Fernandez, Raul; Palma Guizar, Brenda; Rho, Caterina

Article

A sentiment-based risk indicator for the Mexican financial sector

Latin American Journal of Central Banking (LAJCB)

Provided in Cooperation with:
Center for Latin American Monetary Studies, Mexico City

Suggested Citation: Fernandez, Raul; Palma Guizar, Brenda; Rho, Caterina (2021) : A sentiment-based risk indicator for the Mexican financial sector, Latin American Journal of Central Banking (LAJCB), ISSN 2666-1438, Elsevier, Amsterdam, Vol. 2, Iss. 3, pp. 1-27, https://doi.org/10.1016/j.latcb.2021.100036

This Version is available at:
https://hdl.handle.net/10419/336327

Latin American Journal of Central Banking 2 (2021) 100036

Contents lists available at ScienceDirect

Latin American Journal of Central Banking

journal homepage: www.elsevier.com/locate/latcb

A sentiment-based risk indicator for the Mexican financial sector☆

Raul Fernandez, Brenda Palma Guizar, Caterina Rho*

Banco de México, Mexico

ARTICLE INFO

JEL classification:
G1
G21
G41

Keywords:
sentiment analysis
systemic risk
banks

ABSTRACT

We apply sentiment analysis to Twitter messages in Spanish to build a sentiment risk index for the financial sector in Mexico. We classify a sample of tweets from 2006-2019 to identify messages in response to a positive or negative shock to the Mexican financial sector, relative to merely informative ones. We use a voting classifier approach based on three different classifiers: one based on word polarities from a pre-defined dictionary, one based on a support vector machine classifier, and one based on neural networks. We find that the voting classifier outperforms each of the other classifiers when taken alone. Next, we compare our sentiment index with existing indicators of financial stress based on quantitative variables. We find that this novel index captures the impact of sources of financial stress not explicitly encompassed in quantitative risk measures, such as financial frauds, failures in payment systems, and money laundering. Finally, we show that a shock in our Twitter sentiment index correlates positively with an increase in financial market risk, stock market volatility, sovereign risk, and foreign exchange rate volatility.

1. Introduction

In this paper, we use sentiment analysis to build a sentiment index based on tweets in Spanish. The index intends to capture the perception of risk in the Mexican financial system as reflected on Twitter, a social media platform that has gained popularity among mass media, academics, policy makers, politicians, and the general public. To perform the sentiment analysis on tweets, we apply known text mining and machine learning techniques. We extract tweets in Spanish for the entire timeline of Twitter, beginning in April 2006 and ending in June 2019. We select only tweets mentioning Mexican banks, published by verified accounts, and specifically by domestic and international newspapers, news agencies, and rating agencies. Our goal is to select trusted news and comments about the Mexican banking sector and the financial sector as a whole. We find that the Twitter sentiment index improves forecasts of financial market stress and is an effective tool for monitoring financial risk events.

Our analysis develops in three steps. First, we perform a topic analysis to classify the content related to the Mexican financial system. We use the Latent Dirichlet Allocation algorithm to describe the sample of tweets through a set of topics, each represented as a collection of words. We identify some topics, such as financial fraud, money laundering, and failures of online payment systems, that are not traditionally included in financial stress risk indices.

Second, our paper explores alternative techniques that may be suitable for sentiment analysis of social media. To build our sentiment index for the Mexican financial sector, we train three different sentiment classifiers: one that uses word counts, based on the

☆ We are grateful to Liduvina Cisneros Ruiz, Jorge Luis García Ramírez, Fabrizio López Gallo Dey, Calixto López Castañon, Yahir López Chuken, Jorge De La Vega Gongora, Chris Hudson, Lorenzo Menna, Sabino Miranda Jiménez, and Alberto Romero Aranda for helpful comments at various stages of this work. The views expressed in this paper are those of the authors and do not necessarily reflect those of Banco de México or its policy. All errors are our own. Declarations of interest: None

* Corresponding author.
E-mail addresses: rfernandez@banxico.org.mx (R. Fernandez), bpalmag@banxico.org.mx (B. Palma Guizar), crho@banxico.org.mx, caterina.rho@gmail.com (C. Rho).

https://doi.org/10.1016/j.latcb.2021.100036
Received 29 October 2020; Received in revised form 29 May 2021; Accepted 22 June 2021
Available online 3 August 2021

work of Correa et al. (2021), a linear classifier (Baseline for Multilingual Sentiment Analysis model) proposed by Tellez et al. (2017); and one based on neural networks and transfer learning developed by Howard and Ruder (2018). Finally, we combine the three sentiment indices using a voting scheme.

Third, we compare the performance of our index with existing measures of financial stress. We apply local projections (Jordà, 2005) to test the effect of a shock in our sentiment index on a financial market stress index and selected market variables. We do not claim causality on these results because the direction of the causality between sentiment indicators and financial variables is still an open question (Shapiro et al., 2017). When looking over a 26-weeks horizon, a one-standard-deviation shock significantly correlates with an increase in the exchange rate volatility and stock market volatility in the first 10 weeks after the shock. The sentiment index also correlates with increased country risk as measured by the EMBI+ for Mexico. The banking sector risk, proxied by the beta of financial institutions, reacts to a shock in the sentiment index by rising, although the reaction is not significant in the short run. The correlation between the sentiment index and the general financial stress index is positive and significant. Our results align with the work of Shapiro et al. (2017), showing that news sentiment indices improve the forecast performance of financial indicators.

2. Big data analysis in central banks

Central banks and international organizations recently started to enlarge their data sources taking advantage of textual data such as social media content, financial news, or official documents of central banks (financial stability reports, monetary policy reports). New machine learning techniques allow analysis of the increasing volumes of unstructured data. Among the machine learning techniques, text mining has proven to have multiple applications of which sentiment analysis has appeared particularly appealing for financial applications.

In this context, big data techniques found a novel application in analyzing ”soft information,” like sentiment, to monitor financial risk (Nyman et al., 2018), systemic risk (Borovkova et al., 2017), and uncertainty (Baker et al., 2016). A growing collection of literature focuses on studying social media activity, in particular Twitter messaging, on stock market fluctuations in coincidence with decisive events, such as monetary policy decisions (Azar and Lo, 2016). The evidence presented in this literature suggests that social media activity and news content influence financial market agents and can cause a shift in their decisions, leading to changes in market prices (Bukovina, 2016). This may have consequences for the financial sector or the economy as a whole. For this reason, researchers are developing alternative economic and financial indicators based on the analysis of high-frequency unstructured data, especially news or Twitter content (Accornero and Moscatelli, 2018; Angelico et al., 2018; Borovkova et al., 2017).

In the context of financial studies, these data are often used to build financial market indices that replicate the variations in traditional stock market indices, signaling sudden changes in market trends in advance. Borovkova et al. (2017) propose a new Sentiment-based Systemic Risk indicator of the global financial system. They build it by aggregating sentiment in the news regarding the Systemically Important Financial Institutions. They find that their systemic risk indicator anticipates, by as long as 12 weeks, other systemic risk measures such as SRISK or VIX in signaling periods of stress. Shapiro et al. (2017) use machine learning techniques to develop and analyze new time series measures of economic sentiment based on text analyses of articles in financial newspapers from 1980 to 2015. They find that the four news sentiment indices that they developed are strongly correlated with contemporaneous business cycle indicators and improve the forecast performance of standard financial indicators.

An indicator based on tweets about financial news can be used to analyze the sentiment of investors or consumers in response to different shocks. Angelico et al. (2018) use sentiment analysis to show how high-frequency Twitter data can help central banks to complement low-frequency survey-based data when estimating inflation expectations. Other papers apply sentiment analysis to Twitter data to measure the confidence of the general public in the banking sector. Accornero and Moscatelli (2018) use this approach to create an early-warning indicator targeted at evaluating retail depositors' levels of trust. Bruno et al. (2018a) build a dictionary to analyze sentiment in Italian texts, while (Bruno et al., 2018b) apply the dictionary to tweets about selected Italian banks to build sentiment indicators. They find that these sentiment indicators have a positive correlation with some banks' financial variables.

Correa et al. (2017a, 2021) also apply sentiment analysis to the central bank's Financial Stability Reports. In particular, they analyze the relationship between the financial cycle and the sentiment conveyed in these official publications. They build a new dictionary of financial and economic terms, which they use to construct a financial stability sentiment index for 35 countries, from 2005 to 2015. They find that the developments in the banking sector and information about this specific sector are the main drivers of the financial stability index. Moreover, the sentiment captured by their index translates into changes in financial market indicators related to credit, asset prices, and systemic risk. Bruno (2017) conducts a similar analysis on recent Financial Stability Reports issued by the Bank of Italy, while in a recent paper, (Iván and González Pedraz, 2020) build a financial dictionary in Spanish to analyze the sentiment of the Bank of Spain's Financial Stability Reports.

Our paper contributes to this literature by building a financial market risk index based on the sentiment of tweets and news about the main commercial banks in Mexico. We build on the work of Correa et al. (2021), and explore alternative techniques that may be suitable for sentiment analysis in social media, such as a model of neural networks and transfer learning (Howard and Ruder, 2018) and the Baseline for Multilingual Sentiment Analysis model (Tellez et al., 2017). We take inspiration from Shapiro et al. (2017) to test how our Twitter sentiment index performs in comparison with other measures of financial stress and economic uncertainty. We refer to the Financial Market Stress Index developed by Banco de México (Banxico) (Banco de Mexico (2019)) and to selected financial indicators.

Table 1
List of Twitter accounts considered in this study.

Type of source	Name	Type of source	Name
Mexican newspapers	El Financiero	Foreign newspapers	El País
	El Economista		El País ("Americas" edition)
	Reforma		The New York Times (in Spanish)
	Reforma Negocios		Forbes
	Milenio		Forbes Mexico
	La Jornada	Press agencies	Associated Press Latin America
	Excelsior		Reuters, Latin American Edition
	El Sol de México		Xinhua (in Spanish)
	El Universal		AFP (in Spanish)
	La Razon		EFE Mexico
	Diario 24 horas	All-news television	BBC (in Spanish)
	Capital Mexico	Rating agencies	Moody's
	Reporte Indigo		Fitch Ratings
	El Heraldo de México		
	La cronica de hoy		
	SDP noticias		

3. Data

To build the banking risk index, we use Twitter as our data source and Mexican commercial banks' names as our search criteria. We select only tweets that contain the name of at least one Mexican bank or the words "banco," "banca," "bancario" (Spanish for bank, banking). The banking system is at the core of the Mexican financial system; therefore, the health of the financial system is largely determined by how healthy Mexican banks are.

3.1. Extraction of tweets

We use the Twitter Paid Premium Search API, which allows us to extract tweets in Spanish containing Mexican commercial banks' names from April 2006 onward.[1] We limit the extraction to tweets in Spanish because it is the official language in Mexico, and the language that newspapers, rating agencies, and other sources reporting about Mexico are expected to use. English language media (such as those based in the US or UK) often report only major events about Mexico, or as foreign sources, they report events about Mexico with a short delay, although they may reach a broader global audience. By using tweets in English, we may miss information regarding daily events, or events that specifically regard the Mexican financial sector or Mexican banks. As an extension of this analysis, we could take advantage of the tweets in both English and Spanish. However, that is beyond the scope of this paper. The complexity and time cost of setting up a text analysis in two languages at once is significant, especially because of the peculiarities of each language. For this reason, we extract only tweets in Spanish.

We also limit the extraction to verified Twitter accounts of national and international newspapers, news agencies, and rating agencies. Twitter can be viewed as an information source, and when tweets occur in conjunction with traditional news events, more information is spread to the public, in particular to investors who operate in financial markets (Rakowski et al., 2020). We decided to base our analysis on reliable sources that can influence the perception that the public has of banking institutions and the financial sector in Mexico. If the banks are perceived as "healthy" or "solid" by the media, they will likely be perceived as such by financial market players and the public in general. Table 1 lists our media sources.

Limiting the extraction of tweets to the selected accounts allows us to reduce potential noise in our database. To test this hypothesis, we extract all tweets that included any of the chosen keywords for a particular day, and we compared the complete sample (3,004 tweets) with the sample filtered by the selected accounts (34 tweets).[2] We looked at the most frequent words for both samples, and found that among the most popular words in the whole sample, only a few were linked to the topic of interest: "venta" (sale) and "tarjeta" (card, credit card). Other more frequent words were too general to indicate a specific topic ("north," "route," "popular"). Top words for the filtered sample, on the other hand, seemed to be much more related to our topic: "financial," "market," "growth," "director," and "president," all words linked to financial markets, banking, or monetary policy. From the simple reading of the tweets extracted without the filter, we find that many tweets regard the marketing strategies of commercial banks, job offers, users' comments about customer services or their relationship with a certain bank, and events sponsored by banks. This kind of information is not relevant to the focus of this paper. Although the volume of data was drastically reduced, we would amass an excess of information without this filter.

Table 2 shows a selection of sample tweets from our final database built from the selected accounts. For each tweet, we retrieve the tweet content and some other attributes such as the tweet i.d., the publication date and time, the user who published it, the number

[1] We consider that some commercial banks changed their name in the period of interest due to mergers or acquisitions.

[2] We select March 20, 2019 as a representative day because no relevant events were occurring, such as an election day, a change in monetary policy etc., that could bias the results.

Table 2
Sample of extracted tweets included in our database.

Date and time	Text	User	Followers	Country
08/09/2010 19:20	Asigna Moody's calificación de deuda senior a Banamex.	LaRazon_mx	122,751	Mexico
25/11/2011 12:24	El Gobierno indulta al consejero delegado del Banco Santander, Alfredo Sáenz.	el_pais	6,818,004	Spain
17/07/2012 16:31	HSBC de EEUU se disculpa por fallas que permitieron narcolavado.	AP_Noticias	222,131	USA
22/07/2013 16:50	Utilidades de #UBS superan expectativas.	eleconomista	447,505	Mexico
14/01/2014 13:00	#ReformaEnergética: un elemento de cambio en México. Adolfo Acebrás de @UBS ahonda en el tema.	Forbes_Mexico	507,926	Mexico
09/02/2015 14:44	Cómo el banco HSBC "ayudó" a millonarios a evadir impuestos.	bbcmundo	3,163,376	UK
30/09/2016 20:18	El Banco Santander baja su objetivo de rentabilidad por el Brexit. #AFP	AFPespanol	285,893	Uruguay
02/02/2017 17:23	En condiciones actuales, aumento de gasolina sería de 0.5%: Banco Base.	El_Universal_Mx	4,941,610	Mexico
06/06/2018 09:29	TLCAN y aranceles presionan al tipo de cambio, que podría seguir volátil: Omar Taboada, de @Citibanamex y Carlos González, de Monex, en entrevista con @VictorPiz en #AlSonarLaCampana.	ElFinanciero_Mx	1,181,553	Mexico
01/02/2019 00:40	Analistas de Barclays y BNP Paribás advirtieron que inversionistas de Wall Street estén preocupados por la situación de Pemex.	eleconomista	447,506	Mexico

Each entry includes the date and time of the publication of the tweet, the text of the message, the name of the Twitter account, the number of followers and the country where the account was created.

Table 3
Number of tweets posted by national and international sources, 2007-2019.

Year	National sources	International sources
2007	0	47
2008	113	117
2009	582	113
2010	331	102
2011	717	114
2012	2,870	296
2013	2,571	527
2014	2,952	625
2015	2,200	539
2016	1,766	480
2017	1,241	493
2018	2,235	465
2019	1,639	339

of followers of this user, the reactions to the tweet (likes and retweets), and the country of origin of the tweet. The database consists of around 23,000 tweets. The tweet volume at the beginning of the observation period is lower than the one observed towards more recent periods, as Twitter started gaining popularity.

Table 3 shows the number of tweets per year by type of source. All the Mexican newspapers are considered national sources, while all the other accounts are counted as international sources. Most of the tweets we consider in this study come from Mexican sources, and the ratio of foreign tweets to national tweets oscillates between 20 percent and 30 percent in most of the years. The number of tweets we extract per year increases over time and reflects the increase in the use of Twitter by the media.

3.2. Data preprocessing

Since the tweets' main content is text, it is necessary to do some preprocessing before the analysis. We implement the following preprocessing steps, with some variations depending on the specific task or model:

1. We remove tweet-specific elements like hyperlinks, retweets, user mentions, and elements such as stopwords, numbers, and punctuation. This step allows us to drop text that does not add useful information to our analysis;
2. We anonymize banks by masking their names to avoid having banks' names as features in our models;
3. We lemmatize the text to reduce the sparsity of the data;[3] and
4. We convert all uppercase letters to lowercase.

