Text Mining Monetary Policy Statements

A Primer

Zahid Asghar

Table of contents

[1 Introduction 1](#_Toc179448121)

[2 MPS Data 1](#_Toc179448122)

[2.1 Cleaning and Preprocessing 1](#_Toc179448123)

[3 Document Term Matrix 1](#_Toc179448124)

[4 Exploratory Data Analysis 1](#_Toc179448125)

[4.1 Terms appearing in more than 100, 200, 300, 500 and 700 documents 1](#_Toc179448126)

[5 Plotting data 1](#_Toc179448127)

[6 Correlation plot 1](#_Toc179448128)

[6.1 Word Cloud 1](#_Toc179448129)

[6.2 Weighting Scheme 1](#_Toc179448130)

[6.3 Tidytext data table 1](#_Toc179448131)

[7 Exploratory Data Analysis 1](#_Toc179448132)

[7.1 Word counting 1](#_Toc179448133)

[7.2 Relative frequency 1](#_Toc179448134)

[7.3 Semantic analysis 1](#_Toc179448135)

[8 Topic models 1](#_Toc179448136)

[8.1 Heatmap 1](#_Toc179448137)

[9 Topic comparison 2018 vs 2023 1](#_Toc179448138)

[10 Wordfish 1](#_Toc179448139)

[10.1 Plotting Wordfish Score 1](#_Toc179448140)

[10.2 Plotting Wordscores Score 1](#_Toc179448141)

[11 Conclusion 1](#_Toc179448142)

## 1 Introduction

This study explores text data analysis of monetary policy statements issued by State Bank of Pakistan over past 18 to 20 years. Dealing with text is typically not considered as important in training of economics and social sciences for data analysis. This is in direct contrast with how often it has to be dealt with prior to more common analysis, or how interesting it might be to have text be the focus of analysis [Text Analysis in R](https://m-clark.github.io/text-analysis-with-R/intro.html#overview). This paper/tutorial aims at providing a sense of the things one can do with text, and the sorts of analyses that might be useful. In era of ML, AI and NLP, text analysis is becoming more and more important as human may miss some important information and computer-based approaches can process and summarise text more efficiently than humans. Also, textual analysis may extract meaning from text missed by human readers, who may overlook certain patterns because they do not conform to prior beliefs and expectations (Herasymova, 2022). There are many studies which have done in detail textual analysis of central bank statements Shapiro and Wilson (2021). I have mainly followed Benchimol, Kazinnik, and Saadon (2022) guidelines in this paper.

With this motivation, I do textual analysis to monetary policy statements by the State Bank of Pakistan (SBP) which is aimed at understanding the monetary policy stance of the central bank. The monetary policy statement (MPS) is a document that is released by the central bank to communicate its monetary policy stance to the public. The MPS is released after the Monetary Policy Committee (MPC) meeting, which is usually held every two months. The MPS contains information on the current state of the economy, the central bank’s assessment of the economy, and the central bank’s monetary policy stance. Monetary policy statements are an important source of information for researchers, as they provide information on the central bank’s monetary policy stance, which can have implications for the economy. I have extracted monetary policy statements from the State Bank of Pakistan (SBP) website, and used text mining techniques to extract information from these statements.

The primary objective of these statements is to inform and guide economic analysts and other stakeholders involved in advising traders within the financial markets. They provide insights into recent economic developments and anticipate future trends, thereby facilitating informed decision-making. Therefore, it becomes all the important to analyse how effectively SBP communicates through its policy statements. In assessing the effectiveness of these MPS, we employ textual analysis techniques. Unstructured data, often rich in textual content, encompasses a wide array of sources such as news articles, social media posts, Twitter feeds, transcriptions from videos, and formal documents. Its abundance offers fresh opportunities and simultaneous challenges for researchers and research institutions alike. In this paper, I explore various methodologies for text analysis and propose a systematic approach to leverage text mining techniques. Furthermore, I examine potential empirical applications of these methods.

