SENTIMENT ANALYSIS OF THE SPANISH FINANCIAL STABILITY REPORT

2020

BANCO DE ESPAÑA
Eurosistema

Documentos de Trabajo
N.º 2011

Ángel Iván Moreno Bernal and Carlos González Pedraz

SENTIMENT ANALYSIS OF THE SPANISH FINANCIAL STABILITY REPORT

SENTIMENT ANALYSIS OF THE SPANISH FINANCIAL STABILITY REPORT [*]

Ángel Iván Moreno Bernal and Carlos González Pedraz

BANCO DE ESPAÑA

(*) The authors wish to thank José Manuel Marqués, Ricardo Gimeno, Carlos Conesa, Jesús Ibáñez, Adrian van Rixtel and Carlos Pérez for their comments and suggestions, and Alicia Aguilar, José Manuel Carbó, María Teresa González, Roberto Pascual, Jara Quintanero and José Luis Romero for their help in defining and compiling the dictionary. The views and analysis contained in this paper are the exclusive responsibility of the authors and, therefore, do not necessarily reflect those of the Banco de España or the Eurosystem.

Documentos de Trabajo. N.º 2011

2020

The Working Paper Series seeks to disseminate original research in economics and finance. All papers have been anonymously refereed. By publishing these papers, the Banco de España aims to contribute to economic analysis and, in particular, to knowledge of the Spanish economy and its international environment.

The opinions and analyses in the Working Paper Series are the responsibility of the authors and, therefore, do not necessarily coincide with those of the Banco de España or the Eurosystem.

The Banco de España disseminates its main reports and most of its publications via the Internet at the following website: http://www.bde.es.

ISSN: 1579-8666 (on line)

Abstract

This paper presents a text mining application, to extract information from financial texts and use this information to create sentiment indices. In particular, the analysis focuses on the Banco de España's Financial Stability Reports from 2002 to 2019 in their Spanish version and on the press reaction to these reports. To calculate the indices, a Spanish dictionary of words with a positive, negative or neutral connotation has been created, to the best of our knowledge the first within the context of financial stability. The robustness of the indices is analysed by applying them to different sections of the Report, and using different variations of the dictionary and the definition of the index. Finally, sentiment is also measured for press reports in the days following the publication of the Report. The results show that the list of words collected in the reference dictionary represents a robust sample to estimate the sentiment of these texts. This tool constitutes a valuable methodology to analyse the repercussion of financial stability reports, while objectively quantifying the sentiment conveyed in them.

Keywords: text mining, sentiment analysis, natural language processing, central bank communications, financial stability.

JEL classification: C82, G28.

Resumen

En este artículo se muestra una aplicación de la minería de textos para extraer información de documentos financieros y usar esta información para crear índices de sentimiento. En particular, el análisis se centra en los diferentes números del *Informe de Estabilidad Financiera* (IEF) del Banco de España desde 2002 hasta 2019 en su versión en español, y en la reacción de la prensa a este Informe. Para calcular los índices, se ha creado, hasta donde conocemos, el primer diccionario en español de palabras con connotación positiva, negativa o neutra dentro del contexto de la estabilidad financiera. Se analiza la robustez de los índices aplicándolos a distintas secciones del Informe, y usando diversas variaciones del diccionario y de la definición del índice. Finalmente, se mide también el sentimiento de las noticias de los periódicos los días siguientes a la publicación del Informe. Los resultados muestran que la lista de palabras recogida en el diccionario de referencia constituye una muestra robusta para estimar el sentimiento de estos textos. Esta herramienta constituye un valioso instrumento para analizar la repercusión del IEF, y también para cuantificar de forma objetiva el sentimiento que se está trasladando en él.

Palabras clave: minería de textos, análisis de sentimiento, procesado de lenguaje natural, comunicaciones de bancos centrales, estabilidad financiera.

Códigos JEL: C82, G28.

1. Introduction

Among the Banco de España's various publications, the *Informe de Estabilidad Financiera* (Financial Stability Report, hereafter FSR, or IEF by its Spanish abbreviation) stands out as an essential communication tool, not only regarding the risks to the Spanish financial system and to the profitability and solvency of Spanish deposit-taking institutions, but also regarding institutional macroprudential measures and policy. The Report, which has been published since 2002, initially constituted the first chapter of the *Revista de Estabilidad Financiera* (Financial Stability Review), becoming an independent publication in November 2004. Since the very beginning, there have been reports on the IEF in the specialist press, which has published citations of what were considered to be the most important phrases and extracts of the key messages, aiming to summarise its content within the natural space constraints of that medium.

In the realm of central bank publications, various studies have been conducted to analyse the content of different types of communications (Apel and Grimaldi (2012), Born et al. (2014), Correa et al. (2018), and Apergis and Pragidis (2019), among others). However, we have found no reference to analysis of publications in Spanish, nor any reference relating to the impact of the IEF in the written media.

Traditionally, content analysis was a social science discipline, with the main work of the researcher being to classify texts from a qualitative standpoint (Aureli (2017)). With the emergence of computers, initiatives arose to automate content analysis from a more quantitative standpoint, creating what would subsequently become known as text mining, a discipline separate from content analysis. One of the first works relating to sentiment analysis based on searches for words in dictionaries of categories using computers was subtitled "A Computer Approach to Content Analysis" (Stone et al. (1966)). The paper describes the General Inquirer programme, which processed a text and searched for each word, assigning it a category within a dictionary of categories (Stone et al. (1962)).[1] Since then, different approaches to this type of analysis have appeared and the escalation in computing capacity has made it possible to use machine learning techniques based on neural networks.

The different tools and technologies that come under the heading of text mining enable structured and quantitative information to be extracted from the unstructured data present in texts from a set of sources (e.g. reports, news reports, websites, blogs or social media posts). Within text mining, sentiment analysis is a document classification tool[2] that aims to determine the degree of polarity of a text between two extremes or poles, such as positive-negative or strong-weak. In the case of positive-negative polarity, the term “tone” is frequently used to refer to the sentiment of a text (Kearney and Liu (2014)). An index or metric of the documents analysed is thus obtained, which will have two polarities and represents the tone of the document.

Using text mining techniques, this paper proposes a structured and quantitative analysis of financial texts in Spanish, creating a sentiment index that measures the positive-negative polarity of these documents (that is, the degree of optimism or pessimism they reflect). The set of texts analysed comprises all the editions of the Spanish version of the Banco de España’s Financial Stability Report between 2002 and 2019. The exercise is supplemented by applying the same procedure to a set of press articles linked to these reports over the same period.

To calculate the sentiment indices, the paper defines the first Spanish dictionary of words with tonality (positive, negative or neutral) within the field of financial stability. As indicated below, the Spanish language has certain singularities that make using these text mining techniques more

[1] It was programmed with punched cards, initially to be used on IBM 709 series computers, and subsequently on the 7090/7094 series. At that time, the IBM 7090 cost around $3 million and weighed 8 tons (Weik (1961)).

[2] According to the Miner et al. (2012) classification, text mining is a field of knowledge that has seven spheres of activity: document classification, data extraction, natural language processing (NLP), concept extraction, web mining, information retrieval and document clustering. Sentiment analysis is considered to be a field within document classification, but it may also use techniques pertaining to web mining or NLP.

complicated. The paper also contributes to the methodology of sentiment dictionary creation in two ways: first, by incorporating analysis techniques of the level of agreement in the process of annotation or tonality assignment; and second, by developing a specific tool to facilitate this process. A comparative sentiment analysis of press articles relating to the IEF is made to assess the applicability both of the index and the reference dictionary.

As opposed to more complex techniques, the technique used in this paper to determine sentiment by means of searching for words in a tonality dictionary is conceptually intuitive and generally provides good results that are easy to interpret. In addition, it is less constrained by the size of the available corpus. The limited number of texts on financial stability hinders the use of machine learning methods, which require a considerable set of training samples. Accordingly, having an initial analysis such as that presented here may facilitate the work of creating the training sets required by such techniques and may serve as reference for future studies that seek to assess the use of machine learning-based techniques. The main difficulty involved in the technique adopted here lies in the need to have a dictionary of categories. In this case, a dictionary that classifies words into positive, negative and neutral, within the context of financial stability.

Among the works related to textual analysis and central bank communications in the sphere of financial stability, a notable example is Born et al. (2014), which conducts a sentiment analysis of the first chapter of the English-language financial stability reports of 37 countries between 1996 and 2009, using the DICTION 5.0 application.[3] Correa et al. (2018) analyse, for the period 2000-2017, the English-language stability reports of 64 countries, plus those published by the European Central Bank (ECB) and the International Monetary Fund (IMF), in this case based on a specific dictionary of tonalities created by the authors (Correa et al. (2017)), with 96 positive words and 295 negative ones.

[3] A commercial computer-assisted text analysis product (Digitext, Inc. (s.f.)).

Analyses using general dictionaries (such as DICTION) for financial texts offer less precise results than those using specifically tailored dictionaries, as shown both by Loughran and McDonald (2011) and Henry and Leone (2016). As three quarters of the negative words in the Harvard-IV-4 generic tonality dictionary did not have a negative meaning in the context of annual corporate reports, Loughran and McDonald (2011) created a dictionary specifically tailored to financial reports. Moreover, Correa et al. (2018) considered that the financial stability context was also sufficiently different from that of financial reports as to merit the use of a specific dictionary different from that of Loughran and McDonald (2011). In particular, financial stability texts use a series of words that do not necessarily have a negative connotation, but which would be classified by general sentiment dictionaries as negative. The English word "confined", for example, normally has a negative tone, but in financial stability the tone is positive as it is used in connection with limiting negative spillovers (Correa et al. (2018)). Just as there are no sentiment analysis studies on financial stability in Spanish, there are no public tonality dictionaries with terms tailored specifically to the financial context. A direct translation of English dictionaries, as in the case of the dictionary used by Correa et al. (2017), is not the best alternative, given that not all words, when translated, will retain their positive or negative connotation. For example, words that are polysemic in one language and not appropriate to be included in the dictionary may not be polysemic in another language. For this reason, we opted to construct our own specific Spanish tonality dictionary tailored to financial stability.

The next section presents the methodology used to process the texts and create the dictionary. It is followed by definitions of the rules for calculating the different indices proposed for the sentiment analysis. In the results section, the robustness of the sentiment index is analysed, the consistency of the dictionary is examined and the index for the different sections of the Report is calculated. Finally, the index for the press articles on the Report is calculated.

2. Creation of a Spanish dictionary tailored to financial stability

The dictionary was created by assigning tonalities to the words used in the Banco de España's IEF. The text sample analysed includes the various reports, in PDF format, published between 2002 and November 2019, a total of 35 reports.[4]

To analyse the words and sentences, the documents must be processed; Figure 1 depicts the complete process. To preserve the text flow and facilitate the extraction, the commercially available Nuance Power PDF Advanced tool was used, which converts PDF documents to Word format by extracting the text, including text contained in images. This initial conversion retains the original document structure, but in an editable format. As the aim is to analyse the overview and the body of the text separately, so as to capture any possible differences, and as the sample consists of just 35 reports, the relevant sections were selected manually and the textual content was stored in separate files, excluding footers, headers, contents page and title page. Further programmatic steps were taken in an endeavour to reduce the number of processing errors, including correcting hyphens that had been extracted incorrectly and joining up words that had been split with a hyphen at the end of a line. Lastly, the text was changed to lower case to facilitate the analysis. Traditionally, one of the problems associated with processing text in languages other than English is the existence of characters that do not exist in English; in the case of Spanish, the letter "ñ" and vowels with accents or diaeresis. Many programming tools and libraries are geared to the English-speaking market and have problems with these characters. To remedy these constraints, in some cases the accents and diaeresis are eliminated and the "ñ" is converted to "n" or even to "ny". Here, however, no such transformation was necessary as the tools used do not have this limitation.[5]

[4] All the reports may be downloaded in PDF format from the Banco de España's website (https://www.bde.es/bde/es/secciones/informes/boletines/Informe_de_Estab/).

[5] Python 3.7, which was the tool used to programme the process, stores strings of characters in Unicode and allows Spanish characters to be encoded without difficulty. Also, two specific text processing libraries – NLTK and spaCy – were used, neither of which has these limitations.

Once the text had been extracted, the Report was divided into two subsets, one containing the overview and the other the body of the text. The reason for this separation is that, in general, the drafting and discussion processes in the two sections tend to be different; in particular, the overview is generally subject to more intense revision. Once processed, the Reports average 27,554 words: 1,820 words for the overview and 25,734 words for the body of the text. On average the overview has 47 sentences and the body of the text 677. [6]

Before reviewing the words to be included in the dictionary, all the individual words were extracted and the "empty" words, i.e. the most common words that are of no value for the analysis, such as articles, pronouns and prepositions (e.g. "a", "an" "the", "them", "with", "under", "by", etc.), were eliminated. A total of 14,736 terms was obtained, corresponding to 6,111 roots. Words that differ only in gender, number or verb form may have different polarities. In consequence, in our analysis, tonality is defined not at the level of the lexeme or root, but at the level of the lemma or word, with the corresponding inflection. Spanish is a much more inflected language than English. Adjectives and nouns have morphemes of gender and number, whereas in English adjectives do not have morphemes of gender and only nouns have morphemes of number. Spanish verb forms are also more complex. This means that for each lexeme or root, the number of words or lemmas that can be formed, and hence the number whose tonality has to be reviewed, is higher. Certain words may have a positive or negative connotation in some forms and not in others. Following the recommendations of Loughran and McDonald (2011), the connotation of each word is defined deductively, that is, within the context in which it is used; in this case, the IEF.

To facilitate the review, a tool was created that analyses words individually, according to their connotation within the sentences in which they appear (see Figure 2). When a word is selected, the

[6] In text analysis, the basic processing element is the token, which may be any set of alphanumeric characters delimited by spaces or punctuation signs. In practice, it is equivalent to a word, and in this document "word" is used in the sense of token. Likewise, a sentence is understood to be the text contained between two full stops, or failing this a line break.

tool shows the sentences in which it appears in the different reports, and using his/her expert judgment the annotator decides whether the word has positive, negative or neutral tonality. The tool also allows words over which there may be doubt or disagreement to be flagged. In a first expert review, out of all the terms, 3,706 words were identified as susceptible of having tonality. These words were assigned a connotation (positive, negative or neutral) according to the context, and the first annotation or reference annotation was created. Subsequently, the list of words was divided into three groups and distributed between another six annotators, all of whom native Spanish speakers with university-level academic qualifications in economics. Thus, each word in the list of 3,706 words was reviewed by two of these six annotators.[7]

The aim of the annotation process was to create a *gold standard* or basic dictionary that would allow us to calculate sentiment for financial stability texts in Spanish with a high degree of reliability. In this case, drawing on the set of annotations, three phases were established to resolve disagreements and set the gold standard. First, using the tool described, the annotations were compared, one by one, with the other annotations of the same group (phase 1). Subsequently, the annotators of the different groups reviewed the words over which there was still disagreement, so that the opinion of all the annotators was obtained for each of these words (phase 2). Lastly, tonality was assigned by majority or *collective wisdom*, taking into account the initial reference annotation (phase 3).

