, they used semantic orientation and SVM for text classification. Furthermore, to exemplify the usage of indicator, correlation analysis with interest rates and topic analysis with LDA were conducted besides other further analyses.

A supervised machine learning framework presented by Rybinski (2019) to analyze the interactions between official central bank communication of Narodowy Bank Polski and newspaper articles. In the study, dictionary-based method and Wordscores model were used. It is found out that in the areas that central bank is the key decision maker, such as inflation and interest rates, central bank has power to affect media. However, in the areas that central bank is not the key decision maker, the power is limited.

In addition to other central bank communication examinations with text mining, Tumalo and Omotosho (2019) examined the Central Bank of Nigeria's communication by investigating readability, sentiment, and topic modeling with LDA. A similar investigation of the Central Bank of Ghana's communication is conducted by Omotosho (2019).

Benchimol et al. (2022) examined the Bank of Israel's monthly communication and presented a step by step application of text mining methods in R. They examined frequency, word association, sentiment, clustering, LSA, and topic modeling with LDA.

There are several other studies examining other central banks' communication with different text mining methods.

Shirota et al. (2015) studied monetary policy minutes in Japan; Luangaram and Wongwachara (2017) investigated 22 central banks' communication (including the CBRT) in three aspects: readability, topics, and tone; Mathur and Sengupta (2019) studied the monetary policy statements of the Reserve Bank of India; Máté et al. (2021) analyzed effects of central bank communication on sovereign yields for Hungary; Bailliu et al. (2021) used text analytics to predict and interpret the unknown component in the monetary policy of the People's Bank of China.

For the CBRT, there are studies conducted with text mining. Kahveci and Odabaş (2016) investigated the tone of the FED, the ECB, and the CBRT's monetary policy statements for pre and post crisis periods from 2002 to 2015 with Diction 7 software. With this investigation, it is found out that the optimistic tone of the FED has decreased over time while certainty tone has increased and the CBRT and the ECB had no difference in certainty, optimism, and realism tone although the CBRT had a significant increase in the optimistic tone for the last two years.

In his master's thesis, Ermiş (2017) examined the CBRT's monetary policy statements, minutes, and speeches of governors' effects on the Borsa Istanbul 100 index for the period 2006-2017. Text mining methods and decision tree algorithm are applied to those communication tools. The results remarked that the governors' speeches touch on global and financial situation, banking sector, and interest rate policy can affect markets, but in general, the speeches of governors are not effective. When it comes to monetary policy statements and minutes, again, it is found that the effects on markets are weak.

Emekçi (2017) explored, mainly with R programming, the effects of the CBRT's press announcements between 01.01.2006 and 01.01.2017 on financial markets by determining author/author group announcing statements and the identification of the attitudes of author/author group in his Ph.D. thesis. To obtain author group information, clustering methods were used and for the sentiment analysis, word-based and sentence-based sentiment analysis conducted with lexicon–based methods and machine learning respectively. The results indicated that the announcements were written by three different groups and the announcements' stylistics do not change due to chairman or committee member changes.

Moreover, the examination of polarity, i.e., sentiment results, was used to model exchange rate, stock price index, and bond rate. It is found out that the negative announcements have more effects on financial markets than the positive announcements. Thus, it is remarked that the communication tone of the CBRT has important effects on the volatility of stock market, bond market, and exchange rate.

Another study focusing on the CBRT's communication is conducted by Iglesias et al. (2017). They analyzed the statements and minutes for the period of 2006-2017. With Dynamic Topic Model method, selected topics and their changes in time are examined. In addition, sentiment analysis of those topics are conducted based on the lexicon method and the interconnectedness of topics and impacts of communication strategy on the markets are examined. The resulted topics from Dynamic Topic Model were grouped into 7 different groups: foreign exchange and liquidity, inflation non-core, monetary policy, economic activity, labor market, global flows balance of payment, fiscal and structural, and others. The results of the study pointed out that the topics have been changing over time as global conditions change. Economic conditions and capital flow volatility become prominent during the crisis, replacing the traditional discussions on inflation. Moreover, the complexity of monetary policy strategy has increased over time with the rise in capital flow volatility because of financial crisis and macroprudential policies' being complementary to the monetary policy. Also, the inflation topic is increasingly important according to their study. Finally, it is found out that the communication of the CBRT affects the expectations of the future path of monetary policy, while there are fewer certain results for the impacts on inflation and real variables.

A recent study on the CBRT's communication used deep learning, which is not in the scope of this thesis. Kütük (2021) applied statistical natural language processing tools to the CBRT's monetary policy minutes and used deep learning to measure sentiment and estimate the semantic scores of future minutes.

In short, the literature pointed out that text mining is used for extracting descriptive statistics, sentiment, readability-formality scores, and topics of central banks' communication and also for classifying them.

Moreover, after quantifying communication, the relationship with other economic variables and the effects of communication on markets are studied widely. The most used techniques are word-phrase counts; lexicon-based sentiment analysis; unsupervised machine learning methods such as hierarchical clustering, LSA, LDA; and supervised machine learning methods such as Naïve Bayes and SVM. For the communication of the CBRT, there are studies conducted with text mining however there is still much room which is the inspiration of this thesis.

3. METHODOLOGY

In this chapter, the steps and methods applied in this thesis will be introduced. To investigate MPC decisions, widely-used and accepted methods from the literature are used in this thesis. Mainly, methods from "Text Mining for Central Banks" study by Bholat et al. (2015), "Text Mining and Sentiment Extraction in Central Bank Documents" study by Bruno (2016), and "Text Mining Methodologies with R: An Application to Central Bank Texts" study of Benchimol et al. (2022) are used in this thesis.

Text mining process is mainly consisted of the following steps:

1) Collection of information & creating corpus
2) Text preprocessing
3) Transformation
4) Analysis (Sentiment Analysis, Machine Learning, etc.)

The remaining of this chapter is coherent with those steps. After data collection, firstly, the text preprocessing methods will be on stage. Then, document term matrix will be explained. Word clouds and word association will be explained as methods to gain descriptive inference and insights from text. The formulas of readability and formality scores will be described and the methodology of sentiment-polarity analysis will be explained in detail. Finally, supervised and unsupervised learning methods used in this thesis will also be introduced. In addition, since R programming, which is a software for statistical analysis and graphics developed by Robert Gentleman and Ross Ihaka in 1995, is used for analysis; related R packages will be specified in this chapter besides general data mining, data visualization related packages such as "dplyr" (Wickham et al., 2022), "ggplot2" (Wickham, 2021) packages. Throughout the analysis, the 4.0.3 version of R and the 1.3.1093 version of R Studio are used.

3.2 Collection of Information & Creating Corpus

Any written source, e.g., text sources such as books, articles, websites, e-mails, news, etc., can be used as information and collected for text mining.

Corpus is defined as “a collection of natural language (text, and/or transcriptions of speech or signs) constructed with a specific purpose” (Nilsson-Björkenstam, 2013, p.2). In this thesis, to read text files, the 0.81 version of “readtext” package (Benoit et al., 2021) is used.

3.2 Text Preprocessing

Text preprocessing is one of the most important steps of text mining process. In this step, the raw data is prepared for efficient analysis. Generally, text data includes numbers, stop words, dates, etc., which creates noise and complexity. The aim of preprocessing is to prepare unstructured raw data for analysis by converting it into a semi-structured or structured format to increase efficiency. For this conversion and increasing efficiency, trivial and non-informative data extraction, stemming, removing numbers, etc., are used to clean and prepare data.

In addition, some of the preprocessing techniques uniformize the words. For example, lowercasing and stemming are evaluated as a “normalization process” which makes words identical when the same word is written slightly different (Welbers et al., 2017).

Feinerer et al. (2008) indicated that many insignificant stop words or inconvenient formats of raw text can be a problem in analysis, and thus preprocessing, which is the application for cleaning up and structuring input text that is a core component of text mining.

In this thesis, for text preprocessing steps, mainly the 0.7-8 version of “tm” package (Feinerer and Hornik, 2020) is used and in some cases, for further string operations, the 1.4.0 version of “stringr” package (Wickham, 2019) is also used.

3.2.1 Removing Punctuation, Symbols, Digits and White Spaces

By nature, text data can contain many punctuation marks, symbols, numbers, white spaces, dates, and different characters. However, involving those could be a problem and cause misleading analysis, such as character count and frequency analysis. Also, they may increase the volume of data and noise in the data. Moreover, because the aim is getting information from text data, punctuation, symbols, digits, and extra white spaces cannot create additional value.

3.2.2 Lowercase Conversion

All raw data is converted to lowercase to avoid different identifications of the same words in the analysis. For example, “Bank”, “BANK” and “bank” could seem like different words without lowercase conversion.

3.2.3 Removing Stop Words, Other Special Words, and Rephrasing Words

Stop words are defined as “words in a document that are frequently occurring but meaningless, in other words, their discrimination values are very low, and the information carried by those are negligible” (Lo et al., 2005, p.17). “And”, “the”, and “is” are examples of stop words. From the examples, it can be understood that their information value for the content is almost zero.

Kannan and Gurusamy (2014) highlighted that due to high frequency of stop words, their presence creates an obstacle to understanding the content of text documents. To remove stop words, a pre-defined list can be used.

Furthermore, other common words in a corpus or special words having no information for the specific analysis can be removed from the corpus to improve efficiency. Also, some words are rephrased to uniform those words. To illustrate, “ECB” and “European Central Bank” have the same meaning, while their written formats are different. If their spelling is not configured to each other, in the analysis it is not possible to identify them as the same. Thus, one of them could be changed to other spelling.

3.2.4 Stemming

Feinerer et al. defined stemming as “the process of erasing word suffixes to retrieve their radicals” (2008, p.24). The results of stemming do not have to be meaningful words. There may be meaningless roots in the result. For example, “continued”, “continues”, “continue” converted to “continu”. The main aim is to reduce the complexity of the data and increase efficiency.

There are different algorithms for stemming. In tm package, for the English language, Porter’s stemming algorithm is used. It is one of the most common algorithms in use. Porter released this algorithm in 1980 for the first time.

Porter defined the algorithm as “a process for removing the commoner morphological and inflexional endings from words in English” (2006). The aim of stemming does not produce a real word. According to Porter, the aim is to bring different forms of a word together. For the stemming process, there are different rules in the algorithm regarding the endings of the words, and the algorithm is worked accordingly.

3.2.5 Tokenization

Mullen et al. defined the “token” as following: “Segments of text identified as meaningful units for the purpose of analyzing the text” (2018, p.1). In other words, it can be defined as the individual units of meaning in the analysis according to Mohler. (2020) Mohler also indicated that tokens can be words, sentences etc., and defined “tokenization” as the process of breaking text into tokens. In addition, tokens can be n-gram which is contiguous words constituting a phrase (Cheng et al., 2006). N-grams with two words named bigrams, three words trigrams, and so on. For tokenization, the 0.3.2 version of “tidytext” package (Silge & Robinson, 2021) is used.

3.3 Vectorization of Text

Vectorization process is converting raw data to vectors of real numbers. (Jha, 2021) It is applicable for different data types and used long ago. As the result of vectorization, raw data can be presented in vector format.

Similarly, text data is converted to numbers to make further analysis. Chen (2020) defined text vectorization as "the process of converting text into numerical representation". Bengfort et al. (2018) emphasized the importance of vectorization for machine learning applications. They indicated that machine learning algorithms expect input as a two-dimensional array (rows as instances, columns as features) and operate on numerical future space. Thus, the transformation of the text process which is named vectorization or feature extraction is an essential step.

There are different text vectorization methods to represent text data in vector space. Some of those methods consider only the frequency of terms not an order of them and consider only the appearance of text such as bag of words; document term matrix (DTM) etc., and some of them also consider the semantic parsing such as "Word2Vec". For this thesis, DTM and as an extension of that term frequency-inverse document frequency (TF-IDF) will be used.

On the side, this vectorization step can be considered as a transformation step in the text mining process. The 0.7-8 version of "tm" package (Feinerer and Hornik, 2020) is used in this step again.

3.3.1 Document Term Matrix (DTM)

A document term matrix (DTM) is a matrix in which rows correspond to documents, columns correspond to terms and cells are term frequencies. Welbers et al. (2017) stated that DTM is one of the most widespread formats of text representation in a bag of words format. They also underlined that the advantage of DTM usage is about its allowing for analysis with vector and matrix algebra. Also, the use of special matrix formats and sparse matrices, which are memory efficient, are other advantages of DTM.

An example of DTM is below, the grey areas correspond to the term frequency of a related term in the related document:

	Term 1	Term 2	Term 3	Term 4	….	Term n
Document 1						
Document 2						
Document 3						
…..						
Document n						

Figure 1. Document Term Matrix (DTM) Representation

With term frequencies approach, the words with higher frequency are accepted as important.

3.3.2 Term Frequency – Inverse Document Frequency (TF-IDF)

In a corpus, there could be common words which have no importance. Besides, some rare words could be more important. Thus, taking term frequency into account can be misleading. Term Frequency – Inverse Document Matrix (TF-IDF) is useful for this distinction. It is a statistical approach for text data information retrieval, and it is used for the identification of term importance in a corpus (Thakkar and Chaudhari, 2020).

The formula of TF-IDF is defined as follows:

TF-IDF = *Term Frequency (TF) x Inverse Document Frequency (IDF)* where

$$\text{TF} = \frac{Number\ of\ Occurence\ of\ Term\ t\ in\ a\ Given\ Document}{Total\ Number\ of\ Terms\ in\ the\ Document}$$

$$\text{IDF} = \log \frac{Total\ Number\ of\ Documents\ in\ a\ Given\ Corpus}{Number\ of\ Documents\ which\ Given\ Term\ Appears}$$

According to the formula, it can be deduced that if a term occurs in every document of a corpus, its IDF is 0 and if a term occurs in a few documents its IDF is higher. Thus, the importance of very common terms will be lowered which is desirable (Nguyen, 2014). In addition, DTM can be constructed with TF-IDF weighting instead of term frequencies.

3.4 Data Exploration

Before moving into deeper analysis, exploration of data can be informative. For this thesis, beside term frequencies, bigrams, TF-IDF scores from data preparation process, word clouds and word associations are used for data exploration.

3.4.1 Word Clouds

Word clouds are used as a visualization method in text mining. It is a way of text summarization and is used in different contexts to provide an overview by separating words with higher frequency (Heimerl et al., 2014).

Most frequent words appear bigger while the size of words shrinks with lower frequency. Bigrams, trigrams can be represented in word clouds beside unigram words. For word clouds, the 2.6 version of "wordcloud" package (Fellows, 2018) is used.

3.4.2 Word Association

Another common approach is finding word association. For this purpose, the 0.7-8 version of "tm" package (Feinerer and Hornik, 2020) has a function called "findAssocs" which calculates the association for a given term and correlation limit. This function calculates correlations between all terms in DTM and filters the ones higher than the correlation limit. For a given word, the function calculates its correlation with every other word in DTM and gives a score between 0 and 1. The highest score means more appearance of those two words in documents. The function calculates by document level. For the given term, the other terms' association in the document is examined for all documents containing that given term. Documents that do not contain the given term are ignored. (DataCamp, n.d) Furthermore, the association between words can be visualized with association graphs and correlation maps.

3.5 Readability Score

Automatic Readability Index (ARI) is a measurement designed for gauging the readability levels of a text. ARI was introduced by Senter and Smith in 1967 for English texts. The formula of ARI is below:

$$ARI = 4.71 \times \frac{nchars}{nwords} + 0.5 \times \frac{nwords}{nsentences} - 21.43$$ where

nchars is the number of characters, *nwords* is the number of words and *nsentences* is the number of sentences.

From the formula, it is figured out that the complexity of the text is tied to average word and sentence length.

The result of ARI is a predicted United States grade level to fully understand the given text. With an increase in ARI, i.e. the complexity of the text, this grade level becomes higher. Table 1 shows the ARI scores and corresponding United States Grade Levels.

Table 1

ARI Score and Grade Level

Score	Grade Level
1	Kindergarten
2	First Grade
3	Second Grade
4	Third Grade
5	Fourth Grade
6	Fifth Grade
7	Sixth Grade
8	Seventh Grade
9	Eighth Grade
10	Ninth Grade
11	Tenth Grade
12	Eleventh Grade
13	Twelfth Grade
14	College Student

For this thesis, the 2.4.3 version of "qdap" package (Rinker, 2020) is used to calculate automatic readability index without any preprocessing of text through directly raw data by splitting data into sentences.

3.6 Formality Score

Another important feature of text is formality. To measure formality, Heylighen and Dewaele's (2002) formality score (F-score) formula is used. Heylighen and Dewaele (2002) measured formality by focusing on the occurrence of the different classes of words such as verbs, nouns, adjectives etc., for different languages including English.

The formula is below:

$F = 50$ x $(\frac{n_f - n_c}{N} + 1)$ where

$f = \{nouns, adjectives, preposition, article\}$ $\quad n_f = |f|,$

$c = \{pronoun, adverb, verb, interjection\}$ $\quad n_c = |c|,$

$N = \sum (f + c + conjunctions)$

The result of F-score is between 0-100 and a higher score means higher formality. More noun and adjective usage increase the formality as it can be seen from the formula. In their study, Heylighen and Dewaele (2002) found English phone conversations' F-score of 36 and English informational writing' score of 61.

Again, the 2.4.3 version of "qdap" package (Rinker, 2020) is used to calculate F-score without any preprocessing of text through directly raw data in this thesis.

