Sentiments of Power: Executive Board Discontent and IMF Lending Practices

Merih Angin

Department of International Relations & Computational Social Sciences Program, Koç University

Abstract

The variation in the design of International Monetary Fund (IMF) programs, ranging from substantial financial support with minimal conditions to those laden with stringent requirements, has been a subject of active debate. Despite a rich body of literature grounded in both qualitative and quantitative research, the influence of the IMF's Executive Board (EB) remains understudied, largely due to the complexities of systematically quantifying sentiments and disagreements in textual data. This paper addresses that gap by introducing an "Executive Board Discontent" variable, developed using advanced natural language processing (NLP) techniques employed on approximately 50,000 pages of Executive Board meeting minutes that have been labelled by more than 80 human annotators over the course of 5 years. This variable captures the sentiments of IMF Executive Board members regarding program design, offering new insights into the internal dynamics influencing IMF lending decisions. The empirical findings show that discontent within the Executive Board significantly affects the number of conditions attached to IMF programs. While traditional econometric models provide moderate explanatory power, the application of machine learning (ML) models significantly improves the accuracy of predicting IMF loan conditions, underscoring the potential of integrating advanced analytical methods in economic policy research.

Keywords: IMF, Executive Board, Conditionality, NLP, Machine Learning

1. Introduction

The International Monetary Fund (IMF), often described as "the most powerful international institution in history" (Stone, 2002), is indeed the international organization that has garnered the most attention in international relations scholarship, with a rich body of literature examining its influence and operations—particularly regarding its primary mandate of lending to countries in financial distress (Helleiner, 1987; Taylor, 1987; Martin, 1991; Killick, 1995; Vreeland, 2003; Woods, 2006; Momani, 2007; Copelovitch, 2010a; Chwieroth, 2010; Stone, 2011; Rickard and Caraway, 2014; Dreher et al., 2015; Kentikelenis et al., 2016; Nelson, 2017; Lang, 2021; Angin et al., 2023; Metinsoy, 2024). A key issue that continues to be debated in the literature is the question of who governs the IMF and how its programs, which come with a variety of strings attached, are designed.

Previous studies have emphasized the economic conditions of borrowing countries (Bird, 2001; Edwards, 2003), and the influence of powerful member states, particularly the G5, as crucial determinants of IMF lending practices (Stone, 2008, 2011; Copelovitch, 2010a,b; Breen, 2014). Additionally, internal dynamics within the IMF, including the organizational culture, the autonomy of staff, and their confidence in borrowing countries' policies, have been found to be significantly influencing lending decisions and the design of IMF programs (Momani, 2007; Chwieroth, 2010; Nelson, 2017).

Despite a rich body of literature grounded in both qualitative and quantitative research, the influence of the IMF's Executive Board (EB) remains an understudied aspect of the equation. While some of the existing literature discusses the potential impact of Executive Board dynamics (Thacker, 1999; Barro and Lee, 2003; Dreher, 2004), they did not systematically quantify these sentiments and disagreements, likely due to the complexities involved in analyzing text as data. This paper addresses this gap by introducing a novel variable: IMF's EB discontent. Derived from a meticulous annotation of over 50,000 pages of IMF Executive Board meeting minutes, and utilizing an advanced natural language processing (NLP) tool, this variable systematically captures and quantifies the sentiments, disagreements, and alliances within the Board, offering unprecedented insights into the internal factors that shape the stringency and design of IMF programs.

This research makes four key contributions to the literature on the IMF and the study of international organizations (IOs). First, it develops a comprehensive theoretical model for understanding IMF program design and

implementation, integrating both existing variables from the literature and novel ones derived from the analysis of Executive Board meeting minutes. Second, it introduces machine learning (ML)-based models of the IMF lending process, offering high predictive power for program outcomes under various macroeconomic and political conditions. This represents a significant methodological advancement in a field that has traditionally relied on statistical prediction models. Third, the research adopts an eclectic approach, employing a mix of methods—including ML, natural language processing (NLP), and statistical analyses—to account for variation in the terms of IMF programs. Finally, by developing a tool for the automated analysis of Executive Board meeting minutes to extract data on Board views regarding IMF program design, the research significantly enhances the potential for future studies on IO document analysis. The ML framework developed here is also expected to be a valuable tool for political scientists, enabling the application of ML-based modeling to a range of problems in IOs and other domains.