Figure 2 depicts the different phases of the process. Note that the words on which there was no disagreement after phase 1 were not analysed by the other annotators. Accordingly, their tonality does not have the explicit support of the majority of annotators. However, the fact that they were subject to a double review and that the result coincided unknowingly with the reference annotation reduces the level of uncertainty and thus constitutes an acceptable trade-off between accuracy and

[7] The annotators had no previous experience in coding texts or specific linguistic training. They were given an initial training session in the tool and in basic guidelines on how to determine the polarity of words.

annotation effort. The resultant dictionary, after any disagreements had been settled by collective wisdom, contains 376 negative words and 189 positive words. All the others – the majority – have a neutral connotation. Figure 3 shows the words included in the resultant dictionary or gold standard in word clouds, differentiated by their tonality.

As in Correa et al. (2017), words were detected that had different connotations in financial stability compared with their connotations in general tonality dictionaries or even in financial dictionaries. For example, "*morosidad*" ("non-performance") could be understood to have a negative connotation, but in the IEF it is generally associated with "*tasa de morosidad*" ("non-performance ratio") which entails no sentiment per se. The same occurs with "*fallidas*" ("write-offs") or "*dudosas*" ("non-performing"), which are generally associated with the "number of loans" metric.

Regarding particularities of the Spanish language, we found cases of polysemic words that do not have the same polysemy in English. For example, the word "*bien*" ("good"), which could be considered positive, generally appears associated with the expression "*si bien*" ("however"), which in itself has no connotation. The word "*bien*" may also be associated with the concept of ownership, albeit more frequently in the plural (for example, "*bienes inmuebles*" ("immovable property")), and thus in the context of the IEF it has neutral polarity. By contrast, as tonalities are defined by lexeme, we found that on some occasions a word in a specific form may have a connotation in the final dictionary, while other forms of the same word may have a different connotation. For example, the word "absorbed" in English; in Spanish, the feminine form "*absorbida*" or plural "*absorbidas*", which is generally used in reference to absorbed loss(es), was assigned positive connotations, yet the masculine form "*absorbido*" or plural "*absorbidos*" is associated with a variety of concepts and thus has a neutral connotation.

3. Analysis of agreement between annotators

To assess the quality of the dictionary, in this section the level of agreement observed during the process of obtaining the gold standard is presented. The quality of an annotation process is generally measured in terms of accuracy and agreement. Accuracy refers to the level of compliance with specifications and agreement to the level of agreement between annotators. There is a correlation between these two measures, and a measure of agreement is generally considered to be indicative of annotation quality (Wong and Lee (2013)).

In an annotation process, agreement is measured in terms of agreement between annotators or *inter-judge reliability*. To measure the level of agreement in the creation of the basic dictionary, non-parametric Chance-Corrected Agreement Coefficient (CAC) statistics were used which adjust observed agreement by the probability of agreement expected by chance. In addition, as a measure of distance to weight the discrepancies, we use the Euclidean distance, i.e. the absolute difference between the annotated values (see Antoine et al. (2014) and Mielke et al. (2011)).[8]

Zhao et al. (2013) demonstrated the limits of the different CAC type statistics and recommended that different statistics be used on a case-by-case basis, finding no appropriate index for cases in which the annotators seek to be accurate but involuntarily make random assignments. For this reason, in our study three CAC type statistics were calculated in each phase of the annotation process: *Cohen's kappa* (κ) *coefficient*, *Krippendorff's alpha* (α) *coefficient* and *Gwet's* AC_2. These statistics are normally used in medical studies and in psychology, but they are also applicable to annotation processes such as this one (see Gwet (2008), Zhao et al. (2013) and Wongpakaran et al. (2013)). Table 1 presents the statistics for the different annotation phases. Appendix B offers some simplified examples of how the statistics were calculated. Note that Cohen's kappa (κ)

[8] In a categorisation with three levels of polarity (-1, 0 and 1), the binary distance (1 if they are different and 0 if they are equal) and the Euclidean distance coincide in most cases, as the differences between annotators are generally between neutral and either of the other polarities, but rarely between positive and negative.

coefficient only measures agreement between two appraisers; in consequence, there is one value for each combination of two annotators, whereas the α and AC_2 statistics allow for measurement of agreements between two or more annotators. For the calculation of these statistics, taking into account multiple annotators, all the words were considered. Those words over which there was no disagreement in phase 1 (i.e. those with three identical annotations, from the two annotators plus the initial reference annotation) were not assessed by the other annotators, for whom empty values were assigned, since calculation of the coefficients α and AC_2 permits empty values for some of the annotators.[9]

The most common interpretation of the values of κ and α is that agreement is high over 0.8 and moderate between 0.67 and 0.8 (see Krippendorff (2004) and Antoine et al. (2014)). In the case of AC_2, in the absence of a recommended interpretation owing to its newness, the same criterion is usually used as for the other CAC type coefficients. The κ values found in our process are in the interval [0.169, 0.794]. After completing the annotation process, and taking into account the seven annotations, a value of α was obtained equal to 0.454 and an AC_2 agreement coefficient equal to 0.905. In general, most words are assigned to the neutral category, which has a significantly higher number of words than both the positive and the negative category. In the literature on annotations relating to sentiment analysis, the values of the coefficients κ and α (see Antoine et al. (2014)) are usually lower than might be expected. This is known as the first kappa paradox, whereby symmetrical distributions tend to have higher *kappas* than distributions in which one category is prevalent (see Cicchetti and Feinstein (1990) and Callejas and López-Cózar (2008)). In such cases it is advisable to use alternative statistics, such as the AC_2 coefficient which is less sensitive to asymmetrical distributions, i.e. cases in which one of the categories has prevalence over the others (Gwet (2008)). That said, given that Zhao et al. (2013) consider that AC_2 tends to give high results, it is important in our view to include the κ and α values as an additional reference.

[9] All the statistics were calculated using the irrCAC library (Gwet (2019)) for the R programming language.

These differences between the coefficients analysed also reflect the difficulties that arise when establishing a consensus on the polarity of certain words. Specifically, for group A, the coefficients κ and α were especially low, probably owing to annotator bias, in this case caused by the tendency of one of the annotators to polarise annotation for contexts in which the other annotators opted for a neutral classification. The case-by-case comparative review process resulted in higher coefficients, which indicates that this is an effective and relatively inexpensive way to enhance agreement between two annotators. As indicated earlier, in the case of the words that were assessed by all seven annotators, we opted to use *collective wisdom* (i.e. a simple majority) to settle differences, thus minimising the possible annotator bias effect.

Most of the discrepancies observed were between neutral and positive or neutral and negative annotations. Figure 4 illustrates the distance distribution in the annotations by number of words compared with the first reference annotation (R) and compared with the other annotators (i). A distance of 2 indicates a discrepancy between positive and negative polarities, while a distance of 1 indicates a discrepancy between positive or negative polarity and neutral. A distance of 0 represents agreement between annotators. Thus, $A(i) - A(R)$ represents the distribution of differences in assignment of polarity of the words in group A between annotator i and the initial reference annotation R. It is observed that the $i = 2$ group A annotator applied a more flexible criterion when assigning polarity. In any event, there are virtually no discrepancies with distance equal to 2, which also reduces the impact of possible annotator bias.

4. Calculation of the sentiment index for the *Informe de Estabilidad Financiera* (IEF)

Once the dictionary had been created and the level of agreement in its creation analysed, the next step was to analyse the sentiment of a specific financial stability text in Spanish. In this case, the sentiment of the text will be measured as a function of the number of words with tonality appearing in it. For this purpose, first we needed to identify the words in the text that appear in the gold standard dictionary, which determines which words have tonality; second, to define a set of rules for counting the number of words within each category; and lastly, to determine a function that will take us from the number of words categorised to a measure of sentiment or sentiment index. For example, a measure of sentiment would consist in adding together all the words in the text that have a determined (positive or negative) polarity.

Attempts may be made to create rules that seek to identify the sentiment of a sentence considering the multiple possibilities that the grammatical constructions of a language allow. In general, these rules would be very complex and would require a lot of effort to implement in full. In this analysis, the set of rules was simplified as far as possible. Accordingly, the calculation rule adopted was to directly count all the words in the text that were identified as having a given tonality, but taking into account whether the words were negated in the sentence. Thus, if a word was negated, if it had positive tonality it was considered to have a negative connotation, and if it had negative tonality (a double negative) it was considered neutral. This rule reflects the idea that something that is “not good” has a negative tonality, but something that is “not bad” does not necessarily have a positive tonality (see Correa et al. (2018) and Loughran and McDonald (2011)). To determine whether or not a word is negated, a search was made for any of the following words placed no more than three words before the word being analysed: “*menos*” (“less”), “*no*” (“not”), “*nunca*” (“never”), “*sin*” (“without”), “*pérdida*” (“loss”) and “*disminución*” (“decline”). In the count of the number of words with tonality, the presence of intensifiers such as “*muy*” (“very”), “*gran*” (“large”), “*mucho*” (“much”) or “*enormemente*” (“enormously”), which could modify a word’s sentiment charge, was not taken into consideration; rather, all words with polarity were assigned the same weight. Although the rule used to count words is a simple one, it is important to note that during the annotation process tonality was assigned considering the context, so the sentence structure was already taken into account implicitly in the dictionary creation phase.[10]

Once the rule for identifying and counting the words with connotation within a text had been applied, the next step was to define the functions for measuring the overall tonality or sentiment of the text. Thus, for each Report in the sample, identified by its publication date *t*, *Negativity* was defined as the number of negative words over the total words in the text:

$$Negativity_t = \left(\frac{\#\, negative\ words}{\#\, total\ words}\right)_t; \qquad [1]$$

and *Positivity* as the number of positive words over the total words:

$$Positivity_t = \left(\frac{\#\, positive\ words}{\#\, total\ words}\right)_t. \qquad [2]$$

Lastly, *Net Negativity* was defined as the difference between *Negativity* and *Positivity* for the date *t*:

$$Net\ negativity_t = Negativity_t - Positivity_t = \left(\frac{\#\, negative\ words - \#\, positive\ words}{\#\, total\ words}\right)_t \qquad [3]$$

For example, to calculate their sentiment indices, Feldman et al. (2010) and Correa et al. (2018) use some of the measures in equations [1]-[3]. Other content analyses (e.g. Apel and Grimaldi (2012) and Apergis and Pragidis (2019)) define relative measures, that is, instead of using the total number of words (with and without connotation), they measure the gap between negative and positive words compared with the total words with tonality (that is, positive plus negative words), thus obtaining a measure of *Relative Net Negativity*, on a scale of -1 to 1. In our analysis we opted to use this relative measure to define the Financial Stability Report Sentiment Index (*FSSI*)*;* thus, for date t:

$$FSSI_t \equiv Relative\ net\ negativity_t = \left(\frac{\#\, negative\ words - \#\, positive\ words}{\#\, negative\ words + \#\, positive\ words}\right)_t \qquad [4]$$

[10] For example, in the case of the term "*adverso*" ("adverse"), which has negative tonality in the final dictionary, in order to determine that it has mostly negative tonality the annotators will have considered all the cases in which the term is used, taking into account the cases in which the word is used with neutral tonality (for example, in the bigram "adverse scenario", in reference to the bank stress tests).

Note that [3] and [4] are increasing functions as negativity increases; in other words, the higher the value of the $FSSI_t$, the less optimistic the sentiment conveyed. This coincides with the sign criterion used by Loughran and McDonald (2011) and by Correa et al. (2018). In general, the results using either of these two measures should be comparable, since normally the document length is proportional to the total sum of negative and positive words. Any possible differences found between measures [3] and [4] may be on account of the neutral words appearing in more academic sections of the documents and which could, in any event, be ruled out. In their analysis, despite using the number of total words in the denominator, Correa et al. (2018) rule out the sections they consider more theoretical and not related to financial stability, thus reducing the number of words.

5. Results for the overview and the body of the Report

The calculation of the different measures shown in equations [1] to [4] allows the analysis from different standpoints of the variations in tone in the Report over time. Figure 5 presents the results of calculating these measures for the overviews (introductory text), using the reference dictionary and the rules described. To help place the Report and the value of the associated index in its macro-financial context, the charts depict the key economic events (nationally and internationally) occurring during the period analysed (2002-2019). It can be observed that the key events are not distributed evenly across the series, but are concentrated from the onset of the global financial crisis, i.e. from 2008-2009 onwards. For the period 2003-2007, the lack of major events coincides with an expansionary economic phase. The events depicted may, in turn, be grouped into various subsets: those relating to the global financial crisis and the European sovereign debt crisis; those more specific to the Spanish financial system and the Spanish economic situation; and, in recent years, those related to geopolitical events. To complete the analysis, the terms of office of the different Governors of the Banco de España are also flagged.[11] Although this is not the object of the analysis, changes in Governors could potentially affect the tone of the texts (for example, changes on the editorial committee, or on the Executive Commission that approves the Reports).

Chart (a) in Figure 5 shows the Negativity, Positivity and Net negativity (equations [1], [2] and [3], respectively) for the overview section of the IEF from autumn 2002 to autumn 2019. In general, when *Positivity* increases *Negativity* declines, and vice versa, so there is a negative correlation between these indices, while there is a positive correlation between *Negativity* and *Net negativity* (see Table 2, with correlation coefficients between indices). The main advantage of the measures of *Positivity* and *Negativity* (equations [1] and [2]) is that they facilitate the identification of situations in which *Net negativity* and *Negativity* diverge, owing to the effect of *Positivity*; that is, those situations in which the proportion of positive words is significantly different from one period to another, inverting the trend of the *Net negativity* index that takes it into account [3]. One example of an apparently contradictory situation is autumn 2006, when it is observed that, compared with the Spring 2006 Report, *Negativity* (red line) increases at the same time as *Positivity* (green line), resulting in a slight drop in *Net negativity* (blue line). The net index declines owing to the increase in positive words, but at the same time as there is an increase in the number of words with negative tonality.

In the case of the overview, it is observed that *Net negativity* (blue line in Chart (a) in Figure 5) and the *FSSI* (yellow line in Chart (b)) are very similar and highly correlated (over 90%) (see Table 2). The peak for the *FSSI* corresponds to the Autumn 2011 Report, just months before the second bail-out programme for Greece. This Report has hardly any words with positive tonality. By contrast, one of the lowest figures in the series (i.e. one of the figures with the most positive sentiment) corresponds to the Spring 2015 Report, following the launch of the "Draghi Plan".[12]

[11] J. Caruana (July 2000-July 2006), M. Á. Fernández Ordóñez (July 2006-June 2012), L. M. Linde (June 2012-June 2018) and P. Hernández de Cos (June 2018-to date).

[12] This refers to the ECB's quantitative easing monetary policy, launched in 2015 by means of a plan for purchase of commercial banks' assets.

In addition to facilitating analysis of possible discrepancies between Reports, the indices also help to analyse discrepancies within Reports. In our case, as the overviews have been separated from the body of the text, possible differences may be analysed between these two sections of the Reports. The overviews serve as an executive summary and generally present the key points of the report. The body of the text refers not only to the economic situation and the situation of the banking sector, but also to other aspects that may be of academic or general interest.

