Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From Statements to Insights: LDA-Based Analysis of Pakistan's Monetary Policy Statements

Said Abbas
saidabbas600@gmail.com
Quaid-I-Azam University Islamabad
Zahid Asghar
Quaid-I-Azam University Islamabad

Research Article

Keywords: Monetary policy, Textual Analysis, NLP, LDA

Posted Date: April 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9283792/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

This study explores the effectiveness of central bank communication by applying Natural Language Processing (NLP) to the Monetary Policy Statements (MPS) of the State Bank of Pakistan from 2005 to 2024. To quantify the textual features of the policy corpus, we construct a Document-Term Matrix (DTM) and deploy Latent Dirichlet Allocation (LDA), an advanced topic-modeling technique, to identify latent economic themes. By integrating these machine learning techniques with economic indicators, we analyze the complex interplay between monetary policy narratives and macroeconomic outcomes. Our findings uncover hidden textual trends that significantly influence economic decision-making, particularly regarding investment behavior. Ultimately, this research emphasizes that a transparent communication strategy coupled with an optimistic policy tone is critical for building public trust and fostering economic stability. This paper contributes a novel, data-driven perspective to the literature on monetary policy in emerging markets, offering actionable insights for policymakers to enhance communication efficacy.

1 Introduction

For macroeconomic stability and economic growth, monetary policy is the most significant tools available to every country in the world. In essence, it has an effect on the most important macroeconomic indicators, such as inflation, GDP growth, unemployment rates, and exchange rates. Pakistan, like other countries in the world, has a developed body or structure that is considered to be responsible for all of these known things. This body or structure is the State Bank of Pakistan, or SBP, which heavily relies on their monetary policy statement to convey their message or communicate with economic situation, economic measure, and stock and share market expectations. It is usual practice for monetary policy statements to include economic statistics and numbers, policy narratives, and suggestions for market stability in the country. The purpose of these statements is to promote trust in the public realm.

During the course of the last several decades, there has been a consistent evolution in central banks all over the word. These central banks have become more open, accessible, and clear in order to make their operations more understood to the people and stakeholders in the country. The findings of (Gürkaynak, 2005) have demonstrated that statements of monetary policy that have been carefully prepared are universally regarded as being an effective form of communication. It has been hypothesized by Hansen and Hansen, et al. (2016) that the statements of monetary policy papers throughout the past several decades have also undergone changes in terms of the tone of the words and the thoughts that they convey.

On the other hand, there is a great quantity of literature that is available on Pakistan's monitoring policy. This is mostly accomplished via the utilization of technologies that are more technologically advanced and in the field that are well researched by other researchers in the country. According to Khan and Qayyum (2007), the majority of the work that is being done in this field is based on the same three criteria, which are inflation objectives and interest rates. These criteria are developed for the purpose of better understanding and analyzing many other methods that are also significant. At this time, there are a great deal of sophisticated techniques for analyzing monetary policy statements. This gap in the literature highlights the importance of some advanced techniques and the necessity of using them in order to better understand the narrative of monetary policy documents. Therefore, in order to take into account all of the factors, an advanced textual analysis technique is used in this research to analyze the communications of the State Bank of Pakistan (SBP). Studies on this topic are being conducted in Pakistan at a lower rate.

In the discipline of textual analysis, advanced computational methods have revolutionized the understanding of the word through the use of advanced techniques. It is said by (Gentzkow et al., 2019) that there is a substantial quantity of unstructured text data from which researchers may obtain a significant number of themes, tones, attitudes, and patterns. Through the utilization of these contemporary methodologies, such as sentiments analysis, it is now possible to analyze an emotional tone, which is considered to be quantitative. In a similar manner, (Blei et al., 2003) created the Latent Dirichlet Allocation (LDA) algorithm, which provides the latent themes or hidden themes that are present within the text corpora (corpus) of policy statements. Tetlock (2007) conducts analyses based on a variety of documents, including annual reports, media narratives, and investor communications, among others.

Within the context of Pakistan, this particular technique was utilized for the very first time in the working study of the State Bank of Pakistan (SBP), in which they analyzed various monetary policy statements spanning the years 2006 to 2015. The primary emphasis of their investigation was on the fundamental framework of Pakistan's monetary policy. Another piece of study that was published by the State Bank of Pakistan (SBP) was Zaheer et al. (2019), in which the authors studied the minutes of the Pakistani monetary policy committee as well as the voting record of the monetary policy committee.

There is no such research in the literature that completely captures Pakistan's monitoring policy documents by comprehensively capturing all the thematic material of the State Bank of Pakistan. This work focuses on the overall sentiments analysis utilizing various approaches in unsupervised technologies of machine learning. Among them, the most prominent and acknowledged techniques are the ones that are discussed in this study. This gap in the literature was identified by looking at it. This research focuses on NLP and

LDA in order to gain a better understanding of the strategies that have the potential to influence how people see the market and the overall economic outlook of the country.

Objective of the study

Topic modeling for the analysis of monetary policy documents this work emphasizes identifying theme patterns and their progression over time utilizing LDA and bespoke topic modeling tools. This will assist in identifying the primary areas of attention for the central bank (SBP) regarding the enhancement of economic performance, problems, and policy suggestions.

Research Question

What are the dominant and key topics in monetary policy statements and their overtime evolution? And what is the overall thematic trend of Pakistan monitory policy documents?

2 Literature review

Importance of textual analysis is not anymore hidden from the main stream arena of literature. A lot of work have been done in the said area by many researchers of the field .now it’s time to adopt the advance techniques which is probably need of the moment and time. As we have observed in many studies that there are many techniques used from basic to advance this work includes some of the prominent studies which can be included in the review of existing literature.

Bora et al. (2024) Used the application of machine learning techniques of NLP to Federal Reserve staff documents prepared before policy decisions these documents are having different set of information on the basis of which monetary policy is made by using these documents and applying machine learning techniques for the prediction of changes in macroeconomic variables notably like interest rate and monetary policy shocks in the coming policy brief. By examination of these documents and analysis of the content through textual analysis this research finds out the response from the macro variables to the monitory shocks are in the parallel alienation with the theoretical content of the policy documents and insights can be derived from the text of the policy documents before meeting of the committee.

Sharma et al. (2024) have examined Central Bank of Nepal’s (NRB) monetary policy reports for an almost two decades ranging from 2002 to 201 .the main objective was that does the length, tone of the policy documents, predictive nature an d readability of the documents is having any impacts on the economic outcome of the country .they have noted that Central Bank of Nepal’s (NRB) monetary policy reports are noted to be longer during the time of downturns and have quite negative tone and insufficient amount of contents

which predict better prospects of the future. Policy documents are noted not to be easily understandable by the public. They have concluded that sentiments of the Central Bank of Nepal’s (NRB) monetary policy reports says each and every thing about the forth coming challenges and good news in the economy.

The analysis of Kurowski et al. (2023) is quite different from other studies in a sense that the y have that how many time or how frequent there is a debate about financial stability in the reports and other supporting documents of the monetary policy committee. by analyzing all the documents they have found that these topics are discussed in almost all the reports and they are having an average of 2% in the content of policy reports in countries like hungry and Poland .this work 2wa a new contribution to the literature by combing two different sort of documents which are monitory an financial document in post-communist countries.

