WP/20/7

IMF Working Paper

How Do Member Countries Receive IMF Policy Advice: Results from a State-of-the-art Sentiment Index

by Ghada Fayad, Chengyu Huang, Yoko Shibuya, and Peng Zhao

INTERNATIONAL MONETARY FUND

 WP/20/7

IMF Working Paper

Strategy, Policy, and Review Department

How Do Member Countries' Receive IMF Policy Advice: Results from a State-of-the-art Sentiment Index[1]

Prepared by Ghada Fayad, Chengyu Huang, Yoko Shibuya, and Peng Zhao

Authorized for distribution by Rupa Duttagupta

January 2020

Abstract

This paper applies state-of-the-art deep learning techniques to develop the first sentiment index measuring member countries' reception of IMF policy advice at the time of Article IV Consultations. This paper finds that while authorities of member countries largely agree with Fund advice, there is variation across country size, external openness, policy sectors and their assessed riskiness, political systems, and commodity export intensity. The paper also looks at how sentiment changes during and after a financial arrangement or program with the Fund, as well as when a country receives IMF technical assistance. The results shed light on key aspects on Fund surveillance while redefining how the IMF can view its relevance, value added, and traction with its member countries.

JEL Classification Numbers: F33, F42, C88

Keywords: IMF, Surveillance, Economic Policy, Sentiment Analysis, Natural Language Processing

Author's E-Mail Address: GFayad@imf.org; CHuang@imf.org

[1] We are grateful to Sanjaya Panth, Rupa Duttagupta, Daria Zakharova, Ashvin Ahuja, Jeta Menkulasi, Yongquan Cao, Diane Kostroch, Yang Liu, Nujin Suphaphiphat, Alberto Sanchez, and Grace Zimmerman for useful comments, and to Jelle Barkema and David de Padua for excellent research assistance. We also thank several departments at the Fund, namely Fiscal Affairs, Information Technology, Institute for Capacity Development, and Secretary's Departments for facilitating data access, the Knowledge Management Unit for access to the Enterprise Business Vocabulary, and the Office of Risk Management for sharing earlier internal work on sentiment.

CONTENTS

Abstract 2

I. INTRODUCTION 4

II. LITERATURE REVIEW 6
A. Sentiment Analysis Literature 6
B. Traction Literature 7

III. SENTIMENT ANALYSIS 8
A. Dataset and Paragraph Extraction 8
B. Topic Assignment 10
C. Sentiment Estimation 15

IV. PRELIMINARY LOOK AT SENTIMENT INDEX 18

V. ANALYSIS OF SENTIMENT INDEX 22
A. Data 22
B. Regression Analysis 23
C. Applying Model on Executive Directors' Buff Statements 33

VI. CONCLUSION 34

FIGURES
1. Examples of Extraction Rules 9
2. Number of Paragraphs and Countries in the Sample 10
3. Word2vec Visualization 11
4. Confusion Matrix for Topic Assignment 13
5. Topic Composition for Whole Sample and Time-series 14
6. Supervised Learning Overview 16
7. Confusion Matrix for Sentiment Estimation 18
8. Sentiment Composition for the Whole Sample and Time-series 19
9. Average Sentiments Across Income Group Over-time 20
10. Average Sentiments Across Income Group Over-time 21
11. Average Sentiments Across Sectors and Income Groups (2000–18) 22
12. Average Sentiments and Commodity Prices 26
13. Sentiments Across Program Types 29
14. FSAP Effects on Average Sentiments 31
15. Technical Assitance Effects on Average Sentiments 31

TABLES

1. Example of Topic Assignment ____ 12
2. Precision, Recall, and F-score for Topic Assignment ____ 13
3. Precision, Recall and F-score for Sentiment Estimation ____ 18
4. Regression Results for Country Characteristics ____ 27
5. Regression Results for Sectoral Risks ____ 28
6. Regression Results for Fiscal and Monetary Technical Assitance ____ 33
7. Regression Results for Sectoral Crisis Risks ____ 33

APPENDICES

I. Examples of Annotation ____ 35
II. BERT Model in Detail ____ 37

REFERENCES

References ____ 40

I. INTRODUCTION

IMF bilateral surveillance, as practiced through its Article IV consultations, involves continuous monitoring of, and offering advice on, member countries' economic and financial policies.[2] Discussions with country authorities typically cover the macroeconomic situation, the prevailing policy stance, the effects of these on the economy's macroeconomic and financial stability, and the desirable policy adjustments to sustain or strengthen stability. The discussions are concluded after the IMF's Executive Board has considered and endorsed the country report prepared by staff for the Board's consideration. The final published Staff Report for each country includes Fund policy assessments and advice, authorities' views, as well as the views of the Executive Director (ED) representative of that country in a separate "Buff" statement.

While staff report on the status of their recommended policies for member countries, thus far, there is no comprehensive assessment of authorities' overall reception of Fund policy advice, regardless of whether such recommended policies are implemented or not. This paper fills this gap by building the first comprehensive measure of IMF member countries' initial reception of Fund policy assessments and advice over 2000–18, using latest techniques in natural language processing (NLP), and differentiating this reception or "sentiment" across countries' income groups, policy areas, and over time. By doing so, it sheds light on a key dimension of Fund traction with its member countries: how Fund engagement with countries during the Article IV policy consultation cycle influences authorities' views on policy matters.[3] Our paper is thus linked to two strands of literature, the one on traction in the political sciences and the more technical one on sentiment analysis across many fields as discussed in the next section.

The task we set out to do addressed challenges on several fronts. It involved, first, putting together a novel rich dataset from 2000-2018 comprising information on authorities' views in IMF Article IV Staff Reports for all member countries, then, assigning five key macroeconomic topics—fiscal, monetary, financial, real/structural and external—to the extracted paragraphs, and finally training a deep learning model to recognize the nature of sentiments from the description of authorities' views. Our topic model, which combined several techniques, was able to assign the correct topics 89 percent of the time, while our trained deep learning model was able to estimate the correct sentiment (as labeled by the

[2] While engagement with member country authorities is continuous, the Article IV Consultations are done on an annual basis for most member countries as per Article IV of the Articles of Agreement of the IMF with its member countries.

[3] Other important dimensions of Fund traction such as implementation of Fund policy advice is beyond the scope of this paper.

team in a test set) 81 percent of the time, both suggesting very high performance of the model in relation to the related literature.[4]

Our findings suggest that authorities have generally appreciated or had "positive" initial reactions to Fund assessments and policy advice. That said, there are several differences across, time, countries, and issues. We find that on average countries agreed with Fund advice 75 percent of the time, although agreement is relatively lowest in advanced economies, compared to emerging markets and low-income countries (and conversely highest in the last group). Across policies, average sentiments are higher for fiscal, financial and real sectors compared to monetary and external sectors.

We also use panel regression analysis to dig deeper into the relationship between sentiment and the different layers of country structural characteristics and nature of engagement with the Fund. We find that average sentiments are higher in countries with lower IMF quota, smaller economies, those with relatively less open capital accounts, and in governments with more political power and those with more years remaining in office. Commodity exporters' sentiments towards Fund advice, and particularly since the global financial crisis, have moved inversely with commodity price changes, i.e., commodity price decreases tend to be associated with positive sentiment toward Fund advice.

When we look at Article IV consultations with countries that also have IMF-supported financial arrangements, we find that reception of AIV advice is higher during the program period compared to before or after. Reception of Fund fiscal and monetary technical assistance and undergoing financial sector assessment program (FSAP) missions help improve authorities' overall responsiveness to Fund advice. Finally, with risk assessment at the center of the IMF's surveillance mandate, we use internal IMF measures of countries' underlying vulnerabilities in fiscal, external, and financial sectors to gauge how overall sentiment changes with these measures. We find that country authorities are more likely to agree with IMF policy advice when Staff Reports highlight fiscal risks, and less so when Staff Reports point to financial sector risks.

An interesting application of our model, as well as a robustness check, is to run the trained sentiment model on Executive Directors' Buff statements, which are another illustration of authorities' views during Article IV consultations. Our findings are maintained and. in fact, model performance improves, a likely reflection of the higher degree of candor in Buff statements. Overall, our work provides a way forward for Fund staff to systematically track the authorities' overall take on the quality of Fund's engagement with them and use this information to assess the underlying factors and strengthen their engagement strategy in areas where sentiments are less receptive, to gain greater traction of Fund advice.

[4] Our sample consists of just the authorities' views' paragraphs without any information on which country-year report these views are coming from, nor on the context (what specific policy advice was given at the time).

The rest of the paper is structured as follows. Section II reviews the literature and places the important contribution of this paper in perspective. Section III explains how the database was put together, how topic assignments were made, and the sentiment index was built. Section IV documents the sentiment index. Section V discusses the regression results that relate the sentiment analysis to countries' structural and economic conditions, and Section VI presents the robustness checks. Section VI presents our main conclusions from this work.

II. LITERATURE REVIEW

Our paper is related to two strands of literature. The first is from the political science literature on traction or relevance of international organizations such as the IMF. The second is on sentiment analysis which has been used extensively in social science fields. In this section, we briefly review the two literatures.

A. Sentiment Analysis Literature

(i) Finance

The research area that uses sentiment analysis technique most frequently would be Finance. The effects of announcements by the Federal Open Market Committee (FOMC) have been analyzed extensively using sentiment analysis due to the importance of sentiment in the announcements on stock prices. The method used most in the field is (refined) **dictionary approach**.

Lucca and Trebbi (2009) construct sentiment score using the content of FOMC announcements to predict fluctuations in treasury securities. To do this, they use dictionary-based methods: Google and Factiva semantic orientation scores. In the Google/Factiva score, they count how many Google/Factiva search hits occur when searching for phrases plus one of the words from a list of antonym pairs signifying positive or negative sentiment.

Born et al. (2014) extend this idea to study the effect of central bank sentiment on stock market returns and volatility. They construct a financial stability sentiment index from Financial Stability Reports and speeches given by central bank governors. Their approach uses a sentiment dictionary to assign optimism scores to word counts from central bank communications. Most recently, Shapiro and Wilson (2019) propose another dictionary approach to estimate the sentiment in FOMC informal discussion notes. They measure "net negativity" in the discussions based on the use of negative and positive words. The classification of negative/positive words is done by the economics/finance-specific dictionaries of positive and negative words developed by Loughran and McDonald (2011). The dictionary contains thousands of words and includes both common-language terms and terms that are specific to economics and finance. They also use a popular open-source python tool called VADER (Valence Aware Dictionary and Sentiment Reasoner) (Hutto, C., and E. Gilbert (2014)) that allows them to construct an alternative negativity measure.

(ii) Economics

Sentiment analysis has been used in Economics too. Tetlock (2007) employs a dictionary approach to analyze the latent "sentiment" of Wall Street Journal columns, defined along with a number of dimensions such as "positive," "optimistic," and so on. The author used a dictionary called the General Inquirer from the Harvard IV-4 psychosocial dictionary, which provides lists of words associated with each of these sentiment categories. Some of the other papers employ more **statistics-oriented methods** for sentiment analysis. Chinco et al. (2019) apply Least Absolute Shrinkage and Selection Operator (LASSO) in high frequency stock return prediction using pre-processed financial news text sentiment as an explanatory variable. They emphasize the success of LASSO in the out of sample predictions.

(iii) Other fields

Text analysis technique has also been used in other social science fields. Evans and Aceves (2016) and Grimmer and Stewart (2013) offer a comprehensive survey in sociology and political science respectively. Evans and Aceves (2016) survey text mining methodology and provide recommendations on how it can be used as a tool for theory generation in the social theory. Grimmer and Stewart (2013) explain potential pitfalls of using text-mining methods in social sciences.

In academic social science fields, dictionary approach or traditional machine learning methods are still dominant. However, for this project estimating agree/disagree sentiments in IMF Article IV Staff Reports, neither method work. The reason is simple: neither model can take into account paragraph structure/context. As noted in Pang et al (2002), the difficulty in sentiment analysis lies in the structure of the paragraph. When people express a disagreeing sentiment, they often show some small agreement first and then show disagreements, which makes the overall paragraph more nuanced. A human can easily tell that the overall sentiment of the paragraph is disagree, but dictionary or traditional machine learning methods often mislabel those paragraphs since they only count the number of agree/disagree words in the paragraph. In this project, we employ state-of-the-art deep learning model (BERT model) that no longer counts the number of agree/disagree words but takes into account the structure of the paragraph.

