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

Raúl Fernández†, Brenda Palma Guizar‡, Caterina Rho§

November 11, 2020

Abstract

We apply sentiment analysis to Twitter messages in Spanish to build a sentiment risk index for the financial sector in Mexico. Using a sample of tweets that covers the period 2006-2019, we classify the tweets 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.

Keywords: sentiment analysis, systemic risk, banks.

JEL classification: G1, G21, G41

*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, 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

†Banco de México. Email: rfernandez@banxico.org.mx

‡Banco de México. Email: bpalmag@banxico.org.mx

§Banco de México. Corresponding author. Email: crho@banxico.org.mx

1 Introduction

In recent years we have witnessed an unprecedented rise in the production and storage of granular data that cover a broad range of sources, such as social media, online marketing, news websites, transportation services or renting. The availability of novel and rich sources of data represents a key opportunity for policymakers and researchers alike.

The study of unstructured data, such as social media content, is particularly interesting for central banks in the context of financial regulation and supervision. Research showed that consumer sentiment and investor sentiment may affect economic activity and financial markets, suggesting that appropriate sentiment indicators may be useful if incorporated in the analysis of financial stability or systemic risk.

Economic sentiment may affect economic activity according to two mechanisms. On the one hand there is the "animal spirits" hypothesis (Keynes, 1936), stating that consumers and business sentiment can directly affect economic activity (Blanchard, 1993; Benhabib and Spiegel, 2017; Shapiro et al., 2018). In behavioral finance, animal spirits represent the emotions of confidence, hope, fear or pessimism that can fuel growth or cause sudden stops in financial markets (Akerlof and Shiller, 2009). In particular, if sentiment is pessimistic, consumer confidence will also be low, driving down financial markets, and ultimately the economy. Conversely, if sentiment is optimistic, confidence will be high, and markets will rise.

On the other hand, sentiment may be purely informational, containing news about the future states of the economy held by the public but not yet observed in hard data (Barsky and Sims, 2012). In this case, social media sentiment may influence financial markets through the information demand of retail investors. Retail investors may not have access to professional databases like Bloomberg or Thomson Reuters, so they use social media and the Google search engine as a publicly available source of information about market movements (Da et al., 2011; Vlastakis and Markellos, 2012; Ding and Hou, 2015). Sprenger et al. (2014) propose the investment forums of Twitter as an alternative information source for retail investors. These forums are a space for discussion about capital markets for retail investors. While the evidence about the drivers of the correlation between sentiment and economic activity is mixed, it is still possible to take advantage of its correlation for forecasting purposes.

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 literature focuses on studying social media activity, in particular Twitter messaging, on stock market fluctuations coinciding 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 (Borovkova et al., 2017; Accornero and Moscatelli, 2018; Angelico et al., 2018).

Research shows that sentiment indices may help predict not only economic variables; but also financial indicators, even if financial variables react timely to new information (Calomiris and Mamaysky, 2018). Ormerod et al. (2015) show that an emotion index capturing shifts between excitement and anxiety in texts referring to the whole US economy improves the one-quarter ahead consensus forecasts for real GDP growth. The same indices Granger cause the Cleveland and St Louis Indices of Financial Stress. Nyman et al. (2018) show that in the UK sentiment measures and narrative consensus correlate well with economic and financial variables such as the interest rate, FTSE 100 index, and production. Sprenger et al. (2014) study company events via Twitter microblogging forums. They identify good and bad news in a sample of more than 400,000 stock-related tweets. Their findings show that positive ones are often leaked and incorporated into stock prices before the official announcement. On the contrary, the negative ones are predominantly surprising, so the market reaction occurs within a day of the event. Cerchiello et al. (2017) proposed a model to estimate systemic risks combining information on financial markets and financial

tweets, which can help predict the default probability of a bank, conditionally on the others.

Moreover, coming at higher frequencies, sentiment indicators may help policymakers by measuring expectations about core economic indicators, such as inflation and the GDP growth, usually built at monthly or quarterly frequencies. Azar and Lo (2016) show that tweets mentioning the Federal Open Market Committee around FOMC meetings contain information to predict future returns, while Plakandaras et al. (2015) show that investors' sentiment built on a social media sentiment measure has valuable information for future movements of four exchange rates.

Finally, text-based metrics have advantages of cost, timeliness, and scope. They could function like soft data (e.g. surveys), as indicators for policymakers, and inputs into forecasts. Concerning to a Consumer Sentiment survey, extracting news data or social media data is less expensive and timelier. Kalamara et al. (2020) extract timely signals from newspaper text and use them to forecast macroeconomic variables. They find that newspaper text can improve economic forecasts of macroeconomic variables, including GDP, CPI, and unemployment.

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 in Twitter, a social media platform that has gained popularity among mass media, academics, policymakers, 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, specifically of 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.

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 LDA algorithm to describe the sample of tweets through a set of topics, each represented as a collection of words. We identify some topics not traditionally included in financial stress risk indices. The novel topics, such as financial frauds, money laundering, and failures of online payment systems, are associated with a rise or fall in the Twitter sentiment index.

Second, we train three different sentiment classifiers (one based on word counts, a linear classifier, and one based on neural networks) to build a sentiment index for the Mexican financial system. 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 of our Sentiment Index on a financial market stress index and selected market variables. We do not claim causality in these results, because the direction of the causality between sentiment indicators and financial variables is still an open question (Shapiro et al., 2018). When looking over a 26-week horizon, a one standard deviation shock significantly correlates with an increase of the exchange rate volatility and stock market volatility in the first 10 weeks after the shock. The sentiment index also correlates with an increase in country risk as measured by the EMBI+ for Mexico. The banking sector, proxied by the beta of financial institutions, also reacts with a rise to a shock in the Sentiment Index, 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.

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 analyzing 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 the context of financial

studies, it is often used to build financial market indexes that replicate the variations in traditional stock market indexes, 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. (2018) use machine learning techniques to develop and analyze new time series measures of economic sentiment based on text analysis of articles of financial newspapers from 1980 to 2015. They find that the four news sentiment indexes that they developed are strongly correlated with contemporaneous business cycle indicators and improve the forecast performance of standard financial indicators.

A time series of data compiled using Twitter updates of financial news can be used for the analysis of sentiment of investors or consumers in correspondence to shocks happening in different moments. 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 in 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' level of trust. Bruno et al. (2018) build a dictionary to analyze sentiment in Italian texts, while Bruno et al. (2018, a) apply the dictionary to tweets about selected Italian banks extracting sentiment indicators and relate them to some banks' financial variables, finding a positive correlation between them and the sentiment for some of the banks in their sample.

Correa et al. (2017; 2017, a) 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 markets indicators related to credit, asset prices and systemic risk. Bruno (2018) conducts a similar analysis on recent Financial Stability Reports issued by the Bank of Italy, while in a recent paper Moreno Bernal 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 builds on the work by Correa et al. (2017), and it explores alternative techniques that may be suitable for sentiment analysis in social media. We apply the model of neural networks and transfer learning developed by Howard and Ruder (2018) and the multilingual Support Vector Machine model proposed by Tellez et al. (2017).

We take inspiration from Shapiro et al. (2018) 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 Mexico (Banxico) (Banco de Mexico, 2019) and to selected financial indicators.

3 Data

In order to build the banking risk index, we use Twitter as our data source and the Mexican commercial banks' names as our search criteria. We select only the 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 as a whole is in great part determined by how healthy Mexican banks are.

3.1 Extraction of tweets

We use the Twitter Paid Premium Search API that allows us to extract tweets in Spanish that contain the names of Mexican commercial banks from April 2006 onward.[1] We focus on the extraction of 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. Also, English language media (such as those based in the US or UK) often report only major events about Mexico, or as foreign sources, report events about Mexico with a short delay. 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, this is outside the scope of this paper. The complexity and time cost of setting up a text analysis on two languages at once is significant, especially because of the particularities of each language. For this reason, we extract only tweets in Spanish.

We 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 investors (Rakowski et al., 2020). We made this choice to base our analysis on reliable sources, among those 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.

[Insert Table 1 here]

We decide to filter our extraction of tweets using only selected accounts instead of using all messages from the universe of tweets so that the final database may be as clean as possible from potential noise. Without a selection, we would incur an excess of information, and our data would not be as useful for the purpose of our analysis. To test this hypothesis, we extract all tweets from the universe of Twitter for one day and we compare this sample with the sample of tweets extracted only from our selected sources.[2] The total number of tweets extracted for the given day is 3004 for the extraction without selecting accounts, and 34 tweets for the extraction from selected accounts.

Although the amount of information is drastically reduced by our selection, Table 2 shows some interesting results about the relevance of the information extracted in the two cases.

[Insert Table 2 here]

Panel (a) shows the ten most frequent words in the sample extracted without filtering. At a first sight, they are not linked with the topic we are analyzing. Only "venta" (sale) and "tarjeta" (card, credit card) may be linked with banking, and they are only at the 5th and 9th place respectively. Other more frequent words are too general to hint a specific topic ("north", "route", "popular") or they indicate foreign countries ("Colombia"). This result suggests that most parts of the tweets in the general sample are not linked with the topic of financial risk, and may create noise in our subsequent analysis.

Panel (b) compares the ten most frequent words in the selected sample, how many times they occur, and the occurrence of the same words in the non-selected sample. Among the top ten words we find "financial", "market", "growth", "director" and "president"; all words that are linked to the topic of financial markets, banking or policy. Their frequency is not high, but the number of tweets in the selected sample is also very small. These words are not in the top ten of the complete sample, reinforcing the evidence shown in Panel (a).

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

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

We also compare the most frequent words in the non-selected sample with the correspondent words in the selected one (Panel (c)). We find that the most frequent words in the complete extraction that also appear in the selected extraction are very general (verbs, numbers) or occur in the selected extraction in low rankings.

Finally, from the simple reading of the tweets extracted without selection we find that many tweets regard marketing strategies of commercial banks, job offers, comments of users 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. We are aware of the trade-off between the quantity of information and the quality of information, but we find that this preliminary study motivates our choice of limiting the sources of our tweets.

Table 3 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 id, the publication date and time, the user who published it, the number 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, and will constantly increase with future extractions. The tweet volume at the beginning of the observation period is lower than the observed towards recent periods, as Twitter started gaining popularity.

[Insert Table 3 here]

To our knowledge, this is the first paper that builds a financial risk indicator for Mexico starting from sentiment analysis present in Twitter content. This stream of research may be developed in several directions. One possibility is to analyze the characteristics of the Twitter messages in more detail , beyond the text itself. In this paper we use all the available Twitter content that can be extracted using our keywords of interest, without filtering for geographical location of the tweet. It may be possible to build sub-indices of an index, filtering the messages for geographical location, such as domestic tweets versus foreign tweets. This may more clearly allow distinction between external shocks and idiosyncratic ones. We decided not to develop this idea in this paper because not all tweets come with a geographical location. The API we use used in this paper allows us to extract the location of the twitter user when the tweet is published. However, the location is not determined automatically by Twitter. The user has to specify his location, and not all the users present in our sample do that. For this reason, it is hard to filter tweets by location. It may be possible to distinguish external shocks from domestic shocks analyzing the content of the tweets, but in that case, it would be necessary to filter the tweets at the labeling stage, one of the first steps of the sentiment analysis process. We reserve this avenue to future research.

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 stop-words, 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 in order to avoid having banks' names as features in our models;

3. we lemmatize the text to reduce the sparsity of the data;[3]

4. We turn all uppercase letters to lowercase. The following example illustrates the mentioned transformations:

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

3.3 Data exploration

After preprocessing the tweets, we want to conduct an exploratory analysis on the data to better understand the kind of 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 financial risk events brings to the public. Therefore, we need to make sure that the tweets that we extract from Twitter are relevant for our purposes. If the information we extract was not related with topics that are significant for the evaluation of financial risk, our sentiment index would be biased, or even useless. However, analyzing text data manually may be an excessively laborious task: reading a text, and classifying the information it contains is doable when the amount of text analyzed is limited, but it becomes a burden, in terms of time and effort, when you need to analyze a huge amount of textual data. Our final sample contains 23000 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

Topic analysis is a natural language processing technique that automatically extracts meaning from texts by identifying recurring themes or topics in the text corpus. It helps the researcher to organize large sets of data and identify the most frequent topics in a simple, fast and scalable way. For this reason, this technique is used in text analysis to obtain a first description of the data at hand, and it is the best alternative to analyzing the tweets manually. Topic models have been used in the social science literature mostly for descriptive purposes.

Quinn et al. (2010) apply a topic model to congressional speeches to identify which members of Congress speak about which topics. Hansen et al. (2018) analyze minutes from the FOMC meetings to construct communication measures from LDA output. They use these text-based measures to explore how transparency affects monetary policymakers' deliberations.