Chart (b) in Figure 5 shows the sentiment index or *FSSI*, defined as *Relative net negativity* (equation [4]) calculated both for the overview and the body of the Reports in the sample. In addition, the first row of Table 3 shows the correlation coefficients between the different indices calculated, using the texts of the overview and the body of the Reports. In particular, Chart (b) in Figure 5 shows a high positive correlation (0.90) between the *FSSI* of the body of the text and the overview, as is to be expected, considering that the overview constitutes a summary of the rest of the Report. These results confirm the reliability of both the index and the dictionary used. The chart also reflects a higher degree of variability for the index calculated using the overview than for that calculated using the body of the text. The overview, being a shorter text, has fewer words with connotation, and the index is more sensitive to the presence or absence of any of these words. It is also observed that during certain periods (especially between 2003 and 2007), the overview and the body of the Reports presented different *FSSI* levels. However, owing to possible measurement errors, the level of the two indices is less informative than how they evolve. Despite the high correlation between them, there are points at which the direction of the two indices diverges. For example, in the second half of 2017 the index for the overview showed increased optimism (i.e. the index fell) compared with the previous Report, whereas the index for the rest of the Report showed increased pessimism (i.e. the index rose). Specifically, the overview of the Autumn 2017 Report starts as follows: “The latent geopolitical tensions have not prevented the economic recovery, at the international level, from holding on its positive path”. It then continues, with more information on the positive path. By contrast, the body of the Report highlights in several places the political tensions in Catalonia and even has a box on the subject.

These findings show how useful it is to obtain a quantitative variable to summarise the information in a set of texts. The different measures shown allowed us to analyse the changes and development of sentiment in the Report, highlighting the most extreme values or the changes in tendency. The change in sentiment shown is consistent with the key economic events that occurred across the sample. In addition, it was possible to compare the sentiment of different sections of each Report (specifically, the overview and the rest of the document). The findings shown are consistent for both sections; the index enables the focus to be placed on the parts of the Report that present the most striking differences.

6. Robustness analysis: completeness of the dictionary

We conducted two analyses to assess the robustness of the index. First, in a manner similar to Correa et al. (2017), we determined the *FSSI*'s range of variation for the body of the text by randomly removing 5% of the words from the dictionary in 1,000 iterations. Chart (a) in Figure 6 shows the confidence bands for the index obtained through this process of removing words and recalculating the index. In addition, the slope at each point was calculated, to determine not only the range but also the distribution of the *FSSI*'s upward or downward path with the different dictionaries. Accordingly, positive slope values indicated an upward trend and negative values a downward trend. Chart (a) also illustrates the slope distribution at each point, represented using box plot and whisker diagrams. At numerous points, the sample stands fully on one side or the other of the axis, indicating that the positive or negative slope is maintained in all 1,000 dictionaries. Calculating the slopes also helps identify the points at which the sentiment of the text has changed significantly relative to previous texts.

Second, capitalising on the information obtained in the annotation process, a further 1,000 iterations were performed, but this time assigning each word in the dictionary a tone probability that is proportional to the number of annotators who attributed that particular tone to the word. Accordingly, a word classified as positive by six annotators and as neutral by one annotator is assigned a probability of 0.14 of being neutral and of 0.86 of being positive. The results are set out in Chart (b) in Figure 6. The bands are wider in this case, although the value slope ranges generally remain on one side or the other of the axis when the slopes are steep. The colours represent the usual 0-25, 25-50, 50-75 and 75-100 percentiles. The advantage of this approach is that the information obtained in the annotation process can be harnessed without reducing it to a final list of words classified at three levels. According to Wong and Lee (2013), disagreement can be caused by the ambiguity inherent in a classification, and the very fact that the classification is ambiguous is important information in itself. Attempting to force an artificial agreement in such cases may not be the best option. Therefore, we view the probabilities-based approach as an appealing alternative to preserve legitimate disagreement. One important example is the word "crisis". In the dictionary of Correa et al. (2017), crisis has a neutral tone. However, in our dictionary it was classified as neutral by just two annotators, while five annotators considered it negative. This discrepancy may arise from the fact that the word is frequently used as a historical reference to the global financial crisis, although ultimately most annotators did consider that it generally attached a negative tone to the text. Assigning a definitive negative classification to the word would mean losing the fact that two annotators were legitimately in disagreement and regarded the negative tone assignment as unwarranted. A more representative picture of the sentiment perceived by the annotators is obtained by showing confidence bands based on the number of annotators who opted for each classification.

7. Comparison with financial indicators

Having a quantitative measurement of sentiment enables us not only to assess the internal consistency and robustness of the dictionary, but also to compare the index against quantitative financial indicators, to observe the consistency of the sentiment index and to identify how much data it includes on financial stability-related indicators (Correa et al. (2018)). The low number of observations means there are certain limitations to this analysis and the results are more dependent on specific events and small variations in the sample. It should be noted that the objective here is not to measure the ability of the sentiment index to make predictions in relation to the financial indicators.

In particular, the following regressions are analysed to identify how the financial indicators (quantitative variables) explain the time variation of the sentiment index:

$$FSSI_t = \mu_i + \beta_i X_{i,t-h} + \varepsilon_{i,t} \, , \tag{5}$$

where $X_{i,t-h}$ represents the economic variable i analysed in each case with a lag equal to h periods. The results for the coefficients β_i and their standard errors are shown in Table 4 for the contemporaneous regression ($h = 0$) and for the regression with the indicator lagged by two half-year periods ($h = 2$). The latter allows us to see whether the quantitative financial information is included consistently in the text of the Reports, measured using the sentiment index. The variables $X_{i,t}$ considered are a representative sample of the type of variables – relating to the financial stability cycle – that are monitored in the IEF. Specifically, in line with the indicators used in Correa et al. (2018), the following types of variables were considered: macroeconomic and monetary policy variables, such as quarterly GDP change for Spain, the unemployment rate, the short-term interbank interest rate and the short-term notional interest rate; variables relating to the credit cycle, such as the credit-to-GDP gap and the private non-financial debt service ratio; valuation indicators, such as the dividend yield of the IBEX-35 stock market index, the market-to-book ratio for banks, and changes in real housing prices in Spain; and lastly, financial risk variables, such as the credit risk premium in the Spanish banking sector, currency volatility and IBEX-35 volatility.

The results of the regressions are calculated both for the contemporaneous variables and for the two-period (one-year) lagged variables. Although the reliability of the estimators is influenced to some extent by the small size of the sample (36 observations), the results indicate that the sentiment indices are contemporaneously correlated ($h = 0$) with several of the financial stability cycle indicators analysed. In particular, the coefficients are significant for those variables that measure risk, such as the volatility of the IBEX-35 and the banking sector risk premium (represented by the CDS spread of one of its most representative banks). Accordingly, an increase in volatility and the credit risk premium comes with a rise in the sentiment index, i.e. greater pessimism measured in the Report. There is also a strong and significant correlation between the index and changes in Spanish GDP. In this case, the coefficient is negative: a deterioration or negative growth in the economy is linked to a rise in the sentiment index. Significant coefficients are also observed for the valuation indicators. Accordingly, a decline in real housing prices or in the market-to-book ratios of banks is associated with a deterioration of the sentiment index. Lastly, the coefficients for these variables do not change their sign when the regressions are performed with the one-year lagged variables, although their level of significance does decline and some are rendered non-significant.

8. Other rules for calculating the index: the TF-IDF methodology

Some sentiment analyses will typically weight terms with sentiment, meaning that not all words are assigned the same weight. One commonly used weighting technique is Term Frequency-Inverse Document Frequency (TF-IDF) (Manning et al. (2009)). This concept originated in information retrieval algorithms based on search strings and is used to calculate a term's importance within a document. Each term is weighted taking into account two aspects: first, term frequency, i.e. the number of times the term appears in the document analysed, such that the more frequently a particular term appears, the more significant it is; and second, inverse document frequency, which factors in the number of documents in which the term appears within the set of documents under analysis. A term will be considered less relevant in a particular document if that same term appears in many other documents. For example, the word "risk" appears in all IEFs. Therefore, the term will be less relevant in a particular IEF than other words that are specific to that Report.

Based on their analysis, Loughran and McDonald (2011) recommend TF-IDF weighting, arguing that this approach reduces to a minimum the errors caused by the inclusion in the dictionary of words that may be considered empty or irrelevant owing to their high frequency.[13] For their part, Jegadeesh and Wu (2013) maintain that there is no specific reason why a word's frequency of occurrence in a document should be related to the perceived sentiment, suggesting an alternative weighting based on the perceived reaction. This analysis, which focuses on annual reports, was based on market reaction. For our purposes, there is no quantitative value that can be taken as a direct measurement of the perceived reaction to an IEF. Shapiro et al. (2019) run a comparison using newspaper articles, concluding that there is no significant difference between a proportional mechanism and one weighted by TF-IDF. Correa et al. (2017) do not employ any weighting mechanism, since they consider these more useful for large samples of small documents.

Using lemmas to build the dictionary rather than lexemes means that two lemmas with a shared root may have different weightings under the TF-IDF approach, simply because of their different frequencies in the document. As we have noted, Spanish is a highly inflected language. It does not seem reasonable to think that the different frequencies of two words that differ only in their inflection would mean different levels of polarity. Accordingly, for the purposes of comparison

[13] Loughran and McDonald (2011) conduct their analysis based on negative words only, given that they frequently observe positive terms used to describe negative situations. This makes it very difficult for an automated process to capture the true meaning of these positive words.

with the unweighted *FSSI* (equation [4]), we use term weighting based on the frequency of the base lemma rather than the specific lemma. We thus calculate the TF-IDF weighted index through the following difference:

$$FSSI_w = \sum_{T \in negative\ words} w_T - \sum_{T \in positive\ words} w_T \ , \qquad [5]$$

where w_T represents the weight of the term T (with connotation) within the document (Appendix B details how these weightings are calculated). The result does not significantly differ from the unweighted *FSSI* (see the correlations in Tables 2 and 3). As in the case of Shapiro et al. (2019), this result may be indicative of the robustness of the final dictionary.

Although we observe no benefit to using a TF-IDF weighted index in our sentiment analysis, we do regard this weighting as useful for identifying the most relevant words in a specific Report. This is illustrated in the word clouds in Figure 7, where the word size represents the weight of the base lemma in each IEF according to the TF-IDF algorithm. Assigning greater relevance to the specific terms in each Report brings to light trends for key concepts over the various publications.

9. Comparison with press sentiment

Creating a sentiment index allows comparative analyses to be conducted for a set of documents. These comparisons may be between different time periods or dates, between different parts of the same Report or between two texts of different but related origin. This section examines the latter, comparing the indices calculated for the texts of the Report with those calculated for press reports published in the days following the release of the corresponding IEF. Text was extracted from press articles using the Banco de España's in-house news briefings, which select press reports relating to the Banco de España, the financial system and the economic situation. In particular, the news briefings from up to three days subsequent to the publication of the Report were used, analysing those articles that are classified as "news about the Banco de España" and that also mention the IEF. The same text extraction process was followed as for the Report. The dictionary created for the Report was likewise used to calculate the number of words with connotation and the same calculation rules set out in the preceding section were followed.

Rows 2 and 3 of Table 3 show the correlations between the indices calculated for the press reports and the indices calculated for the overview and body of the IEF. For *Negativity*, *Net negativity* and the *FSSI* (unweighted and weighted), the correlation between the news articles and the Report texts is relatively high and significant. In particular, for the *FSSI*, the correlations are 0.66 and 0.61 for the overview and the body of the text, respectively. By contrast, the correlations calculated for *Positivity* are not significant at a confidence level of 5%. Figure 8 illustrates the trend for the *FSSI* for the three sets of texts considered. The press index shows greater variability than the other indices, owing to the limited number of news articles that refer to the Report, particularly the oldest Reports. Broadly speaking, as seen in Table 3, a slightly stronger correlation is observed with the IEF overview than with the body of the text.

Consolidating unstructured textual data (the IEF and news in the press) through a structured numeric metric (e.g. the *FSSI*) helps identify and compare certain points of interest. Looking at Figure 8, some specific dates are worth noting. For example, in the Autumn 2006 Report, *Positivity* and *Negativity* increased with respect to the previous Report (see Figure 5). Compared with the Report, the news in the press shows a sharp rise in negativity. Figure 9 illustrates the differences between the word cloud for the overview and that for the newspapers. The press placed emphasis on the significant growth in non-performing loans to the construction and real estate development sectors (using words such as “danger”, “concern” and “fear”), leaving to one side other more upbeat aspects of the Report (e.g. the “vigour” of the banking business and the “favourable” economic situation despite the elements of “concern”).

Also notable is the Spring 2015 IEF, which records the lowest index values for both the overview and the body of the Report. Here the IEF indices diverge substantially from the news index. In the case of the IEF indices, the launch of the “Draghi Plan” brings an increase in optimism, compared with the greater volatility that had prevailed since the publication of the previous Report, with words such as “recovery”, “improvement” and “favourable”, compared with the lower relative weight of negative words (see the word cloud in Figure 9). For its part, the news index reflects a higher occurrence of negative words and, consequently, greater pessimism. In this case, the press had focused on messages relating to the recommendation to cut expenditure, in particular in terms of the number of bank branches.

A number of issues arise when these discrepancies are analysed. First, they are in line with the lower correlation observed between the index constructed using the IEF texts and the index based on press reports (see Table 3). If only the positivities are considered, the correlation is not significant. Here it is worth noting the difficulty in establishing a criterion to identify whether these correlations with the press reports are high or low, or whether the short-time differences between these indices are related to the Banco de España’s communication strategy. For example, where the sentiment index for the news items is negative, it may be that the committee responsible for the Report preferred to convey the robustness of the system to certain events, leading to a divergence between the indices, without this entailing any loss of transparency in the communication.

Lastly, we believe it is important to emphasise the usefulness of the dictionary in this case. Although the dictionary does not include all of the tonality words used in the press reports, it does function adequately as a proxy to calculate sentiment, even for out-of-sample texts.

10. Conclusions

In recent years, content analysis of central banks’ various communications has gained relevance. Most studies are designed for texts in English and adopt approaches with that language in mind. This paper presents the first sentiment dictionary in Spanish in the area of financial stability. In addition to extending such content analysis and text mining to the Spanish language, the paper gives a detailed description of the procedure employed and analyses the consistency of the sentiment annotations using various non-parametric statistics. This contributes to the literature on such annotation processes in any language. The work involved in annotation tends to be underestimated; it is important to have tools to assist in this task.

Drawing on the dictionary created based on texts extracted from the Banco de España's IEFs between 2002 and 2019, the study analyses various sentiment indices constructed using functions on the number of words with connotation. The resultant indices serve to quantitatively measure the IEF texts, i.e. unstructured data, and have proven to be consistent when calculated for different sections of the Report (the overview and the body of the text). Further, they are robust to variations in the dictionary (whether random or based on the annotation results) and to changes in the methodology used to establish the weightings of each word within the index (equal-weighted or weighted by frequency of occurrence). The indices are consistent with the key macroeconomic events that took place during the sample period, and also contain information on other quantitative financial indicators (e.g. the banking sector's credit risk premium).

Lastly, it is important to note that in this type of analysis, the results may be influenced by structural changes in the composition of the IEF, be they changes on the editorial board or on the Executive Commission that approves publication of the Reports, which may have a bearing on the tone of the text.

Having a dictionary paves the way for sentiment analysis of other financial texts in Spanish, allowing comparisons to be made on a fairly objective basis. Thus, we also calculate indices for news items relating to the IEF and analyse the press reaction to the Report's publication. The results indicate that the list of words included in the reference dictionary is sufficiently consistent to provide a reliable estimator of press sentiment. A greater correlation is observed between the

IEF and the news item indices when they are calculated using only words with negative tone than when they are calculated using only positive words.

