Identifying Financial Crises Using Machine Learning on Textual Data[*]

Mary Chen[†], Matthew Deininger[‡], Seung Jung Lee[§] and Martin J. Sicilian[¶]

October 1, 2020
FIRST DRAFT

Abstract

We use a variety of machine learning techniques on multiple sources of textual data to identify and predict financial crises. One of the challenges in financial crisis management is being able to determine whether a country is in a crisis or not, or what type of crisis a country is in, especially in real time. The onset of a crisis and its severity also has implications for real economic activity and, hence can be a valuable input into macroprudential, monetary, and fiscal policy. The academic literature and the policy realm rely mostly on expert judgment to determine crises. Consequently, the identification of crises and the buildup-phases of vulnerabilities are usually determined with hindsight. Although we can identify and forecast a good portion of various degrees of crises with traditional econometric techniques using readily available market and flow of funds data, we find that textual data helps in reducing false-positives and false-negatives in out-of-sample testing of such models, especially when the crises are considered more severe.

JEL classification: C53; C55; G01.

Keywords: Financial Crises, Machine Learning, Natural Language Processing.

[*] The views stated herein are those of the authors and are not necessarily the views of the Federal Reserve Board or the Federal Reserve System. We are grateful to Arthur Turrell for guiding us through the initial phases of our research. We thank Robin Lumsdaine for a terrific discussion of the paper. We also thank participants at the IF Workshop, Finance Forum, the Textual Analysis in Economics and Finance Research Conference, and the Joint FRB-IMF Workshop on New Techniques and Data in Macro Finance at the Federal Reserve Board, the ECB/ESRB, the 3rd Forecasting at Central Banks Conference at the Bank of Canada, and the 3rd Conference on Financial Stability at the Bank of Mexico for valuable feedback on our analysis. All errors are the authors' and no one else's.

[†] Federal Reserve Board. E-mail: mary.y.chen@frb.gov.

[‡] Federal Reserve Board. E-mail: matthew.d.deininger@frb.gov.

[§] **Corresponding author**, Federal Reserve Board. E-mail: seung.j.lee@frb.gov.

[¶] University of Stanford Law School. E-mail: martinsicilian@gmail.com.

1 Introduction

We use a variety of machine learning techniques on textual data to identify and predict different financial crises. The academic literature and the policy realm rely mostly on expert judgment to determine financial crises. Consequently, the identification of crises and the buildup-phases of vulnerabilities are usually determined with hindsight. This implies that various financial and banking crises data are slow to update and, in many cases, updates occur after many years (Reinhart and Rogoff (2009), Laeven and Valencia (2013)) or not at all (Romer and Romer (2017)).

The identification of financial crises is important in real-time because many of the new macro-prudential policies in place after the Global Financial Crisis (GFC) are activated or deactivated conditional on the state of the financial sector. For example, frameworks for the countercyclical capital buffer (CCyB) are now in place in the majority of advanced economies so that the CCyBs can be activated when vulnerabilities in the financial system are high and released when material risks or a crisis is realized. Moreover, knowing whether a country is in a crisis or not, or what type of crisis a country is in is crucial to policy makers in their financial crisis management.

The onset of a crisis and its severity or type also has implications for real economic activity and, hence can be a valuable input into monetary or fiscal policy. Indicators that identify or predict financial crises are also important for the aforementioned policies to be able to distinguish which phase of the financial or credit cycle a particular financial system is in. For example, (re)activating the CCyB, hiking interest rates, or adopting fiscal austerity measures prematurely when a crisis is not even over may lead to weaker real activity than otherwise.

Thus far, academic papers have relied on developing a set of indicators to identify the severity of market disruptions through financial stress or financial conditions indexes (Brave and Butters (2012)), or help predict financial crises (Drehmann and Juselius (2014), Aikman et al. (2017), and Lee, Posenau and Stebunovs (2018)). However, these papers focus on the onset date of financial crises as the main dependent variable and ignore ongoing crisis states for significant periods of time. In fact, relevant quantitative indicators in these models behave in a way such that either indexes spike at the onset of crises (or at other near-crisis periods of market disruptions), or gradually increase and then peak around the onset of crises and fall sharply afterwards.

In our analysis, we try to build an indicator that signals whether a particular country is in a crisis. Due to the limitations of quantitative variables described above, we use a variety of machine learning techniques on the following textual data: reports from official international organizations (OECD and the IMF) and articles from the media (Thomson Reuters News Archives). These data sources help us develop an indicator that increases and stays elevated for the entirety of the crisis periods. In the context of macroprudential policy, an elevated reading would provide valuable inputs into the decisions regarding when and how long to release the CCyB buffer, for example. Although a decent portion of some types of crisis periods can be identified with traditional econometric techniques using readily available market data, in our analysis, we find that textual data significantly helps in reducing false-positives and false-negatives in out-of-sample testing of such models, especially when the crises are considered more severe.

Textual data could potentially aid in the identification of financial events in two ways: through its descriptive properties, or even through its instigative potential. In order for either explanation to be valid, researchers must be able to extract at least part of the information in the text in an automated fashion.

The descriptive explanation is that the authors of the text convey information about the financial landscape that is relevant to the chance that the events of interest will come to pass. The authors may, for example, use more positive words in good times and more negative words in bad times. An important element of this mechanism by which text can help identify financial events is that the circulation of the text is less irrelevant; it would make no difference if the text were read by millions or no one. This follows the literature that text has useful information as an important source of information for understanding economic phenomena (Gentzkow, Kelly and Taddy (2019)).

Text could also add predictive power in identifying financial events by influencing whether those events happen. This is the instigative, or causal, mechanism and goes much further than being simply descriptive and an additional source of information. For example, Shiller (2017) emphasizes that market participants and economic agents can driven by the narrative, which in turn can be driven by what is written in text. It is not necessary that the text contain any descriptive truth at the time of its publication, as long as the text changes the financial landscape in a predictable way. A classic example of this mechanism is bank runs. Regardless of whether a particular community bank has a cash shortage, an article stating as much published by an influential local paper can become a self-fulfilling prophecy. Note that the effect of the article would likely diminish with a smaller readership (Iyer, Puri and Ryan (2016)).

It is likely that most financially-relevant texts operate to identify financial events in part through both of these mechanisms. OECD Economic Outlook Reports (OECD) have a much lower readership than does the Thomson Reuters News Archive (TRNA), so one might expect TRNA to be relatively more instigative. On the other hand, readers of OECD Reports might have more influence over the financial landscape or policies which affect the landscape. Disentangling the mechanisms is difficult, especially with no reliable readership data. Still, it is possible to draw some conclusions about which mechanisms are more impactful for different text sources, as we will attempt in the results and discussion sections.

After starting with some literature in Section 2, we describe the data and model in Sections

3 and 4, respectively. Sections 5 and 6 show our results for nowcasting and forecasting financial crises. We conclude with a summary of results and a discussion on why different types of text can be useful in identifying and forecasting financial crises in Section 7.

2 Literature

Our research contributes to the literature on detecting financial crises and, more broadly, on using machine learning and textual data to nowcast and forecast aggregate macro and financial conditions.

Drehmann and Juselius (2014), Aikman et al. (2017), Lee, Posenau and Stebunovs (2018), and Cesa-Bianchi, Martin and Thwaites (2019) provide a framework to understand the financial vulnerabilities that lead to financial crises. Brave and Butters (2012) also provide a way to understand how financial conditions can forecast financial crises. The literature points to many different types of vulnerabilities increasing prior to financial crises, especially when they end up being systemic. Related to machine learning, Alessi and Detken (2018) and Bluwstein et al. (2019) provide analyses to show how machine learning can be useful in predicting financial crises. The context of these papers has been mainly on understanding what leads to financial crises as opposed to identifying crises.

Romer and Romer (2017), on the other hand, identifies different severities of financial distress for 24 OECD countries from 1967 to 2012, by reading through each country's OECD Economic Outlooks. This narrative determination of crises are linked to declines in output afterwards, with the variation in the declines driven by the severity and persistence of the financial distress itself. Our study incroporates the approach used in Romer and Romer (2017), but applies machine learning to the OECD economic outlook and other textual data based on the training sample that Romer and Romer (2017) constructed.

Using textual analysis to predict and identify various macro and financial aggregagtes have become more and more popular in recent years. For example, Angelico et al. (2019) uses twitter feeds to understand inflation expectations and Kalamara et al. (2019) uses U.K. newspaper articles to forecast a very wide set of aggregate macro-financial variables such as GDP growth, inflation, and financial vulnerability measures. This new literature points to a expanded set of information that can be used to help understand various macro-financial aggregates. In addition, studies such as Cerchiello et al. (2017) nowcast financial distress at the individual bank-level with textual data. As with our work, all of these papers provide further insight into the determination of macro-financial variables through text.