3.7 Sentiment Analysis

Sentiment analysis is defined as "the computational treatment of opinions, sentiments and subjectivity of text" by Medhat et al. (2014, p.1093) and it is emphasized that sentiment analysis is an ongoing text mining research field. According to Liu (2012), sentiment analysis and opinion mining is the study of analyzing people's sentiments, opinions, attitudes, emotions, evaluations of written language and it is one of the most active research fields in natural language processing. It is studied widely in data mining, web mining and text mining. The applications of sentiment analysis are used in many different areas such as marketing, social media analysis, politics, and so on. To illustrate, consumer reviews, Twitter, news, politicians' speeches are all text data samples for sentiment analysis. As for sentiment analysis methods, it can be indicated that there are different methods and different analysis units. Medhat et al. (2014) summarized the classification of sentiment analysis methods in the figure below:

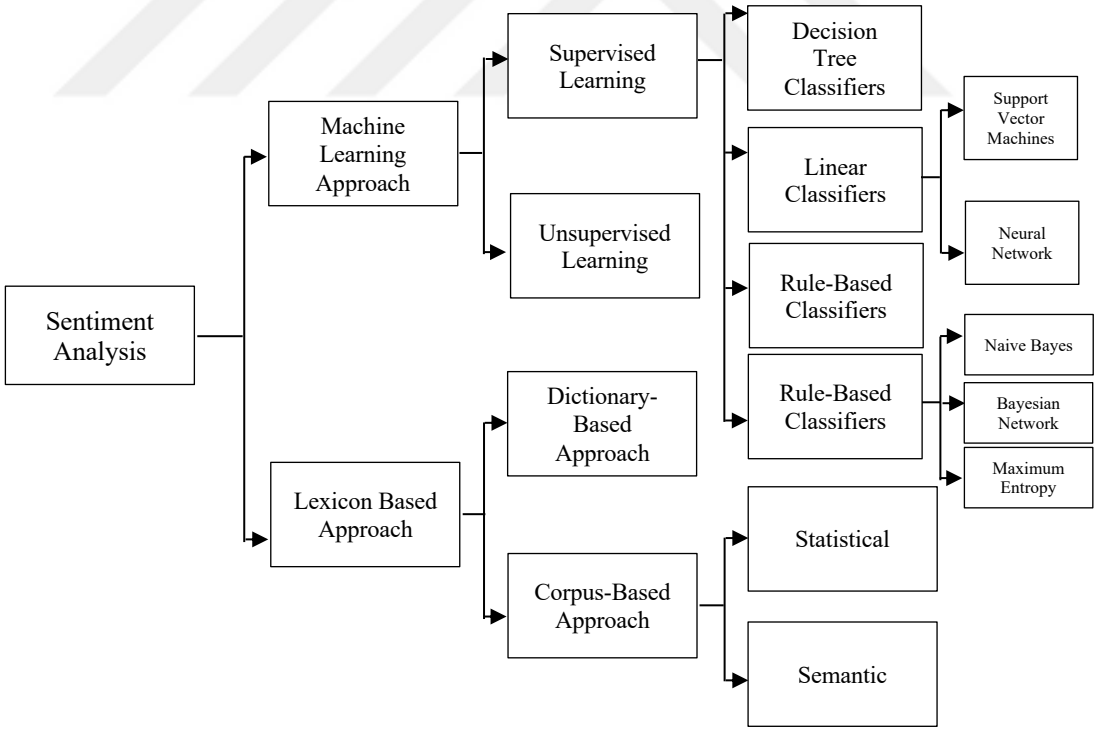


Figure 2. Sentiment Analysis Techniques**.** Adapted from "Sentiment analysis algorithms and applications: A survey." by Medhat, W., Hassan, A., & Korashy, H. (2014). *Ain Shams Engineering Journal*, *5*(4), 1093-1113.

As it can be inferred from Figure 2, there are two main distinct approaches in sentiment analysis: machine learning and lexicon-based approach. Machine learning approach depends on the machine learning algorithms to solve sentiment analysis as a text classification problem by using linguistic, syntactic features while lexicon-based approach depends on a collection of a predefined word list where every word has a polarity score to conduct sentiment analysis (Medhat et al., 2014; Sharma et al., 2020).

In this thesis, dictionary-based approach under the lexicon-based approach is used for sentiment analysis. Dictionary-based approach uses a dictionary consisting of synonyms and antonyms of words (Anees, 2020).

After the construction of dictionary, the words from dictionary are matched with text data. It is a widely used method in sentiment analysis thanks to its simplicity. Moreover, since finding labeled data can be difficult, the dictionary-based method is relatively easy to implement without needing labeled data. However, there are disadvantages of this method, such as the lack of ability to find context-specific words and the importance of choosing an appropriate dictionary.

The dictionary-based analysis starts with a predefined dictionary having words with sentiment polarities. Because constructing a dictionary is a demanding process, there are dictionaries which constructed once and used widely afterward. To illustrate, in the literature the most used dictionary is Harvard Psycho-sociological Dictionary (Benchimol et al., 2022).

Nevertheless, as it is emphasized, dictionary selection is important to avoid misleading results. A dictionary prepared for one profession might not be appropriate for another one. Therefore, Loughran and McDonald (2011) constructed a specific dictionary for financial context. They showed that predefined word lists for other disciplines misclassify common words in financial texts, especially they revealed that three-fourths of the negative words in Harvard Dictionary were not negative in their finance context analysis unit. Thus, they developed negative word lists with five other word lists (positive, uncertainty, litigious, constraining, superfluous) to catch the tone in financial texts. All word lists are in English. In the literature, the Loughran-McDonald (LM) Dictionary is one of the most used dictionaries for textual analysis in finance. In this thesis, LM Dictionary is used as well.

In addition, the level of sentiment analysis can change. Sentence, document, feature, word-level sentiment analysis can be conducted. Both word level and sentence level sentiment analysis are applied in this thesis.

3.7.1 Dictionary-Based Sentiment Analysis at Word Level

Word-level dictionary-based sentiment analysis basically depends on the frequencies of words from predefined word lists from the dictionary. For this application, preprocessing steps are applied, and tokenization is applied at word level.

This analysis adopts "Bag of Words" principles; the frequencies of words are used, word positions are ignored (Welbers et al., 2017).

After getting words and their frequencies from text data, those words are matched with LM Dictionary. Finally, sentiment scores for each document are calculated according to the formula below:

$$Sentiment\ Score\ of\ Document = \frac{Number\ of\ Positive\ Words - Number\ of\ Negative\ Words}{Total\ Word\ Count}$$

for each document

To conduct word-level sentiment analysis, the 0.3.2 version of "tidytext" package (Silge and Robinson, 2021) is used in this thesis.

3.7.2 Dictionary-Based Polarity Scores at Sentence Level

One of the drawbacks of word-based sentiment analysis is its ignorance of valence shifters which are the words changing or intensifying the meanings of words, such as negators and amplifiers. Rinker released "sentimentr" package that has "sentiment" function considers the valence shifters which are negators flipping the sign of polarization of words such as "not", amplifiers (intensifiers) increasing the impact of a polarized word such as "absolutely", de-amplifiers (down-toners) decreasing the impact of a polarized word such as "slightly" and adversative conjunctions dominating the previous statement containing polarized word such as "but".

Rinker indicated that negators appear 20 percent of the time when a polarized word appears in a sentence, and he emphasized not considering valence shifters could have important effects on sentiment. A polarized word means the word has a sentiment polarity in a pre-defined dictionary.

In this method, each paragraph consists of sentences (pi = {s1, s2, …, sn}) and each sentence consists of words (si, j = {w1, w2, …, wn}). After preprocessing, each word in a sentence is compared with the selected dictionary. Positive words are scored as +1, negative words are scored as -1. By default, these polarized words' four words before and two words after are controlled as valence shifters.

After this control, if there is an amplifier, the sentiment increased by 80 percent, and if there is a de-amplifier the sentiment decreased by 80 percent. If there is a negation word in the sentence, the sentiment is multiplied by -1. The formula for polarity score for each sentence is below:

Polarity Score of a Sentence = $\frac{c'_{i,j}}{\sqrt{w_{i,jn}}}$ where

$c'_{i,j}$ *is the sentiment of the sentence*

$w_{i,jn}$ *is the number of total words in the sentence*

$$Sentiment\ Document = Mean(Polarity\ Score\ of\ Sentences\ in\ the\ document)$$

for each document (except the sentences with 0 polarity score)

For sentence-level polarity scores, the 2.9.0 version of "sentimentr" package (Rinker, 2021) and the 1.0.7 version of "qdapDictionaries" package (Rinker, 2013) are used.

3.8 Machine Learning with Text Data

Machine Learning (ML) is concerned with how to build computer programs which are automatically improve with experience (Radovanović and Ivanović, 2008). There has been an upward trend in the usage of machine learning applications in different fields. Text mining is one of those fields where techniques from machine learning are used.

There are different categorizations of machine learning. One of the important divisions of those learning methods is: supervised learning and unsupervised learning.

To define in general terms, unsupervised learning finds groups and structure in data without previously labeled examples, while supervised learning uses previously labeled examples to capture structure and make conclusions (Radovanović and Ivanović, 2008).

In this thesis, different methods of unsupervised and supervised learning are used.

3.8.1 Unsupervised Learning with Text Data

The algorithms of unsupervised learning explore hidden patterns by investigating similarities and differences in data. The most common techniques of unsupervised learning can be listed as clustering such as k-means, hierarchical clustering, topic modeling, association algorithms, dimensionality reduction algorithms such as principal component analysis and neural networks.

For this thesis, hierarchical clustering and topic modeling analysis are performed under unsupervised learning.

3.8.1.1 Hierarchical Clustering

Clustering aims “organizing a collection of data into clusters, such that items within a cluster are more similar to each other” (Grira et al., 2005, p.1). In their study, Grira et al. (2005) also emphasized that the similarity can be expressed in different ways according to the aim of the study, domain-specific assumptions and previous knowledge.

Clustering is used in different areas such as pattern recognition, information retrieval, data mining, and so on. Also, it is used in different sectors from marketing to social media. Furthermore, there are different clustering methods. Partitioning methods, hierarchical clustering (HC), density-based clustering, and model-based clustering are examples of clustering methods.

Within the scope of this thesis, HC is used for clustering. HC aims to get clusters' hierarchy which is named as dendrogram that shows the related clusters. HC creates the hierarchy of clusters based on a proximity measure and HC methods need a pairwise dissimilarity definition (Serra and Tagliaferri, 2019).Euclidean distance, Jaccard distance, and cosine similarity are examples of dissimilarity definitions calculating the distance.

In addition, a linkage method should be selected for HC process. With similarity measures distance between observations can be obtained, and with linkage methods the calculation of the distances and similarities between clusters can be obtained.

Since there are many cases in clusters, calculating only Euclidean distance for example, is not sufficient. The final aim is to find most similar clusters among others in each iteration. Because the aim is to find the similarity between clusters, selecting the right linkage method is substantial. (Jarman, 2020). There are several linkage methods such as single-linkage, complete-linkage, average-linkage, Ward and so on. For this thesis, Ward linkage is used. Ward method considers every possible union of cluster in each step and targets to combine two clusters in a group when the variance is minimum (Vijaya et al., 2019).

Moreover, there are two main strategies for establishing HC: divisive and agglomerative. The divisive method is a top-down approach, splitting large clusters into smaller ones; while the agglomerative method is a bottom-up approach iteratively merging small clusters into larger ones (Grira et al., 2005; Serra and Tagliaferri, 2019). Agglomerative clustering is the most used method, and it starts with n clusters containing one observation. Two most similar clusters are merged in each step and consist of a new level of hierarchy until the last two clusters are merged at the top (Serra and Tagliaferri, 2019).

The result of hierarchical clustering is represented as a tree diagram, dendrogram. An example of a dendrogram can be seen below:

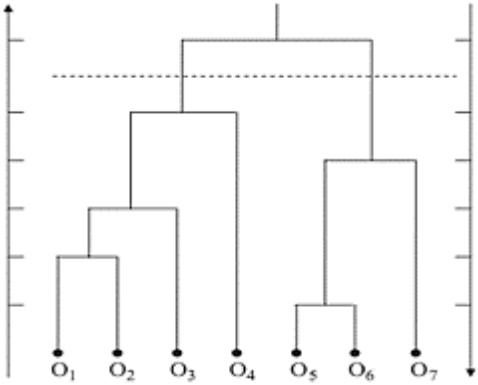


Figure 3. Dendrogram Example. Adapted from "A comparative agglomerative hierarchical clustering method to cluster implemented course" by Sembiring, R. W., Zain, J. M., & Embong, A. (2010). *Journal of Computing 2*(12).

In this thesis, an application of agglomerative HC is on the stage for document clustering. Document clustering is defined by Fung et al. (2009) as "an automatic grouping of text documents into clusters so that documents within a cluster have high similarity in comparison to one another but are dissimilar to documents in other clusters" (p.555). For document clustering, firstly documents are represented in a vector space model which is explained in detail in the 3.3 Section. Again, in the simplest terms, each document is represented with TF vector or TF-IDF weighting in a vector-space model (Steinbach et al., 2000). For this thesis, DTM with TF-IDF weighting is used. After vector-space representation, the similarity between documents is measured. For this measurement, Euclidean distance is used in this thesis. Euclidean distance is the ordinary distance between two points. (Shah and Mahajan, 2012).

For this thesis, the 4.0.3 version of "stats" package (R Core Team, 2020) and the 1.0.7 version of "factoextra" package (Kassambara and Mundt, 2020) are used for constructing HC and visualization.

3.8.1.2 Topic Modeling with Latent Dirichlet Allocation (LDA)

Topic modeling is a probabilistic model for discovering the main topics in the documents. The basic idea behind the topic modeling is treating documents as mixtures of topics and viewing each topic as a probability distribution of words. Topic model views each topic as a collection of words and each document as a set of topics with different proportions depending on the term frequency (Yang and Zhang, 2018). Chauhan and Shah (2021) emphasized the importance of the topic model by underscoring the need for tools to understand a big pool of text and they emphasized that probabilistic topic modeling explains the giant collection of documents by reducing them to a topical subspace.

Latent Dirichlet Allocation (LDA) is described by Blei et al. in 2003 and it is a commonly used technique for text mining. The assumption of LDA is Bag of Words which referred that the order of words that is not important and words exchangeable in a document (Ponweiser, 2012). This assumption results in the DTM representation.

Ponweiser (2012) explained the LDA's generative model with the steps below:

1. "For each topic: Decide what words are likely
2. For each document:
    a. Decide what proportions of topics should be in the document
    b. For each word:
        i. Choose a topic,
        ii. Given this topic, choose a likely word (generated in step 1)". (p.15)

The 0.2.12 version of "topicmodels" package (Grün et al., 2021) is used for LDA in this thesis.

3.8.2 Supervised Learning with Text Data

Different from unsupervised learning, supervised learning trains algorithms with previously labeled data. The most common tasks of supervised learning are "classification" that separates data and "regression" that fits data (Sarker, 2021).

The algorithms of supervised learning need external assistance, and thus data set is divided into train and test datasets; then algorithms learn patterns from training data set and apply test data set for prediction or classification (Mahesh, 2020). Common algorithms of supervised learning are decision trees, Naïve Bayes, SVM, regression analysis.

Throughout the analysis of this thesis, classification task of supervised learning is used with wordscores, Naïve Bayes and SVM algorithms for text classification. Text classification is defined by Ikonomakis et al. (2005) as the task of classifying a text under a predefined category. It is used in many areas such as spam filtering, fraud detection, sentiment analysis and so on.

Moreover, the performance of those algorithms will be compared with a confusion matrix. There are different classification algorithms performance evaluation methods and the discussions in the literature, in the scope of this thesis confusion matrix will be used. Confusion matrix is defined as "table that can be generated for a classifier on a binary data set and can be used to describe the performance of the classifier" (Navin and Pankaja, 2016, p.76).

It is used for binary classification problems and both Naïve Bayes and SVM are binary classification algorithms. An example of confusion matrix can be seen below.

	Predicted: No	**Predicted: Yes**
Actual: No	True Negative (TN)	False Positive (FP)
Actual: Yes	False Negative (FN)	True Positive (TP)

Figure 4. Confusion Matrix Example

There are two classes in the binary classification and the interested class is mainly called as positive. By considering this, true negative means both prediction and actual value are no, i.e. negative classes are predicted correctly. False positive means while actual value is negative class, prediction is positive and false negative is vice versa. True positive means both prediction and actual value are positive. From these measures different performance measures are calculated such as accuracy, precision, prevalence and so on. For this thesis, accuracy measurement will be used. Accuracy shows the correctness frequency of a classifier and has a value between +1 and -1. The best accuracy value is +1. Moreover, sensitivity and specificity metrics are analyzed.

$$Accuracy = \frac{TP+TN}{TN+FP+FN+TP}$$

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Specificity = \frac{TN}{TN+FP}$$

Furthermore, since there is limited small data set, cross-validation method is applied to use as many as observations both in test and train sets. Refaeilzadeh et al. (2009) defined cross validation as a statistical method for evaluation and comparison of the learning models by dividing data to train and test. Cross validation can be seen as a resampling method.

There are different cross validation techniques such as k-fold cross validation, hold-out validation, leave one out cross validation and so on. In data mining and learning, the most common one is 10-fold cross validation (k=10) (Refaeilzadeh et al., 2009). In this thesis, this technique will be used. The process of k-fold cross validation starts with dividing data into k equal portions, i.e. folds and with training – testing is performed with k iterations. Each time a different fold is used for validation and the remaining k-1 folds are used for training.

Additionally, as it is explained, there is a need for labeled data to perform supervised learning. For labeled data, i.e. train set, the MPC labels from the study of Demiralp et al (2012) are used since the authors Demiralp, Özlü and Kara worked with CBRT while scoring the decisions and their study is the only possible labeling for MPC decisions. In their study, they labeled MPC decision statements between February 2002 and July 2010 according to tightening, no change and easing inclination from MPC statements by scoring the statements from +2 to -2. In this thesis, Demiralp et al (2012)'s study's labels for between January 2006 and July 2010 are used.