B. Traction Literature

The political science literature offers three ways (with associated indicators) through which the performance or effectiveness of international organizations (IOs) can be evaluated from the initial stage of policy formulation to the final stage of problem resolution : (i) "output" (process-based performance indicators reflecting the effort to change behavior); (ii) "outcome" (implementation of the IO's advice as a result of behavioral change by targeted actors); and (iii) "impact" (achievement of policy action) (Gutner and Thompson, 2010). In practice, studies assessing the performance of IOs employed "output" indicators given inherent difficulties in measuring "outcome" and "impact". Even though a policy "output"

would not be a sufficient indicator of whether an IO is a successful problem solver, it is a necessary first step in studying the performance of an IO (Tallberg et al 2016).

The definition of Fund traction in its periodic surveillance reviews mainly relied on *authorities' response to past policy advice* whereby reporting on authorities' implementation of past Fund advice is a requirement for Fund IV Staff Reports. More recently, there has been greater recognition of the importance of *the extent to which the authorities engage with the Fund on its advice*, and on *enhancing the policy dialogue with the authorities.* Empirically, internal studies on Fund traction have mostly relied on stakeholder surveys or targeted interviews, while external assessments of Fund traction mainly consisted of case studies (Momani (2006) and Edwards and Senger (2015) on Canada and U.S. respectively) that found limited traction of Fund surveillance, and surveys with more positive findings on Fund traction in Custer et al. (2018).

This paper is the first comprehensive assessment of member countries' reception of Fund policy advice over the last two decades. It applies sentiment analysis on authorities' views in Article IV Staff Reports using latest deep learning techniques. In doing so, it significantly broadens both strands of literature with important implications for Fund surveillance going forward. The sentiment index we develop is easily computable and monitorable across time, countries and policy areas, and can serve as the important starting point for systematic and comprehensive monitoring of Fund traction in the future.

III. SENTIMENT ANALYSIS

This section discusses A) the construction of the dataset on authorities' views; B) how topic assignments were made, and C) the inputs in building the sentiment index. Readers who are keener on the results of the diagnosis can skip this section and move to Section IV.

A. Dataset and Paragraph Extraction

We use Article IV Staff Reports as an input of the deep learning model. Our database, which is the first comprehensive repository of IMF Article IV Staff Reports, includes about 2600 reports for all member countries from 2000–18.[5] The sample also includes reports that are combined Article IV consultation and program review.[6] This allows us to see how

[5] Our AIV SRs come from two sources; (i) 73 percent come from IMF Communications department (COM) database and (ii) the rest come from the Institutional Repository. Article IVs from COM database are publicly available, while those from Institutional Repository include *for official use only* items.

[6] Countries that enter into a financial arrangement with the Fund are still required to go through the AIV consultation process but the AIV cycle in that case will differ from the standard twelve-months cycle of surveillance countries, depending on type of arrangement and on whether program is on trach or not. Though missions and corresponding reports may be separate, Article IV consultations are often combined with use of fund resources papers (program request and program reviews).

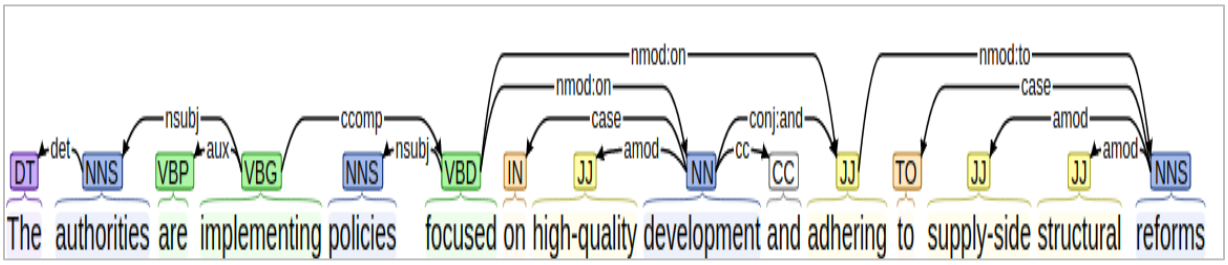
authorities' sentiment toward Fund advice evolves as a country moves in and out of Fund programs.

To be able to construct an authorities' sentiment index, we had to extract relevant paragraphs from Article IV Staff Reports. Here "relevant" paragraphs refer to paragraphs that contain authorities' views toward Fund policy assessments and advice. For the well-structured Staff Reports with sections and headers, we extracted paragraphs from the section called "Authorities' views", which contain all the authorities' views toward Fund advice. For the less structured reports, we had to come up with a systematic extraction rule of authorities' views.

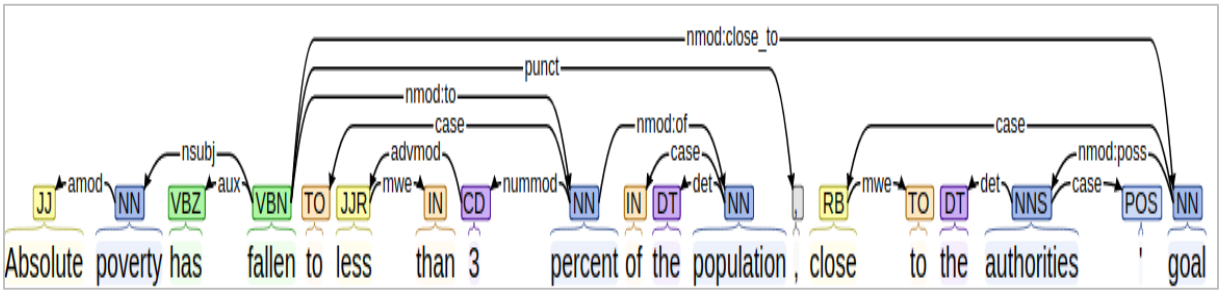
The extraction rule we used is to extract those paragraphs where the word "authorities" used as a **Noun Subject**. We give two example sentences below. The first sentence is the case where the word "authorities" is used as a Noun Subject, i.e., the authorities are expressing their views. We want to include theses paragraphs in our sample. On the other hand, the second example sentence is the case where the word "authorities" is used as a **Noun Modifier**, i.e., the authorities are the subject in the Fund staff's views. We do not want to include theses paragraphs (Figure 1).

Figure 1. Examples of Extraction Rules

Authorities as the "Noun Subject"—include



Authorities as the "Noun Subject"—exclude

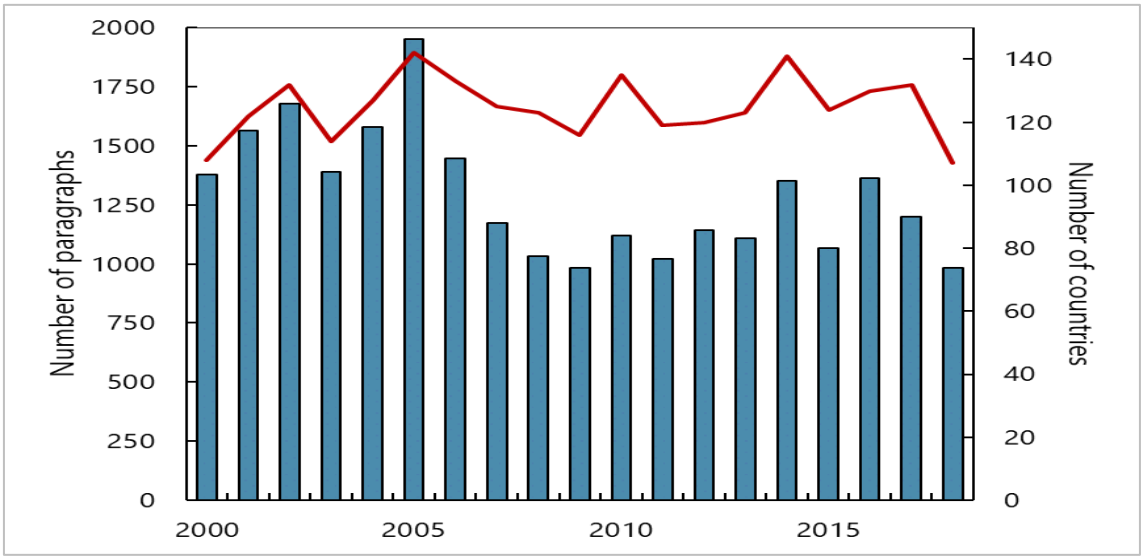


Source: IMF Staff Reports; dependency parsing, and visualization is made using spaCy2 (Honnibal and Montani 2017)

The number of paragraphs and countries in each year in our sample is shown Figure 2. The blue bars show the number of extracted paragraphs and red line shows the number of countries in our sample. There is a time variation in both number of paragraphs and countries, but the number of paragraphs ranges from 1000 to 2000, and the number of

countries lies in the range of 100 to 150. The time variation is partially explained by the fact that the official restriction on number of pages in Article IV Staff Reports was introduced in early 2000s. The restriction makes the sample smaller in recent years. When we conduct an analysis in section V, we present estimated sentiments over year, country, topic pair, i.e. averaging the multiple paragraphs per sector, so the time variation in number of paragraphs should not affect much for our main results.

Figure 2. Number of Paragraphs and Countries in the Sample



Source: IMF Staff Reports and authors' calculations

B. Topic Assignment

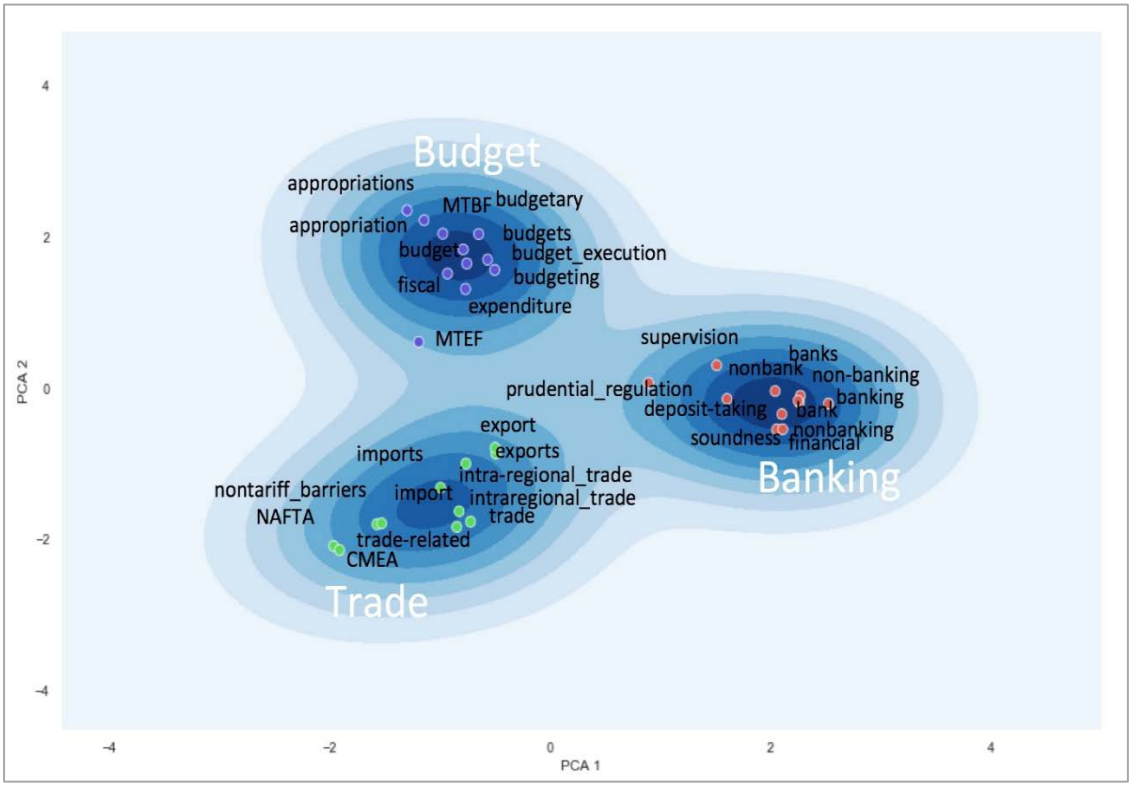
This section explains how we assigned a *topic* to each paragraph. For the extracted paragraphs from Article IV Staff Reports, we assigned one of the following five topics: external, financial, fiscal, monetary and real sectors. We used dictionary method for the topic assignment. We first made a dictionary of sector-specific words and then applied the dictionary to choose a topic for each paragraph by the frequency of sector-specific words.

