Text Mining On Complex Text Sources: Analysis Of The Final Considerations Of The Governor On The Annual Reports Of The Bank Of Italy From 2008 To 2017

Paolo Dalena

July 10, 2019

Contents

Preface to the English version **2**

1 Introduction **2**

1.1 Annual Reports of the Bank of Italy . . . 3

1.2 Final remarks by the Governor . . . 3

1.3 Text mining on complex text sources . . . 4

2 Objectives **6**

2.1 Lexical differences . . . 6

2.2 Most used terms . . . 6

2.3 Positive or negative approach . . . 6

2.4 Recurring Topics . . . 6

3 Methods **7**

3.1 Exporting text . . . 7

3.2 Lemmatization, categorization and cleaning . . . 8

3.3 Analysis of lexical differences . . . 8

3.4 Word Cloud Creation . . . 8

3.5 Sentiment analysis . . . 9

3.6 Topic modeling . . . 9

4 Results and discussion **10**

4.1 Parts of speech distribution . . . 10

4.2 Lexical characteristics . . . 11

4.3 Wordcloud by periods . . . 12

4.4 Comparison cloud and commonalty cloud . . . 13

4.5 Positivity of the lemmas 16
4.6 Topics of the parts of the text 18

5 Conclusions **19**

Acknowledgements **20**

References **20**

Preface to the English version

This is the English version of my Bachelor Thesis **TEXT MINING SU SORGENTI DI TESTO COMPLESSE: ANALISI DELLE CONSIDERAZIONI FINALI DEL GOVERNATORE SULLE RELAZIONI ANNUALI DELLA BANCA D'ITALIA DAL 2008 AL 2017** that I wrote for the Bachelor in Statistical Sciences degree I received on July 10, 2019 at Alma Mater Studiorium - University of Bologna.

You can find the link to the Github page of my workflow (with all the R scripts useful for the analysis and for creating the provided graphs) here and the official published documents here.

Everything, for obvious reasons, is in Italian, so this online book is born to allow more people to understand my work.

Please note that the text has been automatically translated, so it's quite likely to find any linguistic (and other) inaccuracies.

If you find anything wrong, please report it here. I will be happy to help you understand or (more likely) correct the errors.

1 Introduction

In the era of the digital revolution, of frenetic computerization and of the growing and almost annoying presence of the words Intelligence and Artificial in our daily lives, the application of automatic techniques to accompany the comprehension and interpretation of official texts is indisputably topical. In fact, such methods present almost boundless potentialities and their use, although not in-depth, cannot but constitute a good practice that is likely to be interesting.

We have chosen to study the Final Considerations of the Governor on the Annual Reports of the Bank of Italy referring to the years 2008 to 2017, analyzing their style and language, the themes dealt with and the opinions expressed. The description of the data and the reasons behind this choice, together with the presentation of text mining techniques, are dealt with in the following paragraphs of this chapter. In the following one, the objectives of the study are explained. In the third chapter, the methods by which the analysis was carried out are presented in detail. The fourth chapter contains the exposition of the results obtained correlated to their interpretation. In the conclusion section, the study is summarized and critical issues are described.

All the functions created and used, together with the data analyzed, the graphs created, the software used, all the code necessary for the analysis organized by topic, and everything that was useful to complete this work is easily and freely available through my GitHub, in the folder tesi.*(accessible via https://github.com/PaoloDalena/tesi)*

1.1 Annual Reports of the Bank of Italy

The Bank of Italy is the central bank of the Italian Republic, with its headquarters in Rome and secondary offices and branches throughout Italy. It is a public-law institution governed by national and European regulations. It is an integral part of the Eurosystem, consisting of the national central banks of the euro area and the European Central Bank. The Eurosystem and the central banks of the EU Member States that have not adopted the euro make up the European System of Central Banks It pursues purposes of general interest in the monetary and financial sector: the maintenance of price stability, the stability and efficiency of the financial system and other tasks entrusted to it by national law. The Bank's functional and governance structure reflects the need to rigorously protect its independence from external influences, an essential prerequisite for the effective performance of its institutional action. National and European regulations guarantee the autonomy necessary to pursue the mandate; this autonomy is backed up by stringent duties of transparency and publicity. The Bank is accountable to the Government, Parliament and the public for its work through the dissemination of data and information on institutional activities and the use of resources.(1)

The publications of the Bank of Italy reflect the activities carried out by the Institute. They are of an economic-financial, historical and legal nature and are all free of charge and available online. Among these, the one that best provides a concise presentation of the country's economic situation is undoubtedly the Annual Report. The latter is published every year at the end of May and contains an in-depth analysis of the main developments in the Italian and international economy in the previous year and in the first months of the current year and is accompanied by a statistical appendix that is only available online. It is also the subject, in a public meeting not limited to Participants, of Considerations by the Governor of the Bank of Italy.

This publication can be defined as a true analytical and informative consultancy on the state of the economy that the Bank of Italy offers to constitutional bodies in matters of economic and financial policy. By virtue of this, it is easy to understand the importance of the Annual Report and how much its content constitutes a perfect summary framework to be used as the basis for a study of the Italian economic situation and its evolution over time.

1.2 Final remarks by the Governor

As already specified, the Annual Report of the Bank of Italy is discussed at a public meeting. On the occasion of its circulation, therefore, the Governor of the Bank of Italy presents the so-called Final Considerations.

The Governor of the Bank of Italy has the task of representing the bank with third parties, presiding over the meeting and informing the Italian government on foreign or domestic financial matters. Until before the introduction of the Euro, he was also responsible for national monetary policy. This function is exercised collegially together with the other central banks of the Euro area.

The Governor is appointed by decree of the President of the Republic, on the proposal of the President of the Council of Ministers, after deliberation by the Council of Ministers, having heard the opinion of the Superior Council of the Bank of Italy. The procedure also applies to the revocation of the Governor. His office, which until 2005 had no term limit, lasts six years and can be renewed once.(2)

Focusing on the period of interest to us, the Governor in office for the period 2005-2011 was Mario Draghi, who currently holds the prestigious position of President of the European Central Bank, who was succeeded on November 1, 2011 by Ignazio Visco, currently in office.

The Final Considerations referring to 2008, 2009 and 2010, therefore, were drafted by Draghi, while the subsequent seven in analysis (published up to May 2018, therefore referring to the period 2011-2017) are the result of the work of the current Governor. These publications, available free of charge online, provide a concise and more "human" opinion of the overall economic picture of the country. They are, in fact, actual comments that the Governor is called upon to make in order to take stock of the past year. For these and other practical reasons, related to the need to use texts as free as possible of multimedia content, we have

chosen to analyze the Final Considerations of the Governor of the Bank of Italy referring to the ten years between 2008 and 2017. This will provide a clear idea of the evolution of Italy's economic situation in the reference period from how the latter is presented in the Annual Reports.

1.3 Text mining on complex text sources

Data mining is the set of techniques and methodologies that aim to extract useful information from large amounts of data, through automatic or semi-automatic methods (such as machine learning).

Text mining, also called text data mining or (in some ways incorrectly) text analytics, is a particular form of data mining in which the data consists of natural language texts, in other words "unstructured" documents. Text mining combines language technology with data mining algorithms. The goal is the same: the extraction of implicit information contained in a set of documents.

This discipline, also known by the acronym TM, therefore, deals with the search, analysis and thematic classification of information contained in documents. Unlike most of the data with which statistics works, in textual documents the information is present in the form of free text (sentences and words) and only to a small extent as structured text (tables, graphs, etc.). It must also be said that much of the communication, hence information exchange, between human beings takes place through unstructured documentation (books, newspapers, conversations).

TM is a multidisciplinary field based on different sets of techniques, grouped under the name of information retrieval, data mining, machine learning, statistics, and computational linguistics. Figure 1.1 shows the interactions between text mining and some of the areas presented. (3)

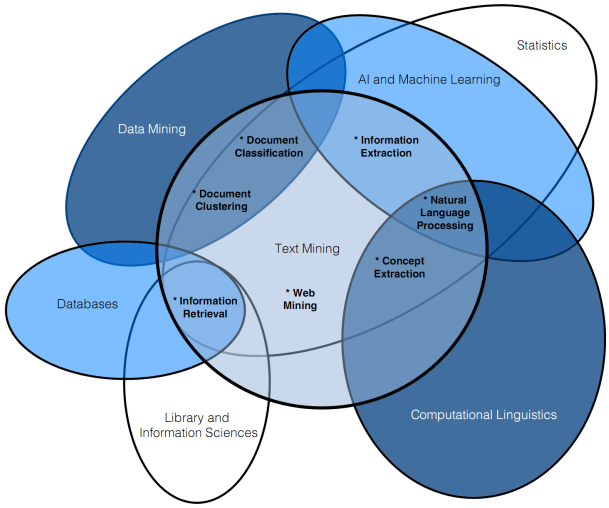

Figure 1: Interaction of Text Mining with other research fields

Given the wide range of possible applications of the technique in question, providing an exhaustive and universal description of how the TM process takes place is a particularly difficult task. However, it is possible to recognize some steps:

- Data collection: first step that involves the collection and selection of documents that may be useful for analysis.

- Text pre-processing: in which the raw text is adapted into analyzable text. Specifically, pre-processing and cleaning operations are performed to detect and remove anomalies so that the true essence of the

available text can be captured and also simply to reduce the size of the data. At this stage, procedures of:

  - Tokenization, which allows breaking a sequence of characters into units (usually words or phrases) called tokens.

  - Filtering, in order to remove unnecessary parts of the text.

  - Lemmatization, which causes the various inflected forms of a word to be grouped together in such a way that they are analyzed as a single entity.

  - Derivation, which is the process by which a form (theme or word) is created from a pre-existing root or word.

- Application of text mining techniques: this is the phase of greatest interest, in which textual data (keywords, concepts, verbs, nouns, adjectives, etc.) are extracted using techniques based on different algorithms. Among these, the most popular are:

  - Text Categorization: represents the beginning of the text analysis process by assigning predefined categories to tokens;

  - Information Extraction: is a technique that extracts meaningful information from a large amount of text. Usually this information is taken from unstructured and/or semi-structured machine-readable documents and transformed into structured information.

  - Information retrieval: represents the set of techniques used to manage the representation, storage, organization and access to objects containing information such as documents, web pages, online catalogs and multimedia objects. It is also used by Google and Yahoo search engines to extract documents from a web search.

  - Clustering: is an unsupervised process of classifying text documents into similar groups called clusters. In a cluster, sets of text that relate to the same topic or identical keywords are grouped together.

  - Text Summary: This is the problem of creating a short, accurate, and fluent summary of a longer text document.

  - Sentiment analysis: also known as opinion mining, this method is used to extract subjective information from content. Just as the term suggests, it has to do with emotion, sentiment. Basically, it is applied to understand the emotional response of a subject in a context.(4)

The software and programming languages that allow the implementation of TM techniques are different and with different characteristics. In this discussion, as we will see in detail later, the analysis was conducted almost entirely on R.

In conclusion, it should be added that the Governor's Final Remarks are only available online(5) in PDF (Portable Document Format). This format is undoubtedly the most widely used format for disseminating publications available on the web, as it allows words to be organized into columns, graphs, and tables making it easier for humans to read. However, what makes it easy for humans, it makes impossible for machines to use directly.

TM algorithms, in fact, are directly applicable only if the data source is simple. This is the case when, for example, you have text files (.txt format), i.e. documents that contain only letters, numbers, punctuation marks, spaces and other printable symbols. With PDFs, given the presence of text formatted in different ways, images, graphics, tables or any other type of multimedia content, the situation becomes very complicated

(this is why we talk about complex text sources). Therefore, it is necessary to carefully extract and transform the data, which will be discussed in detail later.

2 Objectives

2.1 Lexical differences

In the study of the Governor's Final Remarks, a first approach to understanding the evolution of the situation can begin with an analysis of lexical differences over time. It makes sense, in fact, to study how and if the ways in which people express themselves and construct sentences have changed in the ten years of reference.

In order to observe these peculiarities, it is necessary to study the distribution of the different parts of speech (nouns, adjectives, verbs, adverbs, pronouns, etc.), but it would also be useful to have more synthetic measures, such as the average length of sentences or the number of terms not repeated within the document. These last two data, in fact, could respectively provide us with relevant information about the complexity of the treatment, from the point of view of the subordination of the propositions (the longer the average sentence, the greater the subordination) and of the lexicon (the more words are not repeated in a text, the more lexically rich it will be).

2.2 Most used terms

It is interesting to observe the presence of recurrent terms within the single documents or of the same documents grouped by periods, in order to study possible correlations between the economic situation of the moment under examination and the words most present. For example, it is natural to expect that in the years following the crisis of 2008-2009 the term "crisis" is strongly present in the documents.

A further objective of interest is to identify common words between the different documents over the years, so as to observe a trend or recurrent terms over time.

2.3 Positive or negative approach

By means of opinion mining, it is possible to study whether words in documents are predominantly linked to positive or negative emotions.

In this way it is possible to understand how the general situation of the Italian economy is described and whether or not the tones of this approach have changed over time.

2.4 Recurring Topics

Thanks to the potential of topic modeling, a technique that makes it possible to create probabilistic models that, through the analysis of words characterizing texts, identify the topics dealt with in a document, it will be possible to observe any recurring themes over the years.

As a result, it will be possible to examine how the discussion in the Annual Reports has changed over time and in what direction, but also which are the most important objects of analysis, as these will coincide with those that need to be talked about most often, i.e. the most present.

3 Methods

3.1 Exporting text

As already mentioned, the data subject to this analysis are only available in complex form, i.e. in PDF format. It was therefore necessary, first, to search for software that offered the best performance in accurately exporting text and, second, to automate the process.

In order to understand the particular capabilities of the different methods for importing text from PDF so that we could understand which one is the most accurate and, in particular, the most suitable for our needs, a literature review was conducted on the subject.

Once the software was found, its operation was automated by creating an R package containing the functions useful to the cause.

3.1.1 Literature Review

The literature review, conducted on the archives of arXiv2*(Archive of scientific articles in physics, mathematics, computer science, quantitative finance, and biology, accessible via https://arxiv.org/)* and ACM (Association of Computer Machinery) Digital Library *(Collection of all ACL publications, accessible via https://dl.acm.org/)*, allowed to study a publication(6) that offers a perfect overall synthesis of the existing software until then (June 2017) divided according to their different features and capabilities.

Thirteen different programs were evaluated, chosen to exclude those with similar functioning, plus one (PdfAct(7), initially called Icecite) presented in the article. Of these, the characteristics were analyzed in terms of identification of paragraph boundaries, correct reading order, and semantic roles of terms; translation capabilities of ligatures, diacritical marks, and hyphenated words; and possible output formats.

Once the capabilities of individual software were understood, the analysis in the publication shifted its focus to performance evaluations. We examined the possible errors in word or paragraph recognition on the output of an archive made up of about twelve thousand scientific articles randomly taken from arXiv, in PDF format. After a careful evaluation of the characteristics and performances of the different softwares, we concluded that the most suitable for our purpose is PdfAct, the one proposed by the authors of the article. The latter is a Java library capable of recognizing and separating LTBs (Logical Text Blocks) according to a rule-based approach that analyzes distances, positions and fonts of characters.

3.1.2 Automating text extraction with R

Following the literature review, some functions have been created, grouped in a larger R package called corecage , which allow to automatically perform the data extraction process entirely through R. These functions are able to create simple files in .txt format containing the text accurately extracted from the Governor's Final Remarks (in PDF format), simply providing the path of the folder where the PDFs are present and the path of the folder where you want to organize the new text files.

Although the software used is a Java library and therefore does not lend itself to being used directly in an R environment, it is possible to execute all the commands comfortably within the latter. *(As a string is executed directly as a System command thanks to the system base function. For further clarification on the functions used, please refer to the R help provided in the corecage package documentation.)*

This made it possible to have simple files containing only the terms present in the body and in the headings of the paragraphs of the Governor's final considerations. These 10 files (one for each year of interest) were, therefore, used for the creation of the corpus of documents on which the subsequent analyses were carried out, using the tm(8) library for text mining on R.

3.2 Lemmatization, categorization and cleaning

In order to accurately perform the process of reducing a flexed form of a word to its canonical form, called lemma, we used an external tool called TreeTagger. The latter is a tool that allows to annotate words contained in texts of various languages with the appropriate grammatical category and lemma and was developed by Helmut Schmid as part of the TC project at the Institute for Computational Linguistics at the University of Stuttgart. (9)

In order to use this tool in an R environment, we used a library called koRpus, available on CRAN, which offers several useful services for text analysis, including a wrapper for TreeTagger, and the support library for the Italian language (koRpus.lang.it).(10)

Starting from the output provided by this tool, it is therefore possible to reduce the words coming from the same lemma to a single entity and organize them according to their grammatical category. We will be able, therefore, to observe how the different parts of speech are distributed in our documents.

Once we obtained the lemmas organized by grammatical category, the so-called stopwords were removed from the data. The latter are the most common words of a language (such as articles or conjunctions), which are usually more present in a text and could create problems in the analysis. For obvious reasons, moreover, the terms that recur in these particular texts under analysis have been eliminated and therefore, specifically, words like “consideration”, “final”, “governor”, “bank”, “Italy”, etc. Finally, in order to obtain an exhaustive list of the stopwords of the Italian language, we used the stopwords function implemented in the tm package, which provides a list of terms to be removed for different languages, including Italian. (11)

3.3 Analysis of lexical differences

Using the describe function of the koRpus library it is possible to observe various descriptive statistics on the data resulting from the application of lemmatization with TreeTagger. These include several indices describing the number of characters in the documents (all characters, no spaces, only letters, etc), the number of words and sentences and their average length.(12)

Moreover, through a special function(13) of the same package, it was possible to calculate the various MTLD indices (Measure of Textual Lexical Diversity). These indices are clear indicators of the lexical richness of the text, as they are calculated from the ratio between the number of unique terms present in a text and the total number of words within it.*(Such a description of how the calculation occurs is decidedly simplistic. For further clarification, see McCarthy, Philip M., and Scott Jarvis. “MTLD, Vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment.” Behavior Research Methods 42, no. 2 (May 1, 2010): 381-92)* In fact, the increase in the number of words that are not repeated in a document corresponds to the use of a wider vocabulary, synonymous with greater lexical richness.

3.4 Word Cloud Creation

The lemmatized, categorized, and cleaned data were then organized according to the frequency with which they appear within the various documents, so that word clouds could be constructed using the R packages wordcloud(14) and wordcloud2(15).

Within the word clouds, then, terms will appear larger the more frequently they appear in the Governor’s Final Considerations. Underlying this weighting is the simple idea that in a document, the more frequent words are, the more important and meaningful they tend to be to the content.

The word cloud is a type of chart commonly used to succinctly visualize the content of a speech or set of documents, and can provide insights into understanding and interpreting the content of texts. From a statistical point of view, it is equivalent to a univariate frequency bar graph. Compared to this type of graph, the wordcloud certainly makes it more difficult to quantify the relative frequency of words, however, it has the advantage of allowing a visualization that immediately captures the relevance of words.

Furthermore, the comparison cloud and the commonalty cloud were used to compare documents within the corpus. In the former, word size is defined according to the different word presence rates within each document: the comparison cloud highlights, therefore, the differences. The commonality cloud, on the other hand, highlights words common to all documents. In this second case, the size of the word is a function of its minimum frequency across documents. So if a word is missing from any document it has zero size (i.e. it is not displayed).(16)

The word clouds of the comparisons were constructed by using the specific functions comparison.cloud and commonality.cloud of the wordcloud library, while the overall word clouds were created thanks to functions contained in the wordcloud2 package, which offers greater freedom in terms of graphic style.

For practical reasons, related to the difficulty of interpreting ten different clouds, we chose to group the documents in three periods. The first includes the years 2008 to 2010, the second those from 2011 to 2014, and the third those from 2015 to 2017. Comparisons were also constructed from the data thus divided.

Finally, because the items are organized into grammatical categories, it was possible to construct the word clouds by limiting analysis to only one of these types at a time. Given the limited information contained in secondary parts of speech, it was repeatedly chosen to include only nouns in the lemmatized wordclouds.

3.5 Sentiment analysis

In order to link to the words contained in the data their polarity, positive or negative, it is necessary to use a dictionary of opinion word (called lexicon), that is a real list of adjectives, nouns, verbs and adverbs to which are associated the emotions, and therefore the opinions, that they reflect. For example, the word "sun" will be associated with the emotion of joy and a positive opinion, while the word "abandonment" will be associated with a feeling of sadness and a negative polarity.

However, finding a lexicon properly constituted for the Italian language is an arduous task. It is undoubtedly easier to find one in English. For the following analysis, therefore, the polarities of the terms have been extracted from the Italian translations of two different dictionaries.

The first one is the subjectivity lexicon by Janyce Wiebe(17), professor at the Department of Computer Science and Intelligent Systems Program of the University of Pittsburgh, and has been used through some adaptations of the functions contained in the sentiment library. The second is the NRC emotion lexicon(18), provided by Saif M. Mohammad, Senior Research Scientist at the National Research Council Canada. The latter is accessible, in English, through the syuzhet library(19), so it has been used through adaptations of the functions contained in this package.

3.6 Topic modeling

The application of topic modeling algorithms allows us to identify the topics of each individual section that goes to make up an entire document. Thus far in the analysis, we have considered the data as a corpus consisting of ten texts, each containing the lemmatized terms of the Governor's Final Considerations. Given the goal of searching for recurring topics within the individual publications, it was necessary to reorganize the texts.

Ten different corpora consisting of each individual publication were created. Therefore, if up to now we have worked with a corpus of ten documents, from now on we will analyze ten corpora consisting of one document each.

In order to recognize the topics characterizing the sections of each publication, we used, by means of the LDA function present in the topicmodels library(20), a particular probabilistic model of the text called Latent Dirichlet Allocation (LDA). This is a very elaborate technique that lends itself to countless applications. To name a few, it is the method by which the results of a Google search are ordered by relevance, but also the one used by Amazon for clustering customers based on their purchases.(21)

The output provided by this function is loaded with information, including many of complicated interpretation. In order to simplify its understanding and use in line with our objective, it has been used in such a way as to obtain the four lemmas that are most likely to be part of four topics that make up the text. The choice of how many parts to divide the documents into and how many terms to study is arbitrary.

These words, since they are the ones most likely to be found together in one section of the publication rather than in another, will give a precise indication of the subject they refer to. It will be possible, therefore, to reconstruct the themes treated within the different texts and, consequently, to observe whether there are some that are repeatedly taken up over the years.

4 Results and discussion

4.1 Parts of speech distribution

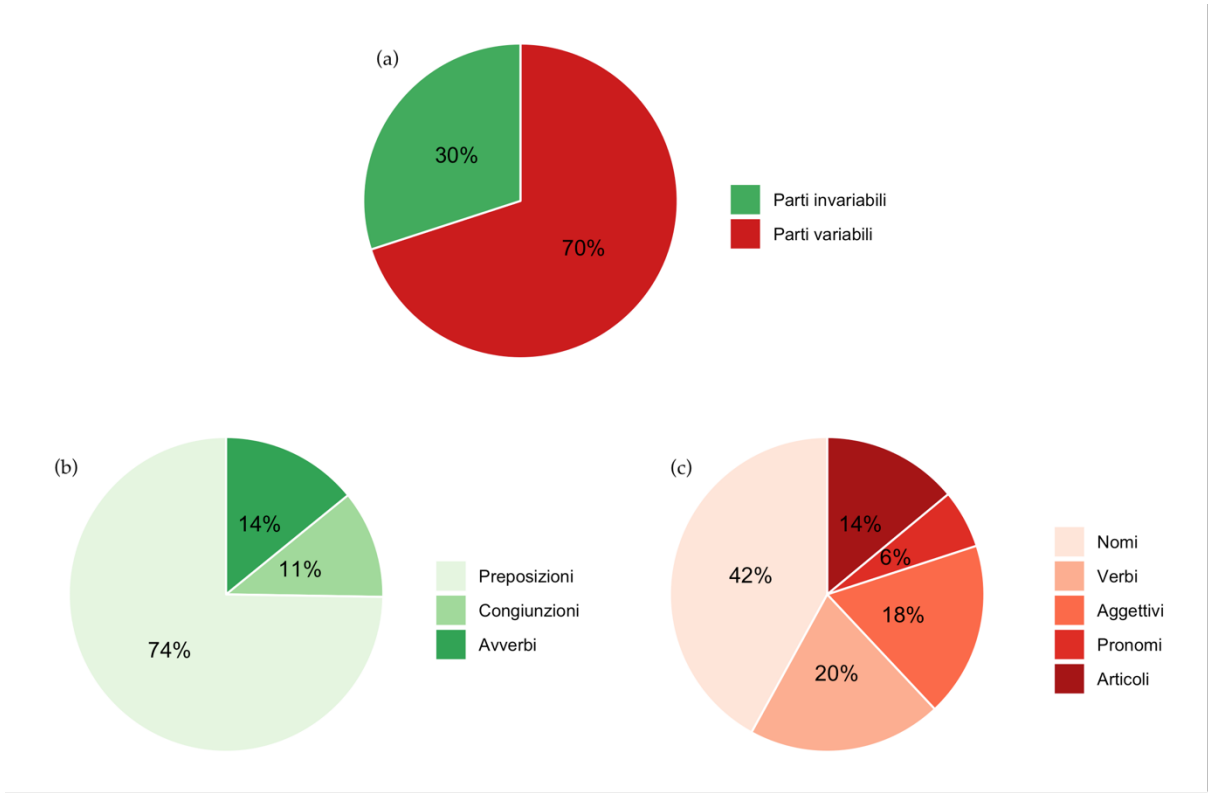

Figure 2: Average distribution of parts of speech (a), especially invariables (b) and variables (c)

Given the large number of terms present within the documents, the different distributions by year of the parts of speech are almost identical. Therefore, analyzing their variations over the period of time considered makes little sense. It is interesting, instead, to observe the average distribution of the nine parts of speech of the Italian language. The latter is represented in Figure 2.

Graph (a) shows that 30% of the data consists of invariable parts of speech, that is, prepositions, adverbs, conjunctions and interjections. In particular, as is clear from graph (b) which shows the distribution limited to this section of the data, most of the invariable parts of speech are prepositions (74%), followed by adverbs (14%) and conjunctions (11%). It is important to note the total absence of interjections in the observed texts. The reasons for this result certainly lie in the fact that exclamations express a particular emotional attitude of the author, which cannot be found in official publications such as those under analysis, to which

a formal register is appropriate.
Considering, instead, graph (c) of Figure 2, we can observe the average composition of 70% of the treatises, made up of the variable parts of speech. Even in this case there is one type, that of nouns, which is much more present than the others: it makes up almost half (46%) of the total. Verbs, adjectives and articles, on the other hand, are present in almost similar proportions, amounting respectively to 20, 18 and 14% of the total of the variable parts. Last we find pronouns, which contribute only 6%. The results observed, characterized by a rich distribution of nouns at the expense of adjectives, together with the absence of interjections, are in line with the textual typology analyzed. In fact, these particularities reflect those which, in linguistics, are proper to an expository text with scientific language, therefore objective. Moreover, we expect the latter to be denotative, therefore limited to the explicit and referential meaning of the word, without any freedom of interpretation.

4.2 Lexical characteristics

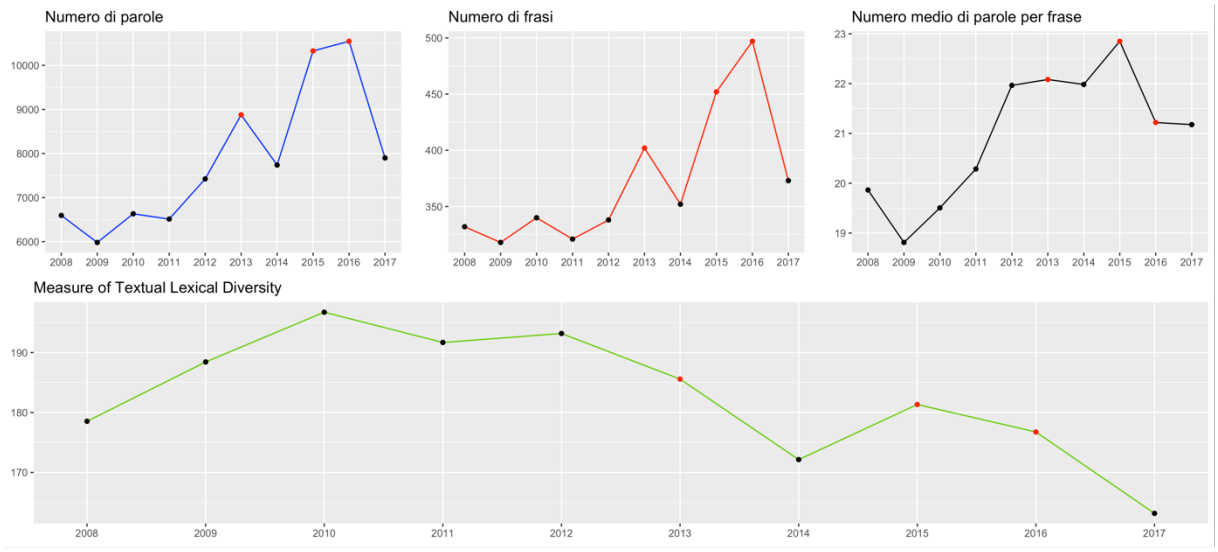

Figure 3: Overview of lexical features of the data over the reporting period

Figure 3 provides an overview of the number of total words and sentences, the average number of words per sentence, and the MTLD index of the publications for each year of the 2008-2017 period. From the first two graphs, we easily notice how, as time goes on, the Governor's Final Remarks tend to be longer and more comprehensive. The total number of words and sentences, in fact, after a minimum reached in 2009 (5982 words in 318 sentences), tends to increase in the interval up to 2017. Three peaks are easily observed, one of which is less significant in 2013 and two very relevant ones related to the years 2015 and 2016, which present, respectively, a number of words equal to 8877, 10328 and 10546 (against a ten-year average of 7853) and a number of sentences equal to 402, 452 and 497 (average equal to 373). The three observations from these three years are marked with a red dot in the graphs in Figure 3.

Shifting our attention to the graph regarding the average number of words per sentence, we observe a positive trend here as well, but a more stable one. We note, in fact, a jump between 2011 and 2012, in which we vary from sentences averaging 20.29 words long to sentences averaging almost 22 (21.96) words, but fairly stable values for the periods before and after the break. Again, the minimum is reached in 2009 (18.81) and the maximum in 2015 (22.85), compared to a total average of 20.97. The documents referring to the years 2013 and 2016, despite being the longest ones, show average values related to the length of the sentences that are little different from the average of the period 2012-2017, equal to 21.88.

The fourth graph in Figure 3 shows the values of the MTLD index, which reflects the lexical richness of the texts. We observe very different results from those presented so far: if previously we noticed an increasing

trend as time goes by, now the index of interest is distributed with a descending trend over the observation period. In fact, after a peak reached in 2010, where an MTLD index equal to 196.73 is recorded, with the passing of the years the value of the data has decreased until the minimum observed in the last year of interest, equal to 163.17.

Focusing attention on the overall comparison between the periods 2008-2010, when Mario Draghi was in office, and 2011-2017, when the Governor was Ignazio Visco, we can draw some interesting conclusions. It is clear, in fact, that in the last period of Draghi's tenure, there was a preference for more streamlined treatises, with fewer words and shorter sentences on average, in favor of a richer lexicon. The documents drawn up by Visco, on the other hand, are more exhaustive, composed of longer sentences on average and characterized by a lower lexical richness.

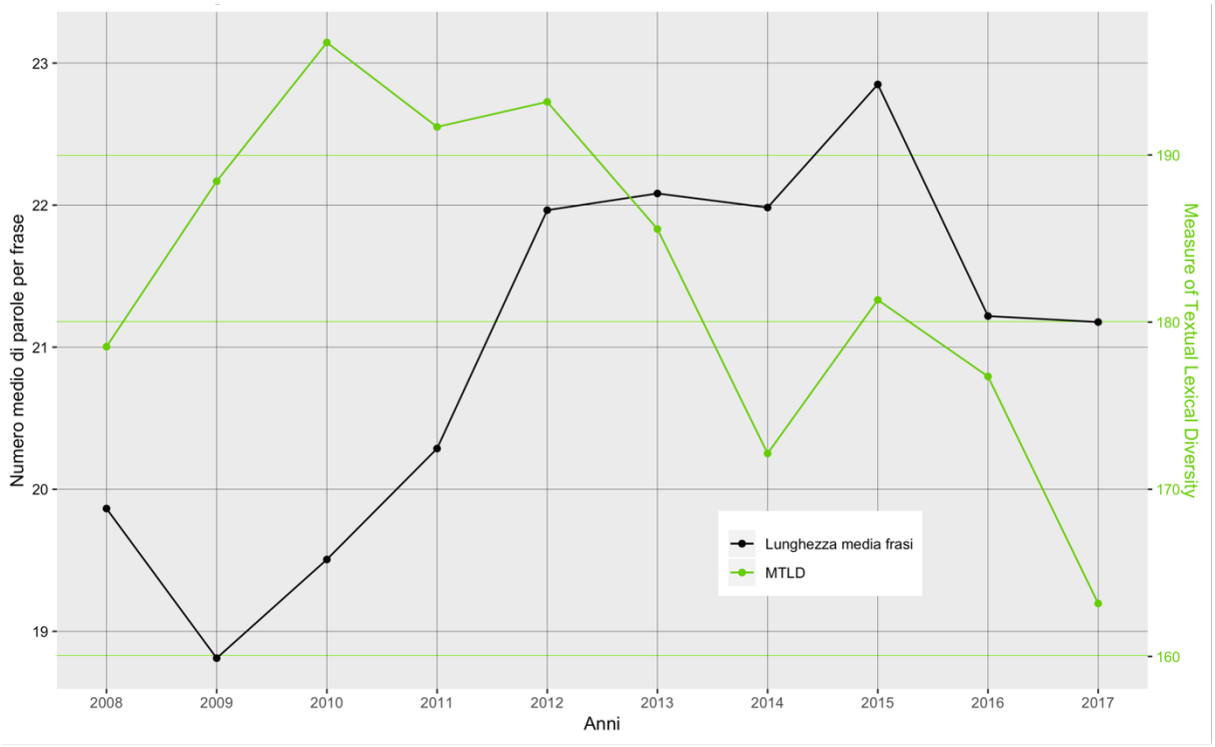

Figure 4: Comparison between average sentence length and lexical richness index

Moreover, if we study the comparison between the average length of sentences and the MTLD index, available in Figure 4, we come to further interesting results. The graph compares the trend of the average number of words per sentence, in black, and that of the lexical richness index, in green. It is easy to see how the two trends are opposite. That is, as time goes by, the sentences are, on average, longer and the terms present within them, on average, less sought after. Adopting a linguistic view of the data, we can conclude that longer and more complicated sentences, and therefore probably texts characterized by more subordination, correspond to the use of a less extensive vocabulary.

4.3 Wordcloud by periods

Figure 5 shows the cloud of words made up of the headwords present in the Final Considerations of the Governor in 2008, 2009 and 2010. It can be seen immediately that the most important words, i.e., those that recur most in the texts and, therefore, those that appear largest in the graph, are crisis, system, market, power and enterprise. It is clear that the publications reflect the situation of the Italian economy in the period of reference, in which Italy was in times of absolute crisis. It is likely, therefore, that they were talking about a crisis involving the market, the enterprise, or even the entire system. In the cloud we also find terms

Figure 5: Wordcloud lemmatized referred to the period 2008-2010

that relate to the conditions underlying the economic situation of the time, i.e. the high level of debt in relation to GDP, the low or absent economic growth and the lack of credibility of those in power.

Shifting our attention, instead, to the word cloud referring to the years from 2011 to 2014 depicted in Figure 6, we can draw different conclusions. Although terms linked to the crisis are still very present, a situation that we also find in the graph previously described, the importance of the term European provides a different interpretation. It can be seen, in fact, that in the period of reference, financing provided by the European Union played a primary role in lifting the country out of the crisis. In addition, the relevance of the supervisory activity carried out by the European Central Bank in the euro area, necessary to maintain financial stability and, therefore, to improve the general situation of the economy, has increased. Finally, the help provided by the Community is also underlined by the importance acquired by terms such as credit and fund, given that the monetary contribution is provided through various European Funds.

The word cloud referring to the years 2015, 2016 and 2017, shown in Figure 7, also reflects the situation of the Italian economy in that period. In fact, with the passage of time, aid from the European Union continued to be used, but the situation did not improve. This has led to the public debt (words most present in the treatments considered) increasing to historic highs(22). It is likely, therefore, that it constituted a primary problem of the country and, therefore, of ample treatment in the final considerations of the Governor.

4.4 Comparison cloud and commonalty cloud

Figure 8 and Figure 9 allow us to observe the differences and similarities in the terms present in the documents.

The comparison cloud (Figure 8) describes the lemmas that characterize one period rather than another, that is, it presents a term as belonging to one of the groups only if the latter is significantly more present in the reference class than it is in the others. Referring to our example, this means that the fact that the words crisis and system are linked to the first period reflects that, although they are present in all three groups of documents, they are more frequent in the years 2008, 2009, and 2010. It is interesting to note the absence, respectively in the second and third groups, of the terms European and public, which are fundamental for the conclusions drawn above. Finally, lemmas related to supervisory activity and debt remain discriminating.

Figure 6: Wordcloud lemmatized referred to the period 2011-2014

Figure 7: Wordcloud lemmatized referring to the period 2015-2017

Figure 8: Lemmatized comparison cloud of the three periods

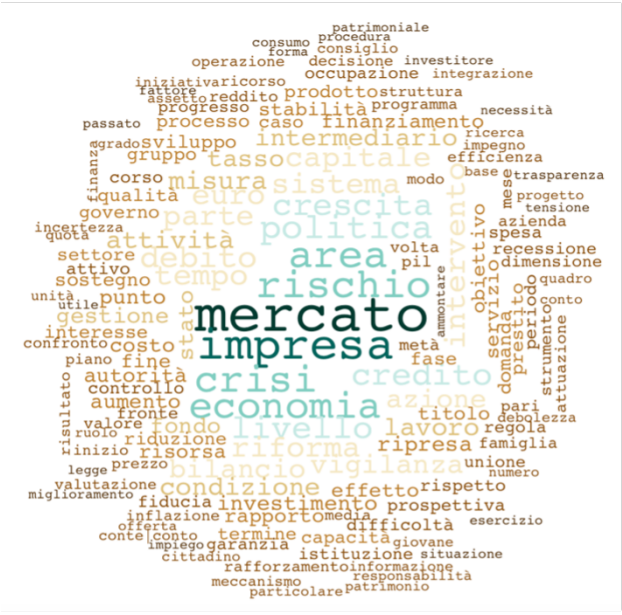

Figure 9: Lemmatized commonalty cloud

The commonalty cloud (Figure 9), on the other hand, presents the words that are most common to all documents, the larger the size of the minimum frequency between documents. In the graph, therefore, we find the general topics that the publication deals with, such as the market, the company and the economy, but also the budget, the politics, the growth of the Italian area.

Finally, it is interesting to note that words such as crisis, risk, credit or debt are frequent throughout the documents. This is evidence of the fact that the period analyzed was characterized by a turbulent economic situation.

4.5 Positivity of the lemmas

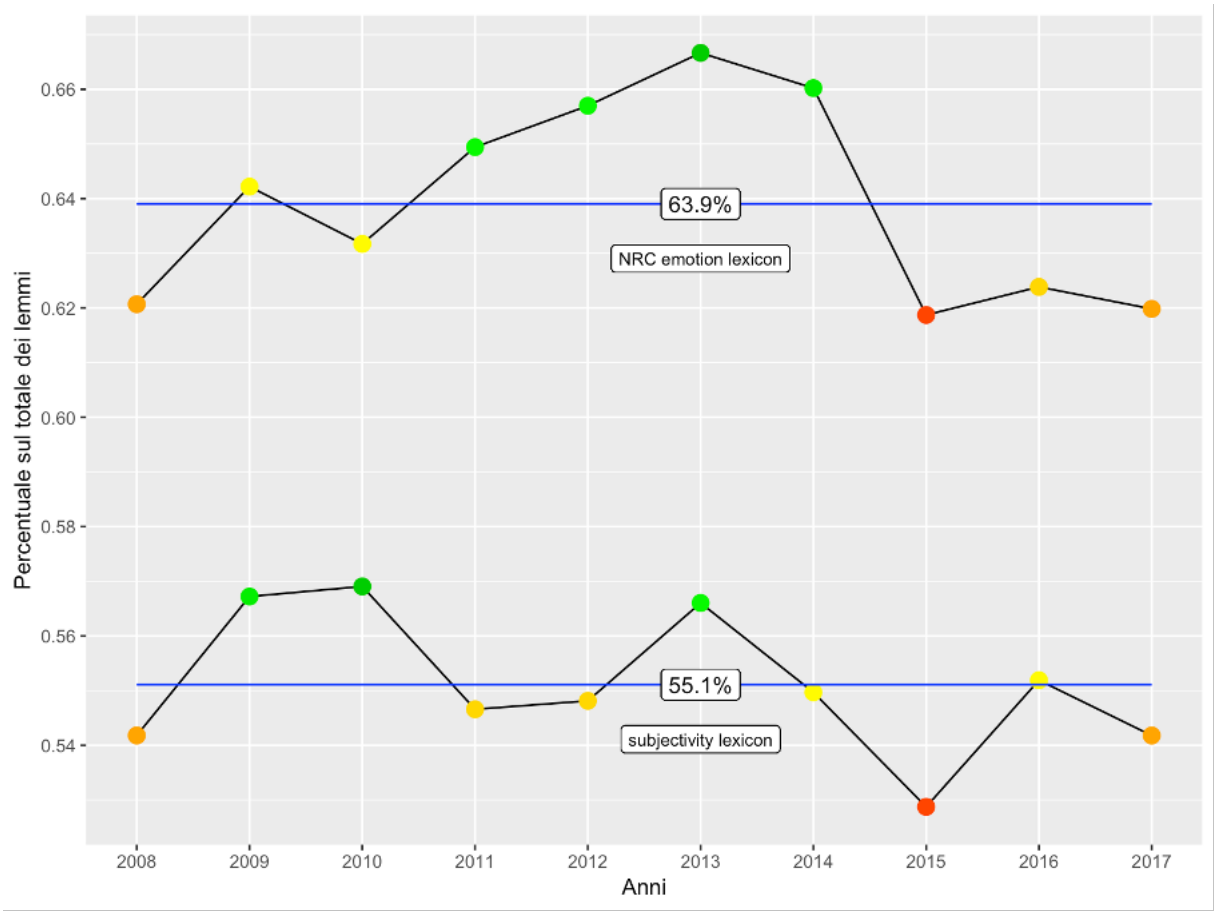

Figure 10: Concentration of positive lemmas: comparison of results obtained with different lexicons

As described in the previous chapter, the opinion mining procedures were performed by using two different lexicons. In Figure 10 it is possible to observe the comparison between the concentrations of positive lemmas within the ten documents obtained from the two dictionaries. It is important to note that all observations are arranged above the 50% threshold, meaning that most of the lemmas in the publication, despite being used objectively to discuss topics that are not always happy, reflect a positive opinion of the author. This observation is much more relevant in the results obtained through the use of the NRC emotion lexicon. The latter, in fact, present much higher concentrations of positive lemmas, as can be seen by comparing the averages: 63.9% of the data obtained with the NRC emotion lexicon, 55.1% with the subjectivity lexicon. Moreover, as regards the trends, it can be seen that these are similar and characterized by a peak of positive polarity, reached in 2013, and an absolute minimum, recorded in the year 2015 (to which corresponds, therefore, the highest concentration of negative opinion).

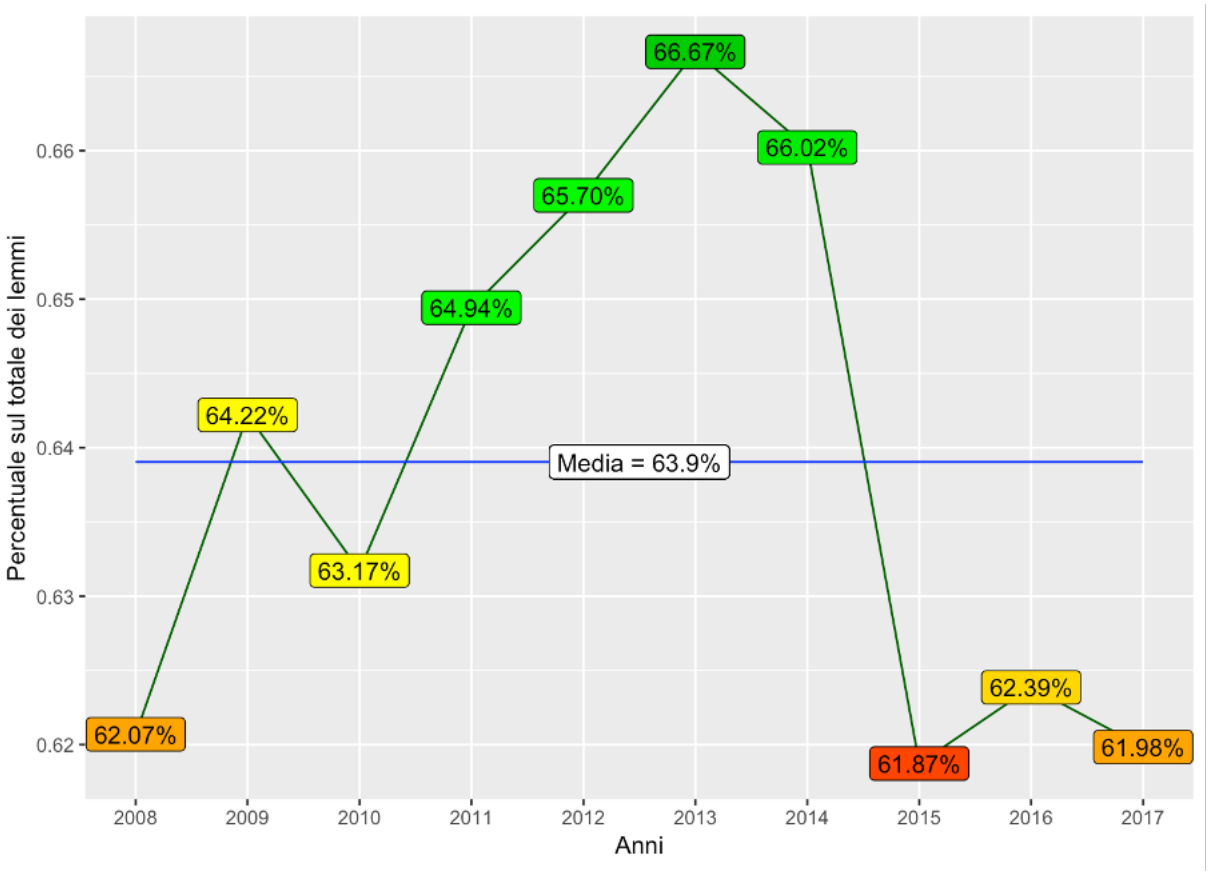

Figure 11: Trend in the concentration of positive lemmas (NRC emotion lexicon)

Of the two dictionaries, the one provided by the National Research Council Canada performed better in terms of recognizing the polarity of terms. It makes more sense, then, to analyze in detail the results obtained from this dictionary, shown in Figure 11.

There is little variation in the performance, with a range of just under 62% to almost 67%, which is little different from the overall average of 63.9%. Moreover, it is easy to distinguish three periods characterized by a similar concentration of positive lemmas. The first includes the years 2008, 2009 and 2010 and is characterized by results close to the overall average, especially for the years 2009 and 2010; the average number of observations in this period is, in fact, 63.2%. The second period, on the other hand, corresponds to the years 2011 to 2014 and shows concentrations decidedly above the intermediate value. In fact, the relative average is 65.8%. Finally, the publications of 2015, 2016 and 2017 are characterized by a significantly lower presence of positive lemmas: the average value referred to these last three years is 62.1%.

4.6 Topics of the parts of the text

Anni	Topic 1	Topic 2	Topic 3	Topic 4	Anni	Topic 1	Topic 2	Topic 3	Topic 4
2008	impresa	partecipante	italiano	finanziario	**2013**	pubblico	finanziario	credito	partecipante
	potere	finanziario	credito	mercato		potere	vigilanza	bancario	esercizio
	crisi	vigilanza	bancario	intervento		politica	nazionale	capitale	riserva
	pubblico	direttorio	crisi	internazionale		economico	autorità	impresa	funzione
2009	pubblico	finanziario	finanziario	mercato	**2014**	potere	mercato	attività	Risoluzione
	crisi	crisi	partecipante	bancario		inflazione	bancario	area	Crisi
	punto	dovere	filiale	intermediario		pubblico	vigilanza	rischio	Mercato
	pil	mercato	economia	europeo		crescita	finanziario	impresa	Europeo
2010	impresa	regola	primo	finanziario	**2015**	impresa	nuovo	potere	crisi
	pubblico	rischio	capitale	pil		potere	europeo	europeo	intervento
	italiano	politica	crescita	crisi		pubblico	vigilanza	finanziario	bancario
	spesa	crisi	grande	sistema		economia	finanziario	deteriorare	europeo
2011	mercato	primo	credito	mercato	**2016**	potere	lavoro	politica	debito
	finanziario	finanziario	rischio	finanziario		grande	occupazione	economia	pubblico
	potere	dovere	intermediario	europeo		pil	sistema	investimento	crisi
	pubblico	azione	bancario	bancario		europeo	produttivo	area	mercato
2012	pubblico	vigilanza	impresa	area	**2017**	impresa	impresa	finanziario	debito
	produttivo	attività	bancario	condizione		potere	crescita	potere	pubblico
	potere	nazionale	credito	euro		mercato	economia	deteriorare	finanziario
	attività	direttorio	dovere	europeo		spesa	attività	relazione	potere

Figure 12: Four headwords most likely to be part of the four topics in each of the ten publications

The application of the LDA algorithm resulted in the lemmas that are most likely to be jointly present within one section of the text rather than another. This means that these four words are the ones that best define the topic they refer to. Therefore, from an analysis of these terms it is possible to reconstruct the four topics addressed in each document and, consequently, to observe if there are any recurrences.

The table in the Figure 12 summarizes the 16 terms of each publication, so that you can get an overview and interpret them better. An example of the output of R is instead provided in Figure 13.

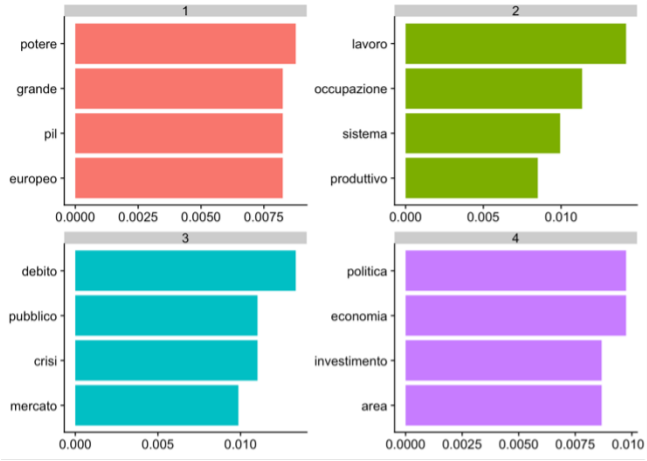

Figure 13: Average distribution of parts of speech (a), especially invariables (b) and variables (c)

The graph refers to the results obtained by analyzing the Governor's Final Considerations regarding the year 2016. It is possible to observe the four lemmas that most characterize a given topic and the probability connected to them of belonging to the topic in question. This means, for example, that the lemmas debt, public, crisis and market are those that with greater probability refer jointly to one of the four topics (3, in the graph). This means, likely, that in the 2016 paper one section deals with the issue of public debt and market crisis.

For each year, therefore, the keywords have been reorganized in the table in the Figure 12. Following a careful analysis, it is possible to recognize different groups of recurrent words. The most present is certainly that referring to the lemmas business and public, often accompanied by power (in red in the table). It can be said, therefore, that in 9 documents out of 10, a section of the text is dedicated to the description of the situation of businesses in the market, or to the problem of the consolidation of public accounts (i.e. the state budget). Another recurring topic is that of banking and financial supervision activities carried out by the Bank of Italy at the national level, present in 5 sections and highlighted in yellow. We also often find topics linked to the central role played by banks in financing the Italian economy. On 4 occasions, in fact, we can observe groups containing the lemmas credit and banking (in purple). Moreover, we observe recurrent topics that characterize only a few documents. Clear examples are those highlighted in green, orange and blue, which we find only in two of the last four years in analysis and refer respectively to the European market in relation to the crisis, the deterioration of financial power and the issue related to public debt. It is also interesting to note how in 2016 the theme of work gains importance (in gray), likely due to the anti-unemployment policies introduced in those years.

Finally, it can be observed how the terms crisis and public debt are present above all in the first and third macroperiod of treatment respectively. Words related to the European market, instead, are frequent throughout the table. This almost perfectly mirrors the results observed through the analysis of word clouds.

5 Conclusions

Following the presentation of the results, it is necessary and proper to reconstruct a common thread within them and discuss the limits of the analysis performed.

The path followed has made it possible to delve into three fundamental aspects of the Governor's Final Remarks: the style and language used, the tones and opinions through which the concepts were presented, and the themes dealt with over the years.

With regard to the first point, the analysis of the average distribution of the parts of speech made it possible to trace the textual typology to which the data belong to that typical of an expository text with scientific and objective language. Moreover, by examining the evolution over time of the length of words and sentences and of the MTLD index, it was concluded that, as the years go by, treatises tend to be more and more exhaustive and characterized by a more restricted vocabulary.

Regarding tones and opinions, opinion mining operations allowed to demonstrate how most of the terms in the treatises are related to positive emotions. Moreover, starting from the trend of polarity over the years, it was possible to reconstruct three periods marked by similar characteristics. The first, consisting of the years 2008, 2009 and 2010, shows positivity in line with the overall average. The intermediate one, from the years 2011 to 2014, is characterized by a significantly higher concentration of positive lemmas. The last interval, consisting of 2015, 2016 and 2017, is instead marked by less positive average terms.

We also find these considerations in the analysis of themes. In fact, the wordclouds and the results of the application of topic modeling show that the topics of greatest discussion in the three periods are, respectively, the crisis, the European market and public debt. These results were confirmed by the observation of the recurrent topics in the treatments over time. Finally, the study of the lemmas present in the commonalty cloud has allowed us to reconstruct the macro-topics that the publication generally deals with, i.e. the market, companies and the economy.

However, it must be said that the conclusions described, although decidedly relevant and in line with what has been the history of the Italian economy in the period of reference, may have been subject to two different forms of distortion.

The first is related to the inevitable loss of information that occurs during automatic text extraction and recognition processes. In fact, the functions, algorithms and software used, despite being unquestionably accurate, are not flawless. Therefore, it can happen, and according to the law of large numbers it has certainly happened, that a word is recognized incorrectly or not at all.

The second, unfortunately of greater potential importance, is connected to the interpretation that has been given to the results. If before, therefore, it was a matter of machine inaccuracy, now we speak of human error. It is natural that the commentary, despite being made from objective data, is subjective, and therefore necessarily different from what another individual would have offered. On the other hand, it is impossible to identify an interpretation that is undeniably right or wrong, or even one that is more right than another. For my part, I can consider myself decidedly satisfied. And that's what's important.

Acknowledgements

There is also a thank you section in the official publication.

These are "personal" things, therefore not concretely relevant to the purpose of this translated publication and, above all, more likely to lose meaning when expressed in another language.

This obviously doesn't mean that I don't dedicate it to my parents or that I don't thank all the people I mentioned in the other document.

It just means that, if you are particularly curious, you will have to go and read them in Italian here.

References

(1) Banca d'Italia, «Banca d'Italia - Chi siamo», consultato 17 giugno 2019, https://www.bancaditalia.it/chi-siamo/index.html.

(2) «Governatore della Banca d'Italia», in Wikipedia, 15 maggio 2019, https://it.wikipedia.org/w/index.php?title=Governatore_della_Banca_d%27Italia&oldid=104906596.

(3) Ramzan Talib et al., «Text Mining: Techniques, Applications and Issues», International Journal of Advanced Computer Science and Applications 7, n. 11 (2016), https://doi.org/10.14569/IJACSA.2016.071153.

(4) «Text mining: il processo di estrazione del testo», Lorenzo Govoni (blog), 16 luglio 2018, https://lorenzogovoni.com/text-mining/.

(5) Banca d'Italia, «Banca d'Italia - Interventi del Governatore», consultato 20 giugno 2019, https://www.bancaditalia.it/pubblicazioni/interventi-governatore/index.html.

(6) Hannah Bast e Claudius Korzen, «A Benchmark and Evaluation for Text Extraction from PDF», in Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries, JCDL '17 (Piscataway, NJ, USA: IEEE Press, 2017), 99–108, http://dl.acm.org/citation.cfm?id=3200334.3200346.

(7) «ad-freiburg/pdfact: A basic tool that extracts the structure from the PDF files of scientific articles.», consultato 12 febbraio 2019, https://github.com/ad-freiburg/pdfact.

(8) Ingo Feinerer et al., tm: Text Mining Package, version 0.7-6, 2018, https://CRAN.R-project.org/package=tm.

(9) «TreeTagger - a language independent part-of-speech tagger | Institute for Natural Language Processing | University of Stuttgart», consultato 20 giugno 2019, https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.en.html.

(10) Meik Michalke et al., koRpus: An R Package for Text Analysis, version 0.11-5, 2018, https://CRAN.R-project.org/package=koRpus.

(11) «stopwords function | R Documentation», consultato 20 giugno 2019, https://www.rdocumentation.org/packages/tm/versions/0.7-6/topics/stopwords.

(12) «Using the koRpus Package for Text Analysis», consultato 21 giugno 2019, https://cran.r-project.org/web/packages/koRpus/vignettes/koRpus_vignette.html#accessing-data-from-korpus-objects.

(13) «MTLD: Lexical Diversity: Measure of Textual Lexical Diversity… in KoRpus: An R Package for Text Analysis», consultato 21 giugno 2019, https://rdrr.io/cran/koRpus/man/MTLD.html.

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

(15) Dawei Lang e Guan-tin Chien, wordcloud2: Create Word Cloud by «htmlwidget», version 0.2.1, 2018, https://CRAN.R-project.org/package=wordcloud2.

(16) Fabrizio Alboni e Ignazio Drudi, «Materiale didattico fornito per l'insegnamento Utilizzo Statistico di Banche Dati Economiche Online, Laurea in Scienze statistiche, Università di Bologna», 2018-2019.

(17) «Janyce Wiebe/Jan Wiebe», consultato 22 giugno 2019, https://people.cs.pitt.edu/~wiebe/.

(18) «NRC Emotion Lexicon», consultato 22 giugno 2019, https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

(19) Matthew Jockers, syuzhet: Extracts Sentiment and Sentiment-Derived Plot Arcs from Text, version 1.0.4, 2017, https://CRAN.R-project.org/package=syuzhet.

(20) Bettina Grün e Kurt Hornik, topicmodels: Topic Models, version 0.2-8, 2018, https://CRAN.R-project.org/package=topicmodels.

(21) «Come funziona LDA - Amazon SageMaker», consultato 28 giugno 2019, https://docs.aws.amazon.com/it_it/sagemaker/latest/dg/lda-how-it-works.html.

(22) «Debito pubblico: come, quando e perché è esploso in Italia», Il Sole 24 ORE, consultato 25 giugno 2019, https://www.ilsole24ore.com/art/debito-pubblico-come-quando-e-perche-e-esploso-italia-AEMRbSRG.