This innovative approach, which integrates ML and NLP techniques with deep qualitative insights, not only addresses a critical gap in the study of international organizations but also makes a significant contribution to computational social science. By introducing the Executive Board sentiment variable, this study opens new avenues for understanding the dynamics of IMF decision-making and paves the way for future advanced studies in the field.

The rest of the paper is organized as follows: The next section outlines IMF decision-making with a brief summary of the existing literature on IMF lending, followed by the hypotheses of the research. Moving forward, the paper explains the research design: the section outlines the methodological approach, describing the data sources, and the techniques employed for analysis. It provides an overview of the statistical methods, natural language processing tool, and machine learning algorithms that constitute the core of our analytical framework. The following section of the paper presents the empirical results. Here, I offer a detailed account of the findings from each methodological strand, discussing the implications of the results in light of the hypotheses. Finally, the conclusion summares the key findings and their potential contributions to the field.

2. Decision making at the IMF

Although the IMF has often been criticized for its "one size fits all" approach (Stiglitz, 2003), a rich body of literature reveals that the IMF's lending policies vary significantly across time and space (Kang, 2007). At the heart of this debate lies IMF conditionality, which emerged as an operating principle in the 1950s, originally focusing on fiscal conditions, credit expansion, and the balance of payments (James, 1996). While IMF conditionality was primarily centered on macroeconomic policies until the early 1980s, it later expanded in both complexity and scope to include structural conditions. These "strings attached" to IMF loans usually require substantial reforms in the borrowing country, leading to politically contentious issues such as labor market reforms, which can spark heated debates at the domestic level. Vreeland (2003)'s work shows that governments may strategically use IMF programs to implement unpopular reforms or to shift blame for economic hardships. However, in extreme cases, implementing these conditions have even contributed to the collapse of regimes, as seen in the downfall of Nicolae Ceaușescu's government in Romania, revealing the profound and far-reaching consequences that IMF conditionality can have on political stability and governance.

The variation in IMF conditionality has, therefore, sparked significant debate among both scholars and policymakers, as the reasons behind these differences are not immediately clear. Stone (2008) argues that US influence is particularly strong when the IMF provides loans to countries of significant political importance. However, in more typical situations, the Fund tends to implement policies that align with the consensus of its most influential members, while enjoying considerable autonomy within its delegated authority. Stone describes this as "informal governance," where US involvement deviates from the formal rules. In a related vein, Copelovitch (2010a) introduces a "common agency" framework, where the G5 countries, exercising "de facto control" over the Executive Board, serve as the "political principal" of the IMF. He argues that differences in preferences among the G5 members account for variations in IMF loan size and conditionality. According to his empirical findings, when the G5 countries' interests are less pronounced, the IMF staff, acting as the agent, has greater autonomy, leading to lending policies that reflect more technocratic or bureaucratic priorities.

In contrast to the state-centric views, constructivist scholars offer a different perspective; Chwieroth (2010) highlights the influence of internal or-

ganizational culture and the beliefs of IMF staff, suggesting that ideational factors significantly affect lending decisions and conditionality design. Momani (2007) also points to the autonomy and expertise of IMF bureaucrats, asserting that staff play a critical role in shaping policy outcomes independent of member state interests. Nelson (2017) adds another layer of complexity by arguing that the IMF's relationship with its borrowers is heavily influenced by economic beliefs and ideational factors, where the IMF staff's confidence in the economic policies of borrowing countries can affect the stringency of conditions imposed.

Copelovitch (2010a)'s work on IMF lending suggests that the rationalist approach, which takes a state-centric view focusing on P-A dynamics, and the constructivist approach, which views bureaucrats as "authorities in their own right," are not mutually exclusive but rather complementary. By the same token, this paper addresses a key shortcoming of the principal-agent (P-A) models—specifically, the need for a more nuanced understanding of agent interests. By challenging the common belief that the IMF consistently favors the strategic allies of the US while leaving Fund staff with predominant leverage over conditionality design in low-income countries, I argue that the reality is more complex, with subtler differences between these cases than is often assumed.