3 Data

3.1 Crisis Data

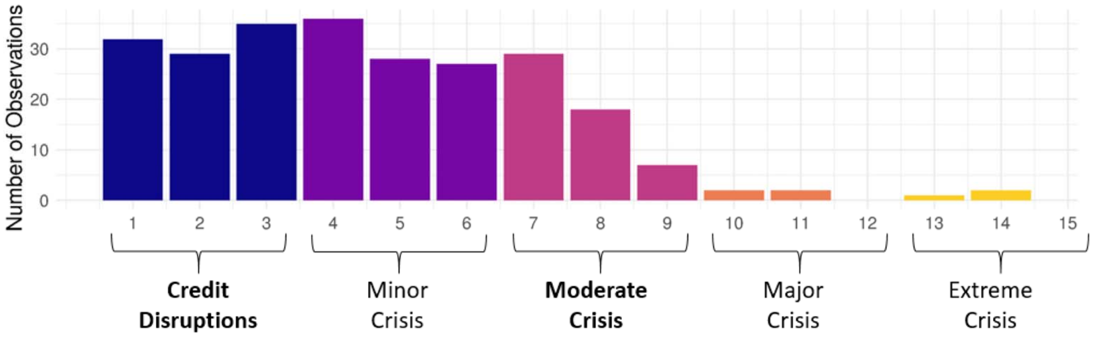
We use a variety of crisis data in our analysis, but rely on two main sources for crisis determination. First, Romer and Romer (2017) provides an ideal set of crisis definitions for our textual analysis. In fact, they define crises with a narrative approach, focused on the OECD Economic Outlook—a roughly 2,000 word quantitative, contemporaneous document—which has been published twice a year since 1967 for all OECD countries (they limit their dataset to only original members). Although mostly focused on the economic environment and real-side forecasts, these documents are read for signs of a rise in the cost of credit intermediation for individual countries. They develop 5 groups of financial distress: credit disruption, minor crisis, moderate crisis, major crisis, and extreme crisis. Each category is also given a minus, normal, or plus. Thus their final scale reanges from 0 to 15. This constitutes a more continuous-like measure of crises. Below are some broader categories of their crisis definitions:

- **Credit Disruptions**: The OECD perceived strains in financial markets, funding problems or other indicators, but did not believe these had any macroeconomic impacts.
- **Minor Crises**: A perception that there are significant problems in the financial sector, a belief that these problems are affecting the credit supply and/or overall economy in a nontrivial way, and not limited to a minor part of the economy; yet not an impact large enough to be damaging to the economy's prospects.
- **Moderate Crisis**: Widespread and severe problems in the financial sector, central to the economy as a whole, but not serious enough that it could be described as the financial system seizing up entirely. Romer and Romer (2017) note that this seems to be roughly the definition other researchers use for binary scales.
- **Major/Severe Crises**: Romer and Romer (2017) looks for the terms "crisis", "dire", "grave", "unsound", "paralysis" in reference to the financial system to classify this level. References to major government interventions contribute to a severe rating.

In our analysis we focus on whether crises are at least minor crises. If we restrict our analysis to major/severe crises, we end up with too few observations to conduct meaningful analysis. Figure 1 describes the various degrees of crises in Romer and Romer (2017).

Our second crisis data is from Laeven and Valencia (2013). We focus on the banking crisis because this provides a higher (monthly) frequency of crisis dates, which are available for a broader

Figure 1: **Romer and Romer (2017) Crises**



set of countries. A banking crisis is defined as an event that meets two conditions: (1) Significant signs of financial distress in the banking system, and (2) Significant banking policy intervention measures in response to significant losses in the banking system. The first year that both criteria are met is the year when the crisis becomes systemic. When the losses in the banking sector and/or liquidations are severe, the first criterion is a sufficient condition. Losses are severe when either (i) a country's banking system exhibits significant losses resulting in a share of nonperforming loans above 20 percent of total loans or bank closures of at least 20 percent of banking system assets or (ii) fiscal restructuring costs of the banking sector are sufficiently high, exceeding 5 percent of GDP. Policy interventions in the banking sector are decided to be significant if at least three of the following six measures have been used: (1) deposit freezes and/or bank holidays, (2) significant bank nationalizations, (3) bank restructuring fiscal costs (at least 3 percent of GDP), (4) extensive liquidity support (at least 5 percent of deposits and liabilities to nonresidents), (5) significant guarantees put in place, and (6) significant asset purchases (at least 5 percent of GDP). Laeven and Valencia (2013) crises are generally considered equivalent to moderate crises (or more severe crises) in Romer and Romer (2017).

3.2 Volatility and Credit-to-GDP Gaps

Our benchmark quantitative model relies on realized volatility similar to the measure used in Duprey, Klaus and Peltonen (2017) and Danielsson, Valenzuela and Zer (2018). This simple model allows us to look at a wide set of countries for a long time period. In addition, market disruptions usually show up in such volatility measures and they feature predominantly in various indicators of financial stress or financial conditions. Throughout the rest of the paper, we refer to this model as the "volatility model."

In addition to this broadly-available market-based measure, we also look at the credit-to-gdp gap as used in Drehmann and Juselius (2014). The credit-to-GDP gap is considered a financial vulnerability metric that has a more forward looking element when it comes to financial stress. We look at this measure more closely when we conduct out-of-sample forecasting exercises.

3.3 Textual Data

As for textual data, we rely on a few sources that are contemporary and publicly available.[1]

First, we rely on the OECD Economic Outlook as in Romer and Romer (2017) described above. The textual data herein provides a large panel of textual data that begins in 1967 for 24 countries from an international organization that are published in a regular semiannual frequency. Topic mining on the entire corpus of text results in potentially informative topics that can be gleaned from the text, as shown in Figure 2. Some selective topics appear to stand out during financial crises for different types of countries. Basic Latent Dirichlet Allocation (LDA) analysis on the full corpus with ten topics yields interesting results. For example, the words in Topic 6 from such unsupervised topic modeling appear to be associated with banking crisis episodes in the United States, Japan, and Turkey, respectively; whereas the words that constitute Topic 7 are associated with recoveries shortly thereafter. Although we do not rely on LDA analysis in our main analysis, such analysis provides a glimpse of the data and shows the potential usefullness of the text in identifying financial crises.

Second, we rely on the media-based Thomson Reuters News Archives (TRNA). The TRNA begins in 1996 and the frequency is to the millisecond that we aggregate to the month for each particular country. Figure 3 provides an overview of the number of non-sports articles (written in English) represented for each country in log scale. The lighter colors represent the countries with a relative abundance of articles in the dataset. Only countries in Africa and the Middle East have relatively fewer non-sports articles that are used in our textual analysis. We begin with about 40 million articles and focus more on the economics and finance related articles, which brings down the count to approximately 20 million. We also use a machine learning tool (newsmap) to build up a framework to tag articles to specific countries. Although the data begins in 1996, we are able to expand the number of countries in our analysis to 62 based on further cleaning of the data.

Finally, we also look at IMF Article IVs, which is also available for a broad set of countries, but the frequency is scattered. Similar to the OECD, however, the documents follow a consistent pattern and potentially focus on financial crises if they do occur in a particular country. IMF Article IVs also have a long time series and can be potentiall used for many more countries; we

[1] Please refer to Appendix A for information on gathering and cleaning the textual data.

Figure 2: **Topic Modeling from OECD Economic Outlooks**

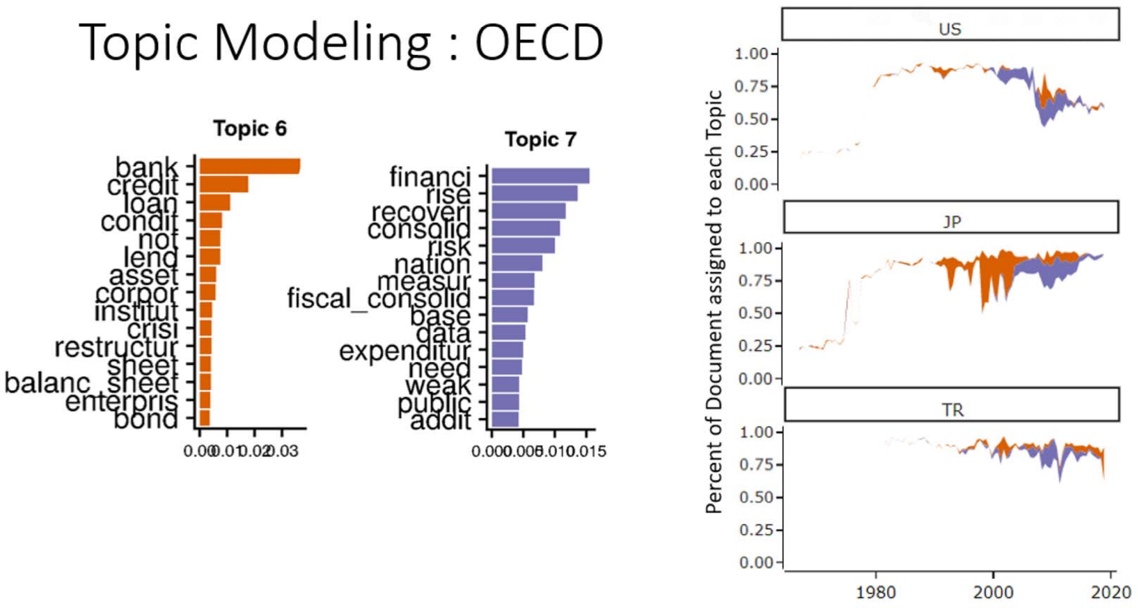
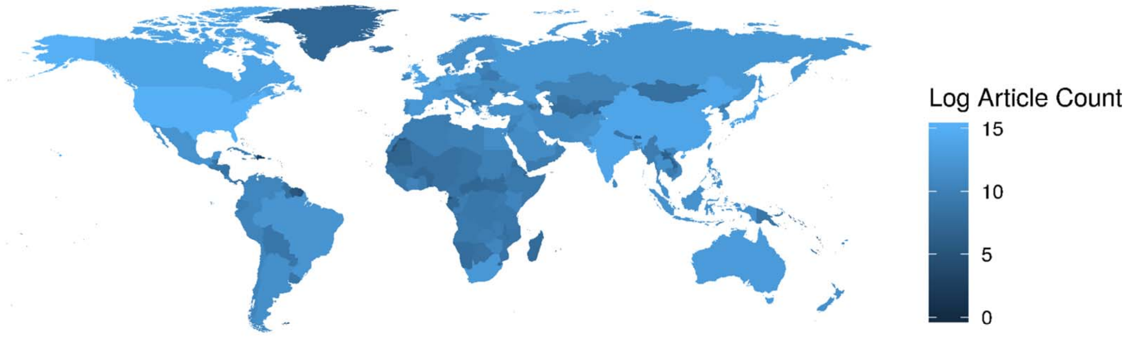