Shah (2023) used Federal Open Market Committee (FOMC) communication content from their speeches press release and meetings minutes .the have preprocessed all the data and content through tokenization and then building a large data set from the available content through all the means of Federal Open Market Committee (FOMC).they have further advance their working by addition of hawkish-dovish mechanism of central bank policy where they have to find the tone of center bank policy communication. Further advancement was mad bay addition of LLM modes in their study .Their main objective was to find out the impact of policy communication on financial markets outcome. Their results highlights the role of FOMC documents on the behavior of market and providing the emerging technique in the understating of the how market reacts to FOMC documents.

Park et al. (2023) adopted textual analysis techniques on the central banks speeches to know about the potential risk in the financial markets due that speeches. They have done this work by direct extraction of the economic key words from the monitory policy committee speeches for the detection of ongoing scenarios in the economy. They found that when monitory policy speeches are closely observed it can yield better mechanism of halting with challenges unlike conventional mechanisms.

Nogueira et al. (2023) examines the central banks communication and monetary policy by utilizing textual analysis technique is a way of their analysis .after analyzing the documents they have said that there is asymmetry in the information of policy documents content. They extend the work by addressing that how it influence macroeconomic variables and financial variable by the insights from the policy documents by the central banks .they concluded that financial market condition reacts the policy documents in both positive in negative ways as per the textual content of the documents .

Cherry et al. (2023) used textual analysis techniques in their research by using Reserve Bank of New Zealand's (RBNZ) communication documents .the main focus of the study

was that how these documents affects the understanding of public on the main objectives of Reserve Bank of New Zealand's (RBNZ) policy like inflation and price stability. They have find out that the key words or frequent terms in the documents are in parallel with the objectives of their policy. they further extend that understanding of the documents is not always easy for the public but after 2018 It is a bit easy with the introduction of monitory snapshots by Reserve Bank of New Zealand's (RBNZ).the main focus of the study was on the content and materials available and their understanding in the public and the find that over time its improved form understanding point of view of the documents through the key words used in the documents.

Cai (2021) introduce this techniques in their work by analysis the content of Federal Open Market Committee (FOMC) and their content and they were in the search to find it potential impact on financial markets .in tier work they have shed light on (FOMC) reports content to now about the expectations of private market agents and the expectation given to them by content of Federal Open Market Committee (FOMC) reports. From the content they work out on different type of bound issued by government and their yield which can be significantly affected by content of Federal Open Market Committee (FOMC) reports. This study is worth importance due to highlighting the importance of computational and machine learning techniques like textual analysis in monetary policy documents and they further uncover the implication of these documents on markets outcomes.

Pandey et al. (2021). Analyzed the monetary policy documents of the Reserve Bank of India (RBI) using textual analysis method .the aim behind the work is the content of the documents and the steps involved in the process of making these documents for the inflation forecast in the economy. They have find out the monetary policy committee use the key macroeconomic variable to forecast inflation in the country but sometimes them fails to address the disinflation scenarios. They consider it is misjudgment of the committee members of monetary policy. They proposed that the Reserve Bank of India (RBI) need to refine the idea and adopt new ways for making and designing monetary policy by including the economic realities of the country.

Ochs et al. (2021) have analyzed that whether monetary tightening is having any true expansionary effect and how the agents interpret shocks through different signals from central bank assessment .they have further analyzed that differentiate the content and the policy actions on the documents. This study have used textual analysis approach in their analysis the study finds out that by applying this new techniques that its clear from the content that monitory policy is not at all expansionary is its said from the economic signals point of view in the recent time by providing clear contents and insights can have true effects on monetary policy communication.

Handlan et al. (2020) this work have specifically focus on communication of the federal reserves in which they have taken documents from Federal Open Market Committee (FOMC) to find the actual impacts and shocks on the price fed funds futures (FFF) on the basis of rate stele by the FED .they have found the variation due to Federal Open Market Committee (FOMC) reports multiple times in the targeted rate by FED .they also found that there are clear response to other variables than fed rate like inflation which aligns with macroeconomic models on standard basis.

Luangaram et al. (2017) examines a large set of the monetary policy documents by applying different techniques of computation and modern linguistics to find out the inflation targets in the markets mostly emerging markets. This work have focused on the main areas which are readability of the documents ,tone and topics mentions in the content of the documents .they have found out that central banks documents are not easy to understand due to the fact that they have used complex jargons which need advances skills and understand of the area. They have found out that central banks mainly focus on inflation in their reports. The tone of the documents says all about changes in the interest rate and economic growth.

Apel et al. (2012). This work have taken into account the impact of central bank communication and monitory policy specifically focusing on the opinion spectrum monitory policy committee members .they have shown that it enhances the applicability and real effectiveness of monitory policy when committee members lies on the same page .It can be clearly observe from the tone and sentiments of the policy documents .they have concluded their work by highlighting the clarity of the document to be understandable which make effective market outcomes.

LDA

Recent technology improvements have enabled economic and financial data analysts to more efficiently derive policy direction from the qualitative information contained in policy papers, such as Monetary Policy Statements. Hidden patterns and correlations within textual data

Topic models are Bayesian statistical frameworks that methodically arrange extensive collections of discontinuous and unstructured textual data, known as a "corpus," based on underlying themes defined as "topics." LDA is an unsupervised machine learning method.

Blei et al. (2003) conducted a significant study. This technique was introduced for the first time in this research; the model functions hierarchically, initially at the document level, subsequently at the topic level, and lastly at the corpus level.

Damane, M. (2022) states that Latent Dirichlet Allocation (LDA) functions as a generative statistical model for topic modelling in natural language processing. The concept asserts that documents have many subjects, each characterized by a distinct distribution of words. This is mostly attributable to the unstructured nature of the text data.

Azqueta et al. (2020) used the Latent Dirichlet Allocation (LDA) model, introduced by Blei, et al. (2003), functions as an unsupervised machine learning approach designed to identify the latent themes inside a collection of texts. This approach utilizes a generative probabilistic model to ascertain the word distribution characterizing a subject, while annotating articles with a distribution of themes.

du Plessis et al. (2024) introduces statistical models into textual analysis, namely known as Word Scores and word Fish. The models are defined so that Word Scores is supervised. Statistical models and word fish unsupervised. This work is regarded as significant primarily because it is the first of its kind to include statistical models into textual analysis.

3 Methodology

This chapter is comprised of data collection techniques and sources for our thesis and the methodology that is used to perform the analysis and how the study is proceeded. This work diligently obtained the Pakistan Monetary Policy Reports from the State Bank of Pakistan (SBP) website (https://www.sbp.org.pk/m_policy/mon.asp), published bi-monthly —six times annually. But this frequency changes in some of the years due to some unknown reasons. If we look into the documents, in some of the years we found only two documents, some are having three, but most of the years are having 6, so we will not disturb the frequency, and all the documents are used as it is in our thesis. We have obtained documents spanning almost two decades, 2004–2024, for our work. Initially the documents were in pdf form, and the requirements of our analysis are in text format. For converting MPS documents, an online application called pdf mate was used throughout the process.

The data set includes 115 documents covering the bi-monthly frequency of the documents showing the overall evolution hidden themes in the policy documents.

