Eskici, Hatice Burcu; Kocak, Necmettin Alpay

Article

A text mining application on monthly price developments reports

Central Bank Review (CBR)

Provided in Cooperation with:
Central Bank of The Republic of Turkey, Ankara

Suggested Citation: Eskici, Hatice Burcu; Kocak, Necmettin Alpay (2018) : A text mining application on monthly price developments reports, Central Bank Review (CBR), ISSN 1303-0701, Elsevier, Amsterdam, Vol. 18, Iss. 2, pp. 51-60, https://doi.org/10.1016/j.cbrev.2018.05.001

This Version is available at:
https://hdl.handle.net/10419/217320

Central Bank Review 18 (2018) 51–60

Contents lists available at ScienceDirect

Central Bank Review

journal homepage: http://www.journals.elsevier.com/central-bank-review/

A text mining application on monthly price developments reports

Hatice Burcu Eskici [a], Necmettin Alpay Koçak [b, *]

[a] Hacettepe University, Turkey
[b] Hacettepe University, Cankaya, Ankara, Turkey

ARTICLE INFO

Article history:
Received 10 April 2018
Received in revised form
12 May 2018
Accepted 12 May 2018
Available online 24 May 2018

JEL Classification:
D71
D83
E52
E58

Keywords:
Central bank
Communication
Reports
Text mining
Cluster analysis

ABSTRACT

Text mining analysis provides big opportunities for economic research. Underlying natural language processing techniques allow us to read the monthly price developments reports (MPDR) of the Central Bank of the Republic of Turkey (CBRT) and to analyse the words, to explore topics and clusters inside. Previous literature on CBRT documents has focused on making word clouds, measuring the sentiments and therefore it is limited with text documents. This study sets out to close this gap and extends the text mining analysis to measure the statistical consistency of the MPRDs with the annual consumer price index (CPI) inflation figures for Turkey. In this study, we showed that MPDRs contain intensifying references to core-groups/sectors in evaluation of inflation as well as they are interested in the tendency of inflation rather than its level. We also showed that how the clusters of MPDRs are significantly consistent with the annual CPI inflation figures from statistical point of view.

1. Introduction

The concept of an independent central bank, which has been enacted in the change in the law of the Central Bank of Republic of Turkey (CBRT) dated 5/5/2001 has brought transparency and accountability with it. Afterwards, it has been necessitated to increase the communication channels in the direction of enhancing the transparency of CBRT due to implicit and explicit inflation targeting strategies. The market agents are carefully examined all the communication channels used by CBRT and their economic and financial positions are determined according to the information they receive from these channels due to the global economically events such as crisis, turbulence, etc. especially occurred after the year 2008.

The impact of monetary policy tools on inflation realizes through “monetary transmission mechanism” (Christiano and Eichenbaum, 1992). The “expectations channel” in this mechanism is mostly affected by the communication channels of CBRT and the perceptions them create in the markets. Since the year 2005, CBRT has implemented the explicit inflation targeting, and it has been in a managerial position in the perceptions of the economic community with its communication channels (Kara, 2008). The communication channels of the CBRT can be divided into two parts (CBRT, 2016). The first is the Governor's speeches and the announcements about decisions of Monetary Policy Committee (MPC). The second is the research and publications which includes the reports, the monetary and exchange rate policy texts, the social media, the live webcasts, the press briefings, the Central Bank Review, the Research Blog, books, booklets and factsheets.

Among these communication channels, it can be suggested that the most influential in the short run is the Governor's speeches which is about the implementation of monetary policy, and the summary of the MPC meetings which reflects the decisions on short-term interest rates and the CBRT's thoughts on the economic situation. On the other hand, the reports under research and publications can be suggested as an important input to market agents in determining theirs medium-long term positions. The most important ones of them: the Inflation Report which presents current inflation targeting framework of the CBRT, the Monthly Price

* Corresponding author.
E-mail addresses: hburcueskici@gmail.com (H.B. Eskici), alpaykocak@hacettepe.edu.tr (N.A. Koçak).
Peer review under responsibility of the Central Bank of the Republic of Turkey.

https://doi.org/10.1016/j.cbrev.2018.05.001

Developments Report (MPRD) which analyses monthly inflation developments to be interpreted by the public in a healthier manner, and the Financial Stability Report which the CBRT publicly shares its views and assessments on the financial sector. A closer look at the CBRT's reports reveals that many reports and researches are either irregular or have been published at intervals of 3 months or more (for example, the Inflation Report is published for 3 months and the Financial Stability Report for 6 months). The most frequent and regularly published report in terms of the frequency of publications is the MPDR.

Text mining allows to measure and analyse documents using natural language processing techniques (Feldman and Sanger, 2006), it can also be defined as the ability of a machine to analyse, understand, and generate text. In text mining, all types of the text information regardless of the format can be transformed to numbers indexed for each of documents, thus unstructured text data become the structured data which are similar to be used in economic research. With the transformation of unstructured data into structural data, the framework of economic analysis is also expanding. However, we will deal with not only the text mining analysis of the comments made in MPDRs but also measuring the consistency of them with the annual consumer prices index (CPI) inflation figures over time in this paper. We will analyse the MPDRs for a long time-span. By examining the MPDRs, we will try to extract frequent-words and to identify topics, and to assign MPDRs into clusters according to the words used in them. We will have some considerations about the consistency of the MPDRs with the annual CPI inflation figures.

The rest of this paper is organized as follows. We start by giving an overview of studies in the literature which mostly examines the monetary texts with limited them text mining analysis. Then, we introduce the dataset and provide a short description of the text mining approach used in this paper. In the results part, we first present the results of the text mining analysis steps such as filtering, topic identifying and clustering. Then, we present the findings of analysis of variance in which the clusters of MPRD and annual CPI inflation figures are examined. It is concluded with some suggestions for further research.

2. Related studies

Since the texts of monthly reports has not been extensively studied in the literature, the review in this paper is mostly based on the quarterly and annual reports. Blinder et al. (2008) and Acosta (2015) studies have a very good compilation of how the text mining analysis is used in the examination of the central bank communication policies. Previous literature shows us that it has been used text mining techniques in so many different forms to analyse the documents, however these findings obtained from the analysis of text mining have been rarely used to obtain new findings in advanced analyses. As highlighted in Iglesias et al. (2017), the big empirical challenge for the text mining analysis is to transform the texts into significant variables which they can be used in another model or analyse.

