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

Raúl Fernández†, Brenda Palma‡, Caterina Rho§

April 27, 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 considering whether they reflect a positive or negative shock on Mexican banks, or whether they are merely informative. We compare the performance of three classifiers: the first based on word polarities from a pre-defined dictionary, the second on a Support Vector Machine Classifier and the third on Neural Networks. We find that the Support Vector Machine classifier has the best performance of the three we test. We also compare this proposed Sentiment Risk Index with existing indicators of financial stress based on quantitative variables. We find that this novel index captures the effect of sources of financial stress that are not explicitly reported in quantitative risk measures, such as financial frauds, fails in payment systems and money laundering. We also show that a shock in the Twitter Sentiment Index increases stock market volatility and foreign exchange rate volatility, having a significant effect on overall financial market risk, especially for the private sector.

Keywords: sentiment analysis, systemic risk, banks.

JEL classification: G1, G21, G41

*We are grateful to Liduvina Cisneros, Jorge Luis García, Fabrizio López Gallo, Calixto López Castañon, Yahir López Chuken, Jorge De La Vega, Lorenzo Menna, Sabino Miranda, and Alberto Romero 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.

†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

Recent years saw the rise of production and storage of an unprecedented amount 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 has been an opportunity for policymakers.

The study of unstructured data (the so-called Big Data), such as social media content, is particularly interesting for central banks in the context of financial regulation and supervision. A growing literature focuses on studying social media activity, in particular Twitter messaging, on stock market fluctuations in coincidence with key events, such as monetary policy decisions (Azar and Lo, 2016). Others focus on the analysis of sentiment embedded in news regarding specific topics, such as financial risk (Borovkova et al., 2017).

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. This may have consequences for the financial sector, or for 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). These indicators complement existing quantitative indicators in two ways. On the one hand, they may reflect new information that traditional indicators do not measure explicitly. On the other hand, coming at higher frequencies, they may help policymakers by measuring expectations about core economic indicators, such as inflation and the GDP growth, that are usually built at monthly or quarterly frequencies.

In this paper, we use sentiment analysis to build a Mexican Sentiment Index based on tweets in Spanish. The index intends to capture media perception of risk in the Mexican financial system, which we think greatly influences the perception of financial stability of both investors and consumers. In order to perform the sentiment analysis on tweets, we apply known text mining and machine learning techniques.

Our contribution is threefold. First, we use Latent Dirichlet Allocation to analyze the topics of the tweets that are associated with a rise or fall in the Twitter Sentiment Index. The topic analysis shows that our index is able to capture sources of potential financial risk that are not traditionally included in financial stress indicators, such as financial frauds, money laundering, and failures of online payment systems. These events may cause negative shocks in terms of reputation for the banks or credibility of the banking system.

Second, we test three models to predict the tweets' sentiment and build the sentiment-based indicator for the Mexican financial sector. We start with a dictionary with word polarities as our baseline model. We modify the dictionary proposed in Correa et al. (2017), translating it to Spanish and adapting it to the context of social media interaction. At the best of our knowledge, this is the first financial dictionary in Spanish, specifically built for sentiment analysis in Twitter. We then work with a Support Vector Machine classifier (Tellez et al., 2017) and neural networks (Howard and Ruder, 2018). We propose a model which captures the responses of the three proposed classifiers, and determines the final sentiment based on a voting system.

Third, we test how well our index performs in comparison with alternative existing measures of financial stress. We apply local projections (Jordà, 2005) to test the effect of a shock of our Sentiment Index on selected financial variables. Our results show that a one standard deviation shock in the Sentiment Index increases stock market risk (proxied by the volatility of the Indice de Precios y Cotizaciones, IPC) and the Foreign Exchange market risk (proxied by the 1-month volatility of daily FIX rate).

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 are being developed to analyze 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 in advance sudden changes in market trends. Borovkova et al. (2017) propose a new Sentiment-based Systemic Risk indicator of the global financial system. They build it by aggregating sentiment in 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. (2019) 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 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, using sentiment analysis. 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) applies the dictionary to Tweets about selected Italian banks to extract sentiment indicators and relate them to some banks' financial variables. They find a positive correlation between financial variables and sentiment for some of the banks in their sample.

Correa et al. (2017) and Correa et al. (2017, a) also apply sentiment analysis to central bank's Financial Stability Reports. In particular, they analyze the relation between the financial cycle and the sentiment conveyed in these official publications. First they build a new dictionary of financial and economic terms, then they use their dictionary to build a financial stability sentiment index for 35 countries and a period of ten years, from 2005 to 2015. They find that the financial stability index is mostly driven by developments in the banking sector and by information about this specific sector. 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.

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. (2019) to test how our Twitter Risk Index performs in comparison with other measure of financial stress and economic uncertainty. We refer to the Financial Market Stress Index developed by 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 and we assume that these tweets reflect the perception of the Mexican banking system transmitted by the media. The banking system is at the core of Mexican financial system, therefore the health of the financial system as a whole is 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 limit the extraction to verified Twitter accounts of national and international newspapers, news agencies and rating agencies. 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.

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 in 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. The amount of information is drastically reduced by our selection.

However, Table 2 shows some interesting results about the relevance of the information extracted in the two cases. 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 part 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 compete sample, reinforcing the evidence shown in Panel (a).

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, and the country of origin of the tweet, among others. The database consists of around 20,000 tweets, and will constantly increase with future extractions. The tweet volume at the beginning of the observation period is lower that the observed towards recent periods, as Twitter started gaining popularity.

[1] We take into account 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.

3.2 Data Preprocessing

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

1. we remove tweet specific elements like hyperlinks, retweets, user mentions, and elements such as stopwords, numbers and punctuation;
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 letters to lowercase.

The following example illustrates the mentioned transformations:

3.3 Data exploration

After preprocessing the tweets, we conduct an exploratory analysis to cluster the text by topic so we can better understand the data we obtained from the extractions. For this task, we use the Latent Dirichlet Allocation (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. In this application, it allows us to identify six topics that constantly appear in the results:[4]

1. Financial markets (top 15 words: 'ganancia', 'dólar', 'millón', 'aumentar', 'vender', 'bmv', 'multar', 'bono', 'cerrar', 'euro', 'anunciar', 'mayor', 'caer', 'bolsa', 'pérdida')
2. Macroeconomic expectations ('dar', 'mantener', 'señalar', 'crédito', 'alertar', 'economía', 'riesgo', 'banca', 'país', 'destacar', 'impulsar', 'calificación', 'crecimiento', 'pesar', 'decir')
3. Foreign exchange market ('dólar', 'grupo financiero', 'prever', 'comprar', 'vender', 'venta', 'centavo', 'país', 'afore', 'ver', 'tipo de cambio', 'cerrar', 'ventanilla', 'peso')
4. Business activity (operación', 'servicio', 'cliente', 'reportar', 'primero', 'grupo financiero', 'presentar', 'comprar', 'crédito', 'fallo', 'mejor', 'banca', 'sucursal', 'ofrecer', 'digital')
5. Illicit activities and penalties ('cliente', 'dinero', 'poner', 'investigar', 'presentar', 'contar', 'directivo', 'crédito', 'multar', 'opinión', 'pedir', 'acusar', 'oceanografía', 'tarjeta', 'fraude')
6. Financial results ('ganancia', 'previsión', 'reportar', 'centrar', 'primer trimestre', 'fondo', 'prever', 'presentar', 'comprar', 'tasar', 'ligar', 'crecimiento', 'anunciar', 'caer', 'elevar')

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

[4] The translation in English of the original terms in Spanish is the following. Financial markets: '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'. 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'. 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'. Business activity: 'operation', 'service', 'client', 'to report', 'first', 'financial group', 'to present', 'to buy', 'credit', 'verdict', 'better', 'bank', 'branch', 'to offer', 'digital'. 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' . 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' .