3.1 NLP techniques

NLP is a widely used technique in different areas of study that is more often used in unstructured data. This study follows the same process of NLP that was previously used in the literature. Bora et al. (2023) have used NLP for fed documents to find out monetary policy shocks. Ash et al. (2023) focused on techniques in which the main focus is on document representation through word vectors, word embedding, and sequences in models by applying the mainstream algorithms. This work is considerably the

methodological tradeoff between econometric challenges by the integration of textual analysis.

Pre processing

By adopting NLP it is important to include all the preliminary steps to further proceed the process of b analysis .the process includes changing of the document to make it machine readable and enhances the NLP models. Steps involved in the process are as follows

Lower casing:

Conversion of all text in to lower case format to enhance smoothness and consistency of the model like (Monetary to monetary).

$$T_{lower} = \left\{ t_{\{i\}} \ | t_{\{i\}} \notin \mathrm{T}, t_{\{i\}} = \mathrm{tolower}\left(t_{\{i\}}\right) \right\}$$

1

.....

T_{lower} Gives the collection of all the words and terms converted to lower case.
$\mathrm{tolower}\left(t_{\{i\}}\right)$ is used to convert specific tokens to lower case.

Removal of stop words:

This steps involve removal of words like “and” or “the” which make the documents more reliable and easy for machine to process

$$T_{stopwords} = T_{lower} \setminus \mathbf{S}$$

2

.....

$T_{stopwords}$ Provides a set of tokens after the removal of stop words and $\mathbf{S}$ shows us set of special characters.

Removal of punctuation and special characters

This steps is used to remove unwanted characters and punctuations like comma, full stop, Collin and semi Collins etc.

$$T_{punct} = \left\{ T_{punct} | t_{\{i\}} \notin \mathrm{p}, t_{\mathrm{i}} \in T_{stopwords} \right\}$$

3

.....

T_{punct} Is the combination of different terms $t_{\{i\}}$ after the removal of punctuations and special characters.

Tokenization:

This step changes the document in to small words or sentences which is known is units.

$$T_{tokens} = split(T_{punct}, \text{delimiters})$$

4

.....

T_{tokens} Is obtained by splitting T_{punct}through some define delimiters which are in this case spaces, commas and indents in the documents.

Stemming and lemmatization:

This is basically the conversion of word to their base form sometime the word become really short such as “running” to “run”.

$$T_{stem} = \{ stem(t_i) | \ t_i \in T_{tokens} \}$$

5

.....

T_{stem} Is set of tokens obtained by converting tokens into its roots form or base form.

Removing Numbers:

When the process moves forward her comes another part in which all the numbers are removed any numeric form as eliminated from the document.

$$T_{cleaned} = \{ t_{\{i\}} \ | t_{\{i\}} \notin \mathbb{N}, t_{\text{i}} \in T_{stem} \}$$

6

.....

$T_{cleaned}$ it is the cleaned set of tokens after removal of numbers and character which are non in text or linguistic format.

Whitespace Normalization:

Removal of extra spaces are tabs which are there in the documents is completely removed and makes the document in a structured format

$$T_{noramlized} = normalized_cleaned\,(T_{cleaned})$$

7

.....

3.2 Topic Modeling

Topic modeling is used during the process of NLP techniques which is a specialized mechanism and considered is unsupervised method in which is used to I identify the latent themes in a large text corpora and give us hidden patterns. There are many techniques present in the literature for textual analysis but here a specialized technique is used which is known is Latent Dirichlet Allocation (LDA). Ambrosino et al. (2018) have used LDA is a topic modeling technique for the economics literature valuable on a publishing site called jstor to identify the pattern of research approaches in the filled.

Latent Dirichlet Allocation (LDA)

Blei et al. (2003) have introduced revolutionary model in the field of machine leering which was a generative probabilistic model used for a huge data sets or collection of discrete data which is usually considered as text corpora or corpora of the text. Main motive behind this work was to find the hidden or latent patterns present in large set of documents.

The joint probability of a topic $\mathcal{Z}$ and a word w given the document d is:

$$\mathrm{P}\left(w,\ \mathcal{Z}\mid\alpha\ ,\ \beta\right)=\prod_{\mathrm{d}=1}^{\mathrm{D}}P\left(\theta_{\,d}\mid\alpha\right)\prod_{N=1}^{Nd}\sum\ P\left(\mathrm{z}_{dn}|\theta_{\,d}\right)P\left(w_{dn}|\mathcal{Z}_{dn},\beta\right)d\theta_{\,d}$$

8

.....

Where,

- α Presents the parameter of the Dirichlet prior for topic distributions.
- β Shows the parameter of the Dirichlet prior for word distributions.
- $\theta_{\,d}$ Is the topic distribution for document d.
- $\mathcal{Z}_{dn}$ Gives the topic for the n-th word in document d.
- w_{dn} Provides the n-th word in documents.

4 Data analysis and results

Pre processing

Initially, all the documents in PDF format were converted into text (.txt) format using a conversion tool named PDF Mate. Subsequent to the conversion, the text was organized into distinct documents, which were then transformed into a corpus. The corpus

consolidates all documents into a singular text document. To organize the document, the following processes were undertaken to create a coherent corpus.

Stop word removal and lower casing

Prior to the creation of the corpus, the documents were refined to remove extraneous words like "and," "the," and "is." The removal of these words was aimed at improving the clarity of the text. All text was transformed to lower case; for example, the term "Economy" was changed to "economy" following the conversion. The transformation of all upper-case words into their lower-case words was precise enough. It is important to highlight that all these modifications were executed through R programming codes, rather than manually.

Tokenization and lemmatization

The text was converted into tokens. Tokens refer to the conversion of all words into a single unit, a process known as tokenization in the field of machine learning. This method transforms the documents into an appropriate format while ensuring clarity and precision in the content. Lemmatization involves converting words into their base dictionary format. For instance, the word "running" will be simplified to "run”. The process of converting all words into their base dictionary format is known as lemmatization of the text.

Corpus creation

All of the above steps are part of the process of corpus creation. For the corpus to be smooth and understandable, we need to take all the preliminary steps that make it smoother, noise-free, and easily understandable for the machine. In this case, we use R programming throughout the process.

Bigram

Natural language processing uses a bigram to represent words that are closely related or adjacent in the corpus. We can extend it by referring to it as an n-gram when we have multiple sequences. For example, if there are two sequences, then n = 2. Bigram works in NLP as a pipeline. The probabilities of the words present in the text shape these pipelines. Probability formula for bigram

$$P\left(w2 \mid w1\right) = \frac{count(w1, w2)}{count\left(w1\right)}$$

9

.....

Where we have

- P (w1 | w2) are the probabilities of w2 given w1.
- Count (w1, w2) refers to the count of bigram (w1, w2).
- Count (w1) refers to count of (w1).

By doing all the process the table below represents bigrams of words.

Table 1
Most Frequent Policy Bigrams

Bigram	Frequency
private sector	675
current account	540
foreign exchange	386
account deficit	350
fiscal deficit	340
banking system	329
exchange rate	306
sector credit	300
external current	260
cpi inflation	248

Source

Created by author using R Studio.

The above table was generated on the above formula using textual analysis techniques in R programming. Given the presence of two words, "monetary" and "policy," the formula generates a bigram. The number 675 represents the frequency with which "private" and "sector" appear together in the documents. Now all these words are created on the same formula mentioned earlier. Table 1 displays the top 10 bigrams from the analysis. Visual representation of these bigrams in the form of a co-occurrence network enhances their understanding.