From this point of the view, the closest papers in the literature are Lucca and Trebbi (2009) and Hansen and McMahon (2015). In the study of Lucca and Trebbi (2009), they applied computational linguistic tools to Federal Open Market Committee (FOMC) statements and measure the effects of them on the macroeconomic variables. They used a vector autoregression framework to measure the effects of the variables constructed by using FOMC statements on the macro-economic variables.

Similarly, Hansen and McMahon (2015) constructed some variables, using Latent Dirichlet Allocation (LDA) for topic modelling and dictionary methods to measure the sentiments, from FOMC texts to represent the dimensions of the monetary policy of Federal Reserve (Fed). Then, they examined the effect of these variables on macro-economic variables with a factor augmented vector autoregression approach.

Of course, the issue of central bank's communication has been studied not only for Fed's texts but also for other central banks'. In the study of Jansen and Haan (2010), they tried to assess the consistency of European Central Bank (ECB) Communication using Word scores method and they suggest that ECB's communication during the years between 2002 and 2009 of European Monetary Union (EMU) as internally consistent but sometimes was flexible. Also, Ranaldo and Rossi (2010) studied effects of Swiss National Bank announcements on the financial market.

Two studies attempting to analyse CBRT's its communication policy with text mining have been found in the literature as far as we can detect. These are the works of Kahveci and Odabas (2016) and Iglesias et al. (2017). However, these works have been limited to the analysis of texts only, and there is no attempt to make an advanced analysis using the findings of text mining.

Kahveci and Odabas (2016) argue that their study differs from the literature in terms of the period of analysis examined and the software used. In their studies comparing the communication policies of the Fed, ECB and CBRT, they found that the optimistic tone of Fed has decreased over time while its certainty tone has increased. The authors say that there is no significant tone difference in the certainty, the optimism and the realism of ECB and CBRT, however they emphases that there is a significant increase of the optimistic tone for the CBRT's statements of the years 2014 and 2015.

The main contribution of the work done by Iglesias et al. (2017) is seen as an improvement of a computational approach to incorporate dynamics and topic evolution by applying not only LDA algorithm, but also Dynamic Topic Models. From topics of the minutes and statements published by CBRT, they inferred that the communication policy of CBRT have been changing over time as the global economic conditions have changed substantially, and the complexity of monetary policy strategy of CBRT has increased as the financial crisis amplified capital flows volatility. In Iglesias et al. (2017), they highlighted that the focus of communication has strengthened in the recent period (last two years) and the tone confirms the strategy of providing price stability.

3. Data and methodology

3.1. Data

MPDRs are published by CBRT within one working day, following the announcement of price statistics by the Turkish Statistical Institute. These monthly reports have important roles since they include CBRT's initial remarks on the inflation figures and aim to contribute to a sounder interpretation of the monthly inflation developments by the public during the period between the announcement of the official price statistics and the MPC meeting. These reports mainly mention the developments in the CPI inflation, while at least the developments in the Producer Prices Index (PPI) inflation are mentioned.

In this paper, we examined the MPDRs reports published between June-2006 and Jan-2018. A total of 138 reports[1] those are available for this time interval have been analysed. These reports

[1] According to time span, there should be 140 MPDRs were published, but we successfully downloaded 138 reports from http://www.tcmb.gov.tr. Two reports in English are missing for the periods April-2009 and July-2012, however they are available in Turkish.

are available in both Turkish and English portable document format and the English version of the reports have been analysed. It is also possible to analyse the reports in Turkish language, but English versions are preferred due to the language of this paper is also same with them. Prior to the text mining analysis, the sections which contain tables, graphs, and images were removed from the reports to make text mining analysis more efficient since it can be negatively affected by the words those are constantly repeated in tables and graphs.

3.2. Methodology

In this part of the paper, we give brief theoretical explanation of the methods used in this study.[2] The reports are first standardized in text mining analysis. They can have in different extensions (word, excel, xml, html). In this case, file types need to be edited after the reports are collected. Keeping the reports in the same file type does not affect the analysis result but can accelerate the analysis process.

Standardized but unstructured text reports can be expressed as structured data by pre-processing techniques (parsing, segmentation, unification of synonyms, removal of ineffective words/stop-words etc.) and then creation of the document-term matrix (DTM) with the vector space model (Feldman and Sanger (2006)). The vector space model represents reports as vectors in an m-dimensional space and helps to express the text data by numbers (Kumar and Chandrasekhar (2012)). For instance, each document (d) is defined by a numerical feature vector (w(d)). It is given in (1):

$$w(d) = (x(d, t_1), \ldots, x(d, t_m)) \tag{1}$$

where $x(d, t_1)$ represents a word or group of words in a document. The simplest method can be used in the construction of DTM, which is an aggregated form of vectors ($w(d)$), is based on whether the corresponding word presents in the reports. In the cells of DTM, the binary variable has a value of zero or one, depending on whether the related word is in the document.

DTM can also be formed by measuring the frequency of terms (words) in the reports, such as weighted frequencies. But in most analyzes, it is not desired to work with such a comprehensive DTM. In that case, DTM must be reduced using some techniques. The term-frequency can be considered after words such as stop-words, conjunctions, pronouns are extracted from the reports. While the frequently used words are kept in the dictionary, others can be excluded from the dictionary. The size of DTM can be reduced by removing sparse words and using synonyms (Weiss et al., 2004). In the frequencies method, frequencies can be used in the cells of DTM instead of the binary variable mentioned above. For instance, if a word repeats 25 times on a document, then this value is in the cell of DTM. Because, the frequency is better in some algorithms. In the method in which frequencies are used, words having a lower frequency than a predetermined threshold value can be excluded, and DTM can be reduced. In addition to a simple threshold value, threshold determination methods such as chi-square, mutual information, odds ratio can be used to reduce the complexity of DTM (Weiss et al., 2004).