This analysis is useful to determine if the collected data is suitable for the task we have at hand, and to uncover the main topics driving the Twitter Sentiment Index. 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 Twitter Sentiment Index has some overlap with the IEMF, but also captures new information, that quantitative financial indicators do not explicitly show. The common sectors that the two indices cover are financial markets and foreign exchange market. We interpret the topic "Macroeconomic expectations" as an indicator of country risk. Topics 4 and 6 (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 being captured by other indicators. We detect sentiment about customer services, digital services, and online payment systems, including bugs. Additionally, our index captures new information within topic 5, "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.

This evidence suggests that a Mexican Banking Risk Index built on sentiment analysis of tweets may complement existing indicators for detecting financial stress.

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 subsamples of 100 tweets to 37 professionals that label them according to the message they transmit regarding the level of risk in the Mexican financial system and/or the Mexican banks.

The "risk" we want to measure with this Index is the banking risk from the point of view of regulatory institutions or the banks themselves. Most of the times 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 privileged the systemic risk consideration, so that we consider the tweet as reporting news that decrease the banking risk. The labeling criteria is the following:

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

- Lower risk (corresponding to positive sentiment): tweets 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).

These criteria were shared with the professionals who participate in the labeling process. Each tweet is classified by at least 2 professionals 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 classifier

We choose three different models to build the sentiment classifier for the tweets. The first one 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, an SVM Classifier (or other linear classifier) can be 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.

We split our labeled data 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. 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.

4.1 Dictionary with word polarities

Correa et al. (2017) built their financial stability dictionary using words from the Financial Stability Reports (FSRs) of 62 countries, plus the European Central Bank and the International Monetary Fund, 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 (Banxico) 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.[5] 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 construction of sentences in Spanish is different from that in 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 the paper highlights 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 a FSR.

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

To find potential keywords that are specific to the universe of Twitter news in Mexico, we use the 2000 previously labeled tweets. We apply the TF-IDF weighting scheme[6] to identify the most relevant terms for the Negative and Positive categories and we include this extended vocabulary in our original dictionary with the correspondent word polarities.

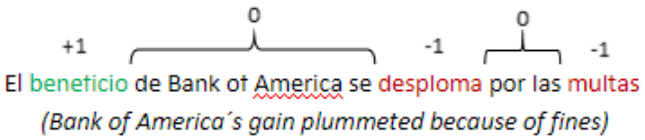
Table 4 presents an extract of the words in the original CKJM dictionary that appear more frequently in the English version of Banxico's FSRs, an extract of the more frequent Spanish words used in Banxico's FSRs and the most frequent negative words used in our sample of tweets. Most of the words used in the English and Spanish versions of the FSRs are the same, 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.

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 correspondent 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.



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 the previous labeling of data. 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 test the performance of this method on the whole sample of labeled tweets since a training step is not required here. Results are discussed in section 4.4.

[6] TF-IDF is a commonly used tool in Natural Language Processing (NLP). 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)

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 the majority of the literature focuses on social media analysis in English, this approach can be used to create a first approximation to a sentiment classifier on any given language.

The main contribution of Tellez et al. (2017) 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. B4MSA applies text-transformations to documents in a corpus and then creates a vector representation of the corpus using the TF-IDF weighting scheme, which serves as input for any linear 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.[7]

The preprocessing step for this model is done by the model itself. We apply the default preprocessing for Spanish language, which is similar to that applied for the LDA model, but with the specific difference that it includes n-grams in the vectorized data. As an additional step to the default preprocessing, we anonymize bank names.

After the text preprocessing and vectorization, we train the Linear SVM Classifier with 90% of our labeled tweets and test it on the remaining 10%. Results are discussed in section 4.4.

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.

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[8] which was trained on Wikipedia articles in Spanish.

Stage two consists in fine-tuning 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

[7] We tried also with a non-linear kernel, but we obtained better results with the linear one. To reduce the high dimensionality of text data, the linear kernel is the more suitable option.

[8] 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/ulmfit-spanish/29715/24

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 datasets. In order to compare the models' performance, we compute the following metrics: accuracy, balanced accuracy, and f1 score. Table 5 shows the results.[9] Higher accuracy reflects better classification of the positive, negative and neutral tweets by the model. When looking at the results for the B4MSA and ULMFiT models, we find the gap on training and test sets accuracy to be reasonable (11-12 points). 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.

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 on both the training and test sets. Its accuracy is in fact 64 percent and 59 percent respectively, much lower than the SVM model and the neural Networks one. B4MSA and ULMFiT models have comparable accuracies, around 75 percent for the test set and 85 percent for the training one, even if B4MSA is slightly more accurate.

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 the models. Again, the B4MSA and ULMFiT results are really close, and considerably outperform the Dictionary results. Since the Dictionary method is based on a previously defined set of words and simple word counts, it is not surprising to see these results. Despite this, it is useful to keep this method a baseline for its simplicity to implement.

4.5 Sentiment by voting

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. This classifier uses a "majority of votes" approach to determine the final sentiment. Since there are three classifiers, at least two must be in agreement for a tweet to receive a polarity. Whenever there is no agreement, the tweet is categorized as neutral. Table 6 shows an example for each case.

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 tweets and negative tweets, and we scale the index by the number of positive and negative tweets:

$$Sentiment\ Index_t = \frac{negative\ tweets_t - positive\ tweets_t}{negative\ tweets_t + positive\ tweets_t} \tag{2}$$

[9] Not all performance metrics can be applied to all models.

With t indicating the time span of interest (a day, week, month or year). Higher values of the sentiment index indicate higher risk in the banking 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, taking into account 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 soccer teams, for instance). Other than the polarity of the tweets, another possible source of information available from our extraction is the importance of the tweet for the Twitter users. We can assume that a tweet that receives more reactions (more retweets, or more likes) contains news that are more important to the public. Considering these two points, the inclusion of neutral tweets in the index, and the importance of each tweet to the Twitter users, we build other three versions of the baseline sentiment index.

The first does includes in the denominator the neutral tweets:

$$Sentiment\ Index_t^2 = \frac{negative\ tweets_t - positive\ tweets_t}{negative\ tweets_t + positive\ tweets_t + neutral\ tweets_t} \tag{3}$$

The second variation weights each tweet by the number of reactions (both retweets and likes) received:

$$Sentiment\ Index_t^3 = \frac{rn * negative\ tweets_t - rp * positive\ tweets_t}{negative\ tweets_t + positive\ tweets_t} \tag{4}$$

Where rn is the number of reactions to negative tweets and rp is the number of reactions to positive tweets. The third variation weights each tweet by the number of reactions and includes the neutral tweets from the denominator:

$$Sentiment\ Index_t^4 = \frac{rn * negative\ tweets_t - rp * positive\ tweets_t}{negative\ tweets_t + positive\ tweets_t + neutral\ tweets_t} \tag{5}$$

Table 7 presents how correlate is the sentiment index when computed with the four different estimators, and in the four different versions. 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, but increases when we weight the tweets by the number of reactions. As a comparison, Shapiro et al. (2019) find a correlation of 34 percent between the different model used to build their Sentiment Indices.