Co-occurrence network

Co-occurrence network represents nodes and the nodes shown in yellow colour represents a particular word in the document. And the edges (line drawn between to nodes) represents the connection of these two nodes which in other term know as a bigram .this

network shows the relationship between tow word as shown in Fig. 1. In Fig. 1 top 150 word selection is depicted in the form of bigrams

Frequent terms from DTM

As discussed above the frequent terms in the documents which are obtained from Document term matrix having unique characteristics are shown as bellow in Table 2 .the word percent is disused with frequency of 4045 which shows that document are having more discussion about the word percent it general refers percent of interest rate in MPS. Second highest frequent word is inflation MPS always discuss inflation inn relevance interest rate in their documents.

Table 2
Top 10 Most Frequent Unigrams in SBP Corpus

Term	Frequency
percent	3246
inflation	2357
rate	2142
sbp	1934
billion	1831
growth	1817
sector	1602
increase	1592
fiscal	1422
government	1297
Source: Created by author using R Studio	

Source

Created by author using R Studio

Above bar chart shows us the same results as depicted in the above table which are having the highest number time the word used is percent then the word inflation .all these main words depicts the intention of the documents that what will be the scenario.

Source

Created by author using R Studio

Document term matrix

Document term matrix is widely used phenomena in textual analysis for quantifying the frequency and terms present in collection of documents .in this research it is used to quantify the term and frequencies used in the analysis of monitory policy document (MPS). Table 3 depicting Document term matrix of MPS documents the following tables is having the given classification .Rows of Document term matrix represents documents present in the corpus and columns represents frequency of the terms.

Table 3
Document Term Matrix (DTM)

Document ID	billion	growth	inflation	percent	rate	sbp
doc__1	6	8	10	16	8	15
doc__10	6	14	18	13	8	0
doc__100	4	8	13	12	5	8
doc__101	3	11	18	18	6	0
doc__102	9	7	20	16	4	0

Source: Created by author using R Studio.

Above word cloud is also the visual interpretation of DTM frequent terms. these visualization help in understanding the real issues and things for which central banks are working .from the above world cloud stake holders can decide that where the country is heading .

Customize topic modeling Gibbs

Customized topic modeling utilizes various sampling techniques, with the Gibbs sampling method being the most commonly used. Porteous et al. (2008) explained the importance of Gibbs sampling for finding the hidden themes in the text through LDA. This research adopts the Gibbs sampling of LDA in the analysis of monetary policy.

LDA indicates the presence of numerous themes in the texts, along with additional disturbing topics that will surface during the process (Blei et al. 2003). Their inherent programming utilizes Gibbs sampling, a Markov Chain Monte Carlo (MCMC) technique, to obtain the main themes.

Topics (themes) are selected from the text of monetary policy documents utilizing the above-discussed machinimas; the below plot will help in the number of topics for our analysis.

Figure 4 **Topic Modeling Gibbs (Customize)**

Source

Created by author using R Studio.

Griffiths et al. (2004) indicate that a lower separation in the bottom matrix of Fig. 4 leads to improved results when analyzing the number of topics at 8. The lower matrix functions as the matrix for maximization, and through the application of the previously discussed techniques, it can be concluded that a reduced number of topics produces superior outcomes. Eight themes have been chosen for the analysis, which will be utilized in two distinct figures to facilitate distribution.

Top terms

Top terms by topic are I identified in each figure. In the first Fig. 5 is showing topic 1–4 and Fig. 6 is showing topics from 5–8.

Figure 5 suggests in each topic a distinct and unique idea which is obtained from monetary policy documents .in the first Fig. 4 key topics are discussed topic 1 is having full focus on themes like "fiscal," "percent," "inflation," and "year". Theses word is suggesting the discussions on fiscal measures, inflation outlook of the country and financial details.

Topic 2 focuses on "inflation," "sector," "billion," and "account." Which confirms the availability of discussion effects and their financial repercussion .while topic 3 giving terms like "rate," "policy," and "monetary. "Which gives information about central banking and their strategies. Topic 4 highlights terms "growth," "increase," "price," and "development." Which shows the discussion on growth drivers, price stability and development initiatives.

Figure 4. **8 Top Terms by Topic (1–4)**

Source

Created by author using R Studio.

As the discussion is done on the above terms and topics, the same will be done for Fig. 6, in which topics are highlighting terms like percent," "year," "growth," and "credit." It proposes future forecasts and projections of economic performance, the accessibility of credit in the economy, and, above all, it presents future economic trends from an economic perspective. Topic 6 introduces concepts such as inflation, price, deficit, and fiscal. Which is a clear indication of higher inflation and financial imbalances, which can be seen from the budget.

Topic 7 provides terms like "inflation," "monetary," "policy," and "rate.” In essence, it involves the interplay between two primary economic variables: inflation control and price control, which are crucial for the country's economic stabilization. Therefore, in topic 8, concepts such as "monetary," "bank," "reserve," and "policy" are introduced. The state bank plays a crucial role in managing reserves, implementing monetary policy, and, above all, ensuring economic stability for the nation.

Source

Created by author using R Studio.

Gamma values distribution across all documents

The given Fig. 7 represents the gamma distribution, which is known as the proportion of topics across all the monetary policy documents. The height of the box plot in the given figure represents the gamma value for each individual topic. In this scenario, a higher height of the box plot indicates a higher prevalence of gamma values within the documents. The width of the box reveals the characteristics of variability in the gamma values within the MPS documents. When the width of the box increases, it suggests the importance of the topic in all the documents. The dots present on the line above each box represent the outliers in the documents. These are box plots that highlight instances where the gamma values are either extremely high or extremely low in all the valuable documents.

In the provided graph, topic 1 (V1) exhibits a high gamma value, and by adjusting the width, it displays a range of variability across all documents. Upon examining topic 3 (V3), it is evident that its gamma values are higher than those of V1 and V2. Through a thorough analysis of the box plots, it becomes evident that the priorities of the monetary policy committee are changing over time in various discourses.

Source

Created by author using R Studio

5 Discussion

Monitory policy is consider to be the main factor in shaping the path for any economy .from broader perspective monetary policy is used in different ways like money supply management of interest rate, credit availability and inflation control in the country. By looking to all these things monetary policy documents provides a broader over view for the stake holder to take their future decision.

Throughout this research the main focus was on the sentiments analysis for finding the tone of monetary policy then uncovering the key areas or theme on which state bank of Pakistan is focusing on from the text after analysis priority or key Ares can be observed from the policy documents by using LDA techniques of topic modeling via Gibbs sampling method.

Thematic Structure of Monetary Policy Documents

The application of LDA is not hide from any one. From LDA this study has found that how the state bank of Pakistan in their monitory policy targeting key areas in the core of their policy. This research ha find out that text of monitory policy have mainly focused on some of the following key are s in the economy and the discussion is mostly revolving around the following areas

Inflation management

All the of the above is mainly the application of LDA from which these topics are obtained .Most frequently the word inflation is used in the documents of monitory policy (MPS).all the documents have highlighted and focused on the inflation throughout the analysis. And from monetary policy perspective it is evident that inflation and interest rate is closely connected. The analysis suggests that to combat inflation SBP rely on interest rates.