This paper focuses on the primer of extracting information from unstructured data, and the potential applications of text mining in the context of monetary policy statements. Quantitative analysis of text data is a rapidly growing field, and the methods and techniques used in this paper are not exhaustive. These methods are in extensive use in political science, sociology, linguistics and information security but are not in wide use in economics and finance in Pakistan. Nevertheless, there is a growing interest in the use of text mining in economics and finance, and this paper aims to provide a starting point for researchers interested in text mining and its applications in economics and finance. Recent advances in open source software and the availability of large text datasets have made it easier for researchers to apply text mining techniques to their research. As text data is usually unstructured, therefore, it is important that a reproducible and systematic approach is used to extract information from text data. The principal goal of text mining is to capture and analyze all possible meanings embeded in text. Text mining transform unstructured data into structured data, and to extract information from text data. Text mining is a rapidly growing field, and has applications in a wide range of fields, such as information retrieval, natural language processing, and data mining. Moreover, it analyzes- patterns and trends in the text data,the sentiment of text data, classify text data, to categorize the text data among many other functions.

This study aims at providing a systematic approach to text mining, and to demonstrate the potential applications of text mining in the context of monetary policy statements. Monetary policy and fiscal policy are two of the most important tools that governments and central banks use to manage the economy. Monetary policy refers to the actions taken by the central bank to influence the money supply and interest rates in the economy. The central bank uses monetary policy to achieve its objectives, such as price stability, full employment, and economic growth. The central bank uses a variety of tools to implement monetary policy, such as open market operations, discount rate changes, and reserve requirement changes. The central bank communicates its monetary policy stance to the public through monetary policy statements. Therefore, it is important to analyze these statements to understand the central bank’s monetary policy stance. Text mining techniques can be used to extract information from monetary policy statements, and to analyze the central bank’s monetary policy stance. This paper/tutorial aims to provide a systematic approach to text mining, and to demonstrate the potential applications of text mining in the context of monetary policy statements.

The paper is organized as follows. Section 2 provides an overview of text mining and its applications in economics and finance. Section 3 provides a systematic approach to text mining, and Section 4 provides an overview of the potential applications of text mining in the context of monetary policy statements. Section 5 concludes the paper.

## 2 MPS Data

I do text analysis using topic modeling, sentiment, and linguistic analysis to the Monetary Policy Statements (MPS) of the State Bank of Pakistan from 2005-2024 to capture the focus, tone, and clarity of monetary policy communications. Overall I have total 81 MPS documents from 2005 to 2024 used in this analysis. I have included all MPS shown on [SBP website link](https://www.sbp.org.pk/m_policy/mon.asp). The data is in PDF format and I have used the pdftools package to extract the text from the PDF files. All these statements are stored as corpus. There are other forms of data storage in R such as tibble and dataframe but I have used tm package to store data as corpus. The text is then cleaned and pre-processed to remove any unwanted characters and symbols. The text is then tokenized and converted to a document term matrix to analyze the text data. tm package is used to clean and preprocess the text data, and to create the document term matrix. tidyverse and tidytext packages are used to analyze and visualize the text data.

### 2.1 Cleaning and Preprocessing

The text data is cleaned and prepossessed to remove any unwanted characters and symbols. Text cleaning (or text processing) makes an unstructured set of texts uniform across and within and eliminates idiosyncratic characters or meaningless terms. Text cleaning can be loosely divided into a set of steps as shown below. Stop words are the most common words in a language, such as ‘the’, ‘is’, ‘at’, ‘which’, and ‘on’. However, before removing stop words, all words are converted to lowercase to make the text uniform using the command corpus <- tm\_map(corpus, tolower. These words are often removed from the text data because they do not carry much meaning. Numbers are also removed from the text data using tm\_map(corpus, removeNumbers) in most of the cases. One final step in data cleaning is to stem the words. Stemming is the process of reducing words to their root form. For example, the words ‘running’, ‘runs’, and ‘ran’ are all reduced to the root form ‘run’.

Below is the text left from corpus after removing the stop words.

monetary policy committee monetary policy statement’s …..

## 3 Document Term Matrix

The tm package is used to stem the words in the text data. Once we have cleaned and preprocessed the text data, we can convert the text data to a document term matrix (dtm). The goals of mapping corpus onto a dtm is twofold: the first is present the topic of each document by the frequency of semantically significant and unique terms, and second is to use it for future data analysis.

Now we create a matrix with term frequencies. First six observations of the document term matrix are shown below.The value in each cell of this matrix is typically the word frequency of each term in each document.

accordingly account accounts achieve activity adjust   
 12 334 27 35 139 4

## 4 Exploratory Data Analysis

Given a dtm with reduced sparsity, we can now perform some exploratory data analysis. The first step is to find the most frequent terms in the dtm.