5.1 Visualization

In order to visualize the results, we build an interactive dashboard using Dash, a productive 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 Banking Sentiment Index along the period of analysis, and a word cloud with the most popular terms during the selected period. This may help understand abnormal changes in the Sentiment Index. Figure 1 shows a screenshot of the dashboard, displaying the wordclouds for January 2019, when Fitch downgraded Pemex Issuer Default Ratings. The risk increase due to this event is caught by the index and the wordclouds highlight as negative words "Pemex", "calificación" and "Fitch". The bigger is a word in the wordcloud, the more important it is in its respective category.

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 SVM consistently signals higher risk that the others . The Neural network Sentiment Index broadly follows the SVM Sentiment Index, except on a period from mid-2015 to mid-2016. The Sentiment index based on voting as expected stands in between the original three indices.

Figure 3 shows the Sentiment Index by voting with monthly frequency. Figure 3 (a) presents the main index,

where we do not consider neutral Tweets in the denominator, and Figure 3(b) presents the index computed considering neutral Tweets. The peaks of the two indices are very similar. An increase in the Sentiment Index corresponds to an increase in risk. We labeled each peak of risk considering 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.

In the first part of the sample, from January 2011 until December 2015, most of the news that increase the risk of our sentiment indicator correspond to events that increase the 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 billions of 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 Oceanografia one that again hit directly Mexico and its financial system. The oil services company Oceanografia was accused of a fraud that also involved the Mexican subsidiary of Citi bank, Banamex. The loan scandal costed more than $500 million to Citigroup.

The second period, from 2016 to June 2019, is characterized by shocks linked to macroeconomic, political and systemic shocks, such as the US elections in November 2016, Mexican elections, the earthquake that hit the country in September 2017, volatility on financial markets and domestic economic slowdown due to uncertainty in November 2018 and June 2019 respectively.

5.2 A cumulative 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 the weekly frequency, it is quite noisy, as depicted in Figure 2. Ideally, we would like to have a smoother cumulative sentiment index, where we consider previous shocks that sum up to the cumulated risk. We can consider the baseline Twitter Sentiment Index as noisy observations of the actual unobserved sentiment.

We take inspiration from Borovkova et al. (2017) and we filter the series to extract a meaningful signal from the data. We apply the Christiano-Fitzgerald band-pass filter (Christiano and Fitzgerald, 2003), that is indicated to smooth high frequency data (such as daily, weekly or monthly).

The Christiano-Fitzgerald filter assumes that the data are generated by a random walk, and even though this assumption is generally false for the most part of time series in economics, they find that it is nearly optimal. The Christiano-Fitzgerald filter dominates in term of an optimality criterion both the HP filter (Hodrick and Prescott, 1997) and the Baxter-King filter (Baxter and King, 1999).

Our goal is to filter the sentiment index series from the high frequencies, to eliminate the excessive noise. We want to filter exclusively the high frequencies, enlarging the band up to 100 years. The Christiano-Fitzgerald filter becomes a sort of low-pass filter.[10] We compute three versions of the business-cycle index with the lower bound fixed at 1 year, 6 months and 3 months.

6 Descriptive results

Following Shapiro et al. (2019), we test the goodness of the Twitter Sentiment Index comparing it with our reference measure of financial risk computed for Mexico: Banxico's Financial Market Stress Index. The IEMF is computed

[10] 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.

weekly by Banxico, and it includes 33 financial indicators. The variables were selected according to their importance in the Mexican financial market so that they show a volatile behavior during periods of financial stress. Therefore, 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 determine financial market stress. On the other hand, we use "soft", qualitative data (news and opinions reported in social media), and we apply algorithms that interpret the sentiment of this information.

Our hypothesis is that the Sentiment Index that we find would be correlated with the reaction of financial markets, reflected in the IEMF.

As shown by the topic analysis and suggested by the peaks of sentiment in Figure 3, the risk measured by the Twitter Sentiment Index is due to different kind of shocks to the financial sector: financial, macroeconomic, political and reputational. The reputational risk, in particular, is not explicitly measured by the IEMF, even though reputational shocks for the banking sector should be reflected by stock market prices.

We build two sub-indices of the general Twitter Sentiment Index, this time dividing the sample of tweets in those that are classified as bringing reputational risk by the LDA algorithm and all the others. We follow the same methodology that we use for the general sentiment index.

Figure 4 shows the general Sentiment Index, the Reputational Index and the Non-reputational one compared with the IEMF over the period 2011-2019. We present the results of the smoothed indices using the Christiano-Fitzgerald filter with the band starting at the 1-year frequencies (panel (a)) and at the 6-months frequencies (panel (b)).[11] It is possible to distinguish the 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 follows more closely the non-reputational one, and their trend is more in line with the IEMF than the Reputational Index.

We compute the correlation of the IEMF with the Non-reputational Index and the general Sentiment Index, to test if the Non-reputational Index may be closer to the IEMF than the general index, which comprises also all the reputational risk that the IEMF does not detect. Column 1 of Table 8 shows the correlation between the IEMF and the baseline Sentiment Index computed on the complete sample of tweets. Column 2 shows the coefficients of the correlation between the IEMF and the Non-reputational Index. In all cases the Non-reputational Index is more correlated with the IEMF than the general Index. The dictionary model show a higher correlation when we consider the whole time period starting from 2011 to 2019, being correlated with the IEMF for the 11 percent (complete Index) and 15 percent (non reputational index). Column 1 show that as soon as more data are available, the SVM and Neural network model increase their correlation from 7 percent and 4 percent to 9 percent and 12 percent, with a higher increase in correlation for the neural network model. Column 2 shows correlations comparable of higher for all models. in particular, the sentiment Index by voting is the one that shows the higher correlation, reaching more than 17 percent in the case of the general index and more than 19 percent for the non reputational index.

As a robustness check, we perform the same correlations using the alternative models of the Sentiment Index (Table 9). However, the correlation between the IEMF and these alternative variations is lower than those presented for the baseline Index both in the case of the general Sentiment Index and the Non-reputational one. For this reason, we will consider only the Sentiment Index built by voting in the baseline version for the rest of our analysis.

The evidence presented in Figure 4 and Table 8 suggests that our intuition is correct. The Non Reputational Sentiment Index is based on the same information that is relevant to measure financial market stress risk. It can be considered an alternative indicator of systemic risk that can complement quantitative indicators of financial risk.

[11] The third version of the index, with the band starting at 3-month frequencies, is still relatively noisy. For brevity the results have been omitted.

We also compute correlations between the IEMF and the filtered indices, in the three versions: the general Twitter sentiment Index, the reputational index and the non-reputational one. Table 10 details the results. Column (1) shows the baseline case, where the sentiment index is not filtered; columns (2) to (4) show the filtered index using different lower bounds: 1 year, 6 months and 3 months respectively. In all cases the non-reputational index is positively correlated with the IEMF, and the correlation is significantly different from zero. Due to the increase in the volume of Tweets, the correlations became stronger after 2015 than in the sample comprising data from 2011 onward. The general Sentiment Index is positively correlated with the IEMF, even if the correlation is lower in absolute value. This is due to the reputational component that results not correlated, or even negatively correlated with the IEMF index. If we look at Figure 4, we see that the peaks of the filtered weekly Reputational index coincide with the expected peaks in the general Index (end of 2012, May 2014). These peaks are not reflected in the IEMF. However, the Non-reputational Index seems to follow more closely the IEMF index. This result suggests that we may divide our general Index in two parts: the Non-reputational index may be seen as an alternative to the IEMF as systemic financial stress index, built using alternative data and methodologies. The Reputational Index may be considered, instead, a separate indicator, that signals peaks of reputational risk for Mexican banks and the Mexican financial system.