Interest Rate Policy as a Stabilization Tool

Interest rate was extracted from the corpus in a significant amount .the results openly suggests that the key tool for stabilization in the economy is interest rate which is always used .there are other factors like macroeconomic variables such is demand side pressure to reduce inflation is also significantly used in the MPS. Now as it can be seen that reduction of interest rate is linked with economic growth.

6 Conclusion

The State Bank of Pakistan issues monetary policy statements, and this study explored the hidden themes and topics of and tone of these statements to uncover the overall impact, identifying the key areas of focus. All the processes were taken out with the application of computational and machine learning approaches like NLP and latent thematic analysis through LDA topic modelling, from which the key themes of monetary policy statements (MPS) were observed. The analysis done on the MPS statements from 2005–2024 in which all the text of the documents was analyzed

Topic modelling (LDA) was used, which uncovers key themes on which monetary policy statements (MPS) strongly focus. The study identified major themes such as inflation,

interest rates, and a few others, suggesting that the primary goal of monetary policy is to promote stability in the country by using optimistic language and tone in their documents.

MPS need to transparently highlight key economic issues and concerns and provide clear guidelines for the public through their documents. Clear acceptance of challenges to the economy should be accepted to maintain the credibility of the documents among the public.

MPS document should provide information and concentrate communications on domestic economy. In the economy there are multiple sectors it should provide sector wise information and guidelines. According to these guidelines with clear information it will be easy for each sector to take decisions.

MPS needs to provide proper linkage of development goals with monetary policy like connecting infrastructure growth and other development initiative which should be in the public interest. MPS should use easy language in their documents for border understanding to ensure the readability of the documents for non-technical audience of the economy.

These recommendations are made in such way to make monitory policy more effective and understandable among any sort of audience. This will help to achieve economic stability and growth targets in more efficient way by bringing market and public in same line of understanding.

Future work

Future directions for the current work can be carried out by utilizing causal inference and attention mechanisms; however, transformer-based NLP models will be needed for the integration of macroeconomic data and the monetary policy document.

Declarations

Declaration of competing interest

Regarding the study done, the findings, and the publication of this work, the authors of this manuscript declare that they have no conflicts of interest.

Supplementary Data and Analysis Repository

The given link provides all the codes step by step used throughout the analysis.

https://github.com/saidabbas600/Textual-analysis-using-Lda

Author Contribution

I certify that each of the mentioned authors made a substantial contribution to the creation and composition of this piece.

Data Availability

The textual datasets (Monetary Policy Statements from 2005 to 2024) analyzed during the current study are publicly available and were sourced from the official website of the State Bank of Pakistan. The processed data, including the Document-Term Matrix (DTM) and the code used for the Latent Dirichlet Allocation (LDA) analysis, are available from the corresponding author upon reasonable request.

References

1. Apel, M., & Blix Grimaldi, M. (2012). The information content of central bank minutes (No. 261). Sveriges Riksbank working paper series.
2. Azqueta-Gavaldon, A. (2020). Text-Mining in Macroeconomics: the wealth of words.
3. Ambrosino, A., Cedrini, M., Davis, J. B., Fiori, S., Guerzoni, M., & Nuccio, M. (2018). What topic modeling could reveal about the evolution of economics. *Journal of Economic Methodology*, *25*(4), 329–348.
4. Ash, E., & Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*, *15*(1), 659–688.
5. Blei, D. M., Ng, A. Y., & Edu, J. B. (2003a). Latent Dirichlet Allocation Michael I. Jordan. In Journal of Machine Learning Research (Vol. 3).
6. Bora, S., Gan Aruoba, ˘, & Drechsel, T. (2024). Identifying Monetary Policy Shocks: A Natural Language Approach *. http://econweb.umd.edu/
7. Cai, D. (2021). Corpus Studies on the Use and Framing of Metaphor in Popular Economic Discourse.
8. Cai, Y., Camara, S., & Capel, N. (2021). It’s not always about the money, sometimes it’s about sending a message: Evidence of Informational Content in Monetary Policy Announcements. http://arxiv.org/abs/2111.06365
9. Cherry, R., & Tong, E. (n.d.). Analytical Notes Reserve Bank of New Zealand Analytical Note Series Words of the RBNZ: Textual analysis of Monetary Policy Statements.
10. Damane, M. (2022). Topic Classification of Central Bank Monetary Policy Statements: Evidence from Latent Dirichlet Allocation in Lesotho. Acta Universitatis Sapientiae, Economics and Business, 10(1), 199–227. https://doi.org/10.2478/auseb-2022-0012

11. du Plessis, E. (2024). Can Text-Based Statistical Models Reveal Impending Banking Crises? Computational Economics. https://doi.org/10.1007/s10614-024-10594-5
12. Gentzkow, M., Kelly, B., & Taddy, M. (2019). American Economic Association Text as Data. Source: Journal of Economic Literature, 57(3), 535–574. https://doi.org/10.2307/26787457
13. Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (n.d.). Integrating Topics and Syntax.
14. Gürkaynak, R. S. (2005). Using Federal Funds Futures Contracts for Monetary Policy Analysis Using federal funds futures contracts for monetary policy analysis *.
15. Handlan, A. (2022). Text Shocks and Monetary Surprises: Text Analysis of FOMC Statements with Machine Learning. https://sites.google.com/view/amy-handlan.
16. 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
17. Khan, M. A., & Qayyum, A. (2007). Trade, financial and growth nexus in Pakistan (No. 2007, 14). Economic Analysis Working Papers.
18. Kurowski, Ł., & Smaga, P. (2024). Communicating financial stability in monetary policy reports: a text-mining experiment in postcommunist countries. Central European Management Journal, 32(1), 57–75. https://doi.org/10.1108/CEMJ-04-2022-0054
19. Luangaram, P., & Wongwachara, W. (2017a). More Than Words: A Textual Analysis of Monetary Policy Communication.
20. Nogueira Ferreira, L. (2023). Essays on Central Bank Communication and Monetary Policy.
21. Ochs, A. C. R. (n.d.). CAMBRIDGE WORKING PAPERS IN ECONOMICS CAMBRIDGE-INET WORKING PAPERS A New Monetary Policy Shock with Text Analysis.
22. Pandey, A., Shettigar, J., & Bose, A. (2021). Evaluation of the Inflation Forecasting Process of the Reserve Bank of India: A Text Analysis Approach. SAGE Open, 11(3). https://doi.org/10.1177/21582440211033826
23. Park, J., Lee, H. J., & Cho, S. (2023). Hot topic detection in central bankers' speeches. Expert Systems with Applications, 230. https://doi.org/10.1016/j.eswa.2023.120563
24. Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008, August). Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 569–577).
25. Sharma, R., & Kafle, P. K. (2024). Decoding the directives: a textual analysis of monetary policy. *Global Knowledge, Memory and Communication*.