Figure 3: **TRNA Articles by Country**



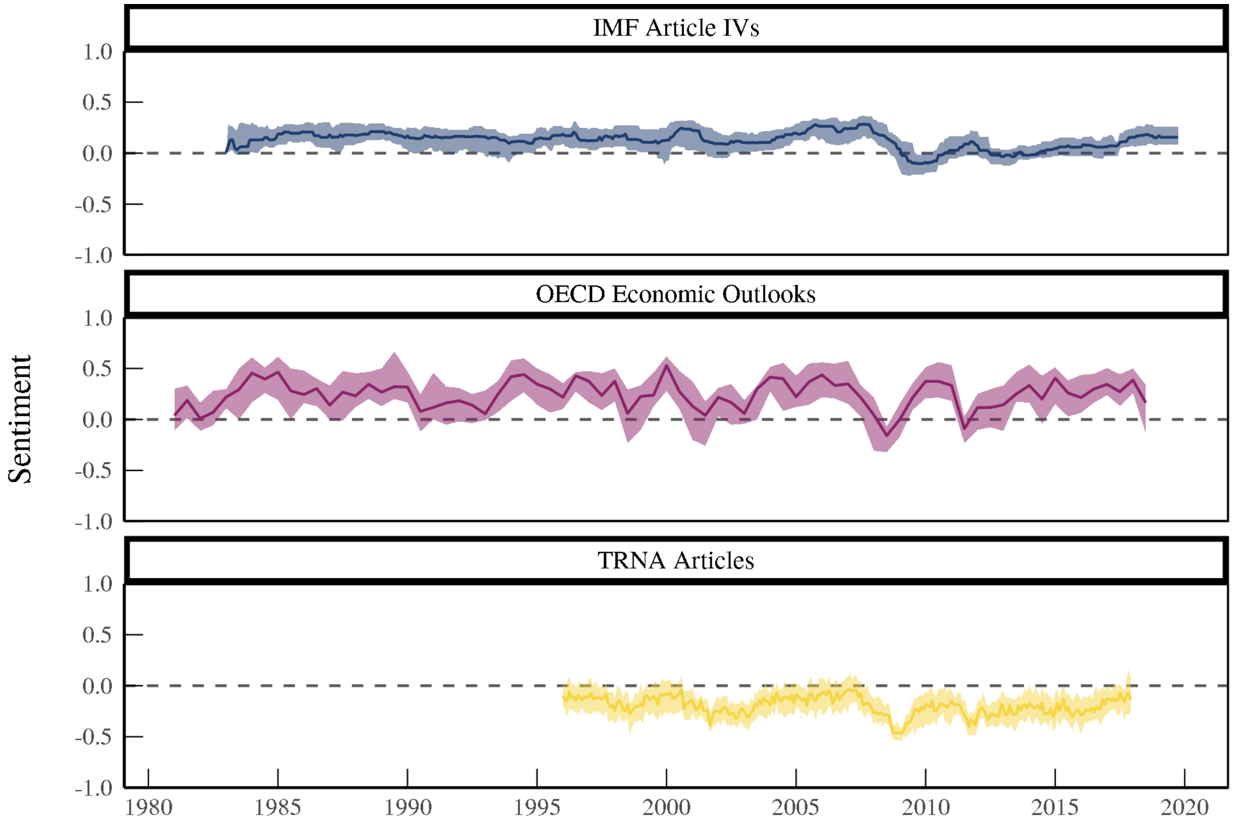
downloaded Article IVs for about 40 countries that goes back to the early 1980s. Table 1 provides a summary of all the data sources, including how many countries (after cleaning) are represented and the time period in which we can conduct analysis in.

Basic sentiment analysis provides some insight on how the different documents could be potentially useful. Figure 4 plots the range of sentiment scores from each source text generated with

Table 1: Summary Statistics of Data Sources

Dataset	Number of Countries Available	Number of Countries Used	Time Period Available	Base Frequency
R&R Crises	24	24	1967-2012	biannual
L&V Crises	118	62	1976-2017	monthly/annual
OECD Text	44	24	1981-2018	biannual
TRNA Text	˜238	62	1996-2017	minutely
IMF Text	˜111	39	1983-2019	annual

Figure 4: **Median Sentiment Scores Using Different Texts**



Showing the Median and 25th and 75th Percentiles across available countries for each text source.

a word dictionary developed by Correa et al. (2017*a*). These plots suggest that these documents have features that may help in detecting financial crises in real time and predicting crises as well as sentiment conveyed in these documents generally dipped during the Great Financial Crisis around the 2008 period.

4 Machine Learning on Textual Data

We rely on supervised machine learning on the aforementioned textual data. First, we create a dictionary of terms to follow based on the Financial Stability Sentiment Dictionary from Correa et al. (2017*a*). These terms are then used to construct a sentiment score according to Correa et al. (2017*b*). We also create a burn dictionary from the OECD Economic Outlook text source from the 1960s to 1980s. Appendix B explains the process. This allows us to follow only terms that exist throughout the sample and minimizes look-ahead bias.

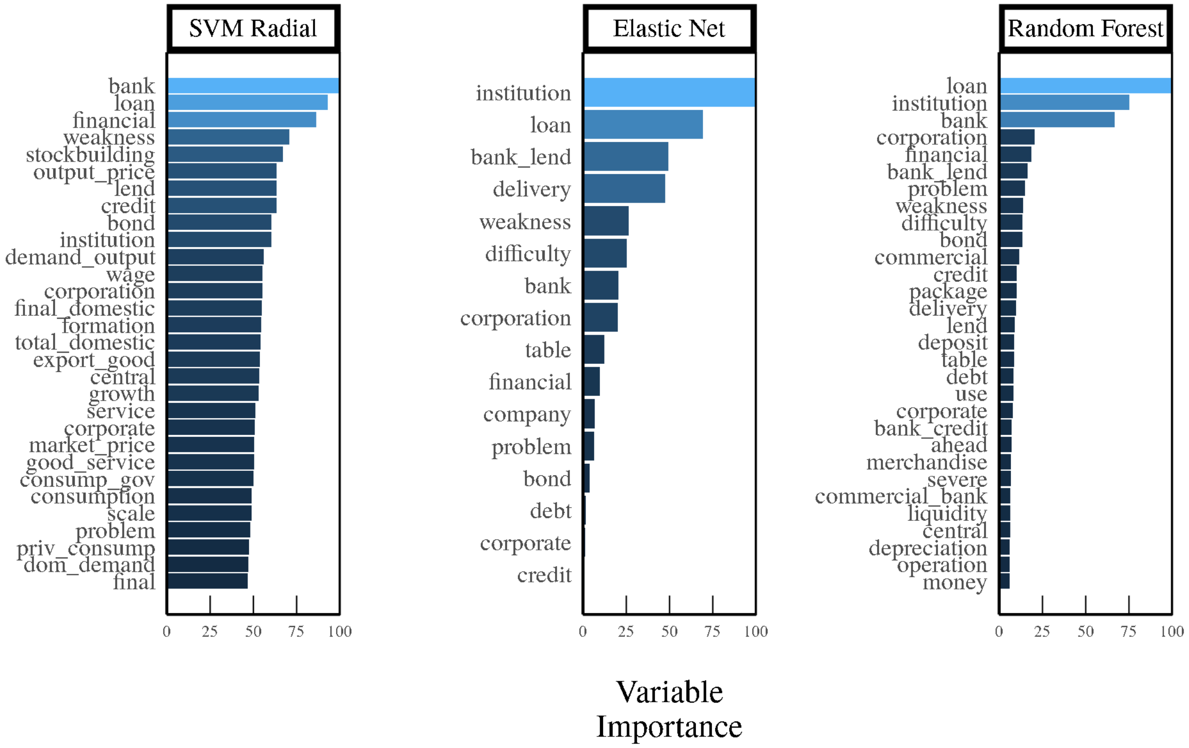
We split the time dimension into a training set and testing set to avoid data-leakage. We decide to split at 2005 so that our out-of-sample prediction results are based on how well our model does in identifying/predicting crises during the GFC period. In other words, we use data prior to 2005 to train the data for creating and tuning the models. Tuning of the models is done by splitting the training data into multiple folds, also along the time dimension. We create ever-increasing validation sets of the data, always predicting one year forward, training on all of the data to that point. This is our way of ensuring we respect the time-series nature of the data. By comparing performance on these validation sets we can select optimal hyper-parameters and train a final model on the whole training data using those parameters.

The main machine learning model we use are Support Vector Machines (SVMs), but we also use GLMNET (based on Elastic Net, Ridge, and Lasso), Random Forests, Neural Nets, Adaptive Boosted Forests, Extreme Random Forests, etc. All the results we show are based on SVMs because the SVM results, though at times inferior to other models, most consistently outperform the other models. However, whichever models we use, the results are alarmingly similar.

We then test the data on the post 2005 data and use the area under the Reciever Operating Curve (ROC), or AUROC statistics, to assess performance out of sample. In addition to this "chunk" method, in which we train the model up until 2004, then look at out-of-sample properties after 2004, we also provide results, we also show results for the "expanding horizon" model, where we re-estimate as we move along the time dimension. For identifying financial crises, if one believes there is a stable relaionship, either method should work. If we believe there is a time-varying relationships between text and financial crises, a rolling-windoe method may be superior. It turns out that the rolling-window results are very similar to the expanding sample results. Therefore, we only report the "chunk" and the "expanding" results in this paper.

In order to see if text data adds value to out-of-sample predictions, we use the baseline logit model based on realized volatility calculated from daily stock return data scaled by country. We also compare Machine Learning models with logit models using sentiment scores. For predicting financial crises, we also report results of logistic regressions using the credit-to-GDP gap as one of

Figure 5: **Variable Importance for R&R Minor Crises on OECD Text**



our regressors.