We believe this analysis also serves as a basis and reference for potential alternative approaches in future, including analysis with additional rules, the exploration of thematic sub-indices or the use of statistical machine learning models that better factor in the contextual use of each word so as to achieve greater precision.

Appendix A. Inter-judge agreement calculation

The general form of the various chance-corrected agreement coefficients (CAC) used in the study is the following: [14]

$$\text{CAC} = \frac{P_a - P_e}{1 - P_e}, \qquad \text{[A.1]}$$

where P_a is the probability or percentage of observed agreement and P_e is the probability or percentage of agreement by chance. For two annotators, Cohen's coefficient κ and Gwet's coefficient AC_2 differ in how they estimate the probability of agreement by chance P_e, while Krippendorff's coefficient α also differs in its approach to estimating P_a.

The following example looks at two annotators recording the sentiment of ten words, classifying their polarity as positive (+), negative (-) or neutral (○):

	X	Y
1	○	○
2	○	○
3	○	○
4	○	○
5	○	○
6	○	○
7	+	+
8	+	-
9	+	+
10	-	-

The two annotators classify the ten words in the same way with the exception of one word, which annotator X considers to have positive polarity and annotator Y negative polarity. The above information can be summarised in the following contingency table, C_{XY}:

		Y			
		-	○	+	
X	-	1	0	0	1
	○	0	6	0	6
	+	1	0	2	3
		2	6	2	10

[14] An equivalent definition for these coefficients in terms of disagreement is as follows:

$$\text{CAC} = \frac{P_a - P_e}{1 - P_e} = 1 - \frac{D_o}{D_e},$$

where $D_e = 1 - P_e$ is the weighted percentage of disagreement attributed to chance and $D_o = 1 - P_a$ is the weighted percentage of observed disagreement.

The diagonal represents the number of words for which both annotators agree on each polarity, while the off-diagonal elements represent the various disagreement combinations.

Calculation of the agreement coefficients also allows second-level agreements or partial agreements (i.e. between negative or positive and neutral) to be taken into account. Thus, weighted probabilities can be calculated adjusting for the level of agreement. Our analysis uses linear weightings, according to the following weight matrix, W:

		Y		
		-	○	+
X	-	1	0.5	0
	○	0.5	1	0.5
	+	0	0.5	1

Calculation of Cohen's coefficient *κ*

Cohen's coefficient κ (Gwet (2008)) is determined as follows:

$$\kappa = \frac{P_a - P_{e|\kappa}}{1 - P_{e|\kappa}}, \tag*{[A.2]}$$

where the probability of agreement P_a is the weighted sum of agreements divided by total annotations. In other words, it is the sum of the elements of the contingency matrix C_{XY} weighted by the elements of the weight matrix W. In the example, it would be: $P_a = \frac{1+6+2}{10} = 0.9$.

$P_{e|\kappa}$ is calculated as the weighted probability of agreement by chance. To this end, the probabilities that the values of the contingency matrix have been arrived at by chance are obtained, assuming that the probabilities of annotators X and Y assigning an annotation (+, -, ○) are independent, i.e. $P_X \cap P_Y = P_X P_Y$. The probability that Y assigns a particular annotation is calculated by dividing the number of times that Y has assigned that annotation (the totals of each column in the matrix C_{XY}) by the total number of annotations (10). Weighting these probabilities of agreement by chance with the elements of the linear weight matrix W, gives $P_{e|\kappa}$.

In our example, as summarised in the contingency matrix C_{XY}, the matrix of probabilities by chance would be:

		Y		
		-	○	+
X	-	0.02	0.06	0.02
	○	0.12	0.36	0.12
	+	0.06	0.18	0.06

Multiplied by the weights: $P_{e|\kappa} = 0.02 + 0.36 + 0.06 + 0.06 \times 0.5 + 0.12 \times 0.5 + 0.12 \times 0.5 + 0.18 \times 0.5 = 0.68$. Finally, the resultant Cohen's coefficient would be: $\kappa = \frac{0.9-0.68}{1-0.68} = 0.6875$.

Calculation of Krippendorff's coefficient α

The general form of Krippendorff's coefficient α (Gwet (2011)) is defined as:

$$\alpha = \frac{P_{a|\alpha} - P_{e|\alpha}}{1 - P_{e|\alpha}} \qquad [A.3]$$

In this case, the value of $P_{a|\alpha}$ is calculated in a similar fashion to the case of Cohen's coefficient, but it always results in a slightly higher value. When there are no blank annotations, $P_{a|\alpha}$ is determined by:

$$P_{a|\alpha} = \left(1 - \frac{1}{nr}\right) P_a + \frac{1}{nr}, \qquad [A.4]$$

where n is the number of words that will be annotated and r is the number of annotators. In our example: $P_{a|\alpha} = \left(1 - \frac{1}{10\times2}\right) \times 0.9 + \frac{1}{10\times2} = 0.905$.

For this coefficient, the weighted probability by chance $P_{e|\alpha}$, based on the annotation results, is calculated by first determining the probabilities of any word being included in a category (+, -, ○). In this case, it is useful to view the annotations in the form of a table of agreement, where each column represents a category and each row a word, and indicating in each cell the number of annotators classifying that word in that particular category. As in the case of the previous coefficient, to determine $P_{e|\alpha}$ the probabilities by chance are weighted using the linear weight matrix W.

The probability of a word being assigned to a specific category is determined by the number of times that the category has been assigned divided by the total number of assignments (20 in our case). In our example, the table of agreement would be as follows:

	-	○	+
1	0	2	0
2	0	2	0
3	0	2	0
4	0	2	0
5	0	2	0
6	0	2	0
7	0	0	2
8	0	0	2
9	1	0	1
10	2	0	0
Total	3	12	5
Probability (π)	0.15	0.60	0.25

The matrix of probability by chance is determined based on the above table, using the same

		Y		
		-	○	+
X	-	0.0225	0.09	0.0375
	○	0.09	0.36	0.15
	+	0.0375	0.15	0.0625

and the weighted probability $P_{e|\alpha}$ for this coefficient: $P_{e|\propto} = 0.0225 + 0.36 + 0.0625 + 0.09 \times 0.5 + 0.09 \times 0.5 + 0.15 \times 0.5 + 0.15 \times 0.5 = 0.685$. Lastly, the resultant Krippendorff's coefficient would be: $\alpha = \frac{0.905-0.685}{1-0.685} = 0.698$.

Calculation of Gwet's coefficient AC_2

The formula for the coefficient AC_2 is:[15]

$$AC_2 = \frac{P_a - P_{e|AC}}{1 - P_{e|AC}} \qquad \text{[A.5]}$$

[15] The general form of Gwet's coefficients AC_1 and AC_2 is the same. The difference lies in that AC_1 uses the identity matrix as weighting matrix and therefore does not take into account second-level agreements.

P_a is calculated in Gwet's coefficient in the same way as in Cohen's coefficient. However, the weighted probability by chance, $P_{e|AC}$, is defined as:

$$P_{e|AC} = \frac{T_w}{q(q-1)} \times \sum_{k=1}^{q} \pi_k (1 - \pi_k), \qquad [A.6]$$

where T_w is the sum of the weights of the weighting matrix W, q is the number of possible classifications and π_k is the probability that an annotator assigns a classification of k to a specific observation, i.e. those calculated in the table of agreement used to calculate Krippendorff's coefficient α (Gwet (2014)).

In our example, with two annotators classifying in three levels (+, -, ○) and a sum of weights of 5: $P_{e|AC} = \frac{5}{6} \sum_{k \in \{+,\circ,-\}} \pi_k (1 - \pi_k)$, with π_k being the probabilities of the table of agreement: $\pi = (0.15;\ 0.60;\ 0.25)$. Therefore: $P_{e|AC} = \frac{5}{6}\big(0.15(1 - 0.15) + 0.6 \cdot (1 - 0.6) + 0.25 \cdot (1 - 0.25)\big) = 0.4625$. And Gwet's coefficient would be: $AC_2 = \frac{0.9 - 0.4625}{1 - 0.4625} = 0.81395$.

As shown in Figure A.1, the formula of the general form of the chance-corrected agreement coefficients (CACs) causes the value of the coefficients to decline sharply when P_e is high. This implies that values may vary significantly depending on the method used to calculate P_e, and that high P_e values will increase the likelihood of low CAC values.

Appendix B. TF-IDF weighting

In a TF-IDF weighting, the *term frequency* $TF_{T,d}$ is defined as the number of times that the term T appears in the document d. Thus, the more times a particular word appears in the document, the more significant it is. The *document frequency* DF_T is defined as the number of documents that contain the term T. To weight the relevance of a term, and where N is the total number of documents, the *inverse document frequency* of a word T is defined as:

$$IDF_T = \log \frac{N}{DF_T} \qquad [B.1]$$

Therefore, the TF-IDF relevance is defined as the product of the *term frequency* and the *inverse document frequency*:

$$TF\text{-}IDF_{T,d} = TF_{T,d} \times IDF_T \quad [B.2]$$

Since it does not seem reasonable that a term appearing in a document 20 times is indeed 20 times more important than another term that appears just once, it is common to use a weighting based on the logarithmic term frequency, as follows:

$$WTF_{T,d} = \begin{cases} 1 + \log TF_{T,d} & if\ TF_{T,d} \geq 1 \\ 0 & all\ other\ cases \end{cases}, \quad [B.3]$$

leading to a weighted relevance of:

$$WTF\text{-}IDF_{T,d} = WTF_{T,d} \times IDF_T, \quad [B.4]$$

thus reducing the importance of terms that appear repeatedly in a document. This weighting does not factor in the length of the document, given that a relevant term is likely to appear more frequently in longer documents. Other term frequency weighting mechanisms not only use the logarithmic scale but also take into account document length by means of the average term frequency. These are called “log ave weightings”, defined as:

$$WTF_{T,d} = \begin{cases} \frac{1+\log TF_{T,d}}{1+\log(a_d)} & if\ TF_{T,d} \geq 1 \\ 0 & all\ other\ cases \end{cases}, \quad [B.5]$$

where $a_d = \mathrm{ave}_{i \in d}(TF_{i,d})$ is the average frequency of all terms i in document d (excluding empty words). Alternatively:

$$a_d \equiv \mathrm{ave}_{i \in d}(TF_{i,d}) = \sum \frac{n_d}{u_d}, \quad [B.6]$$

where n_d is the number of words in document d and u_d is the number of words after duplicates are excluded. In light of their analysis, Loughran and McDonald (2011) recommend this log ave term frequency weighting.

Accordingly, as well as the unweighted *FSSI* (equation [4]), in our analysis we also calculate a sentiment index (equation [5]) taking into account the weighted relevance w_T of each term T included in the final dictionary, defined as:

$$w_T \equiv WTF\text{-}IDF_{T,d} = \begin{cases} \frac{(1+\log(TF_{T,d}))}{(1+\log(a_d))} \log \frac{N}{DF_T} & if\ TF_{T,d} \geq 1 \\ 0 & all\ other\ cases \end{cases}, \qquad [B.7]$$

where N is the total number of reports under analysis, DF_T is the number of reports with the word T, $TF_{T,d}$ is the number of times that the word T appears in Report d, and a_d is the average term frequency of Report d.

Appendix C. Considerations on the annotation process

For the purposes of future annotation work, in this Appendix we document the various comments on the process made by the annotators involved in the study. Since the annotation process was new to the entire team, some annotators noted that their criteria changed as the annotations progressed. Some annotators were initially more inclined to select a specific tone (positive or negative), but this inclination waned as the process advanced, and they began to opt for neutral when the tone was not clearly defined. The opposite was true for other annotators, who initially preferred the neutral option unless a very clear tone was evident, and subsequently looked for a trend in the examples that would allow a term to be classified as positive or negative.

Since the word selection was based on lexeme or root frequency, despite the classification being performed at the lemma level, on occasions there were just one or two examples of certain lemmas. In such cases, the variability in assigning polarity was influenced by the annotator’s ultimate judgement. An annotator who considered the word family to have a clear pattern might opt to assign the same polarity to that word. Other annotators tended to assign neutral polarity in such cases, although this criterion also changed over time.

Where a certain word always appeared together with other words that consigned the sentences to the same polarity, that word was generally assigned the polarity identified in the sentence. In many cases, the polarity depended on the modifiers that appeared together with the word to be classified (e.g. in relation to a rising or falling trend). A neutral classification was selected in such cases, albeit identifying the possibility of adding additional rules to take into account the modifiers when classifying these words. Words that generally appeared in chart titles or captions were assigned neutral polarity.

Appendix D. List of dictionary words

Positive words

absorbidas
abundancia
abundante
acomodaticia
acomodaticias
acomodaticio
afiance
afianzado
afianzamiento
afianzando
ágil
alcista
alcistas
aliviadas
aliviado
aliviando
aliviar
aliviará
aliviaron
alivio
amortigua
amortiguación
amortiguador
amortiguan
amortiguar
amortiguarlos
amortiguarse
apoyada
asentarse
atenuación
atenuados
beneficiándose
beneficiar
beneficiara
beneficiarán
beneficiarían
beneficiaron
beneficiarse
beneficien
beneficioso
benigna
benignas
benigno
benignos
bienestar
buen
buenas
buenos
calma
calmar
capaces
capaz
cómoda
contención
desendeudamiento
dinamismo
dinamizador
disfrutan
eficaces
eficaz
eficiente
eficientes
equilibrada
equilibrado
excelente
excelentes
expandió
expansiva
favorable
favorablemente
favorables
favorece
favorecen
favorecido
favorecieron
fortaleciéndose
fortalecimiento
fortaleció
fortaleza
fortalezas
ganancia
ganó
holgada
holgadamente
holgadas
holgado
holgados
mejora
mejorada
mejorado
mejoran
mejorando
mejorándose
mejorar
mejoraron
mejorase
mejores
mejoría
mejorías
mejoró
mitiga
mitigaban
mitigado
mitigar
mitigaron
normalidad
normalizado
normalizados
normalizando
normalizándose
normalizar
normalizó
oportunidades
optimismo
ordenada
ordenado
positiva
positivamente
positivas
positivo
positivos
progreso
progresos
propicias
propicio
reaccionado
reactivación
reactivándose
reafirmando
recuperación
recuperado
recuperan
recuperando
recuperándose
recuperar
recuperara
recuperaron
recuperarse
recuperase
recuperen
recuperó
reequilibrando
reequilibrar
reforzado
reforzándolo
reforzará
reforzaron
reforzó
remontado
renovado
resisten
resistido
restablecer
restableciendo
restablecimiento
restaurar
revalorizaba
revalorizaciones
revalorizado
revalorizaron
revalorizarse
revalorizó
revitalización
revitalizar
robusta
robustas
robusto
robustos
saneada
saneado
saneados
sanearon
satisfactoria
satisfactoriamente
sólida
sólidas
solidez
sólido
solvente
solventes
sostenibles
suaves
suavizarán
superada
sustenta
tranquilidad
vigorosamente
vigoroso