26. Shah, A., Paturi, S., & Chava, S. (2023). Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis. http://arxiv.org/abs/2305.07972
27. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance, 62(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x
28. Zaheer, S., Khaliq, F., Ahmed, W., & Omer, M. (2016). The Euro and the Battle of Ideas (Vol. 15). http://www.sbp.org.pk/research/bulletin/

Figures

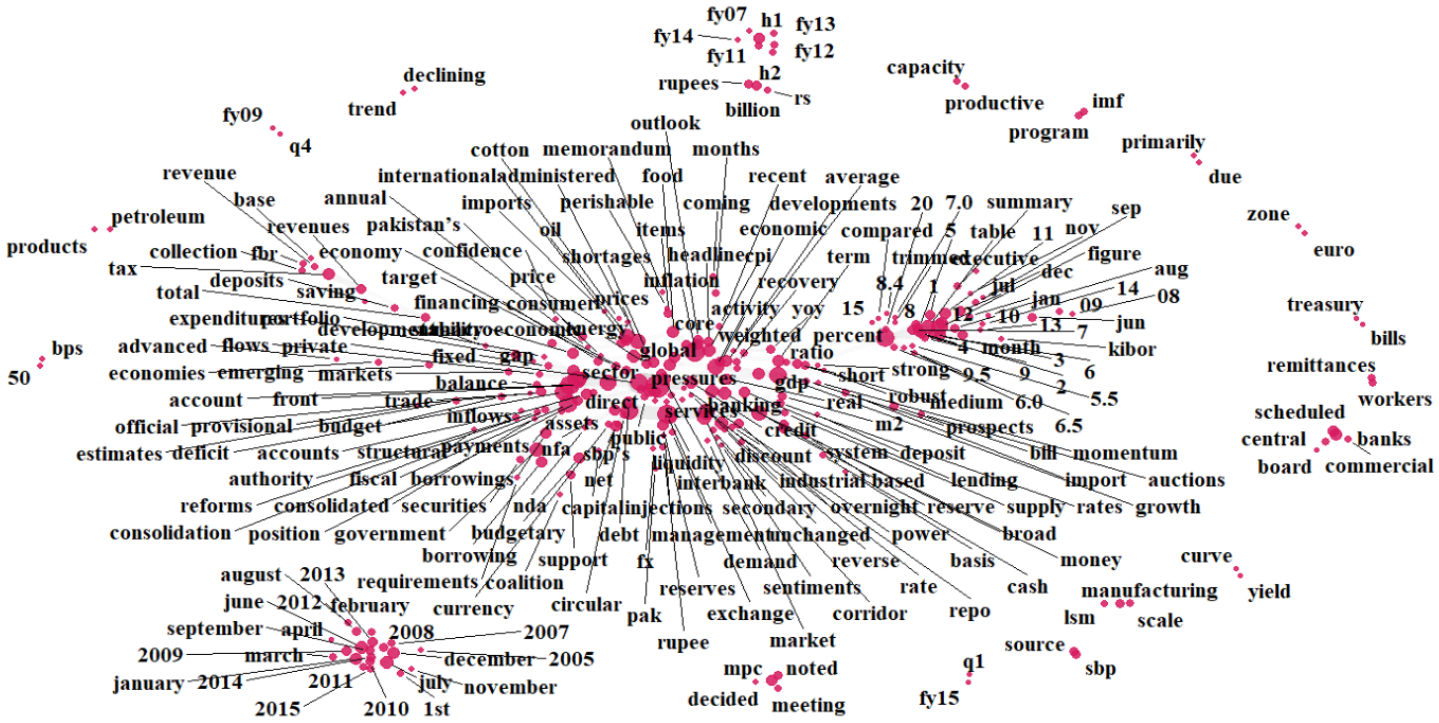

Figure 1