Decision making at the IMF begins with the staff's preparation of proposals. Once prepared, these proposals cannot be amended by the Executive Board, as the recipient country signs the Letter of Intent (LoI) prior to the corresponding Executive Board meeting. Although the Executive Board retains the authority to reject a proposal, which would then be sent back for renegotiation, this has not happened in recent memory (Stone, 2011). Stone argues that the extensive "authority" delegated to IMF management effectively weakens the Executive Board and exacerbates "information asymmetries" within the institution (Stone, 2011). Executive Directors, with the exception of the representative from the recipient country, do not participate in country missions or in the negotiation process for programs. Taking the operation of the IMF into consideration, using a P-A model can be considered as the most natural way of capturing the mechanisms behind the decision-making at the Fund.

Given the complexity of P-A problems in the design of IMF programs, how can we best trace the mechanisms creating variation in conditionality? As Copelovitch (2010a) acknowledges, statistical analyses provide limited understanding of the actual decision-making processes within the Fund (p.

27), especially when Executive Board dynamics are considered. The Executive Board is at the core of the IMF's influence. Chaired by the Managing Director of the Fund, the EB consists of 24 Executive Directors. The G5 countries—the five largest contributors to the IMF (the US, the UK, France, Germany, and Japan)—appoint their own Executive Directors, while China, Saudi Arabia, and Russia also have their own seats. The remaining seats are elected by separate constituencies formed by the other member states. This arrangement gives the G5 substantial power on the Board, which they can use to advance their national interests, as indicated by the abovementioned studies. To fully understand the influence of the IMF Executive Board, however, it is essential to analyze the meeting minutes to capture the sentiments that may shape program design. This paper undertakes exactly that task, systematically examining these minutes to reveal how the Board's sentiments influence the formulation of IMF programs.

As mentioned earlier, conventional wisdom suggests that the IMF eases conditionality when a strategic ally of the G5 is borrowing. However, in some cases, we observe numerous conditions attached to loans, which contradict the preferences of the principal. This suggests that IMF staff, due to information asymmetries, may have the leverage to push for more stringent conditions. As Vaubel (1986) argues, international bureaucrats, like their national counterparts, seek to maximize their power through budget size, staff, and discretionary freedom, benefiting from an information monopoly that politicians rely on. Similarly, recipient-country bureaucrats may gain more autonomy during severe economic crises, which can influence program design in two ways: directly, as the severity of the crisis typically correlates positively with the number of conditions, and indirectly, through the consensus between IMF staff and recipient-country technocrats on the benefits of those conditions. Consequently, we should expect increased conditionality when such a consensus exists.

In this context, we consider what Gutner (2005) terms "antinomic delegation" as a potential cause of principal-agent problems in IMF program design. Principal-agent theory acknowledges the presence of "multiple or collective principals" in bureaucracies, whose goals may sometimes conflict. "Antinomic delegation" refers to the delegation of conflicting or complex tasks that are difficult to institutionalize and implement. In such cases, performance problems cannot be solely attributed to "agency shirking" but may also arise from the challenges agents face in implementing goals that are difficult to specify and balance. I argue that the number of conditions attached

to a program is proportionate to the degree of antinomic delegation. Designing a program to save a country in deep economic crisis is a complex task, and when the Executive Board instructs the IMF staff to sign an agreement regardless of the challenges, it poses risks for the institution. If the program fails, the legitimacy of the Fund—which has been periodically questioned since the collapse of the Bretton Woods System—could be further damaged. To mitigate this risk, IMF staff may design stringent programs with many conditions to ensure that even if the program fails, they are not criticized by an Executive Board that expressed discontent with the design of programs.