Negative words

abrupta
abruptas
abrupto
abruptos
abusivo
acentuaban
acentuadas
acusados
adversa
adversas
adverso
adversos
afrontan
afrontarían
agotamiento
agravada
agravado
agravamiento
agravando
agravar
agravará
agravarían
agravó
agudas
agudizado
agudizamiento
agudizara
agudizaran
agudizaron
agudizó
agudo
agudos
altibajos
amenaza
amenazados
amenazan
amenazar
amenazas
anómala
arrastrado
asimétricos
ataque
ataques
atonía
atravesando
atraviesan
bache
brusca
bruscas
brusco
bruscos
colapsados
colapso
complejidades
complejo
complica
complicaciones
complicada
complicadas
complicado
complicados
complicando
complicar

complicarían
contagiadas
contagiado
contagiaron
contagie
contagio
contagió
contracción
contracciones
contractiva
contrae
contraerse
contrajo
contraproducentes
contrayendo
contrayéndose
convulso
costosa
costosas
costoso
costosos
crisis
cruda
dañar
dañaría
daño
débil
débiles
debilidad
debilidades
debilita
debilitada
debilitado
debilitamiento
debilitan
debilitar
debilitó
débilmente
decepcionante
decepcionantes
decepcionaron
deficiencias
deficiente
deficitaria
delicada
depresión
deprimidos
deprimirían
desaceleración
desastres
desconfianza
desencadenamiento
desequilibrada
desequilibrio
desequilibrios
desestabilizadores
desfavorable
desfavorablemente
desfavorables
destrucción
destruyendo
desvaneciendo
deteriora

deterioraban
deteriorada
deterioradas
deteriorado
deteriorando
deteriorándose
deteriorar
deteriorarse
deteriorase
deterioro
deterioró
difícil
difíciles
dificulta
dificultad
dificultada
dificultades
dificultado
dificultando
dificultándose
dificultar
dificultaría
dificultarían
disfunción
disfunciones
drástica
drásticas
drásticos
dudas
empeora
empeorado
empeoramiento
empeoramientos
empeoran
empeorando
empeorar
empeoró
endurecido
endureciéndose
endurecimiento
endurecimientos
erosión
erosionado
erosionar
escalada
escándalos
escasísima
estallar
estallido
estancada
estrangulamiento
estrangulamientos
evaporarse
excesivo
excesivos
falta
fatiga
frágil
frágiles
fragilidad
fragilidades
fragmentación
frenazo

frustró
grave
gravedad
gravemente
graves
guerra
impactará
inadecuados
incapaces
incapaz
incertidumbre
incertidumbres
incierta
inciertas
incierto
inciertos
inconveniente
indefinición
indeseado
ineficiencia
ineficiencias
inestabilidad
inestable
inestables
insostenible
insuficiencia
insuficiente
insuficientes
intervenida
intervenidas
intervenir
invalidar
inviabilidad
inviable
inviables
irregular
lamentablemente
lastrada
lastradas
lastrado
lastrar
lastre
lastró
lenta
lento
mal
mala
malas
merma
miedo
negativa
negativamente
negativas
negativo
negativos
obstáculo
oscilaciones
padece
padecían
pánicos
peligro
peligros
penalizado

peores
pérdida
perjudica
perjudicadas
perjudiciales
perjuicios
persistencia
persistente
persistentes
persistieron
perturbaciones
perversos
pesimismo
pesimista
pobre
precipicio
prematura
preocupación
preocupaciones
preocupado
preocupados
preocupante
preocupantes
presión
presiona
presionaban
presionada
presionadas
presionado
presionados
presionan
presionando
presionar
presionará
presionaría
presionaron
presiones
problemas
problemática
problemáticas
quebrar
quebró
quiebra
quiebras
ralentice
ralentiza
ralentización
ralentizar
ralentizara
ralentizarse
ralentizó
rebaja
rebajadas
rebrote
recaída
recalentamiento
recesión
recesivas
recrudecían
recrudecidos
recrudecieron
recrudecimiento
recrudeció

rémora
rescatadas
rescatar
resentido
resentirse
restaron
restringiendo
resurgido
resurgimiento
retraimiento
retrasa
retroceder
retrocedieron
retrocedió
retroceso
retrocesos
revés
secuelas
sensible
serias
serio
serios
severa
severas
severo
sobrecalentamiento
sombras
súbita
sufren
sufrida
sufridas
sufrido
sufridos
sufriendo
sufrieran
sufrieron
sufrió
sufrirán
suspensión
temor
temores
tensión
tensiona
tensionaban
tensionado
tensionamiento
tensionando
tensionaron
tensiones
titubeante
traumática
truncada
turbulencia
turbulencias
urgencia
virulencia
volátil
vulnerabilidad
vulnerabilidades
vulnerable
vulnerables

References

Antoine, J.-Y., J. Villaneau and A. Lefeuvre (2014). "Weighted Krippendorff's alpha is a more reliable metrics for multicoders ordinal annotations: experimental studies on emotion, opinion and coreference annotation", *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics,* Gothenburg, Sweden.

Apel, M. and M. B. Grimaldi (2012). *The Information Content of Central Bank Minutes,* Sveriges Riksbank Working Paper No 261.

Apergis, N. and I. Pragidis (2019). "Stock Price Reactions to Wire News from the European Central Bank: Evidence from Changes in the Sentiment Tone and International Market Indexes", *International Advances in Economic Research,* 25(1), pp. 91-112.

Aureli, S. (2017). "A comparison of content analysis usage and text mining in CSR corporate disclosure", *The International Journal of Digital Accounting Research*, pp. 1-32.

Born, B., M. Ehrmann and M. Fratzscher (2014). "Central bank communication on financial stability", *Economic Journal*, pp. 701-734.

Callejas, Z. and R. López-Cózar (2008). "Influence of contextual information in emotion annotation for spoken dialogue systems", *Speech Communication,* 50, pp. 416-433.

Cicchetti, D. and A. Feinstein (1990). "High agreement but low kappa: II. Resolving the paradoxes", *Journal of Clinical Epidemiology,* 43(6), pp. 551-558.

Correa, R., K. Garud, J. M. Londono and N. Mislang (2017). "Constructing a Dictionary for Financial Stability", *IFDP Notes*.

- (2018). *Sentiment in Central Banks' Financial Stability Reports,* International Finance Discussion Papers, No 1203.

Digitext, Inc. (s.f.). *DICTION is a computer-assisted text-analysis (CATA) programme,* retrieved in 2019 from https://www.dictionsoftware.com/.

Feldman, R., S. Govindaraj, J. Livnat and B. Segal (2010). "Management's tone change, post earnings announcement drift and accruals", *Review of Accounting Studies,* 15, pp. 915-953.

Gwet, K. L. (2008). "Computing inter-rater reliability and its variance in the presence of high agreement", *British Journal of Mathematical and Statistical Psychology,* 61, pp. 29-48.

- (2011). "On Krippendorff's Alpha Coefficient", *Communication Methods and Measures*.

- (2014). *Handbook of inter-rater reliability,* Advanced Analytics, LLC, 4th ed.

- (2019). "irrCAC: Computing Chance-Corrected Agreement (CAC)", retrieved from https://CRAN.R-project.org/package=irrCAC.

Henry, E. and A. J. Leone (2016). "Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone", *Accounting Review,* 91, pp. 153-178.

Jegadeesh, N. and D. Wu (2013). "Word Power: A New Approach for Content", *Journal of Financial Economics,* 110(3), pp. 712-729.

Kearney, C. and S. Liu (2014). "Textual Sentiment in Finance: A Survey of Methods and Models", *International Review of Financial Analysis,* 33, pp. 171-185.

Krippendorff, K. (2004). "Reliability in Content Analysis: Some Common Misconceptions and Recommendations", *Human Communication Research,* 30(3), pp. 411-433.

Krippner, L. (2015). *Term structure modeling at the zero lower bound: a practitioner's guide,* Palgrave-Macmillan.

Loughran, T. and B. McDonald (2011). "When is a Liability Not a Liability? Textual analysis, Dictionaries and 10-Ks", *The Journal of Finance,* 66, pp. 35-65.

- (2016). "Textual Analysis in Accounting and Finance: A Survey", *Journal of Accounting Research,* 16, pp. 1-11.

Lowe, W., K. Benoit, S. Mikhaylov and M. Laver (2011). "Scaling Policy Preferences from Coded Political Texts", *Legislative Studies Quarterly,* 36.1, pp. 123-155.

Manning, C. D., P. Raghavan and H. Schütze (2009). *An Introduction to Information Retrieval,* Cambridge University Press.

Mielke Jr., P. W., K. J. Berry and J. E. Johnston (2011). "Robustness without rank order statistics", *Journal of Applied Statistics,* 38(1), pp. 207-214.

Miner, G., J. Elder, A. Fast, T. Hill, R. Nisbet and D. Delen (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications,* Academic Press.

Shapiro, A. H., M. Sudhof and D. Wilson (2019). *Measuring News Sentiment,* Federal Reserve Bank of San Francisco Working Paper 2017-01.

Stone, P. J., D. C. Dunphy, M. S. Smith and D. M. Ogilvie (1966). *The General Inquirer: A Computer Approach to Content Analysis,* Cambridge, MIT Press.

Stone, P., R. Bales, J. Namenwirth and D. Ogilvie (1962). "The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information", *Behavioral Science,* 7(4), pp. 484-498.

Weik, M. H. (1961). *A Third Survey of Domestic Electronic Digital Computing Systems,* Maryland, Ballistic Research Laboratories, Aberdeen Proving Ground.

Wong, B. T. and S. Y. Lee (2013). "Annotating Legitimate Disagreement in Corpus Construction", *International Joint Conference on Natural Language Processing,* Nagoya, Japan.

Wongpakaran, N., T. Wongpakaran, D. Wedding and K. L. Gwet (2013). "A comparison of Cohen’s Kappa and Gwet’s AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples", *BMC Medical Research Methodology,* 13(61).

Zhao, X., J. S. Liu and K. Deng (2013). "Assumptions behind Intercoder Reliability Indices", *Communication Yearbook,* 36, pp. 419-480.

Tables

Table 1

Analysis of measures of agreement in the dictionary creation process

Phas e	Annotators a_1, a_2			Annotators a_3, a_4			Annotators a_5, a_6			Annotators a_1-a_6		Annotators a_0-a_6	
	α	κ	AC_2	α	κ	AC_2	α	κ	AC_2	α	AC_2	α	AC_2
(1)	.044	.162	.356	.372	.379	.832	.544	.544	.868	.282	.711	.372	.787
(1b)	.278	.322	.599	.621	.625	.894	.794	.794	.940	.527	.824	.500	.840
(2)	.217	.253	.625	.199	.217	.819	.169	.169	.872	.166	.636	.170	.652
(3)	-	-	-	-	-	-	-	-	-	.404	.902	.454	.905

Note: The statistics or measures of agreement calculated are: Cohen's *kappa* (κ), Krippendorff's *alpha* (α) and Gwet's coefficient AC_2. See Appendix A for technical details on how these non-parametric statistics are calculated. A value of agreement over 0.8 is considered high and between 0.67 and 0.8 is considered moderate (Krippendorff (2004), Gwet (2008)). In phase 1, the initial list of words (3,706 lexemes annotated by a reference annotator R) is divided into three groups of equal size (P_A, P_B and P_C) and each group is reviewed by two annotators. In the event of disagreement, a second round of reviews is held between the two annotators of each group to settle the differences (phase 1b, which only takes into account words over which disagreements arise, $P_{A'}$, $P_{B'}$ and $P_{C'}$). The first and second rows of the table show the statistics for phase 1 and phase 1b. Phase 2 includes only the subset of words in each group over which there is still disagreement after phase 1b ($P_{A'}$, $P_{B'}$ and $P_{C'}$); these words are reviewed by the four annotators of the other groups (a_1 and a_2 review $P_{B'}$ and $P_{C';}$ a_3 and a_4 review $P_{A'}$ and $P_{C'}$, and a_5 and a_6 review $P_{A'}$ and $P_{B'}$). In phase 3 all the annotations are combined, which explains why the calculation is only for all the annotators (a_1-a_6) plus the reference or initial annotator (a_0).

Table 2

Correlation coefficients between the different indices calculated for each text

	Overview	Body	Press reports
Positivity vs. *Negativity*	-0.53**	-0.22	-0.18
FSSI vs. *Net negativity*	0.95***	0.95***	0.84***
FSSI vs. $FSSI_W$	0.90***	0.85***	0.63***

Note: Each column shows the correlations (Pearson coefficients) between two different indices calculated for the same text (overview, body of report and press reports). The indices used are: *Positivity*, *Negativity*, *Net negativity*, *FSSI* and $FSSI_W$ (weighted by the TF-IDF algorithm), corresponding to the equations [1], [2], [3], [4] and [5], respectively. The asterisks *, ** and *** denote the significance of the correlation coefficient at the confidence levels of 5%, 1% and 0.1%, respectively.

Table 3

Correlation coefficients between the indices calculated based on the overview, the body of the text and the press reports

	Positivity	*Negativity*	*Net negativity*	$\boldsymbol{FSSI}$	$FSSI_W$
Overview vs. body of text	0.52***	0.90***	0.91***	0.90***	0.73***
Press reports vs overview	0.26	0.50**	0.62***	0.66***	0.59***
Press reports vs. body of text	0.21	0.43**	0.52**	0.61***	0.44**

Note: Each column shows the correlations (Pearson coefficients) between the same index calculated for two different texts. The indices used are: *Positivity*, *Negativity*, *Net negativity*, *FSSI* and $FSSI_W$ (weighted by the TF-IDF algorithm), corresponding to the equations [1], [2], [3], [4] and [5], respectively. The asterisks *, ** and *** denote the significance of the coefficient at the confidence levels of 5%, 1% and 0.1%, respectively.

Table 4

Regressions between the indices and different financial indicators

Indicator	*FSSI Overview*		*FSSI Body*	
	Contemporaneous	*Lagged (two half-year periods)*	*Contemporaneous*	*Lagged (two half-year periods)*
Changes in GDP	-0.38***	-0.15	-0.30***	-0.14**
	(0.11)	(0.13)	(0.06)	(0.08)
Unemployment rate	0.02	0.02	0.02**	0.01
	(0.02)	(0.02)	(0.01)	(0.01)
Short-term interest rate	-0.07	-0.04	-0.03	-0.01
	(0.07)	(0.08)	(0.03)	(0.04)
Notional interest rate	-0.05**	-0.04	-0.02	-0.01
	(0.03)	(0.03)	(0.02)	(0.02)
Credit-to-GDP gap	-0.004*	-0.004	-0.002	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)
Private non-financial debt service ratio	0.01	0.04	0.02	0.04**
	(0.03)	(0.04)	(0.01)	(0.02)
Changes in housing prices	-0.02***	-0.02	-0.02***	-0.01**
	(0.01)	(0.01)	(0.00)	(0.01)
Dividend yield	0.14***	0.11*	0.10***	0.08**
	(0.05)	(0.07)	(0.03)	(0.04)
Market-to-book ratio	-0.69***	-0.42*	-0.41***	-0.24*
	(0.14)	(0.27)	(0.08)	(0.15)
Risk premium - banks	0.26***	0.15*	0.19***	0.11**
	(0.06)	(0.10)	(0.02)	(0.05)
Volatility USD/EUR	0.03	0.01	0.02**	0.01
	(0.02)	(0.03)	(0.01)	(0.01)
Volatility IBEX-35	0.03***	0.01	0.02***	0.01*
	(0.01)	(0.01)	(0.01)	(0.01)

Note: Shown are the OLS regression coefficients for the *FSSI*, calculated for the overview and body of the texts in relation to various indicators. The variables considered are: q-o-q changes in Spanish GDP (%); the unemployment rate in Spain (%); the short-term interest rate yield (three-month EURIBOR, %); the short-term notional interest rate (%) calculated according to Krippner (2015); the credit-to-GDP gap (%), defined as the difference between the credit-to-GDP ratio and its long-term trend; the private non-financial debt service ratio (%), calculated based on debt service costs (principal repayments and interest payments) as a proportion of income; changes in housing prices (%), measured as the logarithmic changes in the last year in the BIS real residential property price index for Spain; the dividend yield of the IBEX-35 (%); the market-to-book ratio for a representative Spanish bank (Banco Santander); the credit risk premium in the Spanish banking sector, approximated by the log of the CDS spread of Banco Santander; currency volatility (%), measured as the implied volatility of one-month at-the-money USD/EUR options; and the quarterly (90 day) realised volatility (%) of the IBEX-35 share index. The standard errors for each coefficient are shown in brackets and are calculated using the Newey and West (1987) estimator and the correction factor T/df for small samples, where T is the number of observations and df are the degrees of freedom. The asterisks *, ** and *** denote the significance of the coefficient at the usual confidence levels: 10%, 5% and 1%, respectively. The data sources are Bloomberg and the BIS statistics page.