3.8.2.1 Wordscores

Wordscores was presented by Laver et al. in 2003 and it is a popular method in text analysis. It is a pioneering method of automated content analysis, and it scores or assigns policy positions of documents according to word counts and known document scores (Lowe, 2008). In this method, "reference texts", whose positions or scores are known used as training set for estimating the unknown positions of "virgin texts" (Bruinsma and Gemenis, 2019).

The process starts with estimating the scores for each word type in reference texts and combining these wordscores into virgin texts, and the third step is rescaling the scores of virgin texts to compare them with reference texts more easily (Lowe, 2008). The scores of words are based on word frequency.

In this method selecting reference texts are very important. Laver et al. (2003) underscored the importance of the reference text selection and presented three guidelines for selecting reference text: reference texts should use the same lexicon with virgin texts, ideal positions of reference texts are extreme positions, reference texts should contain as many different words as possible.

Wordscores method is generally criticized by the sensitivity to reference text and the usage of the assumption of Bag of Words which causes its being lack of catching contextual differences only depending on words' appearance and frequency.

To apply wordscores the 0.9.4 version of "quanteda.textmodels" package (Benoit et al., 2021) and the 3.2.0 version of "quanteda" package (Benoit et al., 2021) are used.

3.8.2.2 Naïve Bayes

Naïve Bayes classifiers are known as being efficient and simple. The probabilistic model behind these classifiers is based on Bayes' theorem and the naïve part comes from the assumption that the futures in a dataset are mutually independent (Raschka, 2014). Bayes' theorem is represented with the function below:

$$P\ (A|B) = \frac{P(B|A)\ x\ P(A)}{P(B)}$$

where *P(A) and P(B) are probability of events A and B without regarding each other, P(A|B) is the probability of A conditional on B and P(B|A) is the probability of B conditional on A*

In Naïve Bayes classification, A is categorical outcome events and B is a series of predictors (Zhang, 2016). For text classification, independency means that all words are independent of each other in a document (Mocherla et al., 2017).

Mochearla et al. (2017) also summarized the processes as following: training set is used for building a vocabulary from words in corpus and calculating probabilities for words' belonging to class, afterward those probabilities are used for classifying test documents by applying Bayes rule and class with higher probability becomes decision class.

Thanks to its efficiency and easiness of implementation, Naïve Bayes classifiers are used in many areas as well as text classification problems such as spam detection.

To apply Naïve Bayes the 6.0-90 version of "caret" package (Kuhn et al., 2021) is used.

3.8.2.3 Support Vector Machines (SVM)

SVM is firstly introduced by Vapnik in 1995 and it is proved that it is a strong algorithm for classification and function estimation (Wei et al., 2012). Joachims introduced its usage in text classification in 1997. The main aim of the SVM algorithm is to find a hyperplane in N-dimensional space (N is the number of features) classifying data distinctly (Gandhi, 2018). While separating data there are many possible hyperplanes but the aim is to find one providing maximum distance between points in two classes. SVM generally works in binary classification problems.

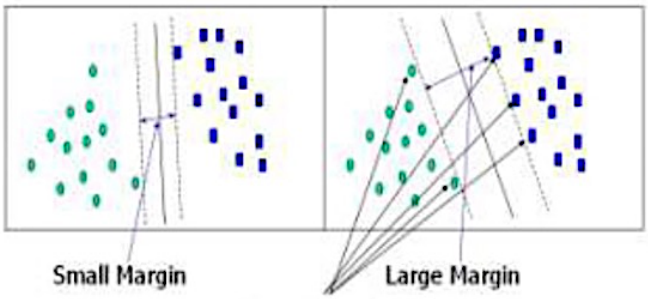


Figure 5. SVM Hyperplane example. Adapted from "A Review on Support Vector Machine for Data Classification " by Bhavsar, H., & Panchal, M. H. (2012). *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET).*

Bhavsar and Panchal (2012) explained that there are two parallel hyperplanes on each side of the hyperplane separating the data. This separating one is the hyperplane maximizing the distance between two parallel ones.

An assumption is that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier. According to them, there can be a classification of linearly separable and non-linear separable data using SVM and SVM has two research areas kernel methods and large margin classifiers. Moreover, SVM is a strong tool in classification, regression, and outline detection. In the literature, SVM's outperformance in comparison to Naïve Bayes is observed in different studies. (Basu et al., 2003)

To apply SVM the 6.0-90 version of "caret" package (Kuhn et al., 2021) is used.

4. DATA, APPLICATION, AND RESULTS

In this chapter, the steps explained in the Methodology part will be applied to the data set and the results will be evaluated. For this aim, the main data set will be introduced, and data preprocessing steps will be performed. After preprocessing, a transformation of text to DTM will be shown and descriptive statistics, which are word frequencies, word clouds and word associations from the main data set will be on the stage. Then, the time series of readability score and formality score will be presented. Afterwards, polarity scores based on word and sentence will be calculated and sentiment will be examined. Finally, selected unsupervised and supervised learning methods will be applied to the main data set.

4.1 Data Set and Creating Corpus

The MPC decision statements are the statements of the CBRT for announcing the rate decisions following an MPC meeting on the same day. It also includes the rationale behind the decision and general assessment of economic developments with forward guidance. The dates of MPC meetings are announced at the beginning of the year, but there could be extraordinary MPC meetings. Currently, 12 pre-scheduled MPC meetings are announced by the CBRT. In the past few years, the number of pre-scheduled MPC meetings was 8.

For this thesis, 172 MPC decision statements between January 2006 and December 2020 are examined. The distribution of MPC decision statements by years can be seen in Table 2.

Table 2

Number of MPC Decision Statements by Years

Year	Number of MPC Decision Statements
2006	13
2007	12
2008	12
2009	12
2010	12
2011	13
2012	12
2013	12
2014	13
2015	12
2016	12
2017	8
2018	9
2019	8
2020	12

The MPC decision statements were obtained from the website[2] of the CBRT. All "pdf" files were converted to "txt" files and each file was named with the same format (YearMonthDay_MPC, for example: 20060124_MPC). Before starting to analyze, the generic parts at the beginning of the decisions about the MPC members, date, and the information part in the final parts of the decisions about the publication of MPC discussion's brief summary were removed. These text files were read into R and a corpus from those text files was created.

[2] www.tcmb.gov.tr

4.2 Data Preprocessing

Data preprocessing steps were applied to the corpus to prepare it for analysis. In this thesis, generally, analyses are performed after the preprocessing and cleaning steps.

Some example sentences from the corpus can be seen below:

Example Sentences:

"The Monetary Policy Committee (the Committee) has decided to set the short-term interest rates at the following levels:

a) Overnight Interest Rates: Marginal Funding Rate has been kept at 9.25 percent and borrowing rate has been kept at 7.25 percent.

b) One-week repo rate has been kept at 8 percent.

c) Late Liquidity Window Interest Rates (between 4:00 p.m. – 5:00 p.m.): Borrowing rate has been kept at 0 percent, while lending rate has been increased from 12.25 percent to 12.75 percent."

4.2.1 Removing Punctuation, Symbols, Digits and White Spaces

The punctuations, symbols such as parenthesis, numbers, and extra spaces were removed from the text.

Step 1 - Example Sentence After Removing Punctuation, Symbols, Digits and White Spaces:

"The Monetary Policy Committee the Committee has decided to set the short term interest rates at the following levels\n Overnight Interest Rates Marginal Funding Rate has been kept at percent and borrowing rate has been kept at percent\n One week repo rate has been kept at percent\n Late Liquidity Window Interest Rates between pm pm Borrowing rate has been kept at percent while lending rate has been increased from percent to percent"

4.2.2 Lowercase Conversion

All text data in the corpus was converted to lowercase.

Step 2 - Example Sentence After Lower Case Conversion:

“the monetary policy committee the committee has decided to set the short term interest rates at the following levels\n overnight interest rates marginal funding rate has been kept at percent and borrowing rate has been kept at percent\n one week repo rate has been kept at percent\n late liquidity window interest rates between pm pm borrowing rate has been kept at percent while lending rate has been increased from percent to percent”

4.2.3 Removing Stop Words, Other Special Words, and Rephrasing Words

In this step, the stop words in English were removed from the data. Moreover, after the observation of common words in the data, "percent", "will", "may", "via", "pm", "am" were also removed since they are similar to stop words in this analysis’ context. In addition, because “cbt” and “cbrt” were used for the same meaning “cbt” was converted to “cbrt” to avoid different representations. After removing stop words and other special words, again, an extra white space removal was applied.

Step 3 - Example Sentence After Removing Stop Words, Other Special Words, and Rephrasing Words:

“monetary policy committee committee decided set short term interest rates following levels overnight interest rates marginal funding rate kept borrowing rate kept one week repo rate kept late liquidity window interest rates borrowing rate kept lending rate increased”

4.2.4 Stemming

According to Porter’s stemming algorithm, stemming was performed.

Step 4 - Example Sentence After Stemming:

“monetari polici committe committe decid set short term interest rate follow level overnight interest rate margin fund rate kept borrow rate kept one week repo rate kept late liquid window interest rate borrow rate kept lend rate increas "

4.2.5 Tokenization

Tokenization was applied by “word” level. Each word is represented in a row.

Step 5 - Tokenization:

Word

monetari

polici

committe

committe

decid

set

short

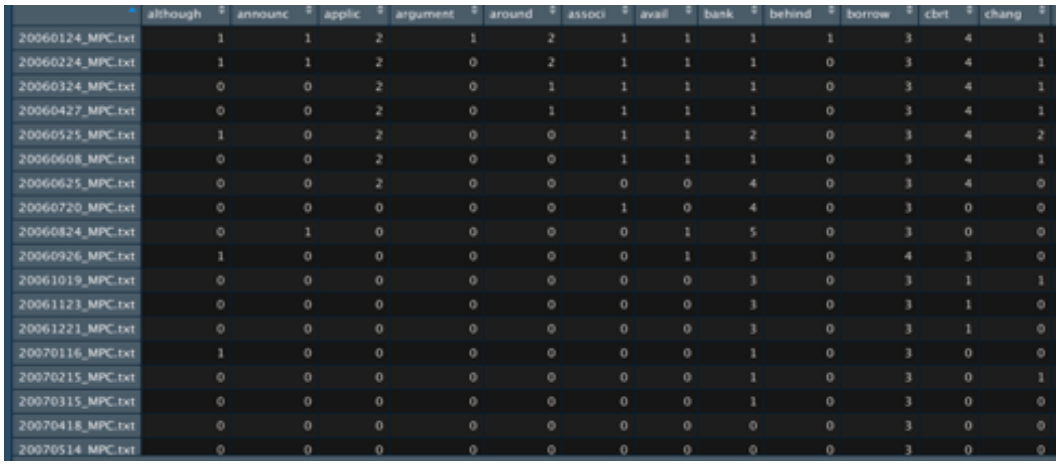
term

interest

rate …

4.3 DTM with Term Frequency and TF-IDF

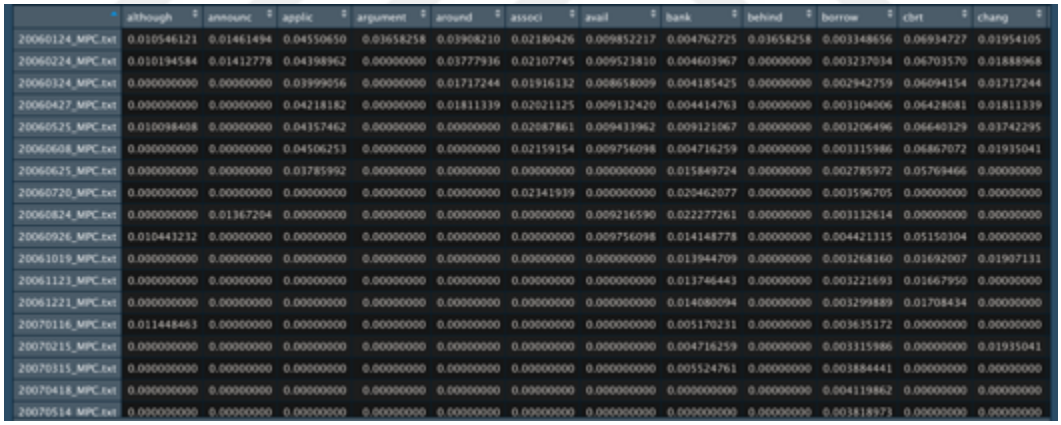
To vectorize corpus and represent text data in matrix, after all preprocessing steps, DTM was created with both term frequency and TF-IDF weighting. Figure 6, a screenshot taken from R program, shows a part of the DTM with term frequency and Figure 7 shows a part of the DTM with TF-IDF.



	although	announc	applic	argument	around	associ	avail	bank	behind	borrow	cbrt	chang
20060124_MPC.txt	1	1	2	1	2	1	1	1	1	3	4	1
20060224_MPC.txt	1	1	2	0	2	1	1	1	0	3	4	1
20060324_MPC.txt	0	0	2	0	1	1	1	1	0	3	4	1
20060427_MPC.txt	0	0	2	0	1	1	1	1	0	3	4	1
20060525_MPC.txt	1	0	2	0	0	1	1	2	0	3	4	2
20060608_MPC.txt	0	0	2	0	0	1	1	1	0	3	4	1
20060625_MPC.txt	0	0	2	0	0	0	0	4	0	3	4	0
20060720_MPC.txt	0	0	0	0	0	1	0	4	0	3	0	0
20060824_MPC.txt	0	1	0	0	0	0	1	5	0	3	0	0
20060926_MPC.txt	1	0	0	0	0	0	1	3	0	4	3	0
20061019_MPC.txt	0	0	0	0	0	0	0	3	0	3	1	1
20061123_MPC.txt	0	0	0	0	0	0	0	3	0	3	1	0
20061221_MPC.txt	0	0	0	0	0	0	0	3	0	3	1	0
20070116_MPC.txt	1	0	0	0	0	0	0	1	0	3	0	0
20070215_MPC.txt	0	0	0	0	0	0	0	1	0	3	0	1
20070315_MPC.txt	0	0	0	0	0	0	0	1	0	3	0	0
20070418_MPC.txt	0	0	0	0	0	0	0	0	0	3	0	0
20070514_MPC.txt	0	0	0	0	0	0	0	0	0	3	0	0

Figure 6. An Example of DTM with Term Frequency from R

DTM in Figure 6 can be interpreted as following: In January 2006 MPC decision statement, “although” is used for one time while it is not used in March 2006 MPC decision statement.



	although	announc	applic	argument	around	associ	avail	bank	behind	borrow	cbrt	chang
20060124_MPC.txt	0.010546121	0.01461494	0.04550650	0.03658258	0.03908210	0.02180426	0.009852217	0.004762725	0.03658258	0.003348656	0.06934727	0.01954105
20060224_MPC.txt	0.010194584	0.01412778	0.04398962	0.00000000	0.03777936	0.02107745	0.009523810	0.004603967	0.00000000	0.003237034	0.06703570	0.01888968
20060324_MPC.txt	0.000000000	0.00000000	0.03999056	0.00000000	0.01717244	0.01916132	0.008658009	0.004185425	0.00000000	0.002942759	0.06094154	0.01717244
20060427_MPC.txt	0.000000000	0.00000000	0.04218182	0.00000000	0.01811339	0.02021125	0.009132420	0.004414763	0.00000000	0.003104006	0.06428081	0.01811339
20060525_MPC.txt	0.010098408	0.00000000	0.04357462	0.00000000	0.00000000	0.02087861	0.009433962	0.009121067	0.00000000	0.003206496	0.06640329	0.03742295
20060608_MPC.txt	0.000000000	0.00000000	0.04506253	0.00000000	0.00000000	0.02159154	0.009756098	0.004716259	0.00000000	0.003315986	0.06867072	0.01935041
20060625_MPC.txt	0.000000000	0.00000000	0.01785992	0.00000000	0.00000000	0.00000000	0.000000000	0.015849724	0.00000000	0.002785972	0.05769466	0.00000000
20060720_MPC.txt	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000	0.02341939	0.000000000	0.020462077	0.00000000	0.003596705	0.00000000	0.00000000
20060824_MPC.txt	0.000000000	0.01367204	0.00000000	0.00000000	0.00000000	0.00000000	0.009216590	0.022277261	0.00000000	0.003132614	0.00000000	0.00000000
20060926_MPC.txt	0.010443232	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.009756098	0.014148778	0.00000000	0.004421315	0.05150304	0.00000000
20061019_MPC.txt	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.000000000	0.013944709	0.00000000	0.003268160	0.01692007	0.01907131
20061123_MPC.txt	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.000000000	0.013746443	0.00000000	0.003221693	0.01667950	0.00000000
20061221_MPC.txt	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.000000000	0.014080094	0.00000000	0.003299889	0.01708434	0.00000000
20070116_MPC.txt	0.011448463	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.000000000	0.005170231	0.00000000	0.003635172	0.00000000	0.00000000
20070215_MPC.txt	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.000000000	0.004716259	0.00000000	0.003315986	0.00000000	0.01935041
20070315_MPC.txt	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.000000000	0.005524761	0.00000000	0.003884441	0.00000000	0.00000000
20070418_MPC.txt	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.000000000	0.000000000	0.00000000	0.004119862	0.00000000	0.00000000
20070514_MPC.txt	0.000000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.000000000	0.000000000	0.00000000	0.003818973	0.00000000	0.00000000

Figure 7. An Example of DTM with TF-IDF from R

As it is explained in detail in Methodology, TF-IDF shows the importance of a word given in a document or corpus. The higher the TF-IDF, the rarer the word which can give more information about the document or corpus. DTM in the Figure 7 shows the TF-IDF values of each word.