There are 82 documents and 3278 columns. The most frequent top 6 and bottom 6 terms are shown below.

[1] 82 12

term frequency  
percent percent 922  
inflation inflation 913  
policy policy 612  
growth growth 525  
monetary monetary 520  
sector sector 423

term frequency  
current current 392  
deficit deficit 336  
account account 334  
rate rate 307  
expected expected 288  
decided decided 86

### 4.1 Terms appearing in more than 100, 200, 300, 500 and 700 documents

The terms appearing in more than 100 are as follows.

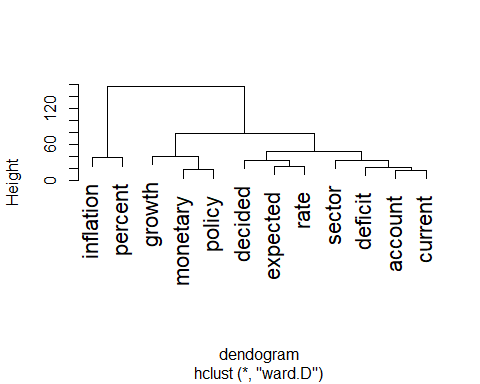
2 percent

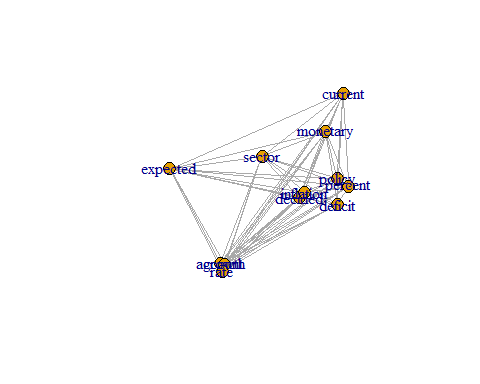
## 5 Plotting data

percent inflation policy growth monetary sector current deficit   
 922 913 612 525 520 423 392 336   
 account rate expected decided   
 334 307 288 86

word freq

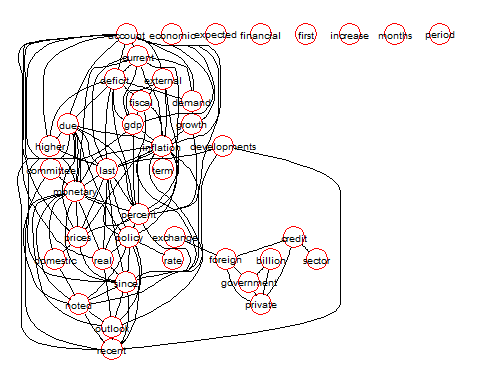
Dendrogram is another way to visualize the data. The dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters without specifying the number of clusters beforehand.





## 6 Correlation plot

One of the most intuitive way to visualize relationships between terms is to with correlation maps. Based on a certain *ad-hoc* criteria, correlation maps show how some certain terms relate to each other.



### 6.1 Word Cloud

The wordcloud package is used to visualize the text data. The word cloud is a visual representation of the frequency of words in the text data. The size of the word in the word cloud is proportional to the frequency of the word in the text data. The wordcloud package is used to create the word cloud. The word cloud is created using the document term matrix. The word cloud is used to identify the most frequent words in the text data.

[1] 82 34



### 6.2 Weighting Scheme

Another weighting scheme - term frequency/inverse document frequency is given here to create word clouds. The term frequency/inverse document frequency is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents.

[1] 82 1169

jul noted borrowings coronavirus committee global   
 0.2880723 0.2817451 0.2701409 0.2453007 0.2324261 0.2030051   
 word freq  
jul jul 0.2880723  
noted noted 0.2817451  
borrowings borrowings 0.2701409  
coronavirus coronavirus 0.2453007  
committee committee 0.2324261  
global global 0.2030051

Document term matrix is a matrix that contains the frequency of words in the text data. The rows of the matrix represent the documents, and the columns represent the words. The matrix contains the frequency of each word in each document. The document term matrix is used to analyze the text data, and to identify patterns and trends in the text data. The tm package is used to create the document term matrix. After cleaning and preprocessing the text data, the text data is tokenized and converted to a document term matrix. The document term matrix is then used to analyze the text data. The goal of dtm is two fold. The first is to present the topic of each document by the frequency of semantically significant and unique terms, and second, to position the corpus for future data analysis. The term frequency-inverse document frequency (tf-idf) is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. The tm package is used to create the document term matrix.