3.3.2 Analysis with LDA

We use the LDA algorithm (Blei et al., 2003; Bruno et al., 2018), 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.

As an example, let us think of a text analysis on a newspaper that contains three sections: politics, economics, and sports. The LDA algorithm is able to identify words that are used often together, and group them. If we use an LDA model to find three topics in our sample newspaper, we would get three groups of words. The first would contain the words "parliament, elections, politician, decision...", the second group would contain the words "firms, economy, production, inflation..." and the third group the words "swimming, championship, racket, ball...". The researcher may assign a label to each group of words to describe each topic. In this case it is easy to understand that the newspaper has three sections: politics, economics and sports.

A certain degree of subjectivity is unavoidable in the interpretation of the topics. The model automatically divides the corpus of documents into group of words (that may be overlapping), but the interpretation of this result and the labeling of the topics is the researcher's responsibility. The user is also responsible for choosing the number of topics to be inferred from the collection of documents. There are some indicators to compare the performance of different models and support the user in this task.[4] Depending on the objective of each analysis, the interpretability

[4] Among the indicators to evaluate the performance of the LDA as a topic model there are topic coherence indicators (for example,

(or coherence) of the topics may be a primary criterion, which is the case for our analysis. Following the newspaper example, setting the number of topics equal to one, the LDA would not classify the words at all: every word would go in one single group, and it would be impossible to interpret this result. On the other hand, ten topics may be too many for understanding the newspaper structure: too detailed information would be redundant. LDA is a model that allows for descriptive metrics of the data to be built, and depending on the specific research question at hand it may be calibrated in different ways.

We apply the LDA model to the totality of our sample of tweets. We fit the model varying the number of topics, from a minimum number of 5 to a maximum of 15. To measure topic coherence, we first apply the UMass score (Mimno et al., 2011), that is specifically designed for LDA. Intuitively, the UMass score measures how much, within the words used to describe a topic, a common word is on average a good predictor for a less common word. The higher the score, the more coherent the topic is. When computing the UMass coherence score for models with different numbers of topics, the coherence increases with the number of topics. This is a natural consequence of the UMass measure: its goal is to group the words in the most coherent way possible, and this automatically increases the granularity of the results. However, there is a trade-off between the number of final topics included in the model and the interpretability of the results (see Chang et al. 2009; Blei 2012). In particular, Hansen et al. (2018) choose the interpretability criteria over a formal model selection criteria to select the final number of topics for their LDA model.

The general theme of our tweets is very specific. We select only tweets that deal with the Mexican financial sector, so the number of topics that we can find in this sample is relatively limited. When we add topics to our LDA model, the UMass score rises, but above a certain threshold the number of topics becomes too high and it is difficult to interpret the results. For this reason, we also apply the interpretability criteria to select the final number of topics to include in the LDA model.

After several iterations using different numbers of topics, we identified six topics that constantly appeared in the results[5]:

1. Financial markets (top 15 words: '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')

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.

[5]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'.

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')

In order to name the topics and to minimize the degree of subjectivity when doing it, 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 get 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 per topic.

We use the LDA model for descriptive purposes, to provide us with an idea about the structure of our textual data, but it is still important that the topics are reasonable in the context of financial stability. The first topic, "Financial markets", is defined by words that are usually used to describe stock market movements. The acronym of the Mexican Stock Market, BMV, is the sixth most relevant word in the group, suggesting that financial markets are at the core of this word group.

The second group, "Macroeconomic expectations", contains words that suggest a linkage with the macroeconomy and systemic risk: 'economy', 'risk', 'bank', 'country', 'rating', 'growth', 'credit'. These words are part of tweets that contains news about macroeconomic expectations reported in commercial bank's policy notes.[6]

The third group, "Foreign exchange market" contains words that describe exchange rate movements: 'dollar', 'to forecast', 'to buy', 'to sell', 'sale', 'exchange rate', 'to close', 'counter', 'peso'. This group of tweets mainly reports news about the daily exchange rate, appreciation or depreciation of the peso with respect to other currencies, or the extent to which the exchange rate is set by commercial banks versus the FIX exchange rate.

The fourth group, "Business activity", report words that are linked to the area of business operations of a bank, such as: 'operation', 'service', 'client', 'financial group', 'to buy', 'credit', 'bank', 'branch', 'to offer', 'digital'. The tweets in this group report banking operations, mergers and acquisitions, new offers to customers, and news about the digitalization of banking services.

The fifth group regards "financial results" and the set of most relevant words is not so self-explanatory as in the other cases. The first fifteen words in the list seem more neutral, with the exception of 'achieve', 'first quarter', 'to expect', 'to present', 'to announce', 'to fall', 'to raise'. These words suggest that the topic is linked with the realized or expected financial results of banking institutions. Checking the words in the context of the tweets, we find a confirmation of our first intuition.

Finally the sixth topic, "Illicit activities and penalties" contains the words 'client' 'money', 'to investigate', 'to fine', 'to charge', 'oceanografía', 'card', and 'fraud', that suggest a connection with financial frauds and police investigations on the bank's conduct. The world "Oceanografía" refers to a specific financial scandal occurred in 2014.

We compare the six LDA topics with Banxico's Financial Market Stress Index (Indice de Estrés de los Mercados Financieros, IEMF, Banco de Mexico, 2019) components.

The IEMF index has weekly frequency and it 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 common sectors that the IEMF and the tweets cover are financial markets. 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 in the "credit institutions" component of the IEMF. However, Twitter data provides information on certain details of the business activity that is not

[6]The main commercial banks (such as BBVA and Citigroup)) produce some information material for their stakeholders regarding general economic forecasts, and this is what is reported in the tweets.

being explicitly captured by the IEMF. We detect sentiment about customer services, digital services, and online payment systems, including bugs. Additionally, our data capture new information within topic 6, "Illicit activities and penalties". This topic comprises news about money laundering activities, tax evasion, banking scandals, online frauds and penalties to banks because of illicit activities.

We consider financial frauds and money laundering as negative shocks for the reputation of the bank, both where the bank is considered to be headquartered in Mexico and 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 fails 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 in 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 framework. Moreover, Banco de Mexico 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, disrupt IT systems and cause failure 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 the confidence in the financial system, especially if we consider a cyberattack to a systemically important bank.

3.4 Data labeling

We create a sample of labeled data which serves to train the models and compare their performance. We take a random sample of 2,000 tweets from our database and we assign juxtaposed sub-samples of 100 tweets to 37 professionals, working at the Directorate General of Financial Stability in Banxico, that label them 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 proxy 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 for the banking system or the 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 prioritized the systemic risk consideration, so that we consider the tweet as reporting news that decrease the banking risk. The labeling criteria to categorize each tweet is the following:

- Higher risk (corresponding to negative sentiment): tweets in which content 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.

- Lower risk (corresponding to positive sentiment): tweets in which content 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, crimes of small entity (bank robberies to a specific branch).

An important note regards three special kinds of news that we ask the volunteers to manage with special attention. The first group is the group of tweets containing 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 (Cetorelli and Goldberg, 2011; Reinhardt and Sowerbutts, 2015; Buch et al., 2019). 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 (Tripathy, 2020; Morais et al., 2015; Alcaraz et al., 2019). 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 headquartered bank, but not on the Mexican subsidiary. For instance, news about BBVA in Spain or Citigroup in the US may also have an impact in Mexico. News about a subsidiary of BBVA in Peru may have a 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 that regards 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), 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, and to the Mexican banking system through cross-country contagion. A tweet that reports news about the denial of the World Bank 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 a potential 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 the business community. The tweet is considered neutral by default, unless it reports an explicit positive or negative judgment. The rationale is to maintain the sample as unbiased regarding day-to-day political decisions or business strategies. If a judgment is explicit, it comes from our set of news, and 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 we 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 2 volunteers using the values of 1 for "Higher risk", -1 for "Lower risk", and 0 for the "Neutral" category.

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 of negative tweets, 26 percent of positive tweets and the remaining 42 percent of neutral tweets.

4 Sentiment classifiers

We choose three different models to build the sentiment classifier for the tweets. The models we choose are based on the three main frameworks used in text analysis: bag of words (or dictionary approach), Support Vector Machines (SVM) and neural networks. Our first approach replicates the Correa et al. (2017) methodology 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 mainly focuses 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 by a language model and a classification layer on top.

Each model has advantages and disadvantages. The dictionary model is the simplest one. The sentiment of the tweet is computed as a word count of the positive and negative words that compose the tweet. If the majority of the words contained in the tweet is negative, the sentiment is negative, and vice versa. On the one hand, the classification process is very intuitive, once you can use a dictionary crafted for your kind of research question and domain. On the other hand, this is the most rigid method. The use of a set dictionary does not allow an evolution of the language, or the incorporation of new topics in the discussion. This model is better suited to analyze documents with a specific structure, that does not change in the short run, such as the policy reports of Central banks (Correa et al., 2017; Moreno Bernal and González Pedraz, 2020), or the minutes of institutional meetings (Hansen et al., 2018), although there are also examples of dictionary approaches used to analyze twitter messages (Bruno et al., 2018).

The second model, based on a SVM and a specific preprocessing for the Spanish language, is more complex than the dictionary method, but also more precise and flexible in the classification process. The SVM is an algorithm that is particularly suitable for analysis in high dimensional spaces, such as textual data. This model has a good balance between complexity and flexibility. It can interpret the words of each tweet in a more precise way than the dictionary method, thanks to its more precise preprocessing and the specific classification algorithm it applies.

The third model is the most complex, but also the most flexible. Neural networks are based on a set of interconnected algorithms that analyze the input data in subsequent layers, and are able to find hidden relationships that another simpler model cannot detect. The model we apply in this paper is able not only to organize and classify the words contained in each tweet, but also to understand the structure of the tweet itself and of its language. A neural network is the most flexible model because it can be trained to understand the relationships between the words in a text. We further discuss the advantages and disadvantages of each model and their characteristics in the correspondent sections.

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 idea is that integrating multiple models, known in machine learning as ensemble methodology, can help to build a model with enhanced predictive performance (Rokach, 2010). Our classifier uses a majority voting rule to determine the final sentiment. A voting rule is a simple ensemble methodology that could help in making the classification more robust. Among the voting rules there are three possibilities: unanimous voting, simple majority and, plurality voting. If the classifier outputs are

independent, then it can be shown that majority voting is the optimal combination rule (Polikar, 2012). Since our classifier based on majority voting comprehends 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. Table 4 shows an example for each case.

[Insert Table 4 here]

4.1 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 for training 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 between 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 to labeling data, this is a pretty useful alternative that we chose as our baseline methodology.

We use Correa et al. (2017) financial dictionary, which 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 financial specific dictionaries proposed in the literature. The dictionary contains 391 words, of which 96 are positive and 295 are negative.

Although Correa et al. (2017) 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 CKJM dictionary and Banxico's FSRs language, we perform text analysis on the FSRs published by Banxico in English from 2006 to 2016.[7] We find a correspondence of 58 percent between CKJM dictionary and the words used in Banxico's FSRs.

Second, CKJM dictionary is in English, while our focus is on tweets in Spanish. We translate CKJM dictionary from English to Spanish, controlling for semantic differences. The correspondence between our translation of 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 differ from English.

Third, we are not applying the financial stability dictionary to FSRs, but to tweets. CKJM dictionary is specifically tailored for the context and structure of FSRs and Correa et al. (2017) highlight the importance of adapting a dictionary to the specific context where the text analysis will be performed. Although we focus our search on reliable sources and we 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 universe 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 that are labeled as positive or negative. We apply the TF-IDF weighting scheme to the two sub-samples of tweets to identify the most relevant terms used in the tweets of each category.[8] We labeled as "negative" (or "positive") the most relevant words that appear in the negative (or positive) tweets. Finally, we include these words in our original dictionary with the correspondent word polarities.

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

[8] 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 5 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 words used in our sample of tweets.

[Insert Table 5 here]

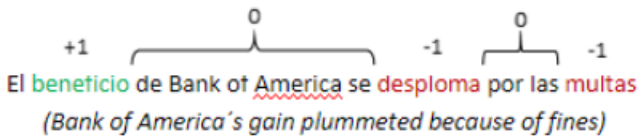
Most of the words used in the English and Spanish versions of the FSRs are similar or exactly the same ('volatility' and 'volatilidad', 'stable' and 'estable', 'contagion' and 'contagio'), suggesting that the sentiment detected by the original CKJM dictionary and our translation of it is comparable, and that it may be a useful tool to analyze the text of the Mexican FSRs. In addition, we find some new words that are especially relevant in the social media context, but they are not commonly mentioned in the FSRs or in the CKJM dictionary ('fine', 'investigation', 'manipulation').