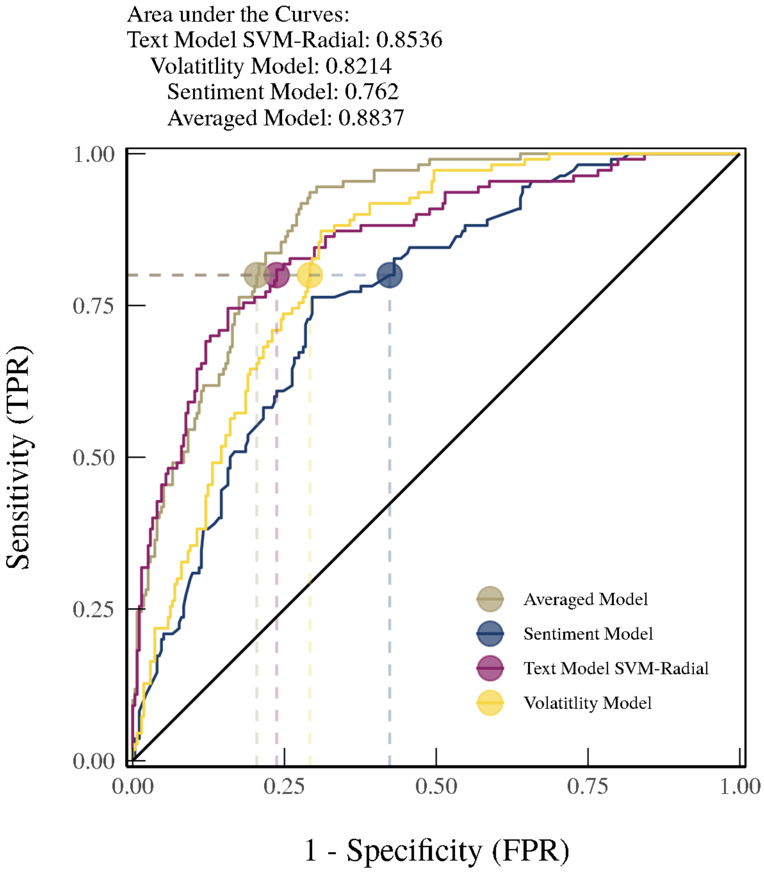
5 Nowcasting Results

5.1 Identifying Romer and Romer (2017) Crises using OECD Text

Figure 5 describes the variable importance for our first exercise, which tries to identify Romer and Romer (2017) minor or more severe crises using OECD text (score of 4 or more). The variable importance is based on the percentage AUROC gains from including a particular word in your model. We can see that the word "bank" and "loan," with its various interactions provide valuable insight into detecting a crisis or not.

As shown in Figure 6, out-of-sample identification for the post 2004 period ROC curves (from 2005 to 2012) suggests that the model using the text-based machine learning approach outperforms the volatility-only model, which in turn outperforms the sentiment score-based model. Here we

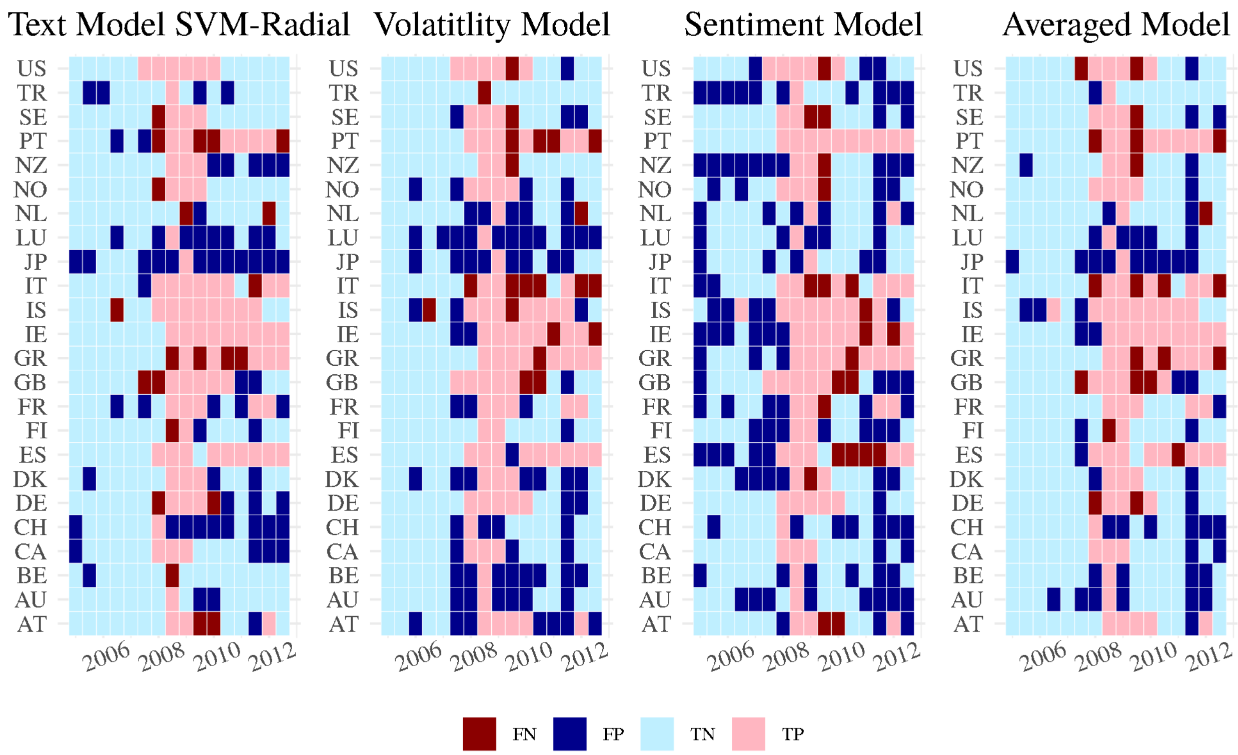
Figure 6: **Results for R&R Minor Crises on OECD Text**



show only the results from the SVM model, but, as mentioned above, the results hold regardless of the specific machine learning model. When we average the different models, we see further improvements in the AUROC.

As shown in the confusion matrix in Figure 7, this is because volatility, sentiment, and the textual model all convey slightly different information in identifying minor or more severe crisis based on false positives or false negatives given an 80 percent true positive rate. “FN,” “FP,” “TN,” and “TP” stand for false negative, false positive, true negative, and true positive observations, respectively. Therefore, the lighter the confusions matrix, the better out-of-sample performance is recorded for a particular model (or average of models). The averaged model is a model where we simply average the probability estimate outputs from each individual model. For example, whereas the textual model, volatility model, and the sentiment model all provide many false positives and false negatives individually for many countries, the simple-averaged model produces only one false

Figure 7: **Confusion Matrix for R&R Minor Crises on OECD Text**



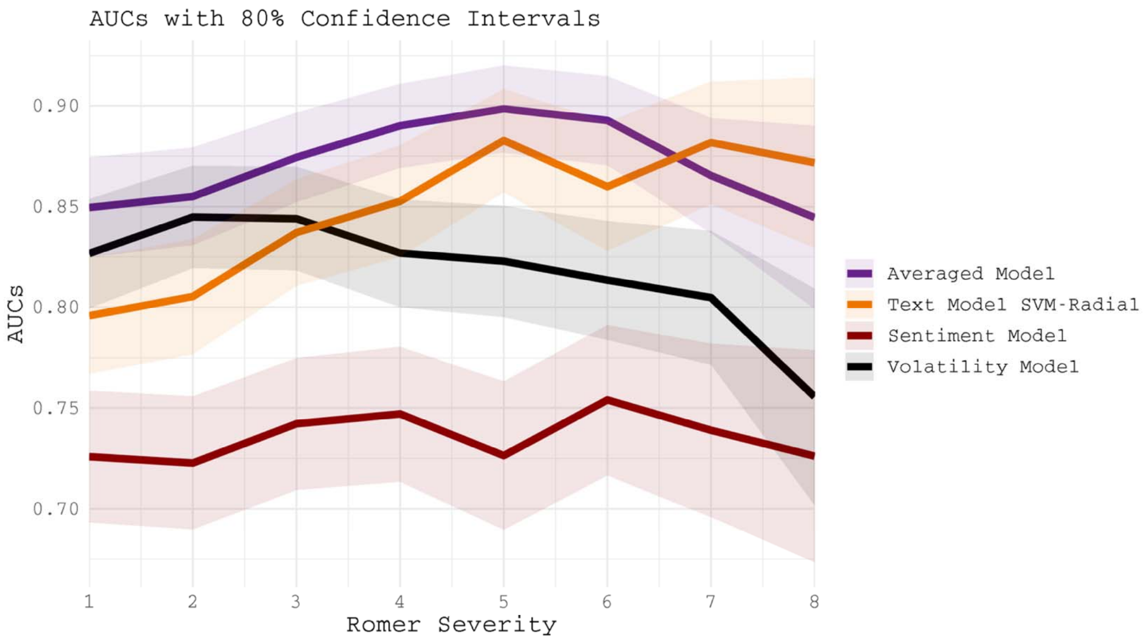
positive or false negative in countries like Turkey, Norway, France, and Austria. That said, the model output from the SVM sometimes work perfectly (without any dark colors) without error for countries like the United States, Ireland, and Spain. The sentiment model is only perfect for Portugal; whereas the volatility model does not perform perfectly for any country out of sample.

In fact, we find that, as in Figure 8, when as we vary the degree of crises we want to identify (starting from minor credit disruptions to more severe crises), the average model generally outperforms any of the other models and consistently outperforms the volatility-only model, especially when the severity of the crisis increases.[2] The area under the ROC curve peaks about at 0.90 when we try to out-of-sample nowcast a Romer and Romer (2017) crisis severeity of 5 or more. As severity of the crisis we are trying to identify increases, the information content in the OECD Economic Outlook text outperforms even the average model, as the signal to noise ratio deteriorates rapidly for the volatility model.

