
11th Biennial IFC Conference on “Post-pandemic landscape for central bank statistics”

BIS Basel, 25-26 August 2022

Using the press as a real-time economic confidence indicator¹

Juan Pablo Cova M, Hugo Peralta V and María del Pilar Cruz N,
Central Bank of Chile

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

Using the press as a real-time economic confidence indicator¹

María del Pilar Cruz N, Hugo Peralta V and Juan Pablo Cova M, Central Bank of Chile.

Abstract

Through the application of textual border analysis tools, this work presents the construction of a high-frequency indicator, generated in real time, based on the computerized reading of the news from the main print media from January 2015 to December 2020 in Chile. This indicator captures the emotional tone of economic and opinion news by making use of an extensive —and novel— purpose-built dictionary in Spanish. This lexicon of words was subjected to challenging statistical tests of robustness, complemented by tests of predictive precision. The latter were carried out by comparing the degree of similarity between a classification by automated means and another by manual means in a random sample. The economic application shows that the constructed indicator has a high correlation with confidence indicators based on surveys, and a high predictive capacity in the face of shock phenomena hitting the economy.

Keywords: in Text mining, nowcasting, dictionary.

JEL classification: C22, C53, C82.

¹ The authors wish to thank Professor Stephen Hansen for his valuable comments during the internal seminar organized by the Central Bank of Chile “Unstructured Data Analysis for Macroeconomics and Monetary Policy”, held on 13 September 2021. We also thank Pablo Filippi and Carlos Medel for their valuable insights to this research, Consuelo Edwards for her editing services, *Observatorio Tecnológico* for their support in the application of text mining tools, Cristián López for his support in hardware services, Valentina Vásquez for her research assistance, as well as Alejandra Rozas, Carlos Madeira, Carlos Medel, Consuelo Edwards, Daniel Pérez, Erika Arraño, Felipe Lozano, Joaquín Pérez, José Manuel Salinas, Josué Pérez, Juan Pablo Ugarte, León Sanz, Luis Felipe Villanueva, Marcos Correa, Mario González, Miguel Musa, Natalia Oyanedel, Pablo Filippi, Sebastián Becerra, Sebastián Silva, whose valuable contribution made it possible to generate a set of manually tagged news articles. The opinions and conclusions expressed in this document are the sole responsibility of the authors and do not necessarily represent the views of the Central Bank of Chile. Emails: mcruz@bcentral.cl, pilarcruz@yahoo.com, hperalta@bcentral.cl, jcova@bcentral.cl. This is a shortened version of the paper, prepared for this Conference. The full version can be found [here](#).

Table of contents

Using the press as a real-time economic confidence indicator.....	1
1. Introduction.....	3
2. Synthesis of related research.....	3
3. Characterizing the news to create a sentiment index.....	5
4. Constructing a dictionary to read the news.....	6
4.1. Methodology for creating a dictionary.....	6
5. Constructing a set of manually tagged news.....	9
6. Constructing a dictionary to read the news.....	9
7. IS-News results and relationship with economic indicators	10
8. Conclusions.....	15
References.....	17

1. Introduction

Since the invention of the printing press, the written media have been an essential mechanism of information, thanks to their ability as a credible source for keeping up to date judgments and visions about the future.

The explosive development of text information by written media has accelerated the need to identify their contents in order to integrate them effectively with economic phenomena. In response to these challenges, computational text analysis has emerged with great force in recent years, through which relevant information can be extracted from unstructured texts and be transformed into structured information. The increase in computer processing capacity has made it possible to create algorithms applicable to language and to considerably reduce processing time and costs.

This technical progress, has in turn, led to important innovations in nowcasting models. Official activity statistics take time to generate and compile, so their availability has lags. The literature in this field (Banbura, Giannone, & Reichlin (2011)) has so far been oriented towards predicting the trajectory of economic variables through the combined use of activity indicators with business and consumer survey indicators. However, more recently, studies such as Kalamara, Turrell, Redl, Kapetanios, & Kapadia (2020) and Thorsrud (2016) have proven that the performance of these models can be improved by considering the news in the press, due to its real-time availability and its predictive value in periods of economic stress.

The main contribution of this work focuses precisely on generating a high-frequency indicator built in real time based on the computerized reading of the news in Chile, with predictive capacity about the evolution of economic conditions, the business cycle and the trajectory of confidence levels in the economy. It uses a database of about 935 thousand pieces of news contained in six major newspapers in the country for the period 2015-2020.

Another contribution is the generation of a Spanish dictionary built from the same news database, which is extensive in the number of tagged words, comparable to its English similes, complex in its variety of grammatical forms, tested on a sample of manually tagged news articles and robust to the statistical testing on news.

The paper is organized as follows. Section 2 contains a synthesis of the related research; section 3 presents a characterization of the IS-News and its transformation into a database suitable for text mining; section 4 presents the construction of the dictionary for reading the news; section 5 presents the construction of a manually tagged set of news; section 6 describes the construction of the IS-News indicator; section 7 shows the results of the IS-News and their relationship with economic indicators, to close with the main conclusions in section 8.

2. Synthesis of related research

Sentiment Analysis is a text mining tool that allows to obtain qualitative data in real time, without resorting to population surveys. It is a field of research that performs a computational treatment of opinions and feelings contained in texts. Its specific application on printed news is relatively recent and its accelerated diffusion is

explained by its contribution to economic analysis and modelling, its ability to anticipate changing conditions and its low production costs compared to surveys (Shapiro, Sudhof, & Wilson (2020)).

From these text analysis tools, two areas of research can be identified that address different objectives: one that detects “intensity” and one that detects “tone”. Intensity measures are based on the number of times certain words appear in the text, and has been extensively disseminated by Baker, Bloom, & Davis (2016) through the Economic Policy Uncertainty (EPU) index. This index counts the number of times words such as uncertainty or recession appear in the analyzed text, similar to what has been done by Altig et al. (2020), Cerda, Silva, & Valente (2016) and Becerra & Sagner (2020). The tone measures, on the other hand, called “Sentiment Analysis” capture the underlying sentiment of optimism or pessimism in texts, using broad tagged lexicons and also automated or machine learning tools. The research area analyzing this paper detects the tone of the texts in the news.

Due to the relevance of using databases with a marked tonal polarity, an important part of sentiment analysis research uses documents with a high subjective and judgmental content, as is the case with blogs, social networks, or product reviews. The press, meanwhile, does not contain this same density of tone as it tries to give the impression of objectivity (Balahur et al. (2013)). In these types of texts, the recount of judgments and polarity is often found in the form of third-party opinions or else in news stories that call upon the opinion of third parties. The editorial line may choose to emphasize or moderate the final texture of the message, but ultimately results in a discourse structure that is more complex to analyze. Therefore, success in capturing the polarity of a news article depends on correctly identifying the sentiment implicit in it (“opinion mining”), isolating the sign of the news piece itself (Sabeti & Saad (2017)).

Sentiment analysis saw its first publications in the early 2000s and experienced very accelerated growth a few years later. Using Google Scholar and Scopus citation counts, Mäntylä, Graziotin, & Kuuttila (2018) note that in 2000 there were only 37 publications in this field, rising to around 7,000 in 2016. The authors conclude that what enabled this dizzying boom has been the possibility of analyzing huge volumes of texts with computational text mining tools. Some of the research that seems most relevant to us is reviewed next.

Work published by the Federal Reserve of San Francisco (Shapiro, Sudhof, & Wilson (2020)) develops a time series that captures sentiment derived from news stories drawn from economic newspapers between 1980 and 2015. By generating a model that combines various internationally recognized labelled lexicon, they show that daily news sentiment is a good predictor of survey-based confidence indices.

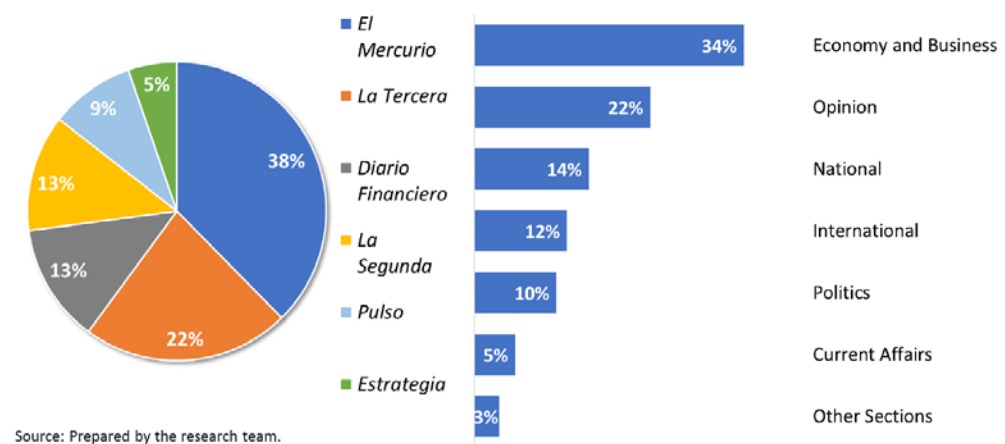
Research by Larsen & Thorsrud (2015) confirms the widespread belief that changes in expectations caused by news is an important autonomous driver of economic fluctuations. Using the main Norwegian business newspapers, the authors identify the topics with the highest predictive power, from which they construct an aggregate index and show that unexpected changes in the index cause significant and persistent fluctuations in markets, especially in credit ones.