To make sector-specific vocabulary list (dictionary), we first tapped into IMF's enterprise knowledge that was built for in-house information retrieval, known as the Enterprise Business Vocabulary (EBV).[7] It contains a six-layer knowledge tree, with each root of the tree being one of the five sectors we want to label, and each node being a subcategory of its parent node. For instance, for monetary sector, we have sub-categories like unconventional monetary policy, inflation targeting etc.

[7] Developed by the Knowledge Management Unit at the Fund.

While the knowledge graph gives us a granular level of what sub-topics that are discussed in each sector at the IMF, these concepts are sometimes abstract, and their exact wordings often do not appear in Staff Reports. To identify the exact phrases, we trained a word2vec model (Mikolov et al., 2013) on IMF documents, and use it to find the top 10 most similar phrases for each sub-category.[8] We then manually went through the extracted phrases to remove irrelevant terms. Word2vec is a technique to represent words in a vector space. It takes a large list of words as an input and produces a few-hundreds dimensioned vector space, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located close to one another. For example, for "unconventional monetary policy" which is an item of the EBV list, word2vec allows us to identify associated words like open-market operation, negative interest rate, quantitative easing, etc (Figure 3).

Figure 3. Word2vec Visualization



Source: IMF Staff Reports and authors' calculations; model is trained using "gensim" (Rehurek and Sojk 2010).

The last step is to simply count the matching phrases or number of sector-specific words used in each paragraph and assign the topic that had the largest number of such words.

[8] Our word2vec model is trained on all externally published IMF documents. We trained the model for 160 iterations with some commonly recommended hyperparameters (300 dimensions and windows size of 5).

Below is an example of a paragraph that has no sector-specific words for external, fiscal and real, one for financial and five for monetary, and was thus assigned monetary as a topic for the paragraph (Table 1). This method was applied to assign topics to the whole sample paragraphs.

Table 1. Example of Topic Assignment

Sectors	External	Financial	Fiscal	Monetary	Real
Keywords		Macro-prudential		QE, inflation target, central bank, BoJ, deflation	
Size	0	1	0	5	0

Source: IMF Staff Reports and authors' calculations

To resolve the "ties" in the above counting approach i.e. when two sectors have equal word counts (10.8 percent of all paragraphs), we train a Support-Vector-Machine (SVM)[9] model by taking advantage of some "natural labels" embedded in some clear subtitles in Buff documents (like "Fiscal Issues").[10] Theoretically, such ML model could offer superior performance by allowing the model to learn different weights for different words/phrases from labels, as well as the interaction among them. For example, although a paragraph in which "interest rate" and "government debt" co-occur may sound like "fiscal policy", it is more likely to be about "quantitative easing" in "monetary policy" if "central bank" is also present.

Model performance

To test the performance of our dictionary method, we manually assigned topics to 200 paragraphs randomly drawn from whole sample.[11] We then compared the assigned topics by our dictionary method to the 'true' topics assigned by a human. Our method achieved 88.8 percent accuracy, i.e., out of 200 paragraphs 88.8 percent of them are assigned correct topics. Figure 4 shows the confusion matrix of topic assignment. We can see from diagonal matrix that, for each sector, we achieved higher than 83 percent precision. Table 2 shows precision (the ratio of correctly predicted observations to the total predictions), recall (the ratio of correctly predicted observations to the all observations in actual class) and F-score

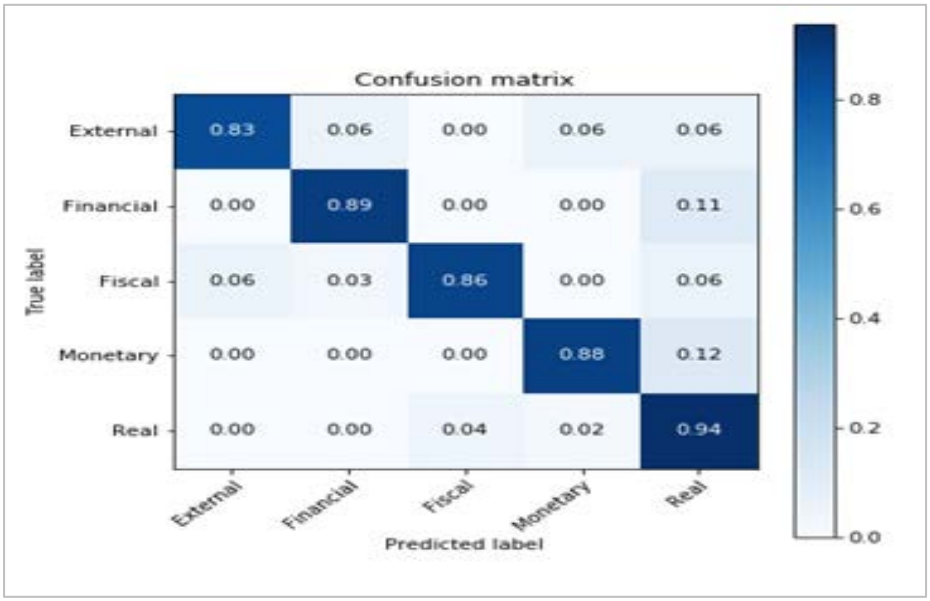
[9] SVM is a supervised machine learning model for classification tasks (in our setting). Compared to linear classification models, SVM can efficiently perform non-linear classification using kernel trick.

[10] Contrary to authorities' views paragraphs in SRs, Buffs have subtitles on topics covered in each paragraph. We use those to train the SVM.

[11] We read 200 paragraphs and determined what kind of topics are discussed.

(combining precision and recall). All index shows high performance of our topic assignment method.

Figure 4. Confusion Matrix for Topic Assignment



Source: IMF Staff Reports and authors' calculations.

Table 2. Precision, Recall, and F-score for Topic Assignment

	External	Financial	Fiscal	Monetary	Real
Precision	0.88	0.95	0.94	0.88	0.81
Recall	0.83	0.89	0.86	0.88	0.94
F-score	0.86	0.92	0.9	0.88	0.87

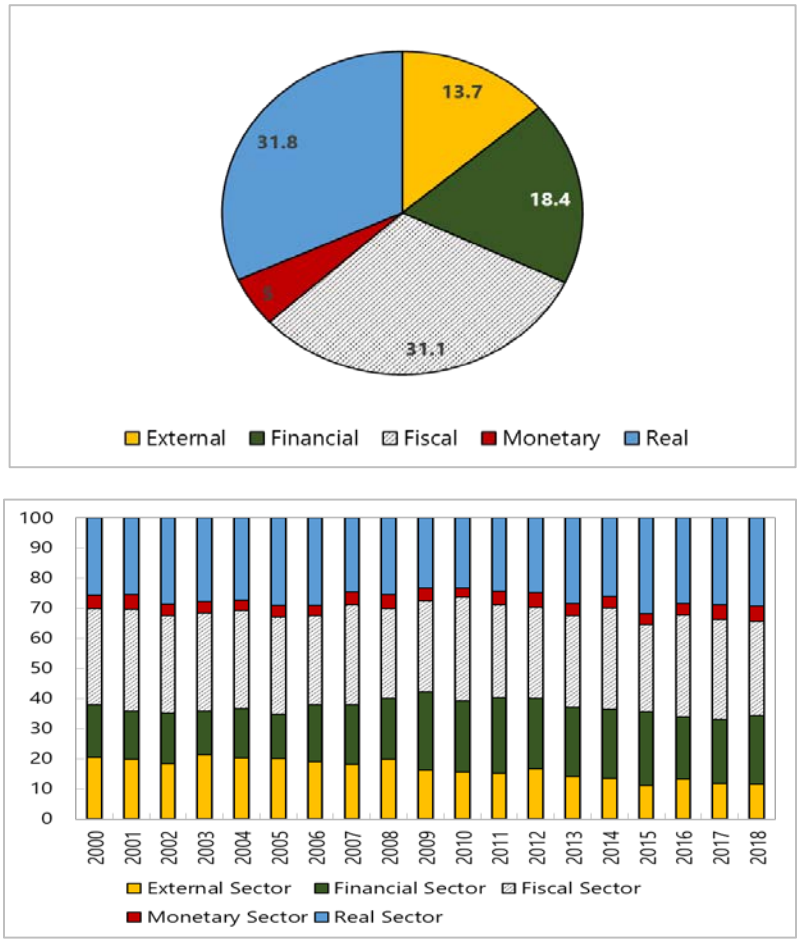
Source: IMF Staff Reports and authors' calculations.

First look at the data

We applied the dictionary method to our whole sample. Figure 5 shows the resulting topic composition of our whole sample. The two largest sectors in authorities' views in Article IV Staff Reports were the fiscal sector and the real sector (both 31 percent). The next largest are financial (18 percent), external (13 percent), and the smallest sector by far is monetary (5 percent). While it is true that, unlike other policy areas, monetary policy is only available as a tool for a subset of countries, this does not explain the relatively low coverage, as when we exclude countries in a monetary union, the topic coverage barely increases.

The topic composition varies over time, especially for the financial sector which has expanded from mid 2000s to 2015, and the external sector which has been on a decreasing trend instead. Fiscal and real are consistently the two largest sectors in any year of our sample, and monetary is consistently the smallest. The relatively steady topic composition over time is a reflection of the standard format of IMF Article IV Staff Reports whereby coverage of those five sectors, and therefore of authorities views on each, is expected/required.

Figure 5. Topic Composition for Whole Sample and Time-series



Source: IMF Staff Reports and authors' calculations.

C. Sentiment Estimation

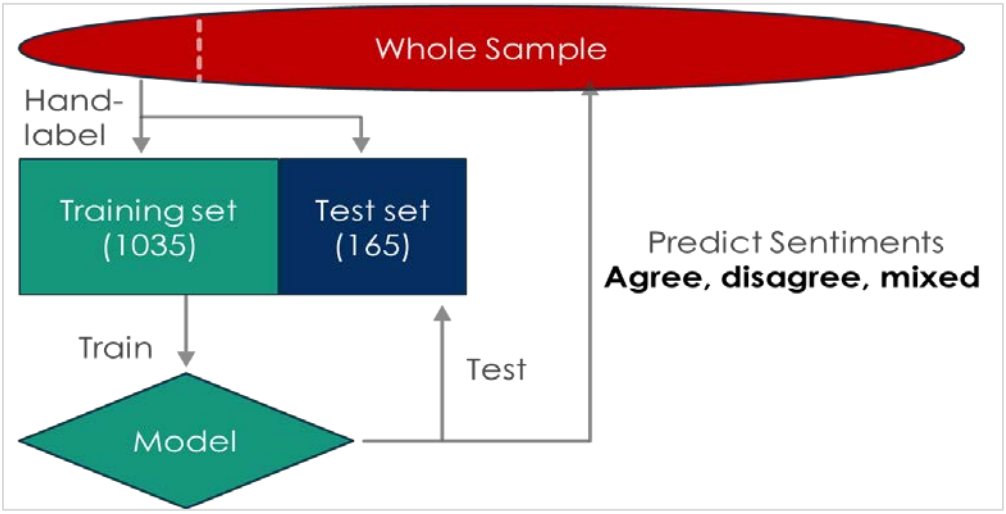
In this section, we describe the steps leading up into our sentiment index. The index captures three levels of sentiments: agree, disagree, and mixed or partial agreement.[12] The sentiment model will in the end be able to assign a sentiment to all extracted authorities' views paragraphs from Article IV Staff Reports. In other words, the paragraphs will be an input into the supervised deep learning model and the model estimates the sentiments for each in the form of a sentiment index. Figure 6 explains how supervised learning works. We randomly draw a small fraction (1200 paragraphs) from the whole sample and hand-label sentiments for those paragraphs. Hand-labeling means that we read the authorities' views paragraphs and determine what kind of sentiment the authorities are expressing there. This is more often than not a difficult task given that the paragraphs summarizing authorities' views can often be significantly nuanced. This is particularly the case for the disagree and mixed categories where paragraphs usually start with broad agreement and end up detailing reasons for disagreement. In other cases, the words agree and disagree are never used.