We can take a closer look at how individual words or bigrams affect the probability of the SVM model based on Local Interpretable Model-agnostic Explanations (LIME) based on Tulio

[2] We only implement the model to major crisis (or more severe) due to data availability for the most severe crises.

Figure 8: **Results for Various R&R Crises on OECD Text**

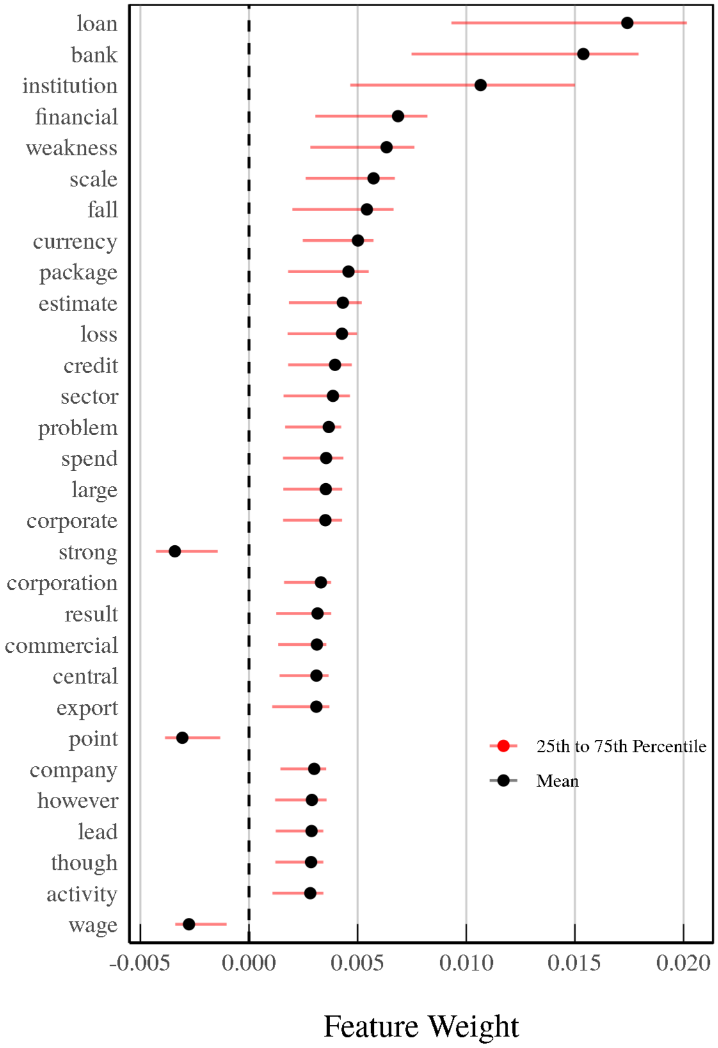


Ribeiro, Singh and Guestrin (2016), which identifies an interpretable model over the interpretable representation that is locally faithful to the classifier model. In Figure 9 notice that the top 30 factors, in terms of feature weights, include many of the important variables listed in Figure 5. Here, however, we can also see how much, on average, they contribute to the probability. For example, the mentioning of "loan" in the OECD Economic Outlook text contributes to approximately an increase of about 2 percentage points in the probability that currently we are are a minor or more severe crisis. In contrast, words such as "strong" are associated with a decrease in this probability, on average.

LIME provides a method to unveil the black-box nature of machine learning algorithms by running a model with and without the words of interest. To provide a more specific example, Figure 10 shows how the OECD Economic Outlook text for the first half of 2008 in the United States looks like with darker red words contributing more (and darker blue contributing less) to the probability that the machine learning algorithm thinks that the United States is in a crisis state. Again, the word "bank" is quite dark, implying that the term is helping to signal that the United States is in a crisis.

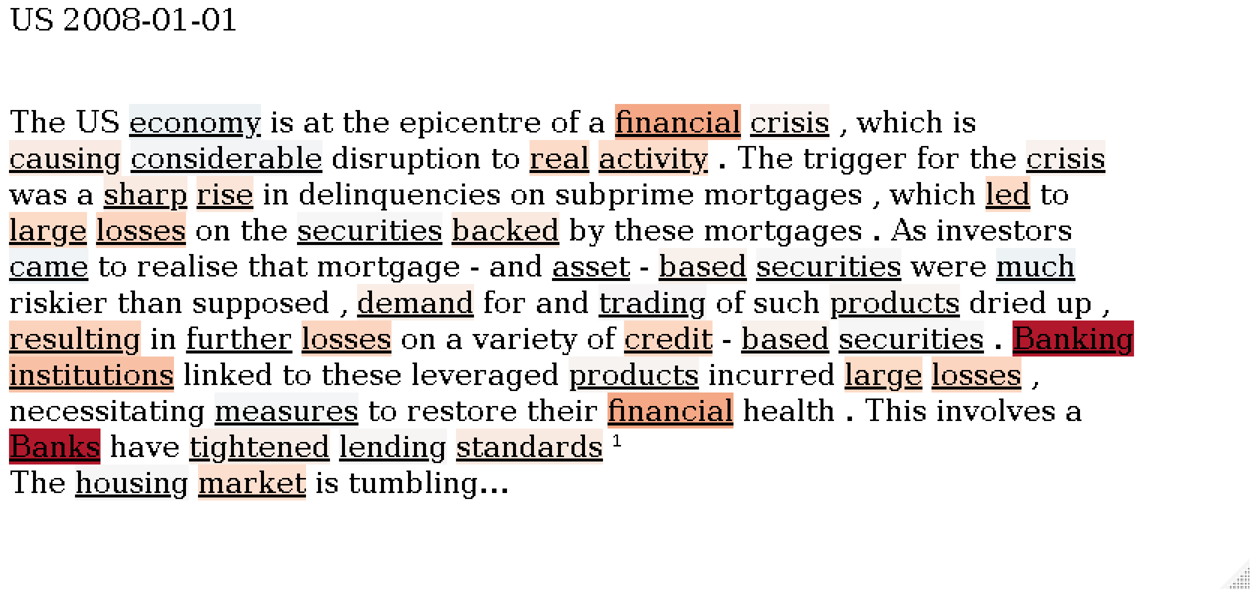
These are quite interesting, but intuitive results. To the extent that, usually, the OECD economic outlooks focus on forecasts for real economic activity, if the documents are inundated with

Figure 9: **Top 30 Average Feature Weights; Minor Crises on OECD Text**



references to the banking sector, loans, and other financial developments, this is quite unusual in normal circumstances and conveys that the financial sector is weighing on real economic activity. This is exactly what Romer and Romer (2017) tries to identify; whether problems in the financial sector are having any real effects.

Figure 10: **US OECD Economic Outlook with LIME Highlighting**



US 2008-01-01

The US economy is at the epicentre of a financial crisis , which is causing considerable disruption to real activity . The trigger for the crisis was a sharp rise in delinquencies on subprime mortgages , which led to large losses on the securities backed by these mortgages . As investors came to realise that mortgage - and asset - based securities were much riskier than supposed , demand for and trading of such products dried up , resulting in further losses on a variety of credit - based securities . Banking institutions linked to these leveraged products incurred large losses , necessitating measures to restore their financial health . This involves a Banks have tightened lending standards [1]
The housing market is tumbling...

5.2 Identifying Laeven and Valencia (2013) Crises using TRNA Text and Combinations of Models

Figure 11 shows the results from a similar exercise that takes advantage of the wider cross section of countries in two datasets; Laeven and Valencia (2013) crisis data, which is considered closely related to "moderate" crises in Romer and Romer (2017), and the TRNA data, which has more than twice the cross section of countries in the OECD data. Here, the simple average of the volatility and sentiment models work best and the textual model vastly underperforms. The confusion matrix in Figure 12 illustrates this in greater detail for the various countries in the sample. The average model outperforms all the individual models, again pointing to different types of information conveyed in the different numerical and textual data.

Finally, we can also bring in IMF Article IVs to enhance our out-of-sample forecasting. As a template, we can choose the models that work best and combine different models. In our next example, we combine the SVM and sentiment models based on trying to predict Moderate crises according to Romer and Romer (2017) using OECD Economic Outlook text, the SVM and volatility models based on trying to predict banking crises according to Laeven and Valencia (2013) using TRNA text, and finally an SVM model predicting banking crises using IMF Article IV text. Figure 13 illustrates the results and shows that the Area under the ROC curve can be pushed up to 0.92.