Finally, Cruz, Peralta, & Ávila (2020) used computational linguistics to analyze the Business Perceptions Report of the Central Bank of Chile (IPN) and generated an index with high correlations with business and economic confidence indicators.

3. Characterizing the news to create a sentiment index

The formulation of the IS-News is based on the reading and processing of the main printed daily newspapers in the country, of national and international coverage, namely: *El Mercurio*, *La Tercera*, *Pulso*, *La Segunda*, *Diario Financiero*, and *Estrategia*². Regional print media³ were excluded because their coverage is not nationwide, as well as those with lower periodicity (weekly or bimonthly), because the index is based on high-frequency data. The database for the construction of the index comes from information generated by *NexNews*.

Figure 1: News by media and sections, 2015-2020 (filtered base, % of total news by media)



The number of news articles in the selected media totals 935 thousand for the period under study (between 2015 and 2020). Two events of great impact on the Chilean economy are included in this period: the social crisis of October 2019 and the arrival of the pandemic in March 2020. This base was purged (see section 3.1), reducing its size to 417 thousand news articles, which include the sections Economy and Business, National, International, Politics, Current Affairs and Opinion (Figure 1).

In this new base, the "Economy and Business" and "Opinion" sections are the most relevant, as they account for an average of 56% of the total news, while the "National", "International" and "Politics" sections each represents around 12%. By media, news in *El Mercurio* and *La Tercera* account for 60% of the total base, while *Diario Financiero*, *La Segunda*, *Pulso*, and *Estrategia* make up the remaining 40% (Figure 1).

² In late 2016, this paper began publishing only on digital media, so from then on it is deleted from the database.

³ Relevant media to construct databases for regional sentiment indexes.

4. Constructing a dictionary to read the news

The most used techniques to perform Sentiment Analysis (SA) are, on the one hand, Machine Learning techniques, which encompass a wide range of statistical models capable of learning from massive databases and generate text tone predicting models. This technique's predictive potential makes it particularly advantageous, but its quality depends on the training set, which needs to be voluminous and complex to learn about the unit lexicons of the language, as well as about both simple and complex sentences.

A second group of commonly used techniques to perform SA contains those that use previously tagged lexicons. This approach is based on a selection of terms that reflect a clear sentiment orientation and can be tagged. Unlike Machine Learning, it is not based on algorithms, but on semantic dictionaries containing terms classified with valences. Words with positive valences are used to collect desired states, and those with negative valences are used to collect undesired ones. In this way words that are contained in the tagged lexicon are detected and added together according to the corresponding polarity. These are the techniques used in this research.

4.1. Methodology for creating a dictionary

The methodology for creating a dictionary in Spanish consisted of several sequential stages, all of them referenced to news databases in Spanish. No other words from other Spanish dictionaries or from translated English dictionaries were included. In other words, the tagged terms were obtained entirely from the printed news themselves, with the purpose of having a domain-specific dictionary.

Briefly, the stages were the following:

- Use of the *Python spaCy* library, which helps identify the grammatical class⁴ and the lemma of each word, as well as recording the publication where the news appears. With this processing, the size of the original database is multiplied roughly by 100⁵, because each word contains a new set of linguistic information.
- The base was subdivided by randomly selecting⁶ news articles equivalent to 10% of the total base (47 thousand news), in the form of a sampling without replacement⁷. The processing of the first random sample of news pieces identified a total of 9,596 terms and the second random sample, when compared with the first, added barely 325 new terms (3.4%). For this reason, a third sampling was ruled out due to its low probability of adding any new terms. Thus, with 90 thousand randomly collected news articles, the universe of terms originating the tagged dictionary was covered.

⁴ For example, verbs, adjectives, adverbs, nouns, and determiners, plus names of entities or locations.

⁵ The complete news base, including functionalities added by *spaCy*, is around 100 GB in size.

⁶ For the sample to be representative, the random choice includes the dimension of the media, sections, and dates, and it is made in a stratified way, that is, preserving their proportions in the full base.

⁷ Sampling without replacement is that which is carried out without taking back to the base those news articles that were chosen to build the sample.

- Identification of the terms (verbs, adjectives, and adverbs) in the random samples is done with the functions of lemmatization and computational POS tagging by *spaCy*. The complete universe of unique words found in the two samples was 9,921, of which 7,616 were selected for tagging, as 2,305 were removed because of their low frequency⁸.
- Upon completing the process described above, the next step is to tag the set of unique words with positive, negative, or neutral tone by each member of the research team. Following international recommendations (Balahur et al. (2013)), the tagging criteria are standardized, in the sense that only those words with a clear tonality are appraised, and that those terms with two or more diverging meanings or with a diffuse connotation are tagged as neutral.
- Upon completing of the manual tagging process, it was decided to maintain the tags of the words that posted 100% coincidence while, in the cases that presented partial coincidences, it was decided to maintain them only if the discrepant valuation was not the opposite.
- For the rest, where the divergence was bigger, the team proceeded to re-evaluate word by word, analyzing the contexts in which they were used.

The outcome of this methodology was the generation of a tagged dictionary⁹ with valences different from zero, based on printed news, with a total of 5,419 lemmas with their inflections¹⁰, out of which 374 are unique lemmas (257 verbs, 67 adjectives and 50 adverbs) (figure 2).

Figure 2: Main tagged complex dictionaries (with positive or negative valence)

DICCIONARY	FEATURE SPACE	LANGUAGE	LEXICON SIZE (TAGGED WORDS)
Hu-Lui (HL)	Movie Reviews	English	6,789
IS News	News Articles	Spanish	5,419
Harvard General Inquirer (GI)*	General English	English	4,206
Loughran-McDonald (LM)*	Financial Reports	English	2,683
Business Perception Report Central Bank of Chile (BPR CBCh)	Economic and Business Reports	Spanish	774
Financial Stability Report Bank of Spain (FSR BoS)	Economic and Business Reports	Spanish	565
Financial Stability Report Central Bank of Chile (FSR CBCh)	Economic and Business Reports	Spanish	361

* These dictionaries contain additional categories to positive and negative, which means that words can belong to several other different categories such as "degree of uncertainty", "power", "strength", among others. In the LM there are about 1.553 words in other categories which gives rise to a total of 4.236 words tagged in all categories. The whole dictionary contains 26 categories and a total of 11.788 words tagged in all categories. Note that a same word can be counted several times; the word "about" is counted seven times because it is tagged in seven different categories.

⁸ Words with very low frequency, present in fewer than six pieces of news in the sample, may represent an error or be irrelevant in the indicator's calculation.

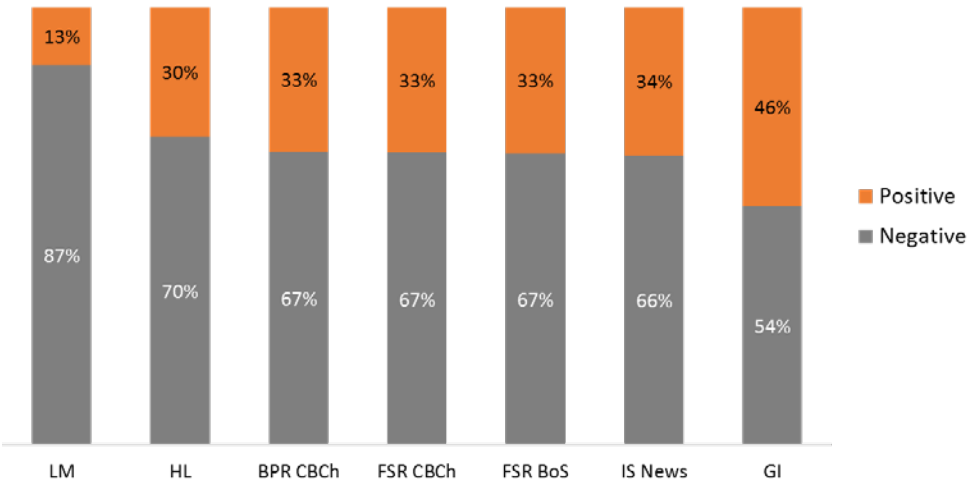
⁹ The dictionary can be found on [here](#).

¹⁰ Inflections are conjugations, plurals, verb tenses, and others. Each verb in the dictionary has associated an average of 20 inflections, for example: *aumentar*, vs *aumentado*, *aumentando*, *aumentó*, *aumentará*, *aumentaría*, and so on.