To make sure our labeling is consistent (across paragraphs and team members), we came up with annotation rules detailing the conditions under which a particular sentiment is assigned for the three categories, and supplemented the rules by examples for each case.[13] We divided the 1200 hand-labeled paragraphs into a training set (1035 paragraphs) and a test set (165 paragraphs). The deep learning model leans how to assign sentiments from the training set, and the test set gives us an idea of how well the model learned and thus performed. If the model performance is good enough, we apply the model to whole sample, and our sentiment index is formed.

[12] We assign a value of +1 to agree, -1 to disagree and 0 to partially agree.

[13] Please see appendix A for more details on our annotation rules.

Figure 6. Supervised Learning Overview



Source: Authors' calculations.

There are two key sources to model success. The first key is precision of hand-labeling. The deep learning model leans how to assign sentiments from training set. If the sentiment labeling in the training set is not precise, the model cannot learn properly. Therefore, and as mentioned before, consistency in hand labels for both training and test set cannot be emphasized enough. All team members hand-labeled the 1200 paragraphs and whenever there were disagreements on how to assign sentiments to certain paragraphs, we discussed until we reached a conclusion.

The second key for the model success is, it goes without saying but, model quality itself. The model has to learn how to assign sentiments to the whole sample from the limited size of the training set. In this paper, we adopt a state-of-the-art NLP solution using the BERT (Bidirectional Encoder Representation from Transformers) model developed by Devlin et al. (2019) from Google AI.

Compared to traditional machine-learning models, BERT's key innovation is in its ability to understand context, i.e. to embed each word as a vector of real numbers *based on its context*, which is key for our sentiment analysis. As mentioned in the literature review, agree/disagree sentiment does not necessarily depend on the frequency of agree/disagree words appearance but the context. This implies several clear advantages well-suited for our task, including:

1. BERT can learn subtleties that are beyond naïve counting of bag of words. For example, authorities may explicitly express agreement but emphasize challenges in timely implementation in many ways. It is very costly to pre-define a dictionary or set of rules that capture all the variations.

2. BERT can take into account the tokens' positions, as well as their long-term dependencies. For example, disagreement is usually preceded by agreement in the paragraph. The model can learn to assign higher weights to the latter half of the paragraph in this case.

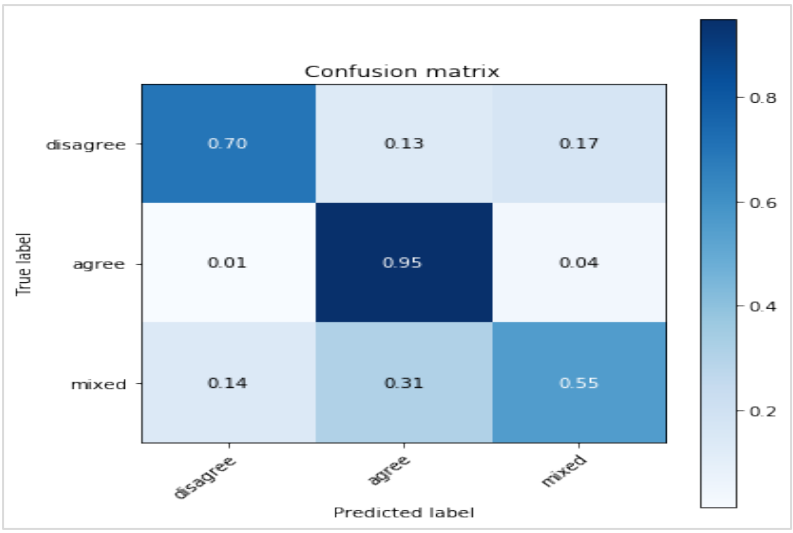
3. BERT can generalize from a relatively small training set like ours. After being pre-trained by Google on BooksCorpus (800M words) and English Wikipedia (2,500 M words), the model already comes with a strong "knowledge" on contextualized embedding. It then only needs to be "fine-tuned" on our small training set to learn a task on sentiment analysis and tap into its "knowledge" to generalize to similar situations.

With these advantages, BERT marked a ground-breaking milestone in the NLP deep learning field. Since its release, it has been widely adopted, and has consistently set new records in most NLP tasks including sentiment analysis.

Model performance

We used 165 paragraphs to compare the assigned sentiments by our supervised deep learning model to true sentiments assigned by us. Our model achieved 81.3 percent accuracy, i.e., out of 165 paragraphs 81.3 percent of them are assigned correct sentiments. Figure 7 shows the confusion matrix of sentiment estimation. The model did an amazing job in predicting agreement (with 95 percent precision), and a good job in predicting disagreement. The model got confused when the true sentiment is mixed, which is natural since mixed contains both agreement and disagreement in one paragraph. Table 3 shows precision, recall and F-score for the sentiment estimation. High precision in any category of sentiments is a good sign for our later analysis. Recall for mixed is very low, as we could have already seen in the confusion matrix. We have to keep in mind that actual mixed paragraphs are more likely to be mis-labeled as agree than disagree, which means a slight upward bias in estimated sentiments.

Figure 7. Confusion Matrix for Sentiment Estimation



Source: IMF Staff Reports and authors' calculations.

Table 3. Precision, recall and F-score for sentiment estimation

	Disagree	Agree	Mixed
Precision	0.81	0.85	0.67
Recall	0.7	0.95	0.55
F-score	0.75	0.9	0.6

Source: IMF Staff Reports and authors' calculations.

Another important evaluation criterion is that the model does not flip between disagree and agree. The most dangerous mistake that a model can make that affects our later analysis result is to mis-label between disagree and agree paragraphs. Our model did a pretty good job in that evaluation criteria: only 5 out of 165 paragraphs are mis-labeled between disagree and agree. That means most of the mistakes are minor ones: between mixed & agree or mixed & disagree.

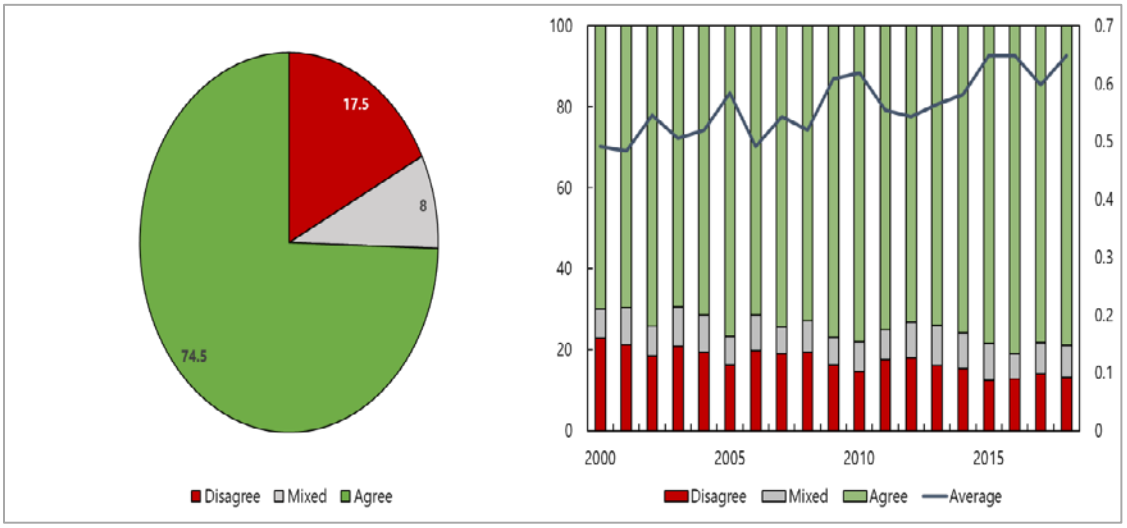
We are now ready to apply the model to the whole sample and produce the sentiment index for authorities' views in Article IVs from 2000–18.

IV. Preliminary Look at Sentiment Index

We now investigate how our sentiment index changes across country groups, time, and policy sectors. Figure 8 shows the sentiment composition for the whole sample during 2000–18. About 75 percent of the paragraphs in our sample show agreement with Fund advice, 18 percent show disagreement and 8 percent are mixed. To look at changes over time, we plotted average yearly sentiments: we first averaged sentiments within country-year pairs and then averaged over countries to eliminate the effect of outlier cases that have huge

number of paragraphs in one year. Time-series average sentiments show slightly increasing trend over time with some time variation, i.e., member countries are reacting more favorably to Fund advice with time.

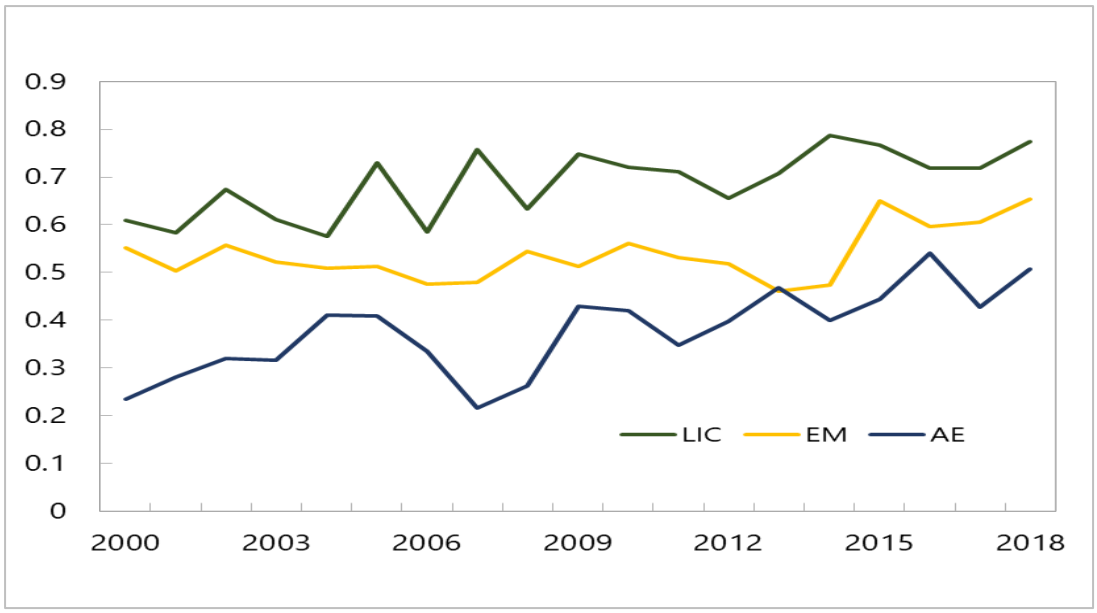
Figure 8. Sentiment Composition for the Whole Sample and Time-series



Source: IMF Staff Reports and authors' calculations.