Figures

Figure 1. Process flow diagram used to obtain the IEF sentiment indices. The process starts with the Nuance Power PDF Advanced 3.0 application which converts PDF documents to Word to obtain an editable format. The overview and the body of the text are then selected manually and copied into text files. These are processed using programming tools: on the one hand, to extract examples to be used in the annotation tool, and on the other, once the tonality dictionary has been created, to calculate the indices and create the corresponding data visualisations.

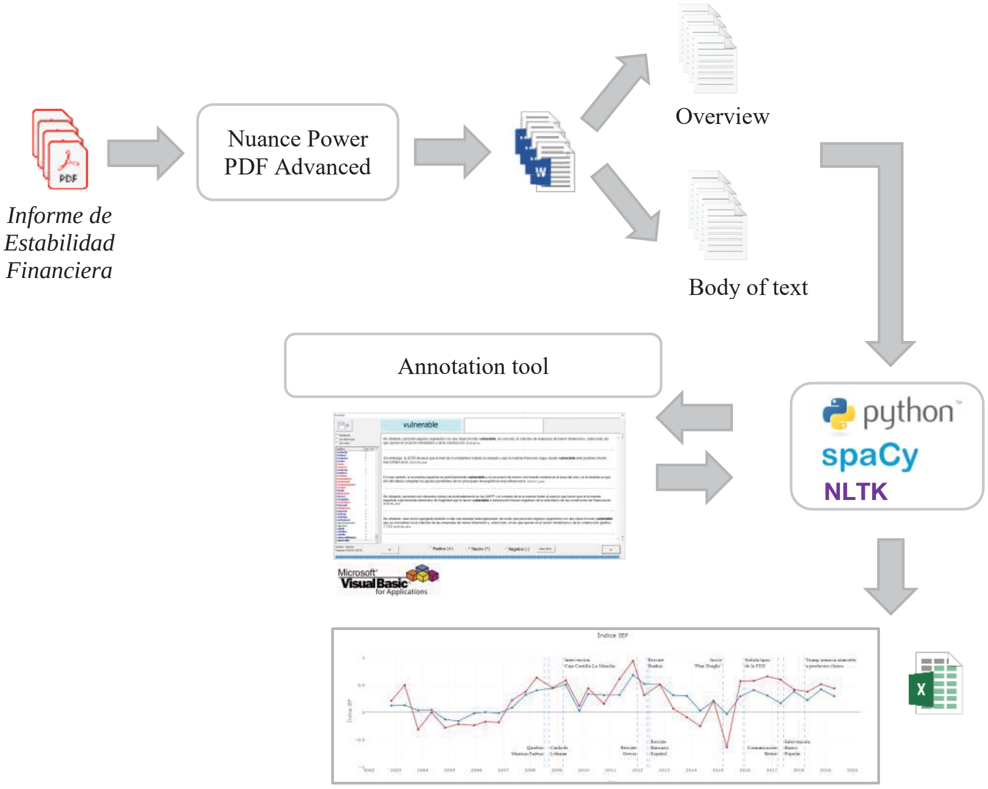

PDF
Informe de Estabilidad Financiera
Nuance Power PDF Advanced
W
Overview
Body of text
Annotation tool
vulnerable
Microsoft Visual Basic for Applications
python
spaCy
NLTK
Índice IEF
X

Figure 2. Phases of annotation process. The starting point is a list of words selected and annotated initially by an annotator (reference annotator, a_0). In phase 1, the list is divided into three groups (P_A, P_B and P_C), each of which is reviewed by two annotators. In the event of disagreement, the annotators of each group review the annotations that differ and change the annotations they deem appropriate (phase 1b). Words with annotations over which disagreement persists (red-shaded areas) are reviewed by the annotators of the other groups (phase 2). Any disagreements remaining in phase 3 are settled by collective wisdom, considering all the annotations (seven different annotators in total, a_0-a_6, with a_0 being the initial reference annotator).

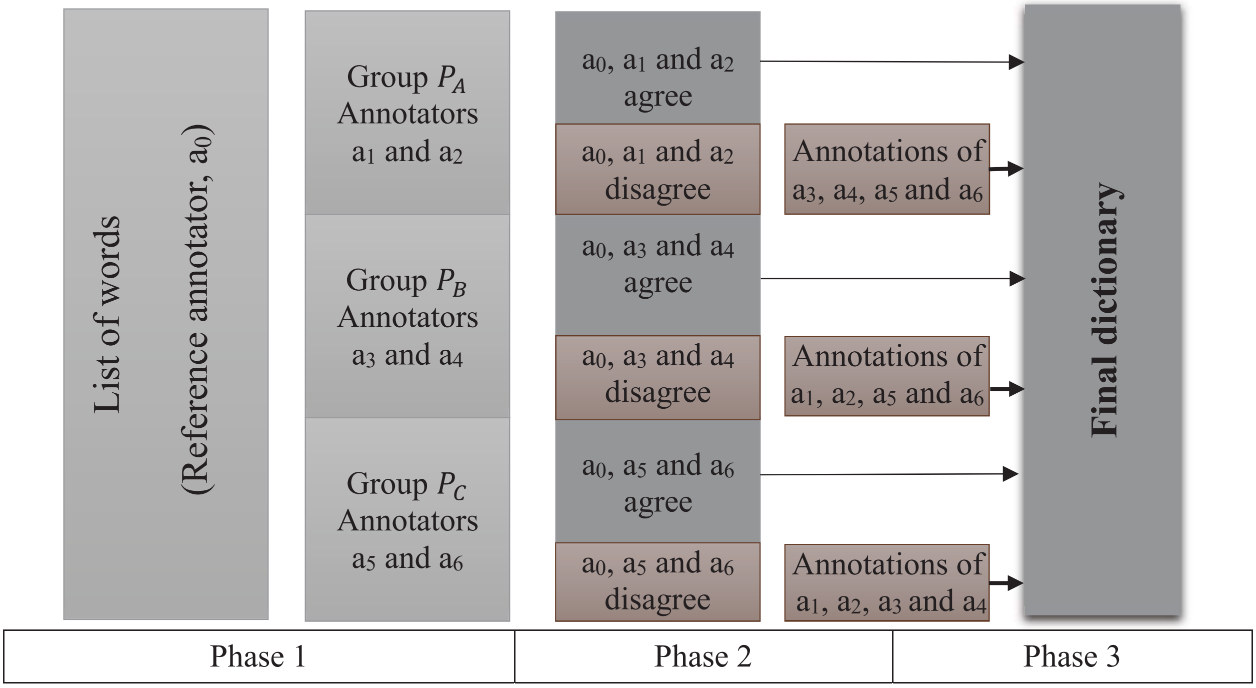

List of words
(Reference annotator, a_0)
Group P_A Annotators a_1 and a_2
Group P_B Annotators a_3 and a_4
Group P_C Annotators a_5 and a_6
a_0, a_1 and a_2 agree
a_0, a_1 and a_2 disagree
Annotations of a_3, a_4, a_5 and a_6
a_0, a_3 and a_4 agree
a_0, a_3 and a_4 disagree
Annotations of a_1, a_2, a_5 and a_6
a_0, a_5 and a_6 agree
a_0, a_5 and a_6 disagree
Annotations of a_1, a_2, a_3 and a_4
Final dictionary
Phase 1
Phase 2
Phase 3

Figure 3. Word clouds in the sentiment dictionary. The bigger the word, the more frequently it appears in the reports analysed. Chart (a) shows all the words assessed by the seven annotators (3,706). In the final dictionary, these words were categorised as: 189 positive, 376 negative and the remainder neutral (see Chart (b)).

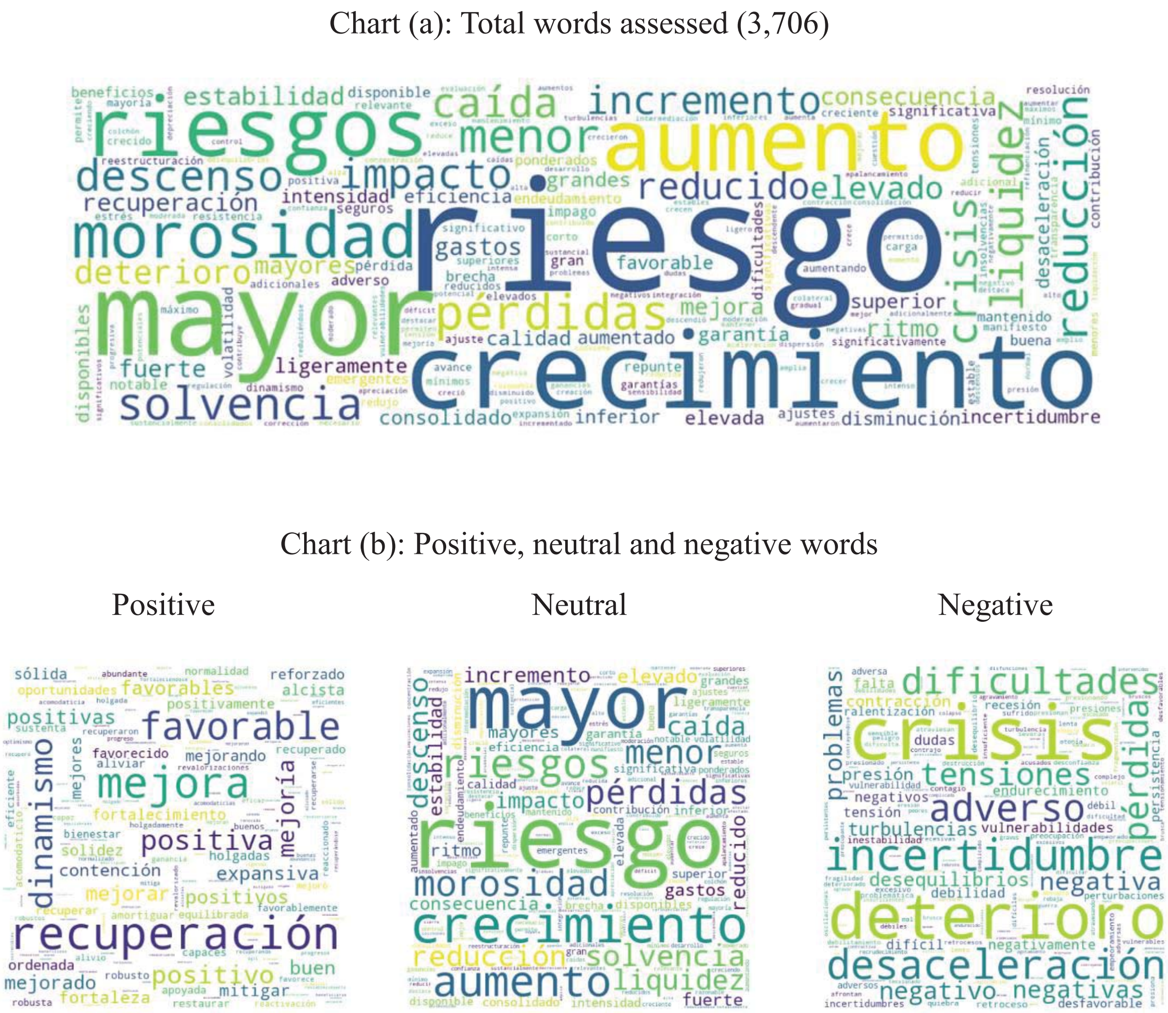


Chart (a): Total words assessed (3,706)

Chart (b): Positive, neutral and negative words

Figure 4. Distance distributions between annotators ($a_i - a_j$; i, j = 0,…,6) after the various phases of the annotation process. Distance may take the values 0, 1 and 2. For each distribution, the corresponding Gwet AC_2 measure is shown in brackets.

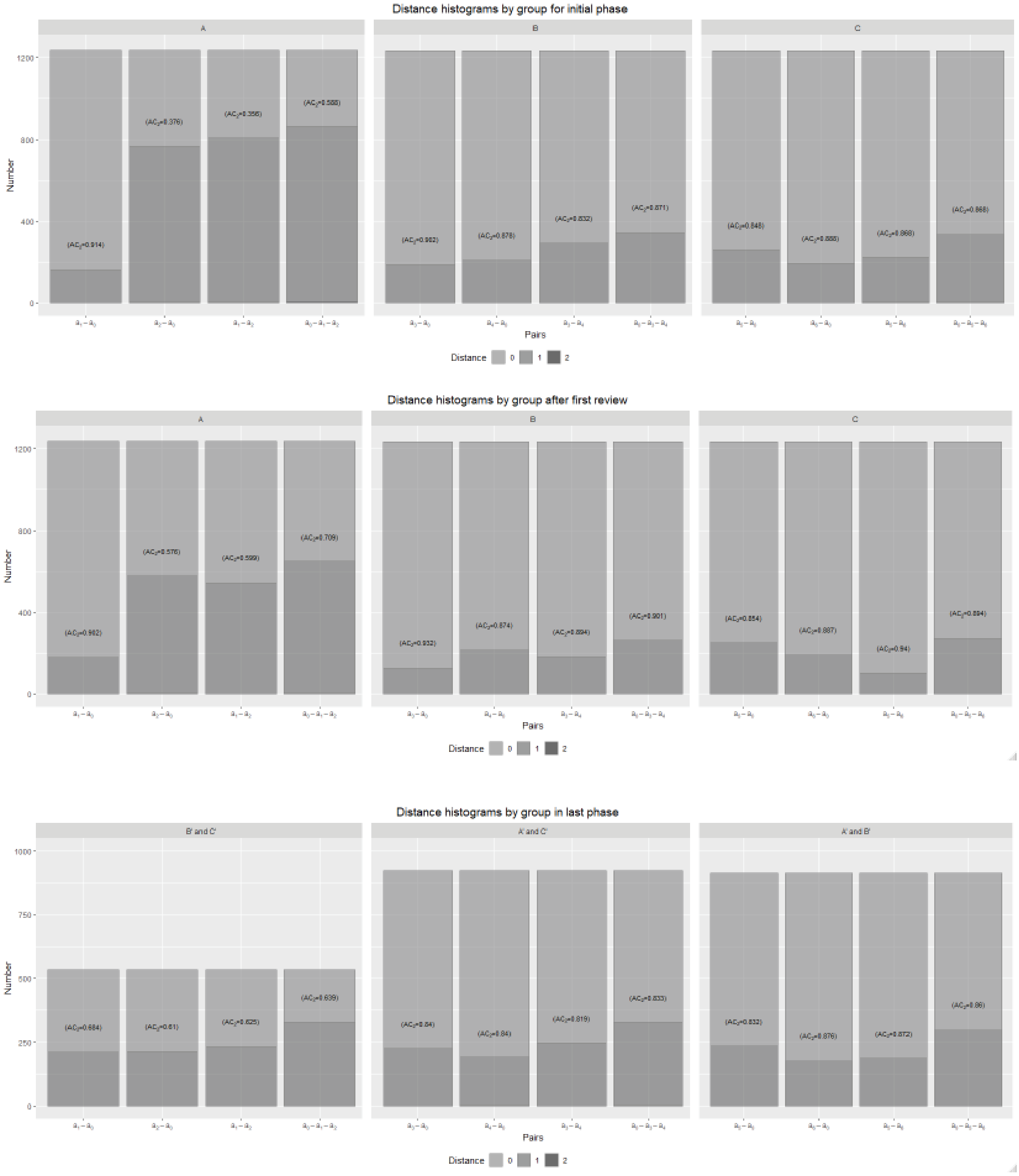


Figure 5. Chart (a) shows the *Positivity*, *Negativity* and *Net negativity* for the overview of the IEF from autumn 2002 to autumn 2019. Chart (b) shows the IEF sentiment index (*FSSI),* calculated as the *Relative net negativity* (equation [4]) for the overview and the body of the Reports in the sample. The grey-shaded bands denote the terms of office of four Banco de España Governors: J. Caruana (July 2000-July 2006), M. Á. Fernández Ordóñez (July 2006-June 2012), L. M. Linde (June 2012-June 2018) and P. Hernández de Cos (June 2018-to date).