The extension of this dictionary compares favourably with other English-language dictionaries, popular in SA, such as the Harvard General Inquirer (GI), which has a length of 4,206 words and is a general-purpose dictionary developed by the Harvard University. It also compares positively with the Loughran-McDonald (LM), which is a little smaller than the GI, with a total of 2,683 words, and which has the particularity of being for the specific domain of economics and finance. The Hu-Lui (HL) dictionary, on the other hand, has a total of 6,789 terms, but because it is created from magazines with movie reviews, it limits its use to the economic-financial area¹¹.

The analysis of the result of negative and positive valences in each of the aforementioned dictionaries reveals that negative tags in IS-News represent 66% of the total number of words with non-neutral valences, similar to what is found in the other two Spanish dictionaries (IPN and IEF), and in the HL (figure 3). The prevalence of negative over positive tags is consistent with the findings of Reis et al. (2015) (section 2).

Figure 3: Tagged words in selected dictionaries (with positive or negative valence)



Yet, it is necessary to bear in mind some limitations associated with its effectiveness to perform Sentiment Analysis (Mechulam Burstin & Salvia Varela (2018)). For instance, correctly identifying and tagging the tone of the words, paying attention to the contexts in which they are used. Indeed, a given word may have a neutral sentiment orientation in an economic context, but negative in a scientific or legal one, or vice versa. Using a sample of corporate financial reports (10-K¹²) in the United States between 1994 and 2008, Loughran & McDonald (2011) show that nearly three-quarters of the words identified as negative by the GI dictionary are typically not considered negative in financial contexts, such as liability, depreciate, tax, or cost.

¹¹ This comparison must take into consideration that the derivations or inflected forms in English are less numerous than in Spanish.

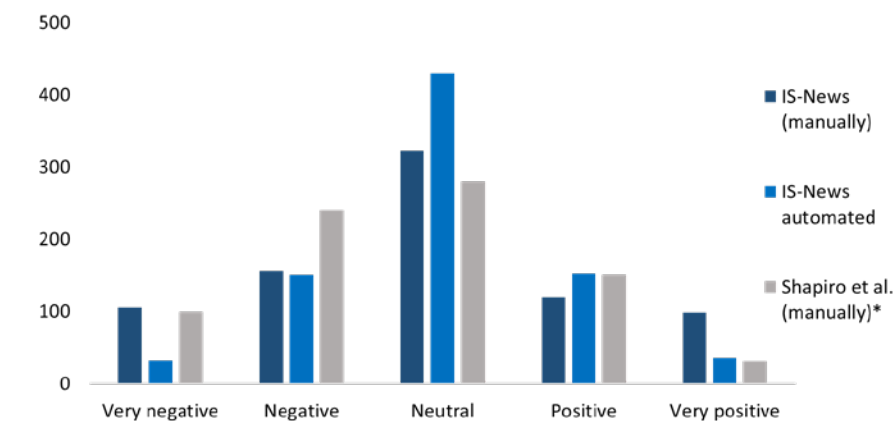
¹² 10-K is a yearly financial report of companies listed in the U.S. stock exchange, containing more details than the one that is delivered every year to stockholders, and is a requirement of the U.S. SEC.

5. Constructing a set of manually tagged news

With the objective of evaluating the predictive capacity of the IS-News dictionary, a manual tagging procedure is being performed on a set of news articles, like the one by Shapiro, Sudhof, & Wilson (2020). This procedure is carried out by surveying a group of 23 CBCh researchers, who manually classify 840 news articles according to the polarity of sentiment they detect from their reading. The 840 news articles are randomly selected from the database used by the research.

Figure 4 shows the histogram of manual classifications, automated news ratings and a comparison with that obtained by a similar manual tagging exercise in Shapiro, Sudhof, & Wilson (2020). The results show that the highest frequency of ratings occurs in neutral news (particularly in the automated rating), although with a bias towards negative categories consistent with international evidence, which shows a greater emphasis on reporting negative news.

Figure 4: Compared histogram of news set tagging (IS-News tagged manually vs automated)



* Adjusted to histogram scale in Shapiro et.al (2020)

6. Constructing a dictionary to read the news

The central objective of this research is to construct a time series, of daily frequency, based on printed news, using the text mining method called Sentiment Analysis.

Two methodologies commonly found in international research are used to calculate the series. One, which we have termed IS-News^{Dicc}, is based exclusively on the use of the dictionary and manual tags assigned by the researchers. The other, which we have termed IS-News^{PMI}, uses a corpus lexicon developed from the dictionary and introduces a measure of probability that a given word is associated with a given sentiment ("Pointwise Mutual Information, PMI").

The most traditional method for generating a sentiment index is based on using a tagged dictionary, which means that the polarity of each word is determined solely by the dictionary. In order to produce the IS-News^{Dicc} series, the following procedures are required:

To create an algorithm to detect polarity for each news article, which adds words with positive and negative tags. Each text is scrolled through, identifying verbs, adjectives, or adverbs, and then looked up in the dictionary. Sentiment words can take the values 1 (positive tone) or -1 (negative tone), modifying words can take values of 1.5 (intensifies the tone) or 0.5 (attenuates the tone) and negation words, a quotient of -1, which reverses the polarity of the next three contiguous words.

To assign tags at the level of each sentence, for which those with assigned polarity are added together, modifiers and negations are applied, and the result is divided by the total number of words in the news article, to avoid longer stories having a greater weight on the sentiment index.

By contrast, *Pointwise Mutual Information* (PMI¹³) methodology provides a measure of the probability that a given word is associated with a given sentiment. Thus, the value assigned manually to each word according to the dictionary is adjusted based on the usage detected throughout the news corpus. For example, if a positive word is found more frequently in positive news, its value will be amplified, whereas if it is more present in negative news, its value will be diminished. Consequently, the application of the PMI algorithm means that the sentiment associated with a news item depends not only on the individual tag of each word, but also on its relative frequency in the rest of the text.

7. IS-News results and relationship with economic indicators

To evaluate the usefulness of the IS-News high frequency indicator, we explore its ability to anticipate confidence indicators such as the Economic Perception Index (IPEC¹⁴), the Monthly Business Confidence Index (IMCE¹⁵) and their respective sub-indexes, the Business Confidence Index (ICE¹⁶), as well as various activity data, in order to identify those with which it is best related. Also, we explore its capacity to anticipate shocks in the economy, whose occurrence over time is also examined.

In addition, we identify which IS-News calculation formula delivers a higher predictive value, i.e., whether the one that uses exclusively the dictionary (IS-News^{Dicc}) or the one that uses the dictionary-based corpus lexicon (IS-News^{PMI}). The results for the period 2015-2020 show the following:

¹³ Re-weighting with PMI is an approximation to methods known as vector space models.

¹⁴ Economic Perception Index (IPEC in Spanish) produced by GfK Adimark Chile. It is calculated by the monthly application of a structured questionnaire to a sample of 1100 persons 18 years old or more, residing in the main cities of Chile.

¹⁵ Monthly Indicator of Business Confidence (IMCE in Spanish) produced by ICARE and Adolfo Ibañez University. It is a synthetic index that is applied to 607 companies in four sectors (Industry, Mining, Trade and Construction) reflecting the weighted sum of those indicators for each sector.

¹⁶ Business Confidence Index (ICE in Spanish) produced by Del Desarrollo University. It measures economic perception from businessmen point of view. It is elaborated from surveys to approximately 300 general managers, business owners or executives, through telephone surveys or emails.

- The Pearson correlation coefficient¹⁷ presents, without exception, significantly higher values when IS-News is calculated using the PMI methodology. This is consistent with the findings of Shapiro, Sudhof, & Wilson (2020) in which the predictive accuracy of the indicator using PMI metrics is better than that generated with any other of the available on-the-shelf dictionaries.
- Correlations rise almost widely when they are lag behind confidence and activity indicators in about four weeks, and they tend to decline when the lag grows to two months, or they are measured contemporaneously.
- The highest correlations of the IS-News^{PMI} were found, with a one-month lag, with the Economic Perception Index and its two subindexes, "situation to purchase home items" and "perception of the firms' current economic situation", which yield Pearson coefficients of 83%, 92%, and 85%, respectively (figure 7). In other words, the IS-News^{PMI} has the best fit with indicators that reflect the population's perception of the present and future economic situation as well as of the propensity to consume.
- Regarding the Economic Perception Index, the IS-News succeeds in capturing with intensity the two shocks occurring in the period analyzed, i.e., the social outbreak of October 2019 and the arrival of the pandemic in Chile in March 2020. Figure 8 focuses specifically on the period in which these shocks occurred (the complete series is presented in figure 7), showing that both events begin to manifest in the IS-News approximately in the chronological month prior to the one that appears in the survey results. The Pearson coefficient for this specific period rises to 92% for the Economic Perception Index and to 95% for the sub-index "situation to buy household items". Something similar occurs during the recovery phase, where the IS-News recovers its level before the confidence indicators do so.