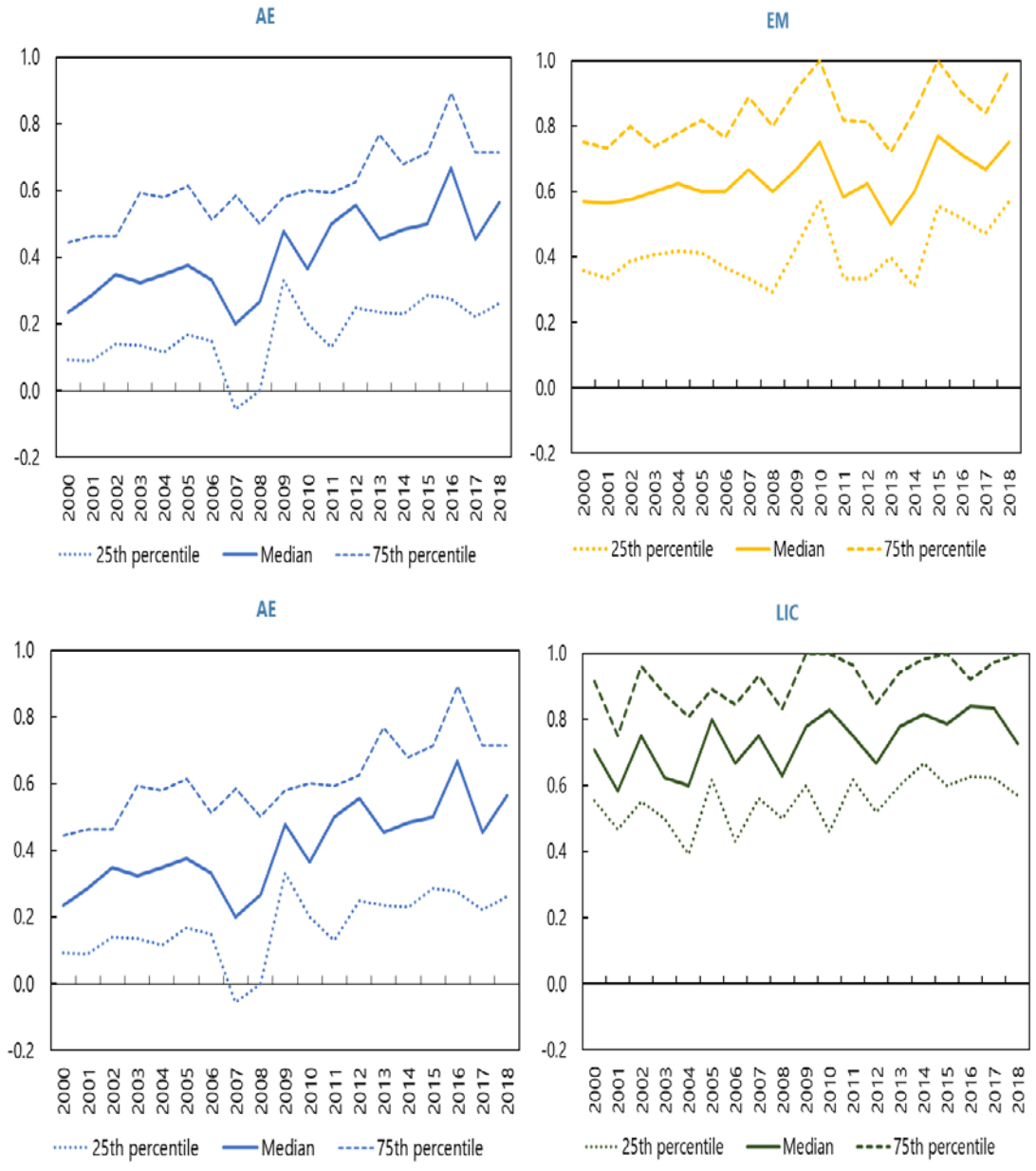
Figure 9 plots average sentiments across three income groups: low income countries (LICs), emerging markets economies (EMs), and advanced economics (AEs) over time. The pattern is striking. The higher the income level is, the lower the average sentiment is. However, averaged data masks important country variations over time. In Figure 10, we plot the interquartile ranges for sentiments across the three income groups, where it becomes clear that there are many instances of AEs' sentiments exceeding those of EMs' and LICs'.

Figure 9. Average Sentiments Across Income Group Over Time



Source: IMF Staff Reports and authors' calculations.

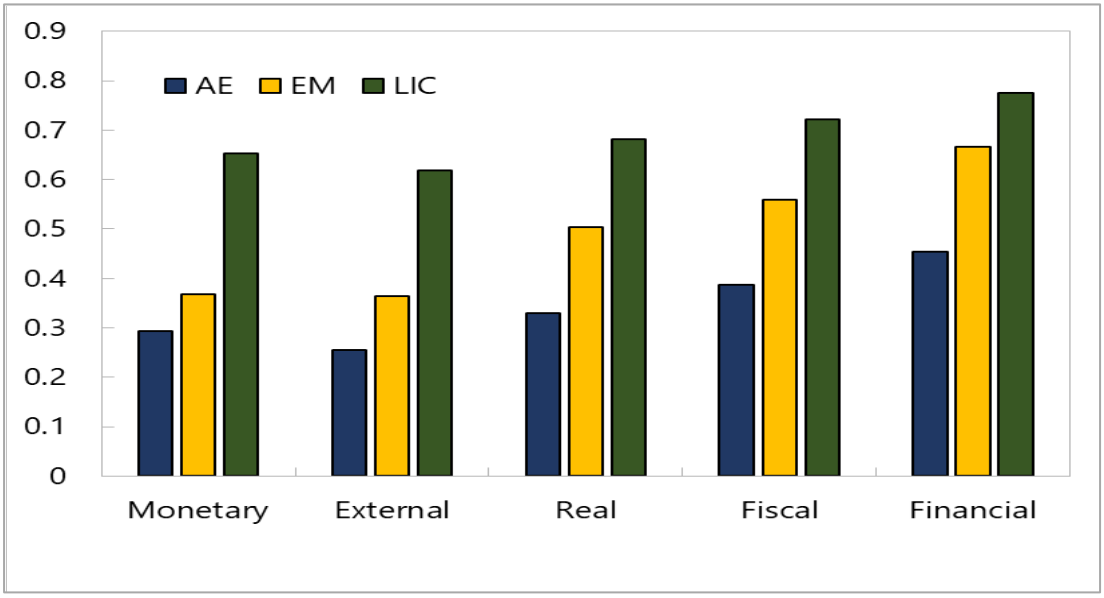
Figure 10. Average Sentiments Over Time – Interquartile Range



Source: IMF Staff Reports and authors' calculations.

Figure 11 shows average sentiments in the whole sample across sectors and income groups. Member countries' authorities express more agreement toward fund policy advice on financial, real and fiscal advice compared to advice on monetary or external sectors. Across the five sectors, it remains true that AEs authorities express the least agreement with Fund staff, followed by EMs and then LICs.

Figure 11. Average Sentiments Across Sectors and Income Groups (2000–18)



Source: IMF Staff Reports and authors' calculations.

In addition to important cross-country and cross-sector variations, our analysis thus far also revealed that global shocks affect countries' sentiments toward Fund advice. First, around the global financial crisis in 2008, there is a huge spike in AE's sentiments in favor of Fund advice following a pre-crisis dip. Second, there is a spike in sentiments following the oil price shock in 2014 (Figure 8). We investigate this further in the regression analysis below.

V. Analysis of Sentiment Index

Our preliminary results so far suggest a deeper look into the relationship between country characteristics and sentiment. More specifically, we investigate how a country's sentiments toward Fund advice changes with its (1) IMF quota; (2) political system and election cycle; (3) share of commodity exports and (4) sectoral aggregate risks. We also investigate how (5) a country's engagement in a financial arrangement with the Fund; and (6) its reliance on Fund fiscal and monetary Technical Assistance (TA) and on Financial Sector Assessment Program (FSAP) missions affect authorities' overall sentiment toward Fund policy advice. Section A lists data sources for the variables used regression analysis. Section B presents the regression results.

A. Data

We use cross-country IMF World Economic Outlook (WEO) data on quota, nominal GDP (current price), commodity exports and commodity prices for the sample period 2000–18. We use Chinn-Ito index as a measure of capital account openness. For political systems, we used several indices from Database of Political Institutions (DPI) 2017 published by the Inter-American Development Bank. We use variables from DPI that show how much power

the country's authorities (executives) have on determining its policies. Specifically, we use data for executive systems, election cycle, and whether the party of executive control all relevant houses. We use the Fund's Monitoring of Fund Arrangements (MONA) database to identify in which year a country was in a program, including on nature and type of program, and internal IMF databases to identify in which year a country received IMF technical assistance in the fiscal and monetary sectors.[14] We use a simple word search on the financial sector paragraphs on authorities views to look for where the word FSAP is mentioned to gauge the effect of having FSAPs on sentiments.

We use sectoral risk assessments from an internal cross-country early warning exercise at the Fund. The exercise identifies for all member countries emerging near-term macroeconomic risks using a bottom-up and consistent approach across all sectors of the economy and across advanced, emerging, and low-income economies. The sectoral risk measures we use cover the fiscal, financial and external sectors for the period 2000–18.

B. Regression Analysis

Country characteristics and Fund relations

In this section, we explore how country characteristics and Fund relations affect its sentiments toward Fund advice. For country characteristics, we look at the effects of a country's IMF quota, its political system, whether the country is a commodity exporter, and its sectoral risk measures. For Fund relations, we will assess how sentiments changes if a country is in program period or not, and whether it has received Fund technical assistance or undergone FSAPs. For the overall analysis, the dependent variable is the sentiment index which varies across countries, years, and sectors.[15] The right-side variables included in the section exhibit variation across countries and years. In our regressions, we include year fixed effects to capture global factors or global shocks affecting countries.[16]

Country characteristics

IMF quota

In Figure 8, we saw that countries with higher incomes tends to express lower sentiments or more disagreements with the IMF. One way to capture this is by looking at a country's IMF quota which determines its maximum financial commitment to the IMF and its voting power,

[14] Fiscal TA data is available for our whole sample period, while monetary TA is available only since 2009.

[15] Where there are multiple paragraphs expressing authorities' views on a particular sector, say fiscal, we average the sentiments across paragraphs to get one sentiment per sector for each country-year observation.

[16] Our first set of regressions include right-side variables that exhibit little time variation (IMF quota, openness, political systems). We therefore do not include country fixed effects in these and remaining regressions. Most of our other results are maintained when country FE are included.

and is usually determined by economic size and characteristics. As quota is increasing with the income level, our previous findings on sentiments across the three income groups suggest that one would expect that countries with larger IMF quota to express more disagreement with Fund policy advice. However, it could also be argued that countries with higher quota could have voiced more of their views internally and have their policy preferences already reflected in Fund policy advice, in which case one would expect more alignment and hence more agreement with Fund policy advice. The first column of Table 4 shows the regression of the sentiment index on country's IMF quota. The coefficient on IMF quota is negative and significant, confirming the former effect dominates.

In the second column of we decomposed IMF quota into its two major determinants: country's GDP and capital account openness as measured by the Chinn-Ito index.[17] The coefficients on both log (GDP) and Chinn-Ito index are negative. Both the size of the country and capital account openness negatively affects authorities' sentiments or their reception of Fund advice.

Political systems

Next, we look at how country's political system affects sentiments. One reason for countries to express disagreements toward Fund advice is the difficulty in implementation. Even when authorities understand the importance of Fund advice, if the policy is politically difficult to implement, they are likely to disagree upfront. Therefore, it is natural to look at how authorities' political power affect sentiments toward Fund advice.
The third column of Table 4 shows that the countries whose authorities' have more power in its political system tend to agree more to our advice. *All house* variable is a dummy that is 1 if the party of executive controls all relevant house, and 0 otherwise. *System* variable is 0 if the political system is presidency, 1 if the executive is assembly-elected president, and 2 if it is parliamentary. Positive coefficient on *All house* variable means that if the executive party has a dominant power in the country, they agree more to IMF. Negative coefficient on *System 2* (parliamentary) indicates that, compared to *System 0* (Presidency, omitted), countries with parliamentary system tends to disagree more with the Fund. The result is intuitive because, in general, it is more difficult to make policy changes when the country has a parliamentary system.

The fourth column of Table 4 shows whether election cycle affects sentiments toward Fund advice. Years left variable is the number of years until the end of the term. The positive coefficient means that the government with more years in current term tend to agree more with Fund advice. When a government has more time until the next election, they are more likely to agree with and even implement IMF policy advice.