3.3. Data exploration

After preprocessing the tweets, we conduct an exploratory analysis of the data to better understand the information we obtained from the extractions. Our final goal is to build a sentiment index based on the negative or positive sentiment that news of potential

[3] Lemmatization reduces inflectional forms and sometimes derivative forms of a word to a common base form (their dictionary form).

financial risk events conveys to the public. Therefore, we need to make sure that the tweets we extract are relevant to our objective. If the information we extract is not related to topics that are significant for the evaluation of financial risk, our sentiment index would be biased, or even useless. However, analyzing text data manually can be an enormous task: reading a text, and classifying the information it contains is feasible when the amount of text analyzed is limited, but it becomes a burden, in terms of time and effort, when a particularly large amount of textual data needs to be analyzed. Our final sample contains 23,000 tweets; the risk of human error in classifying and summarizing this amount of information is too high, and it would be significantly time consuming. For this reason, we apply topic analysis to explore our sample of tweets.

3.3.1. Topic analysis with LDA

Topic analysis is a natural language processing technique that automatically extracts meaning from texts by identifying recurring themes or topics in the text corpus.[4] It helps the researcher organize large data sets and identify the most frequent topics in a simple, fast, and scalable way. We use the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003; Bruno et al., 2018b), a popular topic modeling method, to perform a topic analysis of our tweet database (See section A.1 of the Appendix for more technical details).

After several iterations using different numbers of topics, we identify six topics that constantly appear in the results:

1. Financial markets,
2. Macroeconomic expectations,
3. Foreign exchange market,
4. Business activity,
5. Financial results, and
6. Illicit activities and penalties.

To name the topics and minimize the degree of subjectivity when doing so, we analyze both the collection of words representing the topics and the most representative documents for each topic. Since it is assumed that the documents are a mixture of topics, we can obtain a document-topic matrix indicating the probability of the document belonging to each of the topics. We use this matrix to find the most representative documents (i.e., the most representative tweets) per topic.

We compare the six LDA topics with Banxico's Financial Market Stress Index (Índice de Estrés de los Mercados Financieros, IEMF, (Banco de Mexico, 2019)) components. The IEMF index is published weekly and synthesizes the information of 33 financial variables that have an impact on financial stress. The variables cover six different sources of stress: bond market, stock market, foreign exchange market, derivative market, credit institutions and country risk.

The topics found in our tweets have some overlap with the IEMF, but they also capture new information that quantitative financial indicators do not explicitly show. The IEMF components "bond market," "stock market," and "derivative market" overlap with the topic "financial markets" found in the tweets. The IEMF component "foreign exchange market" corresponds to the "foreign exchange " topic in the tweets. We interpret the topic "macroeconomic expectations" as an indicator of country risk. Topics 4 and 5 (business activity and financial results) may fall under the "credit institutions" component of the IEMF. However, Twitter data provides information on certain details of the business activity that is not being explicitly captured by the IEMF. We detect sentiment about customer services, digital services, and online payment systems, including bugs. Additionally, our data captures new information within topic 6, "illicit activities and penalties." This topic comprises news about money laundering activities, tax evasion, banking scandals, online fraud, and bank penalties related to illicit activities. Appendix A.1.1 presents the most important words present in each topic and more details about the topic selection.

3.4. Data labeling

We create a sample of labeled data that serves to train the models and compare their performance. We take a random sample of 2,000 tweets from our database and assign juxtaposed sub-samples of 100 tweets to 37 professionals, working at the Directorate General of Financial Stability in Banxico, who label the tweets according to the message they transmit regarding the level of risk in the Mexican financial system or to the Mexican banks, following the rules described below.

The "risk" we want to ascertain with this sentiment index is the banking risk from the point of view of regulatory institutions or the banks themselves. Most of the time, the two perspectives coincide. For instance, a tweet about the downgrade of the sovereign rating of Mexico would report a negative shock to the banking or financial system, and it would increase the banking risk both from the point of view of regulators and from the point of view of banks. However, a tweet that reports news about an increase in capital requirements established by the Basel rules, might be negative for banks' profitability but positive from the regulators' viewpoint, because it would increase the resilience of the banking system to negative shocks. In such cases, we favor the systemic risk view, so that we consider the tweet as reporting news that decrease the banking risk. The labeling criteria for categorizing each tweet is as follows:

- Higher risk (corresponding to negative sentiment): tweets with content that reflects negative expectations for the banking sector or the financial system as a whole. Examples are tweets reporting news about lower economic growth, higher volatility of the exchange rate, failures in the IT systems of banks or in online payment systems, safety violations, financial frauds, money laundering operations.

[4] In quantitative linguistic research, a corpus is a set of machine-readable texts prepared for analysis.

- Lower risk (corresponding to positive sentiment): tweets with content that reflects positive expectations for the banking sector or the financial system as a whole. Examples are tweets reporting news about regulatory compliance, comments on the strength of the financial or banking system, higher economic growth.
- Neutral: tweets that are merely informative or that do not contain a clear positive or negative judgment. Examples are tweets reporting news about ordinary business activities of banks; tweets reporting only the daily exchange rate, without any comment or comparison with previous periods; news about changes in the industrial organization of the banking sector; or low-level crimes (bank robberies to a specific branch).

Section A.2 of the Appendix lists more details about our labeling strategy. The final label for each tweet is the mode of the labels we collect for that tweet. Having more than one person labeling the same tweet allows us to control for labeling coherence. The final sample is composed of 32 percent negative tweets, 26 percent positive tweets, and 42 percent neutral tweets.

4. Sentiment classifiers

We choose three different models to build the sentiment classifier for the tweets: a dictionary with word polarities, a Support Vector Machine classifier (SVM), and a neural network. Our first approach replicates the methodology of Correa et al. (2021) based on a previously built financial dictionary with word polarities. This methodology works through word counts. The second model is based on a multilingual language model developed by Tellez et al. (2017). It focuses mainly on text preprocessing and text vectorization. After these transformations, a SVM classifier is trained to perform the classification. The third model is the Universal Language Model Fine Tuning for Text Classification (ULMFiT) developed by Howard and Ruder (2018). This algorithm uses a neural network composed of a language model and a classification layer on top.

Each model has advantages and disadvantages. The dictionary model is the simplest one, and it has the particular advantage of not needing previously labeled data, but the disadvantage of requiring a pre-built dictionary that may not necessarily adjust to the target documents. The user is responsible for choosing the words (attributes) that will determine the sentiment of a given text. On the other hand, the SVM classifier and neural networks can learn from the data, and the attributes used for the classification are inferred during the training process, allowing for a better adjustment to the target corpus. The downside of these two models is the requirement of labeled data, which may be quite expensive and time consuming. We further discuss the advantages and disadvantages of each model and their characteristics in section A.3 of the Appendix.

To classify the tweets, we split our labeled tweet sample into training and test sets. We train each sentiment classification model using the training set, with 90 percent of the labeled tweets, and then compare the models' performance on the test set, the remaining 10 percent of labeled tweets. The training step is not necessary when using the dictionary model, since the tweet sentiment is computed based on word counts of the positive and negative words identified by the dictionary. However, the labeled data, in this case, is useful for measuring the model's performance, and it allows us to compare the performance of the different algorithms.

Finally, to make our classification more robust and increase the average accuracy, we build a sentiment classifier based on the outputs of the previously presented models. The expectation is that integrating multiple models, known in machine learning as ensemble methodology, can help to create a model with enhanced predictive performance (Rokach, 2010). Our classifier uses a majority voting rule[5] to determine the final sentiment. Since our classifier based on majority voting observes three classifiers, at least two must agree for a tweet to receive a polarity. Whenever there is no agreement, the tweet is categorized as neutral.

4.1. Comparison between the sentiment classifiers

The different classifiers are trained and evaluated with the same data set. To compare the models' performance, we compute accuracy, balanced accuracy, and F1 score. For more details, see section A.4 of the Appendix.

All the accuracy measures have a [0, 1] range, where higher scores (close to 1) are preferred. Table 4 presents the results.

Higher accuracy reflects a better classification of the positive, negative, and neutral tweets by the model. If we performed a random classification, the expected probability of a tweet to be assigned to one of our three classes would be 33 percent. If the accuracy of the classifier is higher than this threshold, the model is doing a better job of classifying data than a random classification.[6] When looking at the results for the B4MSA and ULMFiT models, we find that the accuracy in the test set is around 73 percent, which is slightly above the 70 percent accuracies found on Twitter sentiment analysis literature (Zimbra et al., 2018). Similar to what is expected in regression analysis, in both models the accuracy over the training set is higher than that in the test set.[7] The F1 score gives the same results, in line with the test set accuracy.

We also compute the accuracy separately for each class: positive, neutral, and negative. For the comparison between models, the dictionary method is our baseline. Although it performs well, by construction, it cannot adapt to the analyzed documents, the tweets,

[5] A voting rule is a simple ensemble methodology that could help making the classification more robust. The voting rules consist of three possibilities: unanimous voting, simple majority voting, and plurality voting. If the classifier outputs are independent,then it can be shown that majority voting is the optimal combination rule (Polikar, 2012).

[6] To be precise, our sample of tweets has an unbalanced distribution of positive, neutral and negative tweets. Given the subsample of labeled tweets, we may expect a threshold of 32 percent for negative tweets, 26 percent for positive tweets and 42 percent for neutral tweets.

[7] The gap in accuracy between the training and test sets should not be too wide: a wide gap between the test set and training set may be a signal that the model is overfitted, and out-of-sample forecasts may be biased. However, there is no rule of thumb that sets an optimal gap between the accuracy of training and test sets.

Table 4
Accuracy metrics of each classifier.

Model	(1) Dictionary	(2) B4MSA-SVM	(3) ULMFiT	(4) Majority voting
Test set accuracy	0.61	0.73	0.73	0.74
Training set accuracy	0.64	0.86	0.85	0.86
Balanced accuracy	0.61	0.73	0.73	0.74
F1 score	0.61	0.73	0.73	0.74
Accuracy per category				
Positive tweets	0.60	0.63	0.63	0.67
Neutral tweets	0.55	0.75	0.82	0.72
Negative tweets	0.82	0.78	0.69	0.84

The test set and the training set correspond respectively to the 10% and the 90% of the sample of labeled tweets. Accuracy is the ratio of correctly predicted tweets to the total number of tweets. The balanced accuracy is the average of the correctly predicted tweets computed on each class individually. Accuracy spans a [0 1] interval, where higher accuracy corresponds to values closer to 1.F1 score is the harmonic mean of Precision and Recall. See Appendix A.4.

as the other two methods can. For this reason, we expect lower accuracy. Its accuracy is, in fact, 61 percent considering the whole sample of tweets, much lower than the SVM model and the neural networks model (accuracy of 73 percent). It performs remarkably well in the classification of the negative tweets (accuracy of 82 percent), probably because the CKJM dictionary contains many more negative words than positive words. B4MSA and ULMFiT models have comparable accuracies, 73 percent for the test set.

Since our dataset is not balanced (we have more tweets for the neutral category than for the positive or negative ones), we also consider the balanced accuracy for each model. Again, the B4MSA and ULMFiT results are quite close and considerably outperform the dictionary results. The final column of Table 4 presents the performance metrics for the majority voting model. This final classifier maximizes the available information and shows the best performance of the four models. Its general accuracy is 74 percent, the highest, and its accuracy computed for the different classes separately takes advantage of all three models. The accuracy in classifying the negative tweets is comparable to the accuracy of the dictionary model, while for the positive and neutral categories, it is in line with the higher accuracy of the B4MSA and ULMFiT models.

5. Twitter sentiment index

Once the tweets are classified, we proceed to build the Twitter sentiment index. We use the classification given by the majority voting model since it performs better theoretically (Rokach, 2010) and empirically, as stated in Table 4. We base our methodology on Correa et al. (2021).

Instead of the number of positive and negative words in each document, we count the number of positive and negative tweets, and we scale the index by the total number of positive and negative tweets:

$$Twitter\ sentiment\ index_t = \frac{negative\ tweets_t - positive\ tweets_t}{negative\ tweets_t + positive\ tweets_t} \tag{1}$$

with t indicating the time span of interest (day, week, month, or year). Higher sentiment index values suggest higher negative sentiment regarding the banking and financial system.

In the denominator, the baseline Twitter sentiment index considers the positive and negative Twitter messages published in period t. In this way, we normalize the index, considering the variability in the volume of tweets posted in the period of interest. The baseline Twitter sentiment index excludes the neutral tweets because they may introduce some noise to the index. The neutral tweets group may include tweets about banks that give neutral information but also all the tweets that should be discarded because they do not offer relevant information about financial market stress (tweets about events or soccer teams sponsored by a banking group, for instance).

Other than the polarity of the tweets, another possible source of information available from our extraction is the visibility of the tweet to Twitter users. The number of reactions (retweets or likes) a tweet receives may be seen as an indicator of its popularity. Reactions also increase the exposure of tweets, thus augmenting their reach. This may lead to stronger sentiment, positive or negative, given by one single tweet with respect to another. The number of reactions a tweet gets may amplify the sentiment regarding important news: people may retweet more easily news that they find important and for which they feel particularly strong sentiment, either positive or negative. If this is the case, the higher the number of reactions, the stronger the sentiment produced by that specific tweet, and the more important the news content. However, a higher number of reactions may be prompted only by personal curiosity, not by the importance of the news content of the tweet on a systemic level. In this case, the number of reactions of each tweet may be a lower bound for the visibility that the tweet has. Still, it may add noise to the indicator: news that users found interesting but is not important at a systemic level gets a higher weight, and the final index is more biased than the baseline index without weighting for reactions.

Considering these points, the inclusion of neutral tweets in the index, and the potential importance of each tweet to Twitter users, we build other versions of the baseline Twitter sentiment index. The first includes the neutral tweets in the denominator, while the

Table 5
Correlations between the IEMF and different versions of the sentiment index.

Sentiment Index	Correlation with IEMF	p-value
Baseline	0.1342	0.0011
Version 2	0.0576	0.1630
Version 3	-0.0003	0.9934
Version 4	0.0170	0.6808

Baseline: SI computed not considering neutral tweets; Version 2: SI computed considering neutral tweets; Version 3: SI computed not considering neutral tweets and weighting the tweets by the number of reactions to the tweet; Version 4: SI computed considering neutral tweets and weighting the tweets by the number of reactions to the tweet.

second variation weights each tweet by the number of reactions (both retweets and likes) received. Finally, the third version includes both neutral tweets in its denominator and reaction weights. Table 5 presents the correlation between the IEMF (our benchmark indicator of Mexican financial market stress) and each of the sentiment index versions (baseline, including neutral tweets in the denominator, reaction weights, and both neutral and reaction weights combined).

The correlation between the IEMF and the baseline sentiment index is higher than the correlation of each of the other versions of the sentiment index with the IEMF and is the only one that significantly differs from zero. This result suggests that the baseline sentiment index is the best indicator to explain Mexican financial stress; therefore, we will focus our analysis only on this version of the index.

5.1. Visualization

To visualize the results, we build an interactive dashboard that displays a graph with the volume of tweets, broken down by tweet sentiment; a graph showing the Twitter sentiment index along the period of analysis; and a word cloud of the most popular terms used in the tweets during the selected period. This may help understand abnormal changes in the sentiment index.

Figure 1 shows a screenshot of the dashboard, displaying on the right the word clouds for January 2019, when Fitch downgraded Pemex's rating from BBB+ to BBB-.

The risk increase caused by this event is captured by the index, and the word clouds highlight "Pemex," "calificación," and "Fitch" as negative words. The larger a word appears in the word cloud, the more important it is in its respective category. On the left, Figure 1 shows the complete timeline of the volume of tweets extracted and of the sentiment index computed from the tweets. Although we start our extraction in 2006, the year Twitter went online, the graph depicting the volume of tweets shows that, at the beginning of the period, the total number of daily tweets containing one of our keywords (i.e. the names of the banks) is very low.

Over time, the number of tweets increases and, on average, it stabilizes during 2013, with the exception of the occasional spikes. The growth of the tweets about banks follows the growing popularity of Twitter among the general public and the evolution of its use as a communication tool not only by private users, but also by businesses, public and private institutions, newspapers, and media outlets. Even with very few observations, it is possible to compute a sentiment index, as shown in the second graph on the left in Figure 1. Nonetheless, if the number of observations (i.e., the number of tweets) is too low, the index may be biased because it is built on too few observations. For instance, 2006 and 2007 have very few tweets, fewer than 50 for the two years. So, we will truncate the series by starting our empirical analysis in 2008.

Figure 2 shows the four alternative indices computed at monthly frequencies using the baseline model. The Twitter sentiment index scale is set from -1 (minimum risk) to 1 (maximum risk). The Twitter sentiment index computed using the majority voting model consistently signals higher risk compared to the others. The Twitter sentiment index computed using the neural network model broadly follows the voting sentiment index, except during the period from mid-2015 to mid-2016. The raw Twitter sentiment index with weekly frequency is too volatile to be used in a comparison with other, more standard economic indicators. In Section 5.2, we explain in detail how we address this issue.