4.1.1 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 word polarity is considered to be zero. The sentiment score of a tweet is computed as the sum of the word polarities of all the terms in the corresponding tweet:

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

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:



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, the value 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 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 since a training step is not required here. Nonetheless, we test the performance of the dictionary classifier also on the labeled data, the training sample for the other two classifiers, for comparison with the other models. Results are discussed in section 4.4.

4.2 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 specifically built 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). A 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, each of them already classified as belonging to one or the other 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 new tweets to one category of the three categories.[9]

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 the 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 of them specific for preprocessing tweets and preprocessing 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.[10]

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.[11]

4.3 Neural networks and transfer learning

Our third alternative is using deep learning to perform the classification task. Deep learning uses neural networks that estimate non-linear relationships directly from the 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, often researchers 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, made by 23000 elements, is relatively small with respect to NLP standards, where datasets of hundreds of thousands of elements are usually needed.

[9] Technically, the SVM algorithm finds a hyper-plane in a N-dimensional space that maximizes the distance between the data points of two different 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.

[10] N-grams are sequences of n words that are automatically created by the model, and that 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.

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

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 one, leveraging on 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 so it can 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, it can learn well even from a limited volume of labeled data.

ULMFiT consists of three stages. First, we select a pretrained language model which serves 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, the structure of the language, and the 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.[12]

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. There is therefore no need to remove punctuation and stop words, 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 validation 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. Results are discussed in section 4.4.

4.4 Comparison between the sentiment classifiers

The different classifiers are trained and evaluated with the same dataset. To compare the models performance, we compute accuracy, balanced accuracy, and F1 score.

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).

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

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:

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

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

$$F1\ Score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{3}$$

Where *Precision* is the ratio of correctly predicted positive tweets (TP) to the total predicted positive tweets,

[12] 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

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 that actually belongs to the category the classifier was able to label correctly.

All these accuracy measures have a [0, 1] range, where 1 is perfect accuracy and 0 is no accuracy at all. Table 6 presents the results.

[Insert Table 6 here]

Higher accuracy reflects the better classification of the positive, negative and neutral tweets by the model. If we were to perform 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 in classifying data than a random classification.[13] 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 in the 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 the one in the test set.[14] 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, as the other two methods can do. For this reason, we expect a 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 one (accuracy of 73 percent). It performs very 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 very close, and considerably outperform the dictionary results. The final column of Table 6 presents the performance metrics for the majority voting model. This final classifier maximizes the available information and gives 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 the three models. The accuracy in classifying the negative tweets is comparable to the accuracy of the dictionary model, while the positive and neutral categories are in line with the higher accuracy of the B4MSA and ULMFiT models.

4.5 Discussion of the methodology

Even though sentiment analysis models can analyze an immense amount of text, providing timely and useful information, we are aware that a certain degree of subjectivity is unavoidable in the interpretation of the results. It is crucial to have a rigorous approach and a sized and relevant sample.

Concerning the algorithms used to perform sentiment analysis, research shows that textual data have their own specific challenges. Depending on the model used, a classifier may recognize irony, sarcasm, special textual features such as emoticons with various degrees of accuracy. Moreover, not all the models take into account that language evolves. Failing to address these issues could bias the results.

We do our best to circumvent some of these problems. First, we use verified Twitter accounts of national and international newspapers, news agencies and rating agencies as our source of data. This allow us to more easily

[13] To be precise, our sample of tweets has an unbalanced distribution of positive, neutral and negative tweets. Given the sub-sample 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.

[14] The gap on accuracy between the training and test sets should not be too wide: a wide gap between 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 set an optimal gap between the accuracy of training and test set.

select the relevant tweets, that can give information about systemic risk, and to minimize noise. We label the tweets to train and test the models with the help of a group of economists to which we gave detailed instructions regarding the research goal and the logic behind the labeling process. We use alternative machine learning algorithms to take advantage of the strengths of each model, while minimizing their weaknesses. Our results are encouraging, given that the voting classifier is the one with the highest accuracy among the ones we use in this study.

When we perform the topic analysis we rely on previous research and specific statistics, such as the UMass score, to select the optimal number of topics to calibrate the model. There is a trade-off between interpretability and coherence of the topics: to be sure that the information contained in each topic is coherent and relevant to our analysis we label each topic controlling for the most relevant words in each topic and the most relevant tweets in each topic, so that we can take the context where each word is used into account .

5 Sentiment index

Once the tweets are classified, the sentiment index can be built. We base our methodology on Correa et al. (2017).

Instead of the number of positive and negative word of 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:

$$Sentiment\ Index_t = \frac{negative\ tweets_t - positive\ tweets_t}{negative\ tweets_t + positive\ tweets_t}$$

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

The baseline index considers in the denominator the positive and negative Twitter messages published in a period t. In this way, we normalize the index, considering the variability in the volume of tweets published in the period of interest. We exclude the neutral ones because they may introduce some noise in 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 bring relevant information (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 for the Twitter users. The number of reactions (retweets or likes) that a tweet receives may be seen as an indicator of the popularity of the tweet. Reactions also increase the exposure of tweets, thus augmenting their reach. This may lead to a 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 a particularly strong sentiment, either positive or negative. If this is the case, the higher the number of reactions, the stronger the sentiment given by that specific tweet and the more important the news content. However, the higher number of reactions may be given 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, but it may add noise to the indicator. In fact, the final index may result biased if news that users found interesting but that are irrelevant at a systemic level get a higher weight.

Considering these points, the inclusion of neutral tweets in the index, and the potential importance of each tweet to the Twitter users, we build other versions of the baseline sentiment index. The first does include in the denominator the neutral tweets, while the second variation weights each tweet by the number of reactions (both retweets and likes) received. Table 7 presents how correlated the sentiment index is when computed with the four different estimators (dictionary, SVM, neural networks and majority voting), and in the four different versions (baseline, adding neutral tweets in the denominator, and weighting positive and negative tweets).

[Insert Table 7 here]

In all cases we find that the correlation between the sentiment indices computed with different classifiers is high and positive. In the baseline model the correlation between the indices lies in a range that goes from 48 percent, to 77 percent. It decreases in the models that include the noise given by the neutral tweets, as expected, and when we weight the tweets by the number of reactions. As a comparison, Shapiro et al. (2018) find a correlation of 34 percent between the different models that they use to build their sentiment indices.

5.1 Visualization

In order to visualize the results, we build an interactive dashboard using Dash, a Python framework for building web applications. The dashboard 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 with the most popular terms used in the tweets during the selected period. This may help in understanding 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 rating from BBB+ to BBB-.

[Insert Figure 1 here (color should be used)]

The risk increase due to this event is caught by the index and the word clouds highlight as negative words "Pemex", "calificación" and "Fitch". The bigger is a word in the word cloud, the more important it is in its respective category. Figure 1 shows on the left the complete timeline of the volume of tweets extracted and of the sentiment index computed from the tweets. Although we start our extraction from 2006, when Twitter went online, the graph picturing the volume of tweets shows that at the beginning of the period the total number of daily tweets containing one of our keywords (the names of the banks) was 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 regarding banks follows the growing popularity of Twitter for the general public and its evolution as a communication tool not only between private users but also for businesses, public and private institutions, newspapers and media. Notice that, even with very few observations, it is possible to compute a sentiment index, as shown by the second graph on the left in Figure 1. Nonetheless, if the number of observations (the number of tweets) is too low, the index may be biased because it is built on few observations. 2006 and 2007 have very few tweets, less than 50 in total for the two years. For this reason, we will truncate the series, starting our empirical analysis from 2008.

Figure 2 shows the four alternative indices computed at monthly (panel (a)) and weekly frequencies (panel (b)) using the baseline model. The sentiment index scale is normalized from -1 (minimum risk) to 1 (maximum risk). In panel (a) we see that the index computed using the majority voting model consistently signals higher risk than the others. The sentiment index computed using the neural network model broadly follows the Voting sentiment index, except on a period from mid-2015 to mid-2016. Panel (b) shows the raw sentiment index with weekly frequency. As it is, the indicator is too volatile to be used in a comparison with other more standard economic indicators. In paragraph 5.2 we analyze in detail how we address this issue.

[Insert Figure 2 here (color should be used)]

Figure 3 presents the sentiment index built by voting 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 sentiment index corresponds to an increase in risk. We labeled each peak of sentiment according to the keywords in the word cloud of the dashboard, and comparing the keywords with those used in the news of that month. 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 make a comparison to make sense of our results.

[Insert Figure 3 here]

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

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

In July 2012 global financial markets were shaken by the Libor manipulation scandal, while 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 of dollars to US authorities for allowing money laundering from drug cartels from Mexico to its US offices.

The last relevant financial scandal was the Oceanografía one that directly hit Mexico and its financial system during 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.

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 on financial markets and domestic economic slowdown due to uncertainty in November 2018 and June 2019 respectively. In particular, on the 8th of November 2018, the Mexican ruling party proposed a project to reduce or prohibit banking charges for interbank transfers and cash withdrawals. On that day, the price of stocks of Banorte (the second 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 5th, 2019 the credit rating agencies Moody's and Fitch cut Mexico's sovereign debt rating, citing risks posed by Pemex, the national oil company that was heavily indebted, and trade tensions during the ratification process of the trade deal between Mexico, United States and Canada (T-MEC).

5.2 A filtered sentiment index

The sentiment index computed using equation (2) essentially shows the positive and negative sentiment shocks that hit the Mexican banking system in a given period. At weekly frequency it is quite noisy, as depicted in Figure 2. Ideally, we would like to have a smoother cumulative sentiment index that maintains a weekly frequency in the observations, but that shows a more definite trend. We can consider the baseline weekly Twitter sentiment index as noisy observations of the actual unobserved sentiment. Our goal is to extract the trend from the time series of the weekly sentiment index, omitting the noisy high frequency components.

We take inspiration fromBorovkova 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), that is indicated to smooth high frequency data (such as daily, weekly or monthly). It is a filter that 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 the filter they propose dominates the HP filter in terms of flexibility in selecting the frequency bands of interest and possibility of adapting the filter to time series of quarterly, monthly or even higher data frequency. This property is particularly important in our case, since we have weekly data and we are not focusing specifically on studying business cycle frequencies, a case where the HP

filter works particularly well. Our focus is only to filter the high frequencies, while maintaining the lower ones. 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 and King filter is based on a moving average of the data with symmetric weights on leads and lags, so it throws away 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.

To filter exclusively the high frequencies, we enlarge the band of the Christiano-Fitzgerald filter up to 100 years. In this way, the band-pass filter becomes a sort of low-pass filter, that eliminates only the frequencies higher than the lower bound, and it maintains the lower frequencies up to the long run. Ideally the upper bound should go to infinity, but as an approximation we fix it at 100 years.[15]

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 still gives noisy results. As a result, we will focus the rest of the analysis on the filtered series that uses the window 1-100 years when we refer to the filtered sentiment index.

6 Descriptive results

6.1 The Indice de Estrés de los Mercados Financieros (IEMF)

Systemic risk is a multifaceted phenomenon, hard to measure at a uni-dimensional level. To measure systemic risk, one needs to use methodologies that can summarize information coming from many variables in a unique indicator. Examples of such stress indicators are the ones compiled by the Federal Reserve Bank: the St Louis Fed Financial Stress Index (Kliesen and McCracken, 2020), the Chicago Fed National Financial Conditions Index, and 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., 2015). The International Monetary Fund also publishes Financial Soundness Indicators for emerging market countries (IMF, 2003). To obtain a systemic summary indicator, it is necessary to combine market and financial institution's information.

In the case of the Mexican financial system, Banxico elaborates the Indice de Estrés de los Mercados Financieros (IEMF) (Banco de Mexico, 2013), a financial market stress index 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 variables are selected according to their importance in the Mexican financial market so that they show a volatile behavior during periods of financial stress. The IEMF is built using principal components analysis, a non-parametric method that, according to the correlation structure of the variables, computes weights that assign more importance to those variables that contain the most information. The IEMF is updated weekly, and the weights are recalculated at each update. Its coverage starts from January 2005 to the present. The goal of the index is to have a timely, effective measure that captures the level of accumulated risk in the Mexican financial system at a given moment. A higher level of the index indicates a higher financial systemic risk. Due to its construction method (it is built using weekly averages of the variables that compose it) the IEMF is already partially smoothed. For this reason, we do not filter it before comparing it to our smoothed sentiment index.