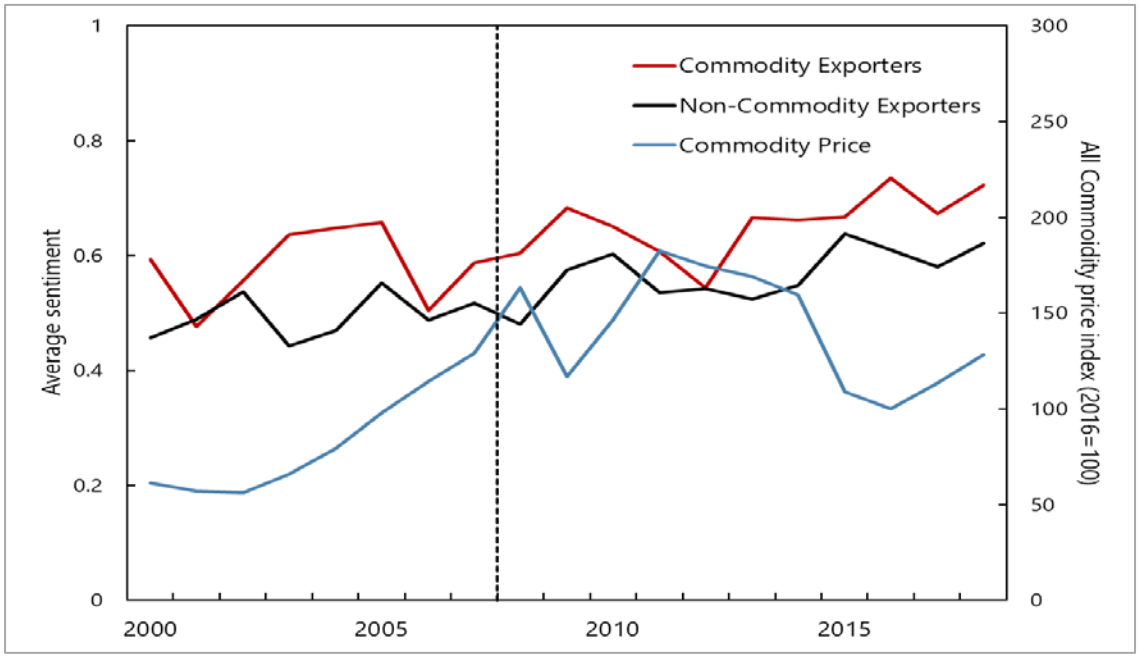
[17] IMF quota is determined by GDP (weight of 50 percent), openness (30 percent), economic variability (15 percent), and international reserves (5 percent).

Commodity price changes

We now investigate how commodity price changes affect sentiments towards Fund advice in commodity exporters vs. non-commodity exporters. Figure 12 shows the simple correlation over time between average sentiments of the two groups and the commodity price index (which includes non-Fuel commodities as well). Our priors are that when oil prices are increasing, net commodity exporters tend to express less agreement with Fund advice during those good times, whereas net commodity importers are more likely to agree with Fund advice as they face the costs of higher oil prices. An interesting pattern emerges: since the GFC, average sentiments across both groups exhibit a negative correlation with commodity prices, a result easier to explain for commodity exporters.

We look deeper into this relationship in the regression, and in doing so we run the regression for the whole sample as well as since 2008. The fifth column of Table 4 shows how commodity exporters' sentiment changes when the commodity price changes, compared to non-commodity exporters. First, for the whole sample, the coefficient on *commodity price* is positive and significant, suggesting that, when the commodity price goes up countries on average express more agreements. On the other hand, the interaction term of commodity price and oil exports is negative, implying that when commodity prices increase, countries with large oil exports tend to disagree more, compared to countries that export less. The overall effect of a commodity price increase on average sentiment is the sum of those coefficients and is a function of oil export intensity. The higher is oil export intensity the more likely that a price increase is associated with disagreements. On average, that overall effect is positive. When we run the regression for the period after 2008, the coefficient on the commodity price becomes negative and significant, and the one on the interaction term remains so, suggesting that the overall effect of a commodity price increase on sentiment is negative, regardless of oil export intensity (column 6). Alternatively, this suggest that following the 2014 negative oil price shock, countries on average agreed more with Fund advice, regardless of whether they are commodity exporters or not.

Figure 12. Average Sentiments and Commodity Prices



Source: IMF Staff Reports and authors' calculations.

Table 4. Regression Results for Country Characteristics

	IMF Quota		Politics		Commodity	
	Full sample					2008–18
	(1)	(2)	(3)	(4)	(5)	(6)
Quota	-0.053***					
Log (GDP)		-0.005**				
Chinn-Ito index		-0.031***				
All house			0.065***			
System 1			-0.036			
System 2			-0.126***			
Years left				0.016***		
Log (Oil Export)					0.071*	0.223**
Commodity Price					0.082***	-0.118*
Log (Oil Export) Commodity Price*					-0.017**	-0.048**
Time FEs	✔	✔	✔	✔		
Observations	8461	7177	5653	5786	5069	2866

Source: IMF Staff Reports and authors' calculations.

Aggregate sectoral risks

In this section, we investigate whether aggregate risks across the fiscal, external, and financial sectors affect countries' reception of Fund advice. The first three columns in Table 5 show the relationship between a risk measure for each sector and the average sentiment (across all sectors). Focusing on the most statistically significant results, fiscal risk has a **positive** effect on sentiment, whereas financial risk has a **negative** effect on sentiment. It means that, in countries assessed by the Fund as having a higher probability of fiscal stress, the authorities are more likely to agree with IMF policy advice, but when the risk is flagged in the financial sector, they are more likely to disagree.

In the fourth column of Table 5, we control for all sectoral risks measures together and find that the result that authorities agree more with Fund advice when we flag fiscal stress and less when we flag stress in the financial sector still holds. We also note the relative magnitude of the coefficients on each type of stress: the large coefficient on financial risk (twice as large as that on fiscal risk) suggests that authorities are more sensitive when we flag financial stress: the decline in average sentiment when doing so is twice as large as the increase in sentiment when fiscal risks are flagged.

Table 5. Regression Results for Sectoral Risks

	Sectoral Risks			
	(1)	(2)	(3)	(4)
External Risk	0.036*			-0.035*
Financial Risk		-0.111***		-0.075***
Fiscal Risk			0.038***	0.036***
Time FE	✔	✔	✔	✔
Observations	8454	8407	8077	7991

Source: IMF staff report and authors' calculations

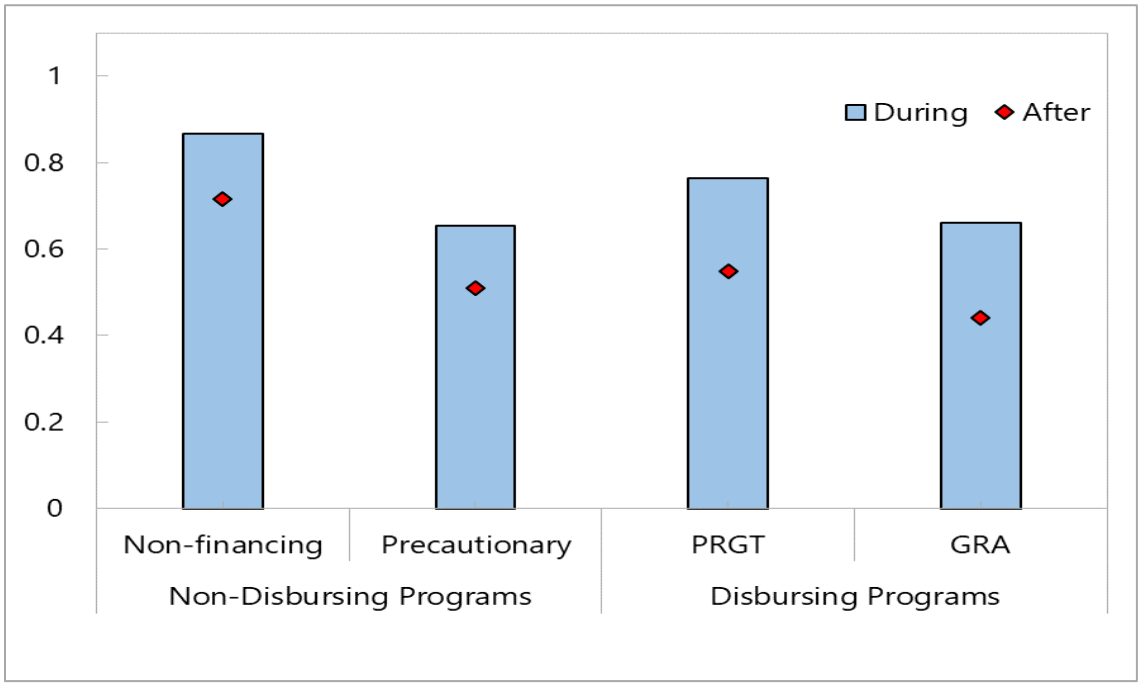
Fund relations

Program countries

Our dataset only includes Article IV Staff Reports during 2000–18 but for the years where some countries enter into a financial arrangement or program with the IMF, we use the combined Article IV and program review/request documents to gauge sentiments captured by the surveillance cycle during and after programs. We exclude pure program review/request Staff Reports. Our priors are to expect more agreement with the Fund during combined Article IV and program missions relative to non-program years.

During our sample period of 2000–18, there were 172 programs over 100 countries. Average duration of a program is 3.4 years. The fraction of precautionary programs out of all programs (program years of precautionary programs/ all program years) was 11.1 percent. Figure 13 shows the sentiments across different program types. Precautionary programs are listed on the left, disbursing programs on the right. On average, sentiments during programs are higher during precautionary arrangements relative to disbursing ones. Moreover, even though sentiments are lower post-program for both types, the decrease in sentiments following disbursing programs is higher than the one observed after precautionary programs.

Figure 13. Sentiments Across Program Types



Source: IMF Staff Reports and authors' calculations.

For program countries' regressions, we only include countries that experienced at least one program during sample period 2000–18. We first looked at, among those program countries, whether countries agree more during programs compared to non-program years. The answer is yes, from the first column of Table 6, where the coefficient on the program dummy is

positive and significant. *Program dummy* variable is 1 in the years the country is in program, and 0 before/after the program.

As for how sentiments differ before and after the program periods, we find that compared to before program period (*before program* dummy is omitted), countries express more agreement *during program*, and more disagreement *after program* (Column 2 of Table 6).[18]

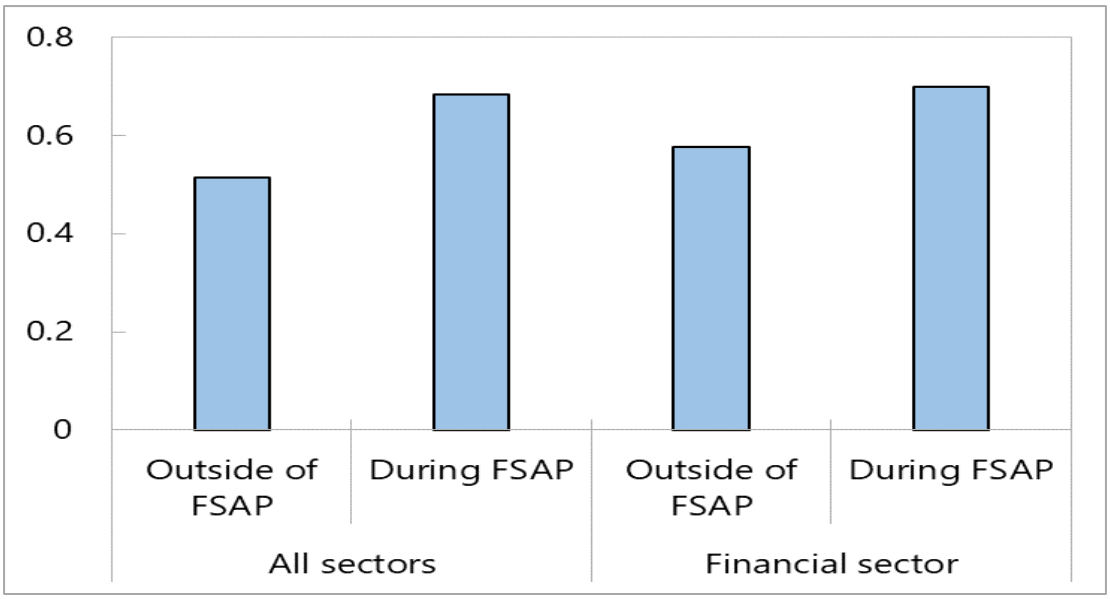
FSAP and technical assistance missions

Many member countries receive technical assistance from the IMF as well as undergo financial sector assessments that look deep into financial sector issues and issue specific financial sector policy recommendations. In this section, we limit our sample to countries that have received technical assistance and FSAPs during the sample period, and compare sentiments during and outside those services.[19] We ask here whether these additional services provided by the IMF increase Fund traction as captured by more positive sentiments. Figures 14 and 15 show that there is indeed an improvement in overall sentiments in countries undergoing FSAPs and receiving Fund technical assistance in the fiscal and monetary sectors. The improvement in sentiment is largest following fiscal sector TA, both in the fiscal sector and across all sectors, followed by improvements in sentiments in countries undergoing FSAPs and receiving monetary TA. We investigate these findings in the regressions below where we regress average sentiments on FSAP and TA dummies separately and interacted with own sector dummies.