The other method (TF-IDF) used to construct the DTM is to arrange the numbers according to the importance of the word instead of determining the frequency of the words in the document. In this study, TF-IDF (full weighting) formulation is used to determine weights. In the following expression, the weight of full weighting for j is:

$$tf - idf(j) = tf(j) \times idf(j) \tag{2}$$

$tf(j)$ is the frequency of the jth term and $idf(j)$ is the logarithmic scale factor for the importance level of the word and is called the inverse document frequency (idf) and is shown in (3).

$$\mathrm{idf(j)} = \log\left(\frac{N}{df(j)}\right) \tag{3}$$

The idf evaluates the number of reports containing the word j (e.g. $df(j)$) and reverses the scale. Thus, if a word is included in more than one document, it is considered insignificant and the scale is reduced. If the word is included in a small number of reports, the scale value will be very high, and it will be significant (Weiss et al., 2004).

Afterwards, text mining methods such as the topic identification and text clustering can be applied on the structural data which the non-structural data are transformed into. Each document consist of topics and each topic consists of words. Topic identification enables to discover the document collection by associating words and reports. It is also possible to define user-topics. It can be summarized as the topics are collections of words that define and describe a main theme. The approach is different from clustering because clustering allocates each document to a unique group while the topic identification assigns a score for each document and term to each topic. Then thresholds are used to control whether the association is strong enough to consider that the document or term belongs to the topic. As a result, reports and words may belong to more than one topic or to none at all.

In this study, text clustering is performed as well as topic identification. The selection between text clustering and text classification depends on whether there is no external information related to texts (Hotho et al. (2005)). In the clustering methods there is no external information on texts, and texts are grouped according to their similarity taking their content into account. The purpose of the clustering method is to create groups those have the same characteristics in themselves but have significant differences among them. In this paper, we applied only text clustering method on the reports. What differs in clustering is that the clusters are not created at the beginning, and it is unknown how to cluster the data according to which characteristics. In order to say that the clusters are valid, it is necessary that the clusters should be interpretable by using the words which are inside themselves.

In this paper, the cluster method is based on the probabilistic approach which uses singular value decomposition. Namely, the expectation-maximization clustering algorithm (EM) is used in the clustering of reports. As explained in Hofmann (2001), it is assumed that a mixture model approximates the data distribution by fitting k cluster density functions, f_h ($h = 1, \ldots, k$), to a data set with a DTM in the expectation-maximization (EM) algorithm. The mixture model probability density function evaluated at point x is given in (4):

$$p(x) = \sum_{h=1}^{k} w_h f_h\left(x \middle| \pi_h, \sum_h\right) \tag{4}$$

where w_h is the proportion of data that belongs to primary cluster h. Each cluster is modeled by a DTM dimensional Gaussian probability distribution is given in (5):

[2] It is utilized SAS® Enterprise Miner(TM) software to analyse the MPDR's in this paper. Detailed information about the software can be found in https://support.sas.com/documentation/cdl/en/emgsj/64144/PDF/default/emgsj.pdf.

$$f_h\left(x\middle|\pi_h,\sum_h\right)=\frac{1}{\sqrt{(2\pi)^d|\sum_h|}}exp\left(-\frac{1}{2}(x-\pi_h)^T\left(\sum_h\right)^{-1}(x-\pi_h)\right) \tag{5}$$

where, μ_h and Σ_h are the mean vector and covariance matrix for each cluster h. EM clustering is an iterative process which includes first obtaining initial parameter estimates and then apply the EM algorithm to find primary clusters and to update parameter estimates. The EM algorithm uses the DTM at each iteration to estimate the model parameters. The algorithm terminates when two successive log-likelihood values differ by an amount or when a maximum of five iterations has been reached. For each observation x in the data set at iteration j, the parameter estimates of EM algorithm are computed as follows:

- Compute the membership probability of x in each cluster h = 1, …,k as given in (6),

$$w_h^j(x)=\frac{w_h^j f_h\left(x\middle|\mu_h^j\sum_h^j\right)}{\sum w_i^j f_i\left(x\middle|\mu_i^j\sum_i^j\right)} \tag{6}$$

- Update the mixture model parameters for each cluster h = 1, …,k as given in (7–9),

$$w_h^{j+1}=\sum_k w_h^j(x) \tag{7}$$

$$\mu_h^{j+1}=\frac{\sum_k w_h^j(x)x}{\sum_k w_h^j(x)} \tag{8}$$

$$\sum_h^{j+1}=\frac{\sum_k w_h^j(x)\left(x-\mu_h^{j+1}\right)\left(x-\mu_h^{j+1}\right)^T}{\sum_k w_h^j(x)} \tag{9}$$

The iterative computation stops if $\left|L(\Phi^f)-L(\Phi^{f+1})<\varepsilon\right|$, where $\varepsilon > 0$, and $L(\Phi)=\sum_k log\sum_{k=1}^h w_h f_h(x|\mu_h,\Sigma_h)$. For EM clustering, the distance between a document and a cluster is the Mahalanobis distance, $\sqrt{[(x-u)'S(x-u)]}$. Here, u is the cluster mean and S is the inverse of the cluster covariance matrix.

4. Results

4.1. Findings of text mining analysis

In this section, we try to present the main findings of the text mining analysis applied to the MPDRs. The DTM which is mentioned in the previous section and obtained as a result of preprocessing applied before text mining, contains very important information. Using the DTM, it is available to search unique words/noun groups over the set of reports. It is given a list represents the most frequent nouns in Table 1, below. Due to space limits of this paper, first ten nouns are given below.

As seen from Table 1, most frequent noun in MPDRs is “good”. Of course, this is not the adjective which is the opposite of “bad”. This is the noun which is a part of the group “goods and services”. Since the word “good” passes the most in the MPRDs, it can be inferred that the goods are the most important source of the change in annual CPI inflation during the years between 2006 and 2018.

The word “food” is the second noun being used for explaining the changes in inflation. This is not surprising because given the fact that food is among the most important triggers of inflation. The third noun is “increase” as a representation of the tendency of inflation in that period. Another finding related to the word “increase” is the increase in the frequency of use of the term “increase” in the periods when the annual CPI inflation increases.

Table 1
Most Frequent 10 Nouns in MPDR's between the years 2006–2018.