The IEMF has a very different nature than the sentiment index that we build in this paper. On the one hand, the IEMF is built using "hard", quantitative variables that prove to have a significant role in determining financial

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

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. Our hypothesis is that the sentiment index would be correlated with the reaction of financial markets, reflected in the IEMF.

6.2 The filtered sentiment index and its sub-indices

As shown by the topic analysis and suggested by the peaks of sentiment in Figure 3, the sentiment measured by the Twitter sentiment index is correlated to different kinds of negative shocks that can hit 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 general Twitter sentiment index, dividing the sample of tweets into those classified as bringing reputational risk according to the LDA algorithm and all the others. We follow the same methodology that we use for the General sentiment index to also compute the two sub-indices.

Figure 4 shows the General sentiment index, the Reputational index and the Non-reputational one compared with the IEMF over the period 2008-2019.

[Insert Figure 4 here (color should be used)]

As in Figure 3, the classification model of our choice is the sentiment index built by majority voting. However, Figure 3 presented the baseline sentiment index, not filtered but computed on a monthly basis. In this case, since the IEMF has a weekly frequency, we present the results of the sentiment index based on the majority voting classifier, with weekly frequency, smoothed using the Christiano-Fitzgerald filter with the band starting at the 1-year frequencies up to 100 years frequencies. It is possible to distinguish two periods where the sentiment index presented in Figure 3 was hit by different news shocks. In 2012 the Reputational index rises until a peak at the end of the year, coinciding with the HSBC scandal. The Reputational index has a second local peak in 2014, during the Oceanografía scandal. After 2015 there are only lower peaks that coincide with news about the development of the past scandals: new evidence about the scandals or a new phase in the judicial process. The General sentiment index more closely follows the Non-reputational one, and their trend is more in line with the IEMF than the Reputational index.

We find that the peaks of the Non-reputational index follow more closely the IEMF peaks as described in the Financial Stability Reports of Banco de Mexico. The Financial Stability Report has published the IEMF among the indicators of systemic financial risk since 2013. In 2011 and 2012 the uncertainty about the Greek default and the default risk of systemic banks in Spain make the IEMF spike; BBVA and Santander are also among the main commercial banks in Mexico. We also find that the bank fragility in Spain and uncertainty about the sovereign default in Greece are news reported in our tweets database. However, we find more tweets about the banking scandals occurring during 2012, so the peak of the reputational sub-index is higher.

In 2013 and 2014 the IEMF reports spikes of financial risk associated with the publication of the minutes of the Federal Reserve. In June 2013 the Fed announced the slowdown in the Quantitative Easing program (QE), and the expected end of the program in October 2014. In our database of tweets we find news about the effects of the announcement of the slowdown of QE 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 sub-index. In 2014 the Oceanografía scandal also happens in Mexico. Most of the tweets in our database comment on this, given the negative effect it had on Citibanamex. The news about Oceanografía spread from February 2014 to August 2014. This scandal, with its negative sentiment, accomplishes the most in our sentiment index and its sub-indices in that year.

During 2015 the IEMF goes through a stabilization first, and later a rise in financial stress, given in part by the end of the asset purchasing program in the previous year, and in part to rising expectations of an increase in interest

rates by the Federal Reserve (as happened in December 2015). In 2015 tweets report news about the depreciation of the peso, weak growth, the increased strength of the dollar and expected international contagion from the interest rates increase in the US and 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 due to the HSBC money laundering scandal. In 2016 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 fall of oil prices and oil revenues in Mexico. Regarding the sentiment index, starting in 2016 the banking scandals and frauds have less weight, so the Reputational 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 NAFTA agreement. In 2018 risk builds up because of the electoral process in Mexico and uncertainty linked to the T-MEC negotiation talks. At the end of 2018, higher volatility and uncertainty on 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 general, sentiment index shows a peak in the second half of 2017, due to news about the September earthquake that hit Mexico, and another one at the end of 2018, due to news about the airport cancellation and the proposal of revising the banking commissions. Finally, in 2019 the risk increases because of uncertainty over the credit perspectives of Pemex and Mexico. In March and June of 2019, Pemex corporate debt and Mexico's sovereign debt suffered a downgrade.

Not all the peaks of the two indices coincide, but we can see that both signal the main news. Also, the two indices are moving in the same direction, presumably, due to common causes. In other words, the sentiment index based on news is capturing information of importance for the systemic risk, and the news reports events that affect financial risk as measured by other indicators.

To test if there is a significant correspondence between our Sentiment Index and the IEMF, we compute the correlation of the IEMF with the Non-reputational sentiment index, the Reputational index and the General index, according to the majority voting model and the other three classifiers. Given the evidence in Figure 4, we expect a more positive correlation of the general Sentiment Index and the Non-reputational sub-index with the IEMF than the sub-index built on reputational tweets.

Column 1 of Table 8 shows the correlation between the IEMF and the different unfiltered indices, computed on the sample of tweets starting from 2008 to 2019. Column 2 shows the coefficients of the correlations between the IEMF and the filtered sentiment indices.[16] In all cases the filtered version of the index is more correlated with the IEMF than the non-filtered one. The filtered Voting sentiment index is the one that shows the highest correlation with the IEMF, reaching a significant positive correlation of more than 40 percent in the case of the General index and more than 49 percent for the Non-reputational index. The Reputational index is not significant, or it is negatively correlated with the IEMF, signaling that the Non-reputational index may be the one that contains more information regarding systemic risk.

[Insert Table 8 here]

As a robustness check, we perform the same correlations using the alternative models of sentiment. However, the correlation between the IEMF and these alternative variations is lower than those presented for the Voting sentiment index. The index obtained using the SVM classifier is the one that presents a closer correlation to the Voting sentiment index: the Non-reputational index is significantly correlated with the IEMF by 39 percent and the General sentiment index based on SVM has a significant and positive correlation with the IEMF of 22 percent. The correlation between the filtered Dictionary index and the IEMF is higher than the correlation between the filtered SVM index and IEMF for the General index, but lower in the other cases. The correlation between the

[16] When we mention the "filtered index" we will refer always to the version computed using the 1 year -100 years band.

IEMF and the index built on Neural Networks has the opposite sign than our expectations. The correlation between our sentiment indices and the financial index of reference, the IEMF, are in line with the findings in Shapiro et al. (2018). In their paper, they compute correlations between the sentiment measures they build and various economic outcomes, among them the S&P500, corresponding to the IPC for Mexican data. The correlations in Shapiro et al. (2018) vary between 2 percent and 47 percent In particular, the S&P500 is correlated with the sentiment measures by at most 22 percent.

The evidence presented in Figure 4 and Table 8 suggests that our intuition is correct. The Non-reputational sentiment index, built using textual sources, is correlated with the indicator of financial market stress, constructed with quantitative variables. The data and the methodologies that we use to build the 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.

As a robustness check, we compute regressions of the IEMF on the Voting sentiment index in the non-filtered and filtered versions (Table 9).

[Insert Table 9 here]

The regression confirms the results 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 last three columns, where the indices are taken alone. The non-reputational 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 is decreasing, while the reputational index is increasing due to the Oceanografía scandal. Finally, the coefficient of the complete sentiment index is always positive and significant, both in the regression with all the indices (column 5) and taken alone (column 8). An increase in 1 percent of the total sentiment index is correlated with an increase of 0.54 percent of the IEMF.

7 Predictive accuracy

We take inspiration from the work of Shapiro et al. (2018) to test if the Twitter sentiment index contains predictive information on the IEMF or specific financial market indicators. We refer in particular to six variables that we use as proxies for 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 Tesorería 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 (Indice 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 behavior and sentiment of other investors.

The 1-month FIX exchange rate volatility is our proxy for foreign exchange market risk. A higher exchange rate volatility increases the 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 like the S&P 500) 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 comprised between 0 and 1.

Finally, we use the JP Morgan Emerging Market Bond Index Plus (EMBI+) for Mexico as an indicator of country risk. 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: higher EMBI+ corresponds to higher risk.

Table 10 presents the correlations between the selected variables and the three versions of the filtered Sentiment Index, in line with our previous results. The General index, the Reputational, and the Non-reputational ones correlate with the expected sign with the variables considered. The Non-reputational one positively correlates with each financial variable, as expected, and the correlation is higher than 25 percent in most of the cases, with the exchange rate volatility and the EMBI+ being the variables with the highest correlation (37 percent 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 14 percent and 13 percent respectively.

[Insert Table 10 here]

These results are in line with previous literature: Shapiro et al. (2018) find correlations between their sentiment indices and the growth rate of the S&P500 index in a range of 6 percent to 22 percent in absolute value. The Reputational index and the financial variables are correlated with a negative sign, as expected from the analysis of the correlations between the IEMF and the sentiment indices. The General sentiment index correlates positively with each variable. The correlation coefficient is a bit lower than the non-reputational index, because the General index includes both the effect of the Non-reputational index and the Reputational one.

To explore whether our Twitter sentiment index has predictive power about 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 less restrictive. We stress that we do not want to claim causality on these results. As stated by Shapiro et al. (2018), even if the correlation between sentiment indicators and financial variables exist, the direction of the causality is still not clear. 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 news sentiment index and y_j itself:

$$y_{j,t+h} = \alpha_j^h + \beta_j^h SI_t + \sum_{i=1}^{n} \gamma_{j,i}^h SI_{t-i} + \sum_{i=1}^{n} \delta_{j,i}^h y_{j,t-i} + \varepsilon_{j,t+h} \tag{4}$$

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 General sentiment index as our baseline. We select the number of lags according to the Schwartz Bayesian Information Criteria (SBIC), 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 fit of the models. To keep the models comparable, we compute the SBIC for three models: an AR(1), an AR(4) and an AR(12).[17] In all cases, first we compute the model where we include only the dependent variable y_j and its lags, then we compute the same model considering both y_j and the sentiment index as an exogenous variable. We calculate the Information Criteria 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,

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

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.

[Insert Figure 5 here (color should be used)]

The first does not include the sentiment index, the second has the General sentiment index, 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, General 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 General sentiment index. These results imply that the sentiment index improves the forecasting ability of a model that considers only the dependent variable.

Finally, we use local projections according to Equation (4) to analyze the impact of a one standard deviation shock of the filtered General Twitter sentiment index on each of the variables of interest. A positive shock (a shock that increases the sentiment index) would be a shock that is positively correlated with the negative sentiment about financial markets and banks, so a shock that may increase financial market risk. The results in Figure 6 confirm this hypothesis. A one standard deviation shock in the sentiment index correlates with an increase of the IEMF, and the rise becomes significant after three weeks. The effect on the IEMF reaches its peak after 20 weeks, starting to decline thereafter.

[Insert Figure 6 here]

Figure 7 presents the impulse response functions of a one standard deviation shock of the sentiment index on the selected financial variables. A positive 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. There is also a significant increase in country risk as measured by the EMBI+ for Mexico. It rises in the moment of the shock, reaching 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 affected by a shock in the Twitter sentiment index. The banking sector, proxied by the beta of financial institutions, also reacts with an increase to a shock in the sentiment index, although the reaction is not significant in the short run.

[Insert Figure 7 here]

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.

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

[Insert Figure 8 here]

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 General 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.

[Insert Figure 9 here]

Finally, we test if using different financial variables as proxies for the different market risks we obtain 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 of derivative market risk, the annual growth of FIX exchange rate as a proxy of foreign exchange market risk, the annual yield of IPC as a proxy of 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.

[Insert Figure 10 here]

The banking funding rate spread reacts positively to a one standard deviation shock of the non-reputational sentiment index, 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 growth of exchange rate 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 in 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.

8 Conclusion

Our paper contributes to the growing literature that apply sentiment analysis on 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 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 NLP techniques to analyze the sentiment of Twitter messages, and we 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 potentially have an impact on banking risk and financial risk. 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 different 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 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, although the reaction is not significant in the short run.