[18] We chose not to distinguish program type in the regression due to the small fraction of countries with precautionary arrangements compared to disbursing ones in our sample.

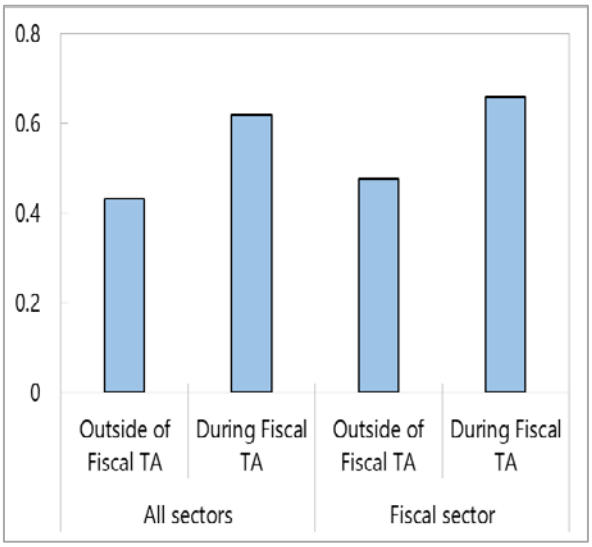
[19] We do not compare countries that received TA and FSAPs vs. those who have not, as there are very few countries that have not received those services during the sample period. If instead we compare for every year average sentiments in those who have and those who haven't received TA and FSAPs separately, we find that sentiments are indeed higher in those who have.

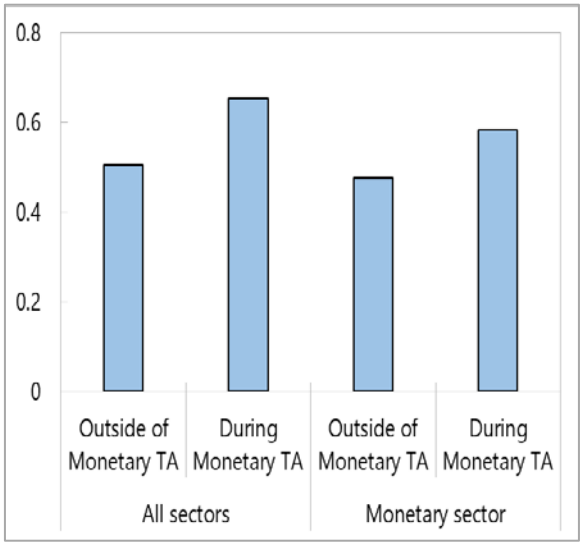
Figure 14. FSAP Effects on Average Sentiments



Source: IMF Staff Reports and authors' calculations.

Figure 15. Technical Assistance Effects on Average Sentiments





Source: IMF Staff Reports and authors' calculations.

Column 3–4 of Table 6 show that FSAP leads to more positive sentiments. We also investigate whether sentiments improve specifically in the financial sector or more generally. The interaction coefficient is not significant. This could be interpreted as FSAP improving traction across all surveillance sectors and is not just limited to the sectors they target. We find similar results on monetary and fiscal TA in Table 7.

Table 6. Regression Results for Programs and FSAPs

	Program		FSAP	
	(1)	(2)	(3)	(4)
Program Dummy	0.168***			
During Program		0.128***		
After Program		-0.117***		
FSAP Dummy			0.174***	0.076
Financial Sector Dummy				0.072***
*FSAP * Financial Sector Dummy*				0.054
Time Fes	✔	✔	✔	✔
Observations	4232	4232	7425	7425

Source: IMF staff report and authors' calculations.

Table 7. Regression Results for Fiscal and Monetary Technical Assistance

	Fiscal TA		Monetary TA (2009–2018)	
	(1)	(2)	(3)	(4)
TA Dummy	0.179 ***	0.018***	0.148***	0.154***
Fiscal Sector Dummy		0.058*		
Monetary Sector Dummy				-0.027
TA Fiscal Sector Dummy*		-0.006		
*TA * Monetary Sector Dummy*				-0.057
Time FEs	✔	✔	✔	✔
Observations	8676	8676	3956	3956

Source: IMF Staff Reports and authors' calculations.

C. Applying Model on Executive Directors' Buff Statements

An interesting application of our model as well as a robustness check is to run our trained sentiment model on buff statements. For brevity, we do not include the results in this paper but are reassured that our analysis holds on a different source of authorities' views.

We first identified all Buff statements that were issued ahead of discussions on Article IV consultations, as recorded by IMF's internal calendar database. We then selected the paragraphs that contain "authorities" as a subjective noun. For each paragraph, we applied the earlier methods and models to estimate its topic and sentiment. Specifically, we assign the topic with the largest number of matched words/phrases. For cases where more than more than one topic has equal number of matches, we apply the Support-Vector-Machine model to assign the topic with highest probability.[20] We apply our trained BERT model to estimate the paragraphs' sentiments. Model accuracy increases to 85 percent, which attests to the strength

[20] The approach gives us 75 percent overall accuracy, partly due to mixed paragraphs in the text (for example, external and monetary policies are often discussed in the same paragraph).

of the model as well as to the higher degree of candor in ED buff statements compared to in Staff Reports' authorities' views paragraphs.[21]

Our results generally hold, with the Buff model showing slightly higher overall agreement with Fund advice over the last 2 decades from Article IV Staff Reports (82.5 percent compared to 74.5 percent), as well as relatively more stable average sentiment across time compared to a slightly increasing trend in the recent years in authorities' sentiment as captured by their views in Article IV Staff Reports.

VI. CONCLUSION

IMF member countries' reception of Fund advice during Article IV discussions is an important indicator of Fund traction. To our knowledge, this is the first paper to provide a comprehensive analysis of this dimension of traction, overcoming the inherent difficulty of analyzing text (not numbers) across thousands of published reports over time. By using latest natural language processing techniques from Google 2018, we construct the first sentiment index across member countries over the last two decades. Our index confirmed some standard priors, which were thus far conventional wisdom rather than empirical findings, as well as new insights on traction of Fund advice.

Our index will be key to monitor an important aspect of traction of Fund policy advice at country-specific and cross-country levels, is easily tractable and can be updated on a regular basis. For instance, the index can be used to help country teams identify where traction of their policy advice needs further work to be more effective, and how these sentiments compare across different country groups and different policy sectors. The index can also be used to highlight cross-country and regional dimensions of traction of Fund advice. While this paper focused on building the index and assessing some of the issues at an aggregate level, we leave these other, equally relevant questions, to future research.

[21] The AIV and Buff datasets do not fully overlap, probably due to the required coverage of authorities' views for all sectors in Article IV Staff Reports whereas EDs can be selective about which sectors to focus on. When we compare their intersection, i.e. when there is a Buff and Article IV sentiment for each sector-year-country, 85 percent of the time our model assigned same sentiments for both (for each country-year-topic combination), and 15 percent of the time the sentiments assigned for each are opposite.

Appendix I. Examples of Annotation

Agree

"The government **agreed** that the phasing out of quota restrictions on rice importation with adequate support for farmers is important for poverty reduction …"

"The authorities **are firmly committed t**o statistical improvement. They recognized the importance of sound macroeconomic statistics…"

"The ECB has madl NPL resolution **a policy priority, but faces hurdles…**"

"The authorities **agreed with staff** that the moderate expansion of the German economy is likely to continue. … **While they concurred with staff** that inflationary pressures would be muted this year, **the Bundesbank underscored that prices were likely to accelerate** owing to pass-through from exchange rate depreciation ... The authorities **agreed** that the current account surplus would decline only gradually. They also **concurred that** public debt is well-anchored ..."

Mixed

"The authorities **shared staff's assessment that** the economy was adjusting in the right direction. **They broadly concurred with** the near-term macroeconomic and financial sector outlook … **Nevertheless, the authorities considered international reserves to be adequate, based on their preferred metric**, at close to the target of 5 months of official merchandise imports associated with non-FDI activities is a significantly lower amount than those used in the staff's metric ..."

Disagree
"The authorities **did not agree** with the characterization of their external position as substantially stronger than warranted by …"
"**While authorities agreed** that fiscal policies should aim at ensuring long-term sustainability and be fully transparent, **they did not see a strong case for** introducing a medium term …"
"The authorities would have preferred to anchor fiscal discipline with a hard constitutional balance rule, but **there was no political consensus for this route …**"
"The authorities **view the exchange rate as close to the** value implied by fundamentals and long-term averages. **They saw few signs of misalignment**, with the weker euro reflecting …"

Appendix II. BERT Model in Detail

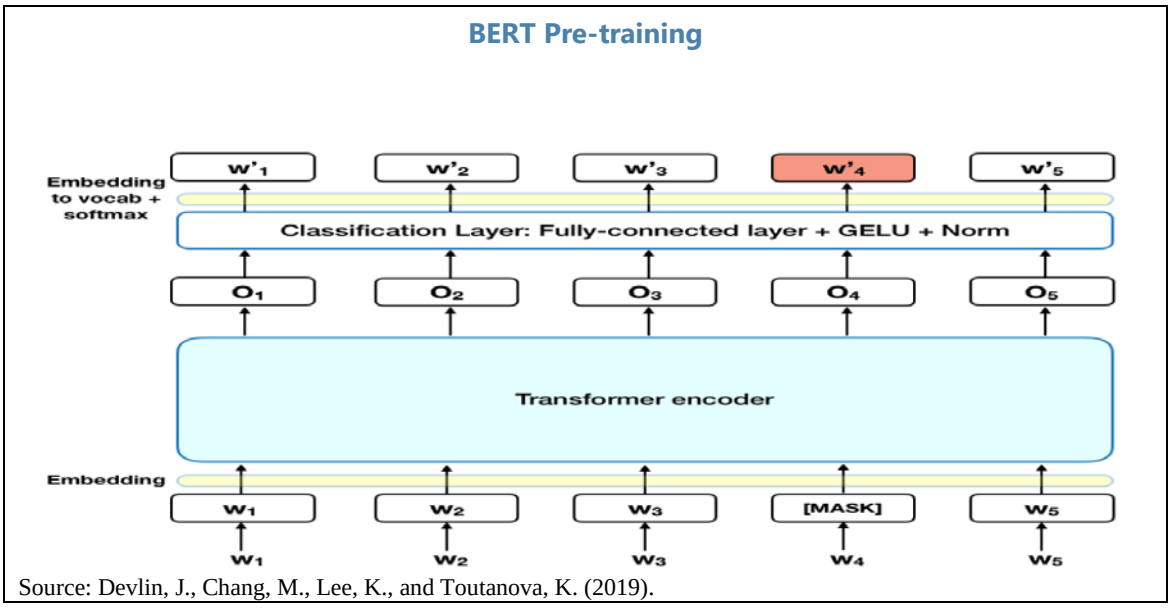
BERT has been widely considered as a milestone in deep learning NLP because of its superior performance in a wide range of NLP tasks. It combines two of the most import findings in modern deep learning NLP; "Language model pre-training" (unsurprised learning in a large unlabeled corpus) and "Attention mechanism" (dynamically assign attention weights to important parts of the document to capture long-term dependency).