Figure 11: **Results for LV Banking Crises on TRNA Text**

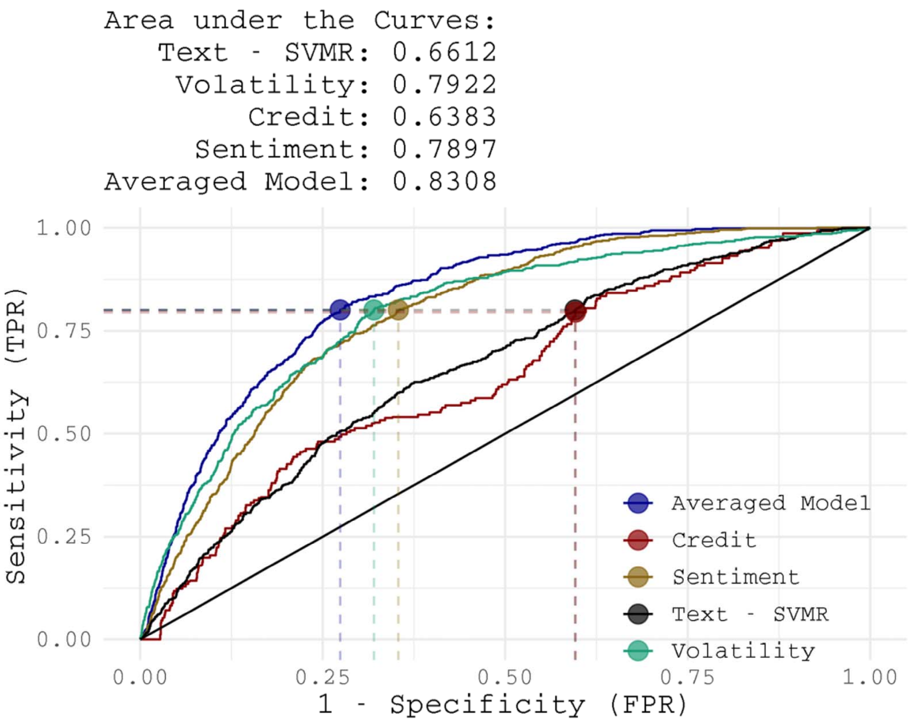


Figure 12: **Confusion Matrix for LV Banking Crises on TRNA Text**

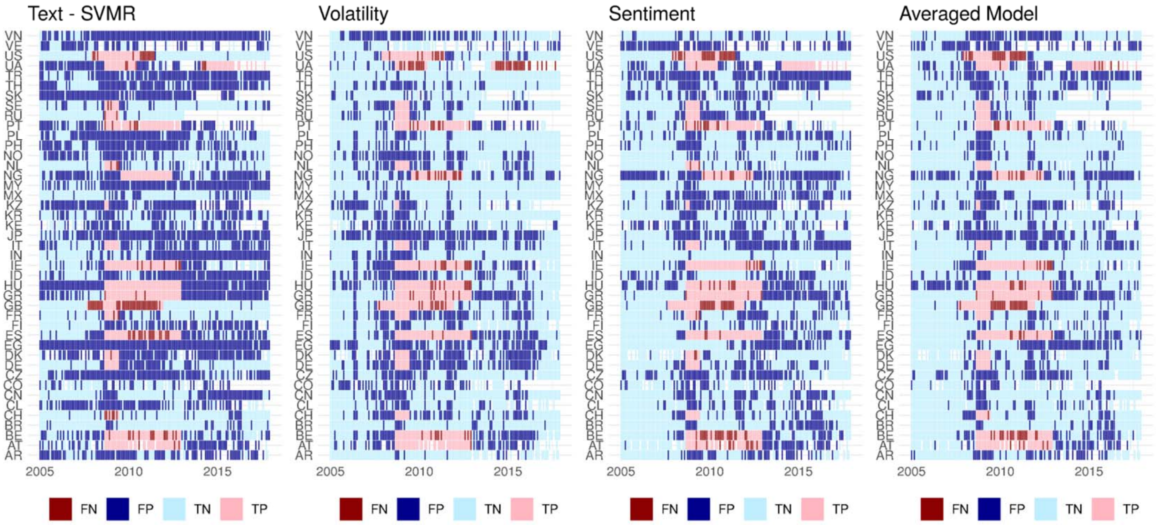
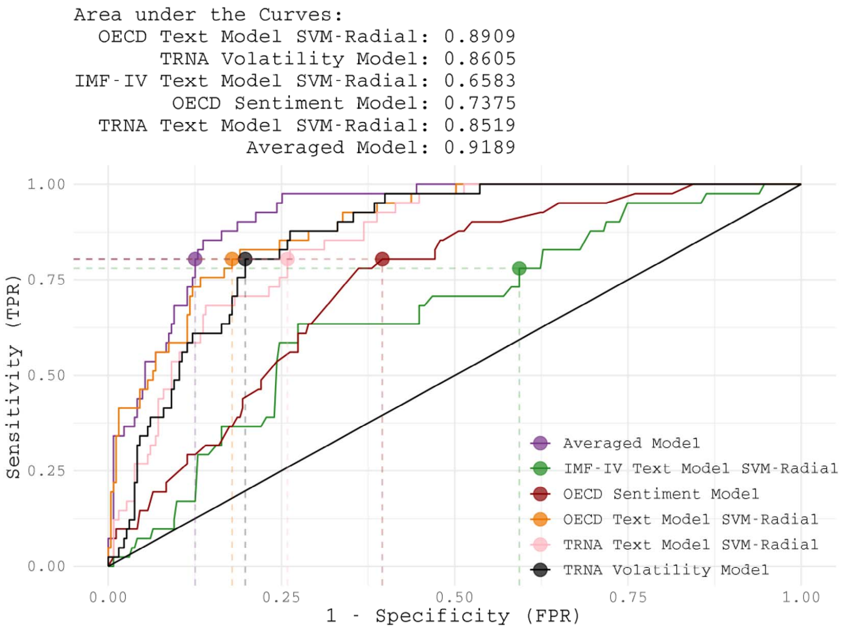


Figure 13: **Results for Combined Model**



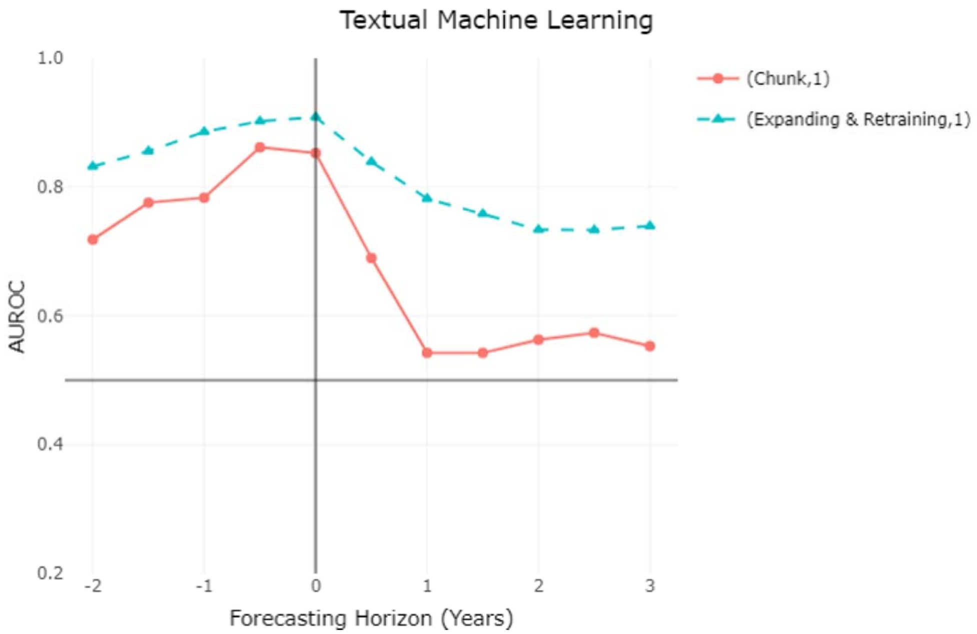
Although this is not statistically significantly different from the OECD SVM model, it is still 0.03 greater and could be more meaningful with additional optimal weighting of the different models.

6 Forecasting and Backcasting

Next, we undertake forecasting and backcasting exercises. Again, we train a model up to 2004 and show out-of-sample results for predicting and backcasting Romer and Romer (2017) Minor (or more severe) crises using OECD text and, comparatively, using TRNA text). For all models, we forecast crises x periods ahead or behind. We find that the volatility model performs quite poorly in backcasting Romer and Romer (2017) Minor (or more severe) crises. The OECD text model is quite dominant in backcasting as far back as two years, and keeps its edge into the nowcasting period. Yet when forecasting even six months ahead, OECD faces a steep drop-off while volatility remains at a peak. Forecasting one year ahead, the OECD model has an AUC of just 0.55 as indicated in the red line in Figure 14.

We find that OECD text models outperform the simple OECD sentiment score model throughout the backcasting and nowcasting period. Interestingly, the performance decline into the forecasting period is not as severe for the sentiment model. Their performance is almost equal when

Figure 14: **Results for R&R Minor Crises on OECD Text**



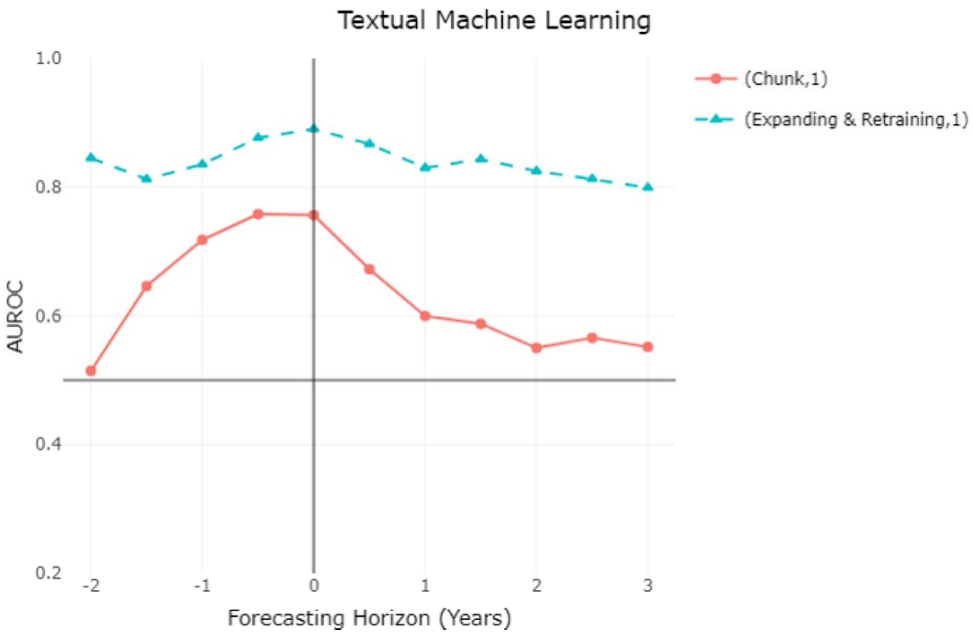
forecasting six months out, and the sentiment model outperforms the text model in later periods.

