Board of Governors of the Federal Reserve System

IFDP Notes

June 2017

Constructing a Dictionary for Financial Stability

Ricardo Correa, Keshav Garud, Juan-Miguel Londono-Yarce, and Nathan Mislang

Constructing a Dictionary for Financial Stability[1]

In a recent paper analyzing the sentiment of central bank communications (Correa, Garud, Londono, and Mislang, 2017), we constructed lists of words conveying positive and negative sentiment—a dictionary— that is calibrated to the language of financial stability reports (FSRs). We then used this financial stability dictionary to generate a financial stability sentiment (FSS) index by comparing the total number of positive and negative words in the text of the FSRs. We found that the sentiment index was mostly driven by developments in the banking sector, and that central banks' sentiment was related to the financial cycle and deteriorates just prior to the start of a banking crisis. In this note, we explain in detail how we made word-level choices in our dictionary. In the note, we also consolidate our lessons from this process into a framework for thinking about dictionary construction.

As suggested by Loughran and McDonald (2011) and Henry and Leone (2016), a dictionary should be tailored to the context in which it is being applied, as the sentiment conveyed by words might depend on that context. For example, the word "leveraged" has a different meaning for financial stability ("leveraged loans") than for marketing ("we leveraged our position"). We build a custom dictionary for financial stability because existing dictionaries are not intended to be used in the context of central bank communications.

One challenge of building a dictionary is validation. Specifically, it is important to verify that a dictionary captures the intended sentiment, but this task can be difficult. Broberg et al. (2010) examine the literature on applied content analysis and provide a detailed list of recommended procedures to enhance the validity of a dictionary. Their recommendations coincide with the process we follow to construct our dictionary in that we inductively determine a list of frequently used words and assign independent researchers to examine randomly sampled sentences that contain these words.

The Value of Dictionaries

Dictionaries are useful for measuring the sentiment of documents that do not have a clear alternative method for classification. Other methods for sentiment analysis, such as machine learning, are powerful, but they are supervised models, which require a pre-classified training sample. A training sample is a set of documents that are manually classified (supervised model) as conveying a particular sentiment or tone (e.g., positive or negative). The training sample serves then as input for the machine learning algorithm to determine whether other related documents convey a particular sentiment. It is not always feasible to create a training sample because of, for example, resource limitations or complications such as hindsight bias. Dictionaries can also be used as a robustness check to alternative models for measuring sentiment.

Our discussion focuses on the financial stability dictionary, which classifies words into two categories: positive or negative. We refer to these categories as the sentiment of a word, although our conclusions also apply to dictionaries that measure other concepts, such as in Dougal et al. (2012) who measure *Wall Street Journal* columnists' bullishness in financial reporting. Dictionaries are simplified models for estimating sentiment, which is often calculated as a ratio of word counts. In the case of the FSS index, the ratio is calculated as the proportion of negative to positive words in the text:

$$FSS = \frac{\#\ Negative\ Words - \#\ Positive\ Words}{\#\ Total\ Words}.$$

The online appendix in Correa, Garud, Londono, and Mislang (2017) contains the financial stability dictionary. Table 1 here presents a detailed description of the dictionary's word distribution. Our dictionary contains 96 positive and 295 negative words. When a word is not included in the dictionary, it is implicitly considered neutral in the sense that it does not convey sentiment in a financial stability context. Often, words that appear to convey sentiment are omitted from our dictionary, as their use in FSRs is overwhelmingly descriptive or part of an economic term. For example, "default" generally has a negative implication for financial stability, but most mentions of the word in FSRs occur in the neutral context of discussing "credit default swaps." We include each form of a word separately and do not use stemming, as different forms of a word may carry different meanings.[2] The online appendix also contains a list of 1093 words that we identified as potential sentiment words but that are omitted from the dictionary for being too neutral or descriptive. For applications outside of FSRs, there may be value in expanding the financial stability dictionary with these words or through stemming words in the text to be analyzed (that is, the corpus) and in the dictionary.

Table 1: Financial Stability Dictionary, Word Distribution and Frequency

	Number of words	Word distribution (percent)	Word frequency (percent)
Total financial stability	391	5.38	4.01
Positive words	96	1.32	1.45
Negative words	295	4.06	2.56
Overlap with LM	270	3.72	3.28
Uniquely financial stability words	121	1.67	0.73

Note: This table shows the distribution of positive and negative words in the financial stability dictionary from Correa et al. (2017). The word distribution shows the number of dictionary words as a percentage of all distinct words across all FSRs. The word frequency is the number of times words occur across all FSRs divided by the sum of all words across all FSRs. Also reported is a comparison between the words in our dictionary and the dictionary in Loughran and McDonald (2011) (LM in table). Uniquely financial stability words refers to words not classified in Loughran and McDonald's dictionary.

Methodology for Coding a Dictionary

To investigate the sentiment conveyed by each word, we split into two teams of two independent coders. We first omit stop words such as "and," "the," and "of." We then sort words by frequency and extract the list of words that make up the top 98 percent of the total word count (7,388 words), and we go through this list to identify words that potentially convey sentiment.[3] Next, we classify words inductively by randomly sampling relevant sentences from FSRs. In cases of disagreement, the coders discussed the sentiment conveyed by the word.