Our training procedure

1. Load Pretrained BERT-base Model from Google:

Training a BERT model from scratch is an extremely time-consuming task. Instead of training our own BERT model, we downloaded "BERT-base" model provided by Google. The model is trained on the BooksCorpus (800 million words) and English Wikipedia (2500 million words).

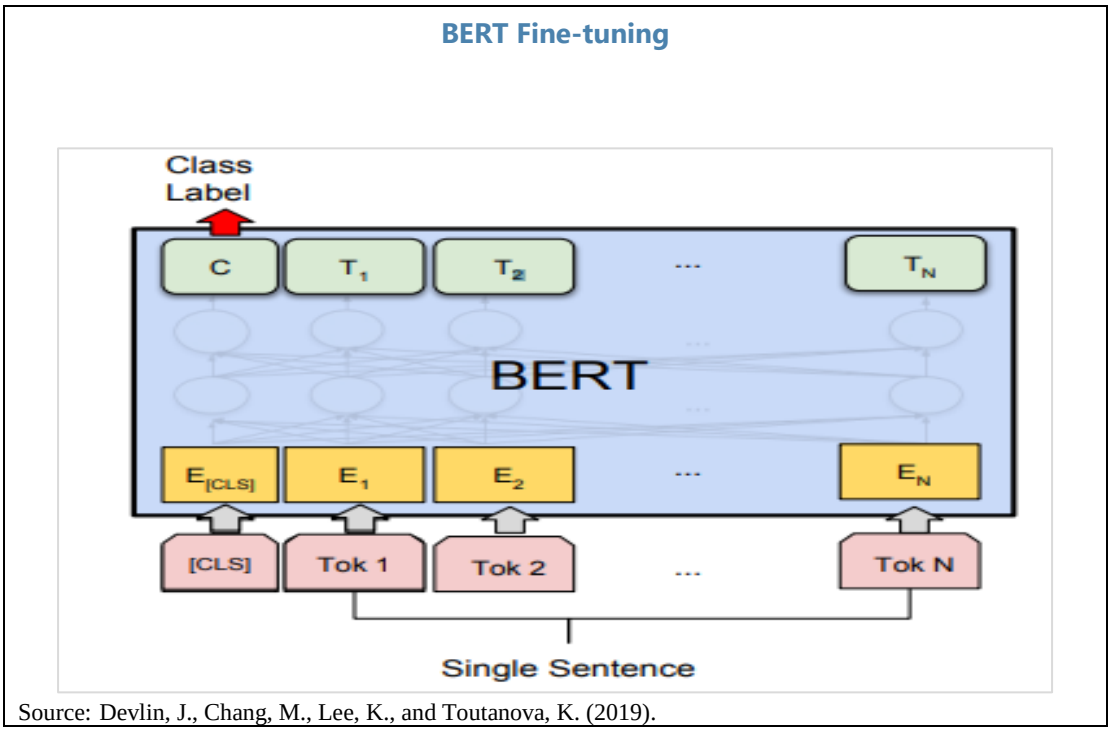
Model is trained in an unsupervised fashion on two tasks: 1) train a language model by using the context words to predict masked words. The idea is to let the model have a general understanding of the words and context. 2) train a classification task to determine whether two sentences are next to each other. The idea of this next sentence prediction task is to train the model to capture long-term dependencies and generate an abstract sentence level embedding for downstream tasks.



Source: Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019).

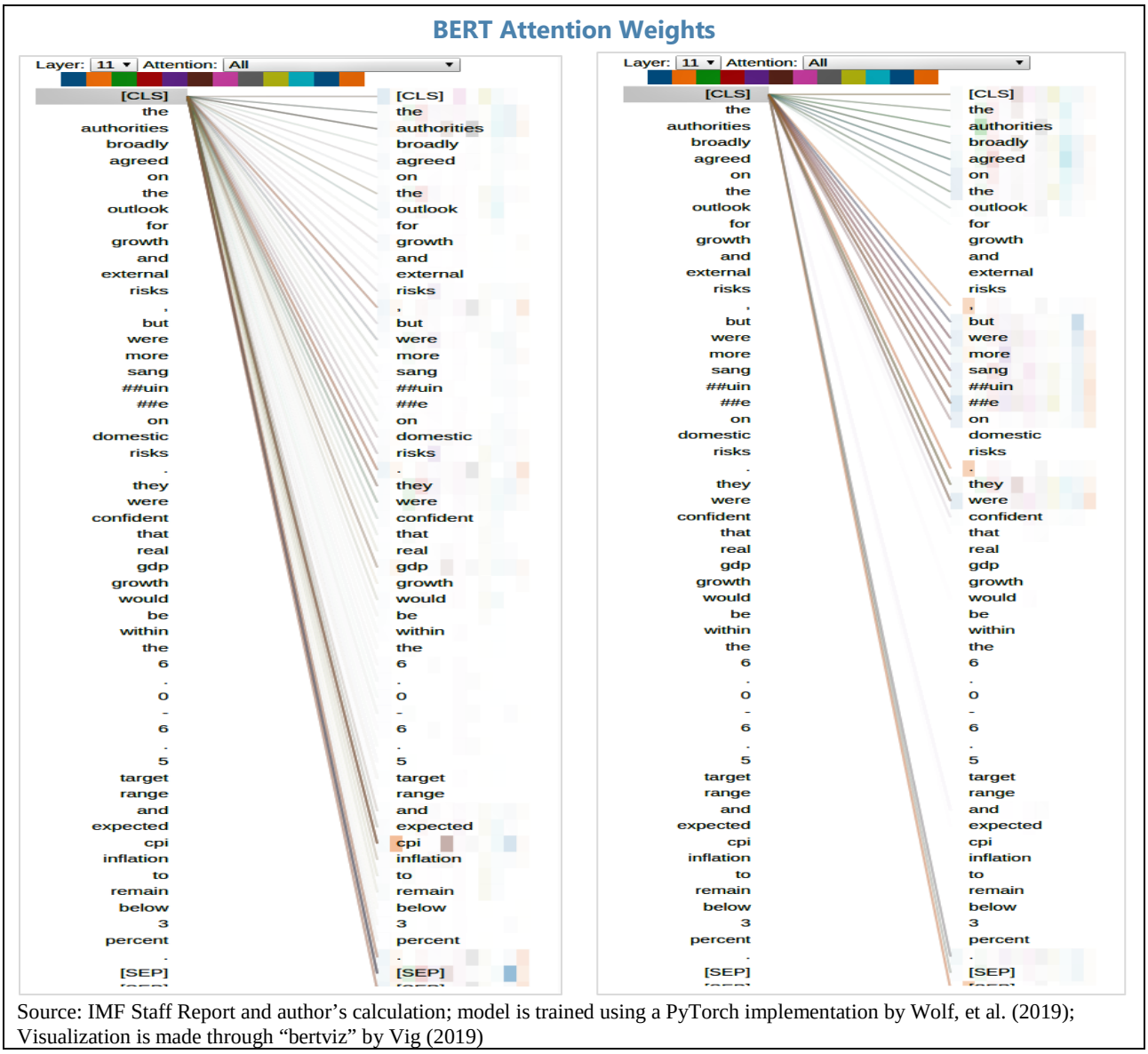
2. Use the pre-trained language model as "prior", fine-tune the model with our sentiment task with the same model architecture:

Simply use pre-trained model without fine-tuning doesn't usually yield good results, as the specific task we are trying to perform (determine authorities' sentiment) is slightly different from pre-training tasks. Also, the contextual information embedded in IMF documents can be very different from books and Wikipedia. As a result, a fine-tuning step using our own labeled training data is essential for the model to adapt to IMF language.



Source: Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019).

Through our own training process, the model learns and adapts attention weights based on our specific sentiment task and use that information for prediction:



Source: IMF Staff Report and author's calculation; model is trained using a PyTorch implementation by Wolf, et al. (2019); Visualization is made through "bertviz" by Vig (2019)

Here, we will show some intuition of how model interpret the sentence, by plotting out the attention weights. We only look at the attention weights of [CLS] token. At high level, [CLS] itself is short for classification as the [CLS] token is often used for classification task and is interpreted as the overall paragraph embedding. Thus, the attention weights of [CLS] token give us a good representation of how the model interprets sentences. We compared a pretrained model from Google (Left) with a model that is fine-tuned by our sentiment classification task (right). We can see a clear difference in terms of which part of the sentence it pays more attention to. The attention weights of the pretrained model seems to spread across the entire sentence, with some heavy emphasis on "CPI". Probably "CPI" appears more on the news. But after fine-tuning with our sentiment task, it learnt to attend more on contents around authorities and contents after "but", which is much closer to how human reads if given a sentiment task.

References

Born, B., Ehrmann, M. and Fratzscher, M. Central bank communication on financial stability. Econ. J. (2014). doi:10.1111/ecoj.12039.

Custer, S., DiLorenzo, M., Masaki, T., Sethi, T., and Harutyunyan, A. (2018). Listening to Leaders: Is development cooperation tuned-in or tone-deaf? AIDDATA- A Research Lab at William & Mary.

Chinco, A., Clark-Joseph, A. D. and Ye, M. Sparse Signals in the Cross-Section of Returns. J. Finance (2019). doi:10.1111/jofi.12733.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In North American Association for Computational Linguistics (NAACL).

Edwards, M.S. and Senger, S., 2015. Listening to Advice: Assessing the External Impact of IMF Article IV Consultations of the United States, 2010–2011. *International Studies Perspectives* 16:3, 312-326.

Evans, J., and Aceves, P., (2016). Machine translation: mining text for social theory. annualreviews.org.

Gutner, T. and Thompson, A. (2010). The politics of IO performance: A framework. *Review of International Organizations*, 5: 227-248.

Grimmer, J. and Stewart, B. M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Polit. Anal. (2013). doi:10.1093/pan/mps028.

Honnibal, M., and Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

Hutto, C., and, E. G.-E. international A. conference on weblogs & 2014, undefined. Vader: A parsimonious rule-based model for sentiment analysis of social media text. aaai.org.

IMF. (2004). Biennial Review of the Fund's Surveillance— Overview; Modalities of Surveillance; Content of Surveillance; and Public Information Notice on the Executive Board Discussion.

IMF. (2008). 2008 Triennial Surveillance Review— Overview Paper.

IMF. (2011). 2011 Triennial Surveillance Review— Overview Paper.

IMF. (2014). 2014 Triennial Surveillance Review— Overview Paper.

Lombardi, D. and Woods, N. (2008). The Politics of Influence: An Analysis of IMF Surveillance. Review of International Political Economy, 15(5): 711-739.

Lucca, D. O. and Trebbi, F. Nber Working Paper Series Measuring Central Bank Communication: An Automated Approach With Application To Fomc Statements. (2009).

Loughran, T. and Mcdonald, B. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. J. Finance (2011). doi:10.1111/j.1540-6261.2010.01625.x

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.; Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. arXiv:1310.4546.

Momani, T. (2006). Assessing the Utility of, and Measuring Learning from, Canada's IMF Article IV Consultations. Canadian Journal of Political Science, 39 (2): 249-269.

Pang, B., Lee, L. and Vaithyanathan, S. Thumbs up? Sentiment Classification using Machine Learning Techniques. (EMNLP, 2002).

Rehurek, R. and Sojka, P. Software Framework for Topic Modellingwith Large Corpora. InProceedings of the LREC 2010 Workshop on NewChallenges for NLP Frameworks, pages 45–50, Valletta, Malta, May 2010.ELRA. http://is.muni.cz/publication/884893/en

Shapiro, A. H. et al. Taking the Fed at its Word: A New Approach to Estimating Central Bank Objectives using Text Analysis Taking the Fed at its Word: A New Approach to Estimating Central Bank Objectives using Text Analysis *. (2019). doi:10.24148/wp2019-02.

Tallberg, J., Thomas S., Theresa S. and Magnus L. (2016). The performance of international organizations: a policy output approach. Journal of European Public Policy, 23 (7): 1077-1096.

Tetlock, P. C. Giving content to investor sentiment: The role of media in the stock market. J. Finance (2007). doi:10.1111/j.1540-6261.2007.01232.x.

Vig, J. A multiscale visualization of attention in the transformer model.arXiv preprint arXiv:1906.05714, 2019