Figure 3 presents the Twitter sentiment index based on the majority voting classifier, with monthly frequency. We focus our analysis on the baseline index, without considering neutral tweets or weighting. We also rescale the index from 0 (maximum positive sentiment) to 1 (maximum negative sentiment): an increase in the Twitter sentiment index corresponds to increased risk. We label each sentiment peak based on the keywords in the dashboard's word cloud, and we compare the keywords with those used in that month's news. We find that the peaks of the Twitter sentiment index correspond to significant events for the Mexican financial system. This is a descriptive analysis, so we are not implying that a peak in the sentiment index causes the event, we only use this comparison to make sense of our results.

At the end of 2008 and during 2009, we see an increase of negative sentiment. This period corresponds to the 2008-2009 global financial crisis. For these two years, the number of tweets is still relatively limited, so we do not see a spike in monthly data. However, when analyzing the content of the tweets, we find negative tweets that refer to the global economic crisis starting in October 2008.

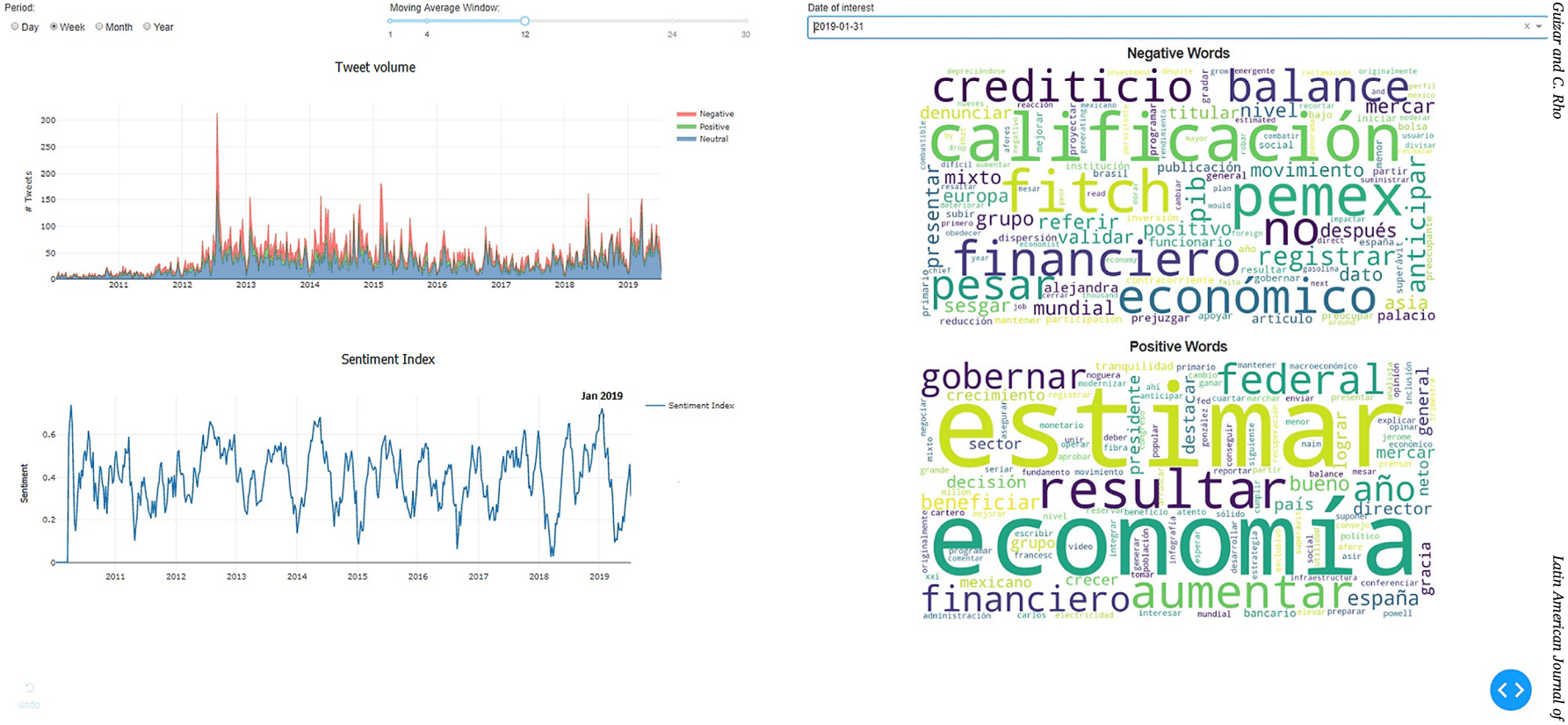

Fig. 1. Example of the sentiment index dashboard. The upper left shows the volume of tweets and their sentiment, the lower left shows the monthly sentiment index, the right shows the word clouds present in negative tweets (negative words) and positive tweets (positive words) in January 2019.

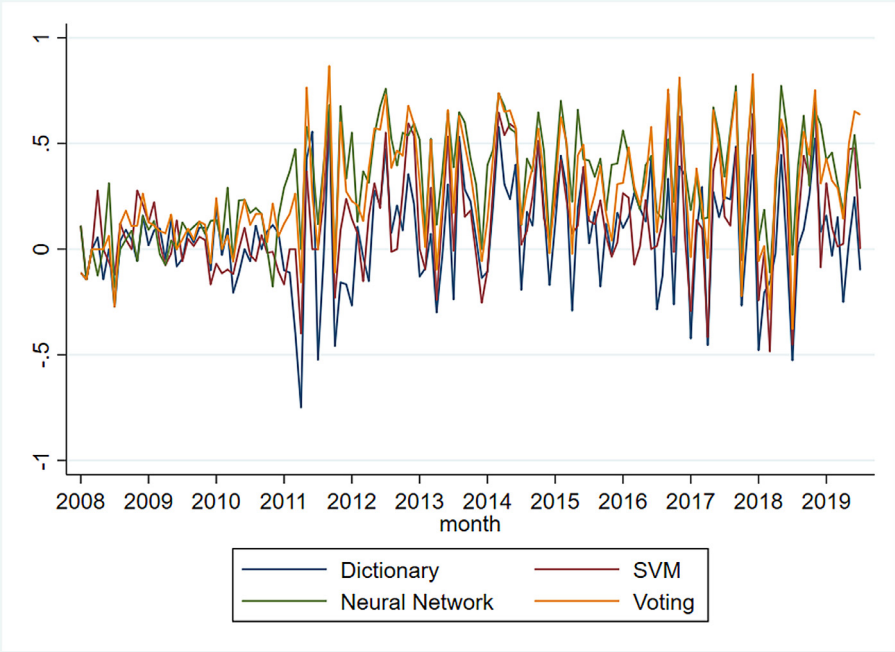

Fig. 2. Comparison of the various Twitter sentiment indices, obtained using the sentiment classification of the four classifiers: one based on the financial dictionary, one based on the Support Vector Machine model, one based on a neural network and a final classifier based on majority voting. Period: 2008-2019, monthly frequency.

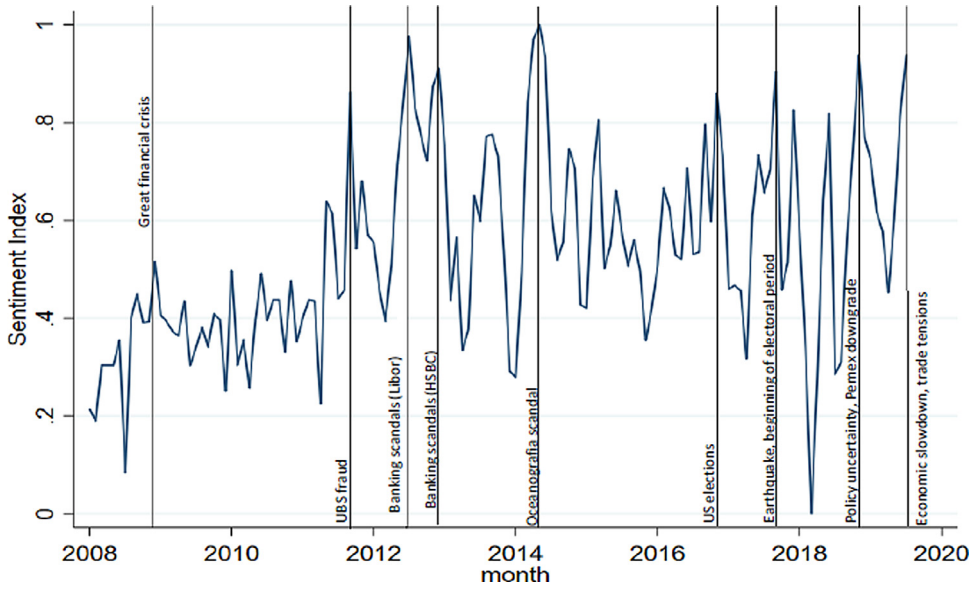

Fig. 3. Twitter sentiment index based on majority voting. The black bars highlight the peaks in negative sentiment toward financial markets in Mexico. Period: 2008-2019, monthly frequency.

From January 2011 until December 2015, most news that increases negative sentiment corresponds to events that increase reputational risk. In September 2011, UBS was involved in fraud due to unauthorized trading by one of its directors. The scandal caused UBS a loss of more than 2 billion US dollars.

In July 2012, global financial markets were shaken by the Libor manipulation scandal, and in December 2012, Mexico was hit by the HSBC money laundering scandal. The global bank had to pay a record fine of 1.92 billion dollars to US authorities for allowing money laundering by Mexican drug cartels in its US offices.

Finally, the Oceanografía scandal affected Mexico and its financial system directly in 2014. The oil services company Oceanografía was accused of a fraud that also involved the Mexican subsidiary of Citibank, Citibanamex. The loan scandal cost Citigroup more than 500 million dollars.

The period from January 2016 to June 2019 is characterized by shocks linked to macroeconomic, political, and systemic shocks, such as the US elections in November 2016, the electoral period in Mexico, the earthquake that hit Mexico in September 2017, volatility in financial markets, and domestic economic slowdown due to uncertainty in November 2018 and June 2019, respectively. In particular, on November 8, 2018, the Mexican ruling party proposed a project to reduce or prohibit banking charges for interbank transfers and cash withdrawals. On that day, the stock price of Banorte (the second largest banking group in Mexico) fell by 11 percent, and Santander stocks fell by more than 9 percent. This news is reflected in our sample of tweets. On June 5, 2019, the credit rating agencies Moody's and Fitch decreased Mexico's sovereign debt rating, citing risks posed by Pemex, the national oil company, which was heavily indebted, and trade tensions during the ratification process of the trade deal between Mexico, the United States, and Canada (USMCA).

5.2. A filtered sentiment index

The Twitter sentiment index computed using Equation (1) essentially shows the positive and negative sentiment shocks that affected the Mexican banking system during a given period. At weekly frequency, it is quite noisy. Ideally, we would like to have a smoother cumulative sentiment index that maintains a weekly frequency in the observations, but which shows a more definite trend. We can consider the baseline weekly Twitter sentiment index to contain noisy observations of the actual unobserved sentiment. Our goal is to extract the trend from the time series of the weekly index, omitting the noisy high-frequency components.

We take inspiration from Borovkova et al. (2017), and we filter the series to extract a meaningful signal from the data. We apply the Christiano-Fitzgerald band-pass filter (Christiano and Fitzgerald, 2003), which is used to smooth high frequency data (such as daily, weekly, or monthly). This filter suits our data better than two other filters widely used in the time series literature: the HP filter (Hodrick and Prescott, 1997) and the Baxter and King filter (Baxter and King, 1999).

In their 2003 paper, Christiano and Fitzgerald show that their proposed filter outperforms the HP filter in terms of flexibility in selecting the frequency bands of interest and the possibility of adapting the filter to time series of quarterly, monthly or even higher data frequency. This occurs because in the case of the HP filter, the smoothness parameter lambda must be chosen both on the basis of the data's frequency (annual, quarterly, or monthly) and to determine the selection of the bands of interest for the filter (in the business cycle literature it is usually the band 8-32 quarters). The selection of the parameter lambda is a key factor widely discussed in the literature, and there is no agreement on a single way to adapt the HP filter for data other than quarterly (Backus and Kehoe, 1992; Baxter and King, 1999; King and Rebelo, 1993; Ravn and Uhlig, 2002). The higher flexibility of the Christiano-Fitzgerald filter suits our case better, because we have weekly data and we do not focus specifically on studying business cycle frequencies, the case for which the HP filter works better and has been studied more deeply. Our focus is simply filtering the noise resulting from higher frequencies, not conducting a business cycle analysis.

In comparison with the Baxter-King filter, another well-known band-pass filter, the main advantage of the Christiano-Fitzgerald filter is that, by construction, it exploits the entire data set. The Baxter-King filter is a fixed-length symmetric filter, based on a moving average of the data with symmetric weights on leads and lags, so it discards a given set of data at the beginning and at the end of the series depending on the lead-lag length defined by the researcher. The Christiano-Fitzgerald filter we apply in this paper is a more general filter. It is not applied to a fixed rolling window of data, but to the full sample for each observation, and the weights on the leads and lags are allowed to differ. This allows us to filter the full data set, without discarding any element at the beginning or end of the series. The last element of the series is estimated using a one-sided filter, allowing real-time estimates.[8]

To filter only the high frequencies, we enlarge the band of the Christiano-Fitzgerald filter to 100 years. In this way, the band-pass filter becomes a sort of low-pass filter that eliminates only frequencies higher than the lower bound, and it maintains the lower frequencies over the long term. Ideally, the upper bound should be infinite, but as an approximation we fix it at 100 years.[9]

We compute three versions of the filtered sentiment index with the lower bound fixed at 1 year, 6 months, and 3 months. The filtered series resulting from the Christiano-Fitzgerald filter with the lower bound fixed at 3 months and 6 months is still noisy. Therefore, we will focus the rest of the analysis on the filtered series that uses the window 1 year-100 years when referring to the filtered sentiment index.

[8] Christiano and Fitzgerald (2003) show that the real-time performance of the filter they propose is higher than the performance of the HP filter, another widely used option for real-time estimates. We run a robustness check in which we filter the Twitter sentiment index and its sub-indices with the HP filter, using alternative values of lambda. The filtered indices obtained using the HP filter procedure are comparable with the filtered indices obtained using the Christiano-Fitgerald filter. These results are available on request.

[9] As a variation, we apply a traditional band-pass filter for business-cycle frequencies that considers the frequencies between 1.5 years and 8 years and we filter only the higher frequencies that last less than 1.5 years. As in the first approach, we use as a lower bound 1 year, 6 months and 3 months. The results are very similar to the main analysis. They are not shown, but are available on request.

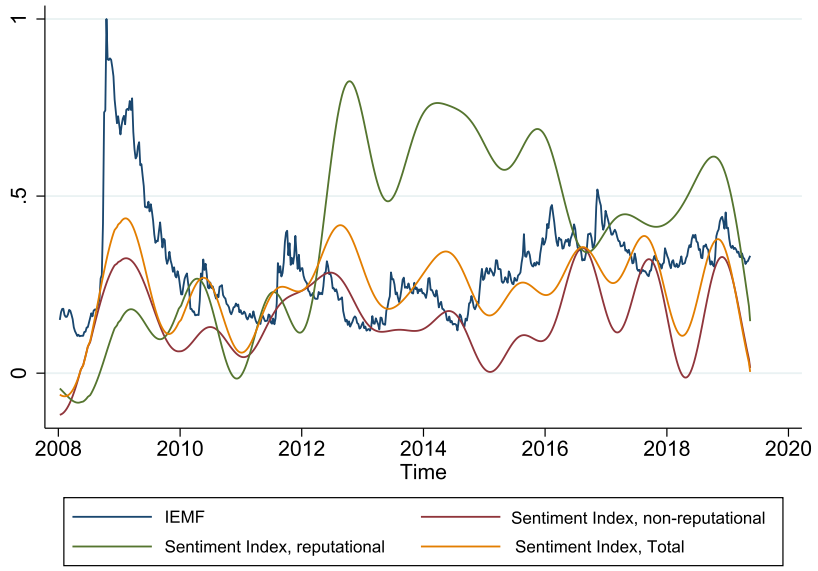

Fig. 4. Comparison of the IEMF and the filtered sentiment indices. The sentiment indices are computed using the majority voting classifier and filtered using the Christiano-Fitgerald filter, with a band of 1 year-100 years.