Co-occurrence network

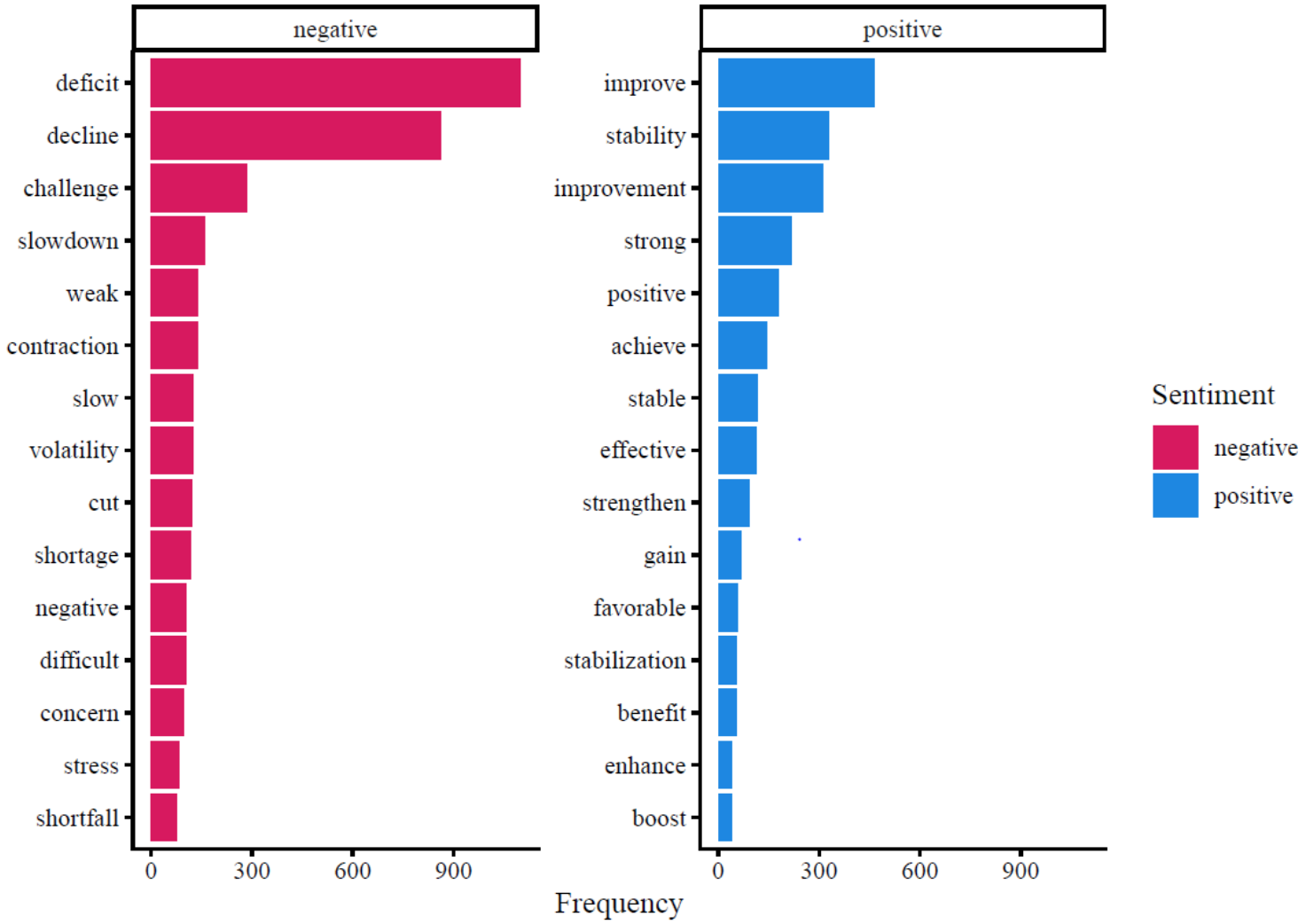

Figure 2

Most Frequent Terms Sentiment Wise

Source: Created by author using R Studio



Figure 3

World cloud

Source: Created by author using R Studio

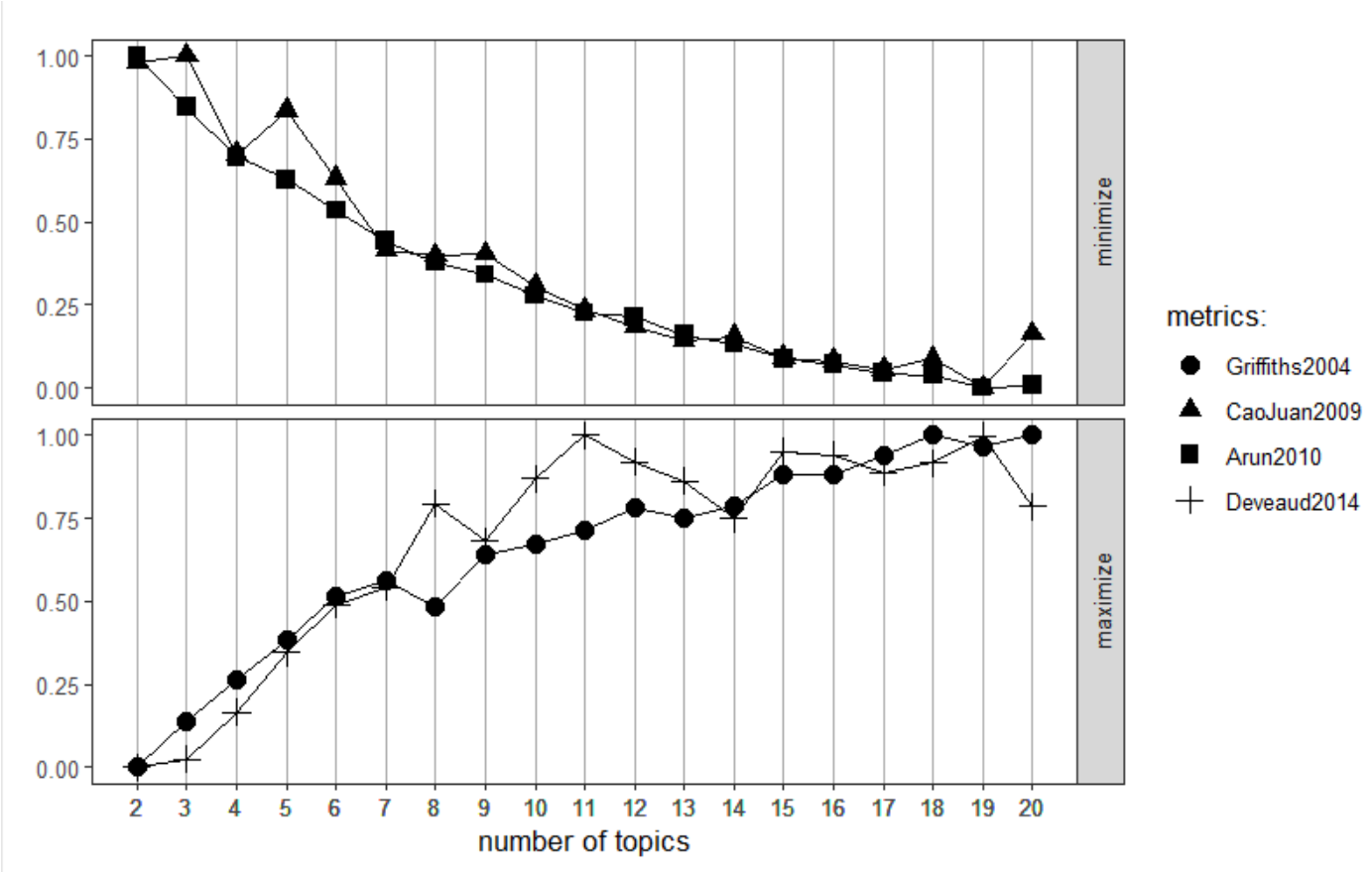

Figure 4

Topic Modeling Gibbs (Customize)

Source: Created by author using R Studio.

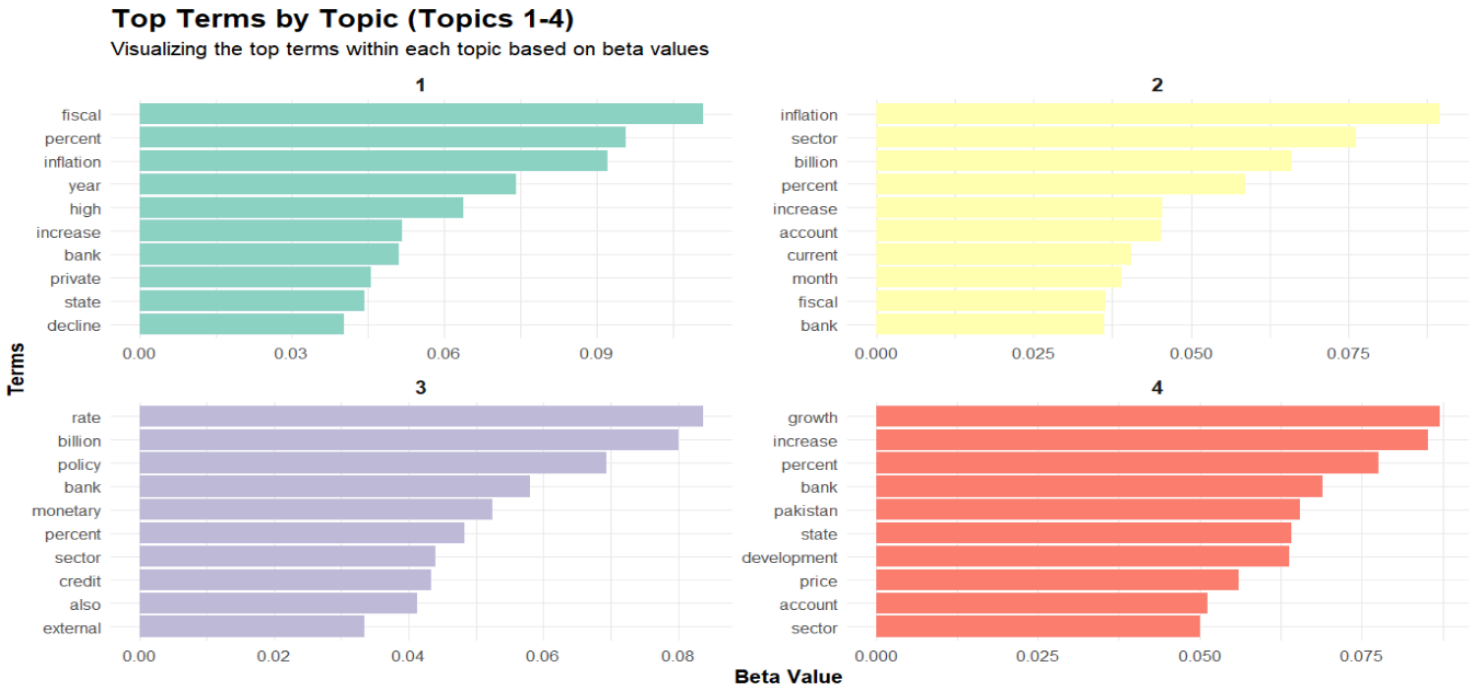

Figure 5

Top Terms by Topic (1-4)

Source: Created by author using R Studio.