---

[2] Stemming is a common dimensionality reduction technique in textual analysis that uses an algorithm to reduce words to a shared root. For example, a stemming algorithm might reduce both "vulnerable" and "vulnerability" to "vulnerab."

[3] The remaining 2 percent of words by frequency amount to 34,579 words, of which 27,219 words are used five or fewer times in all 982 reports. Thus, the lowest 2 percent of words correspond to very specific (often regional) uses of language or are only found in few reports, making them impractical to apply to a broader financial stability context.

A well-defined criterion is important for identifying when a word reliably conveys sentiment. There are three common forms of disagreements in word classification: completeness, bias, and attribution, which we discuss in greater detail below.

Classifying Words: Ensuring Completeness

For a dictionary to be complete, it should include enough words to generate a consistent index. A more complete dictionary will produce a reliable index that should remain stable across repeated trials.

One way to test for consistency is to analyze the sensitivity of the dictionary to the random removal of words. Jegadeesh and Wu (2013) follow this procedure to show that their empirical results are robust to an incomplete dictionary. We perform a similar experiment with repeated trials of randomly removing portions of our dictionary with various intensities. Figure 1 displays the resulting 90 percent bands in our index for select countries. Even if one in five words in the dictionary are misclassified, the confidence bands remain relatively narrow and the contours of the FSS indexes are largely preserved. This evidence implies that a dictionary does not have to be comprehensive to be "complete" and reliable. This evidence also suggests that small mistakes can be made in the classification process.

Figure 1: Confidence Intervals for the FSS Index, IMF, ECB, and Selected Countries

(a) IMF

(b) ECB

(c) Sweden

(d) Korea
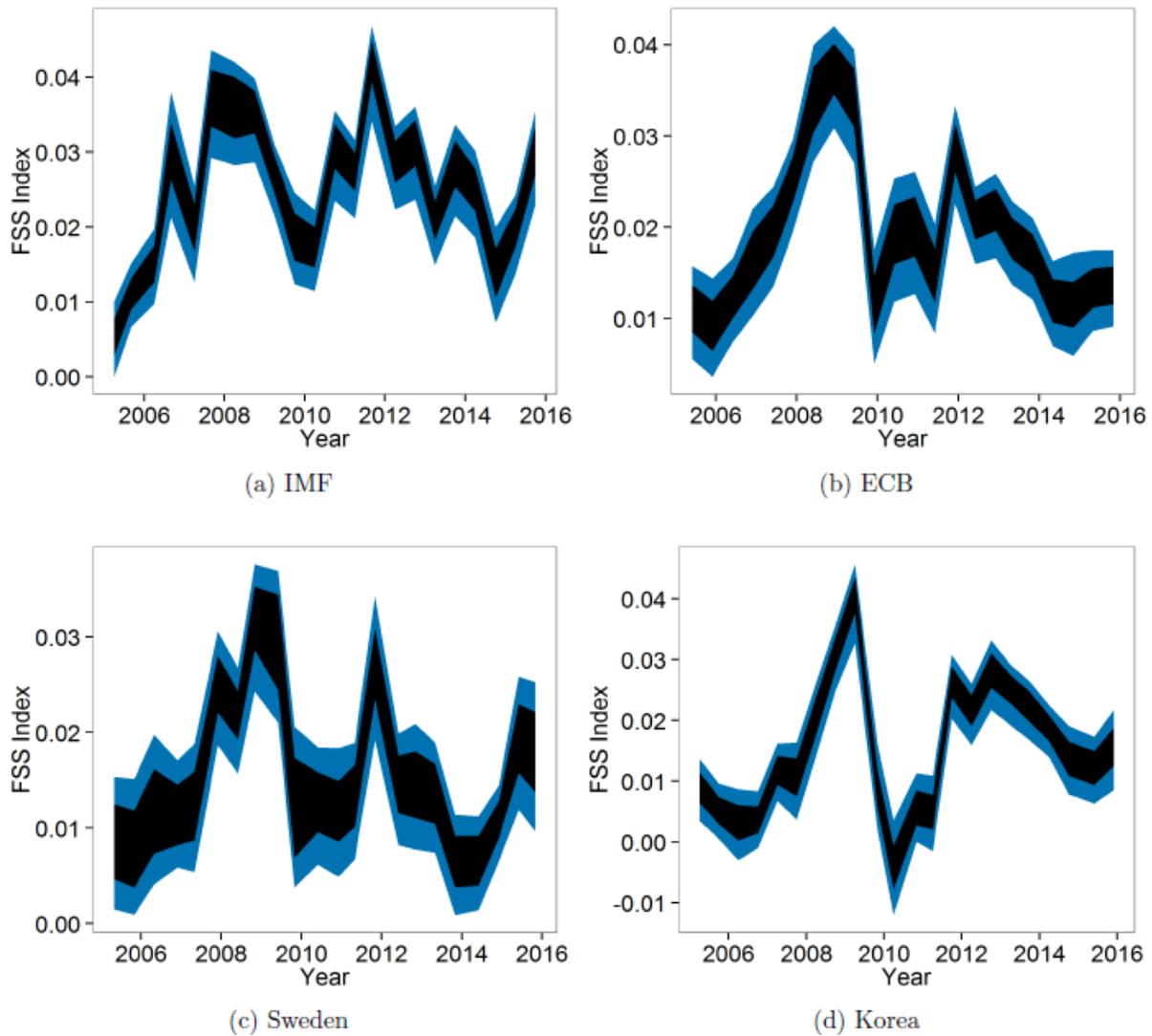
Note: This figure summarizes the results for the sensitivity of FSS indexes to the words in the financial stability dictionary from Correa et al. (2017). The shaded areas show 90 percent confidence intervals calculated by randomly removing 5 (black) and 20 (blue) percent of the words in the dictionary for selected regions (IMF and ECB) and countries (Sweden and Korea). To calculate the intervals, the process of randomly removing words from the dictionary and recalculating FSS indexes is repeated 1000 times.