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

A Tables

Type of source	Name	Type of source	Name
Mexican	El Financiero	Foreign	El País
newspapers	El Economista	newspapers	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	BBC (in Spanish)
	Capital Mexico	television	
	Reporte Indigo	Rating agencies	Moody's
	El Heraldo de México		Fitch Ratings
	La cronica de hoy		
	SDP noticias		

Table 1: Twitter accounts considered in this study

Word	Complete extraction
vía	130
norte	123
cantabria	106
centro	80
venta	77
colombia	76
popular	73
bucaramanga	68
tarjeta	68
día	67

(a) The 10 most frequent words in the complete extraction

Word	Extraction from selected accounts		Complete extraction	
	Order	Frequency	Order	Frequency
director	1	6	34	21
financiero	2	5	48	16
general	3	4	21	28
mercados	4	3	94	7
parte	5	3	23	26
dea	6	3	66	10
vamos	7	3	51	14
crecimiento	8	3	54	13
ser	9	3	3	65
presidente	10	3	31	19

(b) Comparison between the 10 most frequent words in the extraction from selected accounts and the frequency of the same words in the complete extraction

Word	Complete extraction		Extraction from selected accounts	
	Order	Frequency	Order	Frequency
centro	1	80	78	1
ser	2	65	10	3
cuenta	3	62	187	1
dos	4	59	148	1
así	5	58	24	2
mejor	6	55	163	1
bancos	7	46	31	2
hace	8	45	104	1
cómo	9	45	147	1
años	10	44	23	2

(c) Comparison between the 10 most frequent words in the complete extraction and the frequency of the same words in the extraction from selected accounts

Table 2: Comparison between the extraction of Tweets without selection of accounts and the extraction from selected accounts (March 20, 2019)

Date	Text	User	Followers	Country
08/09/2010 19:20	Asigna Moody's calificación de deuda senior a Banamex	LaRazon_mx	122751	Mexico
25/11/2011 12:24	El Gobierno indulta al consejero delegado del Banco Santander, Alfredo Sáenz.	el_pais	6818004	Spain
17/07/2012 16:31	HSBC de EEUU se disculpa por fallas que permitieron narcolavado.	AP_Noticias	222131	USA
22/07/2013 16:50	Utilidades de #UBS superan expectativas	eleconomista	447505	Mexico
14/01/2014 13:00	#ReformaEnergética: un elemento de cambio en México. Adolfo Acebrás de @UBS ahonda en el tema.	Forbes_Mexico	507926	Mexico
09/02/2015 14:44	Cómo el banco HSBC "ayudó" a millonarios a evadir impuestos.	bbcmundo	3163376	UK
30/09/2016 20:18	El Banco Santander baja su objetivo de rentabilidad por el Brexit #AFP	AFPespanol	285893	Uruguay
02/02/2017 17:23	En condiciones actuales, aumento de gasolina sería de 0.5%: Banco Base.	El_Universal_Mx	4941610	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, enentrevista con @VictorPiz en #AlSonarLaCampana.	ElFinanciero_Mx	1181553	Mexico
01/02/2019 00:40	Analistas de Barclays y BNP Paribás advirtieron que inversionistas de WallStreet están preocupados por la situación de Pemex.	eleconomista	447506	Mexico

Table 3: Selected tweets from our database.

Model	Sentiment	Sentiment by voting
A. General agreement		
Dictionary	Positive	
SVM	Positive	Positive
Neural networks	Neutral	
B. Disagreement		
Dictionary	Positive	
SVM	Negative	Neutral
Neural networks	Neutral	

Table 4: Classification of sentiment by majority voting

Most frequent words in English reports			Most frequent words in Spanish reports			Words with the stronger polarity in Tweets		
Word	Polarity	Freq in reports	Word	Polarity	Freq in reports	World	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

Table 5: CKJM dictionary modified

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

Table 6: Models' performance results

	SI dictionary	SI SVM	SI NN	SI voted
Model 1				
SI dictionary	1			
SI SVM	0.5508*	1		
SI neural networks	0.4897*	0.5458*	1	
SI voted	0.6750*	0.7711*	0.7264*	1
Model 2				
SI dictionary	1			
SI SVM	0.5287*	1		
SI neural networks	0.3896*	0.4505*	1	
SI voted	0.6863*	0.7643*	0.6120*	1
Model 3				
SI dictionary	1			
SI SVM	0.4481*	1		
SI neural networks	0.2359*	0.1949*	1	
SI voted	0.5954*	0.7035*	0.4250*	1

Note: *: p-value<0.1; (1): SI computed not considering neutral tweets (2): SI computed considering neutral tweets, (3): SI computed not considering neutral tweets and weighting the tweets by the number of reactions to the tweet.

Table 7: Correlation between alternative sentiment indices

Correlation with IEMF	Not filtered index (1)	Filtered index (2)
SI voted, non reputational	0.1318*	0.4634*
SI voted, reputational	-0.0508	-0.1487*
SI voted, total	0.1342*	0.4008*
SI dictionary, non reputational	0.1214*	0.3739*
SI dictionary, reputational	0.0454	0.0381
SI dictionary, total	0.1327*	0.3578*
SI SVM, non reputational	0.1169*	0.3962*
SI SVM, reputational	-0.0557	-0.1421*
SI SVM, total	0.0897*	0.2254*
SI neural networks, non reputational	-0.0231	-0.2178*
SI neural networks, reputational	-0.1058*	-0.2785*
SI neural networks, total	-0.00390	-0.1866*

p-value<0.1; Filtered index: obtained applying the Christiano-Fitzgerald filter, with band 1 year-100 years neutral tweets.

Table 8: Correlation between sentiment indices and IEMF

IEMF	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
SI voted, non reputational	-0.019	0.036***						
	(0.030)	(0.011)						
SI voted, reputational	-0.032**		-0.014					
	(0.013)		(0.011)					
SI voted, total	0.068**			0.039***				
	(0.033)			(0.012)				
SI filtered, non reputational					-0.150	0.652***		
					(0.125)	(0.052)		
SI filtered, reputational					-0.273***		-0.083***	
					(0.030)		(0.023)	
SI filtered, total					0.977***			0.538***
					(0.136)			(0.051)
Constant	0.296***	0.289***	0.303***	0.284***	0.192***	0.195***	0.329***	0.167***
	(0.008)	(0.006)	(0.008)	(0.007)	(0.013)	(0.010)	(0.011)	(0.013)
Observations	589	589	589	589	589	589	589	589
R-squared	0.028	0.017	0.003	0.018	0.315	0.215	0.022	0.161

Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. Voted sentiment index. Where filtered, it is filtered using CF filter with band 1 year-100 years.

Table 9: OLS regression of IEMF on the sentiment index, majority voting

Sentiment Index	(1) Not 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*
3m swap rate spread	0.126*	-0.134*	0.143*
EMBI+	0.372*	-0.151*	0.333*
3m sovereign bond spread	0.360*	-0.265*	0.202*

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

Table 10: Correlations between the Voting sentiment index and selected market variables, 2008-2019

B Figures

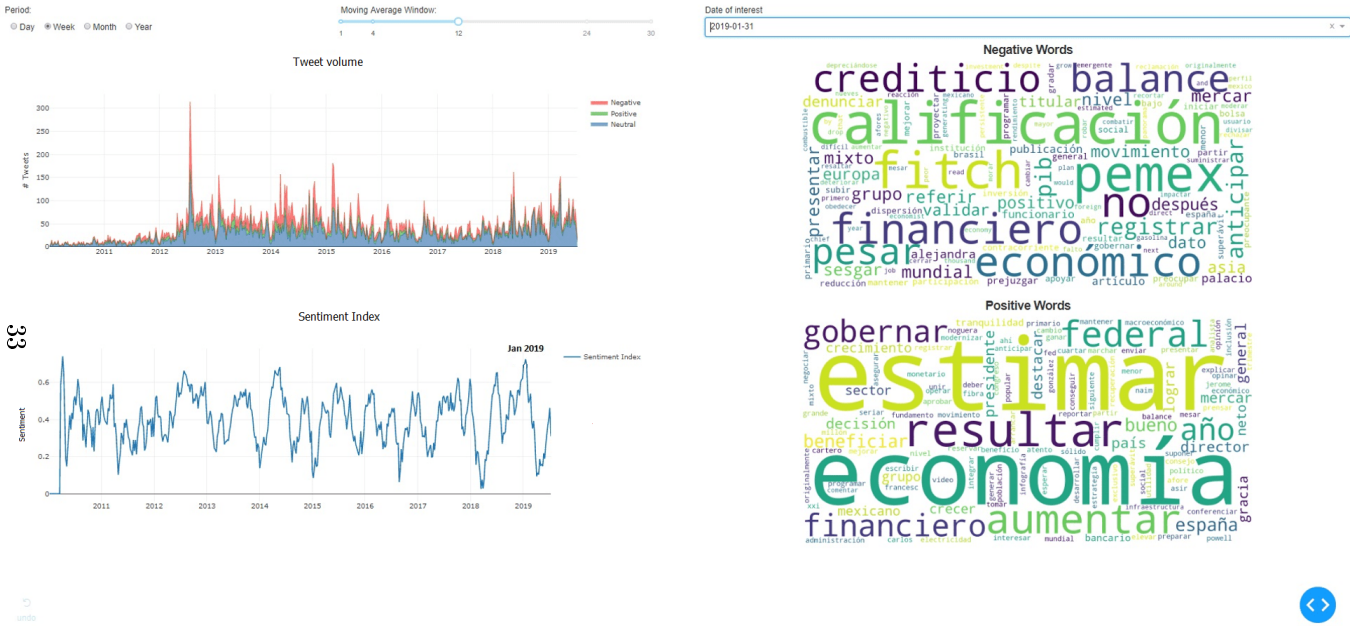
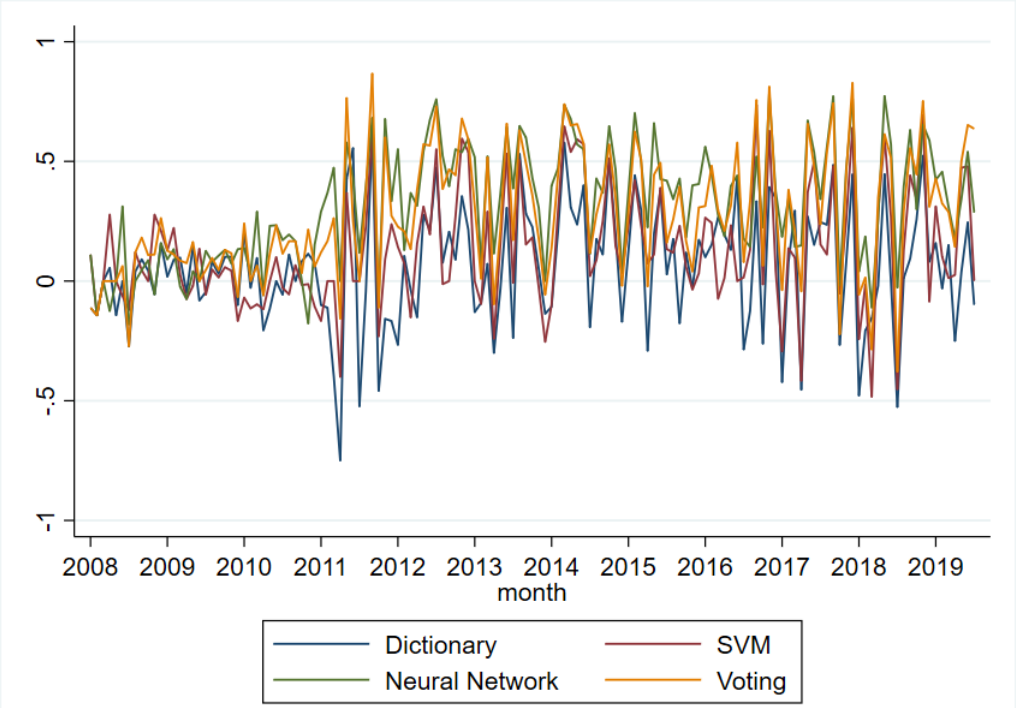
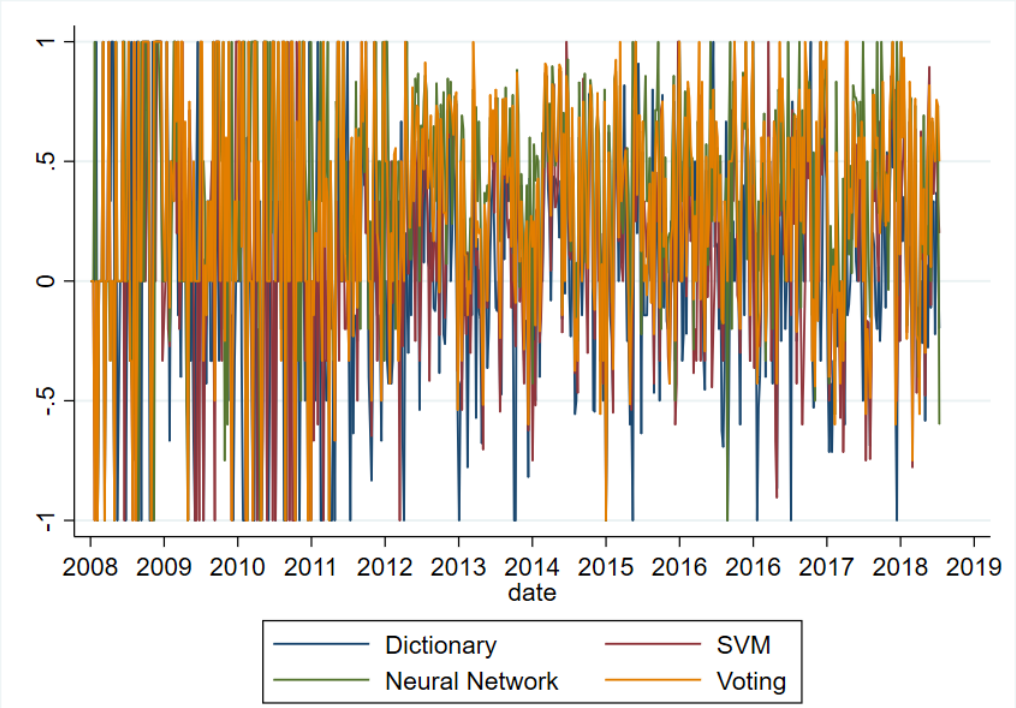