7 Predictive accuracy

We take inspiration from the work by Shapiro et al. (2019) to test if the Twitter Sentiment Index contains predictive information on some of the variables that compose the IEMF Index. We refer in particular to 10 out of the 33 indicators used in building the IEMF:

1. The spread between the 3-month Mexican CETES (Certificado de la Tesorería de la Federación) yield and the 3-month US Treasury bill and the spread between the 10-year fixed rate Mexican Treasury bonds and the 10-year US Treasury bill as indicators of bond market risk;
2. The IPC volatility as indicator of stock market risk;
3. The annual growth of the FIX rate and the 1-month Fix rate volatility, as indicators of foreign exchange market risk;
4. The basis points in the peso-dollar foreign exchange rate swap to buy dollars and the spread between the 5-year swap rate and the 5-year fixed rate domestic sovereign bond as indicators of derivatives market risk;
5. The beta of financial institutions, as indicator of credit institutions risk;
6. The EMBI + computed for corporate risk and the EMBI +, computed for sovereign risk, as indicators of country risk.

Table 11 presents the correlations between the selected variables and the three versions of the Sentiment Indices, being in line with our previous results. We focus on the short sample, from 2015 to 2019, to include the more precise estimates due to the higher amount of information we can extract. The complete Index and the Non-reputational one are correlated with the expected sign with most of the variables considered. Interestingly, the correlation of the complete Sentiment Index with selected variables seems higher, or more significant, than in the case of the correlation of the same index with the IEMF. We explain this with the fact that both reputational and non-reputationial risk may influence single financial indicators, and that this effect may be attenuated in the construction of the IEMF.

To compare the forecasting power of a model that includes one of our Twitter sentiment indices and a model that does not consider them, we report the Akaike Information Criteria (AIC), with measures the fit of a model.

The lower the AIC is, the more forecasting ability the model has. We compute the AIC for the ten selected financial variables.

Figure 5 reports the AIC for our cumulated Twitter Sentiment Index, compared with the baseline model of a VAR where only each variable of interest is included as endogenous variable. In all cases, the cumulated Sentiment Index, shows a higher AIC than the model that does not consider the Twitter Sentiment Index.

Finally, we use the local projections method by Jordà (2005) to analyze the impact of a one standard deviation shock of the cumulated Twitter sentiment Index, on each of the variables of interest. Figure 6 shows the resulting impulse response functions.

The most affected indicators are the volatility of the FIX exchange rate and the volatility of the IPC index that are respectively related to the foreign exchange market risk and the stock market risk. Also the EMBI+ corporate for Mexico, a proxy for country risk, expressed from the point of view of the private sector, is affected by a shock of the Sentiment Index. It rises and stay positively significant for the first 13 weeks. The indicators of bond market risk are not significantly affected by a shock in the Twitter Sentiment Index. Surprisingly, also the indicator of risk for credit institutions, the beta, has a positive reaction to the shock but it is not significant. We explain this with the construction of the beta variable for the IEMF. The variable is based on a sub-index of financial institutions constructed by Morgan Stanley (the MSCI Mexico Financial Index) compared with a market index, also proposed by Morgan Stanley (the MXMX Index). The beta built in this way reflects market actions of big financial groups such as Banorte and Inbursa. However, it may be that big banking groups are less affected by sentiment, since they are "too big to fail", while the Sentiment Index may hit more smaller banks. This is a possibility that we will explore in future work.

8 Conclusion

We propose a Twitter Sentiment Index for Mexico based on sentiment analysis of tweets. We use three different NLP techniques to analyze the sentiment of Twitter messages and we build alternative Sentiment Index indicators (cumulative and non-cumulative).

We contribute to the literature that applies data-driven modeling techniques to the construction of risk indicators in several ways. First, building on the work by Correa et al. (2017), we propose a financial dictionary in Spanish, specifically adapted to text analysis in social media.

Second, we apply state-of-the-art NLP techniques to build alternative versions of the sentiment-based risk indicator.

Third, we use topic modeling techniques (in particular the Latent Dirichlet Allocation) to explore the topics of the tweets that have an impact on our Sentiment Index. The topic analysis shows that our index captures sources of potential financial risk that are not traditionally included in financial stress indicators, such as financial frauds, money laundering, and fails in online payment systems. This paper show that the Sentiment Index can complement indicators of financial risk driven mainly by traditional quantitative indicators of financial risk.

Finally, we assess how well the Sentiment Index correlates with existing measures of financial market risk and selected financial variables. We test the effect of our index on selected financial variables and we find that a shock in the Twitter Sentiment Index increases stock market volatility and foreign exchange rate volatility, having a significant effect on overall financial market risk, especially for the private sector.

A Tables

Type of source	Name	Type of source	Name
Mexican	El Financiero	Foreign	El País
newspapers	El Economista	newspapers	El País (edition Americas)
	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 complete extraction and the frequency of the same words in the extraction from selected accounts

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 estraction from selected accounts and the frequency of the same words in the complete extraction

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 lasituación de Pemex.	eleconomista	447506	Mexico

Table 3: Selected Tweets from our database.

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 4: CKJM dictionary modified

Model	(1) Dictionary	(2) B4MSA-SVM	(3) ULMFiT
Cross validation mean accuracy	NA	0.72	NA
Cross validation sd	NA	0.05	NA
Test set acc.	0.59	0.75	0.73
Training set acc.	0.64	0.86	0.85
Balanced acc.	0.56	0.71	0.71
F1 score	0.58	0.73	0.73
Accuracy per category			
Increase risk	0.44	0.74	0.78
Neutral	0.78	0.78	0.80
Decrease risk	0.47	0.62	0.55

Table 5: Models' performance results

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 6: Sentiment by voting

	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
Model 4				
SI dictionary	1			
SI SVM	0.6585*	1		
SI neural networks	0.5576*	0.6179*	1	
SI voted	0.7382*	0.8288*	0.7853*	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; (4): SI computed considering neutral tweets and weighting the index by the number of reactions.

Table 7: Correlation between alternative Sentiment Indices, 2011-2019

	2011-2019	
	(1)	(2)
SI dictionary	0.1113*	0.1508*
SI SVM	0.0719	0.1198*
SI neural networks	0.0432	0.0713
SI voted	0.0824*	0.1101*
	2015-2019	
SI dictionary	0.1554*	0.1591*
SI SVM	0.0988	0.1242*
SI neural networks	0.1227*	0.1154*
SI voted	0.1717*	0.1907*

Note: *: p-value<0.1; SI computed not considering neutral tweets. Column (1): complete sample; Column (2): only non-reputational tweets.