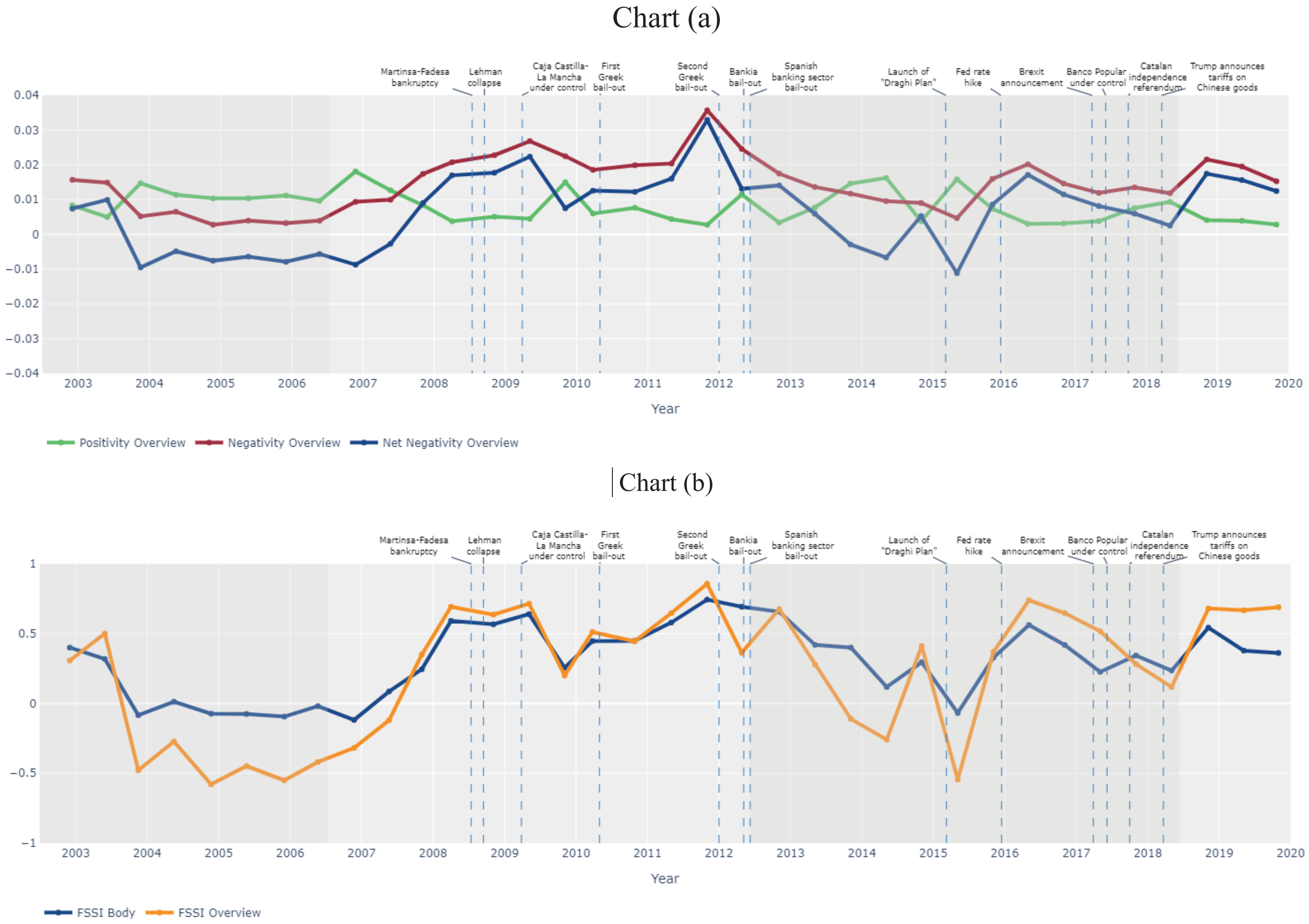


Chart (a)

Chart (b)

Figure 6. FSSI consistency analysis for the body of the Reports from autumn 2002 to autumn 2019. Chart (a) shows the index (red line) and confidence bands (shaded in violet) calculated in 1,000 iterations and randomly removing 5% of the words from the dictionary in each iteration. Both charts also show the slope distributions (index variation between consecutive dates) corresponding to the 1,000 iterations. These distributions are shown in the form of box plot diagrams (in blue).

Chart (a)

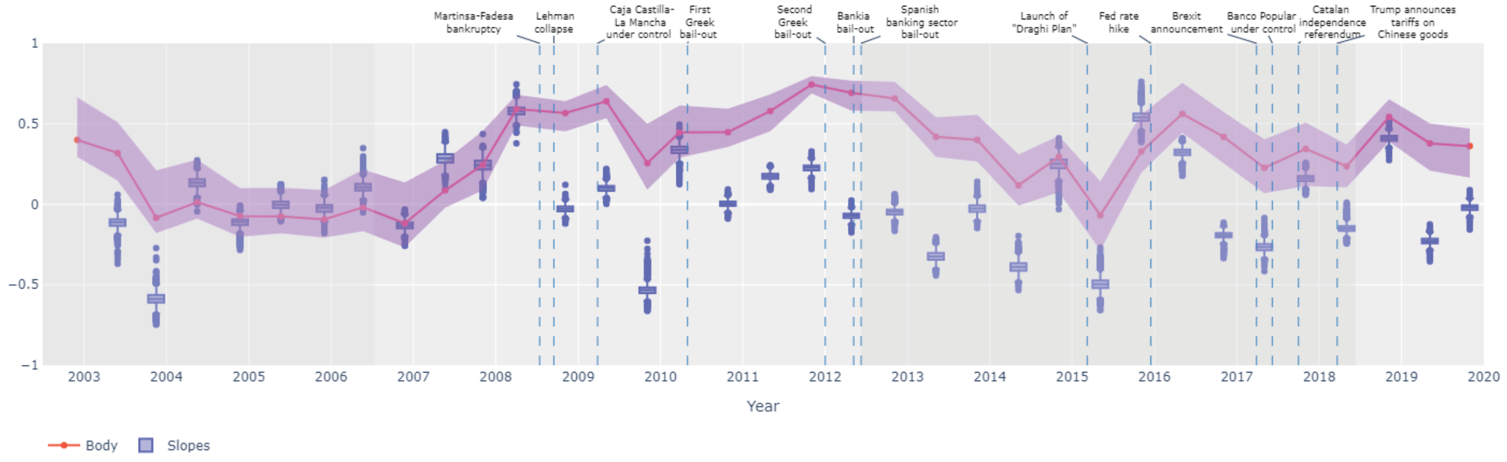


Chart (b)

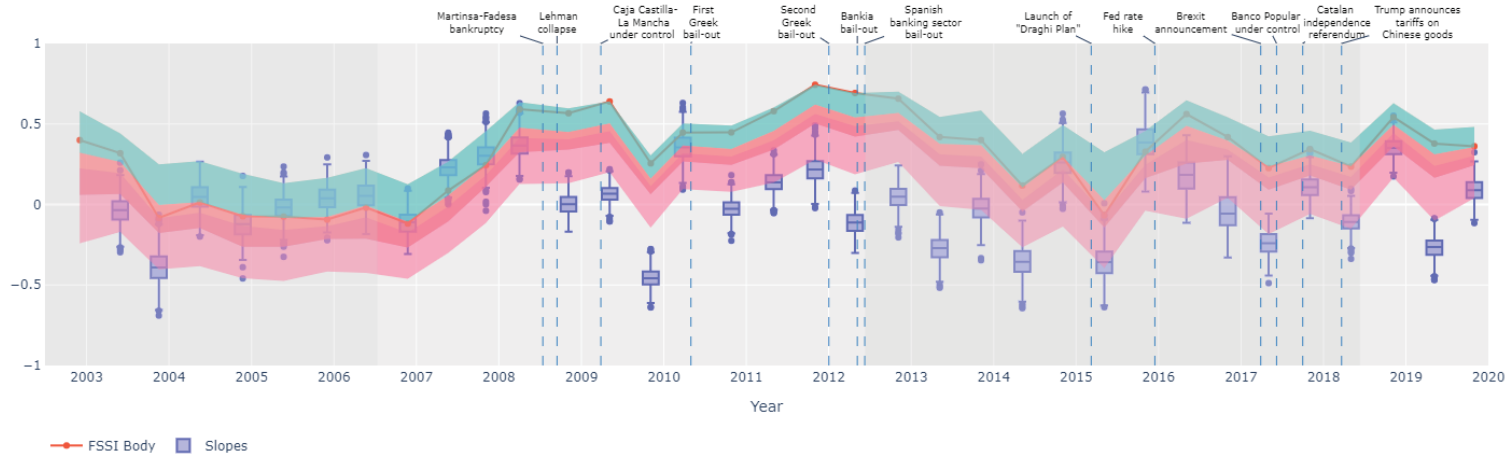


Figure 7. Word clouds for the overview. In each cloud, the size of the word represents the weight of the base lemma in the specific Report using the TF-IDF algorithm. This reveals developments for specific concepts in each Report.

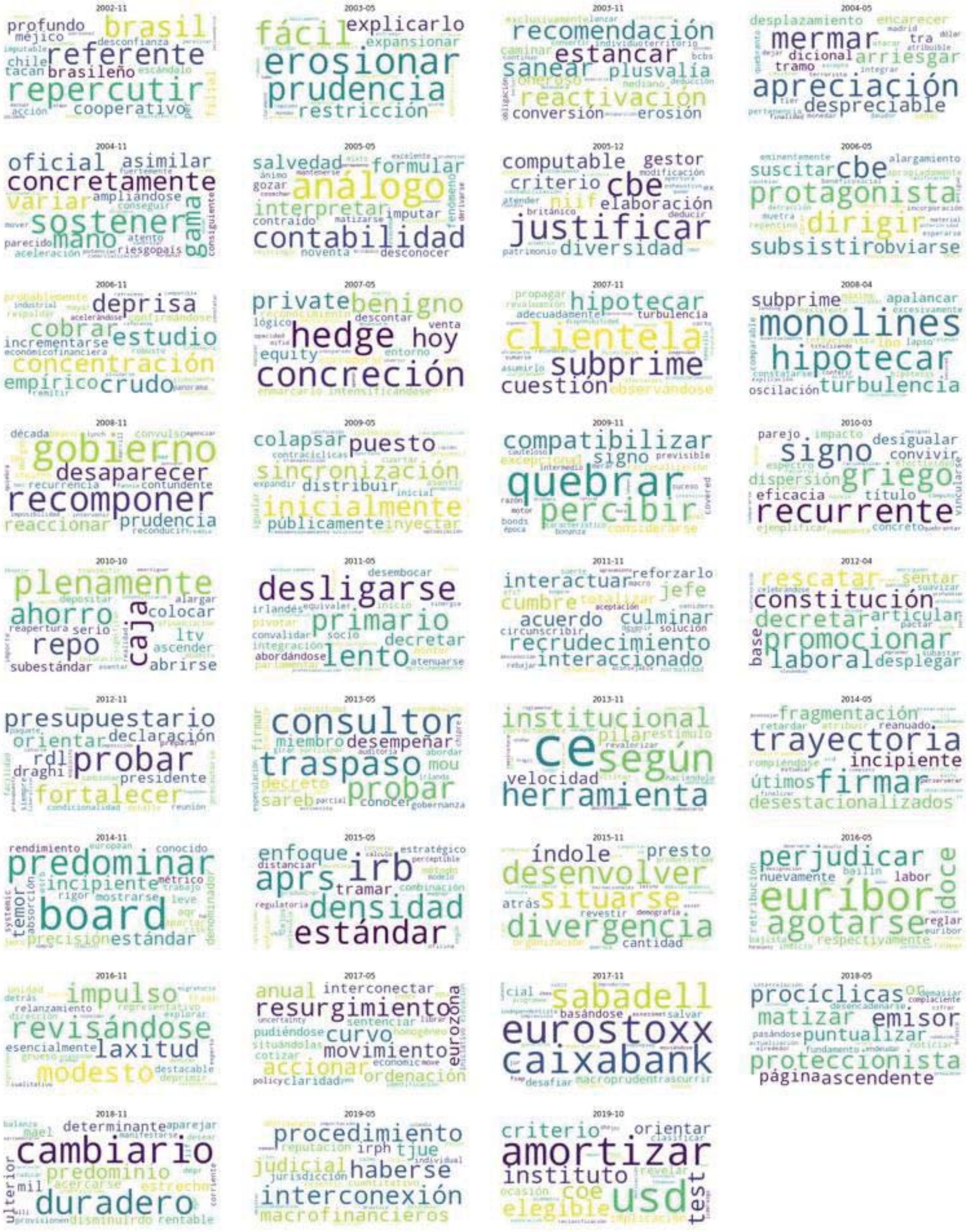


Figure 8. Sentiment calculation for press articles relating to the Reports published between autumn 2002 and autumn 2019. The *FSSI* calculated for the press articles is shown (green line) and compared with the indices calculated using text from the overview and body of the Report (yellow and blue lines).