6. Descriptive results

6.1. The financial market stress index (IEMF)

Systemic risk is a multifaceted phenomenon that is hard to measure at the unidimensional level. To measure systemic risk, one needs to use methodologies that can summarize information coming from many variables with a unique indicator.[10]

Banxico publishes the Índice de Estrés de los Mercados Financieros (IEMF, or Financial Market Stress Index, for its translation in English) (Banco de Mexico, 2013), a financial market stress index with weekly frequency that summarizes in a single variable the information contained in 33 financial variables describing the debt market, the stock market, the foreign exchange market, the derivatives market, credit institutions systemic characteristics, and country risk. The IEMF is built using principal components analysis and its coverage ranges from January 2005 to the present. The index's goal is to create a timely, effective measure that captures the level of accumulated risk in the Mexican financial system at a given moment. A higher index level indicates a higher systemic financial risk.

The IEMF has a very different nature than the sentiment index that we build for this paper. On the one hand, the IEMF is built using "hard" quantitative variables that play a significant role in determining financial market stress. On the other hand, we use "soft" qualitative data (news and opinions reported in social media), and we apply algorithms that interpret the sentiment of this information. We hypothesize that the sentiment index would be correlated with the reaction of financial markets, reflected in the IEMF.[11]

6.2. The filtered sentiment index and its sub-indices

As shown by the topic analysis and suggested by the peaks in Figure 3, the sentiment measured by the Twitter sentiment index correlates to various kinds of negative shocks that can affect the financial sector: financial, macroeconomic, political and reputational. Even though stock market prices might incorporate reputational risk for the banking sector, the IEMF does not measure it explicitly.

We build two sub-indices of the Twitter sentiment index, dividing the sample of tweets into those classified as creating reputational risk according to the LDA algorithm and all others. We follow the same methodology that we use for the Twitter sentiment index to compute the two sub-indices. We define the reputational sentiment index as the sub-index of the Twitter sentiment index built from the tweets that create reputational risk, and we define the non-reputational index as the sub-index of the Twitter sentiment index that contains all the other tweets. Figure 4 shows the Twitter sentiment index, the reputational sentiment index and the non-reputational index, as compared with the IEMF for the period 2008-2019.

[10] Examples of such stress indicators are the St Louis Fed Financial Stress Index (Kliesen and McCracken, 2020), the Chicago Fed National Financial Conditions Index, and the Kansas City Financial Stress Index (Hakkio and Keeton, 2009). Examples in Europe are the Central Bank of Sweden Financial Stress Index (Forss Sandahl et al., 2011) and the European Central Bank Financial Stress Index (Duprey et al., 2017). The International Monetary Fund also publishes Financial Soundness Indicators for emerging market countries (IMF, 2003).

[11] The IEMF is already partially smoothed by construction. For this reason, we do not filter it before comparing it to our smoothed sentiment index.

As in Figure 3, our preferred classification model is the sentiment index based on majority voting. However, Figure 3 presents the baseline sentiment index, which is not filtered, but computed on a monthly basis. In this case, because the IEMF has a weekly frequency, we present the results of the sentiment indices based on the majority voting classifier, with weekly frequency smoothed using the Christiano-Fitzgerald filter with the band starting at the 1-year frequency and extending to the 100-year frequency. It is possible to distinguish two periods in which the Twitter sentiment index presented in Figure 4 was affected by various news shocks. In 2012 the reputational sentiment index increases until it peaks at the end of the year, coinciding with the HSBC scandal. The reputational sentiment index has a second local peak in 2014 during the Oceanografía scandal. After 2015, there are only lower peaks that coincide with news about developments related to past scandals, such as new evidence about the scandals or a new phase in the judicial process. The non-reputational sentiment index follows the Twitter one more closely, and their trends are more in line with the IEMF than the reputational sentiment index.

We find that the peaks of the non-reputational sentiment index follow the IEMF peaks as described in the Financial Stability Reports of Banco de México more closely. In 2011 and 2012, uncertainty about the Greek default and the default risk of systemic banks in Spain make the IEMF spike because BBVA and Santander are also among the main commercial banks in Mexico. We also find that news about bank fragility in Spain and uncertainty about the sovereign default in Greece are reported in our tweet database. However, we find more tweets about the banking scandals occurring during 2012, so the peak of the reputational sentiment index is higher.

In 2013 and 2014, the IEMF indicates spikes in financial risk associated with the publication of the minutes of the Federal Reserve. In June 2013, the Fed announces the slowdown in the Quantitative Easing (QE) program, and the expected end of the program in October 2014. In our database of tweets, we find news about the effects of this announcement in June 2013. However, most of the reaction takes place in 2014. We find a rising number of tweets reporting a decrease in growth expectations for Mexico in the non-reputational sentiment index. In 2014, the Oceanografía scandal also occurred in Mexico. Most of the tweets in our database comment on this, given its negative effect on Citibanamex. The news about Oceanografía spread from February 2014 to August 2014. This scandal, with its negative sentiment, weighs in the most in our Twitter sentiment index and its sub-indices in that year.

During 2015, the IEMF goes through a stabilization first, then experiences an increase in financial stress due in part to the end of the asset purchasing program in the previous year, and to rising expectations of an increase in interest rates by the Federal Reserve, as occurred in December 2015. In 2015, tweets reported news about the depreciation of the peso, weak growth, the increased strength of the dollar, and expected international contagion from the interest rate increase in the US. The non-reputational sentiment index reports a peak in the second part of the year. In the same year the reputational sentiment index has a peak in correspondence of the HSBC money laundering scandal. In 2016, the IEMF shows high financial stress for the entire year, with a peak in the last quarter due to risks linked to external shocks: the electoral process in the US, rising risk of protectionism, low growth in the global economy, and falling oil prices and revenues in Mexico. Regarding the Twitter sentiment index, starting in 2016, the banking scandals and frauds have less weight, so the reputational sentiment index falls. However, we see a rise in the non-reputational index, with tweets reporting on the electoral process and trade tensions.

Financial stress reported by the IEMF decreases in 2017 and 2018 as trade tensions decrease during the renegotiation of the trade agreement with the US and Canada. In 2018, risk increases because of the electoral process in Mexico and uncertainty linked to the USMCA negotiation talks. At the end of 2018, higher volatility and uncertainty in the financial markets are related to domestic factors such as changes in public policies (changes in energy policy, the cancellation of the construction of Mexico City's new international airport). The Twitter sentiment index shows a peak in the second half of 2017 due to news about the earthquake that hit Mexico in September of that year, and it shows another peak at the end of 2018 due to news about the cancellation of the new international airport in Mexico City and the proposed banking commissions reform. Finally, in 2019, the risk increases because of uncertainty over Pemex and Mexico's credit perspectives. In March and June of 2019, Pemex corporate debt and Mexico's sovereign debt suffered a downgrade.

Not all of the two indices's peaks coincide, but both reflect the main news. In addition, the two indices move in the same direction, presumably due to common causes. In other words, the Twitter sentiment index captures information of importance for the systemic risk, and the news reports events that affect financial risk as measured by other indicators.

To determine whether there is a significant correspondence between our Twitter sentiment index and the IEMF, we compute the correlation of the IEMF with the non-reputational sentiment index, the reputational sentiment index, and the Twitter sentiment index, using the majority voting model and the other three classifiers. Given the evidence in Figure 4, we expect a more positive correlation of the Twitter sentiment index and the non-reputational sentiment index with the IEMF than the sub-index built on reputational tweets.

Column 1 of Table 6 shows the correlation between the IEMF and the various unfiltered indices, computed using the sample of tweets for the period 2008-2019. Column 2 shows the coefficients of the correlations between the IEMF and the filtered sentiment indices.[12] In all cases, the filtered version of the sentiment index is more correlated with the IEMF than the non-filtered one is. The filtered voting sentiment index shows the highest correlation with the IEMF, reaching a significant positive correlation of more than 0.40 in the case of the Twitter sentiment index and more than 0.49 for the non-reputational sentiment index. The reputational sentiment index is not significant, or it is negatively correlated with the IEMF, signaling that the non-reputational sentiment index might contain more information about systemic risk.

[12] When we use the term "filtered index" we refer to the version computed using the 1 year -100 years band.

Table 6
Correlation between the alternative sentiment indices and the IEMF.

Correlation with IEMF	Sentiment index, not filtered (1)	Sentiment index, filtered (2)
	Majority voting model	
Non-reputational	0.132*	0.463*
Reputational	-0.051	-0.149*
General	0.134*	0.401*
	Dictionary model	
Non-reputational	0.121*	0.374*
Reputational	0.045	0.038
General	0.133*	0.358*
	SVM model	
Non-reputational	0.117*	0.396*
Reputational	-0.056	-0.142*
General	0.090*	0.225*
	Neural network model	
Non-reputational	-0.023	-0.218*
Reputational	-0.106*	-0.279*
General	-0.0039	-0.187*

* p-value<0.1. Filtered index: obtained by applying the Christiano-Fitzgerald filter with the 1-year-100 years band to the sentiment index based on the majority voting classifier and computed without considering neutral tweets.

As a robustness check, we perform the same correlations using the alternative sentiment models. However, the correlation between the IEMF and these alternatives is lower than those presented for the Twitter sentiment index built using the majority voting rule. The filtered sentiment index obtained using the SVM classifier presents a closer correlation to the filtered Twitter sentiment index built using the majority voting rule: the non-reputational sentiment index is significantly correlated with the IEMF at 0.39, and the Twitter sentiment index based on SVM has a significant and positive correlation with the IEMF of 0.22. The correlation between the filtered Twitter index obtained using the dictionary classifier and the IEMF is higher than the correlation between the filtered Twitter sentiment index obtained using the SVM classifier and the IEMF for the Twitter sentiment index, but it is lower in the other cases. The correlation between the IEMF and the index built on neural networks is negative when we expected it to be positive. The correlation between our sentiment indices and the financial index of reference, the IEMF, are in line with the findings in Shapiro et al. (2017). In their paper, they compute correlations between the sentiment measures they build and various economic outcomes, among them the S&P 500, corresponding to the IPC for Mexican data. The correlations in Shapiro et al. (2017) vary between 0.02 and 0.47. In particular, the S&P 500 is correlated with the sentiment measures by 0.22 at most.

The evidence presented in Figure 4 and Table 6 suggests that our intuition is correct. The non-reputational sentiment index, which was built using textual sources, correlates with the indicator of financial market stress, which was constructed using quantitative variables. The data and the methodologies that we use to build the Twitter sentiment index are different from those used for the IEMF, but the results are similar. The sentiment indicator that we propose could be a useful novel indicator to analyze and forecast financial stress risk. Moreover, a positive contribution of the Twitter sentiment index is that it makes it possible to assess the factors that drive changes in the IEMF index more immediately. As with any structured data series, movements in the IEMF can only be explained by looking at contemporaneous reports of what might drive movements in financial markets. However, as Figure 1 shows, the Twitter sentiment index can explain some of the drivers based on the words and topics of the tweets. This enables us to assess the factors driving the IEMF index. Moreover, one could potentially explain movements in the IEMF with the tweets that explain the sentiment index, when the two indicators move in the same direction.

As a robustness check, we compute regressions of the IEMF on the Twitter sentiment index built using the majority voting rule in the non-filtered and filtered versions (Table 7).

The regression results confirm those obtained with the correlation analysis. When we regress the IEMF on the filtered sentiment indices, the R-squared is higher than in the models where IEMF is regressed on the non-filtered indices. The most interesting results are in the second panel of the table, in columns (2) to (4), where the indices are considered separately. The filtered non-reputational sentiment index is significantly correlated with the IEMF, and the IEMF increases by 0.65 percent when the non-reputational sentiment index increases by 1 percent. The coefficient of the reputational sentiment index is negative and significant, even if it is low in absolute value. This may be explained by the higher proportion of reputational tweets in one year, 2014, when the IEMF decreases, while the reputational sentiment index increases due to the Oceanografía scandal. Finally, the coefficient of the filtered Twitter sentiment index is always positive and significant, both in the regression with all the indices (column 1) and alone (column 4). An increase in 1 percent of the filtered Twitter sentiment index is correlated with an increase of 0.54 percent of the IEMF.

7. Predictive accuracy

We were inspired by the work of Shapiro et al. (2017) to test whether if the filtered version of the Twitter sentiment index contains predictive information for the IEMF or specific financial market indicators. We refer in particular to six variables that we use as proxies

Table 7
OLS regression of the IEMF on the Twitter sentiment index and its sub-indices, obtained by majority voting.

IEMF	(1)	(2)	(3)	(4)
Sentiment index by majority voting				
Non-reputational	-0.019	0.036***		
	(0.03)	(0.011)		
Reputational	-0.032**		-0.014	
	(0.013)		(0.011)	
General	0.068**			0.039***
	(0.033)			(0.012)
Observations	589	589	589	589
R-squared	0.028	0.017	0.003	0.018
Sentiment index by majority voting, filtered				
Non-reputational	-0.15	0.652***		
	-0.125	-0.052		
Reputational	-0.273***		-0.083***	
	-0.03		-0.023	
General	0.977***			0.538***
	-0.136			-0.051
Observations	589	589	589	589
R-squared	0.315	0.215	0.022	0.161

Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. The Twitter sentiment index is obtained by majority voting. When filtered, it is filtered using the Christiano-Fitzgerald filter with band 1 year-100 years. All equations include a constant.

Table 8
Correlations between the sentiment indices obtained by majority voting and selected market variables, for the period 2008-2019.

Sentiment Index	(1) Non-reputational	(2) Reputational	(3) All tweets
Beta	0.268*	0.211*	0.236*
IPC volatility	0.138*	-0.398*	0.038
Exchange rate volatility	0.374*	-0.174*	0.312*
3-month swap rate spread	0.126*	-0.134*	0.143*
EMBI+	0.372*	-0.151*	0.333*
3-month sovereign bond spread	0.360*	-0.265*	0.202*

Note: *: p-value<0.1; filtered sentiment index, for the interval 1 year - 100 years.

of the six types of financial market risk considered in the IEMF. We select our variables, mostly indicators of return volatilities and risk spreads, according to the literature (Hakkio and Keeton, 2009; Holló et al., 2012).

As an indicator of bond market risk, we use the spread between the 3-month Mexican Treasury bill (Certificado de la Tesoreria de la Federación, CETES) yield and the 3-month US Treasury bill. A higher sovereign bond rate relative to a low-risk baseline implies higher rates for all economic agents and higher financial risk.

We use the volatility of the Mexican stock market price index (Índice de Precios y Cotizaciones, IPC) as an indicator of stock market risk. Asset return volatilities tend to increase with investors' uncertainty about future fundamentals or the behaviors and sentiments of other investors.

The 1-month FIX exchange rate volatility is our proxy for foreign exchange market risk. Exchange rate volatility increases exchange rate risk.

As an indicator of derivative market risk, we refer to the spread between the 3-month swap rate and the overnight interbank rate. Swap spreads are indicators of the desire to hedge risk, the cost of that hedge, and the overall liquidity of the market. Larger swap spreads indicate a higher general level of risk aversion in financial markets, and they are indicators of systemic risk.

We use the beta of financial institutions to the IPC as an indicator of credit institutions' risk. Beta is a widely used measure of a stock's volatility to the overall market. The market, as measured by a market index such as the IPC, has a beta of 1. A stock that has higher volatility than the market has a beta higher than 1, and one that is less volatile than the market has a beta between 0 and 1.

Finally, we use the JP Morgan Emerging Market Bond Index Plus (EMBI+) for Mexico as an indicator of country risk. The EMBI+ is a weighted index tracking the rate of return for actively traded and dollar-denominated external debt instruments in emerging markets. The EMBI+ is an equivalent of sovereign spread for emerging economies in that higher EMBI+ corresponds to higher risk.

Table 8 presents the correlations between the selected variables and the three versions of the filtered sentiment index, in line with our previous results. The Twitter sentiment, the reputational, and the non-reputational indeces correlate with the variables considered

with the expected sign. The non-reputational sentiment index correlates positively with each financial variable, as expected, and the correlation is higher than 0.25 in most cases, with the exchange rate volatility and the EMBI+ having the highest correlation (0.37 in both cases). The stock market volatility index and the short-run swap rate have a positive but lower correlation with the non-reputational sentiment index of 0.14 and 0.13, respectively.

These results are in line with the previous literature: (Shapiro et al., 2017) find correlations between their sentiment indices and the growth rate of the S&P 500 index in a range of 0.06 to 0.22 in absolute value. The reputational sentiment index and the financial variables are correlated negatively, as expected from the analysis of the correlations between the IEMF and the sentiment indices. The only exception is the positive correlation between the reputational index and the beta of credit institutions. This indicates that systemic banking risk might be correlated with an increase in reputational risk due to financial fraud or money laundering. The Twitter sentiment index correlates positively with each variable. The correlation coefficient is a bit lower than the non-reputational sentiment index, because the Twitter sentiment index includes both the effect of the non-reputational and the reputational indices.