¹⁷ A measure of the strength of the linear dependence between two quantitative random variables, irrespective of the scale on which they are displayed. It is a good metric when the samples are large, and its distribution follows a normal curve.

Figure 7: IS-News^{PMI} vs indicators of confidence in the economy (IPEC): 2015-2020 (graphic of IPEC with lag +1; Pearson's R calculated using IPEC+1)

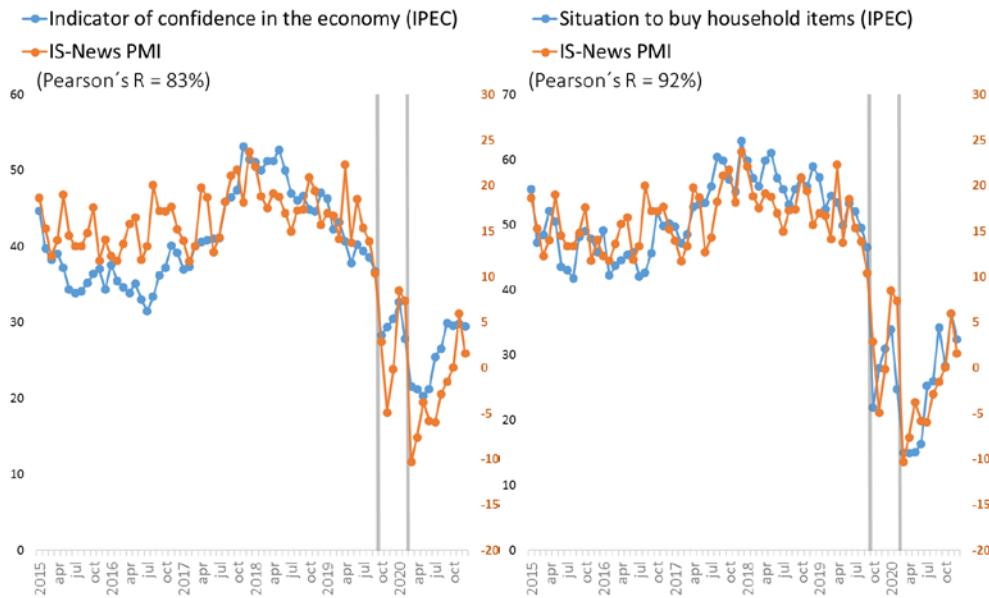
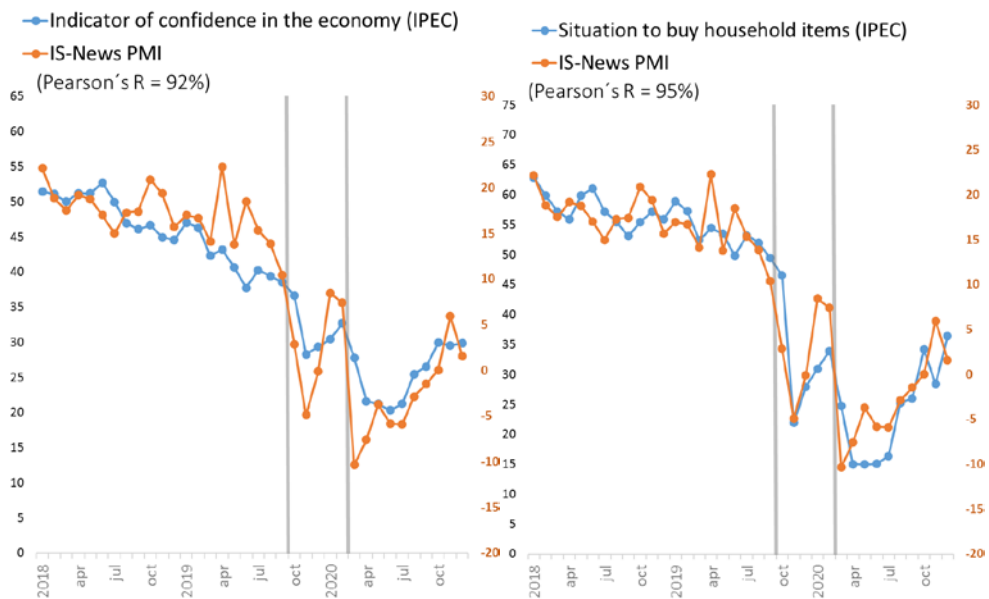


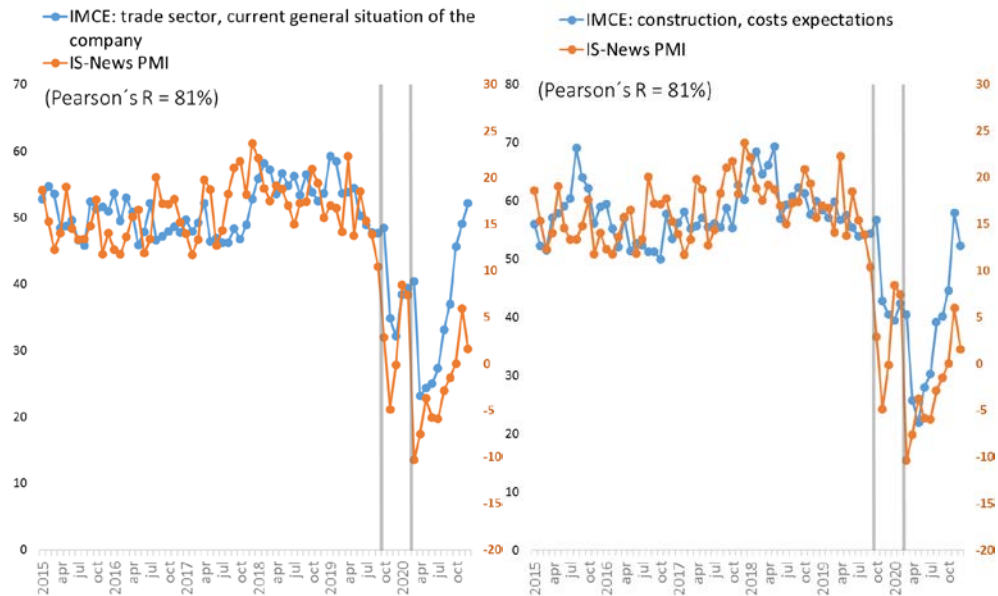
Figure 8: IS-News^{PMI} vs indicators of confidence in the economy (IPEC): 2018-2020 (graphs show contemporaneous series; Pearson's R calculated using IPEC+1)



- When the IS-News^{PMI} is analyzed relative to indicators linked to business confidence surveys, the correlations are weaker than with the Economic Perception Index (IPEC). In fact, the correlation coefficient measured against the Monthly Business Confidence Index (IMCE) and its sectoral indexes, ranges between 50% and 60% and the Business Confidence Index (ICE) stands at 73%. Only a couple of exceptions can be noted in some very specific IMCE sub-indexes in which the Pearson coefficient is higher, one being "general current situation of the firm" in the trade sector, and another being "expected costs" in the

construction sector in which, for both, the correlation coefficient climbs to 81% (figure 9).

Figure 9: IS-News^{PMI} vs business confidence indicators (IMCE): 2015-2020 (contemporaneous series; Pearson's R calculated using IMCE+1)



- When measuring the correlations of IS-News^{PMI} with economic activity indexes, it appears that they are also somewhat lower than those observed with the confidence-in-the-economy index. The Pearson coefficient marks 78% with the headline Imacec, 79% with the non-mining Imacec, and 83% with the services Imacec, all of them displaced one month ahead with respect to the IS-News. In the particular case of the Trade Imacec, the correlation coefficient becomes very low (46%), which has to do with the impulse on consumption during the economy's recovery phase of the pandemic (second half 2020), which is factored into neither the IS-News nor the Economic Perception Index (IPEC) and sub-indexes, which show a rather minor rebound (figure 10).