Noun	Frequency
good	1483
food	1469
increase	1065
energy	860
manufacturing	495
industry	431
oil	417
trend	372
decline	348
transportation	325

Source: Authors calculation.

As a well-known fact that one of the most important triggers of the annual CPI inflation in Turkey is the energy and oil prices, and they are included in Table 1, as expected. The words “manufacturing”, “industry” and “transportation” are also used frequently in MPDRs. It is seen that the word “decline” is used relatively low frequent rather than the word “increase” in line with the trend of inflation in that period.

We also analysed “word groups” in addition to single words. The word groups are the group of word contains two or more words. The associations measured (Euclidean, geometrically, etc.) between the words in the reports are generally used to construct the noun groups. There may be some cases that the word groups can provide more information than single words. Due to space limits of this paper, ten noun groups are given below in Table 2.

As seen from Table 2, the word group “core inflation” is the most used at in MPRD. This is an expected outcome due to CBRT focuses on core inflation indicators rather than raw inflation figures. The second word group used most frequently is the “service inflation” which is seen by CBRT as a key permanent source of change in annual CPI inflation in that period. Among the goods, the word group “durable goods” is the third most frequently passed term in the MPRDs.

The word groups “the exchange rate development”, “fresh fruit” and “unprocessed food” have taken place in the list of most used word groups in MPDR. Table 2 reflects not only the sources of change in the annual CPI inflation but also the information on its tendency. It is understood that the shape of fluctuations in CPI inflation is frequently described using the noun groups such as

Table 2
Most Frequent 10 Noun Groups in MPDR's between the years 2006–2018.

Noun Groups	Frequency
core inflation	1067
services inflation	945
durable good	401
exchange rate development	332
fresh fruit	320
unprocessed food	269
upward change	246
base effect	228
monthly basis	226
seasonal effect	176

Source: Authors calculation.

"upward change", "base effect", "monthly basis" and "seasonal effect". To see together most frequently used single nouns and noun groups in MPRD, a word cloud demonstration is given below (Fig. 1). This word cloud created in R platform (R Core Team, 2017) using "wordcloud" package by Fellows (2014).

In Fig. 1 given above, 297 nouns/noun groups in total are presented. Considering both nouns and noun groups, it seems that the word cloud summarises the findings of Tables 1 and 2. As seen from Fig. 1, most frequently used nouns and noun groups are represented by bigger fonts, less ones are represented by smaller fonts. So, in word clouds, there is positive relationship between the size of the word and its frequency.

At this point of the analysis, we think that it is worth to make a clarification on the relationship between frequencies and weights of the words. In our application, common words such as "inflation", "change", "price" are frequently found at each of MPDRs. These kinds of words are not identifier to discriminate the MPDRS. Because these words have no difference from the words such as "a", "an" or "the", if the aim is to separate the reports into clusters. A general approach is to remove these useless words to prevent them to affect the text mining analysis. That is why we prefer to use the full weighting method.

As given earlier, the full weighting is to arrange the numbers according to the importance of the word instead of determining the frequency of the words in the document. Since the full weighting formulation is used to determine weights in this study, it is tried to illustrate the implication of full weighting approach in DTM. Fig. 2 presents the weights and frequencies of words in the MPDR's. As seen Fig. 2, there is clear the inverse relationship between weights and frequencies as assumed in full weighting approach.

After clarifying the inverse relationship between the frequencies and weights of words, the next is to give the results of topic analysis which produces 27 different topics from the MPDRs. The list of topics and related information are given in Table 3.

Table 3 contains the topic id, the term cut off value, the topic name, the number of words included in the topic, and the number of reports containing the topic. Term cut off rate is the minimum topic weight that a term must have to be used as a term for this topic. The weight of any term that has an absolute value weight less than term cut off rate is effectively set to zero. So, the number of words in Table 3 represents the number of words which have weights above the term cut off rate.

Fig. 1. Nouns and noun groups in monthly price developments reports.
Source: Authors calculation.

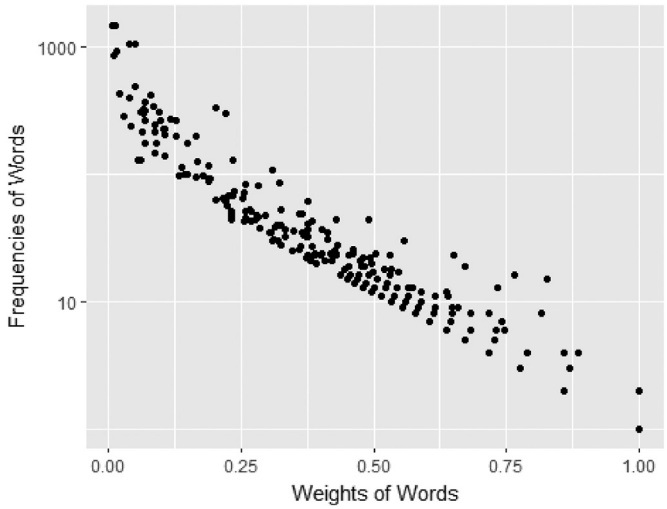

Fig. 2. Relationship between frequencies and weights of words.
Source: Authors calculation.

It is expected that the number of topics should be high. Because, it is unlikely that one or two topics can be expected to extract from an MPRD, which explains the reasons for the change of annual CPI inflation. As it can be seen from Table 3, in 2006–2018 years, the developments in the annual CPI inflation are discussed in the 27 topics in MPDRs.

Among these topics, the most frequently appeared in MPDR's was "consumption, meat, core inflation indicator, low, flat", namely Topic ID-1. This topic contains 50 words (as seen in the third column), and it is mentioned in 27 of the MPDR's (as seen in the fourth column). The second topic is "depreciation, lira, appreciation, catering service, negative outlook, namely Topic ID-26. This topic contains 31 words and is mentioned in 21 of MPDR's. The third most frequently appeared in MPDR's was "sunflower, cotton, falling fruit, soaring price, gold". This topic contains 34 words, and it is mentioned in 20 of the MPDR's.