Table 8: Correlation between Sentiment indices and IEMF

Sample: All tweets		2011-2019	
	(1)	(2)	(3)
SI dictionary	0.0981*	0.0287	0.0310
SI SVM	0.0164	0.0398	0.0181
SI neural networks	-0.0364	0.0218	0.0681
SI voted	0.0249	0.0590	0.0684
		2015-2019	
SI dictionary	0.0135	0.1364*	0.1232*
SI SVM	-0.0436	0.0865	0.0400
SI neural networks	-0.00960	0.0108	0.0565
SI voted	0.0607	0.109	0.1315*
Sample: Not Reputational		2011-2019	
SI dictionary	0.1515	0.0147	0.0138
SI SVM	0.0912	0.0590	0.0317
SI neural networks	0.0342	-0.00470	0.0703
SI voted	0.1029	0.0523	0.0821
		2015-2019	
SI dictionary	0.1459	0.0983	0.104
SI SVM	0.103	0.0730	0.0290
SI neural networks	0.1407	-0.0349	0.0442
SI voted	0.2035	0.0873	0.122

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

Table 9: Correlation between Sentiment Indices and IEMF

	(1)	(2)	(3)	(4)
IEMF				
2011-2019				
SI, Not reputational	0.1101*	0.0368	0.1306*	0.1194*
SI, Reputational	-0.0500	-0.3382*	-0.2569*	-0.2202*
SI, All sample	0.0824*	-0.1754*	0.00980	0.0442
2015-2019				
SI, Not reputational	0.1907*	0.2421*	0.3673*	0.2606*
SI, Reputational	-0.0309	-0.2802*	-0.1213*	-0.101