Figure 10: IS-News^{PMI} vs trade Imacec and confidence-in-the-economy indicator: 2018-2020 (contemporaneous series; Pearson's R calculated using contemporaneous series)

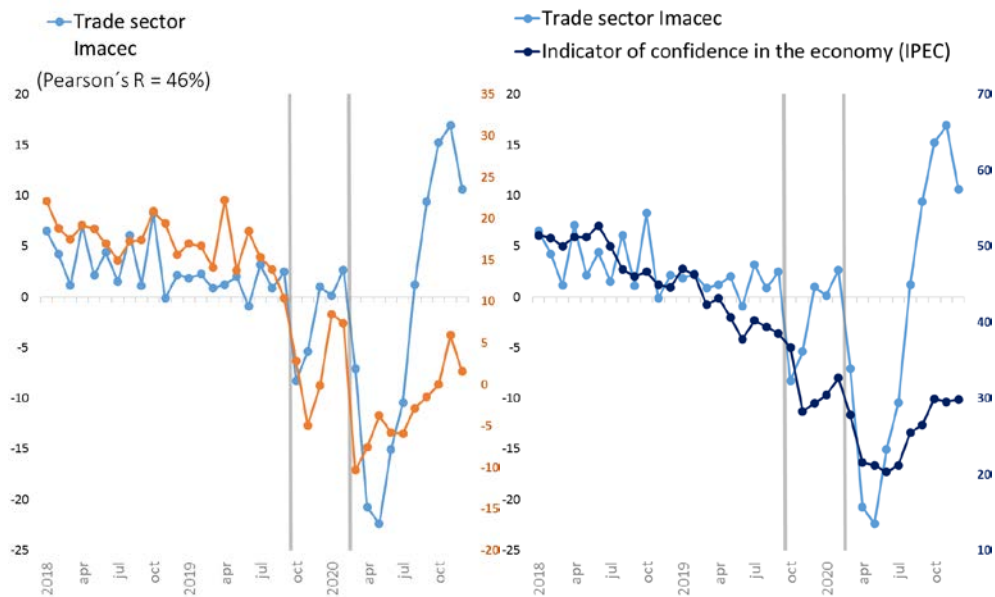
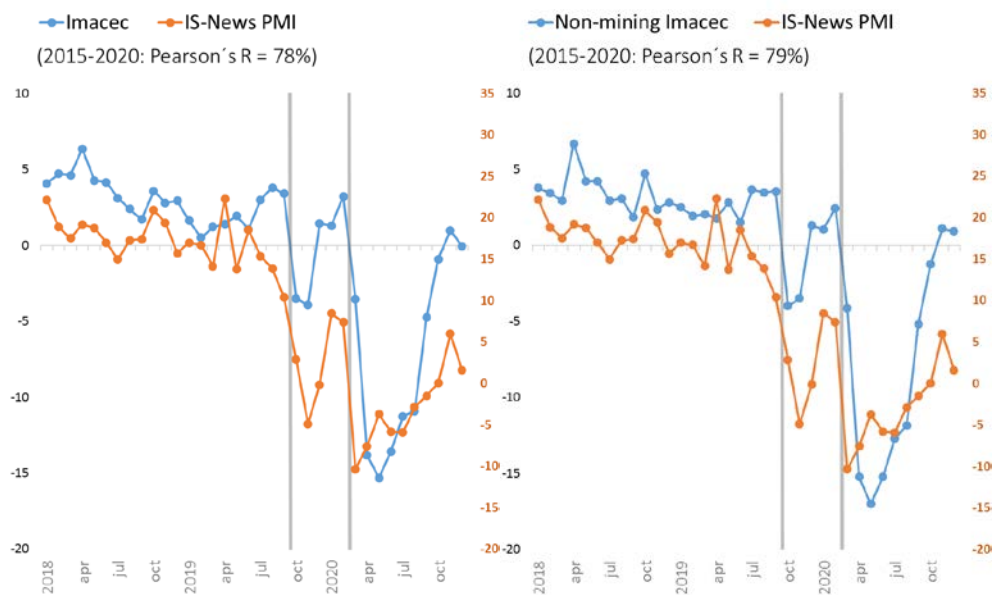


Figure 11: IS-News^{PMI} vs Imacec and non-mining Imacec: 2018-2020 (contemporaneous series; Pearson's R calculated using Imacec+1)

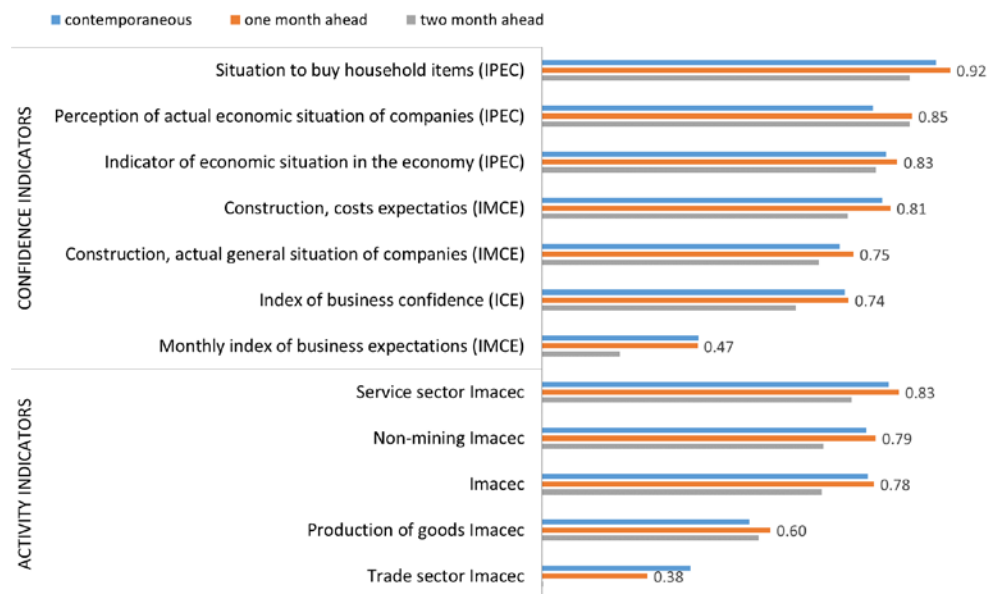


- As with the confidence indicators, the IS-News^{PMI} reflects the impact of the social crisis approximately one month in advance of the Imacec, as is also the case with the sanitary crisis. This short-term predictability condition of the IS-News couples with its real-time availability, which ultimately makes it possible to anticipate by roughly two months what the activity indicators will show (figure 11).

In short, the news index developed for the period 2015-2020 shows a high correlation with indicators of citizens' confidence regarding their personal and the country's situation. Its best fit is achieved with expectations indicators referring to consumption intentions, which depend on greater or lower confidence in the conditions of the economy and their own income, and with their perceived economic stability level that allow sustaining said intended consumption over time (figure 12).

This result is consistent with international research findings, which show that the news contributes to the formation of individuals' expectations, which in turn generate information flows that feed the perceptions surveys. Thus IS-News is capable of recording clearly and in advance the confidence shocks that will be later shown in confidence surveys.

Figure 12: IS-News^{PMI} vs activity and confidence indicators: 2015-2020 (confidence indices in bases; activity indices in 12-month percent variations; Pearson's R)



8. Conclusions

This research uses the latest methodologies to perform sentiment analysis on voluminous news databases. By computationally reading the country's leading economic and opinion newspapers, we have generated a time series that shows a high correlation with confidence-in-the-economy indicators, and somewhat lower correlations with indicators of economic activity.

The goodness of the results obtained is grounded on the quality of the database constructed, and on the use of a Spanish dictionary, generated from the same news database, extensive in the number of tagged words, complex in its variety of grammatical forms, tested on a sample of manually tagged stories and robust to statistical testing. This Spanish dictionary is an important contribution to text mining

research in Spanish, not only because of its size, but also because it allows structuring in various ways the information coming from the written press.

The application of the PMI methodology succeeds in bringing significant improvements to the time series, showing the relevance of incorporating contexts into the process of formulating a manually tagged dictionary. In this sense, it corrects the weaknesses of sentiment analysis methods based solely on dictionaries, allowing the use of a dynamic lexicon, which feeds back as the news base keeps expanding.

In any case, it is important to emphasize that this research made use of the complete database to calculate the PMI, meaning that the valuation of the tagged dictionary words was adjusted according to the usage detected in the entire corpus of news articles. During the period considered, there were two major negative shocks, namely the social outbreak and the pandemic, which may have generated some bias in the application of the PMIs. Therefore, it is proposed here as part of some complementary research, to calculate the optimal and moving PMIs for a time series that uses tagged dictionaries.

The high correlations yielded by the results confirm the international evidence, in the sense that it is possible to computationally capture in real time the level of optimism or pessimism present in the news. There is also evidence that this way of capturing information makes it possible to anticipate confidence shocks, such as those that occurred in 2019 and 2020 in the Chilean economy.

The advantages of having a real-time press indicator are diverse. It provides information that is independent of other sources, has low implementation costs once the methodology has been installed, and can serve as a warning signal of an event of internal or external shock. Moreover, it enables additional applications such as topic analysis, to interpret the unfolding phenomena that accompany the business cycle, as well as the use of bags of words, to obtain intensity measurements.

This notwithstanding, the high correlations between the IS-News and the confidence indicators do not allow inferring that these variables are mutually affecting themselves, meaning that the consumption decisions are affected by the news and vice versa. This would be part of a complementary investigation to this paper.

Finally, the production of the IS-News can serve as a basis for the construction of a model that estimates the way in which sudden confidence-disrupting events detected through the press alter investment and consumption behavior. The early identification of individuals' optimism or pessimism through the IS-News is an intermediate stage between the release of information by the media and the construction of a predictive model about the behavior of economic variables. Modeling this information also represents a way of extending this research and adding value to the fact of having structured information collected from the press.