Given in Table 4, the last result of text mining analysis is about the text clustering. Roughly, MPDR's are clustered in to several clusters using EM method. Table 4 contains four columns namely Cluster, Description, Frequency and Percent. The "Cluster" column shows the number of clusters identified. "Description" column contains the nouns and/or the noun groups which are contained by the clusters. The third column "Frequency" represents how many reports are clustered into. The last column "Percent" represents the distribution of the reports among the clusters.

As seen from the cluster column, number of clusters identified is five. It means that 138 reports are clustered into five clusters. Under the assumption of 20 nouns or noun groups can represent the cluster, the description of columns presents those nouns and noun groups which represents clusters. The widest spread one is 2nd cluster and it represents 51 MPDRs. The second one is 5th cluster and it represents 38 MPDRs. It should be noted that each of the reports is covered by a cluster.

4.2. An extension: consistency of MPDRs with annual CPI data

So far, text mining techniques have been applied to the MPDRs. As a result of all these analysis, 138 reports, which were published from July 2006 to January 2008, were classified into five clusters. For example, these clusters can be used in text classification to identify the new reports to be published in 2018 and so on.

Table 3
The topics identified from the MPDR's between the years 2006–2018.

Topic ID	Term Cut Off Rate	Name of Topic	Number of Words	Number of Reports
1	0.082	consumption, meat, core inflation indicator, low, flat	50	27
2	0.082	exchange rate development, medicine, household equipment, housing, sale	30	12
3	0.082	expenditure, holiday, package tour, religious month, rapid	29	10
4	0.081	methodological change, nondurable good, acceleration, c index, raw milk	26	14
5	0.082	sunflower, cotton, falling fruit, soaring price, gold	34	20
6	0.082	iron, alloy, crop, elevated price, construction	30	9
7	0.082	apparel, health, falling price, shortage, aggregate	38	12
8	0.081	foreign demand, demand, domestic, modest pace, balance	25	9
9	0.082	farm product, natural gas, arrangement, corn, animal product	31	13
10	0.082	annual consumer, passenger, canned vegetable, unfavorable effect, tea	28	13
11	0.082	coke coal, risk, annual increase, disinflation, significant change	30	14
12	0.082	footwear price, annual term, fresh vegetable, chicken price, uncertainty	43	12
13	0.082	stable outlook, culture, cap, international commodity, insurance	38	12
14	0.082	engine, domestic wheat, wheat, mild course, cigarette	27	14
15	0.081	legume, umrah, highway, demand, deterioration	31	8
16	0.081	tax, expiration, cut, electricity, temporary effect	23	12
17	0.082	elevated, repair, maintenance, package tour, care	31	8
18	0.082	phone call, phone, mobile phone, grain, telephone call	23	17
19	0.080	custom, import, duty, cotton, moderate rate	20	10
20	0.082	administered price, electricity, adjustment, equipment, steel price	34	14
21	0.082	education, university tuition rate, bus, tourism, land passenger transport	33	16
22	0.082	supply, shock, shift, base metal price, drop	34	9
23	0.082	home appliance, package tour, electronics, deceleration, consumer electronics	37	14
24	0.082	tomato, other, spike, electricity, leather product	32	8
25	0.082	further upward pressure, tax, decision, health, balance	35	12
26	0.082	depreciation, lira, appreciation, catering service, negative outlook	31	21
27	0.082	sugar, reverberation, momentum, education, college	34	10

Source: Authors calculation.

Table 4
The clusters identified from the MPDR's between the years 2006–2018.

Cluster	Description	Frequency	Percent
1	'methodological change' 'c index' 'annual food inflation' 'nondurable good' 'domestic producer' cost vehicle commodity 'base metal' depreciation 'exchange rate development' 'motor vehicle' aggregate 'automobile price' flat lira 'monthly basis' annual meat 'fresh fruit'	11	7.971014
2	consumption 'mild course' bread 'basic metal' 'domestic producer' fall meat flat depreciation 'monthly basis' trend metal 'capital good' decline lira 'upward change' down gas manufacturing oil	51	36.956522
3	'annual food inflation' cost 'catering service' 'domestic producer' 'c index' depreciation 'motor vehicle' vehicle 'exchange rate development' 'basic metal' lira 'mild course' flat commodity meat petroleum aggregate 'automobile price' trend metal	12	8.695652
4	aggregate 'migs classification' 'footwear price' 'lagged effect' gold 'alcoholic beverage' tobacco 'seasonal effect' down 'automobile price' 'monthly basis' hike 'agricultural price' high bread 'unprocessed food' rent trend fall 'processed food'	26	18.840580
5	'coming month' gold pressure 'agricultural price' 'base metal' 'alcoholic beverage' tobacco low 'migs classification' 'processed food' 'unprocessed food' 'intermediate good' 'footwear price' 'catering service' 'base effect' furniture 'durable good' 'capital good' 'core inflation indicator' transportation	38	27.536232

Source: Authors calculation.

On the other hand, this section is intended to demonstrate whether the clusters derived from MPDRs are consistent with annual CPI inflation. Specifically, in this section, the purpose is to determine whether the identified clusters explain statistically variance of the announced annual CPI figures. It should be noted that historical development of the causes of inflation is out of scope. And this section does not aim to explain the inflation or its causes. Moreover, it is not reasonable to determine the effect of clusters on inflation among the causes of inflation because MPRDs are published after reference period inflation is announced.

At first, it is considered useful to see the interaction between the clusters identified and the annual CPI inflation figures, given in Fig. 3. In Fig. 3, the annual CPI inflation figures are represented by the bars, while the colouring is done according to the clusters obtained from text mining. As it is clear from the figure, the clusters divide the published annual CPI inflation figures into quite distinct periods. Black vertical lines in the chart are added to make the periods between the clusters clear.

Chronologically interpretation of the figure is as followed. The fourth cluster covers the period between July 2006 and April 2008. April 2008 can be described as very early periods of the global economic crisis. Therefore, it can be concluded that MPDR's interpretations in the period up to April 2008 are consistent within itself, which is consistent with the annual CPI inflation figures. The fifth cluster refers to the period between April 2008 and July 2011. Looking at this cluster in terms of the CPI inflation figures, the decline in the CPI inflation figures appears to be predominant, depending on the overall impact of the global economic downturn on demand. In this period, it is understood that MPRDs have similar interpretations.