SI, All sample	0.1717*	0.1221*	0.2675*	0.1943*

Note: *: p-value<0.1; (1): baseline SI (not filtered); (2): SI filtered, 1 year-100 years; (3) SI filtered: 6 months-100 years; (4) SI filtered. 3months-100 years.

Table 10: Correlations between the voter Twitter Sentiment Indices and IEMF

	(1)	(2)	(3)
Sentiment Index	Not reputational	Reputational	All tweets
Spread 3m sovereign bonds	-0.1991*	0.5516*	0.0444
Spread 10y sovereign bonds	0.3684*	-0.5531*	0.0668
IPC volatility	-0.0897	0.1359*	-0.1836*
Annual FIX growth	-0.1446*	0.5249*	0.0690
FIX volatility	0.0884	-0.2064*	0.1060
bps in peso-dollar FX rate swap to buy dollars	-0.1246*	0.5770*	0.1578*
Spread between 5y swap rate and 5y fixed rate sovereign bond	0.0546	0.5119*	-0.0063
Beta of financial institutions	0.3878*	-0.6241*	0.1574*
EMBI + corporate	-0.0217	0.4628*	0.1265*
EMBI + sovereign	0.0426	0.2128*	0.0307

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

Table 11: Correlations between the voter Twitter Sentiment Index and selected market variables, 2015-2019

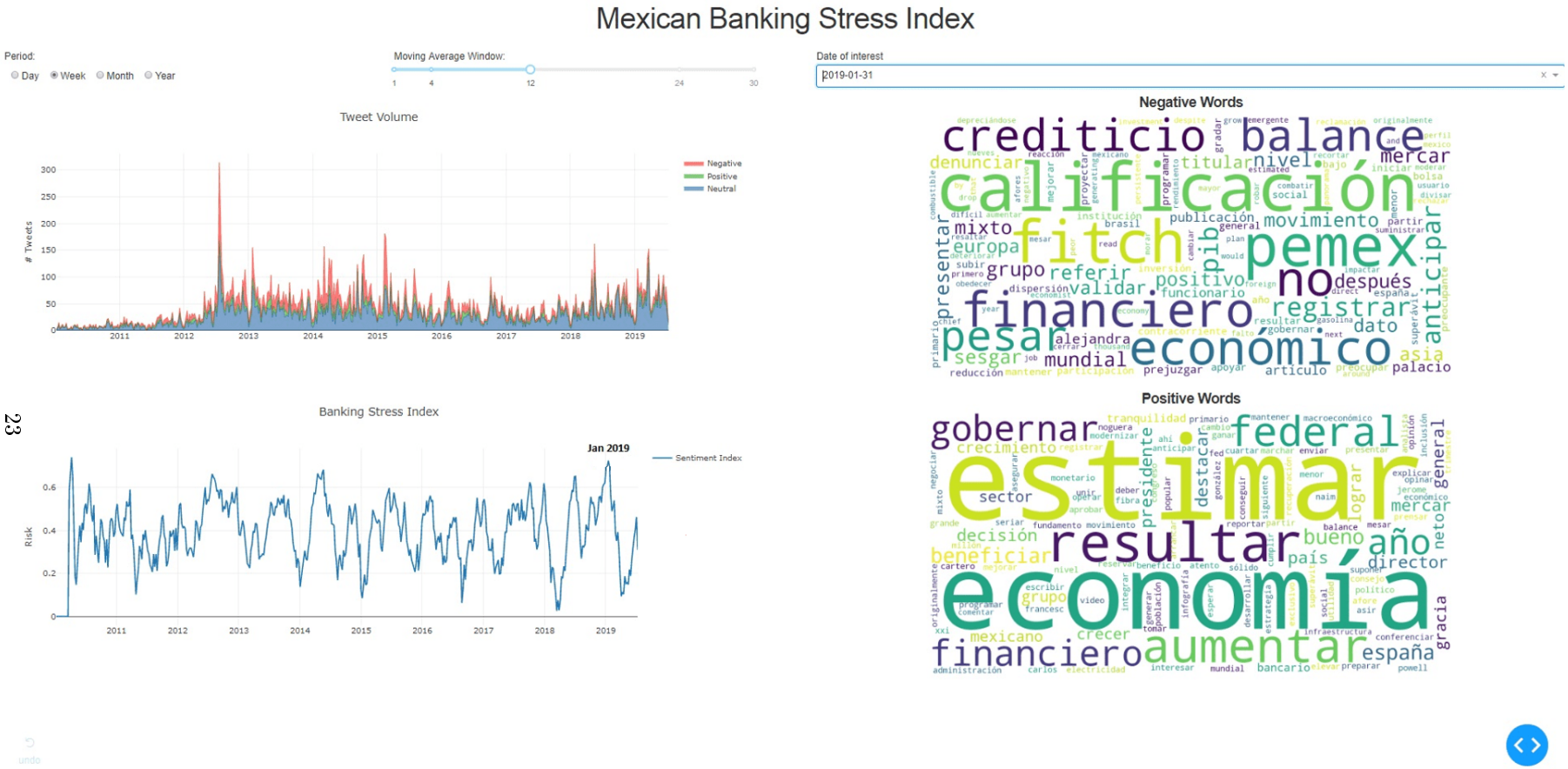
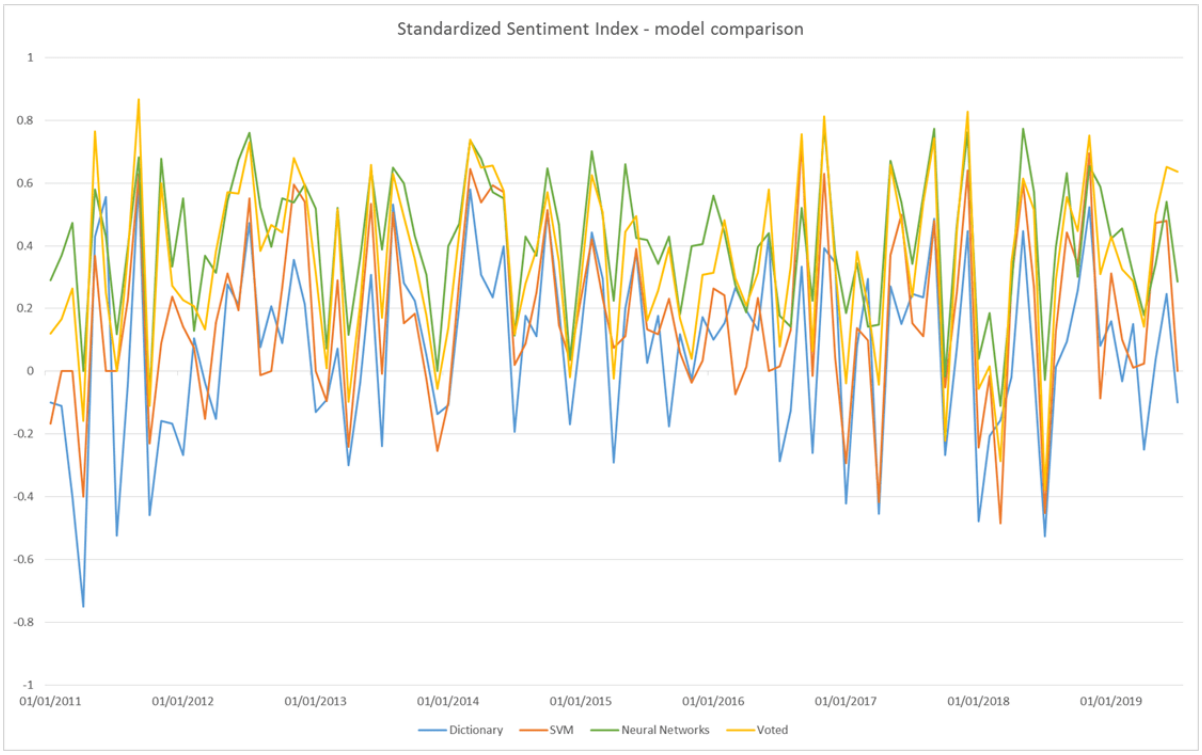
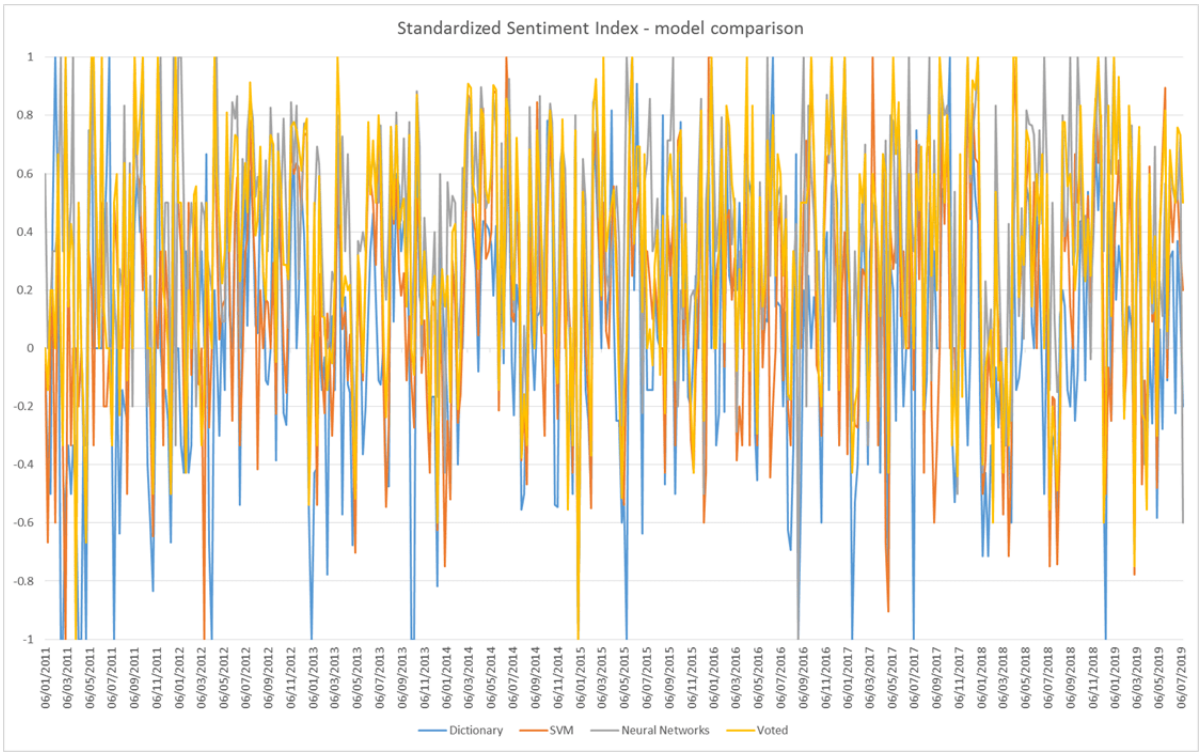