Figure 1: Sentiment index dashboard: volume of tweets, monthly sentiment index, and positive and negative trending words (January 2019)



(a) Monthly frequency



(b) Weekly frequency

Figure 2: Comparison between the four sentiment indices

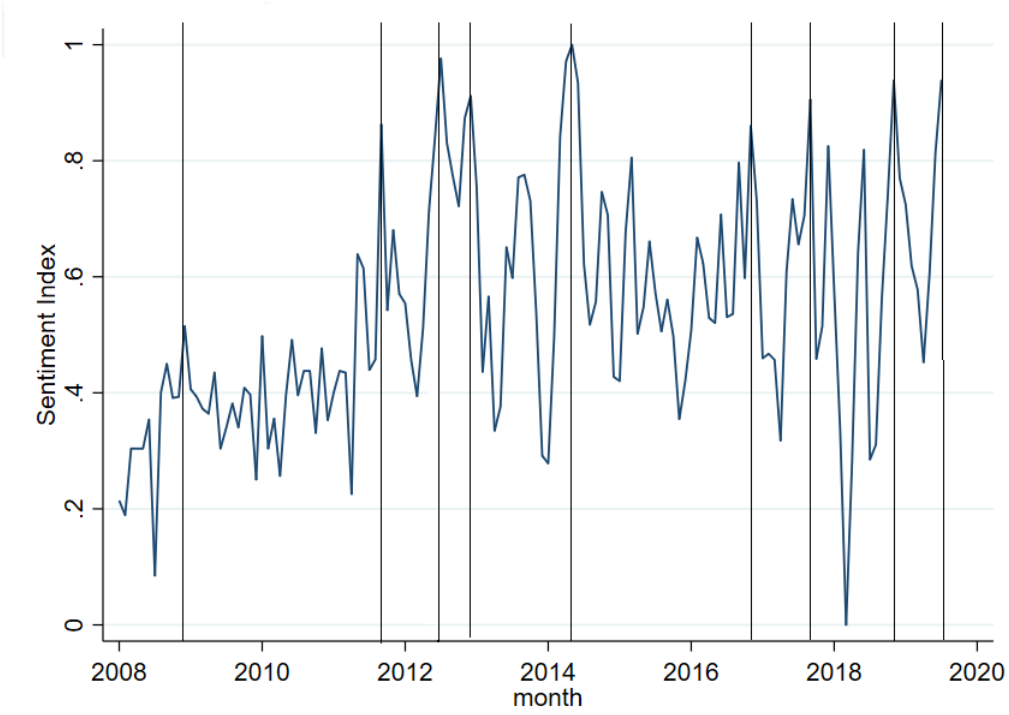


Figure 3: Sentiment index (majority voting), monthly frequency

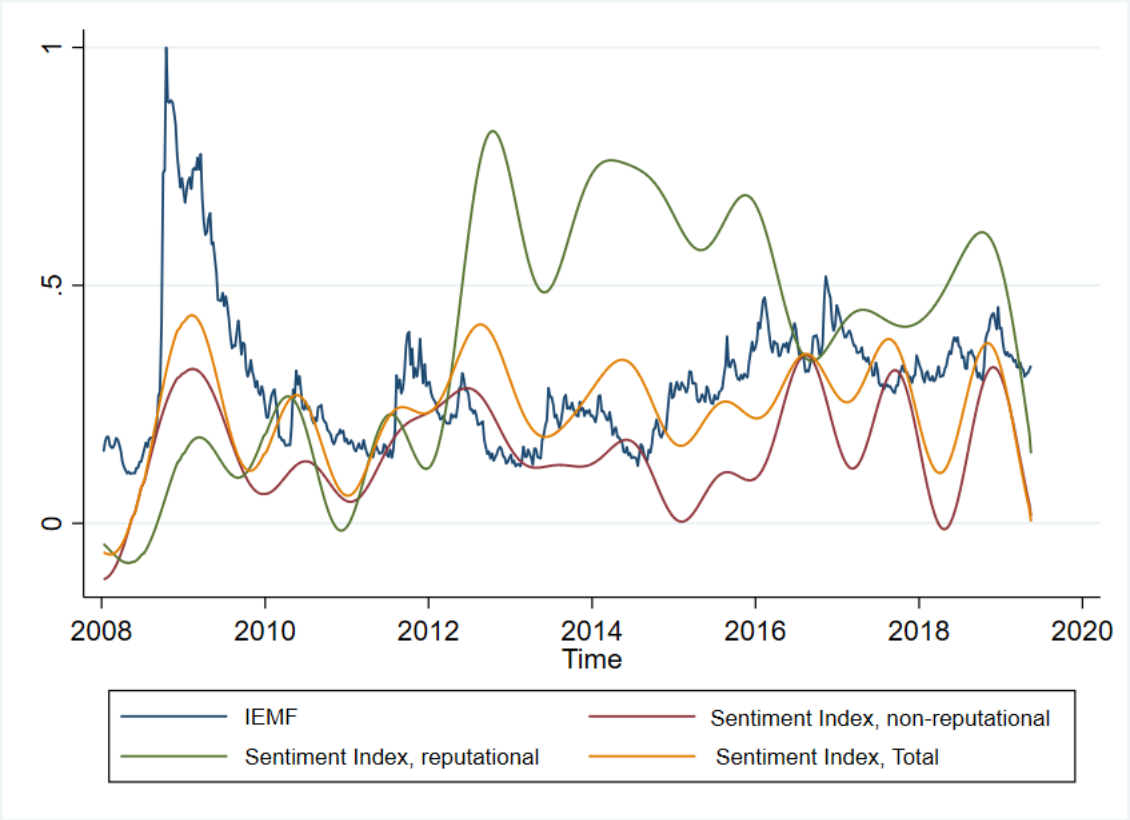


Figure 4: Comparison between the IEMF and the filtered sentiment index. CF filter bands: 1 year-100 years

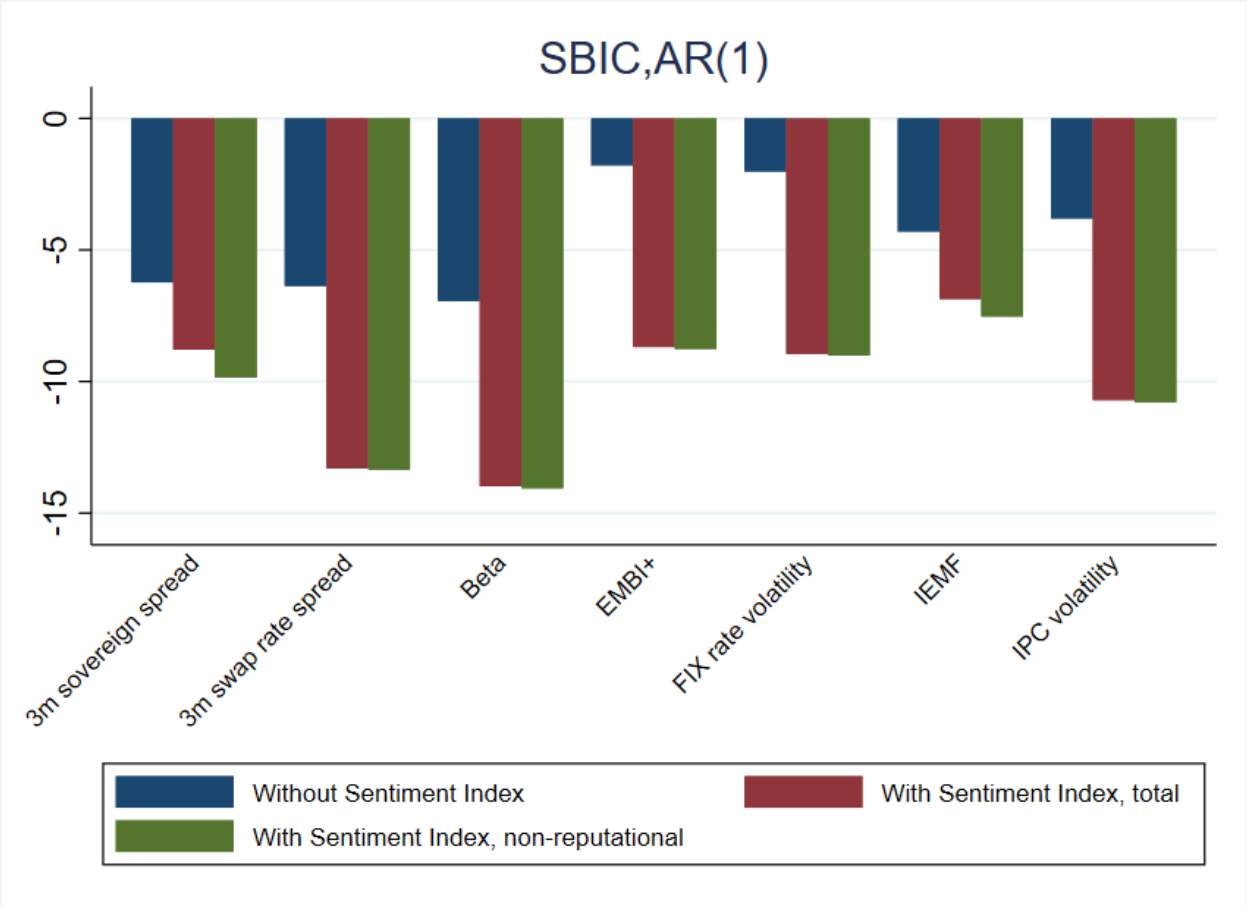


Figure 5: Bayesian information criteria for selected variables. Sentiment index filtered using the 1 year-100 years band

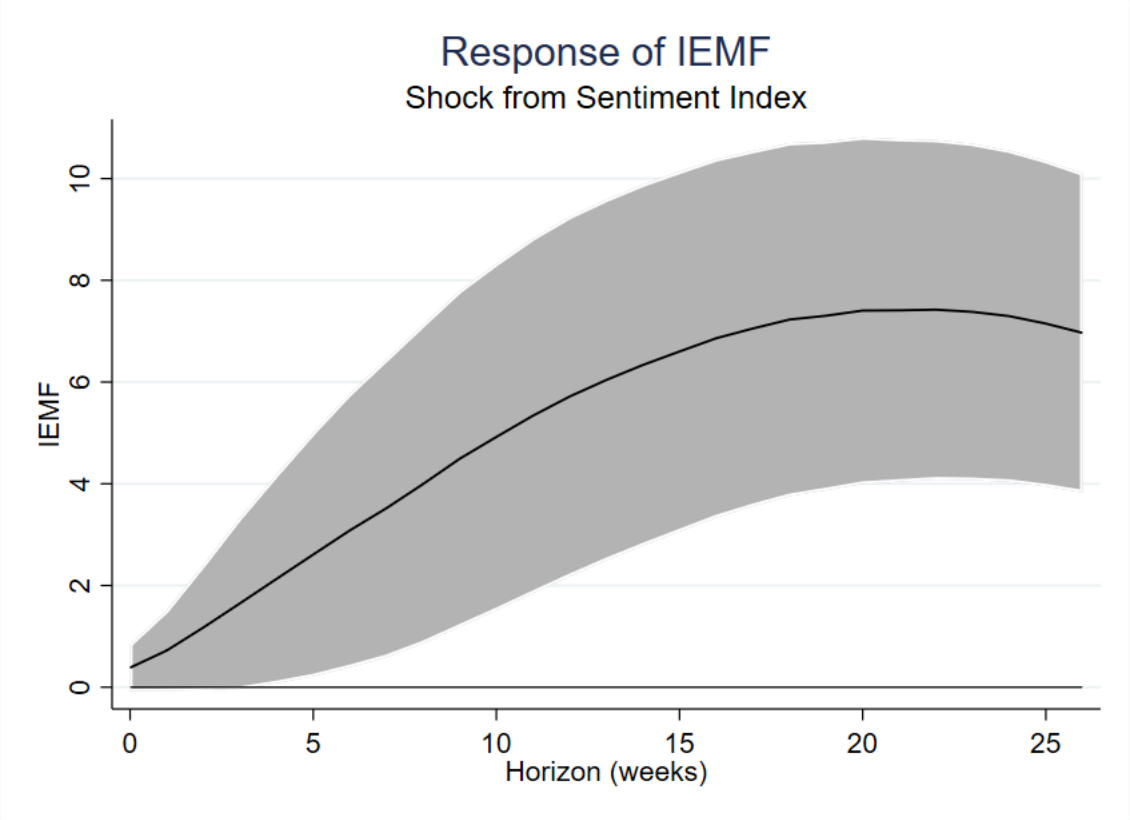


Figure 6: IRFs for IEMF. Impulse variable: General sentiment index, filtered using the 1 year-100 years band