In contrast, TRNA performs better relatively in nowcasting and even forecasting than it does in backcasting as indicated in the red line in Figure 15, relative to the results for the OECD text. Its performance holds up even six months in the future, and then begins a gradual decline. Although volatility works well in this forecasting context, we find that a simple sentiment model outperforms the more complicated machine learning model in almost all forecasting and backcasting horizons.

We also illustrate how retraining on an expanding sample helps in terms of forecasting using only the SVM models. In particular, compared to the OECD text (Figure 14), the TRNA text greatly outperforms even up to a three-year horizon (Figure 15) . This illustrates that the abundance of text helps in learning new relationships as we augment the sample for training the machine learning models.

Finally, using only OECD data, we also look at how the models work when trying to nowcast/forecast/backcast the exact onset of a crisis by not including crisis periods with the exception of the beginning period of crises in our models. Figure 16 illustrates that it is a bit more difficult in forecasting the exact timing of the onset of a crisis, especially from a year to two years out. That

Figure 15: **Results for R&R Minor Crises on TRNA Text**



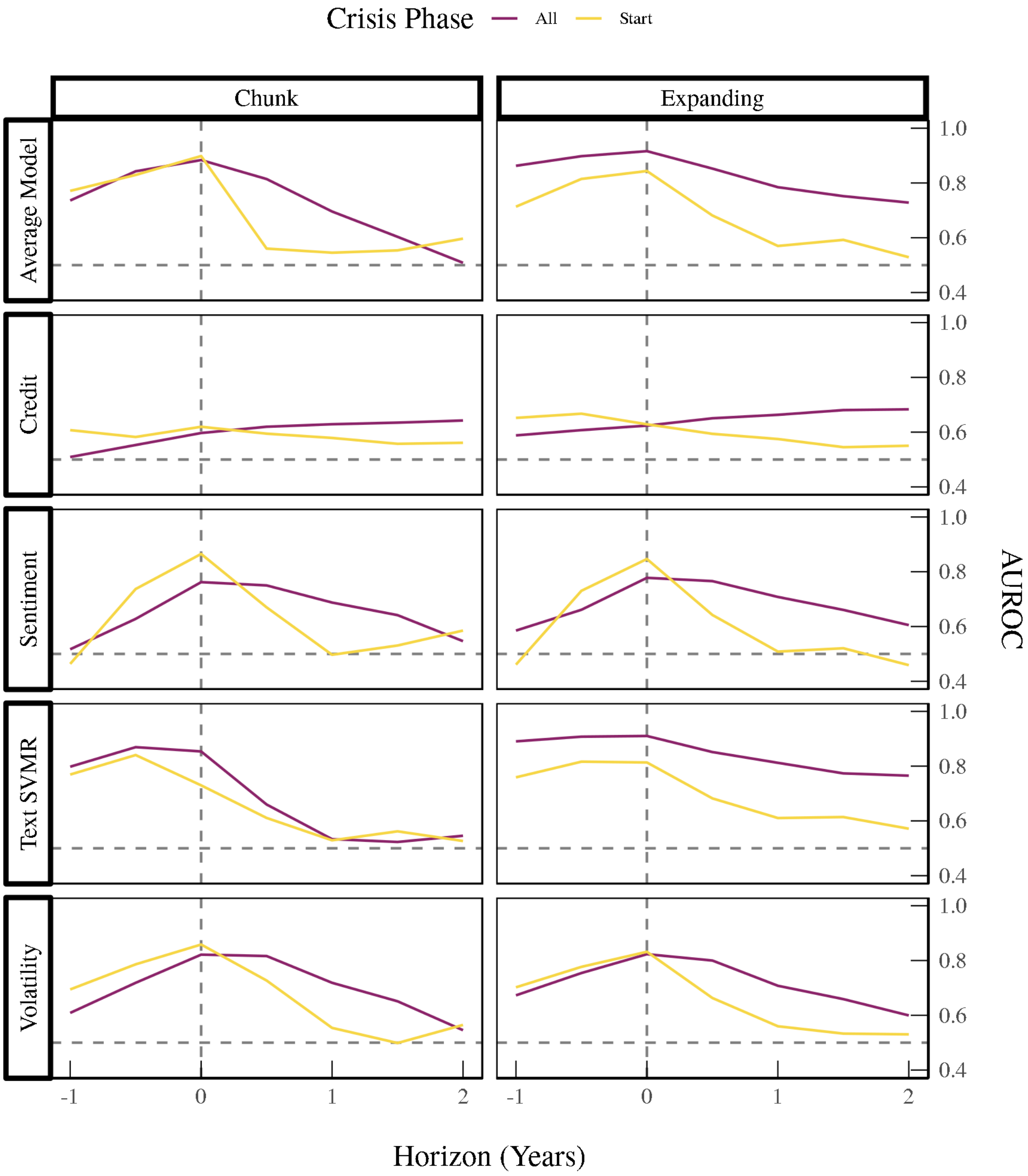
said, the sentiment models perform well in identifying crises whether using the chunk method or the expanding method.

7 Conclusion

We find that text helps in the identification and prediction of financial crises. In particular, different types of text can be trained on different crisis definitions to better identify different severities of crises. As expected, the OECD textual models perform quite well at identifying Romer and Romer (2017) crises. The TRNA sentiment seems to provide a lot of information, with the added benefit of access to a broader set of countries. IMF Article IVs have inconsistent frequencies, but clearly add some information due to their relevance in writing about financial stress or financial crises. A naive combination of sources shows potential improvements over individual text sources. Finally, text data appears to provide statistically significant improvement over a baseline model with just volatility, especially in detecting more severe types of crises.

Our results may also shed light on the channels by which OECD and TRNA text models can

Figure 16: **Crisis Phase: Starts vs. All**



identify financial crises. Recall the two mechanisms outlined earlier: descriptive and instigative. It is natural to think that the descriptive mechanism would be more powerful for backcasting than for forecasting. This is because it is easier to describe the financial landscape of the present or past than it is to describe the financial landscape of the future. This leads us to expect that texts with a relative advantage in backcasting may operate more strongly through the descriptive mechanism. We bolster this expectation with the observation that the instigative mechanism cannot apply to backcasting at all; it is impossible for a text to instigate action in the past.

Taken together, and combined with the result that TRNA performs relatively better in forecasting than backcasting and that OECD is the opposite, these dynamics point towards the conclusion that written news, or at least TRNA, are relatively more instigative than descriptive, compared to OECD outlook text. Of course, there are other explanations. Two text sources operating solely though the descriptive mechanism might differ in forecasting/backcasting power simply because one attempts to describe the past and the other attempts to describe the future. While it is possible that TRNA describes the future while OECD describes the past, we think this is unlikely. The OECD reports are explicitly meant to be outlooks, whereas TRNA focuses almost entirely on present, real-time events.

Understanding the nature of how text helps in understanding financial crises implies the following; various machine learning and econometric methods, applied to a variety of types of crises and using a range of different text sources can potentially help policymakers determine where a particular financial system is in the financial cycle, or more specifically, whether the financial system is weighing down on real economic activity. This, in turn, can be useful for macroprudential policy, monetary policy, and even fiscal policy, as different phases in the financial cycle have different implications for real economic activity.

Appendix A Data Cleaning

A.1 OECD Text Data

OECD Economic Outlooks are downloaded as PDFs from the OECD iLibrary, which provides all of the documents going back to 1967. The PDFs are converted into HTML files, which then allows us to to identify blocks of text, font size, and position of the text on the page. This allows the scanning of the document in order to identify the sections that are about specific countries, these almost always start with a header line for that country, i.e. "CANADA", in large font.

Once we have all of the scanned documents broken into the specific countries, we begin removing numbers, stop words ("the", "and", etc), short words (less than three characters), and converting all text to lower case. We then lemmatize the remaining words, using the R package *textstem* (Rinker, 2018*b*) and the *lexicon* (Rinker, 2018*a*) engine. This combines various versions of a word ("fall" and "fell" and "falling") into one instance of that word ("fall"). This is preferable to stemming, which removes common endings to words, as that misses some peculiarities of the English language.

Finally, we create bigrams from the text that remains. This will join words that are adjacent, for instance "federal" and "reserve" would become "federal_reserve". We keep both the original unigrams and bigrams.

A.2 TRNA Text Data

We extensively cleaned and sorted articles from the Thomson Reuters News Archive, which contains around 66.3 million articles. After dropping all articles written in languages other than English, we are left with 42.5 million articles. We constructed a story chain identifier, which sequentially links articles that are updates to original articles. We only keep the first article in a story chain and discard any articles that are updates, leaving us with 28.1 million articles. Finally, we drop articles that are classified as "repeated works," (ie, obituaries, weekly oil readouts), alerts and headlines with no body text, ending with a final 19.3 million articles.

Because the Archive does not identify the country or countries about which articles are written, we use an algorithm to classify articles. Using the R package *newsmap* (Watanabe, 2018), we first use a seed dictionary of proper nouns that map to countries to make a "first-pas" over the corpus to find other proper nouns to add to the dictionary. All articles are classified again in a second round with the expanded dictionary. The *newsmap* (Watanabe, 2018) package gives each article a score for each country. Higher scores mean the article is more likely to be about a certain country. Articles may have high scores in multiple countries. To illustrate the process, the seed dictionary associates "Paris", "France" and "French" with the country "France." If any of these words are found in an article, the algorithm gives the article a high score for the category “France.” Then, *newsmap* (Watanabe, 2018) finds other words that are common in these articles most likely to be classified as “France,” such as “Sorbonne,” and gives articles that contain the word “Sorbonne” a high score in the category of “France.”