References

- Altig, D., Baker, S. R., Barrero, J. M., Bloom, N., Bunn, P., Chen, S., . . . Thwaites, G. (2020). Economic uncertainty before and during the COVID-19 pandemic. *Journal of Public Economic*, 191, 104274.
- Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *Quarterly Journal of Economics*, 131(4): 1593–1636.
- Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., . . . Belyaeva, J. (2013). Sentiment Analysis in the News. *Proceedings of the 7th International Conference on Language Resources and Evaluation*, (pp. pp. 2216-2220). Malta.
- Banbura, M., Giannone, D., & Reichlin, L. (2011). Nowcasting. In M. P. Clements, & D. F. Hendry, *The Exford Handbook of Economic Forecasting*. New York: Oxford University Press.
- Becerra, J. S., & Sagner, A. (2020). Twitter-Based Economic Policy Uncertainty Index for Chile. Central Bank of Chile, Working paper N°833.
- Cerda, R., Silva, Á., & Valente, J. (2016). Economic Policy Uncertainty Indices for Chile. Latin American Center for Economic and Social Policies (CLAPES-UC), Working Paper.
- Cruz, M. d., Peralta, H., & Ávila, B. (2020). Análisis de Sentimiento Basado en el Informe de Percepciones de Negocios del Banco Central de Chile. Banco Central de Chile, Documento de Trabajo N°862.
- Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2020). Making Text Count: Economic Forecasting Using Newspaper Text. Bank of England, Staff Working Paper No. 865.
- Larsen, V., & Thorsrud, L. (2015). The Value of News. BI Norwegian Business School, CAMP Working Paper Series No 6.
- Loughran, T., & McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, Volume 66, pp. 35-65.
- Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers. *Computer Science Review*, Volume 27, pp. 16-32.
- Mechulam Burstin, N., & Salvia Varela, D. (2018). Construcción de Léxicos Dinámicos para el Análisis de Sentimientos. Universidad de la República, Uruguay. Tesis de grado.
- Reis, J., Benevenuto, F., Vaz de Melo, P. O., Prates, R., Kwak, H., & An, J. (2015). Breaking the News: First Impressions Matter on Online News. *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, (pp. 357-366).
- Saberi, B., & Saad, S. (2017). Sentiment Analysis or Opinion Mining: A Review. *International Journal on Advanced Science, Engineering Information Technology*, pp 1660-1666.
- Shapiro, A. H., Sudhof, M., & Wilson, D. (2020). Measuring News Sentiment. Federal Reserve Bank of San Francisco , Working Paper 2017-01.
- Thorsrud, L. A. (2016). Nowcasting Using News Topics. Big Data Versus Big Bank. *Advances in Short-Term Forecasting*. ECB Workshop. Frankfurt am Main.

USING THE PRESS AS A REAL-TIME ECONOMIC CONFIDENCE INDICATOR*

AUGUST 26, 2022

RESEARCH TEAM:

María del Pilar Cruz N.

Hugo Peralta V.

Juan Pablo Cova M.

(*) The views and conclusions presented in this paper are exclusively those of the author(s) and do not necessarily reflect the position of the Central Bank of Chile or of the Board members.

ELABORATION OF A NEWS SENTIMENT INDEX

OBJECTIVE:

- Additional source complementing economic activity analysis.

ADVANTAGES:

- Real time data.
- High predictability of the economic cycle.
- Low costs compared to surveys.
- Larger population coverage.
- High effectiveness in changing conditions.

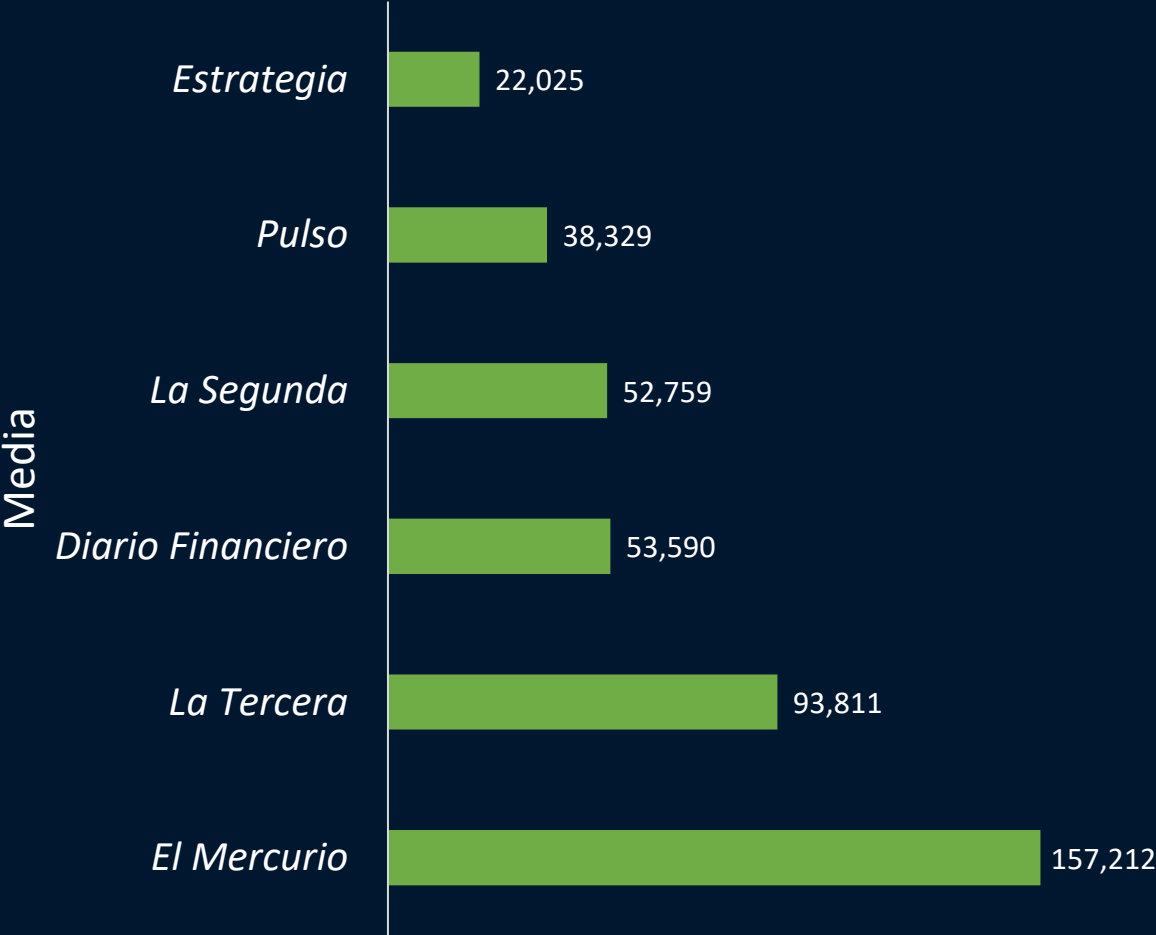
TRANSFORMING DATA INTO (DENSE) INFORMATION USEFUL FOR ANALYSIS

- News database includes main Chilean economic and financial newspaper companies.
- About 935K news articles from 2015 to 2020 were obtained, which were reduced to 417K after the text cleaning process.
- NLP functions from the *spaCy* library in *Python* were used to facilitate the construction of the dictionary and calculation of the sentiment indicator IS-News.