4.4 Data Exploration

Before moving into deeper analysis, different data exploration methods were applied in this step.

4.4.1 Descriptive Statistics

Before stop word removal and stemming, in total, there are 43.819 words in corpus, while there are 1.383 unique words. There are approximately 254.76 words per MPC decision statement. The minimum word count is in May 2018 with 128 words. The maximum word count is in May 2020 with 421 words.

Most frequent words and their frequencies are represented in Table 3. As it is expected, stop words and common words such as “rate” are the most frequent words which show the importance of the cleaning process before analysis.

Table 3

The Most Frequent Words Before Stop Word Removal and Stemming

Word	Frequency
the	4668
and	1220
percent	1069
rate	1050
committee	906
inflation	867
policy	746
interest	651
rates	629
that	609

After stop word removal and stemming, there are 31.197 words in corpus in total, while there are 916 unique words. As it is indicated in Methodology, after standard stop word removal, "percent", "will", "may", "via", "pm", "am", and “iii” were also removed since their added value in this context was insignificant.

Moreover, “cbt” was converted to “cbrt” to avoid double counting. After cleaning, there are approximately 181.38 words per MPC decision statement. The minimum word count is in May 2018 decision with 88 words. The maximum word count is in May 2020 with 321 words.

Figure 8 shows the changes of the word counts over time and the dashed lines point out a governor change.

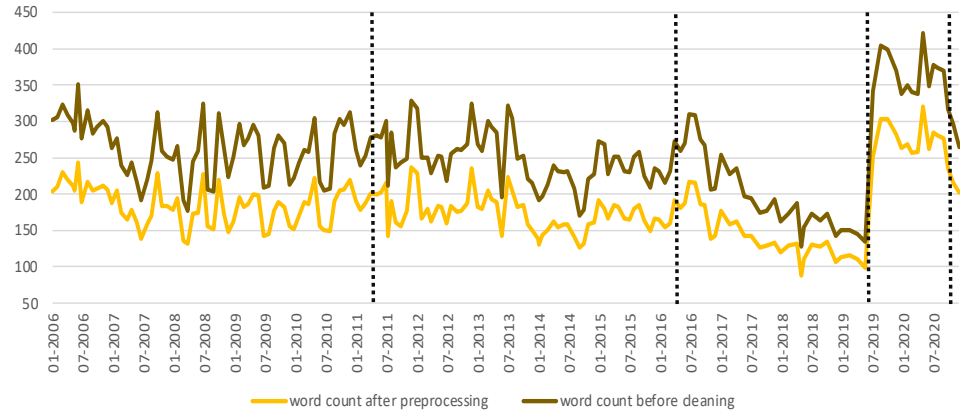


Figure 8. Word Counts

As it can be seen, there is a fluctuated change in the number of words in the MPC decisions. Moreover, there is a decreasing trend from 2017 to 2018 and there is a strong increase in 2019 after governor change. It can be observed that governor changes affect the number of words and hereby the usage of language in the MPC decisions. Especially, after 2016, such change can be seen more clearly.

The final most frequent words are represented in Table 4. Thanks to cleaning steps, stop words and insignificant words are removed and stemmed versions of words are counted.

Table 4

The Most Frequent Words After All Preprocessing Steps

Word	Frequency
rate	1679
committe	906
inflat	867
polici	777
interest	651
monetari	571
demand	430
borrow	427
price	398
continu	395

In addition to word level, the most frequent bigrams are examined. The results are shown in Table 5 and they are consistent with the general announcements in the MPC decision statements.

Table 5

The Most Frequent Bigrams After All Preprocessing Steps

Bigram	Frequency
interest rate	636
monetari polici	454
borrow rate	275
inflat outlook	231
lend rate	231
one week	196
polici rate	190
polici committe	187
new data	171
late liquid	162

Moreover, TF-IDF scores of words can give good descriptive insights about the data. Table 6 and 7 show the words with the highest and lowest TF-IDF scores.

Table 6

Words with the Highest TF-IDF Scores

Word	TF-IDF
kept	1.646
decreas	1.330
cut	1.014
tight	0.981
contribut	0.910
market	0.886
recoveri	0.866
financi	0.859
increas	0.849
stabil	0.803

Table 7

Words with the Lowest TF-IDF Scores

Word	TF-IDF
committee	0
decid	0
monetari	0
one	0
polici	0
rate	0
repo	0
week	0
new	0.008
data	0.014

From Table 6 and 7, it can be referred that the words with lowest TF-IDF scores such as “committee” and “decid” give almost no information about the text since they occur very often (words with 0 TF-IDF score occur all documents in the corpus), while the words with highest TF-IDF scores such as “kept” and “decreas” give important insights about the text since they occur rare.

4.4.2 Word Clouds

Word clouds represent the most frequent words or n-grams in a corpus by showing them with different sizes and colors. Bigger size of a term implies higher frequency. It is an exploration method before deeper analysis with the data is undertaken.

Below, the most frequent 50 words and the most frequent 50 bigrams are shown in Figure 9 and Figure 10, respectively.

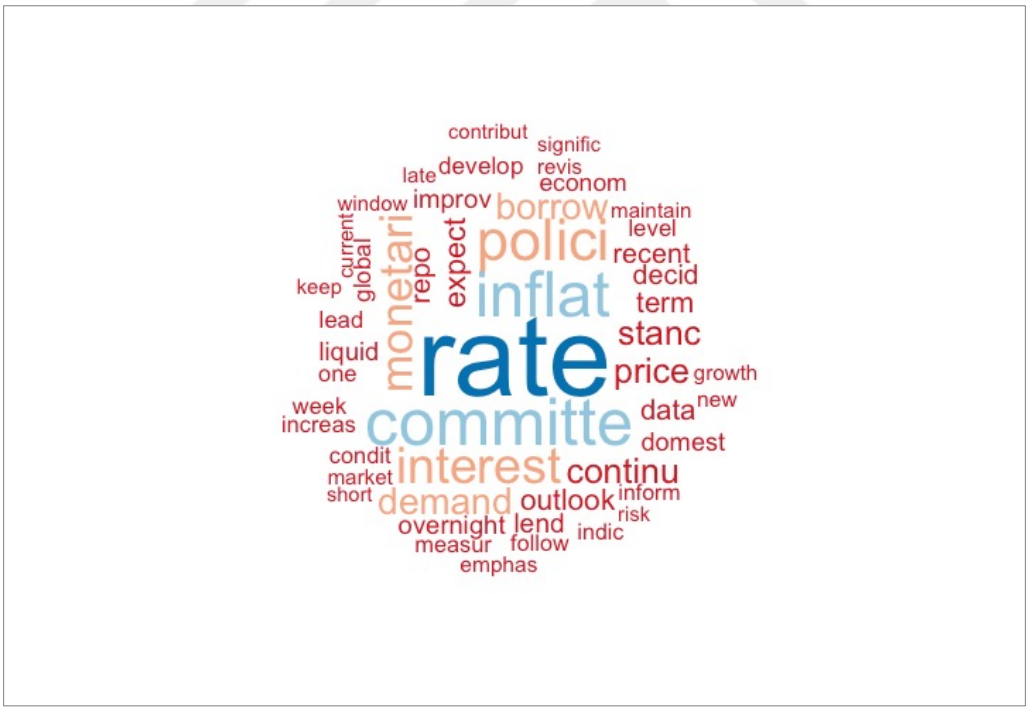


Figure 9. The Most Frequent 50 Words

Figure 10. The Most Frequent 50 Bigrams

As it is expected, word clouds are consistent with descriptive statistics, the most frequent words, and bigrams. Besides, there are more clues about frequent discussions in the decision statements.

4.4.3 Word Association

Word association is another method for exploring data before deeper analysis. It shows the co-occurrence of words in document level.

In Table 8, associations of chosen important words from the decision statements are listed. Associations are calculated from the DTM with term frequency weighting and the correlation limit is 0.5.

Table 8

Word Associations

Word	Associations
inflat	outlook: 0.66 sector: 0.56 service: 0.51
growth	loan: 0.73 macroprudenti: 0.70 contribut: 0.67 reason: 0.66 consum: 0.66 moder: 0.65 respons: 0.64 core: 0.58 stanc: 0.57 account: 0.56 tight: 0.56 cours: 0.54 trade: 0.53 improv: 0.51 margin: 0.50
monetari	stanc:0.70 contribut: 0.66 consum: 0.60 trade: 0.60 loan: 0.56 maintain: 0.52 cautious: 0.51 polici: 0.50
risk	econom: 0.59 disinfl: 0.55 sustain: 0.51 competit: 0.50 protection: 0.50

Association examination can lead to interesting inferences. For example, the association between inflation and service, growth and loan can lead to further research on these issues with more focus.

Moreover, specific topics can be examined, for example, the association with the word "pandem" (stemmed version of pandemic) gives a quick exploration of the data about an extreme situation and its effects on monetary policy.

In the CBRT's decision statements, "pandem" was used in 2020 March to December decisions. The associated words with "pandem" with the correlation limit 0.70 are listed as follows: disinflationari: 0.91, normal: 0.90, diseas: 0.86, sever: 0.83, output: 0.82, phase: 0.80, preval: 0.76, substanti: 0.73, step: 0.72, tourism: 0.72, evolv: 0.71.

4.5 Readability Score

Automatic Readability Index (ARI) was calculated for each decision statement between 2006 & 2020 and the results are shown in Figure 11. In the figure, dashed lines represent the governor changes. For this analysis, without any preprocessing and cleaning, raw data was used by splitting into sentences.

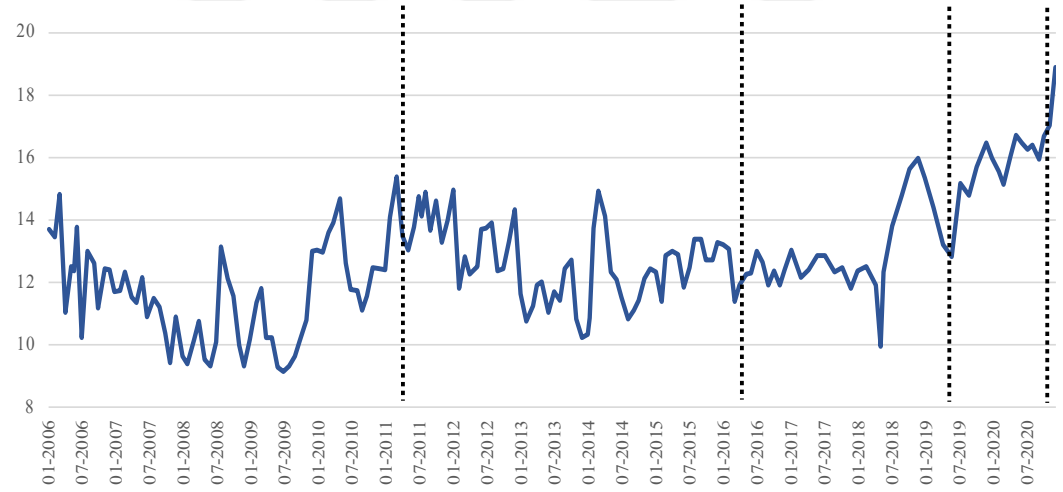


Figure 11. ARI Scores of the CBRT's MPC Decision Statements

Higher score means an increase in the complexity of a text. For the CBRT's decisions, the average ARI score of 172 MPC decisions is 12.62, which is close to college student level. As it can be inferred from Figure 11, there is a fluctuation in the ARI scores during the period 2006-2020. The lowest score was in July 2009 decision with a score of 9.15, while the highest score was in December 2020 decision with a score of 18.89. Since 2019, the ARI score has an increasing trend. To illustrate, in 2020, for the whole year, ARI is more than 15 which is above the college student readability level.

4.6 Formality Score

For the period between 2006-2020, Heylighen and Dewaele's (2002) formality scores (F-score) of the CBRT's MPC decision statements were calculated and the change in years is shown in Figure 12. Again, for this analysis, without any preprocessing and cleaning raw data was used by splitting into sentences.

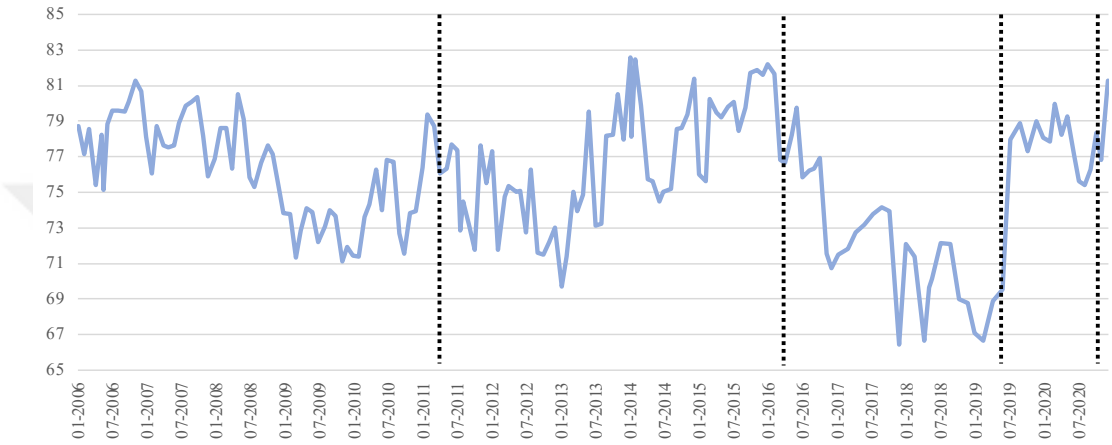


Figure 12. Formality Scores of the CBRT's MPC Decision Statements

F-score changes from 0 to 100 and a higher score means more formality in the language. The average F-score for the CBRT's MPC decision statements is 75.90. The minimum formality score is in December 2017 decision with a score of 66.43, while the maximum formality score is in January 2014 decision with a score of 82.58. Similar to ARI scores, it can be deduced there is highly fluctuated changes in the F-score between 2006 and 2020 from Figure 12. Again, the language changes due to the governor changes may lead to changes in the F-score.

4.7 Sentiment – Polarity Analysis

With Loughran-McDonald (LM) dictionary, a word-based sentiment analysis and a sentence-based polarity analysis were conducted. For both analyses, words with negative and positive sentiment from the dictionary were used. While there are other sentiment groups, they are not considered in the scope of these analyses.

Moreover, for both analyses, “late”, “deficit”, “deficits”, “liquid”, and “cut” were removed from the negative words of the dictionary since those words are used as technical terms in the CBRT’s MPC decision statements.

4.7.1 Dictionary-Based Sentiment Analysis at Word Level

For the word and lexicon-based sentiment analysis, firstly, the data was split into tokens, i.e., words. For this analysis, 31.197 words (repetitive) were obtained after preprocessing. Then, those words’ sentiments were paired with the matching words’ sentiments in the LM dictionary. After getting sentiments, positive and negative words were counted, and the word-based sentiment was calculated according to the formula in 3.7.1.

Figure 13 shows the change in the word-based sentiment of the MPC decision statements in years and dashed lines show governor changes. Positive scores mean an overall positive sentiment in the decision statements and vice versa.

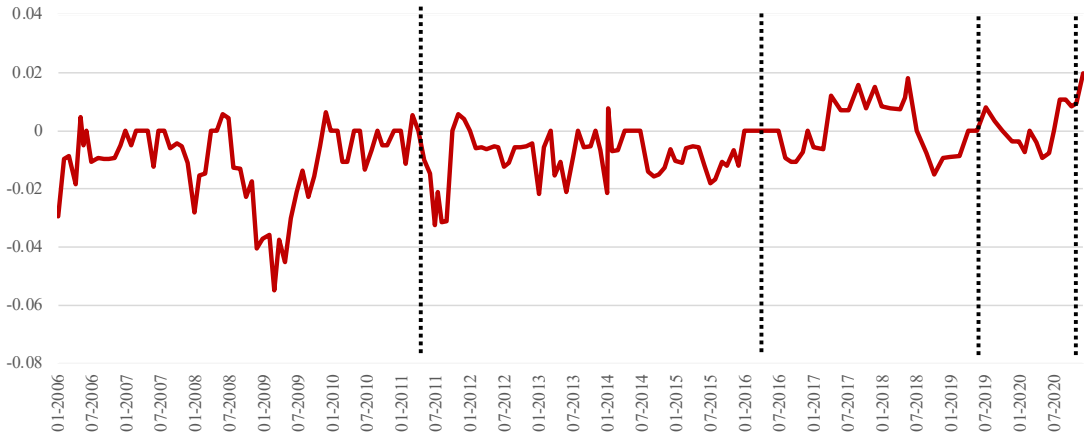


Figure 13. Word Level Sentiment Scores of the CBRT’s MPC Decision Statements

The average word-level sentiment score is -0.0067. The minimum sentiment score is -0.0549 in March 2009 decision. The maximum sentiment score is 0.0198 in December 2020 decision. The sentiment is generally negative for the reviewed period which can be interpreted as the cautiousness of the CBRT for most of the last 15 years. However, there is a movement towards a positive sentiment in the decision statements after 2016.

The most frequent positive and negative words are listed in Figure 14 and Figure 15, respectively.