Figure 1: Banking Risk Index and positive and negative trending words (January 2019)

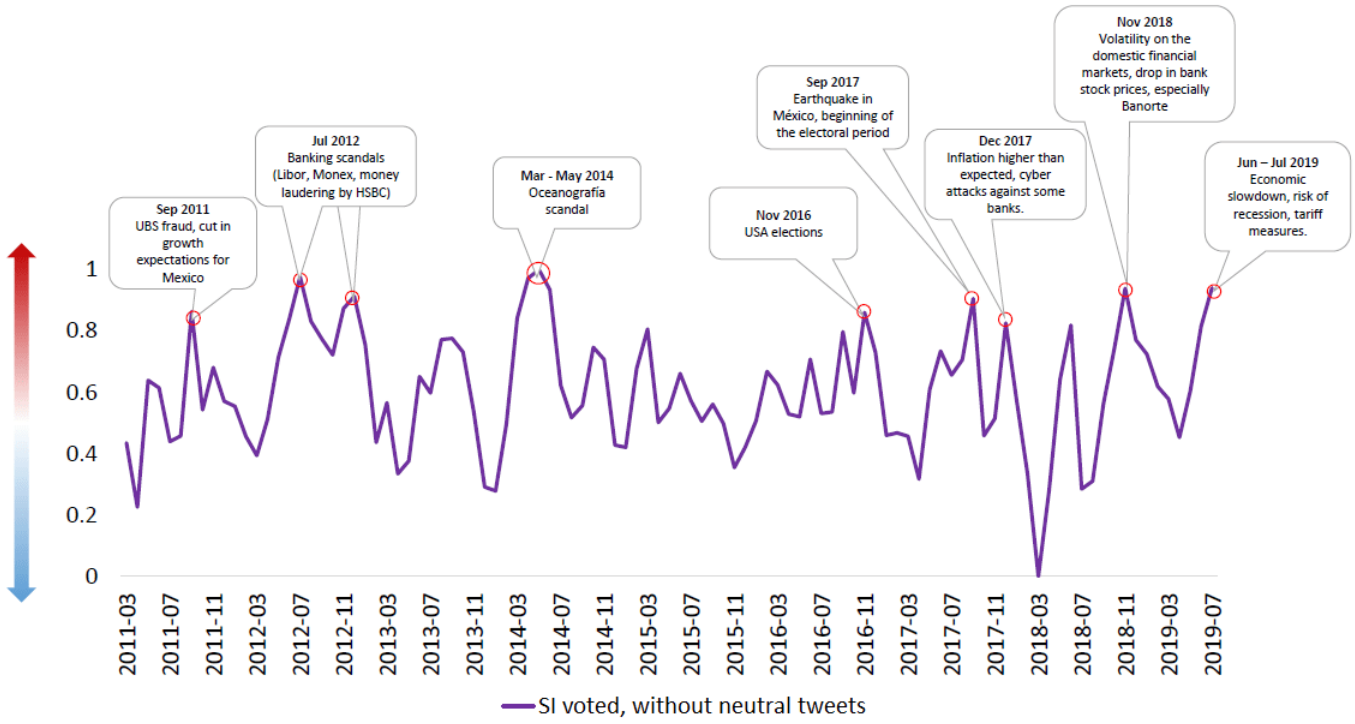


(a) Monthly frequency

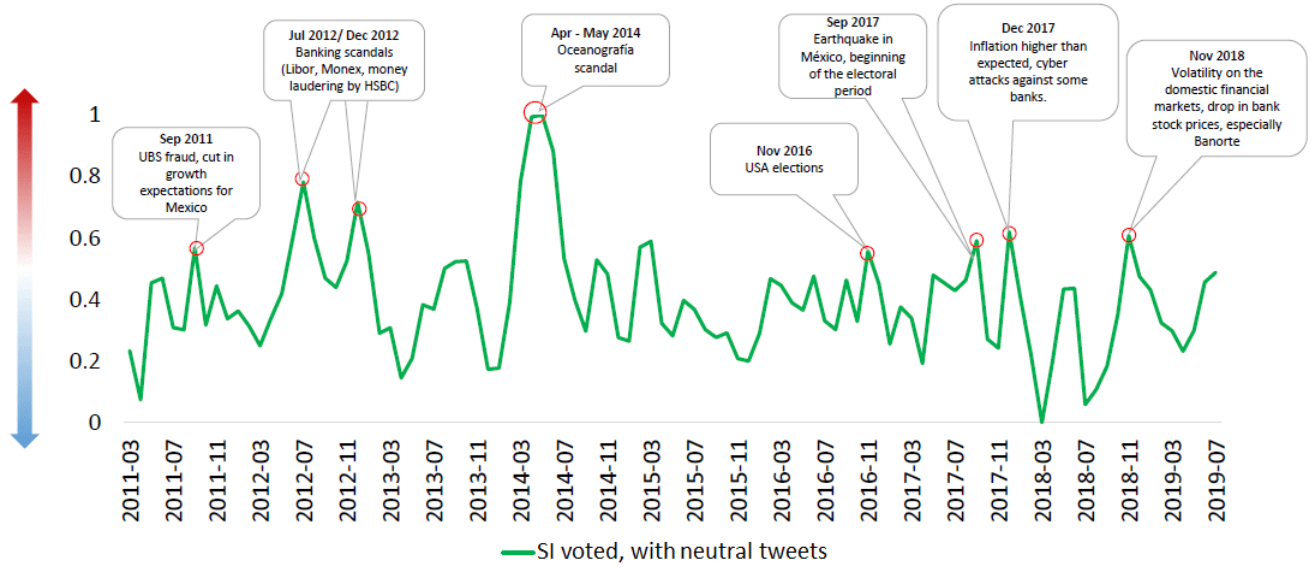


(b) Weekly frequency

Figure 2: Comparison between the four sentiment indices

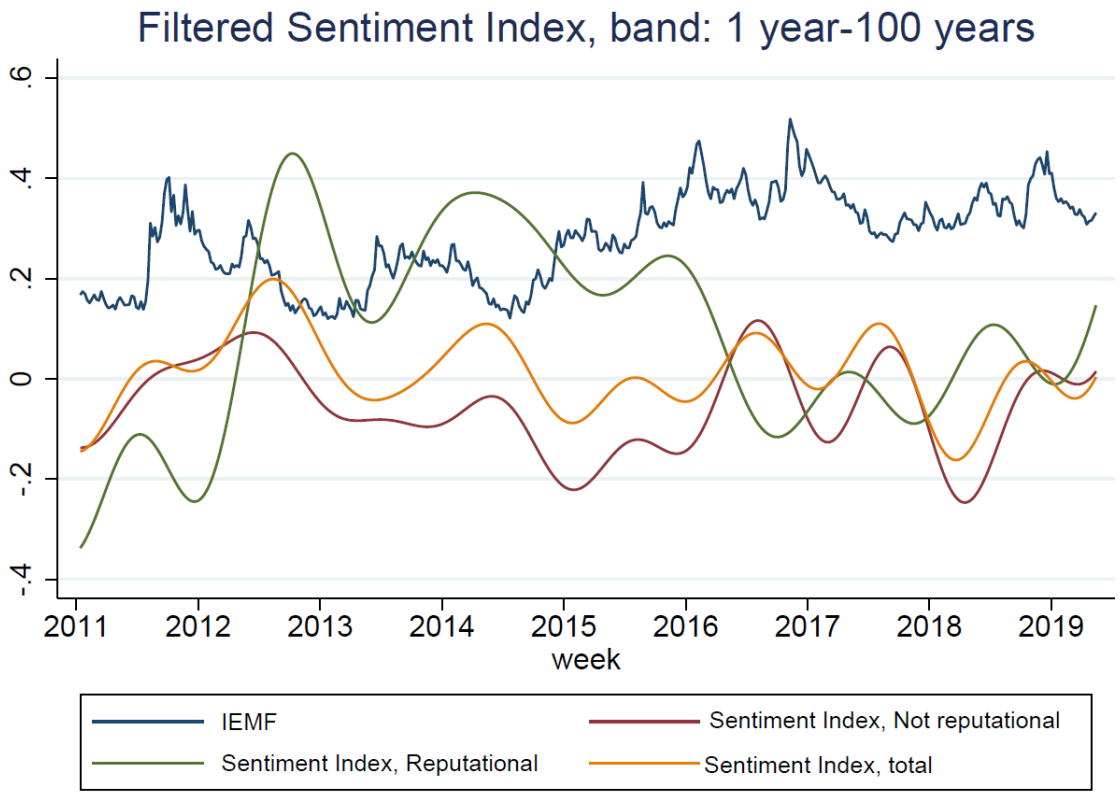


(a) Sentiment Index, computed without neutral tweets

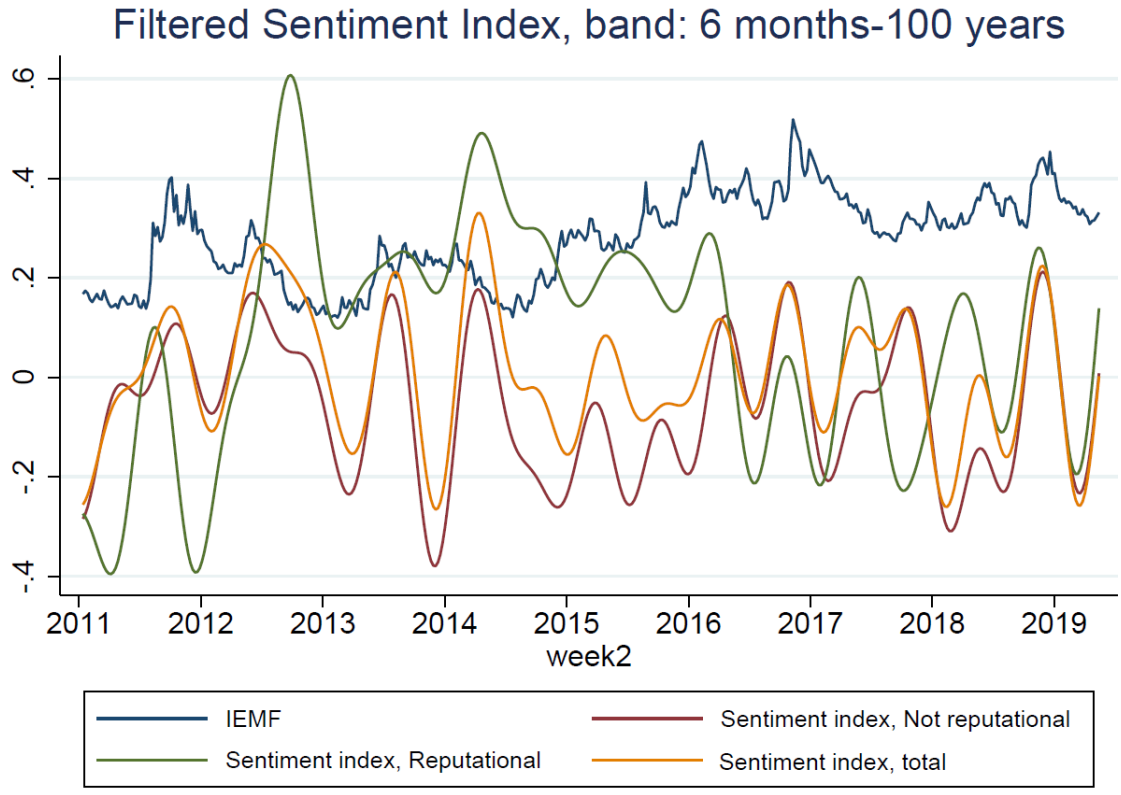


(b) Sentiment index, computed with neutral tweets

Figure 3: Sentiment index (voted), monthly frequency



(a) Sentiment Index, CF filter: 1year-100 years



(b) Sentiment Index, CF filter: 6 months-100 years

Figure 4: Filtered Sentiment Index

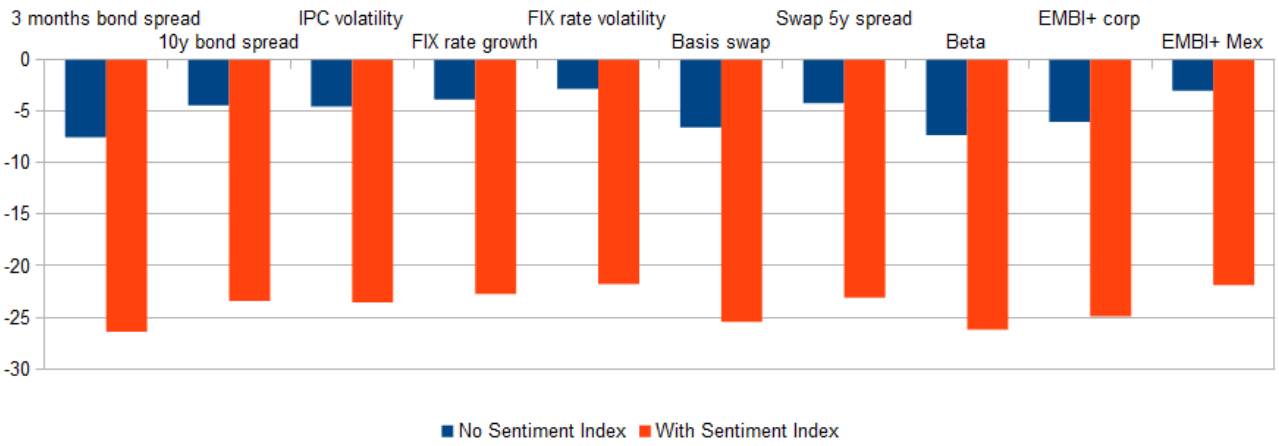


Figure 5: Information criteria for selected variables

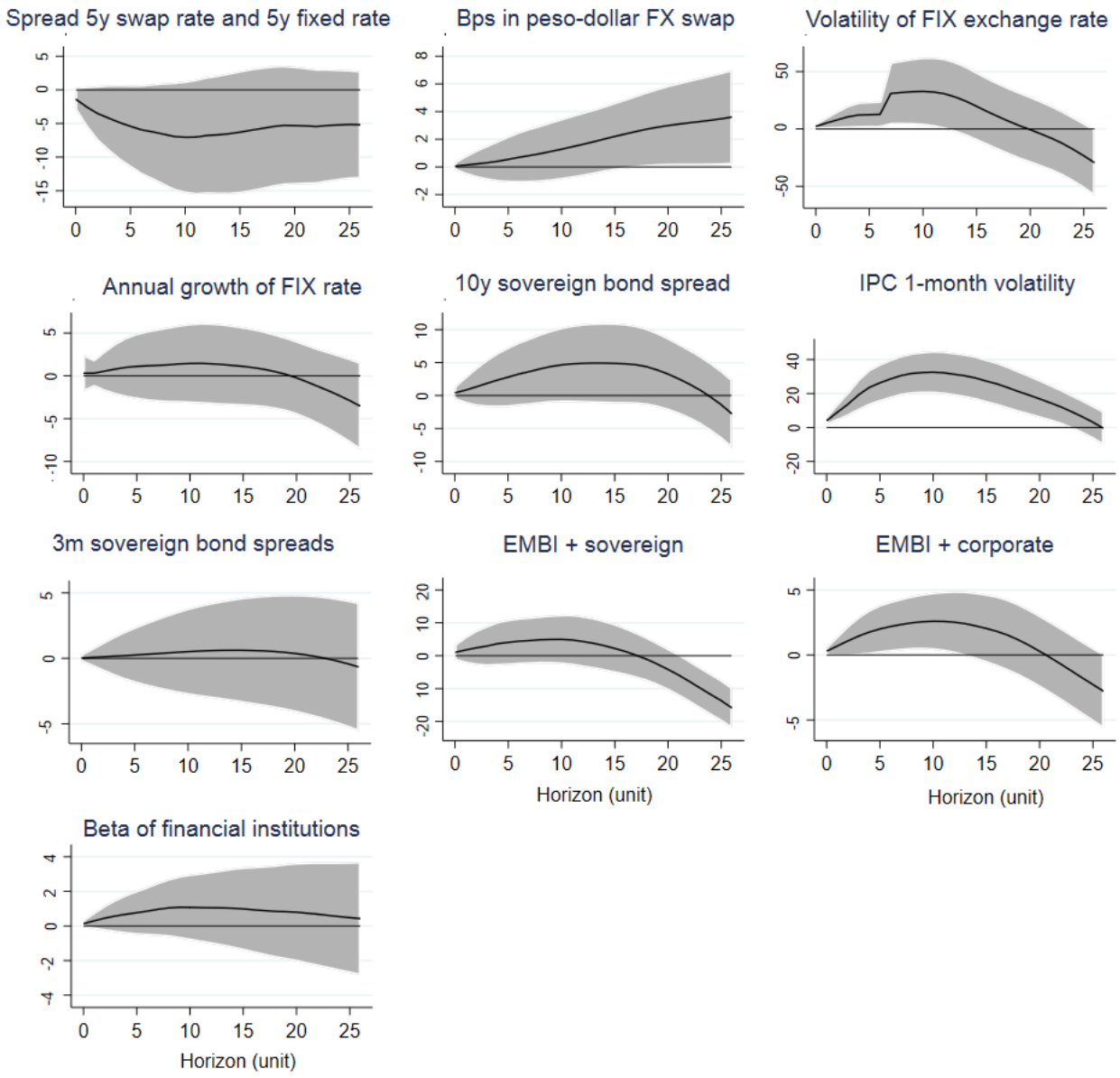


Figure 6: IRFs for selected financial variables. Impulse variable: Sentiment Index, filtered using the 1y-100y band

References

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

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.

Banco de Mexico (2019). Financial stability report.

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.

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.

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.

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.

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.

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

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

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

Shapiro, A. h., M. Sudhof, and D. Wilson (2019, June). Measuring news sentiment. Working Paper 2'217-01, Federal Reserve Bank of San Francisco.

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.