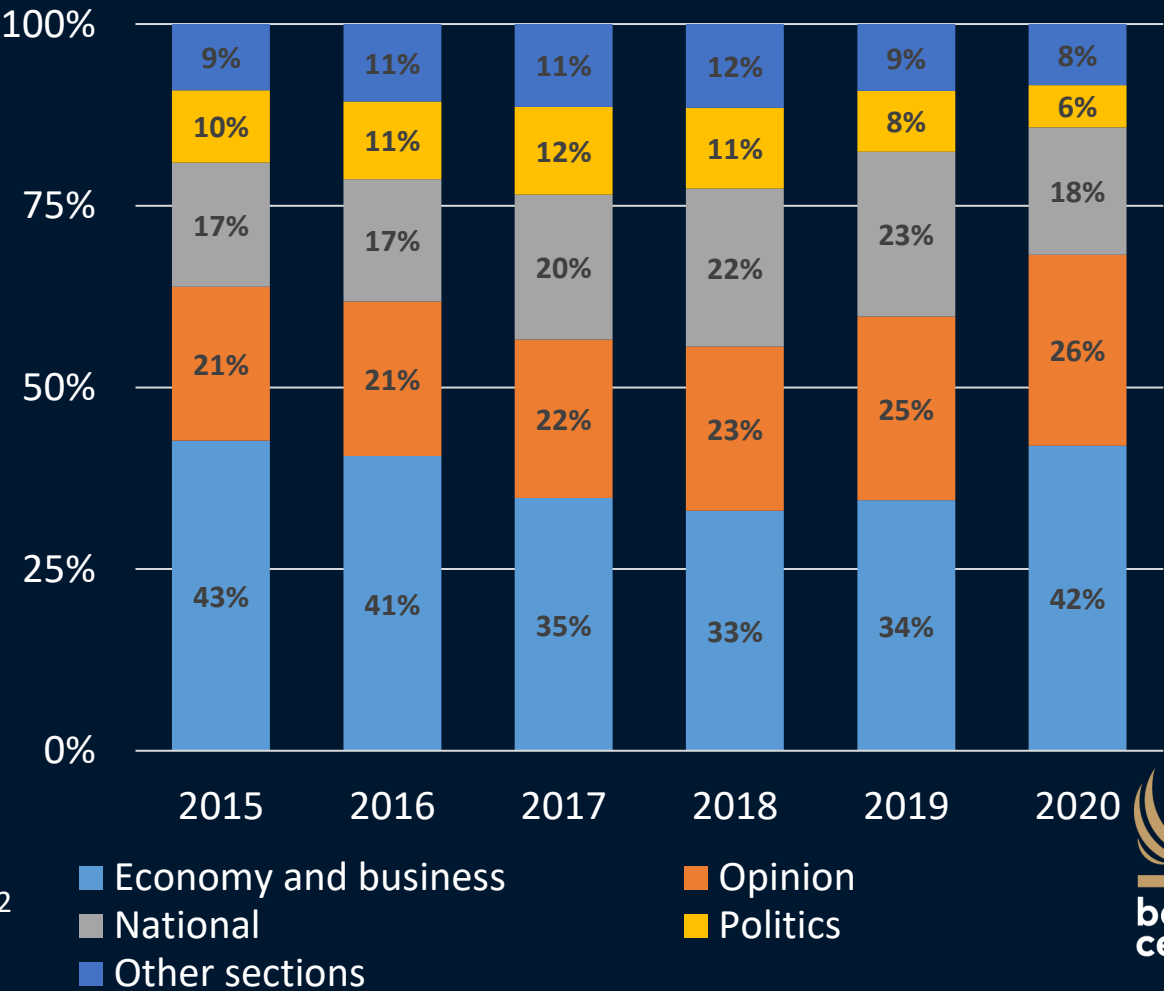


NEWS BY MEDIA AND SECTIONS: 2015-2020

NEWS BY MEDIA: 2015-2020
(filtered base; N° of news articles)



NEWS BY SECTIONS
(filtered base; % of total news articles)



A DICTIONARY

TO READ

NEWS ARTICLES



AN EXTENSIVE NEWS DICTIONARY WAS CREATED

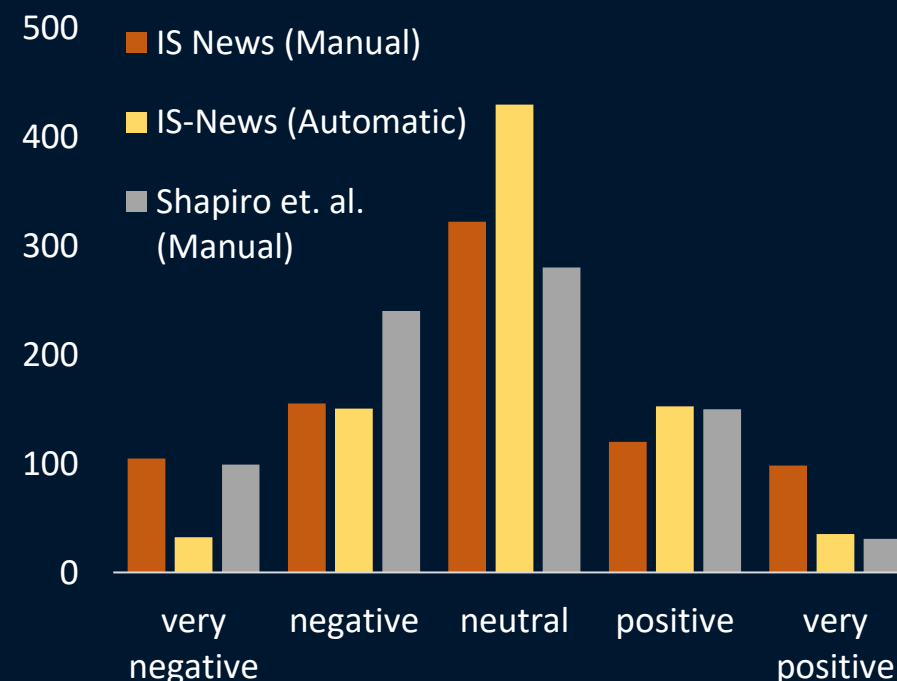
6

- **7,617 unique words were identified from the data** (2,832 verbs, 4,157 adjectives and 628 adverbs)
- **374 words were labeled with 1, -1, 0.5 or 1.5**, equivalent to 5,419 words in their various inflected forms.
- **A random set of news items was manually classified:** testing with automatic labeling; similar results were obtained (Shapiro et. al., 2020).

SIZE OF MAIN LABELED DICTIONARIES

Dictionary	Domain	Language	Labeled words
Hu-Lui (HL)	Movies reviews	English	6,789
Is News	Chilean news articles	Spanish	5,419
Harvard General Inquirer (GI)	General English	English	4,206
Loughran-McDonald (LM)	Companies Financial Statements Report	English	2,683
Business Perceptions Report (BPR) Central Bank of Chile (CBCh)	BPR BCCh	Spanish	744
Financial Stability Report (FSR) Central Bank of Spain (CBS)	FSR CBS	Spanish	565
FSR CBCh	FSR CBCh	Spanish	361

IS-NEWS HISTOGRAM: COMPARING MANUAL V/S AUTOMATIC LABELING



ELABORATING THE IS-NEWS:

ONE INDICATOR - TWO CALCULATION METHODOLOGIES

IS-NEWS WITH DICTIONARY:



- WORD = LABELED BY DICTIONARY

- NEWS SENTIMENT INDEX (ISN):

$$ISN = \frac{\left[\frac{\#positive}{\{1\}} + \frac{\#negative}{\{-1\}} \right] * \frac{\#modifiers}{\{0.5; 1.5\}}}{\#total \text{ words within one piece of news}}$$

- DAILY NEWS SENTIMENT INDEX (ISD):

$$ISD = \frac{\sum ISN}{\#news \text{ per day}}$$

WITH POINTWISE MUTUAL INFORMATION (PMI)



- WORD = PMI

$$PMI(w_i, c) = \log \left(\frac{p(w_i, c)}{p(w_i) * p(c)} \right)$$

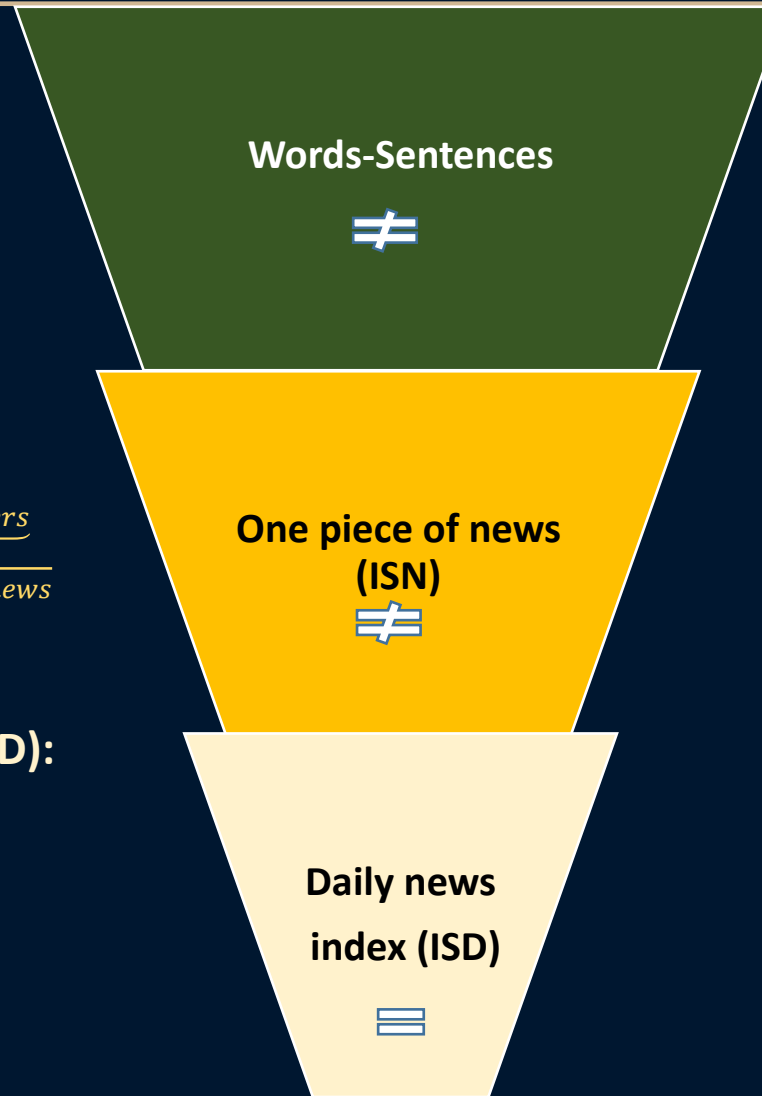
- NEWS SENTIMENT INDEX:

$$ISN^{PMI} = \overline{S(w)}$$

$$S(w_i) = PMI(w_i, POS) - PMI(w_i, NEG)$$

- DAILY NEWS SENTIMENT INDEX:

$$ISD^{PMI} = \frac{\sum ISN^{PMI}}{\#news \text{ per day}}$$



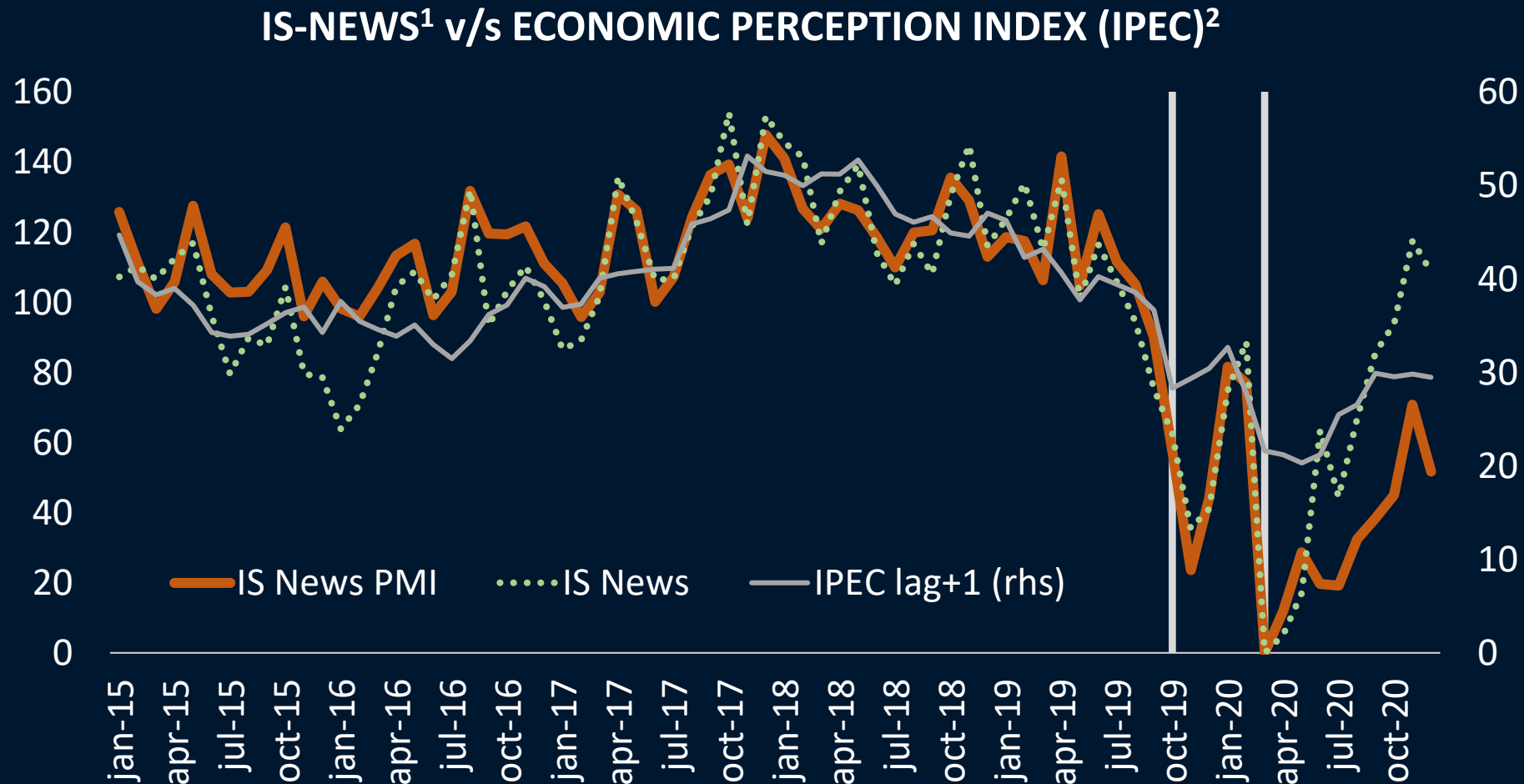


IS-NEWS RESULTS

IS-NEWS HIGHLY CORRELATED WITH CONSUMER SENTIMENT INDEX

IT IMPROVES USING PMI (Pearson's R = 83%)

10



(1) Index 2015-2020 average = 100.

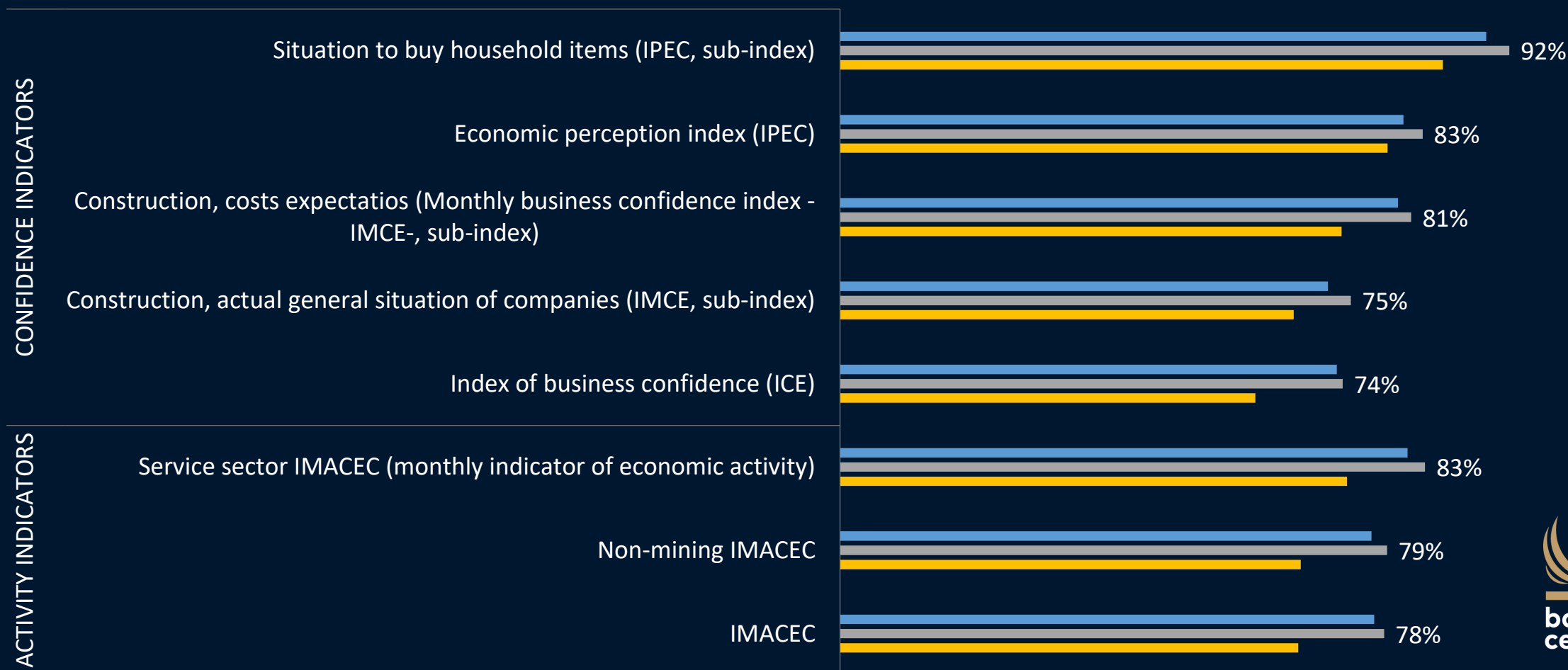
(2) Value above (below) 50 indicates optimism (pessimism).

IS-NEWS CORRELATES WITH CONFIDENCE AND ACTIVITY INDICES

11

PEARSON CORRELATION COEFFICIENT: IS-NEWS^{PMI} vs CONFIDENCE AND ACTIVITY INDICES (confidence indices in bases; activity indices in 12-month percent variations; %)

■ contemporaneous ■ one month ahead ■ two month ahead



1. **IS-News**: high correlations with confidence and activity indices, which grow with the application of the PMI.
2. **Effectiveness**: based on the construction of a dictionary of purpose, and a comparable size to those of greater use in the English language.
3. **Predictability**: anticipates economic shocks in the Chilean economy in a period of around 3 to 4 weeks.
4. **High availability and low cost**: the implementation of a news sentiment index in real time and independent from other data sources.
5. **Text mining in news articles**: many others research applications, such as topic analysis and bag of words to measure intensity.