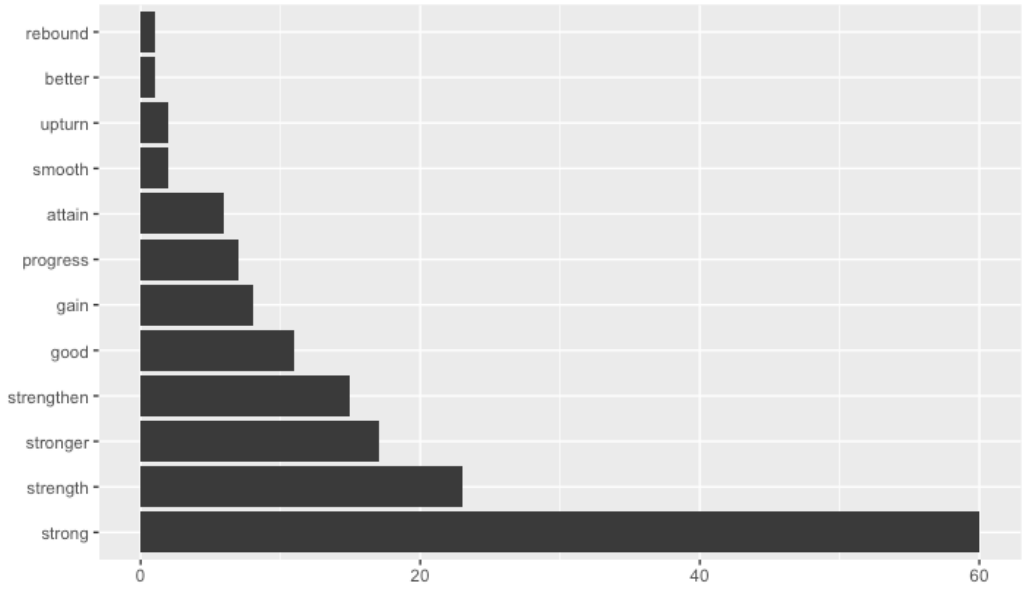


Figure 14. The Most Frequent Positive Words

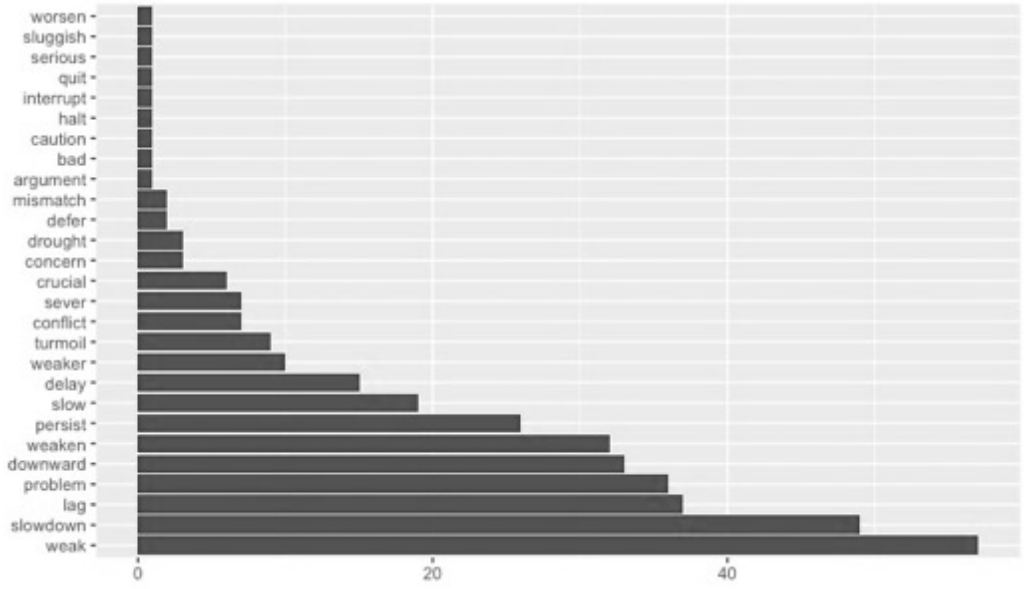


Figure 15. The Most Frequent Negative Words

4.7.2 Dictionary-Based Polarity Scores at Sentence Level

Sentence level polarity scores consider the valence shifters in the sentences which is a criticism of word-level sentiment discussed in the 3.7.2. For this polarity score calculation, firstly, the text data was split into sentences and preprocessing steps were applied. During the preprocessing steps, stop words were updated and the ones containing valence shifters were kept since they were considered calculating polarity scores. Moreover, “less” was added to valence shifters as a deamplification word.

After the preprocessing steps, 2.594 sentences remained and the polarity scores were calculated according to the formula in the 3.7.2. For each MPC decision statement, the average of the polarity scores of sentences, except the sentences with 0 scores, were calculated as the overall polarity score of the statement.

The polarity scores of MPC decision statements in the years and governor changes with dashed lines are shown in Figure 16. Positive polarity scores indicate a general positive sentiment in the decisions and vice versa.

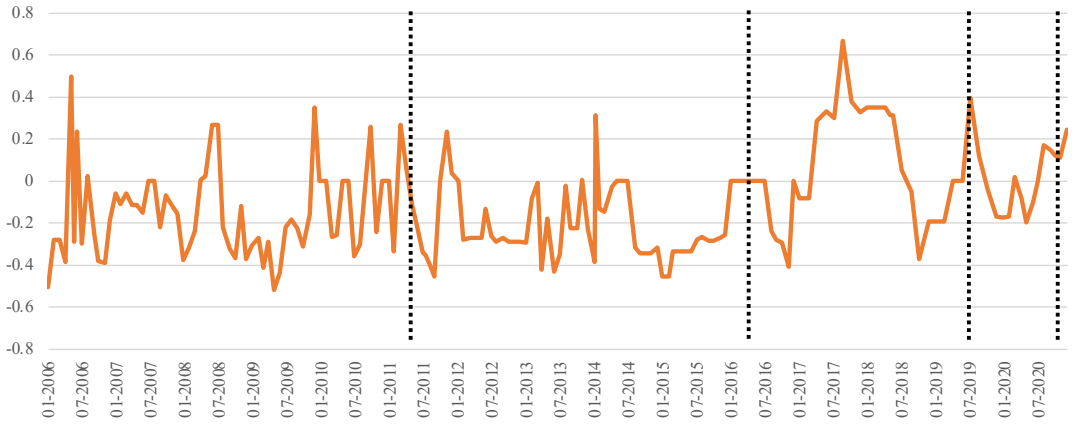


Figure 16. Sentence Level Sentiment Scores of the CBRT’s MPC Decision Statements

The average sentence-level sentiment score is -0.1359. The minimum sentiment score is -0.5185 in March 2009 decision. The maximum sentiment score is 0.6666 in September 2017 decision.

In addition, from Figure 13 and Figure 16, it can be observed that the trends of word and sentence level sentiments do not change. However, sentence-level polarity generally captures the strong movements of the sentiment better.

Similar to the word level sentiment, the MPC decisions' sentiment is generally on the negative side. However, again, an increase in the positive tone can be seen after 2016. In 2019, after the governor change, there is a strong decrease in the sentiment and it is interpreted as a sign of more cautious stance of the CBRT.

4.8 Unsupervised Learning

For unsupervised learning algorithms, HC for document classification and LDA for topic modeling are applied for this thesis. Also, the application processes and results are explained in this section.

4.8.1 Hierarchical Clustering (HC)

Hierarchical clustering aims to get clusters' hierarchy according to a similarity measurement and the results are represented in a dendrogram.

In this thesis, agglomerative HC from a TF-IDF weighted DTM based on Euclidean distance and Ward linkage was used. For HC, the resulted dendrograms show hierarchical clustering. As it is explained in detail in 3.8.1.1, the process starts with n clusters consisting of one observation and two most similar clusters are merged in each step.

Moreover, with the determination of the number of clusters, the clusters can be visualized in dendrogram. There are different technical and heuristic methods for determining the number of clusters. For this thesis, with the domain information and heuristic approach, the number of clusters is determined as 4. The dendrogram is represented in Figure 17.

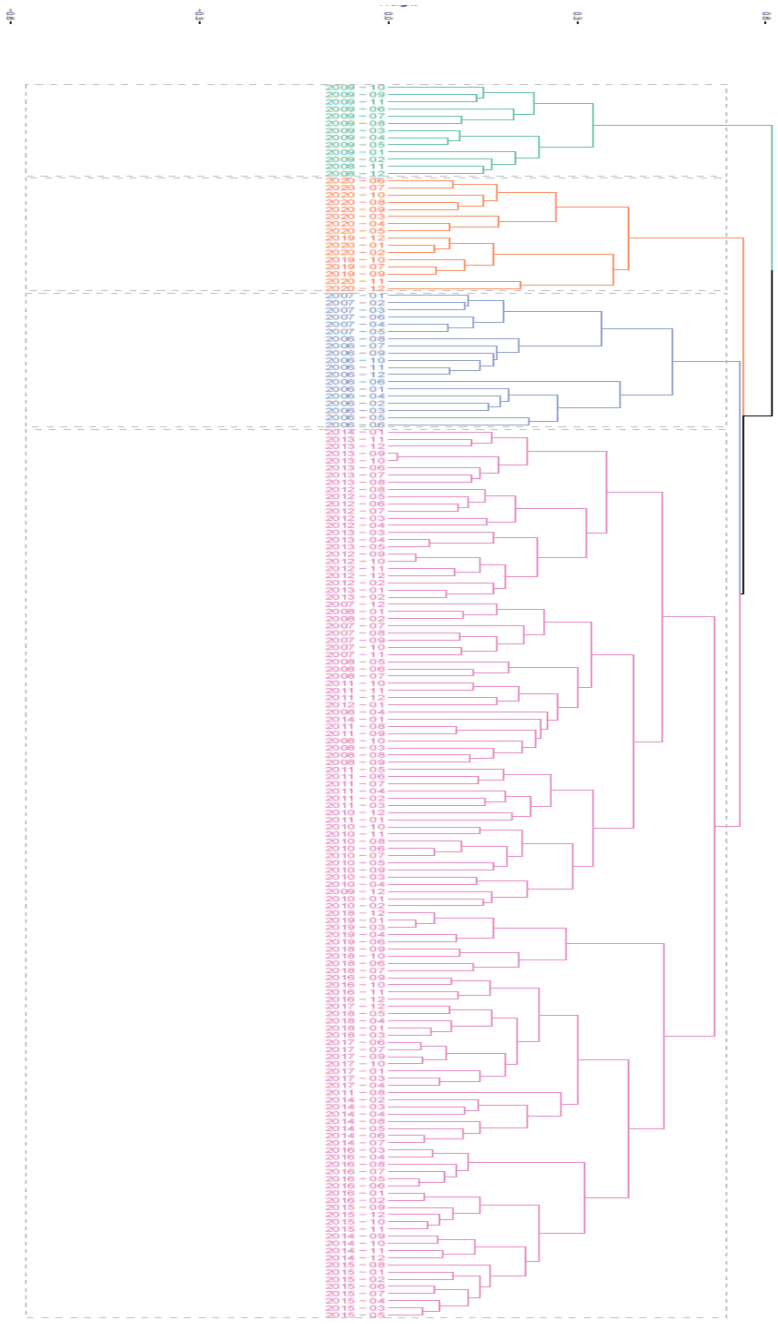


Figure 17. Dendrogram of the CBRT's MPC Decision Statements

Another HC application was based on the same distance and linkage method but a DTM with lower sparsity. Sparsity is the sparse entries proportion in the matrix, i.e., the proportion of zero entries. The raw DTM has a sparsity of 87 percent with 916 terms.

To reduce sparsity, terms with more than 90 percent sparsity are removed and the remaining result is 65 percent sparsity of 262 terms. The dendrogram result of this application is shown in Figure 18.

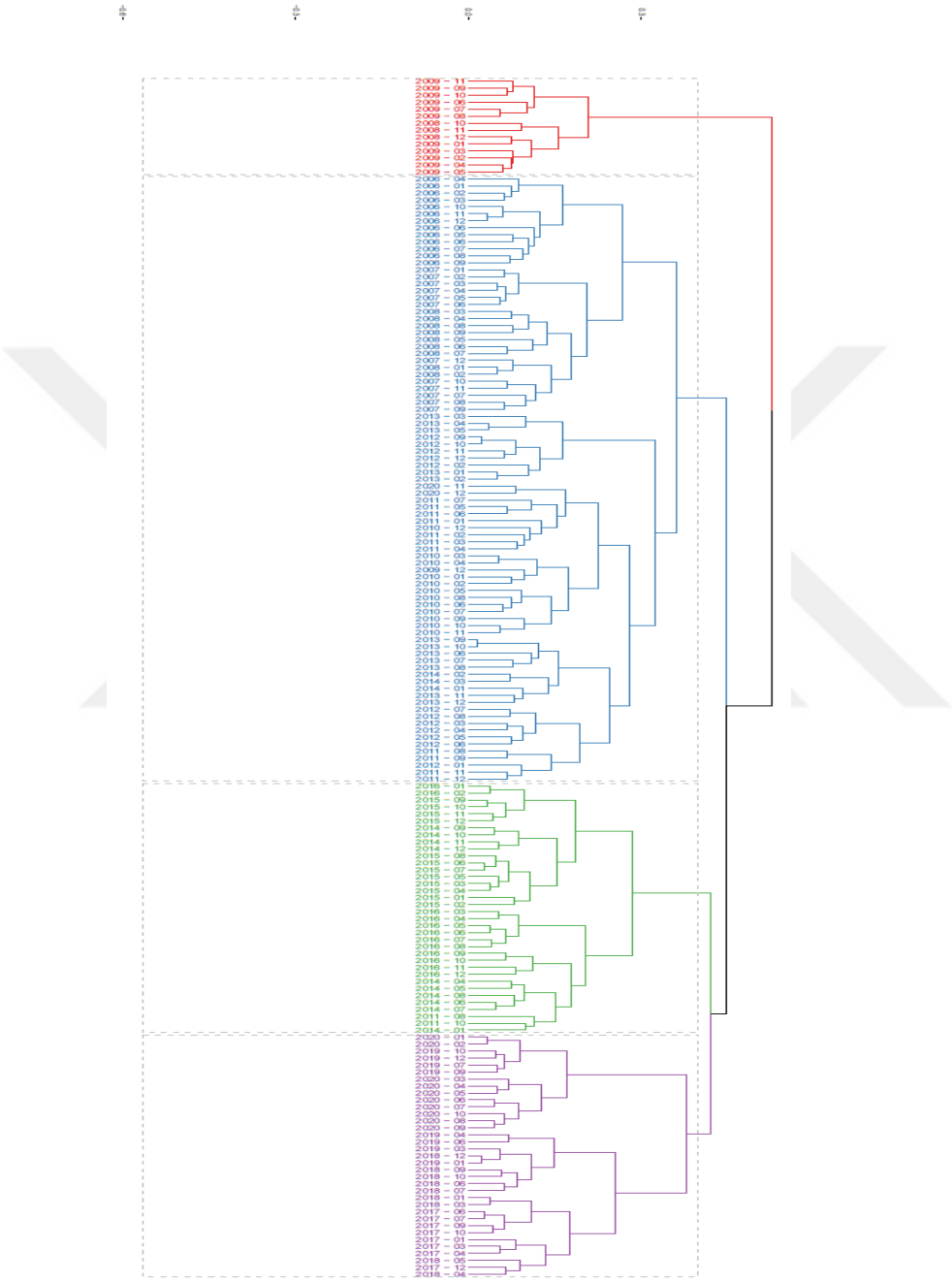


Figure 18. Dendrogram of the CBRT's MPC Decision Statements (DTM with Lower Sparsity)

The clusters of dendrograms consist of similar decision statements within the clusters. Moreover, less similarities are expected between clusters. In both dendrograms, MPC decision statements in 2009 are clustered differently. Thus, it can be deduced that decision statements in 2009 are similar to each other but less similar to the other statements in the reviewed period. To investigate this finding and interpretation, other economic indicators can be examined. In 2009, the effects of 2008 Global Financial Crisis continued in the economy which most probably affected the language of the CBRT and this could be observed in the dendrogram analysis.

Moreover, from Figure 18, it can be inferred that there are clusters containing most of the decisions from 2017-2020 and there is a cluster containing most of 2014-2016. Based on those clusters, it can be deduced that those years' decision statements have different characteristics. To illustrate, 2017-2020 stood out with the rise in the inflation which has affected the economic conditions as well as the decision statements' content.

4.8.2 Topic Modeling with Latent Dirichlet Allocation (LDA)

Topic modeling application was performed with a probabilistic model, LDA, discovering the main topics in the documents. With this analysis, topics in the CBRT's MPC decision statements, words highly related to those topics, and the change of topics in the years are examined.

To apply LDA, firstly, the number of topics should be decided. There are different methods for deciding the number of topics and one of them is deciding institutively beside more technical methods. While deciding the number of topics, being too specific and too general is conflicting two extremes and being able to interpret the results is important (Benchimol et al., 2022). Thus, for this thesis, the assumption is that there are 4 topics in each decision statement, and it is intuitively based on the general outline of the decision statements. Generally, the CBRT announces the rate decision, reviews inflation developments, gives information about the outlook with forward guidance and mentions other developments such as financial stability, current account, growth and so on. While applying LDA, DTM with term frequency weighting was used.

The results of topic modeling are evaluated for the word lists and their related topics. Also, the change of topic shares in the MPC decision statements is examined. As it is indicated in the previous parts, the CBRT announces the rate decision, reviews inflation developments, gives information about the outlook with its stance for forward guidance and mentions other developments such as financial stability, current account, growth and so on in the decision statements. Thus, the results of LDA are tried to be associated with those topics.

The first LDA was directly applied to the raw data after preprocessing and the topics' word lists are shown in Figure 19, which shows highly related words with each topic.