The second cluster roughly points to a relatively long period between July 2011 and September 2016. In this period, the annual CPI inflation figures show moderate increase after the crisis. Although there are different clusters in this period in some periods, it is seen that the number two clusters are mainly effective. It seems that the third cluster is active in a relatively short time interval

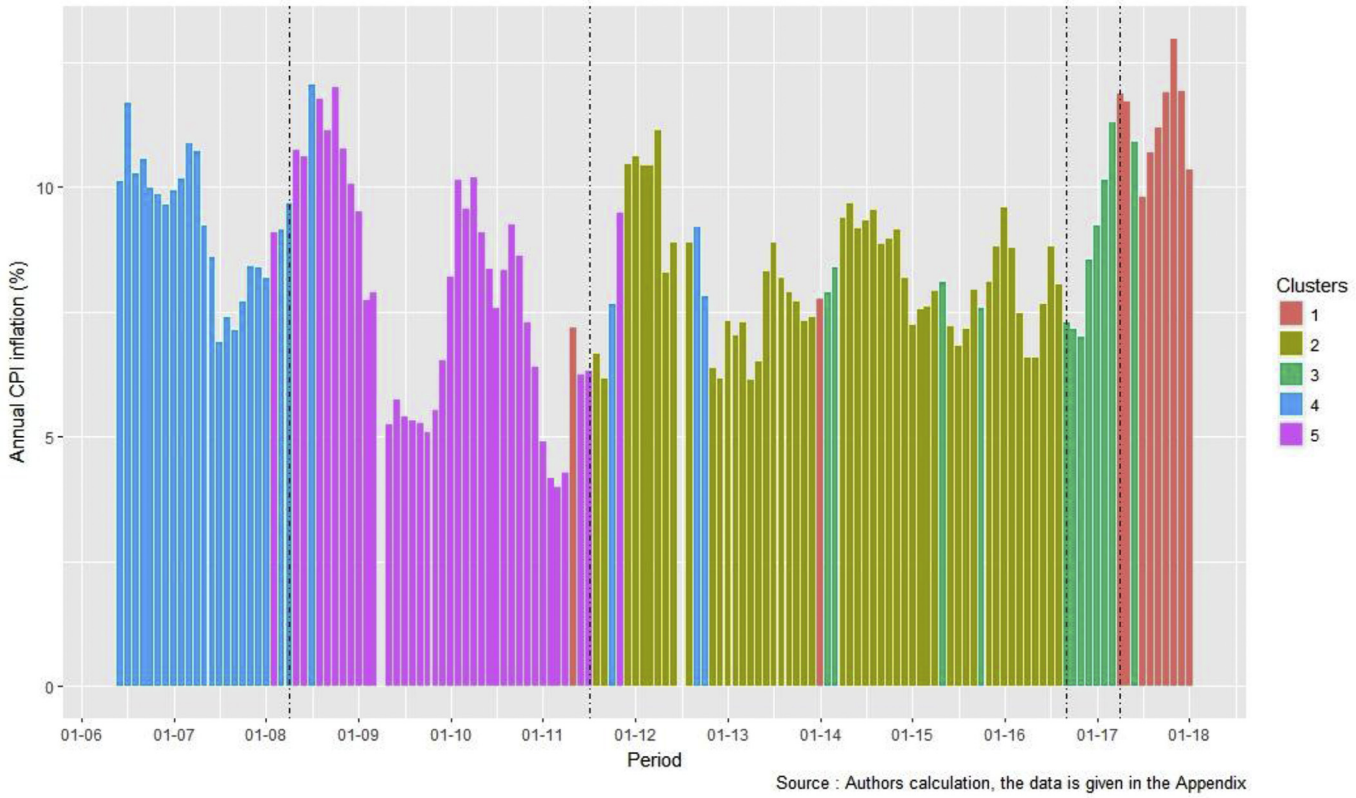

Fig. 3. The annual CPI inflation vs. the clusters.
Source: Authors calculation.

(from September 2016 to April 2017, for a total of 8 months). In this period, annual CPI inflation figures are seen to be in a period of rapid increase.

Therefore, it is understood that the rapid increase in the annual CPI inflation in this period is consistent with the MPDR's report clusters. Finally, first cluster covers the period from April 2017 to January 2018. And in this period, the clusters in the MPDR's reports appear to coincide with higher levels in the annual CPI inflation figures.

Although had been intended to examine in this study, the statistical way of examination of consistency followed by Jansen and Haan (2010) is not appropriate for this study since the clustering approach has been applied. For this reason, an ANOVA test is performed between categorical cluster variables and continuous variables of inflation to examine the consistency of MPDRs instead of correlation/covariance analysis as applied in Jansen and Haan (2010).

If one-way model is applied to investigate the consistency, the possibility of changing the annual CPI figures change by time reduces the power of the one-way ANOVA test. For this reason, we first tried to show whether the annual CPI figures changed over time (2-way-model). Later, we added the clusters to analyse and applied 3-way-model to determine the effects of the cluster variable and its interaction variables on the annual CPI figures. Finally, we compared the 2-way-model and 3-way-model (Chambers and Hastie, 1992) and showed that the 3-way- model in which the cluster variable is used more clearly explained the annual CPI figures (compared to the 2-way-model). The following null hypothesis (2-way-model) will be tested against the alternative (3-way-model).

H_0: Clusters have no effect on annual CPI inflation figures. $CPI_t = \alpha_{1,1}Y_t + \alpha_{1,2}M_t + \alpha_{1,3}Y_t{*}M_t + \varepsilon_{1,t}$ (Model-1)
H_1: Clusters have statistically significant effect on annual CPI inflation figures. $CPI_t = \alpha_{2,1}Y_t + \alpha_{2,2}M_t + \alpha_{2,3}Cl_t + \alpha_{2,4}Y_t{*}M_t + \alpha_{2,5}Y_t{*}Cl_t + \alpha_{2,6}M_t{*}Cl_t + \alpha_{2,7}Y_t{*}M_t{*}Cl_t + \varepsilon_t$ (Model-2)

where t = Jan-2006, Feb-2006, …, Jan-2018. In the null hypothesis, i.e. in Model-1, it is assumed that the annual CPI inflation figures are affected from the year and the month variables and their interactions. In Model-2 given in the alternative hypothesis, annual CPI inflation figures are assumed to be influenced by year, month and cluster variables and their interactions.