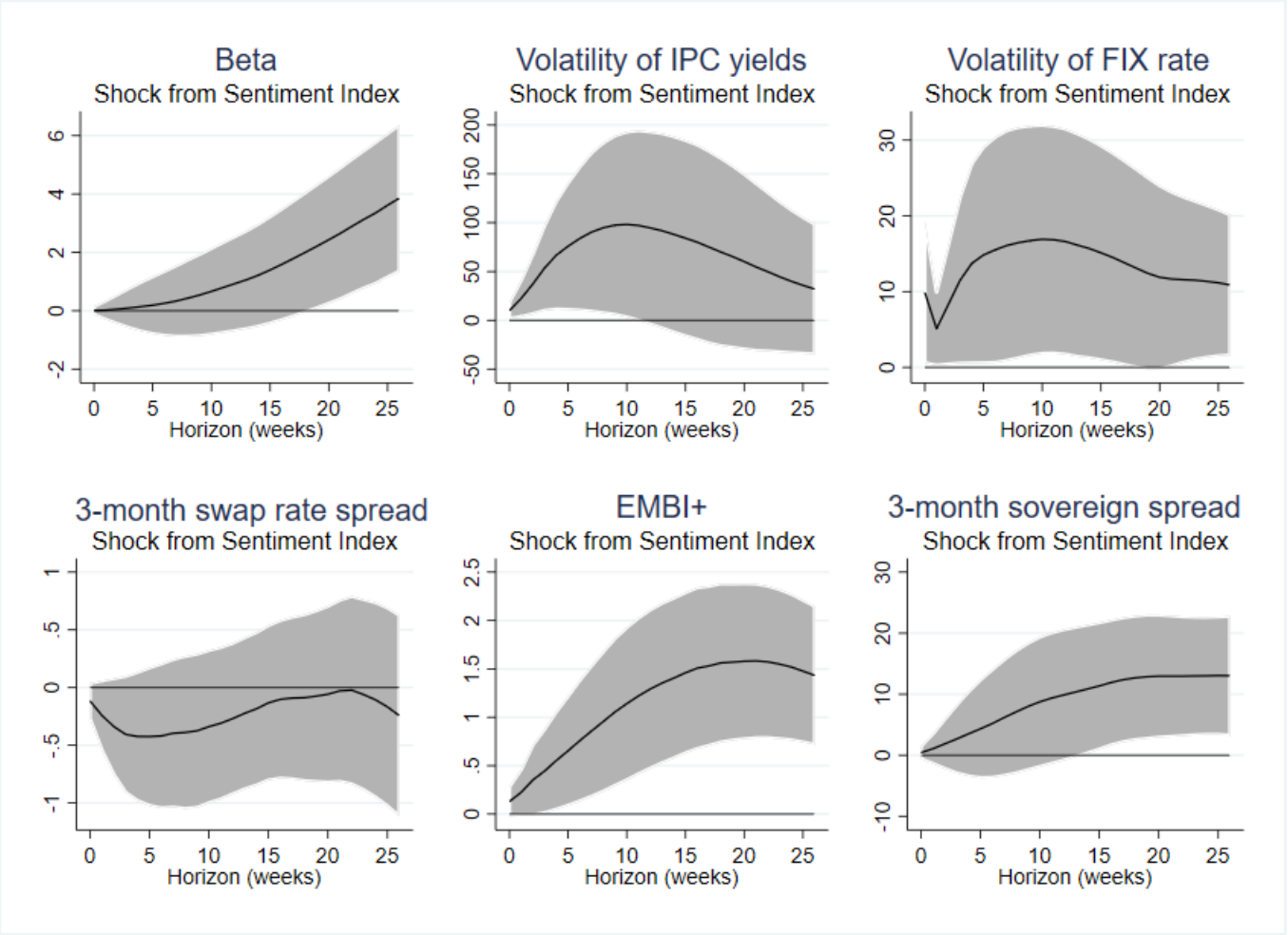


Figure 7: IRFs for the six financial variables. Impulse variable: General sentiment index, filtered using the 1 year-100 years band

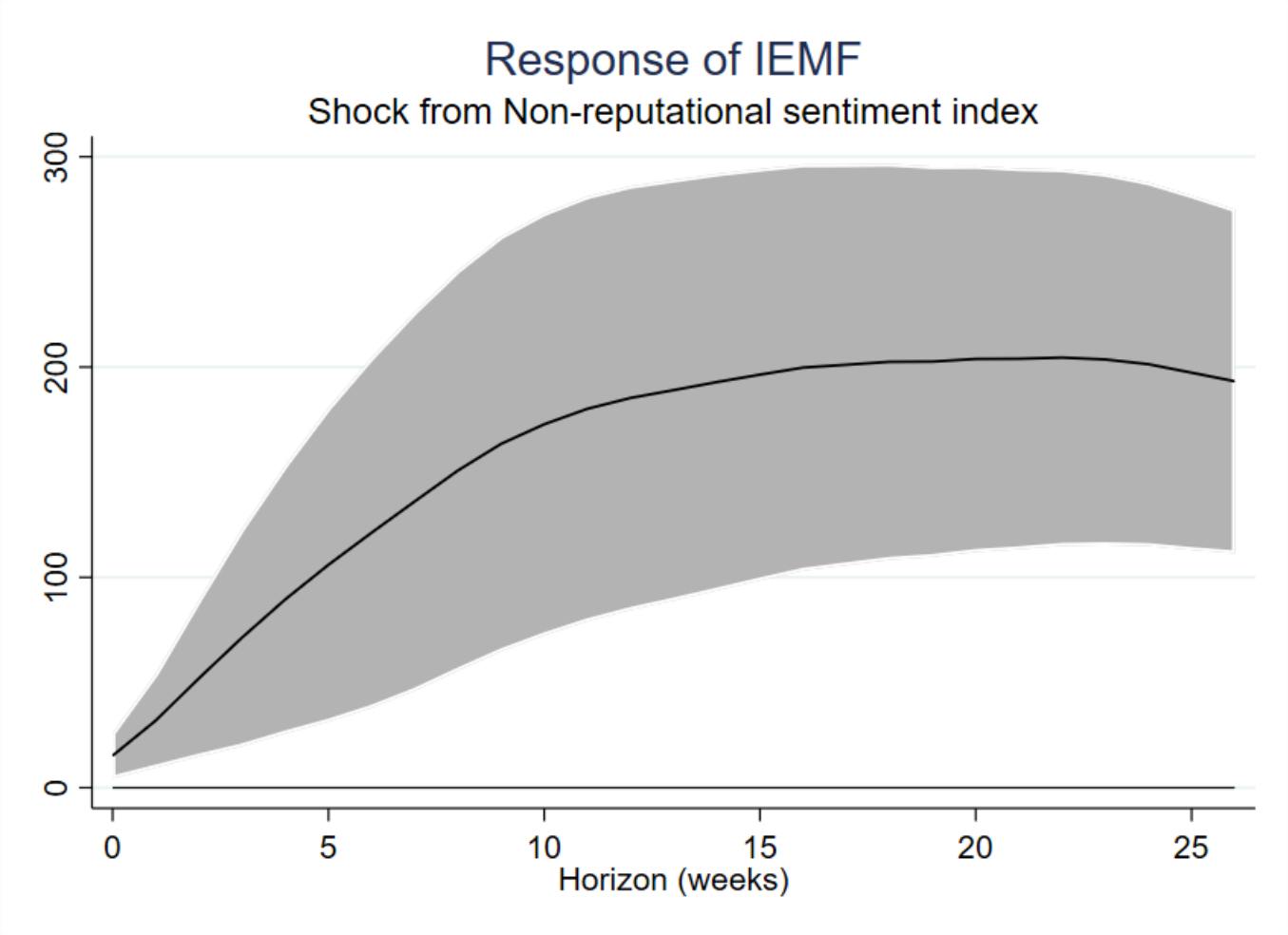


Figure 8: IRFs for IEMF, robustness check. Impulse variable: Non-reputational sentiment index, filtered using the 1 year-100 years band

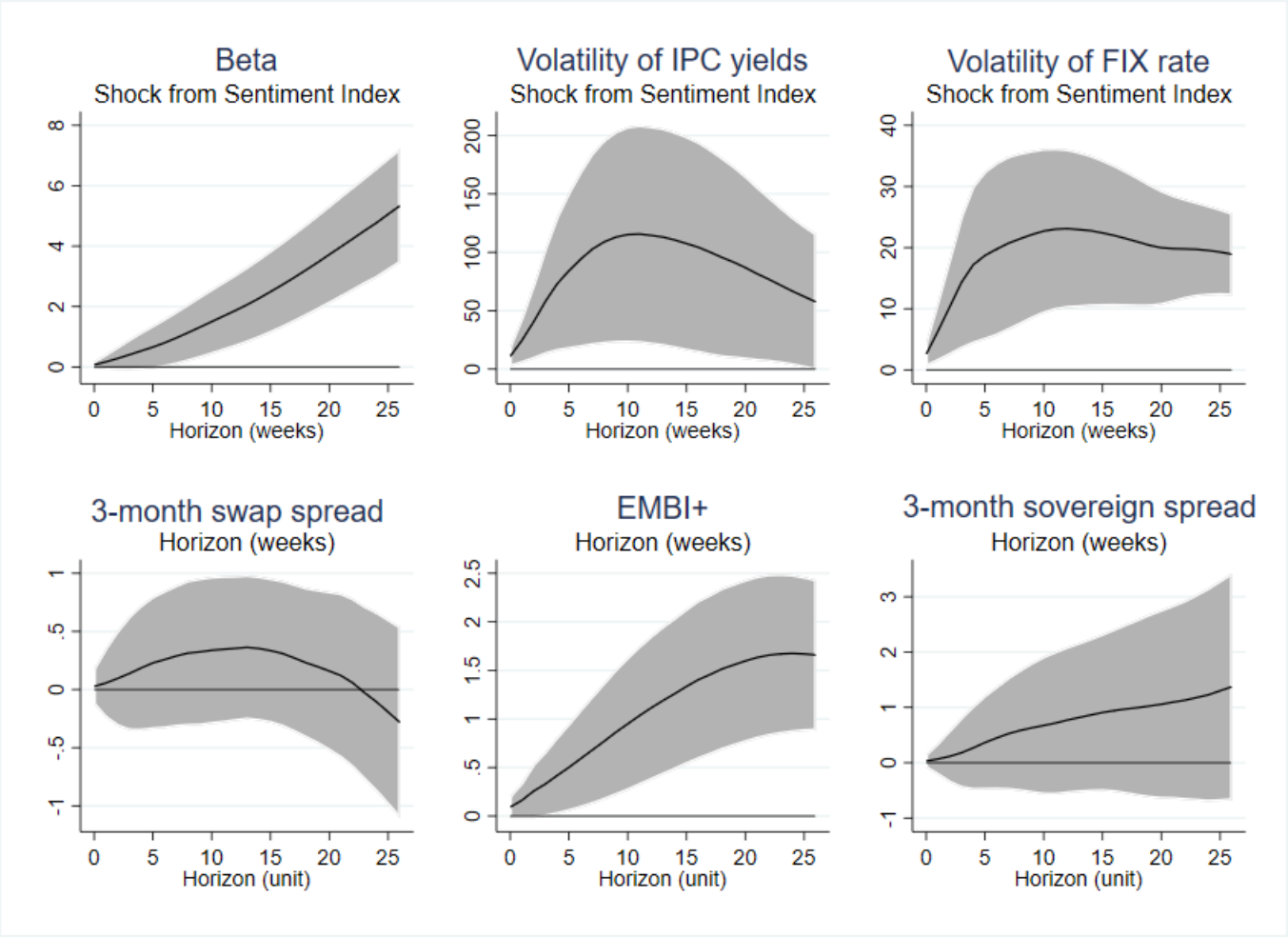


Figure 9: IRFs for the six financial variables, robustness check. Impulse variable: Non-reputational sentiment index, filtered using the 1 year-100 years band