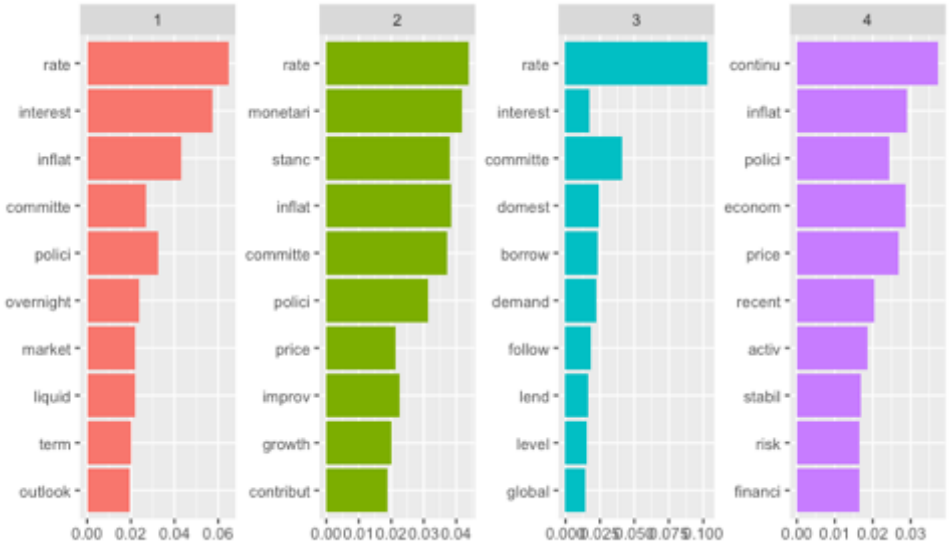


Figure 19. The Most Related Terms for Each Topic

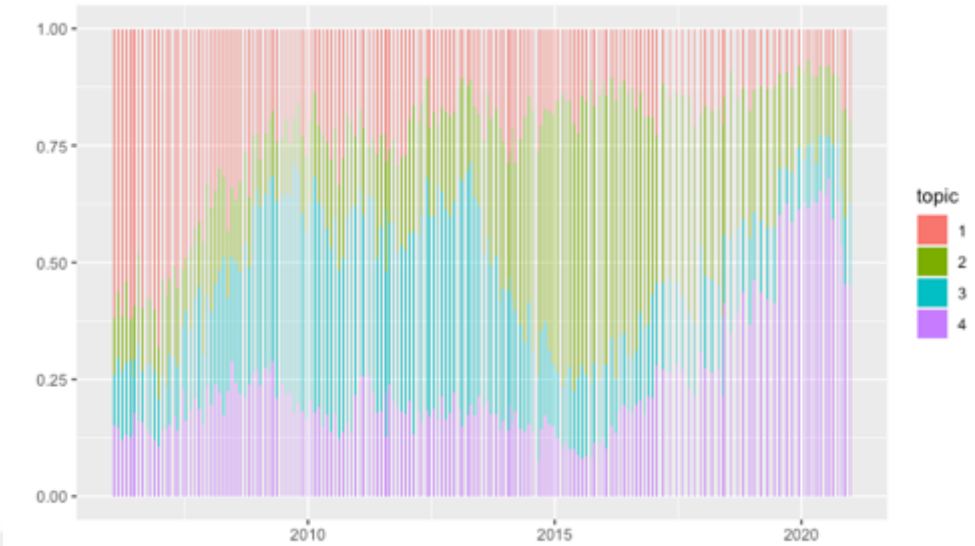


Figure 20. Topic Share Over Years

According to highly related words, the main theme of each topic was tried to be extracted. As it can be seen from Figure 19, one word can be related to more than one topic. Also, the results of LDA can be interpreted as follows: Topic 1 and Topic 3 are about the announcement of rate decision and inflation, Topic 2 is about the stance of monetary policy, Topic 4 is about other issues such as risks, financial stability and so on. The topic shares importantly change over years. From Figure 20, the increase in the share of Topic 4, related to the issues such as risks, financial stability etc., is notable. Moreover, the decrease of Topic 3 in years is well-marked. The decrease of rate decision topic can be interpreted as the share of the direct statements of policy rate has decreased but the explanations about the risks, etc., have increased in years.

To increase the performance of LDA and get a different view on results, "one", "committee", "decide", "monetary", "policy", "rate", "repo", and "week" are removed from the corpus. Those words have low TF-IDF and appear in many documents and topics as it can be seen in Figure 19. After removing those words, the results of topics are represented in Figure 21. With the update of words, the topics' focus is changed as it is expected.

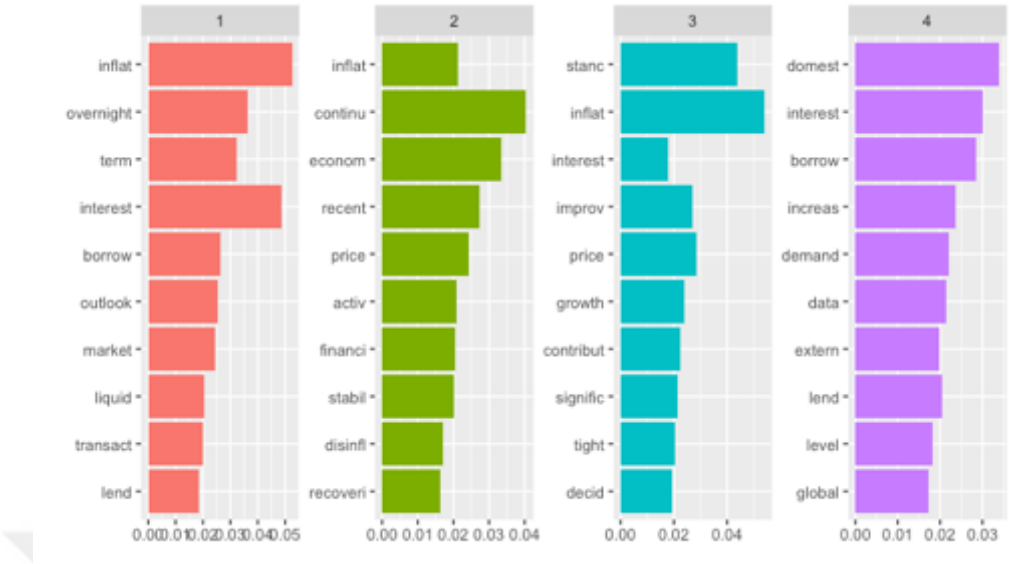


Figure 21. The Most Related Terms for Each Topic after Words Removal

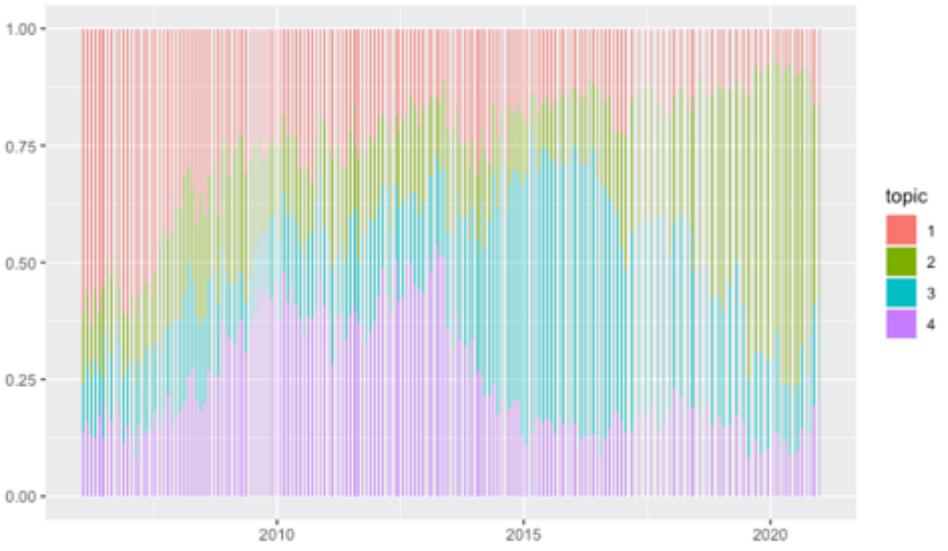


Figure 22. Topic Share Over Years after Words Removal

According to Figure 21, Topic 1 is related to rate decision announcement and Topic 3 is about policy stance. While Topic 2 is related to inflation outlook, Topic 4 is about other issues. However, Topic 2 also has financial stability which implies that Topic 2 also has terms related to other issues. According to Figure 22, the share of Topic 2 has increased over the years while the share of Topic 1 has decreased.

Moreover, as a further improvement, to increase the performance of LDA, the sparsity of DTM was decreased. With the last update, DTM has a sparsity of 88 percent with 910 terms. To reduce sparsity, terms with more than 90 percent sparsity are removed and the remaining result is 65 percent sparsity of 262 terms. The result of this application is shown in Figure 23.

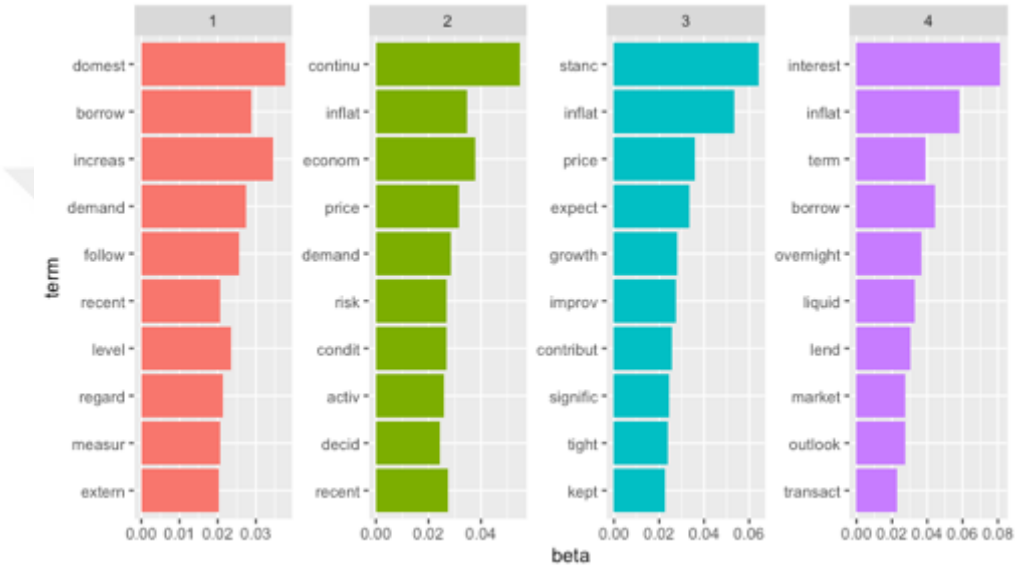


Figure 23. The Most Related Terms for Each Topic with Lower Sparsity

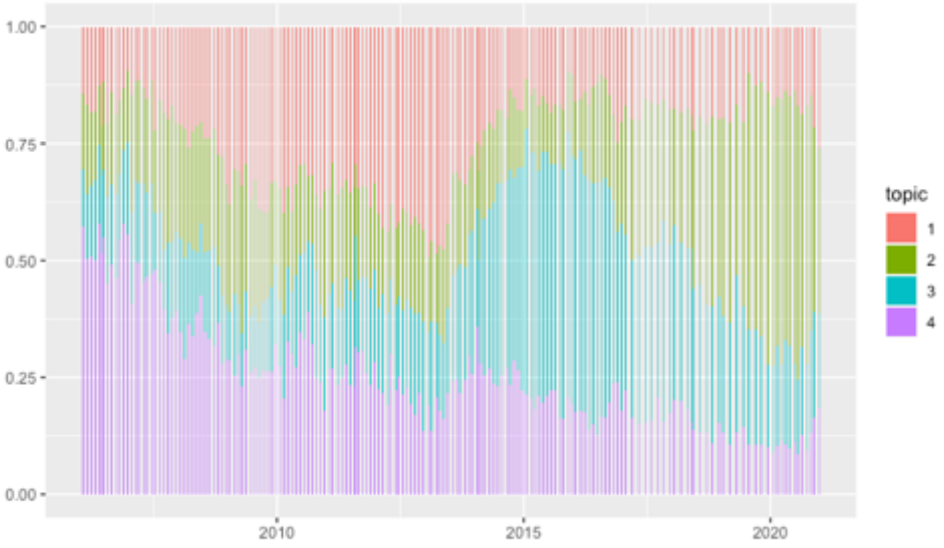


Figure 24. Topic Share Over Years with Lower Sparsity

According to Figure 23, Topic 1 is about other issues such as external issues, Topic 2 is about the inflation outlook, Topic 3 is about policy stance and Topic 4 is related to rate decision. Again, from Figure 24, the share of Topic 4 has decreased over time. However, inflation outlook, Topic 2 has increased its share in decision statements. Moreover, since 2015, the share of other issues has increased while the share of policy stance has decreased.

Overall, it could be inferred that, the share of the rate announcement has decreased in MPC decision statements over the years. In recent years, share of the inflation outlook has increased and the share of other issues has also increased. Moreover, the topic of monetary policy stance has decreased over years.

4.9 Supervised Learning

As supervised learning methods, Word Scores, Naïve Bayes and SVM are applied for this thesis and the detail of the application is explained in this section.

In 3.8.2, it is emphasized that for supervised learning methods algorithms are trained with previously labeled data. In this thesis, the labeled data for supervised learning algorithm applications were selected from Demiralp et al. (2012)'s study on the monetary policy communication effectiveness of the CBRT. In their study, the authors quantified the signal for the next interest rate decision and examined the communication's effect on predictability. For the period of February 2002 and July 2010, they assigned five potential values to each MPC decision statement: +2 strong tightening inclination, +1 weak tightening inclination, 0 signaling no change, -1 weak easing inclination, and -2 strong easing inclination.

4.9.1 Wordscores

With the reference texts, wordscores method was used for estimating scores of MPC decision statements according to reference texts. The application process started with selecting the reference texts and estimating the scores for each word type in reference texts and combining these wordscores into virgin texts.

The reference texts were selected from the Demiralp et al. (2012)'s study, and among the assessments of this study, June 2008 decision with +2 score and March 2009 decision with -2 score were selected as reference texts and their scores were classified as +1 and -1.

After selecting reference texts, the model was applied and virgin texts' scores were estimated. Finally, as a third step, the scores of virgin texts were rescaled for more easy comparison with reference texts. The results of this process are shown in Figures 25 and 26.

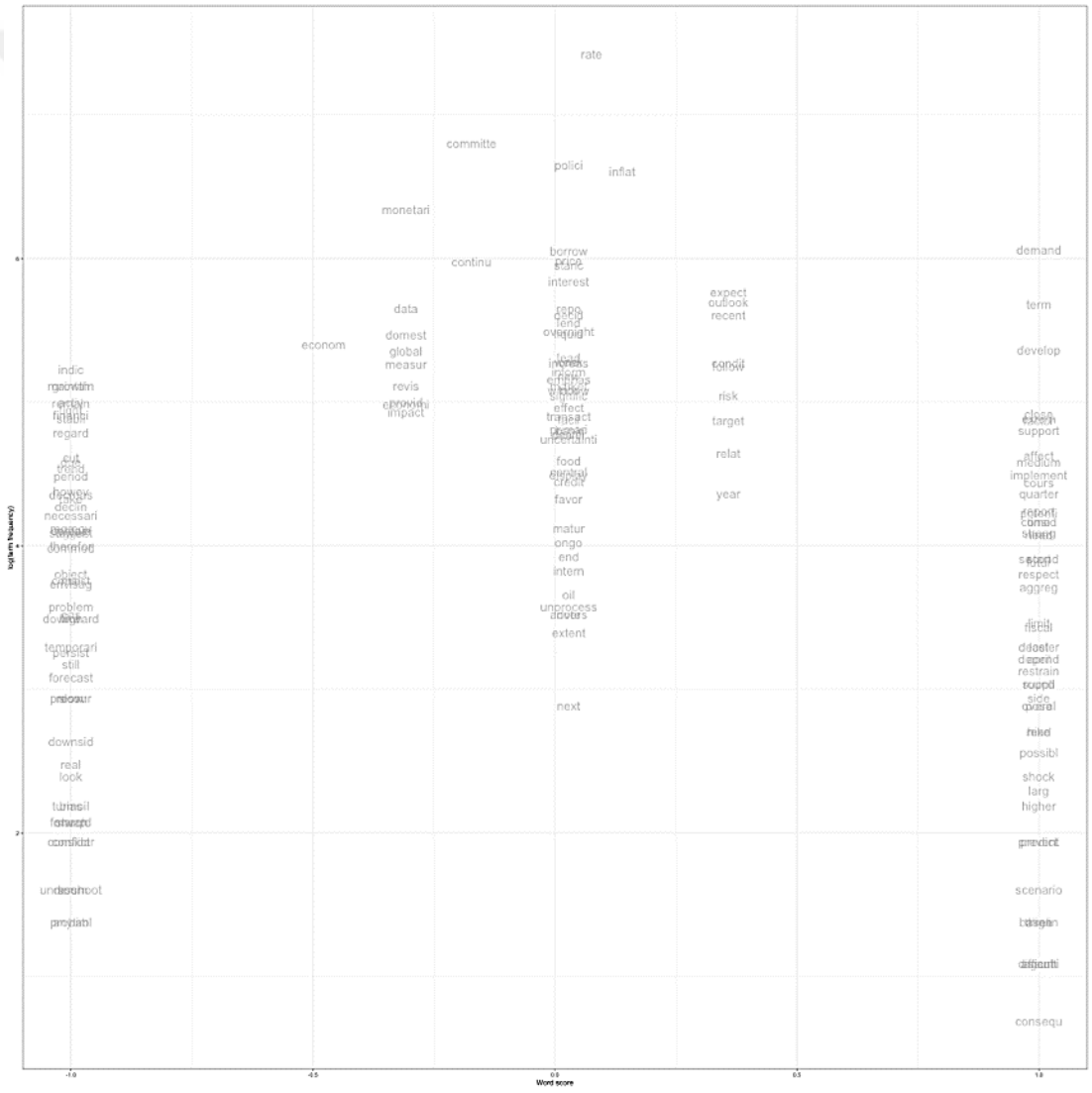


Figure 25. Wordscores and Words

Figure 25 shows the estimated wordscores of words. The neutral and most common words such as “polici”, “rate”, “borrow” and so on are scored as 0. Moreover, while “risk”, “target”, “shock”, “higher”, etc., are related to the strong tightening inclination position, “undershoot”, “downside”, “domest” and “global” etc., are related with the strong easing inclination. Thus, the results of wordscores are consistent with the expectations and such study may help wording choice of the policymakers.