7.1. Local projections

To explore whether our Twitter sentiment index can predict financial stress and financial conditions, we apply the local projections method developed by Jordà (2005). Local projections are similar to the standard vector auto-regression model (VAR), but they are less restrictive. We emphasize that we do not claim causality for these results. As stated by Shapiro et al. (2017), even if the correlation between sentiment indicators and financial variables exists, the direction of the causality remains unclear. However, given that a correlation exists, it may help to improve predictive models of financial market risk.

For each forecast horizon h, with h=0...26 weeks, we run a different regression of a given financial measure y_j on contemporaneous and lagged values of the Twitter sentiment index and y_j itself:

$$y_{j,t+h} = \alpha_j^h + \beta_j^h SI_t + \sum_{i=1}^{n} \gamma_{j,i}^h SI_{t-i} + \sum_{i=1}^{n} \delta_{j,i}^h y_{j,t-i} + \varepsilon_{j,t+h} \tag{2}$$

where y_j represents the variable of interest, SI is the sentiment index filtered using the 1 year-100 years band, and n is the number of lags that each equation contains. We consider the specification that includes the Twitter sentiment index as our baseline, and we select the number of lags according to the Schwartz Bayesian Information Criteria (SBIC), which is considered optimal for the local projection model (Brugnolini, 2018).

To compare the forecasting power of a model that includes our filtered Twitter sentiment index and a model that does not consider it, we report the SBIC, which measures the models' fit. To keep the models comparable, we compute the SBIC for three models: an AR(1), AR(4), and AR(12).[13] In all cases, first we compute the model in which we include only the dependent variable y_j and its lags, and then we compute the same model considering both y_j and the Twitter sentiment index as an exogenous variable. We calculate the SBIC of each model adding one lag at a time, up to 24 lags. The lower the optimal SBIC is, the more forecasting ability the model has, so if the optimal SBIC is lower when the model includes the sentiment index, it means that the sentiment index contains some predictive information about the variable of interest.

Figure 5 reports the SBIC for the AR(1) model.

The first model does not include the sentiment index, the second considers the Twitter sentiment index among the independent variables, and the third incorporates the non-reputational sentiment index. The results are qualitatively similar also when we compute the SBIC for the AR(4) and the AR(12) models. In all cases, the models that include a sentiment index, Twitter or non-reputational, show a lower SBIC than the model that does not include any sentiment index. The model that includes the non-reputational sentiment index seems to have slightly higher predictive power than the model that includes the Twitter sentiment index. These results imply that the Twitter sentiment index improves the forecasting ability of a model that considers only the dependent variable.

Finally, we use local projections based on Equation (2) to analyze the impact of a one standard deviation shock of the filtered Twitter sentiment index on each of the variables of interest. A positive shock (i.e., a shock that increases the Twitter sentiment index) would be positively correlated with negative sentiment about financial markets and banks, so it might increase financial market risk. The results in Figure 6 confirm this hypothesis. A one standard deviation shock on the Twitter sentiment index correlates with an increase in the IEMF, which becomes significant after three weeks. The effect on the IEMF reaches its peak after 20 weeks, and starts to decline afterwards.

Figure 7 presents the impulse response functions of a one-standard-deviation shock of the Twitter sentiment index on the selected financial variables. A positive one-standard-deviation shock significantly correlates with an increase in exchange rate volatility and stock market volatility in the first 10 weeks after the shock. There is also a significant increase in the correlation with country risk as measured by the EMBI+ for Mexico. It increases at the moment of the shock, and reaches a peak of 1.2 standard deviations after 20 weeks. Similarly, the 3-month sovereign bond spread, the indicator of bond market risk, is positively correlated with a shock in the Twitter sentiment index. The banking sector, proxied by the beta of financial institutions, also reacts to a shock in the Twitter sentiment index with an increase, although this reaction is not significant in the short run.

[13] We also compute the AIC criteria with similar results.

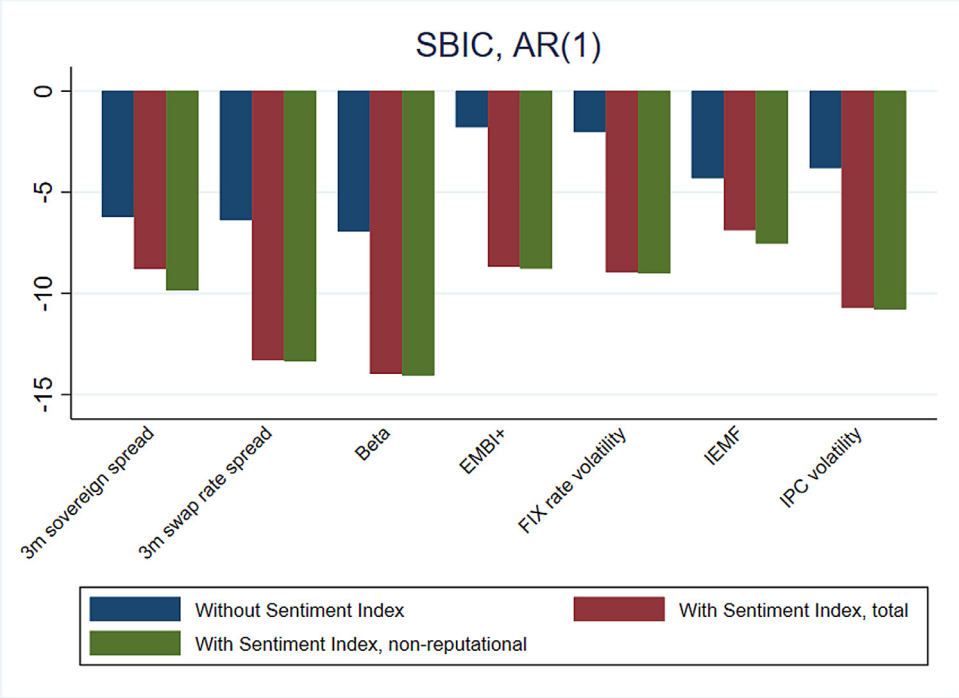

Fig. 5. Comparison of the Schwarz Bayesian Information Criteria computed for a VAR(1) model without any sentiment index among the exogenous variables and two models that include a type of sentiment index as an exogenous variable. The dependent variable of each model is specified on the x axis. The sentiment indices are filtered using the Christiano-Fitgerald filter, with the 1 year-100 years band.

These results show that an increase in the negative sentiment regarding Mexican banks and financial markets is positively correlated with a risk increase in the financial sector as a whole, as measured by the IEMF, and in specific market segments, such as stock market risk, country risk, foreign exchange risk, and the banking sector.

7.2. Robustness checks

As a robustness check, we run the same analysis using the non-reputational sentiment index instead of the Twitter sentiment index. Figures 8 and 9 show the effect of a one-standard-deviation shock on the IEMF and the selected financial variables.

In all cases the reaction of each variable to a shock of the non-reputational index is similar to the previous case. The correlation between the non-reputational sentiment index and the IEMF is positive and significant; it seems stronger than the correlation between the IEMF and the Twitter sentiment index (Figure 8).

Observing Figure 9, we see that the positive correlation between the non-reputational sentiment index is stronger and the effect seems more persistent over time. The only exception is the short-run sovereign bond spread, our proxy for bond market risk, that shows a positive but non-significant reaction to a shock in the non-reputational sentiment index.

Finally, we test whether using different financial variables as proxies for the various market risks obtains similar results to Figure 7. We use the 10-year sovereign bond spread as a proxy for bond market risk, the EMBI+ corporate for Mexico as a proxy for country risk, the spread between 5-year swap rate and 5-year fixed rate sovereign bond as a proxy for derivative market risk, the annual growth of the FIX exchange rate as a proxy for foreign exchange market risk, the annual yield of IPC as a proxy for stock market risk, and finally the spread between the maximum value and the minimum value of daily banking funding rate as a proxy for credit institutions' risk.

Figure 10 shows that the results stay broadly consistent for each kind of market risk, even when we use different financial variables as proxies.

The banking funding rate spread reacts positively to a one-standard-deviation shock of the non-reputational sentiment index, and the effect is significant up to 10 weeks after the shock. The stock market yield reacts negatively to an increase of negative sentiment, reaching a trough after 10 weeks. The exchange rate growth is positively correlated with an increase of the sentiment index, implying that an increase of negative sentiment regarding financial markets is correlated with higher depreciation. The country risk measured from the point of view of the corporate sector reacts positively to an increase of negative sentiment, similar to the case when we used country risk measured as sovereign risk. Our indicator of derivative risk, the 5-year swap rate spread, has a negative but non-significant reaction to a shock in the sentiment index, similar to what we saw in Figure 7 with the 3-month swap rate spread. Finally, the 10-year sovereign bond spread is positively correlated with an increase in the sentiment index, but the effect is not significant.

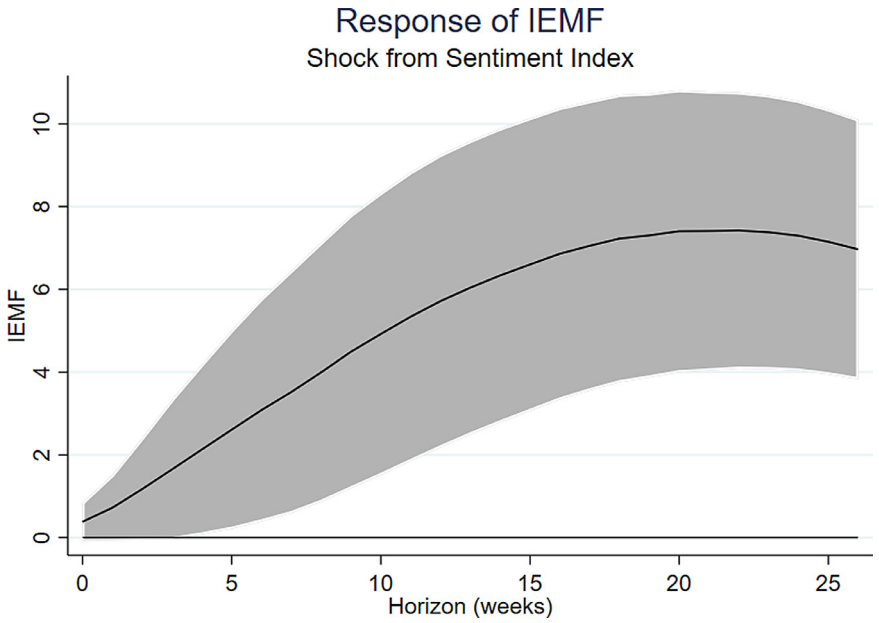

Fig. 6. Impulse response functions of the IEMF to a one-standard-deviation shock of the Twitter sentiment index. The Twitter sentiment index is filtered using the Christiano-Fitzgerald filter with the 1 year-100 years band. Forecasting horizon of 26 weeks.

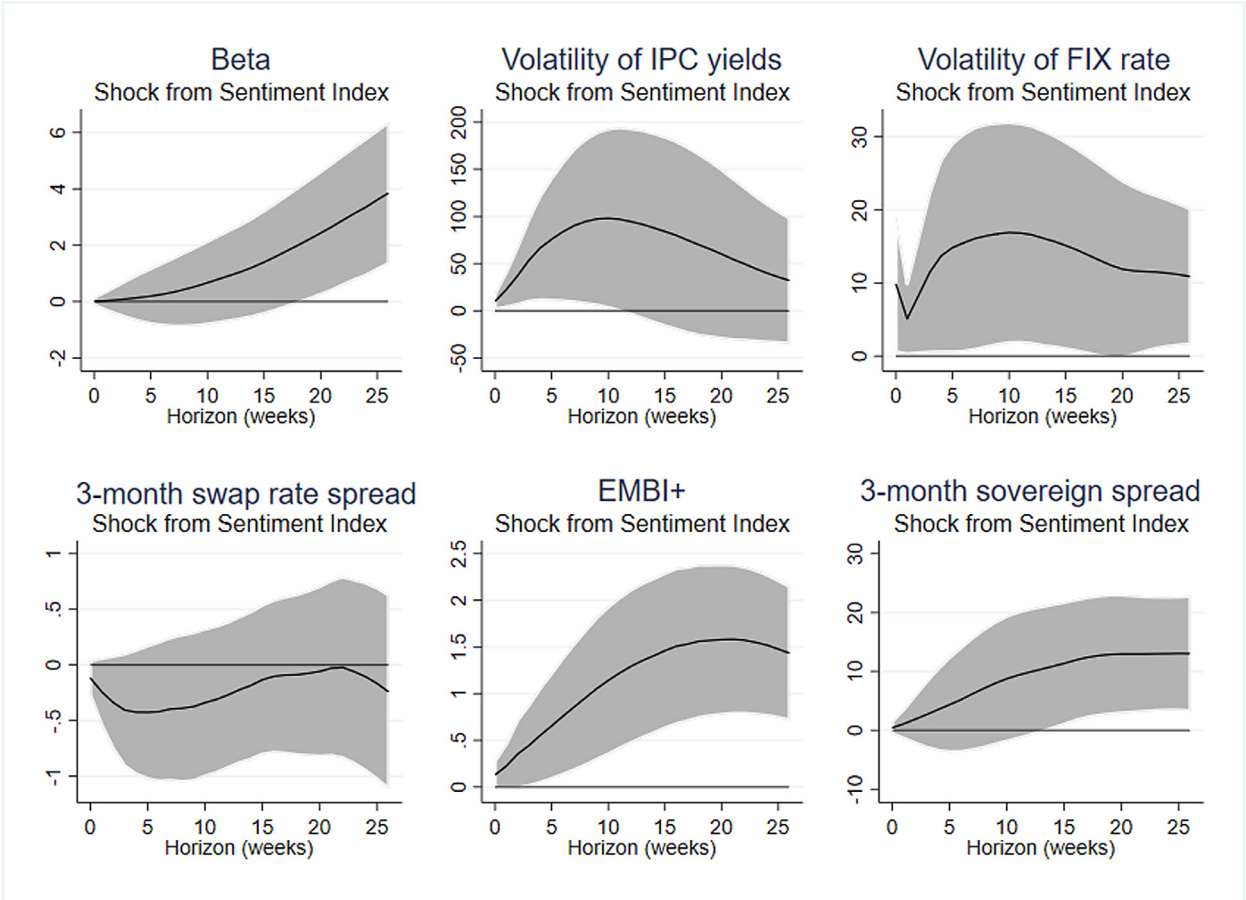

Fig. 7. Impulse response functions of selected financial variables to a one-standard-deviation shock of the Twitter sentiment index. The Twitter sentiment index is filtered using the Christiano-Fitzgerald filter with the 1 year-100 years band. Forecasting horizon: 26 weeks.

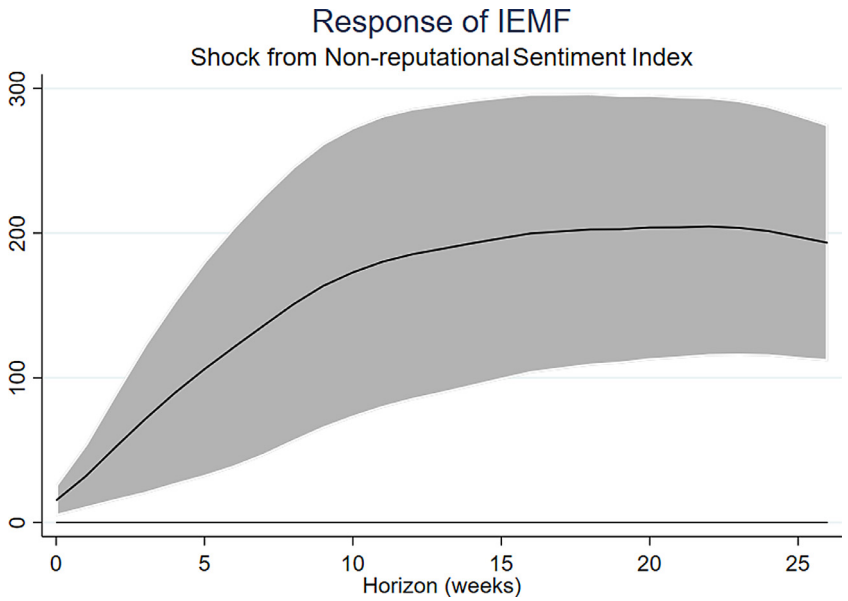

Fig. 8. Impulse response functions of the IEMF to a one-standard-deviation shock of the non-reputational sentiment index. The non-reputational sentiment index is filtered using the Christiano-Fitzgerald filter with the 1 year-100 years band. Forecasting horizon: 26 weeks.