Table 5 shows the model estimation results of Model-1. According to probabilities given in Table 5, it can be claimed that both year and month variables also their interactions have not significant effect on annual CPI inflation figures.

Table 6 shows the model estimation results of Model-2. According to probabilities given in Table 6, it can be suggested that cluster variable and its interactions with year variable are statistically significant at 1% significance level. This shows that the clusters identified significantly effects the changes in annual CPI inflation figures.

To statistically present the significance of Model-2 over Model-1, the ANOVA test results which shows variance table between Model-1 and Model-2 are given in Table 7. According to Table 7, it

Table 5
Model-1 for annual CPI inflation figures with year and month.

Term	Degrees of Freedom	Sum of Squares	Mean of Squares	Test Statistics	Probability
Year (Y)	1	0.0878760	0.0878760	0.0247585	0.8752072
Month (M)	1	0.0105409	0.0105409	0.0029698	0.9566210
(Y*M)	1	0.6670817	0.6670817	0.1879460	0.6653288
Residuals	134	475.6098730	3.5493274	–	–

Source: Authors calculation.

Table 6
The Model-2 for CPI with year, month and clusters.

Term	Degrees of Freedom	Sum of Squares	Mean of Squares	Test Statistics	Probability
Y	1	0.0878760	0.0878760	0.0317300	0.8588989
M	1	0.0105409	0.0105409	0.0038061	0.9509017
Cl	1	28.5986543	28.5986543	10.3262968	0.0016539
Y*M	1	0.7210013	0.7210013	0.2603365	0.6107543
Y*Cl	1	79.9557935	79.9557935	28.8701435	0.0000003
M*Cl	1	0.0580584	0.0580584	0.0209635	0.8851027
Y*M*Cl	1	6.9087577	6.9087577	2.4945888	0.1166684
Residuals	130	360.0346896	2.7694976	–	–

Source: Authors calculation.

Table 7
Comparison between Model-1 and Model-2.

Residual Degrees of Freedom	Residual Sum of Squares	Degrees of Freedom	Sum of Squares	Test Statistics	Probability
134	475.6099	–	–	–	–
130	360.0347	4	115.5752	10.43287	2e-07

Source: Authors calculation.

can be suggested that the variance difference between Model-1 and Model-2 are statistically significant at 1% significance level. This means that null hypothesis is rejected in 99% confidence level. The results show that the clusters and its interactions with the year variable can contribute to explain the variance of annual CPI inflation figures. That means the clusters of MPDRs are consistent with annual CPI figures since the clusters are found as an explanatory information for annual CPI figures.

5. Conclusions and discussion

In the studies so far, it is seen that the texts or meeting summaries published by central banks are analysed by text mining. In these studies, it is understood that especially the documents whose frequency is quarterly or yearly are analysed. Another shortcoming in the literature is that findings from text mining are not used as input in later analyses.

We applied text mining and analysis of variance to the examination of the Monthly Price Development Reports (MPDR) of the Central Bank of Republic of Turkey (CBRT). Our findings can be divided into two parts. First, we found the most used nouns and nouns groups among the reports between years 2006 and 2018. These words generally represent the core inflation measures and some sectors such as service or hotel and accommodation as well as type of goods like durable or non-durable. Given the tendency of inflation in Turkey during the past 10 years, "increase" word to be among the most frequent words are outstanding as an expected finding. In the second part of the findings, it was presented the analysis of the consistency of MPDRs with the actual inflation figures. Specifically, the clusters identified in text mining were used in the 3-way variance model for annual CPI inflation figures. In this sense, this approach may be considered as a relevant analysis to complement the traditional text mining analysis. Because this aspect has been neglected in the literature. As a result, the model in which assumed annual CPI inflation figures are explained by trend, seasonality and the clusters of MPDRs has been found more significant than the model in which assumed annual CPI inflation figures are only explained by trend and seasonality. This result implies that MPDRs are consistent with annual CPI figures.

Several extensions of this paper are warranted in future work. The first is to extend the analysis to other types of CBRT's communication; perhaps quarterly inflation reports and their effects on future inflation figures. This type of analysis may help to extract more messages from quarterly inflation reports. Second, it would be useful to see if there is a time-varying pattern in the summaries of MPC meetings in terms of the words used in. It is thought that it is worth to analyse the time-varying change in the expression of monetary tightening or expansion and the multivariate relationship between them and interest rates or inflation.

Appendix

Table A1
The clusters identified from the MPDR's between the years 2006–2018 and Annual CPI Inflation.