Why is frequency of each word is important? Simple frequency of each word is inappropiate because it can overstate the importance of small words that happen to be frequent. The term frequency-inverse document frequency (tf-idf) is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. tf-idf is defined as follows:

tf(t)= (Number of times term t appears in a document) / (Total number of terms in the document)

A more appropriate way to calaculate word frequencies is to employ the tf-idf weighting scheme. The tf-idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general.

ifd(t)= log\_e(Total number of documents / Number of documents with term t in it)

Conjugating the two gives the tf-idf score for each word in each document. 𝚝𝚏−𝚒𝚍𝚏(𝑡)=𝚝𝚏(𝑡)×𝚒𝚍𝚏(𝑡)

### 6.3 Tidytext data table

Now I shall use tidytext with the help of unnest\_tokens to convert one word per row.

word freq  
jul jul 0.2880723  
noted noted 0.2817451  
borrowings borrowings 0.2701409  
coronavirus coronavirus 0.2453007  
committee committee 0.2324261  
global global 0.2030051  
meeting meeting 0.1984043  
recovery recovery 0.1947314  
covid covid 0.1877514  
floods floods 0.1834692

To reduce dimensionality, we use sparse term-document matrix.

ng   
 2

## 7 Exploratory Data Analysis

With conversion to dtm, exploratory data analysis is performed to identify patterns and trends in the text data.

### 7.1 Word counting

Dictionary-based text analysis is popular approach mainly because its easy to implement and interpret. The dictionary-based approach is based on the idea that the frequency of certain words in a text can be used to infer the sentiment of the text. However, sentiment words from one discipline to another might be different. For example, words used in psychology to express positive sentiments might be different from words used in economics. Therefore, it is important to use a dictionary that is specific to the discipline. The tidytext package is used to count the frequency of words in the text data. The get\_sentiments function is used to get the sentiment words from the dictionary. In this document, I am using Loughran and McDonald dictionary to count the frequency of positive and negative words in the text data.

It is important to be careful in use of words to be positive or negative. For example, the word ‘increase’ is generally considered to be positive, but in the context of inflation, it is considered to be negative. Similarly the word ‘decrease’ is generally considered to be negative, but in the context of inflation, it is considered to be positive. Another example is tight and loose monetary policy. The word tight is generally considered to be positive, but in the context of monetary policy, it is considered to be negative. Similarly, the word loose is generally considered to be negative, but in the context of monetary policy, it is considered to be positive. Therefore, it is important to be careful in use of words to be positive or negative.

Next we use the match function that compares the terms in both dictionary and the text data. The match function returns the position of the first match. If there is no match, the match function returns NA. The match function is used to count the frequency of positive and negative words in the text data.

We then assign a value of 1 to the positive and negative matches. The ifelse function is used to assign a value of 1 to the positive and negative, and measure the overall sentiment for each document by the following formula:

A document is considered to be positive if the score is greater than 0, and negative if the score is less than 0.

### 7.2 Relative frequency

The relative frequency of positive and negative words is calculated by dividing the frequency of positive and negative words by the total number of words in the text.

### 7.3 Semantic analysis

The semantic analysis is performed to identify the semantic orientation of the text data. The semantic orientation is the degree to which a word is positive or negative. The semantic orientation is calculated by dividing the frequency of positive words by the frequency of negative words. The semantic orientation is calculated for each document in the text data.

## 8 Topic models

Topic modeling is a type of statistical model for discovering the abstract “topics” that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently: “dog” and “bone” will appear more often in documents about dogs, “cat” and “meow” will appear in documents about cats, and “the” and “is” will appear equally in both. A document typically concerns multiple topics in different proportions; thus, in a document that is 10% about cats and 90% about dogs, there would probably be about 9 times more dog words than cat words. The “topics” produced by topic modeling techniques are clusters of similar words. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each document’s balance of topics is.