The practical lesson behind this symmetry is an understanding of how words transmit information. Polarized words, which are words that consistently convey positive or negative sentiment, contribute less error to the FSS index and thus contain more information. However, words near the border between neutral and a sentiment category will always be noisy in the dictionary framework. The quality of a particular dictionary is determined by the accurate classification of these noisy words, as polarizing words often share the same classification in independently formed dictionaries.

Classifying Words: Avoiding Bias

Tailoring a dictionary to its context helps standardize many aspects of language use. However, context can be dynamic in a corpus. Substantial changes in how language is used across a corpus may cause the dictionary's measurement error to be biased. In particular, for FSRs, the time period often determines the sentiment conveyed by words. For instance, in the mid-2000s, the word "crisis" usually had a negative connotation, as it was often used to refer to the possibility of a crisis event. After the Global Financial Crisis (GFC), however, nearly all mentions of "crisis" were a reference to that event. If we were to add "crisis" in our financial stability dictionary as a negative word, our FSS index would be higher in the period since the GFC, but it would not convey information about central banks' perceptions of future crisis episodes. Therefore, in our dictionary, the word "crisis" is classified as neutral, as it conveys no sentiment.

Potential time variation in sentiment is one reason why careful thought and validation is important in dictionary formation. Because omissions are equivalent to a neutral classification, the error of biased words is unavoidable. Exploring the effect of word bias would be an interesting extension of research that implements text analysis.

Classifying Words: Attributing Tone

Attribution is the process of identifying where the tone of a sentence originates. This process is done at the sentence level, and it is how the researcher collects evidence for classifying words. Without a well-defined guideline for attribution, independent researchers may form drastically different dictionaries. A comprehensive treatment of this topic requires linguistic expertise, so we briefly describe the consensus we reached while forming the financial stability dictionary.

Our approach to attribution is to determine how each word under examination contributes to the sentiment of the sentence as a whole. If the tone of the sentence changes when the word is removed, we consider that word to convey sentiment. By considering sentiment in this way, we capture how the sentiment of each word interacts within the context in which it is used. We do not consider descriptive terms such as "unemployment" that need to be qualified to contain sentiment ("High unemployment" or "low unemployment"), except in cases where the term is systematically qualified in a certain direction. For instance, the word "delays" could be qualified, as in "delays were negligible," but it is most often used to describe unexpected setbacks in economic recovery or policy implementations. We trust the judgment of the coder and use two independent coders to reduce human errors. For simplicity, we omit words that are difficult to classify, so it is likely that further applications of the financial stability dictionary may benefit by adding words instead of removing them.

Summary and Lessons

Dictionaries that capture word sentiment allow researchers to construct indexes in cases that render other text analysis approaches infeasible. We propose three criterion that may help in the dictionary construction process. During the word classification process, it is important to ensure completeness, avoid biases, and attribute the right tone to each word. Additionally, dictionaries should be context-specific and researchers should always consider the original intent of the dictionary. Using the financial stability dictionary from Correa, Garud, Londono, and Mislang (2017) as an example, we highlight challenges in the construction and application of text analysis dictionaries. Taking these factors into account will facilitate the development of more robust dictionaries and associated sentiment indexes.

References

Correa, Ricardo, Keshav Garud, Juan M. Londono, and Nathan Mislang (2017). Sentiment in Central Banks' Financial Stability Reports. International Finance Discussion Papers 1203.

Dougal, Casey, Joseph Engelberg, Diego Garcia, and Christopher A. Parsons. Journalists and the Stock Market. Review of Financial Studies 25, no. 3 (2012): 639-679.

Henry, E., Leone, A. J., 2016. Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. Accounting Review 91, 153-178.

Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. Journal of Financial Economics, 110(3), 712-729.

Laeven, L., Valencia, F., 2013. Systemic banking crises database. IMF Economic Review 61, 225-270.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance 66, 35-65.

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. (2010). Construct validation using computer-aided text analysis (CATA): An illustration using entrepreneurial orientation. Organizational Research Methods, 13, 320-347.