Period	CPI (YoY, %)	Cluster	Period	CPI (YoY, %)	Cluster	Period	CPI (YoY, %)	Cluster
06–06	10.12	4	05–10	9.10	5	04–14	9.38	2
07–06	11.69	4	06–10	8.37	5	05–14	9.66	2
08–06	10.26	4	07–10	7.58	5	06–14	9.16	2
09–06	10.55	4	08–10	8.33	5	07–14	9.32	2
10–06	9.98	4	09–10	9.24	5	08–14	9.54	2
11–06	9.86	4	10–10	8.62	5	09–14	8.86	2
12–06	9.65	4	11–10	7.29	5	10–14	8.96	2
01–07	9.93	4	12–10	6.40	5	11–14	9.15	2
02–07	10.16	4	01–11	4.90	5	12–14	8.17	2
03–07	10.86	4	02–11	4.16	5	01–15	7.24	2
04–07	10.72	4	03–11	3.99	5	02–15	7.55	2
05–07	9.23	4	04–11	4.26	5	03–15	7.61	2
06–07	8.60	4	05–11	7.17	1	04–15	7.91	2
07–07	6.90	4	06–11	6.24	5	05–15	8.09	3
08–07	7.39	4	07–11	6.31	5	06–15	7.20	2
09–07	7.12	4	08–11	6.65	2	07–15	6.81	2
10–07	7.70	4	09–11	6.15	2	08–15	7.14	2
11–07	8.40	4	10–11	7.66	4	09–15	7.95	2
12–07	8.39	4	11–11	9.48	5	10–15	7.58	3
01–08	8.17	4	12–11	10.45	2	11–15	8.10	2
02–08	9.10	5	01–12	10.61	2	12–15	8.81	2
03–08	9.15	4	02–12	10.43	2	01–16	9.58	2
04–08	9.66	4	03–12	10.43	2	02–16	8.78	2
05–08	10.74	5	04–12	11.14	2	03–16	7.46	2
06–08	10.61	5	05–12	8.28	2	04–16	6.57	2
07–08	12.06	4	06–12	8.87	2	05–16	6.58	2
08–08	11.77	5	08–12	8.88	2	06–16	7.64	2
09–08	11.13	5	09–12	9.19	4	07–16	8.79	2
10–08	11.99	5	10–12	7.80	4	08–16	8.05	2
11–08	10.76	5	11–12	6.37	2	09–16	7.28	3
12–08	10.06	5	12–12	6.16	2	10–16	7.16	3
01–09	9.50	5	01–13	7.31	2	11–16	7.00	3
02–09	7.73	5	02–13	7.03	2	12–16	8.53	3
03–09	7.89	5	03–13	7.29	2	01–17	9.22	3
05–09	5.24	5	04–13	6.13	2	02–17	10.13	3
06–09	5.73	5	05–13	6.51	2	03–17	11.29	3
07–09	5.39	5	06–13	8.30	2	04–17	11.87	1
08–09	5.33	5	07–13	8.88	2	05–17	11.72	1
09–09	5.27	5	08–13	8.17	2	06–17	10.90	3
10–09	5.08	5	09–13	7.88	2	07–17	9.79	1
11–09	5.53	5	10–13	7.71	2	08–17	10.68	1
12–09	6.53	5	11–13	7.32	2	09–17	11.20	1
01–10	8.19	5	12–13	7.40	2	10–17	11.90	1
02–10	10.13	5	01–14	7.75	1	11–17	12.98	1
03–10	9.56	5	02–14	7.89	3	12–17	11.92	1
04–10	10.19	5	03–14	8.39	3	01–18	10.35	1

Source: Authors calculation.

References

Acosta, Miguel, 2015. FOMC Responses to Calls for Transparency. Finance and Economics Discussion Series 2015-60. Board of Governors of the Federal Reserve System (U.S.). https://ideas.repec.org/p/fip/fedgfe/2015-60.html.

Blinder, Alan S., Ehrmann, Michael, Fratzscher, Marcel, Haan, Jakob de, Jansen, David-Jan, 2008. Central Bank Communication and Monetary Policy: a Survey of Theory and Evidence. Working Paper Series 898. European Central Bank. https://ideas.repec.org/p/ecb/ecbwps/2008898.html.

CBRT, 2016. Annual Report. Annual Report Series. Central Bank of Republic of Turkey. http://www3.tcmb.gov.tr/yillikrapor/2016/files/en-tcmb2016.pdf.

Chambers, J.M., Hastie, T.J., 1992. Statistical Models in S. Wadsworth & Brooks/Cole.

Christiano, Lawrence J., Eichenbaum, Martin, 1992. Liquidity effects and the monetary transmission mechanism. Am. Econ. Rev. 82 (2), 346–353. American Economic Association. http://www.jstor.org/stable/2117426.

Feldman, Ronen, Sanger, James, 2006. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.

Fellows, Ian, 2014. Wordcloud: Word Clouds. https://CRAN.R-project.org/package=wordcloud.

Hansen, Stephen, McMahon, Michael, 2015. Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communication. Discussion Papers 1537. Centre for Macroeconomics (CFM). https://ideas.repec.org/p/cfm/wpaper/1537.html.

Hofmann, Thomas, 2001. Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. 42 (1), 177–196. https://doi.org/10.1023/A:1007617005950.

Hotho, A., Nurnberg, A., Paas, G., 2005. A brief survey of text mining. J. Comput. Ling. Lang. Technol. 20 (1). http://media.dwds.de/jlcl/2005_Heft1/19-62_HothoNuernbergerPaass.pdf.

Iglesias, Joaquin, Ortiz, Alvaro, Rodrigo, Tomasa, 2017. How Do the EM Central Bank Talk? a Big Data Approach to the Central Bank of Turkey. Working Papers 17/24. BBVA Bank, Economic Research Department. https://ideas.repec.org/p/bbv/wpaper/1724.html.

Jansen, David-Jan, Haan, Jakob de, 2010. An Assessment of the Consistency of ECB Communication Using Wordscores. DNB Working Papers 259. Netherlands Central Bank, Research Department. https://ideas.repec.org/p/dnb/dnbwpp/259.html.

Kahveci, Eyup, Odabas, Aysun, 2016. Central banks' communication strategy and content analysis of monetary policy statements: the case of fed, ECB and CBRT. Procedia Soc. Behav. Sci. 235. http://www.sciencedirect.com/science/article/pii/S1877042816315737.

Kara, A.Hakan, 2008. Turkish experience with implicit inflation targeting. Cent. Bank Rev. 8 (1), 1–16. https://ideas.repec.org/a/tcb/cebare/v8y2008i1p1-16.

html.
Kumar, Anil, Chandrasekhar, S., 2012. Text data pre-processing and dimensionality reduction techniques for document clustering. Int. J. Eng. Res. Technol. 1 (5). In: https://www.ijert.org/download/475/text-data-pre-processing-and-dimensionality-reduction-techniques-for-document-clustering.
Lucca, David O., Trebbi, Francesco, 2009. Measuring Central Bank Communication: an Automated Approach with Application to FOMC Statements. NBER Working Papers 15367. National Bureau of Economic Research, Inc. https://ideas.repec.org/p/nbr/nberwo/15367.html.
R Core Team, 2017. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Ranaldo, Angelo, Rossi, Enzo, 2010. The reaction of asset markets to Swiss National Bank communication. J. Int. Money Finance 29 (3), 486–503. https://ideas.repec.org/a/eee/jimfin/v29y2010i3p486-503.html.
Weiss, Sholom, Indurkhya, N., Zhang, T., Damerau, F.J., 2004. Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer.