Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Central Bank Communications: information extraction and semantic analysis[1]

Giuseppe Bruno,
Bank of Italy

[1] This paper was prepared for the meeting. The views expressed are those of the author and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

Central Bank Communications: Information extraction and Semantic Analysis.

Giuseppe Bruno*

Bank of Italy, Economics and Statistics Directorate.

Abstract

Central Banks, among other tasks, provide a relevant amount of information for Institutions and market operators. Indeed central banks employ a multiplicity of communication channels to drive market expectations. In this paper we present some methodologies aimed to quantify the information content of official communications and we present their application to the semi-annual publication of the Financial stability report. While these methodologies are quite developed for the English and other highly spoken languages in the world, they are still in their experimental phase for the Italian language. Here the goal is twofold: on one hand we provide a transparent numerical framework to consider sub-unit of an official Central Bank report written in Italian. Moreover it is proposed an analytical tool to gauge the impact of an official document on the public. In the context of reports released by the Bank of Italy, we show how this framework can be employed to numerically characterize and extract their information content.

We deem quite relevant a quantitative evaluation of the impact of these reports in increasing the central bank transparency with the goal of enhancing the effectiveness of its institutional action.

JEL classification: C83, E58, E61.
Keywords: Text Mining, Semantic Analysis, Pointwise Mutual Information, Web search.

1. Bank of Italy, Research Area.
* giuseppe.bruno@bancaditalia.it
This version February 2017
The author wishes to thank Fabio Di Bernardini for his helpful suggestions and unceasing cooperation.
The views expressed are the author's only and do not imply those of the Bank of Italy.

1 Introduction and motivation

For a large class of cases – though not for all – in which we employ the word meaning
it can be defined thus: the meaning of a word is its use in the language game.
– Wittgenstein, Philosophical Investigations A. 43

In the last two decades, the amount of textual information available at everybody fingertips has soared. According to rough estimates 80% of the total amount of web pages on the internet is given by textual or unstructured data. In this paper we take the challenge of adopting and experimenting a methodology for quantifying the linguistic content of some official documents of the Bank of Italy. In particular we take into consideration the Italian version of the Financial Stability report (henceforth *FSR*) . This is a young publication whose first issue came out on 2010. We build a small corpus of all the available issues of the *FSR* and we show how to convert it into a vector space model by employing familiar concept of linear algebra such as eigenvalues and eigenvectors. Among these vector space models, we consider here the *Latent Semantic Analysis* (henceforth *LSA*) model. This is one of the most successful models which is emerged in computational linguistic around 20 years ago (Landauer and Dumais [9] and Landauer et al. [10]). *LSA* was patented in 1998 in the US and it is a widely used technique in Information Retrieval (*IR* and natural language processing for analyzing relationships between a set of documents and the words they contain. Within the framework of the conceptual model of lexical semantics words, statements, chapters and whole documents are represented as high dimensional vectors in the same space. The main advantage of this closed representation is the natural metric induced in the vector space. Therefore in the *LSA* model we can:

1. compute semantic similarity measures between words/documents by exploiting the statistical redundancies in text;
2. compute word neighborhood which are set of words/documents sharing semantic concepts (synonimity);
3. compute text coherence and summary of given documents;
4. answer to multiple choice questions to topics dealt with in the corpus.

We make one step further by taking inspiration from previous works on similar topics such as Lucca and Trebbi [11], Carvahlo et al. [2] and Kawamura et al. [8] and using Web search queries in Italian, we evaluate the public perceptions associated with some keywords associated with the financial stability issues. Although in recent years we have seen a great progress in sentiment classification, it is still challenging to develop a practical sentiment classifier for open applications. Lexical semantic orientation can be measured by complementing the *LSA* with the Pointwise Mutual information (*PMI*) which has proved to be a dominating indicator for sentiment classification (Turney [18]).

The web-derived computation of the *PMI* allow us to measure the people orientation about some relevant theme for the financial stability.

Aside form the previous two economic goals of the paper, here we compare the effectiveness of the most relevant web search engines[1] and the simplicity of two very popular open source software frameworks: **Python** and **R**[2]. The two software packages have been employed for the statistical analysis and for producing the code interacting with the search engines.

In particular, with reference to our corpus composed of the whole set of the Italian version of the Bank of Italy Financial stability reports, this work attempts to answer the following questions:

1) what are the most relevant concepts considered in the documents of the corpus between 2010 and 2016?

2) what is the readability and formality level of each document?

3) how can we measure the impact of the *FSR* on the readers by web searching?

The buoyant literature on these themes (see for example D. Bholat and Schonhardt-Bailey [3] and R. Nyman and Tuckett [14]) confirms the growing interest shown by public institutions and Central Banks on these issues.

The development of web tools for monitoring and extracting sentiment orientation, provides further analytical support fostering a wider adoption of text mining and semantic techniques for improving the statistical accuracy required for assuming well informed decisions.

The paper is arranged in the following way. After this introduction in section 2 we present the basic theoretical concepts behind the text-mining techniques. Section 3 introduces the algebraic concepts behind the Latent Semantic Analysis. Section 4 describes the corpus of the *FSR* for our empirical application. Section 5 presents the results of our semantic analysis carried out on a corpus containing all the 11 editions, from 2010 to 2016, of the Bank of Italy *FSR*. Section 6 presents the methodology employed for evaluating the people orientation with respect to some economic issues through web search. Finally section 7 provides some concluding remarks.

2 Building a Corpus of Text documents

Quantitative analysis of human languages allows to discover common features of spoken or written text. In order to simplify the analysis of written text we have adopted the *Bag-of-words* model (see Salton and McGill [16]). This set-up considers each document as a sequence of different words where the semantic meaning of any statement is conveyed by the word-document co-occurrence while neither grammar nor word order play any determinant role[3]. Therefore this is an orderless representation of the document where we forgo any grammatical relationship among the words. Within the *Bag-of-words* framework each document

[1] Here we have considered Bing, Google and Yahoo which reached about 90% of the market share in 2015.

[2] `www.python.org` and `cran.r-project.org` respectively

[3] A bag or *multiset* is a set-like object where only element multiplicity is accounted for, while the order of its elements is irrelevant.

is represented simply by a vector whose length is the size of the corpus vocabulary and each vector entry is a weighted count of the word occurrences.

Our analysis is based on the text mining methodology which can generally be defined, see for example Hearst [5], as the task of harnessing *a large online text collections to discover new facts and trends about the world itself*. The development of a corpus of textual homogeneous documents is the first step to address any kind of text mining and *LSA* applications. The only determinant factor for the semantic value of a text is the frequency of occurrences of each word. At first glance this seems a pretty unplausible assumption, but for the purposes of information retrieval, comparison and measures of similarity among documents will prove very effective. Our corpus is defined as a set of documents written in the same language. The whole set of different terms constitutes our vocabulary $V = \{k_1, k_2, \ldots, k_N\}$ whose size N is the cardinality of the set. In our examples the terms are words and each one of them is an independent dimension in our vector space [4]. Therefore any word/document is represented by a vector in the space $\mathbb{R}^N$. Since each one of our documents contain a subset of the Vocabulary V their vector representation will be very sparse.

It is evident from this description that the atomic building block of a corpus is a word. Therefore, for processing reasons, all the documents must be converted in plain textual format so that it is straightforward to tally the occurrences of the different words in each document belonging to the considered set. On the other hand, the choice to take into account just the main text imposes the need to give up any consideration about tables, figures and other external elements such as notes, bibliography etc. To this end we proceed by extracting the plain text from the original PDF© or MS-Word© format. Once we convert our original documents into a set of text files we can start the preproprocessing steps, which, depending on the final goals, consist in some of the following tasks:

- lower case conversion and white space removal,
- stopword and number removal,
- stemming or lemmatization,
- special characters conversion or filtering.

The realization of some of these tasks aims to reduce the size of the considered vocabulary allowing to focus on the most relevant topic-determinant words. While tasks like lowercase conversion, number and white space removal are independent of the considered language, for other tasks the language employed in the documents plays a relevant role. Although the analytical instruments for the English language are well developed, software tools for the italian language are yet in their growing phase. In this work we have tested the software capabilities of some of the **R** packages addressing their suitability for carrying out these language dependent tasks on text written in Italian.

The stopword list available in the **tm** package is composed of around 300 tokens which belong to grammatical categories such as determiners, pronouns, conjugated forms of auxiliary verbs. In our empirical analysis we had to complement this list with other italian words with poor information relevance. The stemmer provided by the **SnowballC** package has seemed suitable for the Italian language employed in our official documents[5]. Other preliminary text processing tasks such as synonymous replacement and part of speech tagging are not considered here because out of the scope in our analysis. Once the required preliminary tasks are completed our corpus of text documents is translated into a term by document matrix (henceforth *TDM*) where each entry $a_{i,j}$ gives a weighted frequency of the word i in the document j. This normed vector space representation for the set of our documents constitute the starting point for the semantic analysis described in section 3 .

[4] A single word will be a vector with all zeros but a one in the position of the word in the vocabulary.

[5] SnowballC is a R interface to the C stemmer which implements the Porter's algorithm (see Porter [12] and Porter [13]).

3 Latent Semantic Analysis: algebraic background

When we consider the learning speed with which people stockpile knowledge, we are faced with an apparent puzzle, where it is hardly possible to explain the amount of people's word knowledge by considering the shallowness of their information set. In the past, for example, Landauer and Dumais [9] or Landauer et al. [10] went even way back to Plato's hypothesis that *people must come equipped with most of their knowledge, needing only hints and contemplation to complete it*. Latent Semantic Analysis (*LSA*) tries to set up on more solid foundations the puzzle of excessive learning speed by assuming a multiplicative inference generated by the relationships available with the present stock of knowledge.

The *LSA* methodology was originally suggested by S. Deerwester and Harshman [15] in the framework of automatic indexing and information retrieval (IR). The suitability of *LSA* for different text analytics purposes has been already established in different fields among which we have the just mentioned IR. Adoption of this technique is already significant among central banks, see, for example, Boukus and Rosenberg [1] who analyze the information content in the *FOMC* of the Federal Reserve Board, Hendry and Madeley [6] who carry a similar analysis for the Central Bank of Canada. These two works perform text mining tasks on a corpus of the English language. Carvahlo et al. [2] employs the *LSA* to quantify the informational content in the portuguese version of the statements of the "Comitê de Política Monetária" [6] of the Central Bank of Brazil while Kawamura et al. [8] analyzes the Japanese version of the Monthly Report of the Bank of Japan to check whether the central banks communicate strategically being selective about the type of information they disclose.

Application of the *LSA* methodology rests on the *Bag-of-words* model, where word occurrences and co-occurrences build up the basic information set. The most interesting feature of the *LSA* is its use of the same vector space for both words and documents. In this closed framework a word meaning is not an absolute concept but it is an average of the meaning of all the statements of the corpus including this word. These averages are numerically computable and this quantitative feature is one of the main advantage of the model.

The input element for the *LSA* analysis is a *TDM* matrix as produced with the processing steps described in the chapter 2. Starting with a *TDM*, the *LSA* algorithm can be broken up in three separate steps. The first one consists in rebalancing the relative impact of low- and high-frequency word by applying a weighting scheme to the *TDM*. Among the available weighting schemes we have considered the Tf-Idf (Term frequency-Inverse document frequency) which is very popular in the IR domain[7]. This scheme consists in weighting each nonzero element of the *TDM* in the following way:

$$\omega_{i,j} = wf_{i,j} \cdot \left(\log \left(\frac{m}{df_i} \right) \right) \tag{1}$$

where: $wf_{i,j}$ is the frequency of word i in document j, df_i is the number of documents containing the word i, the total number of documents in our corpus is m, $\log \left(\frac{m}{df_i} \right)$ is the log of the inverse document frequency and $\omega_{i,j}$ is the final value given to word i in document j in the *TDM* matrix. This means we multiply the raw frequency by the logarithmically scaled inverse of document fraction containing the chosen word.

Tf-Idf is a very intuitive weighting scheme which provides a higher weight to words occurring frequently in very few documents (topic determinants words) while giving lower weight to words uniformly present in all the documents of the corpus [8].

In the second step of *LSA* linear algebra kicks in. At this stage we employ the Singular Value Decomposition (SVD) to factorise our weighted rectangular *TDM* matrix in three factors. This decomposition is a generalization to the eigenvalue/eigenvector decomposition by representing each term and document in an orthonormal base. Its importance derives from the circumstance that it can be applied without restrictions to

[6] Monetary Policy Committee.

[7] This scheme provides an higher retrieval accuracy with respect to the simple Tf (Term frequency).

[8] The Idf of a word appearing in every document of the corpus will be zero.

any rectangular matrix. Formally by assuming the weighted *TDM* A_w of size $m \times n$, the SVD decomposition corresponds in determining the three matrices in the right hand side of the following equation:

$$A_w = U \cdot \Sigma \cdot V^T \tag{2}$$

where: U is a $m \times m$ orthonormal matrix containing the left eigenvector, V is a $n \times n$ orthonormal matrix containing the right eigenvector, and Σ is the $m \times n$ rectangular matrix with elements $\sigma_{i,j} = 0$ if $i \neq j$ while the elements $\sigma_{i,i}$ are the singular values. The rows of U contains the word vectors, while the rows of V contains the document vectors. The third and last step of *LSA* corresponds to the dimensionality reduction that is implemented by setting to zero all the singular values below a certain threshold. In our empirical application we considered a very small number of documents [9], therefore we did not carry out this reduction which is required when the number of documents is in the range of hundred of thousands.

The representation of words and documents as vectors in $\mathbb{R}^N$ allows for a straightforward evaluation of a numerical value for the similarity two elements in the vector space. A commonly used measure is the cosine similarity between the two document vectors $\mathbf{x}, \mathbf{y}$ which is given by:

$$\cos(\theta) = \frac{\sum_{i=1}^{N} x_i \cdot y_i}{\sqrt{\left(\sum_{i=1}^{N} x_i^2\right)\left(\sum_{i=1}^{N} y_i^2\right)}} \tag{3}$$

where: θ is the angle between the two document vectors. In absence of weighting all the vector components are positive and $cos(\theta)$ will range from 0, for completely different documents, to 1, for very similar documents. In the more general case of logarithmic weights, therefore $cos(\theta)$ might span its whole range from -1 to 1. Here the similarity concept pertain the use of the same or co-occuring words regardless of their order[10]. In the following sections different applications of this similarity measure will be presented and employed in different circumstances.

4 The Bank of Italy Financial Stability Corpus

For our empirical application we have built a small corpus composed of all the issues of the *FSR*. Bank of Italy rolled out the publication of the FSR in 2010[11]. It started as a yearly publication but in 2012 the report became biannual. The whole report consists of 4 chapters for a total of about 40 pages for issues. Each publication includes an average of 50 graphs, 5 tables and around 10 in-depth information boxes.

To the purpose of our analysis, the documents have been converted in plain text after discarding tables, graphs and other auxiliary elements. Each Report has been split in its component chapters. At the completion of this step we had a corpus of 58 text documents. In the table 1 we show some descriptive statistics about the corpus considered in our experimental set-up. Here we can see a decreasing number of sentences over time with a stable number of word per sentence, characters per word and characters per sentence. These shallow text statistics play a key role for the estimation of the readability and formality of the considered documents. Beside from these general figures, the first statistical analysis carried out on these documents is based on word cloud and heatmaps.

The word cloud is a synthetic picture showing the principal words in the document by resizing their fonts proportionally to their relative frequency. As an example we provide the wordmap for the whole corpus of the 11 issue of the *FSR*. Some of the most relevant words are: *rischio*, *credito*, *liquidità* and *banche* [12]

[9] By splitting in their chapters the 11 *FSR* issues we ended up with a corpus of 58 documents

[10] *LSA* allow to pin down synonyms by harnessing semantic similarity.

[11] Electronic version of the documents are available online at http://www.bancaditalia.it/pubblicazioni/rapporto-stabilita/index.html.

[12] The English translation is risk, credit, liquidity and banks

Table 1

issue	n statem	#word per statem	sd #word	#char per statem	#char per word
2010_1	518	31.30	14.69	182.41	5.83
2011_1	428	32.40	15.29	190.00	5.86
2012_1	295	32.97	16.27	191.99	5.82
2012_2	364	33.18	16.06	192.01	5.78
2013_1	288	32.21	15.56	187.26	5.81
2013_2	317	31.85	15.46	185.60	5.83
2014_1	271	31.52	15.10	181.26	5.75
2014_2	379	34.21	16.64	195.40	5.71
2015_1	266	34.32	14.98	195.94	5.71
2015_2	267	32.21	14.92	183.88	5.71
2016_1	297	32.87	14.94	187.57	5.71

The heatmap is another qualitative summary representation of the *TDM* matrix where the frequency of each word in each document is coded through the color intensity. In the following we provide a heatmap for all the 11 issues where we can see the more frequently used words in these 6 years of the *FSR* publication. A normalized version of the heatmap is shown in fig. 4.2.

Figure 4.1: wordmap for the whole corpus of the *FSR*

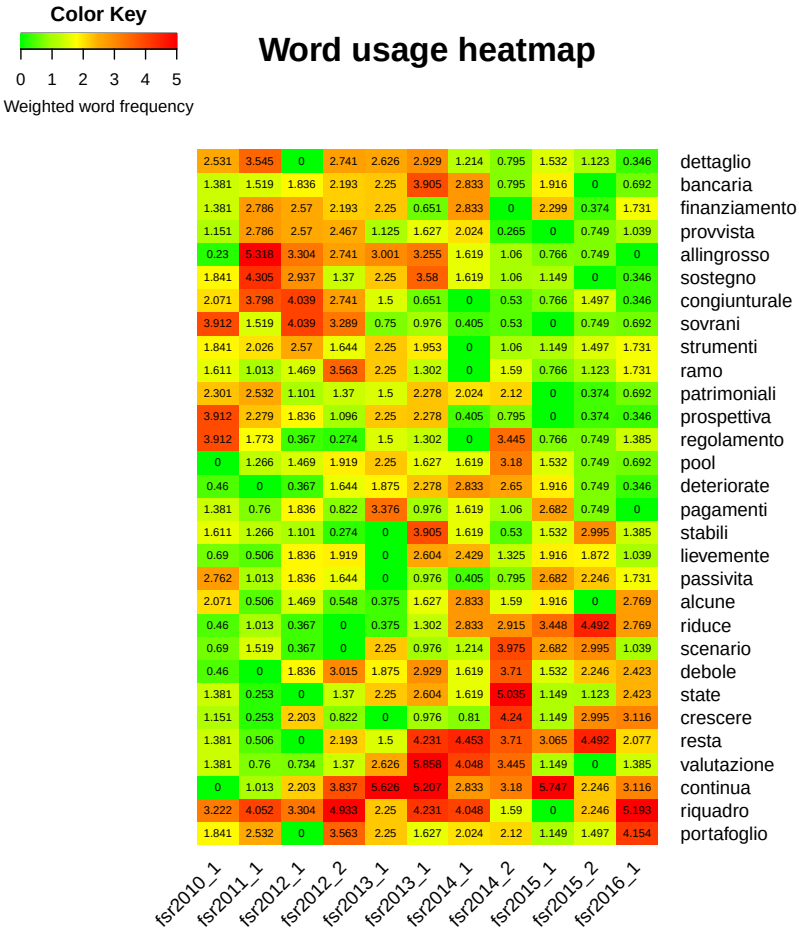


Figure 4.2: Heatmap

From this heatmap we can easily track down the evolution of the more frequently used words over the past six years. It can be seen that in 2011 the words *sostegno* and *congiunturale* [13] were hot topics. In the second half of 2013 we have again *supporto* along with *bancaria* and *valutazione*. In the last issue *portafoglio* [14] becomes a relevant topic. These words constitute a clue in representing some topics. Further investigation with *LSA* might confirm the true role of the frequency of these words in signalling interest in given semantic concepts.

[13] In Italian they are respectively support and short-term

[14] Portfolio in Italian.

5 LSA application with Financial stability reports

In order to check the actual behavior of the available text mining procedures we have taken into account the corpus composed of the 11 issues of the *FSR*. These documents are available in PDF© format at the Bank of Italy web site.

A relevant consideration to take into account here is the language of the corpus. As a matter of fact many computational tools are already quite developed for the English, German and Chinese languages while they are still at a rather infancy stage for the Italian.

The first *LSA* analysis run on the *FSR* corpus has consisted in deriving a coherence measure on each chapter of the different 11 issues. The coherence index is a measure contained in the range $(0 \div 1)$, values above 0.5 signal a semantic similarity among the sentences composing the chapters. The **R** software provides the package **LSAfun** with a *coherence* function computing both a local (statement to statement) and a global coherence. Here we present just the global coherence which is an average value among all the statements in a chapter. In fig 5.1 we show the the evolution of the coherence over the different 58 chapters and the coherence of a 10 statements automatic summary achieved by applying once again the *LSA* methodology as proposed in Gong and Liu [4].

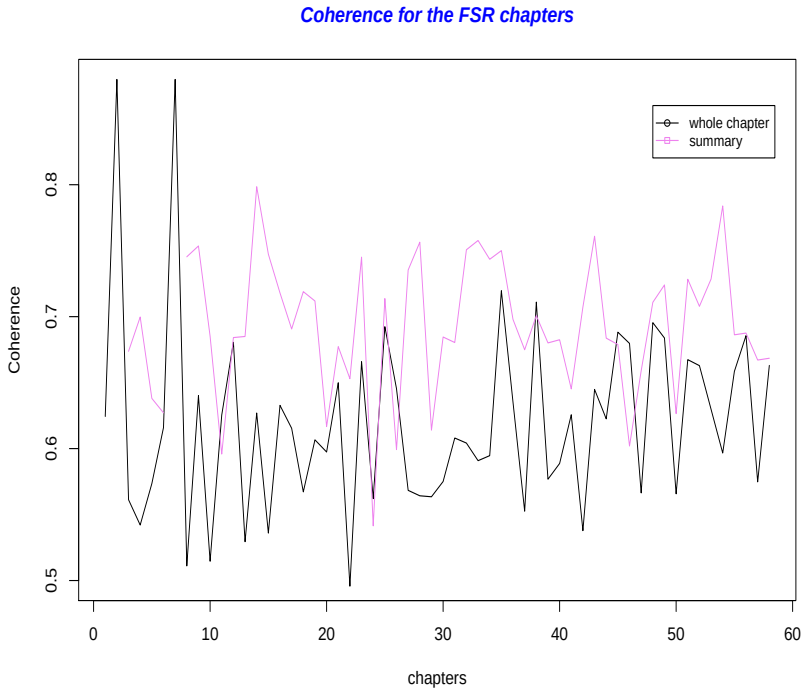


Figure 5.1: Coherence evolution for the *FSR*

In the table 2 we show the average coherence statistics for the whole set of chapters and for a summary of each issue of the *FSR*:

Table 2

variable	mean	std dev
global	0.80	0.030
summary	0.85	0.032

The first row of the table show the average coherence of the whole reports. The second row lists the average coherence of the automatic summary generated as a by product of the *LSA* methodology. These values provide an objective measure of the high coherence level shown by the chapters of the corpus.

As a second application of the *LSA* we have computed the nearest neighbors for the italian translation of the words crisis and stability[15]. For each given word a nearest neighbors is a word with semantically similar meaning. These neighbors are computed by ranking in decreasing order the similarity between the each couple of words composed of the reference and another word of the vocabulary. In the figure 5.2 we show the top five nearest neighbors to the word *crisi* using the Multidimensional Scaling on the similarity matrix of all the available couple of words. The concept of crisis emerges as strictly linked with debt, consolidation and countries. These words having semantically closeness with *crisi* could play the role of either causes or consequences.

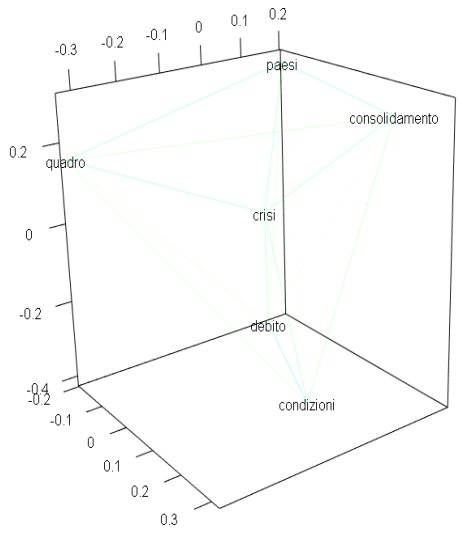


Figure 5.2: Top 5 neighbors for crisi

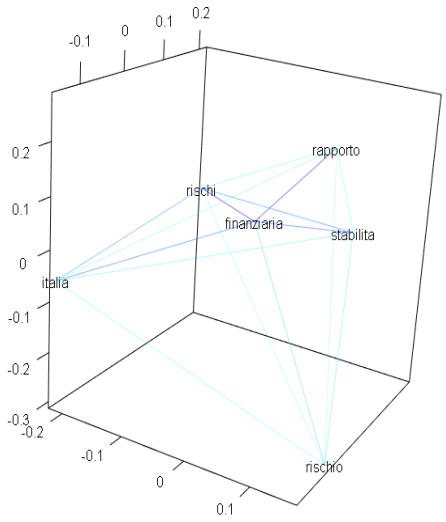


Figure 5.3: Top 5 neighbors for stabilità

The second example is presented in figure 5.3. Here we show the nearest neighbors to the word *stabilità*. This time the *stabilità* concept is obviously closely related with the word *finanziaria*[16] and with the words *rischio/rischi* and *Italia*. The close connection among some neighbor words might sporadically look mysterious, and sometimes words that should be close are not. One possible explanation for this phenomenon could come from a bias due to the too thinly sampled words (small sampling bias).

[15] In Italian they are respectively crisi and stabilità

[16] The two words stabilità and finanziaria constitute the title of the report

5.1 Gauging the Readability for the *FSR*

Evaluating in an automatic way the understandability of a text is a relevant factor for estimating its general public acceptance. Most of the classical readability metrics are linear model of few superficial features of words and sentences such as those shown in chapter 4. These readability indexes are generally given by a linear combination two proxies: a) the word difficulty measured by the number of letters per word, b) the sentence difficulty given by the number of words per sentence. In our empirical application we have chosen to employ the following functions available in the **qdap** R package. It is the following:

1. Automate Readability Index, $ARI = 4.71 \cdot \left(\frac{N_{char}}{N_{words}}\right) + .5 \cdot \left(\frac{N_{words}}{N_{sentences}}\right) - 21.43$
2. Fleisch Vacca test, $FV = 206.0 - 1.0 \left(\frac{N_{words}}{N_{sentences}}\right) - 0.65 \left(\frac{N_{syllables}}{N_{words}}\right)$

In the two formulae we have:

N_{char} is the character count for every word in the text,

N_{words} is the word count for every sentence and

$N_{sentences}$ is the total number of sentences in the text,

$N_{syllables}$ is the total number of syllables in the text.

These two indexes tend to reward the employment of short words and short sentences. ARI is a simple empirical derivation for the English language which provides a useful comparison tool over time and among different documents belonging to the same class. In our experiment we have used the test for a corpus written in Italian. In this case we extend to the Italian language the assumptions made about difficulty of words and sentences in English. The Fleisch-Vacca (FV) index is a modification of the Fleisch-Kincaid measure proposed for the Italian language[17]. This index approximate the readability ease and is generally comprised between 0 and 100. Values around 100 indicate a very simple reading while values below 30 are judged as texts requiring a degree for their understanding. In the pictures 5.4, 5.5, 5.6 and 5.7 we show the readability evolution for four of the past issues of the *FSR*.

In the pictures 5.8, 5.9, 5.10 and 5.11 we show the readability evolution computed with the Fleisch Vacca index for the same issues of the *FSR*.

Values of the ARI index in the pictures are around 20. In this case the measure is upward biased because the ARI Index was originally designed for the English language where words and sentences are on average shorter than their Italian translation by respectively 10% and 40%. Therefore this value represents an upper bound for the true readability. The readability values provided by the Fleisch-Vacca index are all between 30 and 35 meaning that a degree is required for their understanding. The results from the two measures appear slightly different: while the ARI index seems to indicate a readability at the college level degree, the FV measure judges the *FSR* as a text requiring a higher level of education. A possible avenue for further analysis could be the implementation of the GULPEASE index which was devised directly for the Italian language[18]. The final result can be interpreted as a general understandability of the *FSR* by people between an average to higher specialization.

5.2 Measuring the Formality for the *FSR*

Another relevant feature allowing to numerically gauge the degree of the context-dependence of a document is the formality of a document. There are some different definitions for the formality. Here we have chosen the definition suggested in Heylighen and Dewaele [7] where the Formality score is calculated according

[17] This test has been obtained modifying the coefficients provided in the Fleisch Kincaid test of the **qdap** package

[18] GULP is the Gruppo Universitario Linguistico Pedagogico (Linguistic Pedagogical University Group).

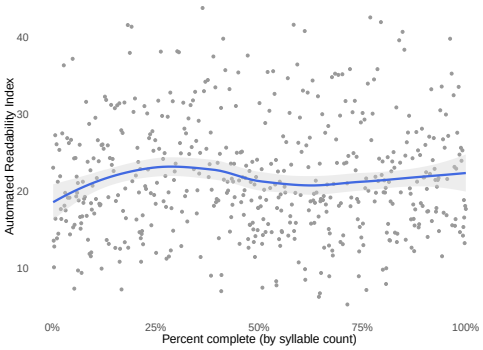


Figure 5.4: Readability of Financial Stability 2010

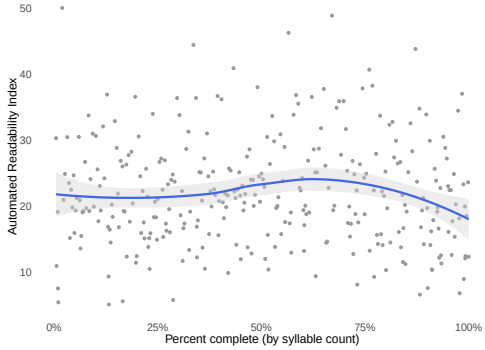


Figure 5.5: Readability of Financial Stability 2013-2

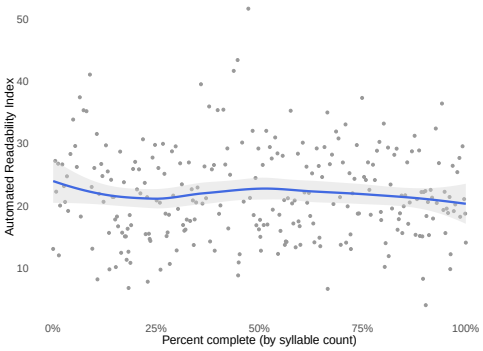


Figure 5.6: Readability of Financial Stability 2015-2

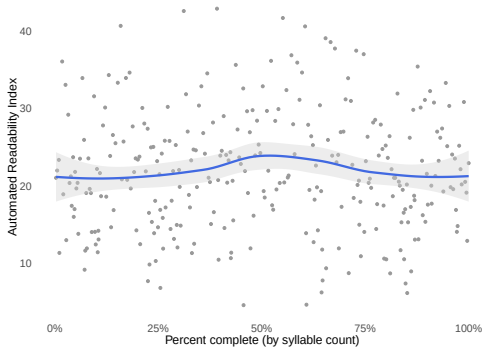


Figure 5.7: Readability of Financial Stability 2016-1

the equation:

$$F = 50 \cdot (\frac{n_f - n_c}{N} + 1) \tag{4}$$

where: $f = \{nouns, adjective, preposition, article\} \qquad n_f = |f|,$

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

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

This Formality score gets higher when statements make more use of nouns and adjectives rather than pronouns and adverbs.

In the picture 5.12, 5.13, 5.14 and 5.15 we show the formality evolution for four of the past issues of the *FSR*,

Reference values for the Formality measure taken from Heylighen and Dewaele [7] indicates values of 70 as highly formal. The last issues of the *FSR* feature a systematically higher values for the Formality index which is in the range $[75 \div 80]$. This result confirms our intuition of highly formal documents leaving few room to individual interpretation. The generally high formality value for the *FSR* signals a substantial absence of semantic ambiguity.

6 The web impact of the *FSR* through Google search

This section tries to answer the third question put forth in the introduction. It describes the methodology employed for evaluating the impact produced by the *FSR* on the web by counting the hits returning a web search suitably defined[19]. This technique, based on the measurement of the similarity of a pairs of phrases, has been already put forward and employed by different authors. Originally proposed by Turney [18] for the

[19]This impact is defined as semantic orientation in Lucca and Trebbi [11]

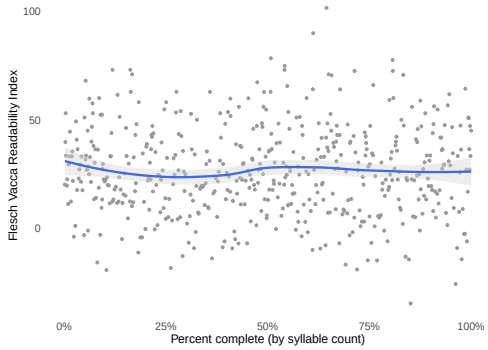


Figure 5.8: Fleisch-Vacca Readability of Financial Stability 2010

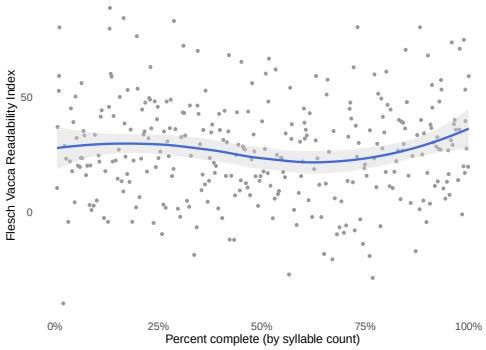


Figure 5.9: Fleisch-Vacca Readability of Financial Stability 2013-2

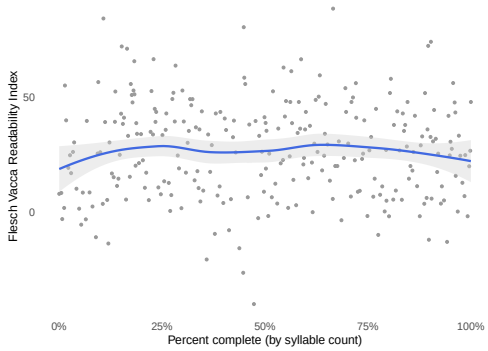


Figure 5.10: Fleisch-Vacca Readability of Financial Stability 2015-2

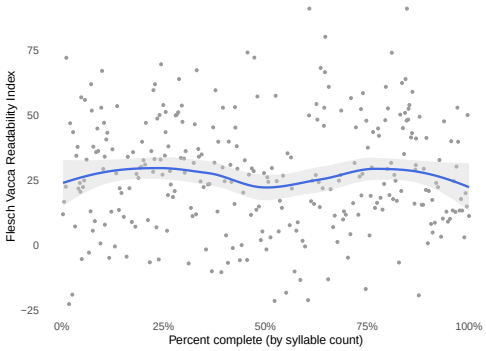


Figure 5.11: Fleisch-Vacca Readability of Financial Stability 2016-1

unsupervised classification of reviews. In the framework of Central Banks communications, to the best of our knowledge, the first paper making use Google hits count to gauge the web reaction about an economic issues is Lucca and Trebbi [11] which derive the relevant web score by applying an information-theoretic based tool. Similar technique are employed by Carvahlo et al. [2] for the evaluation of the information content of the interest rate setting statements of the Central Bank of Brazil and by Kawamura et al. [8] in analyzing the hypothesis of strategic disclosure by the Bank of Japan. The concept of orientation or value judgment towards a statement is based on the comparison of two distances. The first one is the distance between the considered phrase and a positively polarized word (e.g. "stability") and the second is the distance between the same phrase and a negatively polarized word ("instability"). When the phrase appears closer to the positively polarized word we assign a positive orientation to the phrase. On the other hand if the statement appears closer to the negatively polarized word we attribute a negative orientation to the statement. The distance between two words or sentences is assumed to be given by the Pointwise Mutual Information (PMI) which measures the likelihood that the first word/sentence will appear along with the second one. The formal definition of the PMI is given by

$$PMI(\phi_1, \phi_2) = \log \left[\frac{p(\phi_1, \phi_2)}{p(\phi_1) \cdot p(\phi_2)} \right] \tag{5}$$

It can take positive or negative values. It is zero when the two words/sentences are independent. Because of the practical unfeasibility to compute the probabilities considered in equation (5), we approximate them by evaluating the number of *Web search hits* in the web search of the statement of our document associated with the two opposing adjective (antonym). In their work regarding monetary policy, Lucca and Trebbi [11] provide the example of the antonymy "hawkish" versus "dovish". In practice, after having extracted the more sensible statements relating to monetary policy actions, the semantic orientation *SO* is evaluated by estimating the difference between the two *PMI* of the extracted statement and each term of the antonymy. In this paper we extend the Lucca and Trebbi [11] procedure by taking into account the whole set of statements contained in each issue of the *RSF*. In this way we attempt to average the polarity with respect to a given antynomy over all the statements.

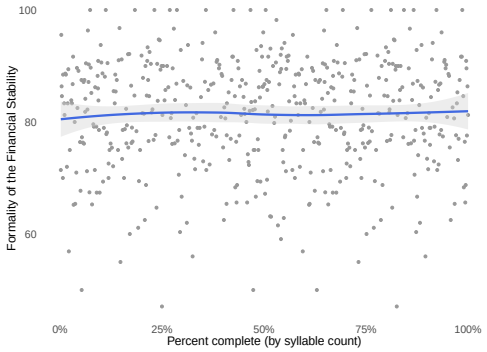


Figure 5.12: Formality of Financial Stability 2010

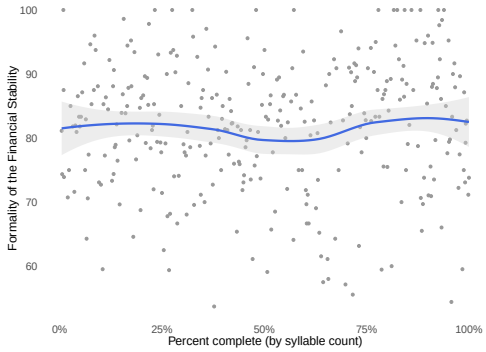


Figure 5.13: Formality of Financial Stability 2013-2

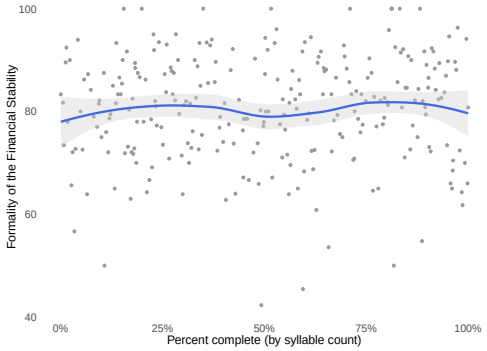


Figure 5.14: Formality of Financial Stability 2015-2

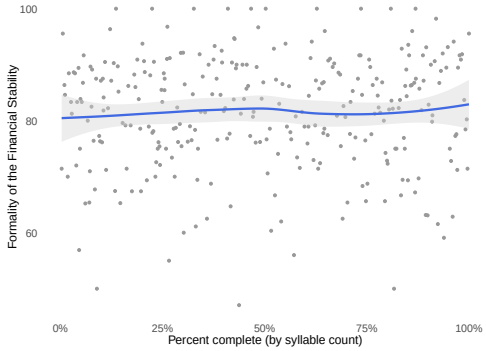


Figure 5.15: Formality of Financial Stability 2016-1

Formally, called $\mathbb{X}$ the statement under scrutiny [20], the *SO* of the statement will be estimated as:

$$SO(\mathbb{X}) = PMI(\mathbb{X}, hawkish) - PMI(\mathbb{X}, dovish) \tag{6}$$

While in the given references the web search has been carried out only on Google, in our investigation we went further on examining the stability of web search against three of the major web search engine: Bing, Google and Yahoo [21]. The details about code employed are in appendix. Here we have found that the number of web search hits are not very stable among the considered search engines.

In our empirical exercise we have considered the semantic orientation relative to each statement of the different *FSR* with respect to the following three antonyms:

1. stabilità/instabilità (stability/instability)
2. crisi/espansione (stability/instability)
3. vulnerabilità/solidità (stability/instability)

As an example we provide here the three pictures for the *PMI* referring to the three antonyms taken into account. In the figure 6.1 we show the *PMI* evolution for three different editions. The initial issue of 2010, the first issue of 2014 and finally the same variables are shown for the first issue of 2016.

[20] As an example, Lucca and Trebbi [11] consider the statement $\mathbb{X} \equiv$ "pressures on inflation have picked up in recent months".
[21] On the 25th of July 2016 Yahoo has been acquired by Verizon

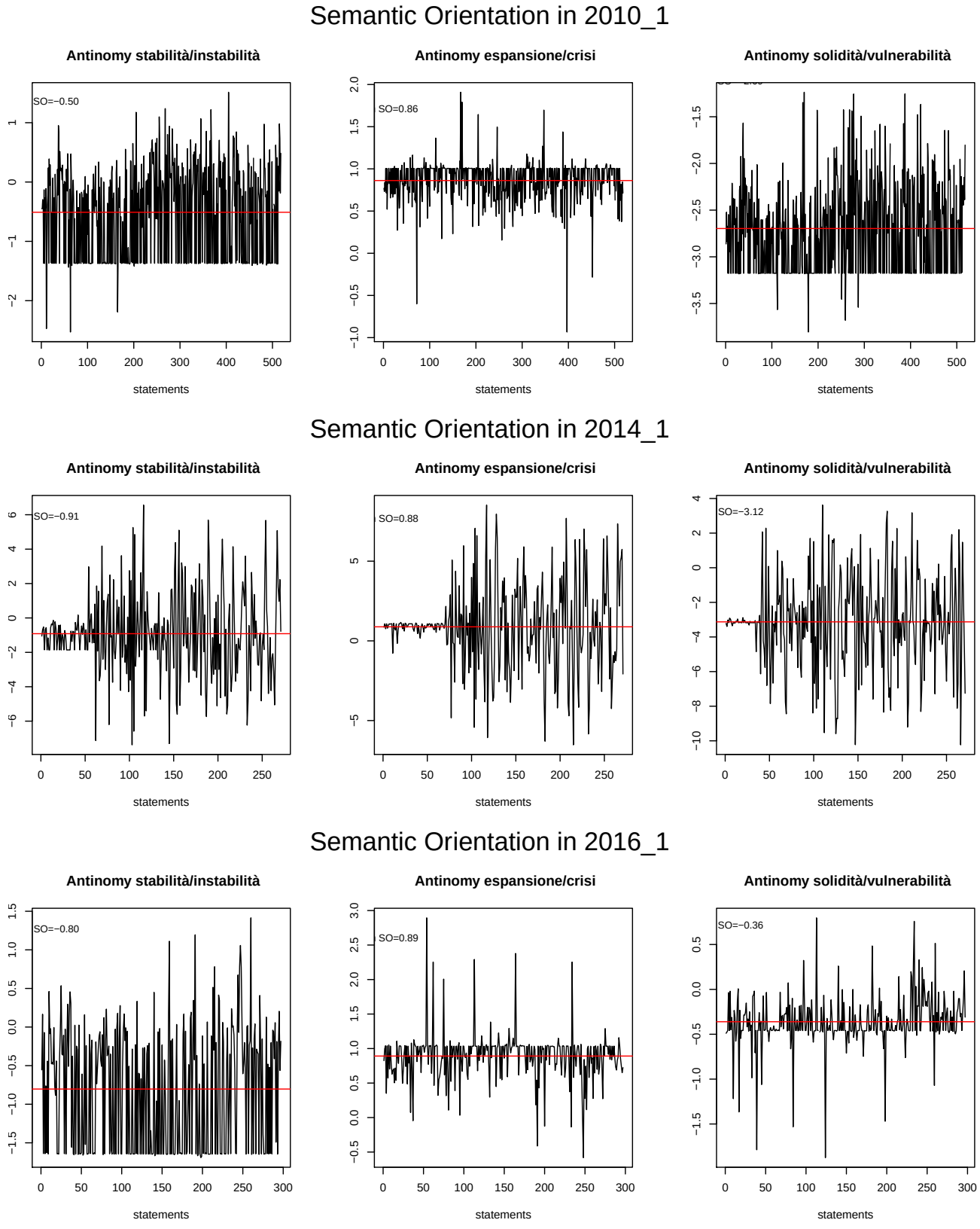

Semantic Orientation in 2010_1
Antinomy stabilità/instabilità
SO=−0.50
Antinomy espansione/crisi
SO=0.86
Antinomy solidità/vulnerabilità
statements
Semantic Orientation in 2014_1
Antinomy stabilità/instabilità
SO=−0.91
Antinomy espansione/crisi
SO=0.88
Antinomy solidità/vulnerabilità
SO=−3.12
Semantic Orientation in 2016_1
Antinomy stabilità/instabilità
SO=−0.80
Antinomy espansione/crisi
SO=0.89
Antinomy solidità/vulnerabilità
SO=−0.36

Figure 6.1

These three pictures can be read along two dimensions. By considering the three antonyms we see that in 2010 and in 2014 there is a strong feeling of vulnerability while in 2016 we see more neutral semantic orientation. By considering time evolution, we see a more confident/positive perception on all the three antonyms. Here we cannot take any definitive position but we believe these analyses could spawn many complementary tools improving the effectiveness of the standard communication means. The same web search has been tested on different web search engines. The results are similar between Bing and Yahoo, while they are significantly different with Google. We interpret this result with the difficulty of tracking down the internal behavior of the different search engines. It seems necessary to further our analysis by cooperating with the Companies providing these web query tools.

7 Concluding Remarks

In this paper we have presented the main computational linguistic methodologies for mining the relevant information from a corpus of documents and evaluating some summary statistics. These methodologies employ a tool-set taken from the IR technology.

We have shown that these technologies can be quite helpful in quickly analyzing huge amount of documents and automatically extracting sentiment or opinion orientation and gauging polarity of these sentiments. By taking advantage of some packages available on the CRAN[22] repository, we have written some *R* procedures implementing different algorithms for text mining and sentiment analysis. These *R* procedures have been applied for the analysis of an homogeneous corpus of documents based of the Bank of Italy *Financial stability report* which started in 2010.

The main conclusions are the following:

1) the *FSR* has shown a high level of local and global coherence;

2) the ARI and the FV readability indexes show a readability which implies approximately a college to higher degree for understanding the texts;

3) the examined corpus of documents, as it could be expected, has shown a quite high average level of the Formality score. This result witnesses the scarcity of room left to ambiguities.

The present strand of research looks quite promising especially for the possibility to quickly provide institutional answers more closely connected to the social emotions and preferences of the different economic agents.

References

[1] Boukus, E. and J. V. Rosenberg (2006). The information content of fomc minutes. *Federal Reserve Bank of New York* (NA), 1 – 53.

[2] Carvahlo, C., C. Cordeiro, and J. Vargas (2013). Just words? a quantitative analysis of the communication of the central bank of brazil. *Revista Breasileira de Economia 67*(4), 443 – 455.

[3] D. Bholat, S. Hansen, P. S. and C. Schonhardt-Bailey (2015). Text Mining for Central Banks. *Centre for Central Banking Studies 33*, 1–19.

[4] Gong, Y. and X. Liu (2015). Text mining for central banks. *Centre for Central Banking Studies- Bank of England 33*, 1 – 19.

[5] Hearst, M. A. (1999). Untangling Text Data Mining. *ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 3–10.

[22] Comprehensive R Archive Network, https://cran-r-project.org

[6] Hendry, S. and A. Madeley (2010). Text Mining and the information content of Bank of Canada Communications. *Bank of Canada Working Paper*, 2–53.

[7] Heylighen, F. and J. Dewaele (2002). Variation in the Contextuality of Language: an Empirical Measure. *Foundation of Science*, 293–340.

[8] Kawamura, K., Y. Kobashi, M. Shizume, and K. Ueda (2016). Strategic central bank communication: Discourse and game-theoretic analyses of the bank of japan's monthly report. *JSPS Working Paper Series* (80), 1 – 34.

[9] Landauer, T. K. and S. Dumais (1997). A solution to platos problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*(2), 211 – 240.

[10] Landauer, T. K., P. Foltz, and D. Laham (1998). Introduction to latent semantic analysis. *Discourse Processes 25*, 259 – 284.

[11] Lucca, D. O. and F. Trebbi (2011). Measuring central bank communication: an automated approach with applications to fomc statements. *NBER working paper* (15367), 1 – 37.

[12] Porter, M. F. (1980). An Algorithm for suffix stripping. *Program 14*(3), 130 – 137.

[13] Porter, M. F. (2006). Stemming Algorithms for various European languages.

[14] R. Nyman, P. O. and D. Tuckett (2015). Measuring financial sentiment to predict financial instability: A new approach based on text analysis. *University College London*.

[15] S. Deerwester, S. Dumais, G. F. and R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science 6*(41), 391 – 407.

[16] Salton, G. and M. J. McGill (1983). *Introduction to Modern Information retrieval*. New York: McGraw Hill Book Co.

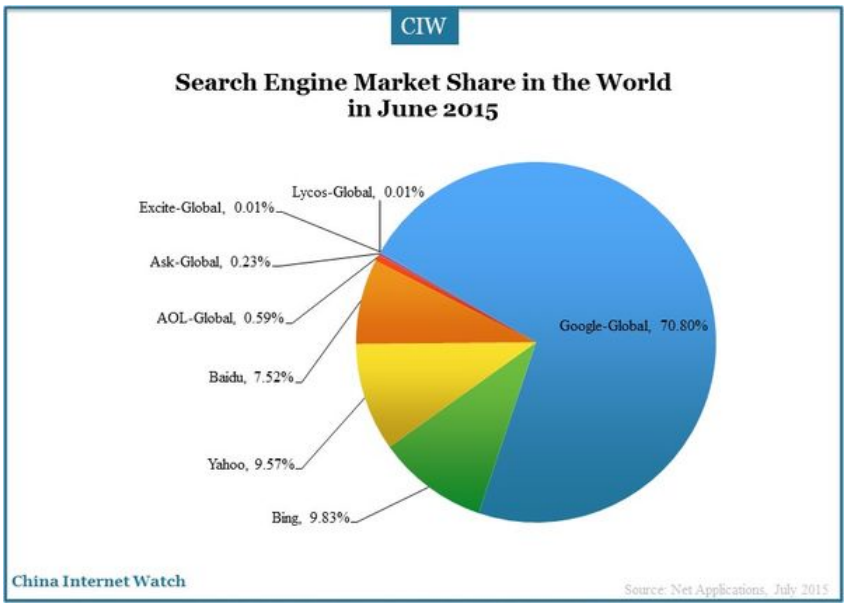
[17] Senter, R. J. and E. A. Smith (1967). Automated readability index. *Aerospace Medical Research Laboratories*.

[18] Turney, P. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of review. *Proceedings of the 40th Meeting of the Association for Computational Linguistics*, 417–424.

Appendices

A Market share distribution for Web search engines.

The following figure shows the market share for the main web search engine in 2015: from this picture we



realize that the search engines we consider cover over 90% of market share.

B Code snippet for computing number of web-hits.

B.1 Google web search

For computing the the web-hits by using the Google search engine we have used the **R** programming environment. the kernel code adopted is the following:

```
require("XML")
require("RCurl")
# Function to compute the number of hits on a given Google search
# Adapted from theBioBucket at http://goo.gl/TXvTxP
GoogleHits <- function(query){
        url <- paste0("https://www.google.com/search?q=", gsub(" ", "+", query))
        CAINFO = paste0(system.file(package="RCurl"), "/CurlSSL/ca-bundle.crt")
        script <- getURL(url, followlocation=T, cainfo=CAINFO)
        doc   <- htmlParse(script)
# Results look like this:
# <div class="sd" id="resultStats">About 10,300,000 results</div>
        res <- xpathSApply(doc, '//*/div[@id="resultStats"]', xmlValue)
        return(as.numeric(gsub("[^0-9]", "", res)))}
```

B.2 Bing and Yahoo web search

For Bing and Yahoo we have employed the Python language [23]. This choice has been taken simply for a lack of R packages for using these two search engines.

[23] Python was first released in 1991 by the dutch programmer Guido van Rossum.

```
# Bing Cognitive Search API
def bingcs(**kwargs):
    """
                Bing query language:  https://msdn.microsoft.com/en-us/library/ff795620.aspx
                Bing CS Search API:   https://msdn.microsoft.com/en-us/library/ff795657.aspx
    """
    KEY="XXXXXXXX"

    import requests
    url = 'https://api.cognitive.microsoft.com/bing/v5.0/search'
    payload = kwargs.copy()
    if 'fields' in payload.keys(): payload.pop('fields')
    headers = {'Ocp-Apim-Subscription-Key': KEY}
    r = requests.get(url, params=payload, headers=headers)
    j = r.json()
    if 'fields' in kwargs.keys():
        try:
            return _pathGet(j, kwargs['fields'])
        except KeyError:
            return 0
    else:
        return j

# Yahoo public search
def yahoo(**kwargs):
    """
    Documentazione parametri Yahoo: https://search.yahoo.com/web/advanced
    research suggestions:  https://help.yahoo.com/kb/search/improve-yahoo-search-results-sln2242.ht
    """
    import requests
    import string
    from bs4 import BeautifulSoup
    url = 'https://search.yahoo.com/search'
    payload = kwargs.copy()
    req = requests.get(url, params=payload)
    soup = BeautifulSoup(req.content, 'html.parser')
    try:
        text = soup.find("div", class_="compPagination").find("span").text
        return ''.join(ch for ch in text if ch not in string.punctuation).split()[0]
    except AttributeError:
        return 0
```

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Central Bank Communications: information extraction and semantic analysis[1]

Giuseppe Bruno,
Bank of Italy

[1] This presentation was prepared for the meeting. The views expressed are those of the author and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

Central Bank Communications: Information extraction and Semantic Analysis.

Giuseppe Bruno[1]

[1]Economics and statistics Directorate
Bank of Italy

IFC - Bank Indonesia Satellite Seminar on Big Data March 21st 2017

Outline

The Questions we are going to address

Extracting information from textual data

- The web provides an enormous warehouse of information. Around $4/5$ of this info is of textual nature.
- What is the impact of the Bank's communications? Can we devise an objective measurement mechanism?
- What is the measure of the sentiment caused by these communications?

[I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind. Lord Kelvin - 1883]

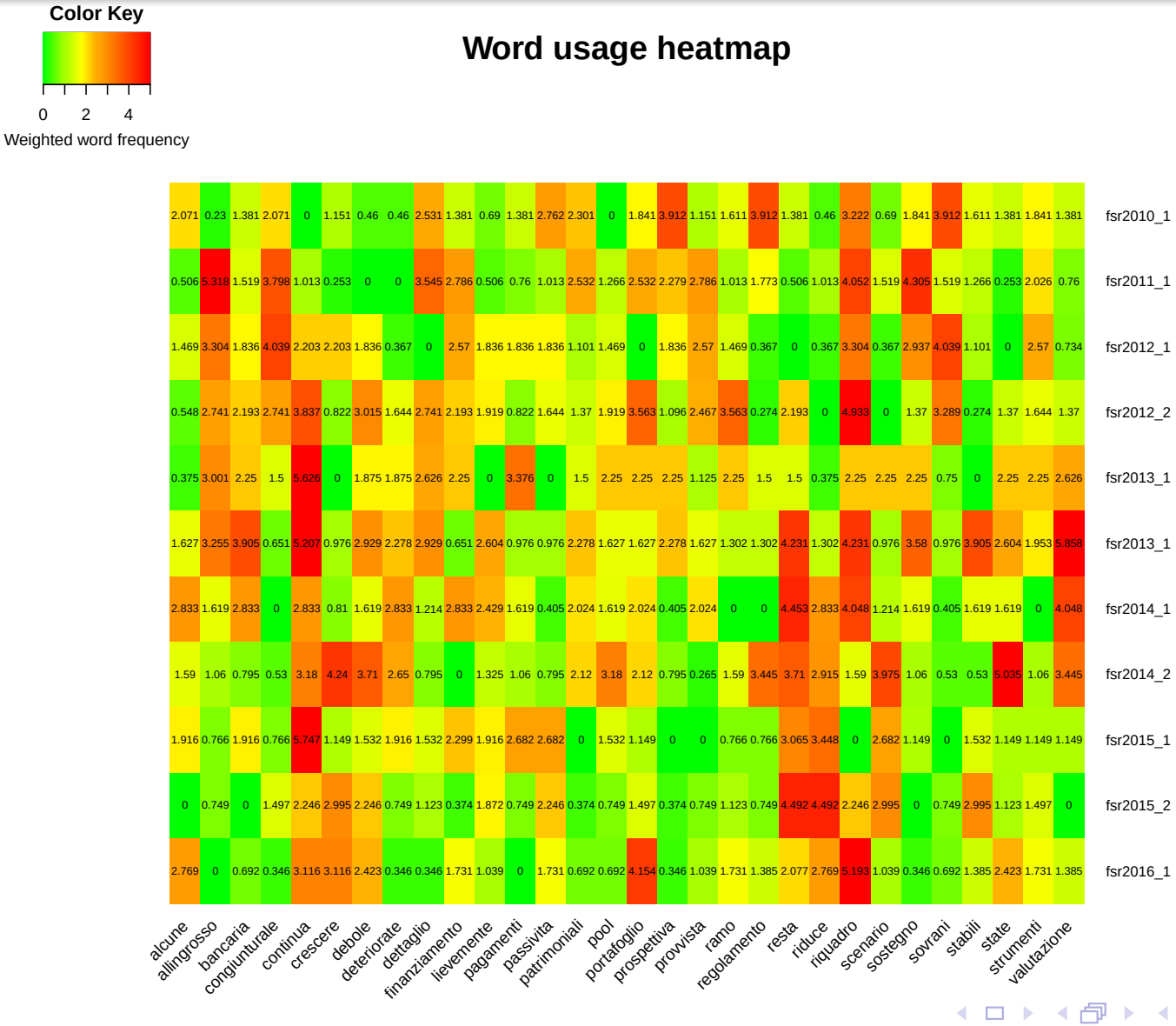
Statistics for the Financial stability report

Some linguistic statistics.

issue	#sentence	#word per sentence	sd #word	#char per sentence	#char per word
2010_1	518	31.30	14.69	182.41	5.83
2011_1	428	32.40	15.29	190.00	5.86
2012_1	295	32.97	16.27	191.99	5.82
2012_2	364	33.18	16.06	192.01	5.78
2013_1	288	32.21	15.56	187.26	5.81
2013_2	317	31.85	15.46	185.60	5.83
2014_1	271	31.52	15.10	181.26	5.75
2014_2	379	34.21	16.64	195.40	5.71
2015_1	266	34.32	14.98	195.94	5.71
2015_2	267	32.21	14.92	183.88	5.71
2016_1	297	32.87	14.94	187.57	5.71

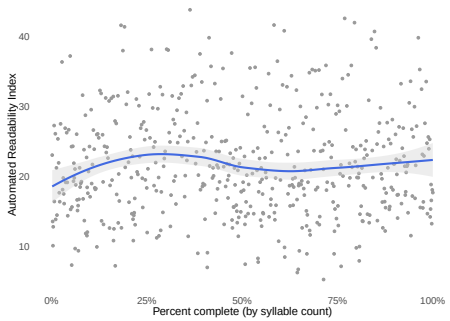
The Financial stability report appeared in 2010. It started as a yearly publication. In 2012 the report became biannual.

It has about 40 *pages, with* 50 *graphs,* 5 *tables and around* 10 *in-depth information boxes.*

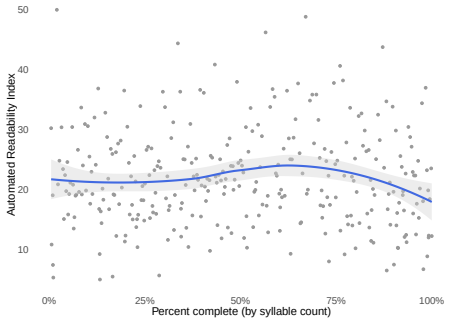

Color Key
0
2
4
Weighted word frequency
Word usage heatmap
fsr2010_1
fsr2011_1
fsr2012_1
fsr2012_2
fsr2013_1
fsr2013_1
fsr2014_1
fsr2014_2
fsr2015_1
fsr2015_2
fsr2016_1
alcune
allingrosso
bancaria
congiunturale
continua
crescere
debole
deteriorate
dettaglio
finanziamento
lievemente
pagamenti
passivita
patrimoniali
pool
portafoglio
prospettiva
provvista
ramo
regolamento
resta
riduce
riquadro
scenario
sostegno
sovrani
stabili
state
strumenti
valutazione

Readability

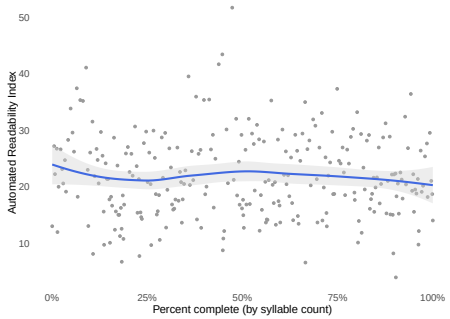
$$ARI = 4.71 \cdot \left(\frac{N_{char}}{N_{words}} \right) + .5 \cdot \left(\frac{N_{words}}{N_{sentences}} \right) - 21.43$$



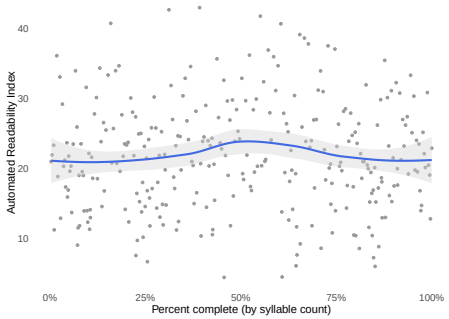
Readability FSR 2010



Readability FSR 2013-2



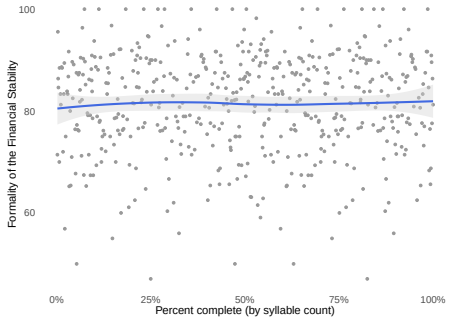
Readability FSR 2015-2



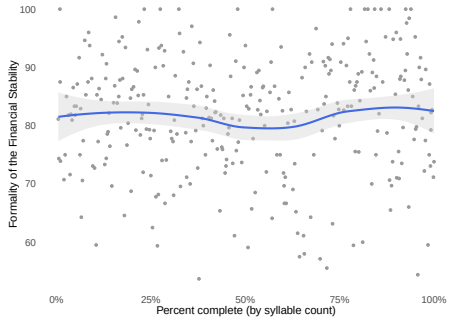
Readability FSR 2016-1

The Formality measure.

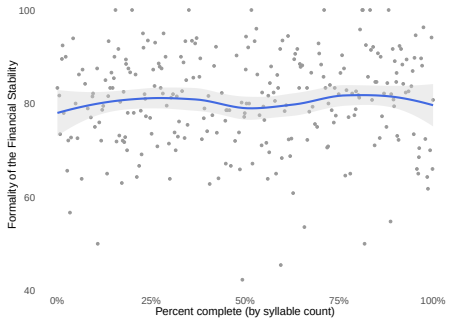
Formality of a statement is defined as the amount of expression that is immutable irrespective to changes of context.



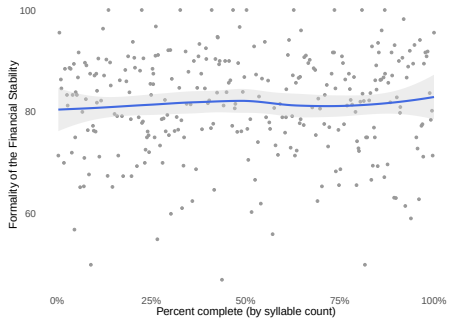
Formality of FSR 2010



Formality of FSR 2013-2



Formality of FSR 2015-2

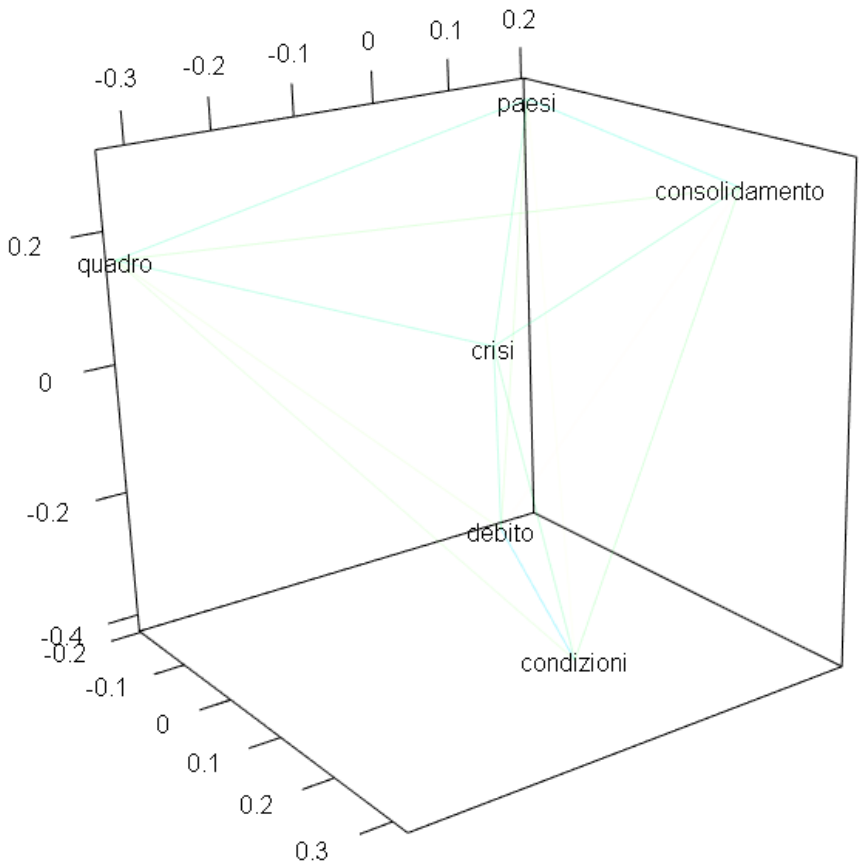


Formality of FS 2016-1

Measuring correlation between words and documents.

After completing the task of building a corpus of documents, it is possible to start the semantic analysis.
Latent Semantic Analysis (LSA) is a methodology for extracting and representing the contextual-usage of words (co-occurrence) for determining the similarity of meaning of sentences by analysis of large text corpora.
LSA methodology is well established and available in software such as Python, R and SAS.

LSA app: words most highly similar with 'crisi'



Semantic Orientation from PMI

Given two events x and y, we have:

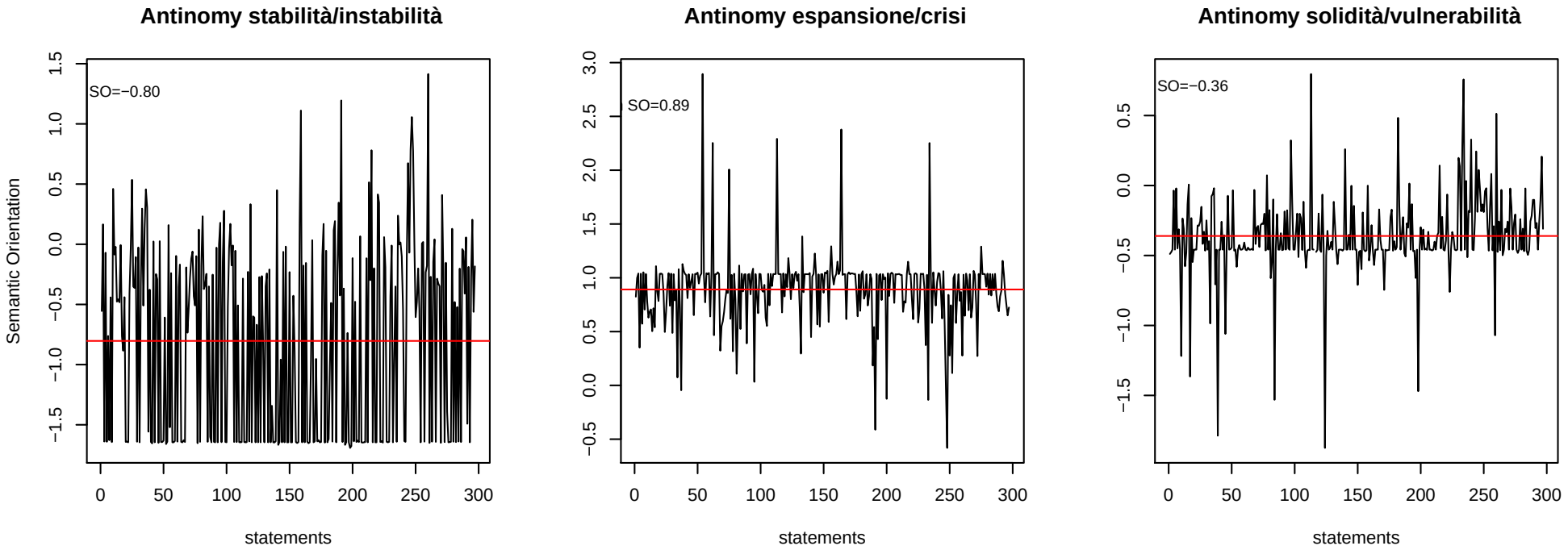
$$PMI(x;y) \equiv \log \frac{p(x,y)}{p(x) \cdot p(y)}$$

PMI measures the degree of statistical independence between x and y. The semantic orientation can be made more robust by employing an array of N antonyms:

$$SO(sent_i) \equiv \sum_{ant_j=1}^{N} \left(PMI(sent_i, ant_j[pos]) - PMI(sent_i, ant_j[neg]) \right)$$

Semantic Orientation in 2016

Semantic Orientation in 2016_1



Concluding Remarks

- We have built a Latent Semantic space to measure similarity among words and sentences in the*FSR*;
- we have evaluated some general characteristics of the *FSR* (readability and formality);
- we have extended the Semantic Orientation in Lucca(2011) by employing all the sentences as search units;
- we have shown a technique for evaluating the sentiment and polarity orientation caused by the text on the Web.

For Further Reading

F. Heylighen and J. Dewaele.

Variation on the Contextuality of Language: an Empirical Measure.

Foundation of Science, 2002.

R. Senter and E.A. Smith.

Automated Readability Index.

Aerospace Medical Research Laboratory, 2010.

D. Lucca and F. Trebbi.

Measuring Central Bank Communication: an Automated Approach with Applications to FOMC Statements.

NBER working paper, 2011.

Thank you for your attention.

Any questions?