### 8.1 Heatmap

The heatmap is used to visualize the frequency of positive and negative words in the text data. The heatmap function takes the frequency of positive and negative words as input and creates the heatmap. Heatmaps can be used to compare the content of each document, side by side, with other documents in the corpus. The word are at the bottom and color intensity shows how frequent the word is in the document. One observes that current account is in document 2 and 5 but not in document 6 and 4 while inflation is in document 6 but not in document 7.

## Topic comparison 2018 vs 2023

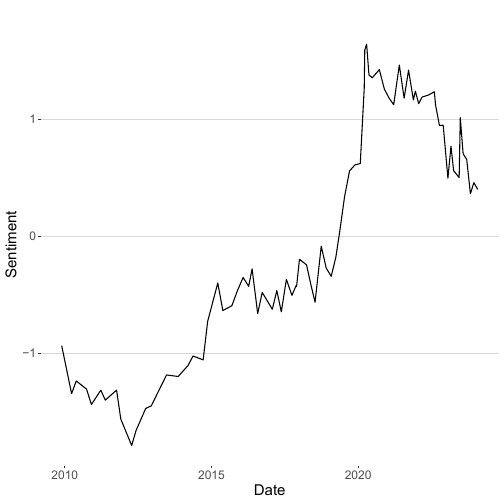
Heatmap for 2018

Heatmap for 2023

## 10 Wordfish

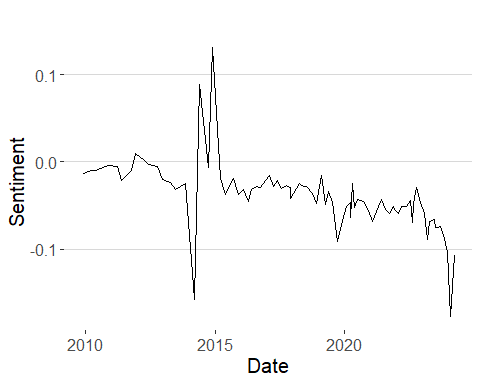
### 10.1 Plotting Wordfish Score

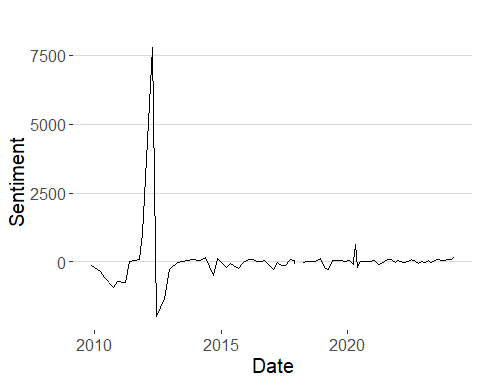
png   
 2



monetary policy committee statement pakistans economic   
-0.31839271 -0.26837970 -0.70257021 0.01878908 0.01161508 -0.19206118   
 growth track achieve highest   
 0.01481043 0.01878908 0.01591574 0.01161508

### 10.2 Plotting Wordscores Score





## 11 Conclusion

I hope that this document will be useful for researchers and practitioners who are interested in text mining and sentiment analysis. I have demonstrated how to use the R programming language to perform text data analysis based on monetary policy statements. The code provided in this document can be used as a starting point for further research and analysis. The R programming language is a powerful tool for text mining and sentiment analysis, and it is widely used in the academic and business communities. I hope that this document will help to promote the use of R in the field of text mining and sentiment analysis. This reproducible document will also serve the purpose of how to automate the process of text data analysis and sentiment analysis. I also demonstrate that R langugage has made it easy to perform text data analysis and sentiment analysis and has become a powerful tool for text mining and sentiment analysis, and it is widely used in the academic and business communities. My initial findings suggest that SBP policy has nuetral tone and it is not biased towards hawkish or dovish. However, the sentiment analysis is based on the text data and it is important to consider the context and the content of the text data.

Benchimol, Jonathan, Sophia Kazinnik, and Yossi Saadon. 2022. “Text Mining Methodologies with R: An Application to Central Bank Texts.” *Machine Learning with Applications* 8 (June): 100286. <https://doi.org/10.1016/j.mlwa.2022.100286>.

Shapiro, Adam Hale, and Daniel J Wilson. 2021. “Taking the Fed at Its Word: A New Approach to Estimating Central Bank Objectives Using Text Analysis.” *The Review of Economic Studies* 89 (5): 2768–2805. <https://doi.org/10.1093/restud/rdab094>.