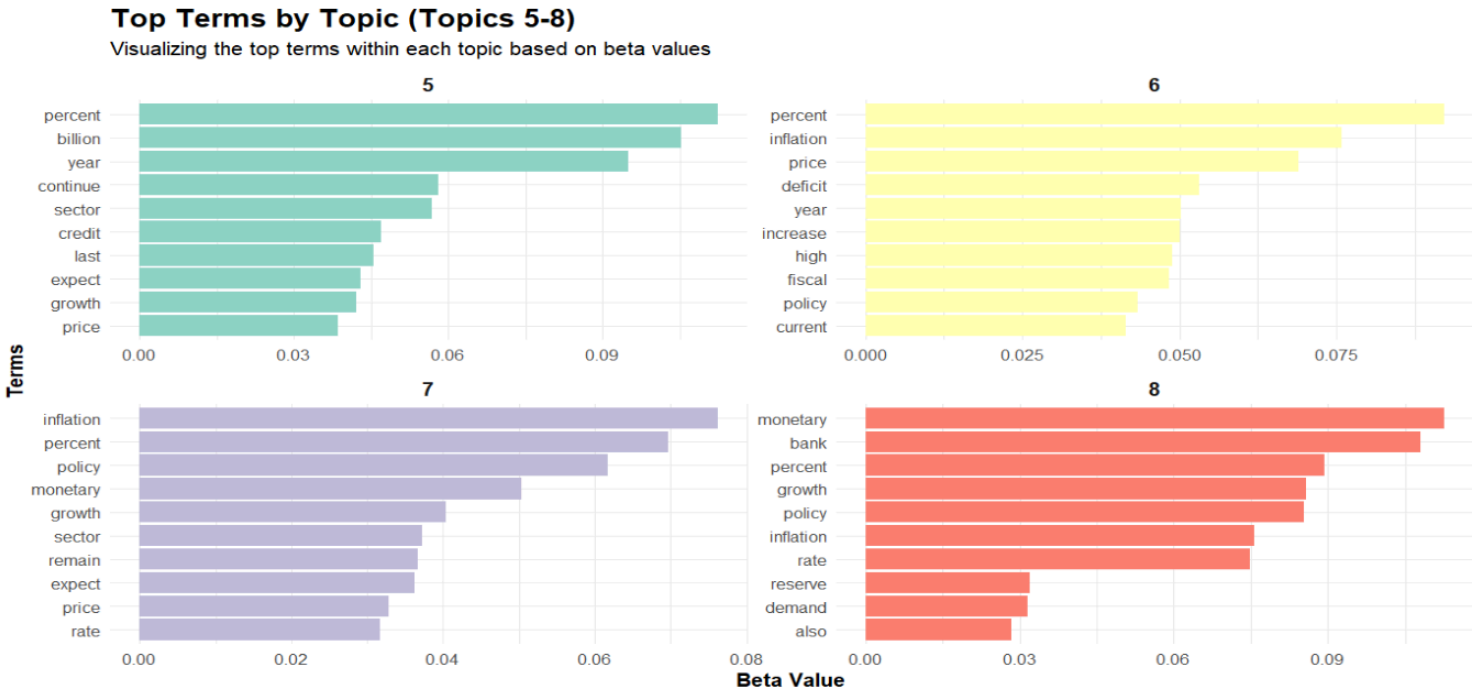

Figure 6

Top Terms by Topic (5-8)

Source: Created by author using R Studio.

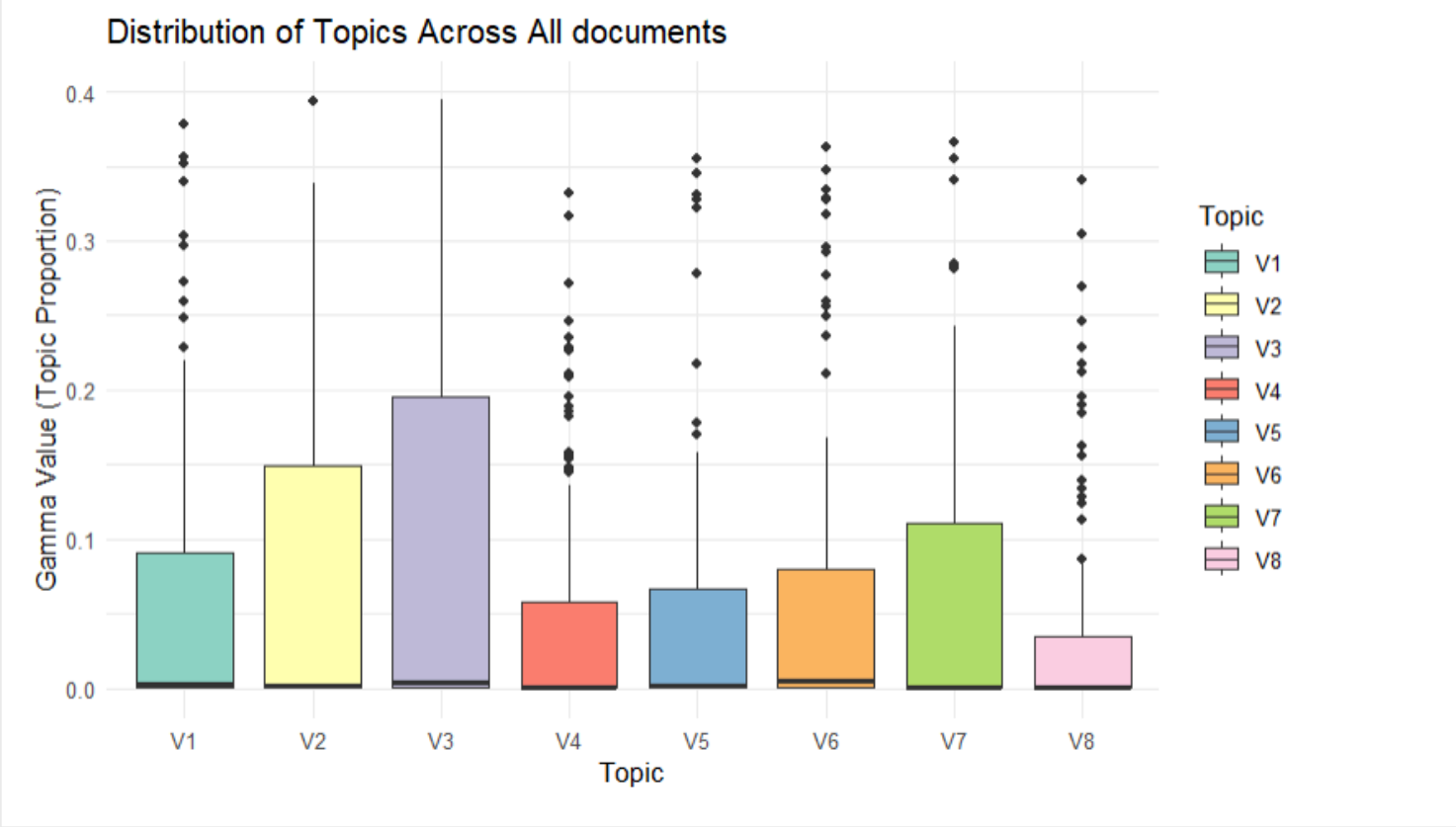

Figure 7

Box Plot of Topics

Source: Created by author using R Studio