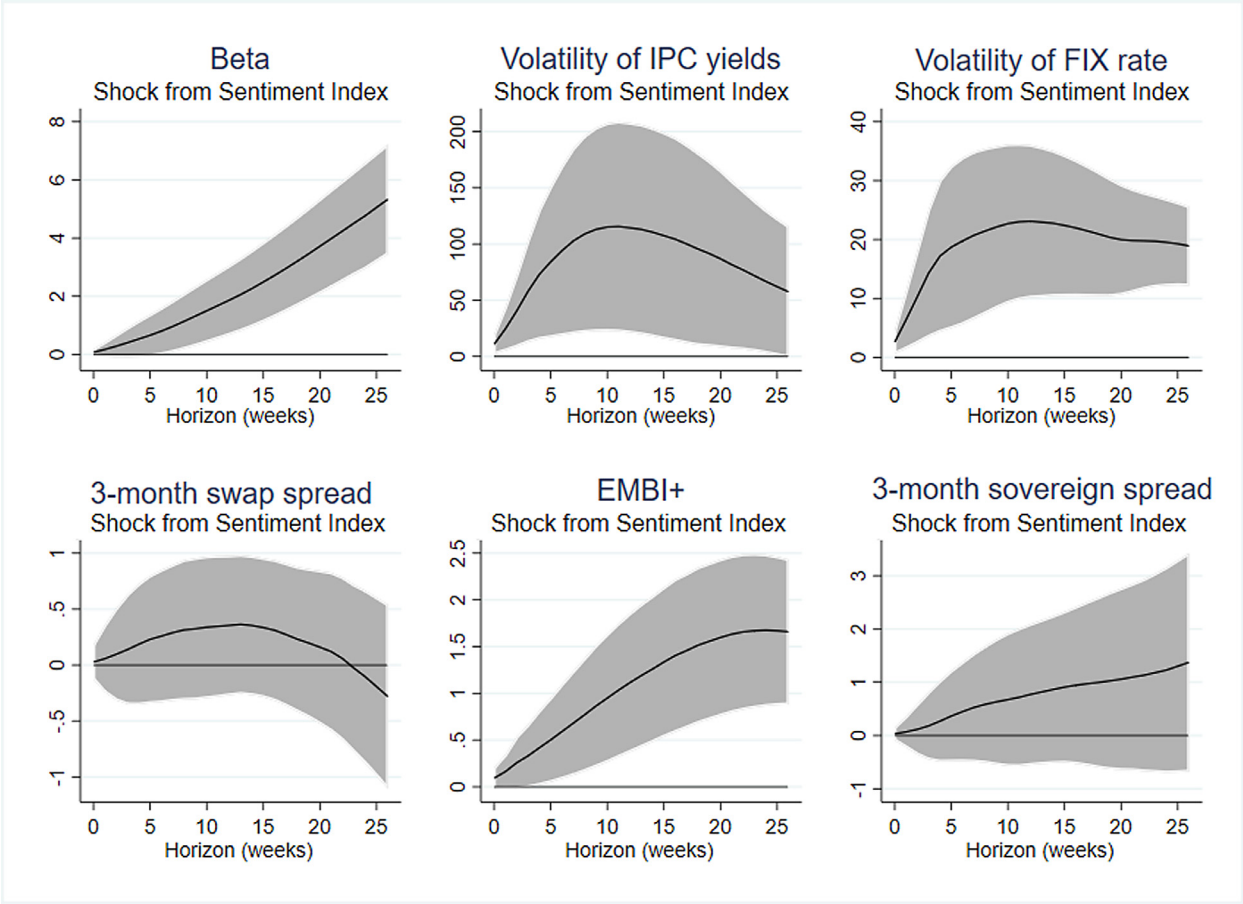

Fig. 9. Impulse response functions of selected financial variables to a one-standard-deviation shock of the non-reputational sentiment index. The non-reputational sentiment index is filtered using the Christiano-Fitzgerald filter with the 1 year-100 years band. Forecasting horizon: 26 weeks.

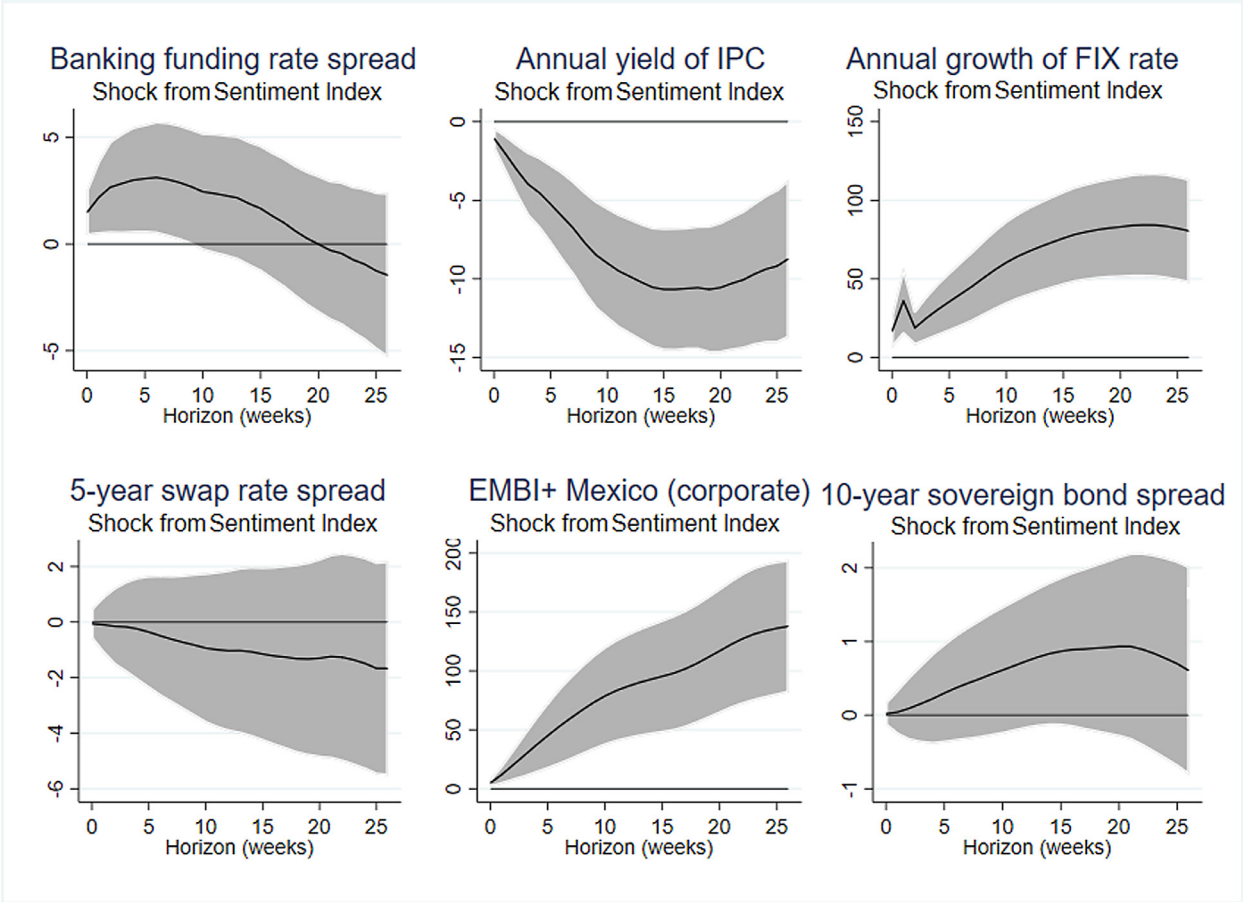

Fig. 10. Robustness check: Impulse response functions of alternative financial variables to a one-standard-deviation shock of the Twitter sentiment index. The Twitter sentiment index is filtered using the Christiano- Fitzgerald filter with the band 1 year-100 years. Forecasting horizon: 26 weeks.

7.3. Banking risk and reputational sentiment

In this paper we focus mainly on the effect of the Twitter sentiment index on the IEMF, which is an aggregate measure of financial market stress. However, it would be interesting to verify whether our Twitter sentiment index, build using exclusively tweets about Mexican commercial banks, may have some power to predict banking sector's soundness.

As a first test, we substitute the risk of credit institutions from to the total IEMF as a dependent variable in equation (2). Figure 11 shows the results. The risk of credit institutions positively correlates with a shock to the Twitter sentiment index or one of its sub-indices (reputational sentiment index and non-reputational one).

Our second exercise is to explore whether the reputational sentiment index has a stronger correlation with the risk of credit institutions during the period of higher reputational risk, as signaled by the reputational sentiment index (from approximately 2011 to the end of 2015). We compare the local projections results that we obtain for this interval (reputational stress period) and those obtained in the non-reputational stress period (2008-2010, 2016-2019). The results are presented in Figures 12 and 13. We perform the exercise using all three indicators of banking risk as dependent variables: the credit institutions' risk component of the IEMF, the beta of financial institutions, and the bank funding rate spread. We find that a one-standard-deviation shock of the reputational sentiment index positively correlates with the three indicators, especially with the credit institutions component of the IEMF. In the high reputational stress period, the correlation is stronger, as expected, while during the period of low reputational stress, the correlation is still positive but lower in absolute value. The results for the bank funding rate spread are in line with the results of the IEMF, while the correlation between the shock of the reputational sentiment index and the beta of financial institutions seems insignificant in the short run, increasing in the period when the reputational index is more in line with the systemic sentiment index.

8. Conclusion

Our paper contributes to the growing literature that applies sentiment analysis to textual data to construct novel indicators for economic and financial analysis. Sentiment indices can help to forecast not only economic variables - for instance, in nowcasting

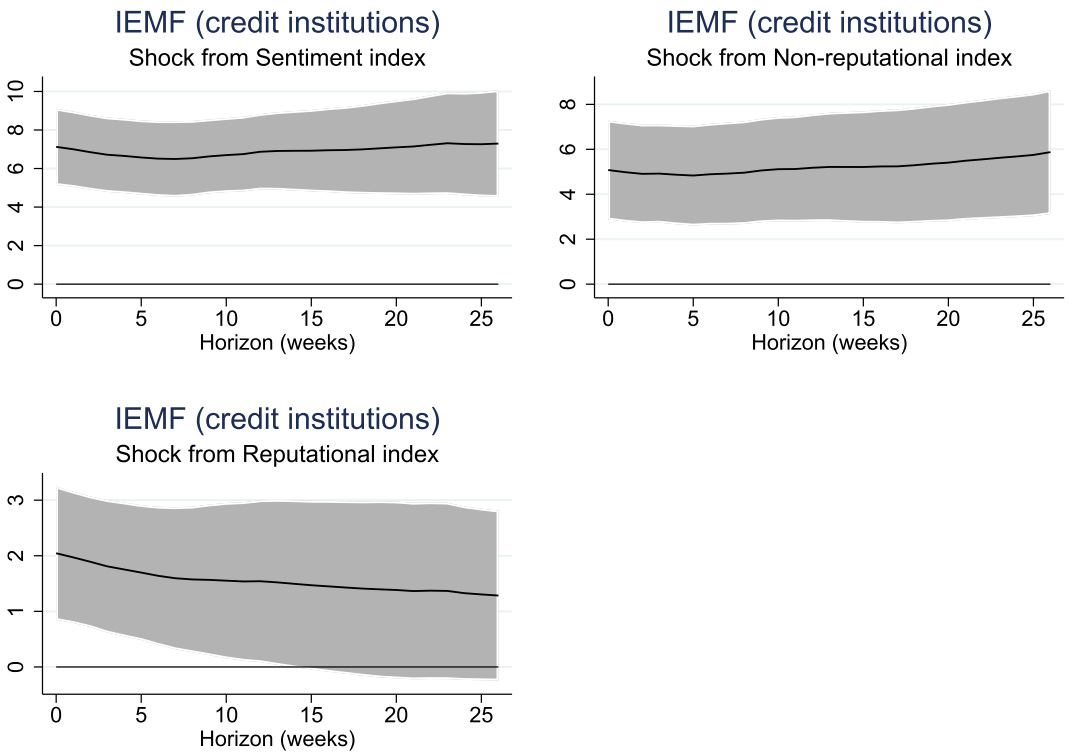

Fig. 11. Comparison between the Impulse response functions of the credit institutions' risk component of the IEMF to a one-standard-deviation shock of alternative sentiment indices. Forecasting horizon: 26 weeks.

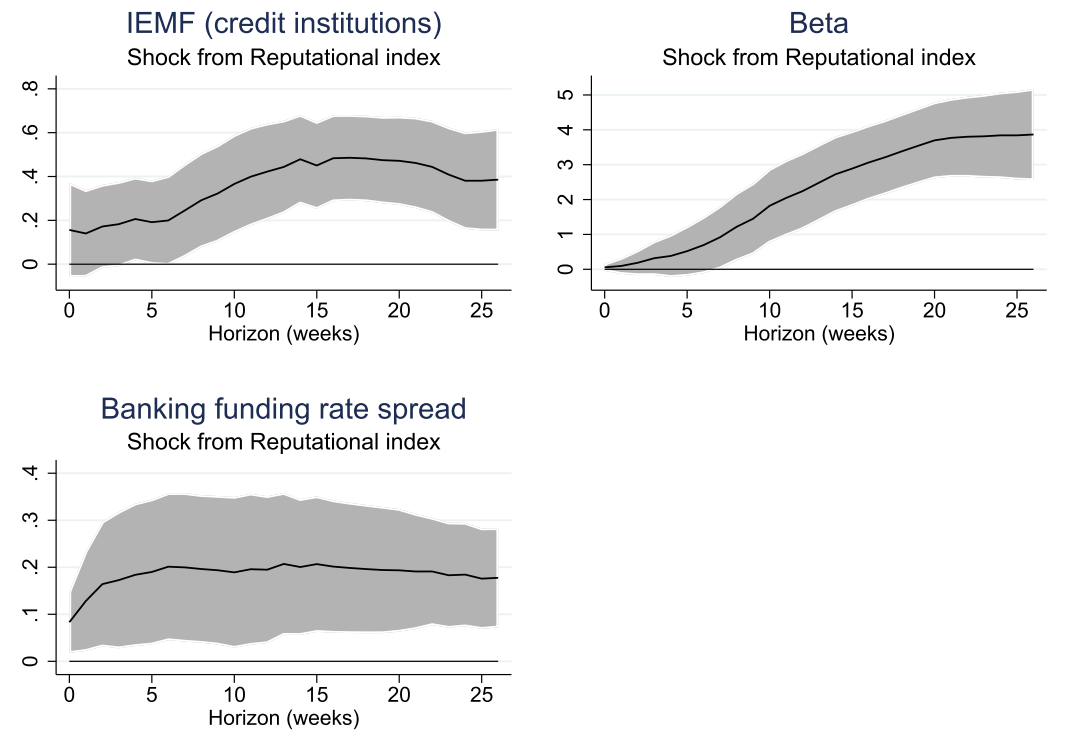

Fig. 12. Comparison between the Impulse response functions of the credit institutions' risk component of the IEMF and banking risk indicators to a one-standard-deviation shock of the reputational sentiment index (low reputational risk period). Forecasting horizon: 26 weeks.