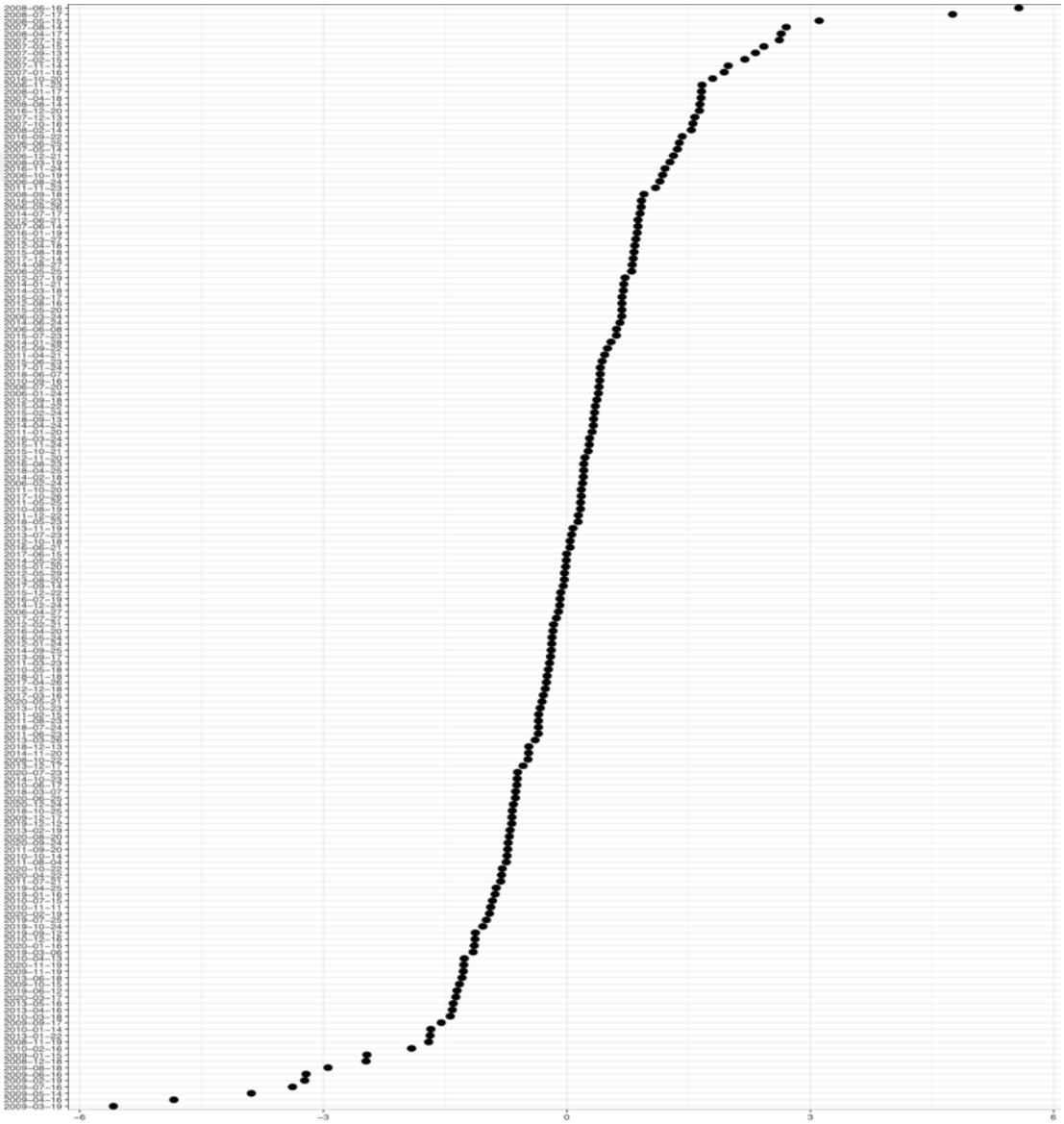


Figure 26. Wordscores and Document Positions

The estimated document positions are shown in Figure 26. June 2008 and March 2009 decision statements were selected as reference texts; thus, they are the extremes which can be found in two extreme points of Figure 26. Other decisions are sorted according to their positions in Figure 26. For example, July 2008 and May 2008 are the closest decision statements to the reference text with strong tightening inclination while April 2009 and May 2009 are the closest decision statements to the reference text with strong easing inclination. Decision statements which are closer to 0 score are the ones without the signal for either tightening or easing.

However, it should be considered that wordscores are based on reference texts and the labeled data is only available until 2010, and in this case, the reference texts are from 2008 and 2009. The change in the language in the MPC decisions may not be captured precisely. Nevertheless, this application could be a useful work for decision makers to locate their decisions' wording with more and recent labeled data.

4.9.2 Naïve Bayes

For Naïve Bayes algorithm, the MPC decision statements between January 2006 and July 2010 were used because the labeled data was available from Demiralp et al. (2012)'s study and the labeled data was used for both train and test data sets to evaluate model performance. The process started with subsetting those 55 ordinary MPC decision statements (one extraordinary MPC decision in this period was omitted from the dataset) and preprocessing. Then, the tags between -2 and 2 from Demiralp et al. (2012)'s study was added to the data set. However, since Naïve Bayes works with two classes, the data labeled as 0 were kept and other cases were classified as 1. Thus, the results classify whether there are any signs of inclination towards either tightening or easing or not in the decision statements. After that, the data was split into 75 percent training data and 25 percent test data. With this proportions, train data set has 42 observations while test data set has 13 observations.

4.9.3 Support Vector Machines (SVM)

For SVM algorithm, a very similar process with Naïve Bayes was conducted. Again, the 55 ordinary MPC decision statements between January 2006 and July 2010 and their labels from Demiralp et al. (2012)'s study were used.

After the subsetting of the MPC decisions and preprocessing, the tags between +2 and -2 were added. Again, since SVM works with binary classes, the data labeled as 0 were kept and other cases were classified as 1. Thus, the results classify whether there is no inclination in the decision statement or there are signs of inclination either tightening or easing. After that, the data was split into 75 percent training data and 25 percent test data. Again, with this proportions, train data set has 42 observations while test data set has 13 observations.

4.9.4 Naïve Bayes and SVM Evaluation

Both Naïve Bayes and SVM models have the same ordinary MPC decision statements period between January 2006 and July 2010, preprocessings, train-test proportions, and two classes of 0 and 1 (no change inclination and any inclination in direction to either tightening or easing). The results of Naïve Bayes and SVM models are compared according to their accuracy in Table 9.

Table 9

Naïve Bayes and SVM

	Accuracy
Naïve Bayes	0.8571
SVM	0.9286

According to accuracy scores, SVM has a better performance in comparison to Naïve Bayes. Besides, the accuracy of Naïve Bayes is also a high score.

Moreover, 10-fold cross validation is applied and the results are listed in Table 10.

Table 10

Naïve Bayes and SVM (10-fold Cross Validation)

	Accuracy	Sensitivity	Specificity
Naïve Bayes	0.6154	1.0	0
SVM	0.7692	0.8750	0.60

From Table 10, it can be seen that SVM has a higher accuracy after 10-fold cross validation method application. Moreover, it can be observed that Naïve Bayes' sensitivity is 1, while specificity is 0, which means the model can predict each category's true positives exactly but cannot predict true negatives.

Moreover, it should be considered that the data set is very small in this application due to the lack of labeled data. However, the importance and the possible contributions of this approach are tried to be exemplified in the scope of this thesis.

5. CONCLUSION & FUTURE WORKS

In this thesis, the CBRT's MPC decision statements for the period between January 2006 and December 2020 are examined in detail with text mining methods. For the application of those methods, firstly, preprocessing techniques were applied to text data. Then, for data exploration, frequencies of words, bigrams and TF-IDF scores of words were examined and thereupon the changes in word counts over years were investigated. It is observed that there are many fluctuations in word counts. Moreover, word clouds and word association methods were applied as other data exploration methods. In addition, readability and formality scores were calculated. In recent years, there is an increasing trend in both formality and readability scores. Furthermore, both word and sentence level sentiment scores of decision statements were calculated. It is found out that sentence level sentiment captures the extremes more. Finally, selected unsupervised learning methods (HC and LDA) for clustering decision statements and topic modeling, and supervised learning methods (Wordscores, Naïve Bayes and SVM) for classification were applied.

The results of those applications mainly point out that different periods can be observed according to trends. The changes in word counts, readability score, and formality score are usually consistent with governor changes. As for sentiment, it is observed that the cautious stance of the CBRT is reflected in the generally negative sentiment of the decision statements until 2016. After 2016, the sentiment has more positive scores. In addition, hierarchical clustering generally captures the important developments in the economy, probably because of their effects on the decision statements. Moreover, the results of topic modeling indicate a decrease in the share of the rate announcement focus in the decision statements, while inflation outlook and other issues have gained greater share in the statements in recent years. Finally, supervised learning methods help to classify the decision statements according to previously labeled data by experts. Although there is a limited data, the application of supervised learning models can be a leading example for policymakers. The results show that SVM has a better performance of classification in the scope of this thesis' limited labeled data.

All in all, the text mining methods in this thesis show that those methods are valuable for investigating central banks' policies and the application to one of the CBRT's monetary policy communication tools, MPC decision statements, leads to important information about the CBRT's communication in recent years. The outputs of the analysis show that text mining methods can be useful for central bankers before designing their policies and communication since those methods are useful for positioning the communication before making it public. To illustrate, before publishing MPC decision statements, central banks may compare the statement with previous decisions in terms of sentiment, clusters and classifications of the inclinations. Furthermore, the outputs of text mining methods may be new inputs to existing models. Moreover, any analysts trying to understand central bank policies and communication can benefit from these methods.

For future works, the relationship between text mining methods' results and other economic variables can be explored. The relationship with inflation, expectations, exchange rate, CDS, interest rates, and stock exchange can be examples of those examinations.

In addition, with more labeled data and different text mining methods & algorithms, text mining applications to central bank texts can be enriched. Firstly, the limited labeled data for the CBRT can be improved and with an increase in the labeled data, different supervised learning methods can be applied with more accuracy. Also, different text mining methods and machine learning algorithms can be applied. To illustrate, in the literature, Dynamic Topic Modeling and Latent Semantic Analysis are used widely.

Moreover, other central bank documents such as monetary policy discussion summaries, Inflation Report, Financial Stability Report and the speeches of governors can be investigated with text mining methods. Also, the tenures of governors' can be compared in detail with the text mining methods.

REFERENCES

Aggarwal, C. C. (2015). *Data mining: the textbook* (Vol. 1). New York: springer. https://doi.org/10.1007/978-3-319-14142-8

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919.

Anees, A. F., Shaikh, A., Shaikh, A., & Shaikh, S. (2020). Survey paper on sentiment analysis: Techniques and challenges. *EasyChair Preprint* No 2389.

Apel, M., & Grimaldi, M. (2012). The information content of central bank minutes. *Riksbank Research Paper Series*, 92. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2092575

Atan, S. (2020) Metin Madenciliği: İmkânlar, Yöntemler Ve Kisitlar. *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 31, 220-239. https://doi.org/10.20875/makusobed.476524

Bailliu, J. N., Han, X., Sadaba, B., & Kruger, M. (2021). Chinese monetary policy and text analytics: Connecting words and deeds. *Bank of Canada Staff Working Paper 2021-3*. https://doi.org/10.34989/swp-2021-3

Basu, A., Walters, C., & Shepherd, M. (2003). Support vector machines for text categorization. *36th Annual Hawaii International Conference on System Sciences,* 2003, 7. https://doi.org/10.1109/hicss.2003.1174243

Başkaya, S., Kara, H., & Mutluer, D. (2008). Expectations, communication and monetary policy in Türkiye. *CBRT Working Paper*, 08/01.

Benchimol, J., Kazinnik, S., & Saadon, Y. (2022). Text mining methodologies with R: An application to central bank texts. *Machine Learning with Applications*, *8*, 100286. https://doi.org/10.1016/j.mlwa.2022.100286

Bengfort, B., Bilbro, R., & Ojeda, T. (2018). *Applied text analysis with python: Enabling language-aware data products with machine learning*. O'Reilly Media, Inc..

Benoit, K., Obeng A., Watanabe, K., Matsuo, A., Nulty, P. & Müller, S. (2021). readtext: Import and Handling for Plain and Formatted Text Files. R package version 0.81. https://CRAN.R-project.org/package=readtext

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., Matsuo, A. (2021). quanteda: An R package for the quantitative analysis of textual data. R package version 3.2.0. https://CRAN.R-project.org/package=quanteda

Benoit, K., Watanabe, K., Wang, H., Perry, O. P., Lauderdale, B., Gruber, J., Lowe, W. & Sindhwani, V. (2021). quanteda.textmodels: Scaling Models and Classifiers for Textual Data. R package version 0.9.4. https://CRAN.R-project.org/package=quanteda.textmodels

Bernanke, B. S., & Mishkin, F. S. (1997). Inflation targeting: a new framework for monetary policy?. Journal of Economic perspectives, 11(2), 97-116. https://doi.org/10.1257/jep.11.2.97

Bhavsar, H., & Panchal, M. H. (2012). A review on support vector machine for data classification. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *1*(10), 185-189.

Bholat, D., Hansen, S., Santos, P., & Schonhardt-Bailey, C. (2015). Text mining for central banks. *Centre for Central Banking Studies,* 33. https://doi.org/10.2139/ssrn.2624811

Binette, A., & Tchebotarev, D. (2019). Canada's Monetary Policy Report: If Text Could Speak, What Would It Say? Bank of Canada, Staff Analytical Note No. 2019-5. https://doi.org/10.34989/san-2019-5

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

Blinder, A., Goodhart, C., Hildebrand, P., Lipton, D., & Wyplosz, C. (2001). How do central banks talk? Geneva Reports on the World Economy 3. *Center for Economic Policy Research*. 1-15. Retrieved from https://www.cimb.ch/uploads/1/1/5/4/115414161/geneva3.pdf

Blinder, A. S., Ehrmann, M., Fratzscher, M., De Haan, J., & Jansen, D. J. (2008). Central bank communication and monetary policy: A survey of theory and evidence. *Journal of economic literature*, *46*(4), 910-45. https://doi.org/10.1257/jel.46.4.910

Boukus, E., & Rosenberg, J. V. (2006). The information content of FOMC minutes. https://doi.org/10.2139/ssrn.922312

Brouwer, N., & De Haan, J. (2021). Central bank communication with the general public: effective or not?. *SUERF Policy Brief, No 57.* Retrieved from https://www.suerf.org/suer-policy-brief/21899/central-bank-communication-with-the-general-public-effective-or-not

Bruinsma, B., & Gemenis, K. (2019). Validating Wordscores: The promises and pitfalls of computational text scaling. *Communication Methods and Measures*, *13*(3), 212-227. https://doi.org/10.1080/19312458.2019.1594741

Bruno, G. (2016). Text mining and sentiment extraction in central bank documents. In *2016 IEEE International Conference on Big Data (Big Data)*, 1700-1708. IEEE. https://doi.org/10.1109/bigdata.2016.7840784

Bruno, G. (2017). Central Bank Communications: information extraction and semantic analysis. *IFC-Bank Indonesia Satellite Seminar on "Big Data" at the ISI Regional Statistics Conference 2017.*

CBRT. (2011). Merkez Bankaları ve İletişim, Türkiye Cumhuriyet Merkez Bankasında İletişim Politikalarının Gelişimi. Retrieved from https://www.tcmb.gov.tr/wps/wcm/connect/TR/TCMB+TR/Main+Menu/Yayinlar/Kitap%2C+Kitapciklar+ve+Brosur/

Chauhan, U., & Shah, A. (2021). Topic modeling using latent Dirichlet allocation: A survey. *ACM Computing Surveys (CSUR)*, *54*(7), 1-35. https://doi.org/10.1145/3462478

Chen, S. (2020). Getting Started with Text Vectorization. Retrieved from https://towardsdatascience.com/getting-started-with-text-vectorization-2f2efbec6685

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, *11*(4), 411-433.

DataCamp. (n.d). *Text Mining with Bag-of-Words in R.* DataCamp Course, Retrieved from https://campus.datacamp.com/courses/text-mining-with-bag-of-words-in-r/adding-to-your-tm-skills?ex=7

De Haan, J., Eijffinger, S. C., & Rybiński, K. (2007). Central bank transparency and central bank communication: Editorial introduction. *European Journal of Political Economy*, *23*(1), 1-8. https://doi.org/10.1016/j.ejpoleco.2006.09.010

Demiralp, S., Kara, H., & Özlü, P. (2012). Monetary policy communication in Türkiye. *European Journal of Political Economy*, *28*(4), 540-556. https://doi.org/10.1016/j.ejpoleco.2012.06.001

Duran, M., Özlü, P., & Ünalmış, D. (2010). TCMB faiz kararlarinin hisse senedi piyasalari üzerine etkisi. *Central Bank Review*, *10*(2), 23-32.