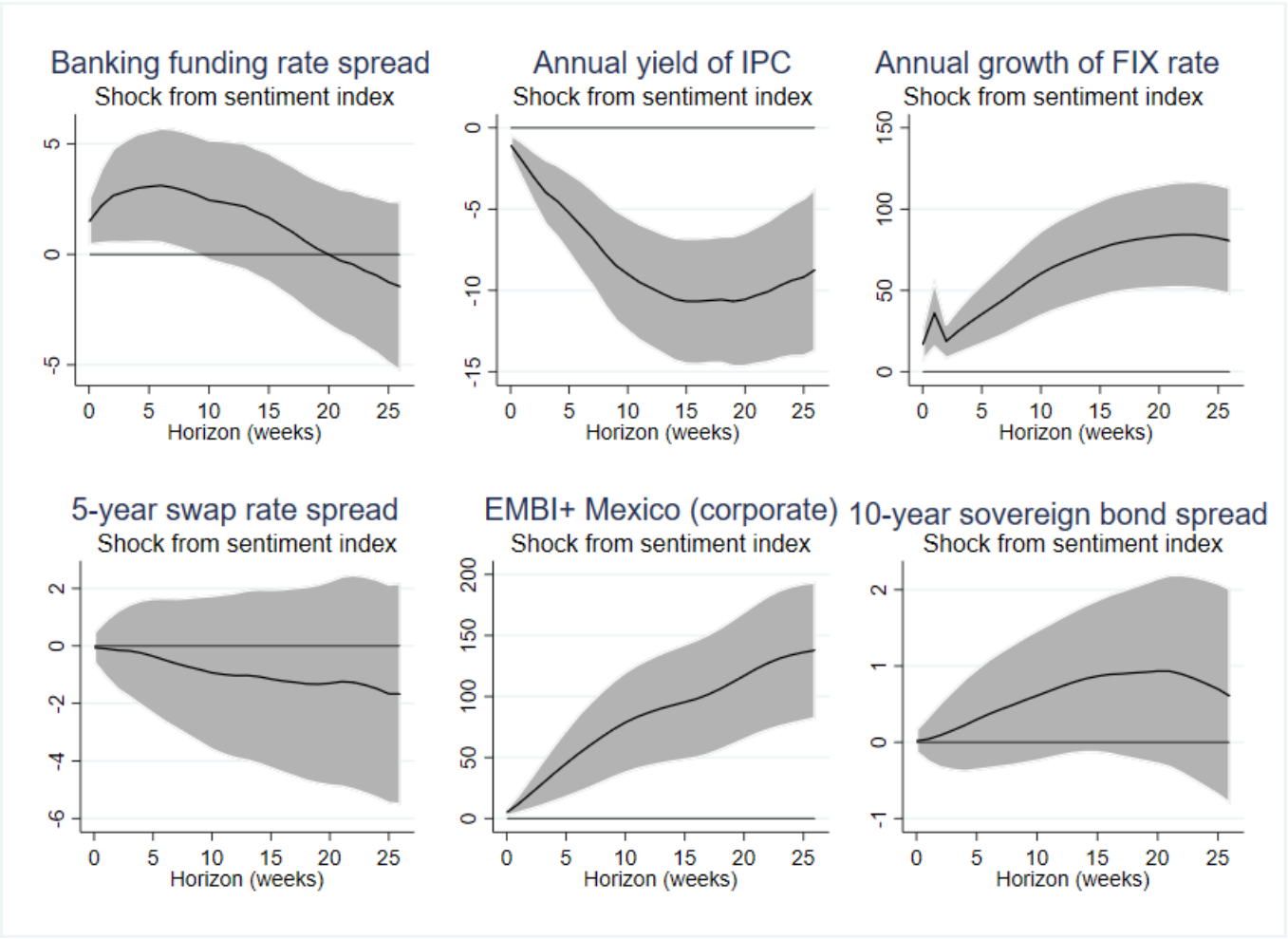


Figure 10: IRFs for the alternative financial variables, robustness check. Impulse variable: General sentiment index, filtered using the 1 year-100 years band

References

Accornero, M. and M. Moscatelli (2018). Listening to the buzz: social media sentiment and retail depositors' trust. Technical report, Bank of Italy.

Akerlof, G. A. and R. J. Shiller (2009). *Animal Spirits: How Human Psychology Drives the Economy and Why It Matters for Global Capitalism*. Princeton University Press.

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

Angelico, C., j. Marcucci, M. Miccoli, and F. Quarta (2018). Can we measure inflation expectations using twitter? Technical report, Bank of Italy.

Azar, P. D. and A. W. Lo (2016). The wisdom of twitter crowds: Predicting stock market reactions to fomc meetings via twitter feeds. *The Journal of Portfolio Management Special QES Issue 42*(5), 123–134.

Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics 131*(4), 1593–1636.

Banco de Mexico (2013). Financial stability report.

Banco de Mexico (2019). Financial stability report.

Barsky, R. B. and E. R. Sims (2012). Information, animal spirits, and the meaning of innovations in consumer confidence. *The American Economic Review 102*(4), 1343–1377.

Baxter, M. and R. G. King (1999). Measuring business-cycles: Approximate band-pass filters for economic time series. *The Review of Economics and Statistics* (81), 575–593.

Benhabib, J. and M. Spiegel (2017). Sentiment and economic activity: Evidence from u.s. states. Working Paper 23899, National Bureau of Econmic Research.

Bholat, D., S. Hansen, S. Pedro, and C. Schonhardt-Bailey (2015). Text miing for central banks. Technical report, 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.

Blanchard, O. (1993). Consumption and the recession of 1990-1991. *The American Economic Review 83*(2), 270–274.

Blei, D. (2012). Probabilitic topic models. *Communications of the ACM 55*, 77–84.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of machine Learning Research* (3).

Borovkova, S., E. Garmaev, P. Lammers, and J. Rustige (2017, April). Sensr: A sentiment-based systemic risk indicator. DNB Working Paper 553, De Nederlandsche Bank.

Boser, B. E., I. M. Guyon, and V. N. Vapnik (1992). A training algorithm for optimal margin classifiers. In *Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152.

Brugnolini, L. (2018). About local projection impulse response function reliability. CEIS Research Paper 440, Tor Vergata University, CEIS.

Bruno, G. (2018). Central bank communications: Information extraction and semantic analysis. Technical report, Bank of Italy.

Bruno, G., P. Cerchiello, J. Marcucci, and G. Nicola (2018). Twitter sentiment and banks' financial ratios: Is there any causal link? Technical report, Bank of Italy.

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

Buch, C., M. Bussierè, L. Goldberg, and R. Hills (2019). The international transmission of monetary policy. *Journal of International Money and Finance 91*, 29–48.

Bukovina, J. (2016). Social media big data and capital markets-an overview. *Journal of Behavioral and Experimental Finance 11*, 18–26.

Calomiris, C. W. and H. Mamaysky (2018, March). How news and its context drive risk and returns around the world. Working Paper 24430, National Bureau of Economic Research.

Cerchiello, P., P. Giudici, and G. Nicola (2017). Twitter data models for bank risk contagion. *Neurocomputing 264*, 50–56.

Cetorelli, N. and L. S. Goldberg (2011). Global banks and international shock transmission: Evidence from the crisis. *IMF Economic Review 59*, 41–76.

Chang, J., S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei (2009). *Reading Tea Leaves: How Humans Interpret Topic Models*, pp. 288–296. Curran Associates.

Christiano, L. J. and T. J. Fitzgerald (2003). The band pass filter. *International Economic Review* (44).

Correa, R., K. Garud, J. M. Londono, and N. Mislang (2017, March). Sentiment in central banks' financial stability reports. International Finance Discussion Papers 1203, Board of Governors of the Federal Reserve System.

Correa, R., K. Garud, J.-M. Londono-Yarce, and N. Mislang (2017, June). Constructing a dictionary for financial stability. Ifdp notes, Board of Governors of the Federal Reserve System.

Da, Z., J. Engelberg, and P. Gao (2011). In search of attention. *Journal of Finance 66*(5), 1461–1499.

Ding, R. and W. Hou (2015). Retail investor attention and stock liquidity. *Journal of International Financial Markets, Institutions and Money 37*, 12–26.

Duprey, T., B. Klaus, and T. Peltonen (2015). Dating systemic financial stress episodes in the eu countries. Technical Report 1873, European Central Bank.

Forss Sandahl, J., M. Holmfeldt, A. Ryden, and M. Stroemqvist (2011). An index of financial stress for sweden. *Sveriges Riksbank Economic Review*.

Hakkio, C. S. and W. R. Keeton (2009). Financial stress: what is it, how can it be measured, and why does it matter? *Economic Review, Federal Reserve Bank of Kansas City 94*(2), 5–50.

Hansen, S., M. McMahon, and A. Prat (2018). Transparency and deliberation within the fomc: a computational linguistic approach. *Quarterly Journal of Economics 133*(2), 801–870.

Hodrick, R. and E. C. Prescott (1997). Postwar u.s. business cycles: An empirical investigation. *Journal of Money, Credit, and Banking* (29).

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

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

IMF (2003). Financial soundness indicators - background paper. Technical report, International Monetary Fund.

Ingo, W. (2011). *Reputational Risk*, Chapter 6, pp. 103–123. John Wiley & Sons.

Jordà, O. (2005). Estimation and inference of impulse responses by local projections. *American Economic Review* (95), 161–182. March.

Kalamara, E., A. Turrell, C. Redl, G. Kapetanios, and S. Kapadia (2020). Making text count: economic forecasting using newspaper text. Staff Working Paper 865, Bank of England.

Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money*. London, Macmillan.

Kliesen, K. and M. McCracken (2020). The st. louis fed' s financial stress index, version 2.0.

Mimno, D., H. M. Wallach, E. Talley, M. Leenders, and A. McCallum (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 262–272.

Morais, B., J.-L. Peydro, and C. Ruiz (2015). The international bank lending channel of monetary policy rates and qe: Credit supply, reach-for-yield, and real effects. International Finance Discussion Papers 1137, Board of Governors of the Feredal Reserve System.

Moreno Bernal, n. I. and C. González Pedraz (2020). Sentiment analysis of the spanish financial stability report. Working Paper 2011, Bank of Spain.

Newman, D., Y. Noh, E. Talley, S. Karimi, and T. Baldwin (2010). Evaluating topic models for digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries, JCDL*, pp. 215–224.

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

Ormerod, P., R. Nyman, and D. Tuckett (2015). Measuring financial sentiment to predict financial instability: A new approach based on text analysis. Papers 1508.05357, arXiv.org.

Plakandaras, V., T. Papadimitriou, and G. Periklis (2015). Forecasting daily and monthly exchange rates with machine learning techniques. *Journal of Forecasting 34*.

Polikar, R. (2012). *Ensamble learning*.

Quinn, K. M., M. B. L., C. Michael, C. M. H., and D. R. Radev (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science 54*(1), 209–228.

Rakowski, D., S. E. Shirley, and J. Stark (2020). Twitter activity, investor attention, and the diffusion of information. *Financial Management*, 1–44.

Reinhardt, D. and R. Sowerbutts (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. *Artificial Intelligence Review 33*(1-2), 1–39.

Shapiro, A. H., M. Sudhof, and D. Wilson (2018). Measuring news sentiment. Working Paper 2017-01, Federal Reserve Bank of San Francisco.

Sprenger, T. O., A. Tumasjan, P. G. Sandner, and I. M. Welpe (2014). Tweets and trades: the information content of stock microblogs. *European Financial Management 20*(5), 926–957.

Tellez, E. S., S. Miranda-Jiménez, M. Graff, D. Moctezuma, R. R. Suárez, and O. S. Siordia (2017). A simple approach to multilingual polarity classification in twitter. *Pattern Recognition Letters* (94). 68-74.

Tripathy, J. (2020). Cross-border effects of regulatory spillovers: Evidence from mexico. *Journal of International Economics 126*, 103350.

Vlastakis, N. and R. N. Markellos (2012). Information demand and stock market volatility. *Journal of Banking and Finance 36*(6), 1808–1821.

Zimbra, D., A. Abbasi, D. Zeng, and H. Chen (2018). The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation. *ACM Transactions on Management Information Systems 9*(2).