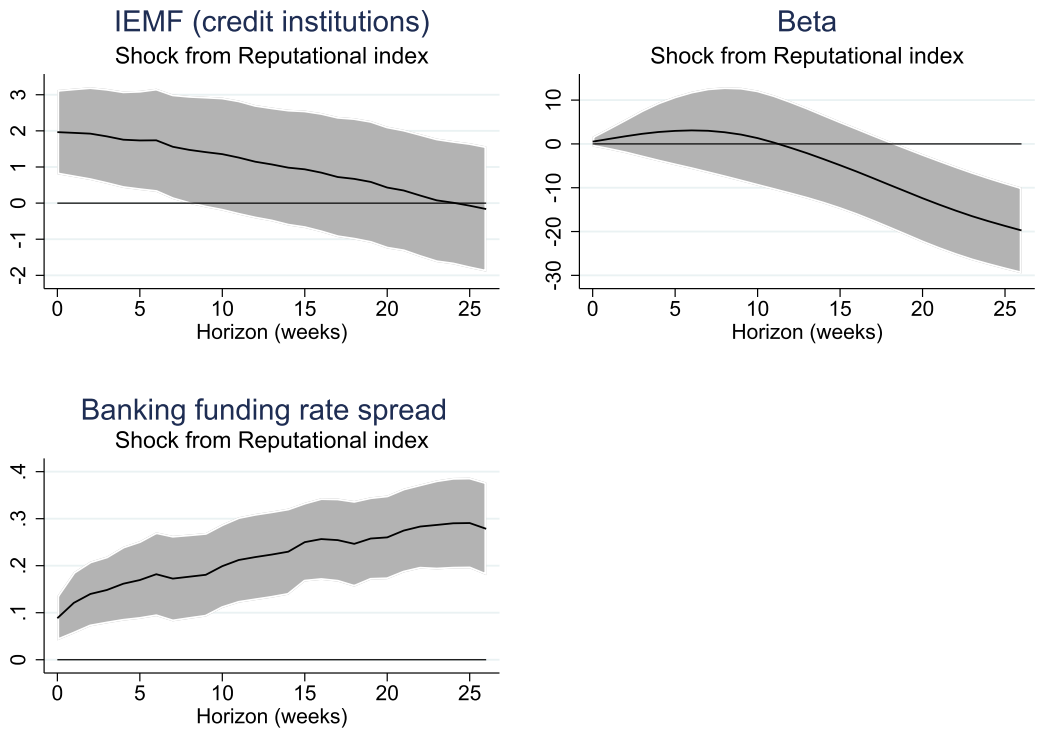

Fig. 13. Comparison between the Impulse response functions of the credit institutions' risk component of the IEMF and banking risk indicators to a one-standard-deviation shock of the reputational sentiment index, 2011-2015 (high reputational risk period). Forecasting horizon: 26 weeks.

exercises - but also financial variables through the information demand of retail investors. In this paper, we propose a new sentiment index for Mexico based on the analysis of Twitter messages. We use three different machine learning models to analyze the sentiment of Twitter messages, and build alternative sentiment indices to inform the analysis of financial market risk.

We first extract tweets in Spanish from Twitter - in the period April 2006-June 2019. We select tweets that report information that may have an impact on banking and financial risks. We use the LDA algorithm to perform a topic analysis to classify the content related to the Mexican financial system, identifying some topics not traditionally included in financial stress risk indices, such as financial frauds, money laundering, and failures of online payment systems.

We consider three sentiment classifiers (one based on word counts, a linear classifier, and one based on neural networks) to build the sentiment index for the Mexican financial sector. Finally, we combine the three sentiment indices using a majority voting scheme.

We apply local projections to test the effect of a shock of our sentiment index on selected market variables. A one-standard-deviation shock in the sentiment index significantly correlates with an increase in exchange rate volatility and stock market volatility in the first ten weeks after the shock. The Twitter sentiment index also correlates with an increase in country risk as measured by the EMBI+ for Mexico. We also find that the banking sector reacts to an unanticipated rise in the sentiment index.

In future research, we plan to develop the analysis further to explore the direction of causality between our Twitter sentiment index and indicators of financial market risk in more detail.

Appendix A

A1. The latent dirichlet allocation algorithm

The LDA algorithm (Blei et al. (2003); Bruno et al. (2018b)), is commonly used for topic modeling. LDA is a generative probabilistic model that facilitates the discovery of abstract topics that occur in a collection of documents. This model assumes that each document in the corpus is modeled as a distribution of topics, and that each topic is modeled as a distribution of words. The goal is to find the most relevant topics that represent the corpus of documents. The output of the model is the distribution of topics over documents and the distribution of words over topics. The model automatically divides the corpus of documents in groups of words (that may overlap), but the interpretation of this result and the labeling of the topics is the researcher's responsibility. When fitting the LDA model, the user is responsible for choosing the number of topics to be inferred from the collection of documents. Once the researcher chooses the number of topics and the model is fitted, the resulting topics can be interpreted based on their most representative words.

A certain degree of subjectivity is unavoidable in the interpretation of the topics, but there are some indicators to compare the performance of different models and support the user in this task. Among the indicators to evaluate the performance of LDA as a

topic model are topic coherence indicators, for example, UMass coherence (Mimno et al., 2011) and the UCI measure (Newman et al., 2010). These indicators are especially helpful for distinguishing whether a topic is semantically interpretable. Intuitively, within the words used to describe a topic, the UMass score measures to what extent a common word is (on average) a good predictor for a less common word. The higher the score, the more coherent the topic is.

A1.1. Topics found with the LDA analysis

We assign labels to the six topics we find with the LDA analysis considering the most frequent words in each topic found by the LDA algorithm and analyzing selected tweets from those contained in each topic. We list the most relevant words for each topic below:[14]

1. Financial markets (top 15 terms: “earnings,” “dollar,” “million,” “to increase,” “to sell,” “bmv” -acronym for Bolsa Mexicana de Valores, the Mexican Stock Market-, “to fine,” “bond,” “to close,” “euro,” “to announce,” “biggest,” “to fall,” “stock market,” “loss”)
2. Macroeconomic expectations (“to give,” “to maintain,” “to signal,” “credit,” “to warn,” “economy,” “risk,” “bank,” “country,” “to emphasize,” “to drive,” “rating,” “growth,” “to weight,” “to tell”)
3. Foreign exchange market (“dollar,” “financial group,” “to forecast,” “to buy,” “to sell,” “sale,” “cent,” “country,” “pension fund,” “to see,” “exchange rate,” “to close,” “counter,” “peso”)
4. Business activity (“operation,” “service,” “client,” “to report,” “first,” “financial group,” “to present,” “to buy,” “credit,” “failure,” “better,” “bank,” “branch,” “to offer,” “digital”)
5. Financial results (“gain,” “forecast,” “to report,” “to achieve,” “first quarter,” “fund,” “to expect,” “to present,” “to buy,” “to value,” “to tie,” “growth,” “to announce,” “to fall,” “to raise”)
6. Illicit activities and penalties (“client,” “money,” “to put,” “to investigate,” “to present,” “to count,” “manager,” “credit,” “to fine,” “opinion,” “to ask,” “to charge,” “oceanografía,” “card,” “fraud”)

We consider financial frauds and money laundering as negative shocks for the reputation of a bank, both regarding a bank headquartered in Mexico or an international bank headquartered abroad with a Mexican subsidiary. Reputational risk is the “risk arising from negative perception on the part of customers, counterparties, shareholders, investors, debt-holders, market analysts, other relevant parties or regulators that can adversely affect a bank’s ability to maintain existing, or establish new, business relationships and continued access to sources of funding” (BIS, 2009, p 19). Adverse events typically associated with reputational risk include ethics violations (such as money laundering operations), safety issues (such as failure in payment systems or online frauds), a lack of sustainability, poor quality, and lack of or unethical innovation (Ingo, 2011). These kinds of activities primarily affect the specific bank that incurred the adverse event, but they also have potential systemic effects, to the extent that the Financial Stability Board and the BIS (BIS, 2017) released specific guidelines describing how banks should include risks related to money laundering within their overall risk management frameworks. Moreover, Banco de México monitors banking cybersecurity and the safety of electronic payment systems as part of its financial supervision duties. Banxico’s Financial Stability Report (Banco de Mexico, 2019) signals that cyber risks can damage financial institutions, disrupting IT systems and causing fail in the service, compromising the integrity of the information managed by the institution, and causing financial losses to the institution or its clients. Additionally, the reputational shock caused by cyberattacks may lower confidence in the financial system, especially considering a cyberattack on a systemically important bank.

A2. Labeling

Other than the basic classification rule that we present in the main text, we apply special care to the classification of various groups of tweets that present common characteristics and may be more challenging to classify as positive, negative, or neutral tweets.

The first group contains tweets reporting news about foreign banks that have subsidiaries located both in Mexico and other countries. It has been widely shown that the banking sector has a significant role in the international transmission of policy shocks and financial risk (Buch et al., 2019; Cetorelli and Goldberg, 2011; Reinhardt and Sowerbutts, 2015). In particular, the banking system in Mexico was affected by foreign shocks, occurred in Spain or the US through the cross-border transmission of the shocks from headquarter banks to branches and subsidiaries during the global crisis (Alcaraz et al., 2019; Morais et al., 2019; Tripathy, 2020). For this reason, we consider that news about the headquarters of foreign banks that hold subsidiaries in Mexico may also affect the Mexican financial sector. However, we consider that news about other subsidiaries or branches of the same banks located in countries other than Mexico may have an impact on the headquarter bank, but not on the Mexican subsidiary. For instance, news about BBVA in Spain or Citigroup in the US may have an impact in Mexico. News about a subsidiary of BBVA in Peru may have a

[14] Original terms in Spanish: Financial markets: “ganancia,” “dólar,” “millón,” “aumentar,” “vender,” “bmv,” “multar,” “bono,” “cerrar,” “euro,” “anunciar,” “mayor,” “caer,” “bolsa,” “pérdida.” Macroeconomic expectations: “dar,” “mantener,” “señalar,” “crédito,” “alertar,” “economía,” “riesgo,” “banca,” “país,” “destacar,” “impulsar,” “calificación,” “crecimiento,” “pesar,” “decir.” Foreign exchange market: “dólar,” “grupo_financiero,” “prever,” “comprar,” “vender,” “venta,” “centavo,” “país,” “afore,” “ver,” “tipo de cambio,” “cerrar,” “ventanilla,” “peso.” Business activity: “operación,” “servicio,” “cliente,” “reportar,” “primero,” “grupo_financiero,” “presentar,” “comprar,” “crédito,” “fallo,” “mejor,” “banca,” “sucursal,” “ofrecer,” “digital.” Financial results: “ganancia,” “previsión,” “reportar,” “centrar,” “primer trimestre,” “fondo,” “prever,” “presentar,” “comprar,” “tasa,” “ligar,” “crecimiento,” “anunciar,” “caer,” “elevar.” Illicit activities and penalties: “cliente,” “dinero,” “poner,” “investigar,” “presentar,” “contar,” “directivo,” “crédito,” “multar,” “opinión,” “pedir,” “acusar,” “oceanografía,” “tarjeta,” “fraude.”

direct impact on BBVA Spain, but it is unlikely that the news would also have an indirect effect on BBVA Mexico. For this reason, we ask our volunteers to consider news about bank subsidiaries not located in Mexico as neutral by default, and to evaluate as positive or negative only news concerning events occurring in Mexico or in the headquarter countries of Mexican banks.

The second group of special news regards economic news about Mexico or the global economy. These kinds of tweets are more common in the topic of macroeconomic expectations, and they report news highlighted by the briefs published by commercial banks in Mexico. The sentiment of these tweets is classified as neutral unless the news directly impacts the Mexican financial system. For instance, a tweet reporting news about how Spain is a risk for the eurozone ("España mayor riesgo para eurozona, Bank of America", tweeted June 28, 2012), is a non-neutral tweet (negative, in this case). This is because Spain being a risk for the eurozone implies that the Spanish country risk is very high, with potential spillovers to the Spanish banking system, as well as the Mexican banking system through cross-country contagion. A tweet that reports news about the World Bank's denial to intervene in the Greek crisis ("Banco Mundial nega sugerencia de involucrarse en Grecia, Banco Base informa" tweeted on June 14th, 2012), is considered neutral. It is potentially negative news for Greece, but it is not immediately clear how it may impact Mexico.

The last special group of tweets are those reporting news regarding protagonists of Mexican or international politics, finance or business community. One of these tweets is considered neutral by default unless it reports an explicit positive or negative judgment. The rationale is to maintain the unbiased sample regarding day-to-day political decisions or business strategies. If a judgment is explicit, it comes from our set of news, not from an unconscious bias of the labeling volunteers. We select tweets published by a broad sample of media, so we expect that we may find partisan judgment but try to minimize this effect.

These criteria were shared with the volunteers who participate in the labeling process. Each tweet is classified by at least two volunteers using the values of 1 for "higher risk," -1 for "lower risk," and 0 for the "neutral" categories. Juxtaposing the sub-samples assigned to the volunteers allows us to obtain more than one label for the same tweet, assuring robustness in the labeling. In case a tweet is labeled with different values, we apply a majority voting approach to assign the final label or personally revise the tweet to select the correct label.

A3. Sentiment classifiers

A3.1. Dictionary with word polarities

The first method we choose for the sentiment classification task is the dictionary with word polarities. This method is particularly valuable because it does not require labeled data to train the model. However, it does require a domain-specific or context-specific dictionary to obtain a reasonable performance. The greatest limitation of this method is its low flexibility to adapt to new data. For instance, if there is a shift in vocabulary or popular expressions along time periods, a dictionary tuned to a specific time period may perform poorly if used to classify information of another time period. Nevertheless, considering the high costs associated with labeling data, we chose this useful alternative as our baseline methodology. We use Correa et al. (2017a) financial dictionary that was built using words from the Financial Stability Reports (FSRs) of 64 institutions published between 2000 and 2015. The dictionary is a refinement of general dictionaries and finance-specific dictionaries proposed in the literature. It contains 391 words, of which 96 are positive and 295 are negative. Although Correa et al. (2017a) tailored their dictionary (from now on, CKJM dictionary) to assess sentiment in a financial stability context, we cannot use it as it is in our sentiment analysis, for three reasons.

First, the FSRs of Banco de México are not included in their sample, so the vocabulary in our data may differ from that in the dictionary. To measure the overlap between the CKJM dictionary and Banxico's FSRs language, we perform text analysis on the FSRs published by Banxico in English from 2006 to 2016.[15] We find a correspondence of 58 percent between the CKJM dictionary and the words used in Banxico's FSRs.

Second, the CKJM dictionary is in English, while our focus is on tweets in Spanish. We translate the CKJM dictionary from English to Spanish, controlling for semantic differences. The correspondence between our translation of the CKJM dictionary and Banxico's FSRs published in Spanish is 50 percent. We expect a lower correspondence than the one obtained between the original dictionary and the FSRs in English, because the two languages have different characteristics and the construction of sentences in Spanish differs from English.

Third, we are not applying the financial stability dictionary to FSRs, but to tweets. The CKJM dictionary is specifically tailored for the context and structure of FSRs and (Correa et al., 2017a) highlight the importance of adapting a dictionary to the specific context in which the text analysis will be performed. Although we focus our search on reliable sources and expect well written tweets, we acknowledge that news reported on Twitter regarding the financial sector may be different from what is reported in an FSR. To find potential keywords that are specific to the context of Twitter news in Mexico, we refer to the sample of 2000 previously labeled tweets. The tweets in this sample have been classified as positive, neutral, or negative by the volunteers that helped in the labeling step (Section 3.4).

We take into consideration only the two groups of tweets labeled as positive or negative and apply the TF-IDF weighting scheme to the two sub-samples of tweets to identify the most relevant terms used in the tweets in each category.[16] We include these words in our original dictionary with the correspondent word polarities. Table A1 presents an extract of the words in the original CKJM dictionary that appear more frequently in the English version of Banxico's FSRs, an extract of the more frequent Spanish words used in Banxico's FSRs, and the most frequent negative and positive words used in our sample of tweets.

[15] We used the Python package pyPDF for PDF content extraction and a word count.

[16] TF-IDF is a commonly used tool in natural language processing. It computes a weight that represents the importance of terms in a collection of documents, considering how many times they appear in multiple documents. See Bholat et al. (2015).

Table A1
CKJM dictionary modified to take into account the most frequent words in our sample of tweets.

Most frequent words in English reports			Most frequent words in Spanish reports			Words with stronger polarity in tweets		
Word	Polarity	Frequency in reports	Word	Polarity	Frequency in reports	Word	Polarity	TF-IDF score
losses	-1	96	morosidad	-1	84	multar	-1	0.0032
contagion	-1	52	volatilidad	-1	80	investigar	-1	0.0027
stable	1	44	estable	1	60	manipulación	-1	0.002
volatility	-1	38	tiempo	-1	60	incumplir	-1	0.0018
adverse	-1	36	contagio	-1	54	blanquear	-1	0.0014
positive	1	36	deterioro	-1	52	solidez	1	0.0019
grew	1	32	mitigar	1	50	impulsar	1	0.0016
recession	-1	32	exposición	-1	42	fortaleza	1	0.0011
contraction	-1	28	incumplimiento	-1	42	sanar	1	0.0005
slowdown	-1	28	cierre	-1	40	garantizar	1	0.0005

First set of words: most frequent words present in the original CKJM dictionary in English and in the FSR of Banco de México written in English. Second set: most frequent words present in the CKJM dictionary translated in Spanish and in the FSR of Banco de México written in Spanish. Third set: most frequent words present in our sample of tweets, according to the TF-IDF weighting scheme.