Ehrmann, M., & Fratzscher, M. (2005). How should central banks communicate?. *ECB Working Paper No 557*. https://doi.org/10.2139/ssrn.850944

Emekçi, H. (2017). *Computational Analysis of CBRT's Policy Statements and Quantifying the Effects on Financial Markets*. (PhD thesis). Graduate School of Social Sciences, Department of Economics, Hacettepe University, Ankara, Türkiye

Ermiş, A. M. (2017). *Türkiye Cumhuriyet Merkez Bankasının iletişim politikalarının değerledirilmesi ve açıklamalarının Borsa İstanbul üzerine etkisinin analiz edilmesi.* (Master's thesis). Social Sciences Institute, Department of Economics, İstanbul Technical University. İstanbul, Türkiye

Eusepi, S., & Preston, B. (2010). Central bank communication and expectations stabilization. *American Economic Journal: Macroeconomics*, *2*(3), 235-71. https://doi.org/10.1257/mac.2.3.235

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM*, *49*(9), 76-82. https://doi.org/10.1145/1151030.1151032

Feinerer, I., Hornik, K. (2020). tm: Text Mining Package. R package version 0.7-8, https://CRAN.R-project.org/package=tm.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of statistical software*, *25*, 1-54. https://doi.org/10.18637/jss.v025.i05

Feldkircher, M., Hofmarcher, P., & Siklos, P. (2021). What do central banks talk about? A European perspective on central bank communication. *Focus on European Economic Integration, Austrian Central Bank,* Q2/21, 61-81.

Fellows, I. (2018). wordcloud: Word Clouds. R package version 2.6, https://CRAN.R-project.org/package=wordcloud

Fung, B. C., Wang, K., & Ester, M. (2009). Hierarchical document clustering. *Encyclopedia of Data Warehousing and Mining,* 2*,* 970-975. IGI Global. https://doi.org/10.4018/978-1-60566-010-3.ch150

Fung, G. P. C., Yu, J. X., & Lu, H. (2005). The Predicting Power of Textual Information on Financial Markets. *IEEE Intell. Informatics Bull.,* 5(1), 1-10.

Gandhi, R. (2018). Support Vector Machine — Introduction to Machine Learning Algorithms. Retrieved from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Geraats, P. M. (2002). Central bank transparency. *The economic journal*, *112*(483), F532-F565. https://doi.org/10.1111/1468-0297.00082

Ghirelli, C., Hurtado, S., Pérez, J. J., & Urtasun, A. (2021). New Data Sources for Central Banks. *Data Science for Economics and Finance*, Springer, Cham, 169-194. https://doi.org/10.1007/978-3-030-66891-4_8

Grira, N., Crucianu, M., & Boujemaa, N. (2004). Unsupervised and semi-supervised clustering: a brief survey. *A review of machine learning techniques for processing multimedia content*, *1*, 9-16.

Grün, B., Hornik, K., Blei, D., Lafferty, J. D., Phan, X., Matsumoto, M., Nishimura, T. & Cokus, S. (2021). topicmodels:Topic Models. R package version 0.2-12, https://CRAN.R-project.org/package=topicmodels

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, *1*(1), 60-76. https://doi.org/10.4304/jetwi.1.1.60-76

Gürkaynak, R. S., Sack, B. P., & Swanson, E. T. (2004). Do actions speak louder than words? The response of asset prices to monetary policy actions and statements. *The Response of Asset Prices to Monetary Policy Actions and Statements (November 2004)*. https://doi.org/10.2139/ssrn.633281

Hansen, S., & McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, *99*, S114-S133. https://doi.org/10.1016/j.jinteco.2015.12.008

Hansen, S., McMahon, M., Prat, A. (2014). Transparency and Deliberation within the FOMC: a Computational Linguistics Approach. *Discussion Papers, Centre for Economic Policy Research* (CEPR) 9994

Hearst, M. (2003). What is text mining. *SIMS, UC Berkeley*, *5*. Retrieved from https://people.ischool.berkeley.edu/~hearst/text-mining.html

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014,). Word cloud explorer: Text analytics based on word clouds. *47th Hawaii international conference on system sciences*, 2014, 1833-1842. IEEE. https://doi.org/10.1109/hicss.2014.231

Hendry, S., & Madeley, A. (2010). Text mining and the information content of Bank of Canada communications. *Available at SSRN 1722829.* https://doi.org/10.2139/ssrn.1722829

Heylighen, F., & Dewaele, J. M. (2002). Variation in the contextuality of language: An empirical measure. *Foundations of science*, *7*(3), 293-340. https://doi.org/10.1023/a:1019661126744

Hong, J. W., & Park, S. B. (2019). The identification of marketing performance using text mining of airline review data. *Mobile Information Systems*, 2019, 1-8. https://doi.org/10.1155/2019/1790429

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *In Ldv Forum*, 20(1), 19-62.

Hughes, P. T., & Kesting, S. (2014). A literature review on central bank communication. *On the Horizon*, 22(4), 328-340. https://doi.org/10.1108/oth-07-2014-0027

IBM Cloud Education. (2020). *Text Mining.* IBM. https://www.ibm.com/cloud/learn/text-mining

Iglesias, J., Ortiz, A., & Rodrigo, T. (2017). How do the EM Central Bank talk? A Big Data approach to the Central Bank of Türkiye. *BBVA Working Paper No. 17/24.*

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS transactions on computers*, *4*(8), 966-974.

Jarman, A. M. (2020). Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method. *Georgia Southern University*.

Jha, A. (2021). Vectorization Techniques in NLP [Guide]. Retrieved from https://neptune.ai/blog/vectorization-techniques-in-nlp-guide

Kahveci, E., & Odabaş, A. (2016). Central banks' communication strategy and content analysis of monetary policy statements: The case of Fed, ECB and CBRT. *Procedia-Social and Behavioral Sciences*, *235*, 618-629. https://doi.org/10.1016/j.sbspro.2016.11.039

Kannan, S., & Gurusamy, V. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16.

Kansu, A. (2007). Para Politikasında Şeffaflık Ve Enflasyonist Beklentilerin Yönlendirilmesi. *Doğuş Üniversitesi Dergisi*, *8*(1), 59-71. https://doi.org/10.31671/dogus.2019.242

Kassambara, A. & Mundt, F. (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version of 1.0.7. https://CRAN.R-project.org/package=factoextra

Kolini, F., & Janczewski, L. (2017). Clustering and topic modelling: A new approach for analysis of national cyber security strategies. *PACIS 2017 Proceedings*. http://aisel.aisnet.org/pacis2017/126

Krukovets, D. (2020). Data science opportunities at central banks: overview. *Visnyk of the National Bank of Ukraine*, 249, 13-24. https://doi.org/10.26531/vnbu2020.249.02

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R. Ziem, A., Scrucca, L., Tang, Y. & Candan, C. (2021) caret: Classification and Regression Training. R package version 6.0-90. https://CRAN.R-project.org/package=caret

Küçükkocaoğlu, G., Ünalmış, D., & Ünalmış, I. (2013). How do Banks' Stock Returns Respond to Monetary Policy Committee Announcements in Türkiye?. Evidence from Traditional versus New Monetary Policy Episodes. *Central Bank of the Republic of Türkiye Working Paper*, 13/30.

Kütük, Y. (2021). Semantic Analysis of the Central Bank of the Republic of Türkiye Communication Reports and Forecasting Model with LSTM. https://doi.org/10.2139/ssrn.3828428

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American political science review*, *97*(2), 311-331.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Lo, R. T. W., He, B., & Ounis, I. (2005). Automatically building a stopword list for an information retrieval system. *Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR)* ,5, 17-24.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance*, *66*(1), 35-65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Lowe, W. (2008). *Understanding Wordscores. Political Analysis, 16(04), 356–371.* https://doi.org/10.1093/pan/mpn004

Luangaram, P., & Wongwachara, W. (2017). More than words: a textual analysis of monetary policy communication. *PIER Discussion Papers*, *54.*

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*, *9*, 381-386.

Máté, Á., Sebők, M., & Barczikay, T. (2021). The effect of central bank communication on sovereign bond yields: The case of Hungary. *Plos one, 16(2), e0245515.*

Mathur, A., & Sengupta, R. (2019). Analysing monetary policy statements of the Reserve Bank of India. https://doi.org/10.2139/ssrn.3383869

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.

Mishkin, F. S. (2004). Can central bank transparency go too far?. *NBER Working Papers 10829, National Bureau of Economic Research, Inc* https://doi.org/10.3386/w10829

Mocherla, S., Danehy, A., & Impey, C. (2017). Evaluation of naive bayes and support vector machines for wikipedia. *Applied Artificial Intelligence*, *31*(9-10), 733-744. https://doi.org/10.1080/08839514.2018.1440907

Mohler, T. (2020). The 7 Basic Functions of Text Analytics & Text Mining. Retrieved from https://www.lexalytics.com/lexablog/text-analytics-functions-explained

Moniz, A., & Jong, F. D. (2014). Predicting the impact of central bank communications on financial market investors' interest rate expectations. *Lecture Notes in Computer Science,* 8798, 144-155. https://doi.org/10.1007/978-3-319-11955-7_12

Montes, G. C., Oliveira, L. V., Curi, A., & Nicolay, R. T. F. (2016). Effects of transparency, monetary policy signalling and clarity of central bank communication on disagreement about inflation expectations. *Applied Economics*, *48*(7), 590-607. https://doi.org/10.1080/00036846.2015.1083091

Mullen, L., Benoit, K., Keyes, O., Selivanov, D., & Arnold, J. (2018). Fast, consistent tokenization of natural language text. *Journal of Open Source Software*, *3*(23), 655. https://doi.org/10.21105/joss.00655

Navin, M. J. R., & Pankaja, R. (2016). Performance analysis of text classification algorithms using confusion matrix. *International Journal of Engineering and Technical Research (IJETR),* 6(4), 75-78.

Neuenkirch, M. (2012). Managing financial market expectations: the role of central bank transparency and central bank communication. *European Journal of Political Economy*, *28*(1), 1-13. https://doi.org/10.1016/j.ejpoleco.2011.07.003

Neuenkirch, M. (2013). Monetary policy transmission in vector autoregressions: A new approach using central bank communication. *Journal of Banking & Finance*, *37*(11), 4278-4285. https://doi.org/10.1016/j.jbankfin.2013.07.044

Nguyen, E. (2014). Chapter 4-Text mining and Network Analysis of Digital Libraries. *Data Mining Applications with R,* 2014, 95-115. ttps://doi.org/10.1016/B978-0-12-411511-8.00004-9

Nilsson Björkenstam, K. (2013). What is a corpus and why are corpora important tools? *Presented at the Nordic seminar: How can we use sign language corpora?* Copenhagen, Denmark, December 12-13, 2013. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-99009

Omotosho, B. S. (2019). Central bank communication in Ghana: insights from a text mining analysis. *Available at SSRN 3526451*. https://doi.org/10.2139/ssrn.3526451

Ponweiser, M. (2012). Latent Dirichlet allocation in R. Retrieved from https://core.ac.uk/display/11007925

Porter, M. F. (1980), An algorithm for suffix stripping, *Program*, 14(3), 130−137. https://doi.org/10.1108/eb046814

Porter, M.F. (2006), Retrieved from https://tartarus.org/martin/PorterStemmer/

R Core Team. (2020). stats: The R Stats package. R package version 4.0.3. https://CRAN.R-project.org/package=stats

Radovanović, M., & Ivanović, M. (2008). Text mining: Approaches and applications. *Novi Sad J. Math*, *38*(3), 227-234

Raschka, S. (2014). Naive bayes and text classification i-introduction and theory. *arXiv preprint arXiv:1410.5329*. https://doi.org/10.48550/arXiv.1410.5329

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *Encyclopedia of database systems*, *5*, 532-538. https://doi.org/10.1007/978-0-387-39940-9_565

Rinker, T. (2013). qdapDictionaries: Dictionaries to Accompany the qdap Package. R Package version 1.0.7. University at Buffalo. Buffalo, New York. https://CRAN.R-project.org/package=qdapDictionaries

Rinker, T. (2020). qdap: Bridging the Gap Between Qualitative Data and Quantitative Analysis. R package version 2.4.3. https://CRAN.R-project.org/package=qdap

Rinker, T. (2021). Sentimentr: Calculate Text Polarity Sentiment. R package version 2.9.0. https://CRAN.R-project.org/package=sentimentr

Rybinski, K. (2019). A Machine Learning Framework for Automated Analysis of Central Bank Communication and Media Discourse: the Case of Narodowy Bank Polski. *Bank i Kredyt*, 50(1*)*, 1-20.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 1-21. https://doi.org/10.1007/s42979-021-00592-x

Sembiring, R. W., Zain, J. M., & Embong, A. (2010). A comparative agglomerative hierarchical clustering method to cluster implemented course. *Journal of Computing, 2(12)*. https://doi.org/10.48550/arXiv.1101.4270

Senter, R. J., & Smith, E. A. (1967). *Automated readability index*. Cincinnati University, OH.

Serra, A., & Tagliaferri, R. (2019). Unsupervised Learning: Clustering. *Encyclopedia of Bioinformatics and Computational Biology*, 2009, 350-357. https://doi.org/10.1016/b978-0-12-809633-8.20487-1

Shah, N., & Mahajan, S. (2012). Document clustering: a detailed review. *International Journal of Applied Information Systems*, *4*(5), 30-38. https://doi.org/10.5120/ijais12-450691

Shahi, T. B., & Pant, A. K. (2018). Nepali news classification using Naive Bayes, support vector machines and neural networks. *2018 International Conference on Communication Information and Computing Technology (ICCICT)*, 1-5. https://doi.org/10.1109/iccict.2018.8325883

Sharma, D., Sabharwal, M., Goyal, V., & Vij, M. (2020). Sentiment analysis techniques for social media data: a review. *First international conference on sustainable technologies for computational intelligence*, 75-90.

Shirota, Y., Hashimoto, T., & Sakura, T. (2015). Topic extraction analysis for monetary policy minutes of Japan in 2014. In *Industrial Conference on Data Mining*, 141-152. Springer, Cham. https://doi.org/10.1007/978-3-319-20910-4_11

Silge, J. & Robinson, D. (2021). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. R package version 0.3.2. https://CRAN.R-project.org/package=tidytext

Soylu, N., Korkmaz, T., & Çevik, E. (2014). Merkez Bankası Faiz Duyurularının Finansal Piyasalara Etkisi. *Business & Economics Research Journal*, *5*(4), 89-118. Retrieved from https://www.berjournal.com/tr/merkez-bankasi-faiz-duyurularinin-finansal-piyasalara-etkisi

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. *Computer Science & Engineering (CS&E) Technical Reports, 00-034.* Retrieved from the University of Minnesota Digital Conservancy, https://hdl.handle.net/11299/215421.

Sumathy, K. L., & Chidambaram, M. (2013). Text mining: concepts, applications, tools and issues-an overview. *International Journal of Computer Applications*, *80*(4), 29-32. https://doi.org/10.5120/13851-1685

Şen-Taşbaşı, A.(2011). Why Should Central Banks Communicate with Public?"Exposuring the Frame" vs."Never Explain, Never Excuse". Retrieved from https://recil.ensinolusofona.pt/handle/10437/5345

Tan, A. H. (1999, April). Text mining: The state of the art and the challenges. In *Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases*, 65-70.

Thakkar, A., & Chaudhari, K. (2020). Predicting stock trend using an integrated term frequency–inverse document frequency-based feature weight matrix with neural networks. *Applied Soft Computing*, *96*, 106684. https://doi.org/10.1016/j.asoc.2020.106684

Tobback, E., Nardelli, S., & Martens, D. (2017). Between hawks and doves: measuring central bank communication. ECB Working Paper No. 2085. https://doi.org/10.2139/ssrn.2997481

Tumala, M. M., & Omotosho, B. S. (2019). A text mining analysis of central bank monetary policy communication in Nigeria. *CBN Journal of Applied Statistics*, *10*(2), 73-107. https://doi.org/10.33429/cjas.10219.3/6

Vijaya, Sharma, S., & Batra, N. (2019). Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon).* https://doi.org/10.1109/comitcon.2019.8862232

Wegrzyn-Wolska, K., & Bougueroua, L. (2012). Tweets mining for French presidential election. *2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN)*, 138-143. https://doi.org/10.1109/cason.2012.6412392

Wei, L., Wei, B., & Wang, B. (2012). Text classification using support vector machine with mixture of kernel. *Journal of Software Engineering and Applications*, *5(12)*, 55-58. https://doi.org/10.4236/jsea.2012.512b012

Weidmann, J. (2018). Central bank communication as an instrument of monetary policy. Retrieved from https://www.bis.org/review/r180511a.htm

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures*, *11*(4), 245-265. https://doi.org/10.1080/19312458.2017.1387238

Wickham, H. (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr

Wickham, H. (2021). ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package version 3.3.5. https://CRAN.R-project.org/package=dplyr

Wickham, H., François, R., Henry, R., & Müller, K. (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.8. https://CRAN.R-project.org/package=dplyr

Woodford, M. (2005). Central bank communication and policy effectiveness. *NBER Working Papers 11898, National Bureau of Economic Research, Inc*. https://doi.org/10.3386/w11898

Yang, S., & Zhang, H. (2018). Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis. *International Journal of Computer and Information Engineering*, *12*(7), 525-529.

Yellen, J. (2012). *Revolution and Evolution in Central Bank Communications.* Haas School of Business, University of California, Berkeley, Berkeley, California, FED. https://www.federalreserve.gov/newsevents/speech/yellen20121113a.htm

Yetkin, Z. Ö. (2005). Merkez Bankalarının Para Politikalarının Tasarımında İletişim Politikalarının Önemi ve Bekleyişlerin Yönetimi. *Uzmanlık Yeterlilik Tezi, Central Bank of the Republic of Türkiye*.

Zhang, Z. (2016). Naïve Bayes classification in R. *Annals of translational medicine*, *4*(12). https://doi.org/10.21037/atm.2016.03.38