We manually classified a random sample of 560 articles and compared these to the scores outputted from the algorithm. We find that, using a score cutoff of 2.25 for the highest-value country for each article, and a score cutoff of 3.5 for the 2nd-5th countries, results in a 90% confidence for the average classification. That is, a random country-article classification in the results is 90% likely to be correct, where correct is defined by the manual verification. In the end, we use a cutoff of 0.0 for the first country, 2.5 for the second and third, and discard the rest. After the process is complete, 87% of the articles are classified with at least one country, 3.6% of which are classified with two countries and almost 1% of which are classified with three countries, with 82% confidence. Using these results, there are 238 unique countries in TRNA.

The TRNA articles are then subset to only articles that contain at least one relevant economics-related word. These words are “financ-,” “stability,” “econ-,” “market,” “investment,” “trade,” “stock,” “sovereign,” “debt,” “bank,” and “assets.” Articles that contain at least one sports-related word is removed. These words are “soccer,” “championship,” “World Cup,” “basketball,” “football,” “baseball,” “sport,” “tennis,” and “olympic.”

Articles are aggregated to a monthly frequency. We filter to only country-months with at least fifty articles, and only countries with at least 75 percent of months meeting this condition.

We remove numbers, stop words (“the”, “and”, etc), short words (less than three characters), and converting all text to lower case. We then lemmatize the remaining words, using the R package *textstem* (Rinker, 2018*b*) and the *lexicon* (Rinker, 2018*a*) engine. This combines various versions of a word (“fall” and “fell” and “falling”) into one instance of that word (“fall”). This is preferable to stemming, which removes common endings to words, as that misses some peculiarities of the English language.

Finally, we create bigrams from the text that remains. This will join words that are adjacent, for instance “federal” and “reserve” would become “federal_reserve”. We keep both the original unigrams and bigrams.

A.3 IMF Article IV Text Data

Staff Reports for Article IV Consultations are downloaded as PDFs from the Archives Catalog from the International Monetary Fund. While there is at most one Article IV Consultation per country per year, there may be multiple iterations of Staff Reports available for download for the same Article IV Consultation. In these cases, we keep only the most complete Staff Report at the latest release date. We use this release date of publication instead of IMF's stated year of the Article IV Consultation in our analysis.

We use Adobe Acrobat's OCR to convert scanned PDFs to searchable text PDFs. The PDFs are converted into HTML files, from which we extract only text. We remove numbers, stop words ("the", "and", etc), short words (less than three characters), and convert all text to lower case. We then lemmatize the remaining words, using the R package *textstem* (Rinker, 2018*b*) and the *lexicon* (Rinker, 2018*a*) engine. This combines various versions of a word ("fall" and "fell" and "falling") into one instance of that word ("fall"). This is preferable to stemming, which removes common endings to words, as that misses some peculiarities of the English language.

Finally, we create bigrams from the text that remains. This will join words that are adjacent, for instance "federal" and "reserve" would become "federal_reserve". We keep both the original unigrams and bigrams.

Appendix B Creating the OECD Dictionary

The OECD dictionary is created from the raw OECD text from 1967 to 1980. This allows us to create a dictionary of words that we will then follow throughout the training and testing sets that will be entirely void of any time leakage. There are seven countries available during the 1967 to 1980 period: Canada, France, Germany, Italy, Japan, United States, and the United Kingdom. To create our dictionary, we require a term be present across the 7 countries during this time period. This helps avoid country specific words in the text. We also require that the tokens be mentioned at least 50 times over the entire time period, across all countries. This helps us remove both infrequent words, but also words that are misspelled, or are very specific to one document. This provides us with a list of terms, from which we manually remove four types of terms: (1) references to time ("year", "summer", "january"), (2) units ("percent", "billions"), (3) section pointers ("outlook", referencing OECD Economic Outlooks, "source"), (4) common typos ("tion", "ing", words are often split, and these common endings get picked up as unique terms).

We are then left with a dictionary of 881 terms.

References

Aikman, David, Michael Kiley, Seung Jung Lee, Michael G. Palumbo and Missaka Warusawitharana. 2017. "Mapping heat in the U.S. financial system." *Journal of Banking & Finance* 81:36 – 64.
URL: *http://www.sciencedirect.com/science/article/pii/S0378426617301012*

Alessi, Lucia and Carsten Detken. 2018. "Identifying excessive credit growth and leverage." *Journal of Financial Stability* 35:215 – 225. Network models, stress testing and other tools for financial stability monitoring and macroprudential policy design and implementation.
URL: *http://www.sciencedirect.com/science/article/pii/S1572308917304291*

Angelico, Cristina, Juri Marcucci, Marcello Miccoli and Filippo Quarta. 2019. Can We Measure Inflation Expectations Using Twitter? Working paper.

Bluwstein, Kristina, Marcus Buckmann, Andreas Joseph, Miao Kang, Sujit Kapadia and Ozgur Simsek. 2019. Credit Growth, the Yield Curve and Financial Crisis Prediction: Evidence from a Machine Learning Approach. Working paper.

Brave, Scott and R. Andrew Butters. 2012. "Diagnosing the Financial System: Financial Conditions and Financial Stress." *International Journal of Central Banking* 8(2):191–239.
URL: *https://ideas.repec.org/a/ijc/ijcjou/y2012q2a6.html*

Cerchiello, Paola, Giancarlo Nicola, Samuel Rönnqvist and Peter Sarlin. 2017. "Deep learning bank distress from news and numerical financial data." *ArXiv* abs/1706.09627.
URL: *https://arxiv.org/abs/1706.09627v3*

Cesa-Bianchi, Ambrogio, Fernando Eguren Martin and Gregory Thwaites. 2019. "Foreign booms, domestic busts: The global dimension of banking crises." *Journal of Financial Intermediation* 37:58 – 74.
URL: *http://www.sciencedirect.com/science/article/pii/S1042957318300615*

Correa, Ricardo, Keshav Garud, Juan-Miguel Londono-Yarce and Nathan Mislang. 2017*a*. Constructing a Dictionary for Financial Stability. Ifdp notes Board of Governors of the Federal Reserve System.
URL: *https://doi.org/10.17016/2573-2129.33*

Correa, Ricardo, Keshav Garud, Juan-Miguel Londono-Yarce and Nathan Mislang. 2017*b*. Sentiment in Central Banks' Financial Stability Reports. International Finance Discussion Papers 1203 Board of Governors of the Federal Reserve System.
URL: *https://doi.org/10.17016/IFDP.2017.1203*

Danielsson, Jon, Marcela Valenzuela and Ilknur Zer. 2018. "Learning from History: Volatility and Financial Crises." *The Review of Financial Studies* 31(7):2774–2805.
URL: *https://doi.org/10.1093/rfs/hhy049*

Drehmann, Mathias and Mikael Juselius. 2014. "Evaluating early warning indicators of banking crises: Satisfying policy requirements." *International Journal of Forecasting* 30(3):759 – 780.
URL: *http://www.sciencedirect.com/science/article/pii/S0169207013001428*

Duprey, Thibaut, Benjamin Klaus and Tuomas Peltonen. 2017. "Dating systemic financial stress episodes in the EU countries." *Journal of Financial Stability* 32:30 – 56.
URL: *http://www.sciencedirect.com/science/article/pii/S1572308917305016*

Gentzkow, Matthew, Bryan Kelly and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57(3):535–74.
URL: *http://www.aeaweb.org/articles?id=10.1257/jel.20181020*

Iyer, Rajkamal, Manju Puri and Nicholas Ryan. 2016. "A Tale of Two Runs: Depositor Responses to Bank Solvency Risk." *The Journal of Finance* 71(6):2687–2726.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12424*

Kalamara, Eleni, Arthur Turrell, Chris Redl, George Kapetanios and Sujit Kapadia. 2019. Making text count: economic forecasting using newspapers. Working paper.

Laeven, Luc and Fabian Valencia. 2013. "Systemic Banking Crises Database." *IMF Economic Review* 61(2):225–270.
URL: *http://www.jstor.org/stable/43302165*

Lee, Seung Jung, Kelly E. Posenau and Viktors Stebunovs. 2018. "The anatomy of financial vulnerabilities and banking crises." *Journal of Banking & Finance* .
URL: *http://www.sciencedirect.com/science/article/pii/S0378426618300864*

Reinhart, Carmen M. and Kenneth S. Rogoff. 2009. *This Time Is Different: Eight Centuries of Financial Folly*. Princeton, New Jersey: Princeton University Press.
URL: *http://www.reinhartandrogoff.com/*

Rinker, Tyler W. 2018*a*. *lexicon: Lexicon Data*. Buffalo, New York: . version 1.2.1.
URL: *http://github.com/trinker/lexicon*

Rinker, Tyler W. 2018*b*. *textstem: Tools for stemming and lemmatizing text*. Buffalo, New York: . version 0.1.4.
URL: *http://github.com/trinker/textstem*

Romer, Christina D. and David H. Romer. 2017. "New Evidence on the Aftermath of Financial Crises in Advanced Countries." *American Economic Review* 107(10):3072–3118.
URL: *http://www.aeaweb.org/articles?id=10.1257/aer.20150320*

Shiller, Robert J. 2017. "Narrative Economics." *American Economic Review* 107(4):967–1004.
URL: *http://www.aeaweb.org/articles?id=10.1257/aer.107.4.967*

Tulio Ribeiro, Marco, Sameer Singh and Carlos Guestrin. 2016. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." *arXiv e-prints* p. arXiv:1602.04938.
URL: *https://ui.adsabs.harvard.edu/abs/2016arXiv160204938T*

Watanabe, Kohei. 2018. *Newsmap*.
URL: *https://doi.org/10.1080/21670811.2017.1293487*