Computing the tweet sentiment

To perform the sentiment classification of each tweet, we use the previously mentioned dictionary with word polarities (WP): a value of 1 for positive-oriented terms and a value of -1 for negative-oriented terms. Positive-oriented terms are all the words that reduce banking risk, and negative-oriented terms are those that increase the banking risk. For all terms that do not appear in the dictionary, the WP is considered to be zero. The sentiment score of a tweet is computed as the sum of the WPs of all the terms in the correspondent tweet:

$$Sentiment\ score\ for\ a\ tweet = \sum_{i=1}^{n} WP_i \tag{3}$$

Where n represents the number of terms in a tweet. We perform these word counts over the tweets as shown in the example. In this case, the tweet is negative, because there are two negative words and only one positive word:

$$\underset{+1}{El\ beneficio}\ \underbrace{\underset{0}{de\ Bank\ of\ America}}\ \underset{-1}{se\ desploma}\ \underbrace{\underset{0}{por\ las}}\ \underset{-1}{multas}$$
$$(Bank\ of\ America's\ gain\ plummeted\ because\ of\ fines)$$

After obtaining the sentiment score for each tweet, we turn the scores into categorical variables. We assign the value -1 to tweets with a negative sentiment score, 1 to those with a positive sentiment score, and keep the value of 0 for tweets with a score of zero. The use of a dictionary is practical and convenient, since sentiment classification can be done without a previous data labeling step. This methodology is especially efficient when the text analysis is performed on a closed set of documents with a specific terminology and a clear interpretation. Although we adapt the CKJM original dictionary to our specific context, this method is not ideal to analyze text messages in social networks because the body of text evolves over time, the language is more informal, and sentiment can be expressed using irony or sarcasm, images like emoticons, hashtags, or neologisms linked to current events. For this reason we explore two other methods for text classification, but keep the dictionary method as our baseline. We could directly test the performance of this method only on the whole sample of labeled tweets because a training step is not required here. Nonetheless, we also test the performance of the dictionary classifier on the labeled data, the training sample for the other two classifiers, for comparison with the other models.

A3.2. Multilingual sentiment analysis

An alternative model for building our sentiment classifier is the Baseline for Multilingual Sentiment Analysis (B4MSA) model developed by Tellez et al. (2017). B4MSA is a Python-based sentiment classifier built specifically to analyze tweets. While most of the literature focuses on social media analysis in English, this approach can be used to classify sentiment of tweets in any given language.

This model is based on a support vector machine classifier (SVM). An SVM classifier (Boser et al., 1992) is a more refined classification model than the one based on the dictionary approach. Unlike the dictionary approach, it does not use a simple dictionary of words with a given polarity as a reference for classification. The algorithm needs a given set of training data already classified as belonging to one of n categories. In our case, the model needs a sample of tweets already labeled as having positive, neutral or negative sentiment. On the basis of the training sample of labeled tweets, the SVM algorithm assigns tweets to one of the three categories.[17]

[17] Technically, the SVM algorithm finds a hyper-plane in a N-dimensional space that maximizes the distance between the data points of two categories. This hyper-plane may be seen as a decision boundary. It is especially useful in high-dimensional spaces, which is why we decided to apply it in this context.

The main contribution of Tellez et al. (2017) to a baseline SVM classifier is to develop an efficient method to select the best text preprocessing techniques according to the language and writing style of the data of interest, specifically tweets. Their model applies two types of preprocessing techniques, some of them similar to those we used in Section 2 and some specific to preprocessing tweets and Spanish words. In particular, B4MSA can effectively process the content of symbols and emoticons, typical features of the twitter language. With respect to the preprocessing steps for Spanish, B4MSA considers cross language features, such as accents, punctuation and case sensitivity, stop words, negations and n-grams.[18] B4MSA applies the preprocessing text transformations to the tweets in our sample, then creates a vector representation of the sample (i.e., text is encoded and represented as a numeric matrix) using the TF-IDF weighting scheme, so that the more relevant words in the sample of tweets (or corpus) have a higher weight. The obtained matrix representation of the corpus serves as input for the classifier. Since text has many words and is often linearly separable, we use a linear SVM classifier like the standard B4MSA setting proposes to perform the sentiment classification.[19]

A3.3. Neural networks and transfer learning

Our third alternative is using deep learning to perform the classification. Deep learning uses neural networks that estimate non-linear relationships directly from data. It can be applied to many problems and contexts, and has been especially successful with computer vision applications and some natural language processing (NLP) tasks. A successful NLP task is characterized by the availability of large amounts of labeled data to train the model. However, researchers often do not have access to such volumes of labeled data, nor the computational resources to process them, which limits the possibilities of NLP. Moreover, NLP classification models struggle when language gets more ambiguous, as often there is not enough labeled data to learn from. Our dataset of tweets, consisting of 23,000 elements, is relatively small with respect to NLP standards, where data sets of hundreds of thousands of elements are usually needed. We decided to use the Universal Language Model Fine Tuning for Text Classification (ULMFiT) method developed by Howard and Ruder (2018), which addresses these challenges.

ULMFiT is built upon the concept of transfer learning. Transfer learning uses a model trained to solve one problem as the basis to solve a second problem related to the first, leveraging the labeled data of some related domain. The original model is fine-tuned to adjust to the target corpus. The fine-tuned model builds on the pretrained language model to reach higher accuracy with significantly less data and computation time than standard models trained from scratch.

The ULMFiT method significantly outperforms existing models and, more importantly, can learn well even from a limited volume of labeled data. ULMFiT consists of three stages. First, we select a pretrained language model to serve as the basis for the sentiment classifier. Intuitively, in this step, the algorithm ”learns the language” of interest. In this way, the algorithm will be able to recognize the patterns, structure of the language, and semantic similarities between words. Since we focus this study on tweets in Spanish, we use Andreas Daiminger’s language model, which was trained on Wikipedia articles in Spanish.[20] In stage two, we fine-tune the language model to fit the target corpus, which in our case is a set of tweets. It is important to emphasize that the preprocessing of the tweets for this model is different from the preprocessing applied for the other models. Since ULMFiT includes a language model as the basis, the expected input follows the natural language structure. Therefore, there is no need to remove punctuation and stopwords or to lemmatize terms. However, it is possible to apply some specific preprocessing to particular tweet elements. For instance, we delete all hyperlinks, since they do not add relevant information; we anonymize bank names, user mentions, and numbers, and we tag hashtags. We then use our whole preprocessed corpus to fine-tune the pretrained language model. Finally, we add a classification layer to the model and use 90% of our labeled tweets as the training set and the remaining 10% as the test set. The training set is the same as the one used for the B4MSA model, and both models are also tested on the same subset.

A4. Definitions of accuracy measures

Accuracy is the ratio of correctly predicted tweets (True Positives + True Negatives) to the total number of tweets (True Positives + True Negatives + False Positives + False Negatives).

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{4}$$

The balanced accuracy is used to deal with imbalanced datasets in binary and multi-class classification problems. It is the average of the correctly predicted tweets computed on each class individually. Consider a model that has to classify observations on two classes, 1 and 2:

$$BalancedAccuracy = \frac{1}{2} * \left(\frac{(TP + TN)_1}{(TP + TN + FP + FN)_1} + \frac{(TP + TN)_2}{(TP + TN + FP + FN)_2} \right) \tag{5}$$

[18] N-grams are sequences of *n* words that are automatically created by the model and can help the sentiment classification. The most used n-grams are sequences of two words (bi-grams). For instance, in the sentence ”The exchange rate between peso and dollar remains stable,” the sequence of two words ”exchange” and ”rate” may be considered as a single element for classification: ”exchange_rate.”’ This bi-gram has a specific meaning that is different from the separate words ”exchange” and ”rate.” For this reason, creating the bigram ”exchange_rate” may improve the classification performance of the model.

[19] We also tried with a non-linear kernel, but we obtained better results with the linear one.

[20] The pretrained model weights were posted on the ULMFiT Spanish fast.ai forum. The original post can be found in the following link: https://forums.fast.ai/t/ulm_t-spanish/29715/24.

Finally, the F1 score is the harmonic mean of Precision and Recall:

$$F1\ Score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{6}$$

Where Precision is the ratio of correctly predicted positive tweets (TP) to the total predicted positive tweets, both correctly and incorrectly (TP + FP), and Recall is the ratio of correctly predicted positive tweets (TP) to the total observations that should have been identified as positive (TP + FN). It computes what percentage of the tweets the classifier was able to label correctly.

References

Accornero, M., Moscatelli, M., 2018. Listening to the buzz: social media sentiment and retail Depositors' trust. Technical Report. Bank of Italy.

Alcaraz, C., Claessens, S., Cuadra, G., Marques-Ibanez, D., Sapriza, H., 2019. Whatever it takes: what is the impact of a major nonconventional monetary policy intervention? Working Paper Series 2249. European Central Bank.

Angelico, C., Marcucci, J., Miccoli, M., Quarta, F., 2018. Can we measure inflation expectations using twitter? Technical Report. Bank of Italy.

Azar, P.D., Lo, A.W., 2016. The wisdom of twitter crowds: predicting stock market reactions to FOMC meetings via twitter feeds. J. Portfolio Manag. Spec. QES Issue 42 (5), 123–134. doi:10.3905/jpm.2016.42.5.123.

Backus, D.K., Kehoe, P.J., 1992. International evidence on the historical properties of business cycles. Am. Econ. Rev. 82, 864–888.

Baker, S.R., Bloom, N., Davis, S.J., 2016. Measuring economic policy uncertainty. Q. J. Econ. 131 (4), 1593–1636. doi:10.1093/qje/qjw024.

Banco de Mexico, 2013. Financial stability report. Technical report. Banco de Mexico.

Banco de Mexico, 2019. Financial stability report. Technical Report. Banco de Mexico.

Baxter, M., King, R.G., 1999. Measuring business-cycles: approximate band-pass filters for economic time series. Rev. Econ. Stat. (81) 575–593. https://EconPapers.repec.org/RePEc:tpr:restat:v:81:y:1999:i:4:p:575-593.

Bholat, D., Hansen, S., Pedro, S., Schonhardt-Bailey, C., 2015. Text mining for Central Banks. Handbooks 33. Centre for Central Bank Studies, Bank of England.

BIS, 2009. Supervisory review process: SRP30 - risk management. Technical Report. Bank of Internatioanl Settlements.

BIS, 2017. Sound management of risks related to money laundering and financing of terrorism. Guidelines. Bank of International Settlements.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. (3) doi:10.5555/944919.944937.

Borovkova, S., Garmaev, E., Lammers, P., Rustige, J., 2017. Sensr: A Sentiment-Based Systemic Risk Indicator. DNB Working Paper 553. De Nederlandsche Bank.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Fifth Annual Workshop on Computational Learning Theory, pp. 144–152. doi:10.1145/130385.130401.

Brugnolini, L., 2018. About Local Projection Impulse Response Function Reliability. CEIS Research Paper 440. Tor Vergata University, CEIS.

Bruno, G., 2017. Central bank communications: information extraction and semantic analysis, Vol. 44.

Bruno, G., Cerchiello, P., Marcucci, J., Nicola, G., 2018b. Twitter sentiment and Banks' financial ratios: is there any causal link? Technical Report. Bank of Italy.

Bruno, G., Marcucci, J., Mattiocco, A., Scarnò, M., Sforzini, D., 2018a. The sentiment hidden in Italian texts through the lens of a new dictionary. Technical Report. Bank of Italy.

Buch, C., Bussierè, M., Goldberg, L., Hills, R., 2019. The international transmission of monetary policy. J. Int. Money Finance 91, 29–48. doi:10.1016/j.jimonfin.2018.08.005.

Bukovina, J., 2016. Social media big data and capital markets - an overview. J. Behav. Exp. Finance 11, 18–26. doi:10.1016/j.jbef.2016.06.002.

Cetorelli, N., Goldberg, L.S., 2011. Global banks and international shock transmission: evidence from the crisis. IMF Econ. Rev. 59, 41–76. doi:10.1057/imfer.2010.9.

Christiano, L.J., Fitzgerald, T.J., 2003. The band pass filter. Int. Econ. Rev. (44) doi:10.1111/1468-2354.t01-1-00076.

Correa, R., Garud, K., Londono-Yarce, J.M., Mislang, N., 2017a. Constructing a dictionary for financial stability. IFDP Notes. Board of Governors of the Federal Reserve System.

Correa, R., Garud, K., Londono-Yarce, J.M., Mislang, N., 2021. Sentiment in central Banks' financial stability reports. Rev. Finance 25, 85–120. doi:10.1093/rof/rfaa014.

Duprey, T., Klaus, B., Peltonen, T., 2017. Dating systemic financial stress episodes in the EU countries. J. Financ. Stabil. 32, 30–56. doi:10.1016/j.jfs.2017.07.004.

Forss Sandahl, J., Holmfeldt, M., Ryden, A., Stroemqvist, M., 2011. An index of financial stress for Sweden. Sveriges Riksbank Econ. Rev.

Hakkio, C.S., Keeton, W.R., 2009. Financial stress: what is it, how can it be measured, and why does it matter? Econ. Rev. Federal Reser. Bank Kansas City 94 (2), 5–50. https://www.kansascityfed.org/documents/432/PDF-09q2hakkio_keeton.pdf.

Hodrick, R., Prescott, E.C., 1997. Postwar U.S. business cycles: an empirical investigation. J. Money Credit Bank. (29). https://EconPapers.repec.org/RePEc:mcb:jmoncb:v:29:y:1997:i:1:p:1-16.

Holló, D., Kremer, M., Lo Duca, M., 2012. CISS - a composite indicator of systemic stress in the financial system. ECB Working Paper 1426. European Central Bank.

Howard, J., Ruder, S., 2018. Universal language model fine-tuning for text classification. Technical Report. Cornell University arXiv:1801.06146v5.

IMF, 2003. Financial soundness indicators - background paper. Technical Report. International Monetary Fund.

Ingo, W., 2011. Reputational Risk. John Wiley & Sons, pp. 103–123 chapter 6. doi:10.1002/9781118266298.ch6.

Iván, M.B.Á., González Pedraz, C., 2020. Sentiment analysis of the Spanish financial stability report. Working Paper 2011. Bank of Spain.

Jordà, O., 2005. Estimation and inference of impulse responses by local projections. Am. Econ. Rev. (95) 161–182 March. doi:10.1257/0002828053828518.

King, R.G., Rebelo, S.T., 1993. Low frequency filtering and real business cycles. J. Econ. Dyn. Control 17 (1/2), 207–231.

Kliesen, K., McCracken, M., 2020. The St. Louis Fed's Financial Stress Index, Version 2.0. https://fredblog.stlouisfed.org/2020/03/the-st-louis-feds-financial-stress-index-version-2-0/.

Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A., 2011. Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 262–272. doi:10.5555/2145432.2145462.

Morais, B., Peydro, J.-L., Roldan-Pena, J., Ruiz, C., 2019. The international bank lending channel of monetary policy rates and QE: Credit supply, reach-for-yield, and real effects. J. Finance 74, 55–90. doi:10.1111/jofi.12735.

Newman, D., Noh, Y., Talley, E., Karimi, S., Baldwin, T., 2010. Evaluating topic models for digital libraries. In: Proceedings of the 10th annual joint conference on Digital libraries, JCDL, pp. 215–224. doi:10.1145/1816123.1816156.

Nyman, R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P., Smith, R., 2018. News and narrative in financial systems: exploiting big data for systemic risk assessment. Staff Working paper 704. Bank of England.

Polikar, R., 2012. Ensemble learning doi:10.1007/978-1-4419-9326-71.

Rakowski, D., Shirley, S.E., Stark, J., 2020. Twitter activity, investor attention, and the diffusion of information. Financ. Manag. 1–44. doi:10.1111/fima.12307.

Ravn, M.O., Uhlig, H., 2002. On adjusting the Hodrick-Prescott filter for the frequency of observations. Rev. Econ. Stat. 84, 371–380.

Reinhardt, D., Sowerbutts, R., 2015. Regulatory arbitrage in action: evidence from banking flows and macroprudential policy. Bank of England working papers 546. Bank of England.

Rokach, L., 2010. Ensemble-based classifiers. Artif. Intell. Rev. 33 (1-2), 1–39. doi:10.1007/s10462-009-9124-7.

Shapiro, A.H., Sudhof, M., Wilson, D., 2017. Measuring news sentiment. Working Paper 2'217-01. Federal Reserve Bank of San Francisco doi:10.24148/wp2017-01.

Tellez, E.S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Suárez, R.R., Siordia, O.S., 2017. A simple approach to multilingual polarity classification in twitter. Pattern Recognit. Lett. (94) 68–74 doi:10.1016/j.patrec.2017.05.024.

Tripathy, J., 2020. Cross-border effects of regulatory spillovers: evidence from Mexico. J. Int. Econ. 126, 103350. doi:10.1016/j.jinteco.2020.103350. http://www.sciencedirect.com/science/article/pii/S0022199620300660.

Zimbra, D., Abbasi, A., Zeng, D., Chen, H., 2018. The state-of-the-art in Twitter sentiment analysis: a review and benchmark evaluation. ACM Trans. Manag. Inf. Syst. 9 (2). doi:10.1145/3185045.