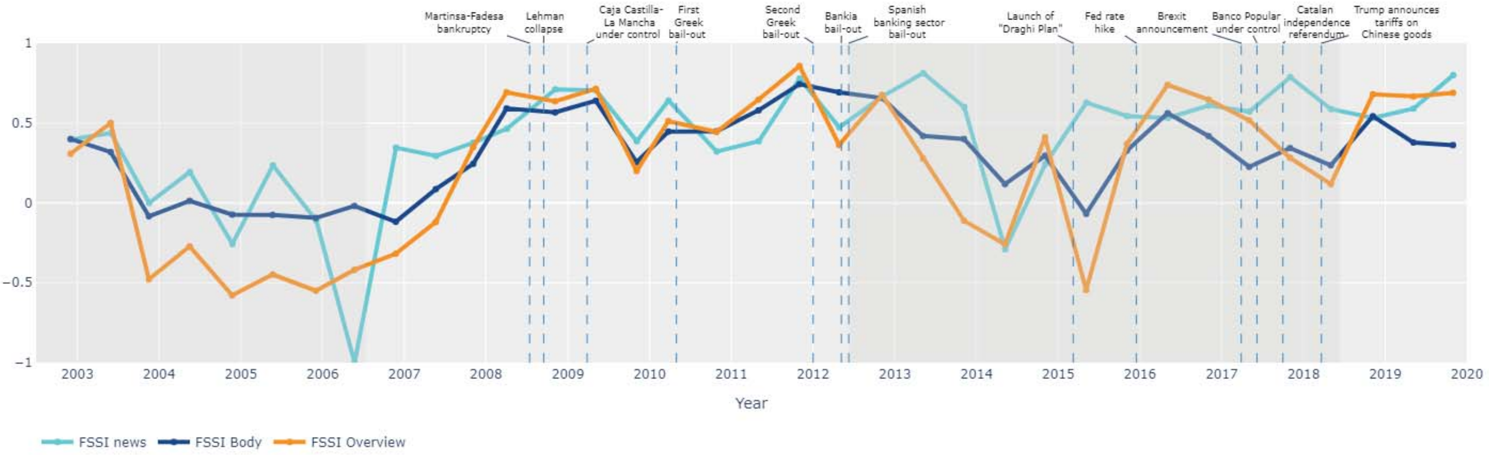


Figure 9. Word clouds based on tonality. Three sets are shown: autumn 2006, spring 2015 and autumn 2019 for the overview (Chart (a)) and for the press articles on the Report (Chart (b)). The size of the words represents their frequency of occurrence in the Report. Words with positive tonality are shown in green and those with negative tonality in red.

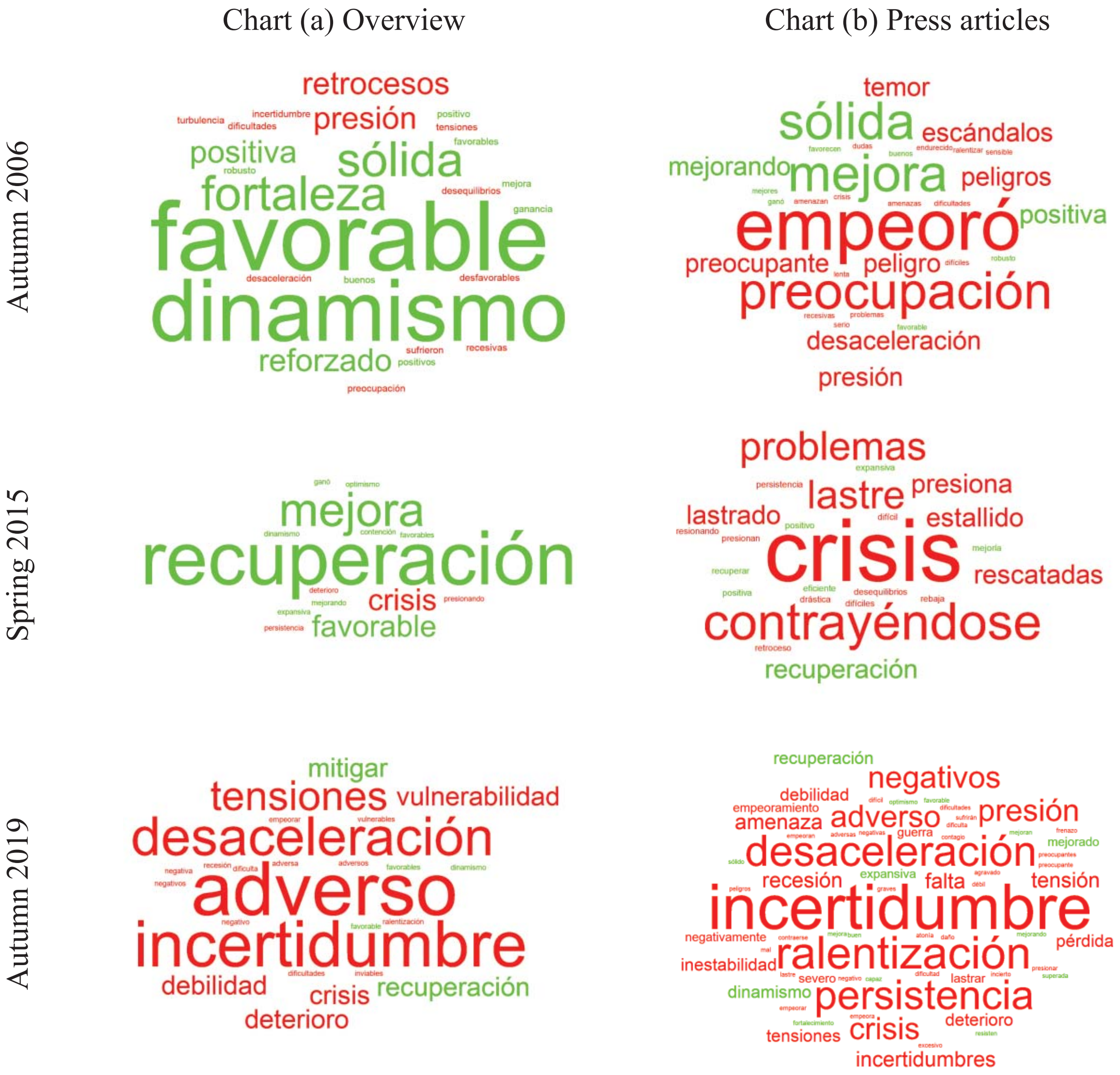
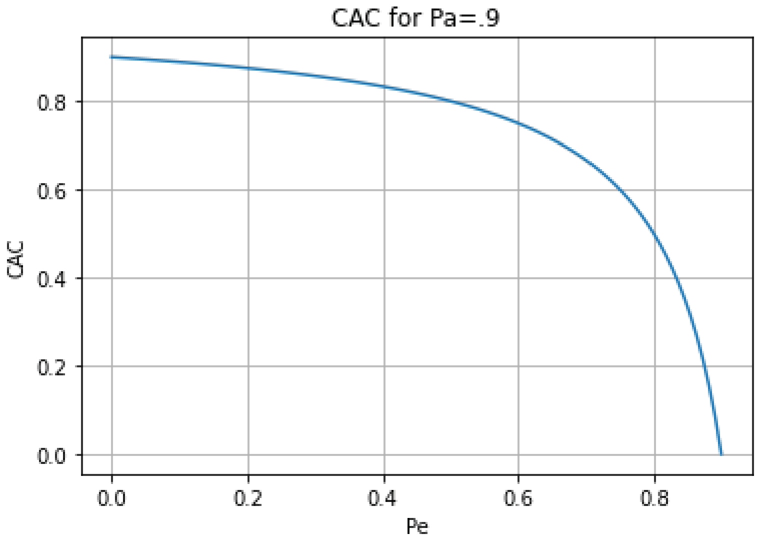


Figure A.1. Change to a CAC type coefficient if $P_a = 0.9$ as a function of P_e, where $CAC = \frac{P_a - P_e}{1 - P_e}$.



BANCO DE ESPAÑA PUBLICATIONS

WORKING PAPERS

1920 LUIS J. ÁLVAREZ, MARÍA DOLORES GADEA and ANA GÓMEZ-LOSCOS: Inflation interdependence in advanced economies.

1921 DIEGO BODAS, JUAN R. GARCÍA LÓPEZ, JUAN MURILLO ARIAS, MATÍAS J. PACCE, TOMASA RODRIGO LÓPEZ, JUAN DE DIOS ROMERO PALOP, PEP RUIZ DE AGUIRRE, CAMILO A. ULLOA and HERIBERT VALERO LAPAZ: Measuring retail trade using card transactional data.

1922 MARIO ALLOZA and CARLOS SANZ: Jobs multipliers: evidence from a large fiscal stimulus in Spain.

1923 KATARZYNA BUDNIK, MASSIMILIANO AFFINITO, GAIA BARBIC, SAIFFEDINE BEN HADJ, ÉDOUARD CHRÉTIEN, HANS DEWACHTER, CLARA ISABEL GONZÁLEZ, JENNY HU, LAURI JANTUNEN, RAMONA JIMBOREAN, OTSO MANNINEN, RICARDO MARTINHO, JAVIER MENCÍA, ELENA MOUSARRI, LAURYNAS NARUŠEVIČIUS, GIULIO NICOLETTI, MICHAEL O'GRADY, SELCUK OZSAHIN, ANA REGINA PEREIRA, JAIRO RIVERA-ROZO, CONSTANTINOS TRIKOUPIS, FABRIZIO VENDITTI and SOFÍA VELASCO: The benefits and costs of adjusting bank capitalisation: evidence from Euro Area countries.

1924 MIGUEL ALMUNIA and DAVID LÓPEZ-RODRÍGUEZ: The elasticity of taxable income in Spain: 1999-2014.

1925 DANILO LEIVA-LEON and LORENZO DUCTOR: Fluctuations in global macro volatility.

1926 JEF BOECKX, MAARTEN DOSSCHE, ALESSANDRO GALESI, BORIS HOFMANN and GERT PEERSMAN: Do SVARs with sign restrictions not identify unconventional monetary policy shocks?

1927 DANIEL DEJUÁN and JUAN S. MORA-SANGUINETTI: Quality of enforcement and investment decisions. Firm-level evidence from Spain.

1928 MARIO IZQUIERDO, ENRIQUE MORAL-BENITO and ELVIRA PRADES: Propagation of sector-specific shocks within Spain and other countries.

1929 MIGUEL CASARES, LUCA DEIDDA and JOSÉ E. GALDÓN-SÁNCHEZ: On financial frictions and firm market power.

1930 MICHAEL FUNKE, DANILO LEIVA-LEON and ANDREW TSANG: Mapping China's time-varying house price landscape.

1931 JORGE E. GALÁN and MATÍAS LAMAS: Beyond the LTV ratio: new macroprudential lessons from Spain.

1932 JACOPO TIMINI: Staying dry on Spanish wine: the rejection of the 1905 Spanish-Italian trade agreement.

1933 TERESA SASTRE and LAURA HERAS RECUERO: Domestic and foreign investment in advanced economies. The role of industry integration.

1934 DANILO LEIVA-LEON, JAIME MARTÍNEZ-MARTÍN and EVA ORTEGA: Exchange rate shocks and inflation comovement in the euro area.

1935 FEDERICO TAGLIATI: Child labor under cash and in-kind transfers: evidence from rural Mexico.

1936 ALBERTO FUERTES: External adjustment with a common currency: the case of the euro area.

1937 LAURA HERAS RECUERO and ROBERTO PASCUAL GONZÁLEZ: Economic growth, institutional quality and financial development in middle-income countries.

1938 SILVIA ALBRIZIO, SANGYUP CHOI, DAVIDE FURCERI and CHANSIK YOON: International Bank Lending Channel of Monetary Policy.

1939 MAR DELGADO-TÉLLEZ, ENRIQUE MORAL-BENITO and JAVIER J. PÉREZ: Outsourcing and public expenditure: an aggregate perspective with regional data.

1940 MYROSLAV PIDKUYKO: Heterogeneous spillovers of housing credit policy.

1941 LAURA ÁLVAREZ ROMÁN and MIGUEL GARCÍA-POSADA GÓMEZ: Modelling regional housing prices in Spain.

1942 STÉPHANE DÉES and ALESSANDRO GALESI: The Global Financial Cycle and US monetary policy in an interconnected world.

1943 ANDRÉS EROSA and BEATRIZ GONZÁLEZ: Taxation and the life cycle of firms.

1944 MARIO ALLOZA, JESÚS GONZALO and CARLOS SANZ: Dynamic effects of persistent shocks.

1945 PABLO DE ANDRÉS, RICARDO GIMENO and RUTH MATEOS DE CABO: The gender gap in bank credit access.

1946 IRMA ALONSO and LUIS MOLINA: The SHERLOC: an EWS-based index of vulnerability for emerging economies.

1947 GERGELY GANICS, BARBARA ROSSI and TATEVIK SEKHPOSYAN: From Fixed-event to Fixed-horizon Density Forecasts: Obtaining Measures of Multi-horizon Uncertainty from Survey Density Forecasts.

1948 GERGELY GANICS and FLORENS ODENDAHL: Bayesian VAR Forecasts, Survey Information and Structural Change in the Euro Area.

2001 JAVIER ANDRÉS, PABLO BURRIEL and WENYI SHEN: Debt sustainability and fiscal space in a heterogeneous Monetary Union: normal times vs the zero lower bound.

2002 JUAN S. MORA-SANGUINETTI and RICARDO PÉREZ-VALLS: ¿Cómo afecta la complejidad de la regulación a la demografía empresarial? Evidencia para España.

2003 ALEJANDRO BUESA, FRANCISCO JAVIER POBLACIÓN GARCÍA and JAVIER TARANCÓN: Measuring the procyclicality of impairment accounting regimes: a comparison between IFRS 9 and US GAAP.

2004 HENRIQUE S. BASSO and JUAN F. JIMENO: From secular stagnation to robocalypse? Implications of demographic and technological changes.

2005 LEONARDO GAMBACORTA, SERGIO MAYORDOMO and JOSÉ MARÍA SERENA: Dollar borrowing, firm-characteristics, and FX-hedged funding opportunities.

2006 IRMA ALONSO ÁLVAREZ, VIRGINIA DI NINO and FABRIZIO VENDITTI: Strategic interactions and price dynamics in the global oil market.

2007 JORGE E. GALÁN: The benefits are at the tail: uncovering the impact of macroprudential policy on growth-at-risk.

2008 SVEN BLANK, MATHIAS HOFFMANN and MORITZ A. ROTH: Foreign direct investment and the equity home bias puzzle.

2009 AYMAN EL DAHRAWY SÁNCHEZ-ALBORNOZ and JACOPO TIMINI: Trade agreements and Latin American trade (creation and diversion) and welfare.

2010 ALFREDO GARCÍA-HIERNAUX, MARÍA T. GONZÁLEZ-PÉREZ and DAVID E. GUERRERO: Eurozone prices: a tale of convergence and divergence.

2011 ÁNGEL IVÁN MORENO BERNAL and CARLOS GONZÁLEZ PEDRAZ: Sentiment analysis of the Spanish Financial Stability Report. (There is a Spanish version of this edition with the same number).

BANCO DE **ESPAÑA**
Eurosistema

Unidad de Servicios Generales I
Alcalá, 48 - 28014 Madrid
E-mail: publicaciones@bde.es
www.bde.es