The following hypothesis is thus surmised:

H1*: The IMF attaches more conditions when the IMF Executive Board is more discontent with the program design.*

Hypotheses 2 and 3 serve as exploratory analyses to investigate additional dimensions of IMF decision-making:

H2*: The IMF grants more stringent conditions to borrowing countries when the scope of the economic crisis is larger.*

H3*: The IMF grants larger loans to borrowing countries when the IMF Executive Board is more content with the program design.*

3. Research Design

To assess these arguments, I utilize the data on IMF conditionality from Kentikelenis et al. (2016). The data are sourced from internal IMF documents—including IMF staff reports, the Letters of Intent (LoI) of borrowing countries, and Memoranda of Economic and Financial Policies (MEFPs)—that collectively contain detailed information on IMF program approvals, conditionality, and policy implementation. This allows us to analyze the Fund's behavior at different phases of an IMF program. Specifically, we assess the favorability of IMF lending along two variables: (1) loan size; and (2) types of policy conditions imposed by the IMF. We transform the original data into country-year observations. Our dataset includes a total of 640 observations with 126 borrowing countries between 1976 and 2017.

Dependent Variables

The empirical analysis centers on two key outcomes. The first dependent variable is the total loan disbursement amount (in millions of SDR), divided by the IMF quota of the borrowing country, agreed upon in year *t*. The IMF quota formula is a weighted average of GDP (weight of 50 percent), openness

(30 percent), economic variability (15 percent), and international reserves (5 percent).

We then analyze the factors influencing the number and types of policy reforms attached to the IMF programs. Policy conditions are defined in a recipient country's Memoranda of Economic and Financial Policies (MEFP), which accompanies the Letter of Intent at the start of an IMF program. The types of conditionality in Fund programs vary in specificity, content, and monitoring requirements, leading to differences in how binding they are on the recipient country (Copelovitch, 2010a). Performance criteria, often based on numeric variables, are crucial indicators of whether a program is functioning as intended. Failure to meet these criteria requires IMF Executive Board approval to continue the program and access future credit tranches. Structural performance criteria, unlike numeric ones, reflect changes such as the enactment of laws or administrative actions. Compliance with both performance criteria is mandatory for successful program reviews. Prior actions are measures a country must take before the Executive Board approves a loan or completes a review. Structural benchmarks, which are qualitative, serve as "markers" to assess the progress of an IMF program during a review. IMF conditionality covers a wide range of policy reforms. This research adopts a disaggregated approach to examine conditionality by policy category, such as labor-related or fiscal conditions.

Main explanatory variables

- **EB discontent:** The arduous part is measuring the Executive Board's sentiment based on the Executive Board meeting minutes, since it requires us to form a variable based on our own intuition. In this project, we develop an automated analysis technique for EB meeting minutes, which relies on NLP to extract the EB members' stances on IMF programs, including the discontent they express, level of solidarity between board members from different countries, and the views of G5 members on the program design. Further information on this NLP tool and the creation of the variable can be found in the Methodology section.

- **Economic hardship**: Based on earlier literature on IMF lending, we also expect economic crisis to be associated with tougher programs. We measure economic hardship by a standard macroeconomic measure of the IMF recipient country, i.e., GDP growth. Data for GDP growth are obtained from the World Bank.

Controls

The vector of controls include factors that are likely to influence IMF lending. As mentioned earlier, the existing literature suggests that major IMF shareholders' allies usually receive larger loans with less stringent conditions. To account for shared security interests, the models use a dummy to indicate whether the recipient country is an ally of any G5 country, based on data from the Correlates of War (COW) Formal Alliance dataset.

Based on earlier literature on IMF lending, we also expect economic crisis to be associated with tougher programs. We measure economic hardship by standard macroeconomic measures of the IMF recipient country. These include GDP growth and GDP per capita. Data for GDP growth and GDP per capita are obtained from the World Bank.

The economic relationship between G5 members and IMF recipients is also an important consideration. Countries that are vital to the economic interests of G5 members, such as key export markets, are likely to receive more favorable treatment because the negative effects of strict policy conditions can impact foreign economies, especially when there is significant economic interdependence. The models control for these shared commercial interests by including the recipient country's total imports and exports (logged) with G5 countries. Data are obtained from COW Bilateral Trade dataset.

To account for the influence of political institutions in the recipient country, the models include a control for the level of democracy, measured by the Polity score. Countries with higher levels of democracy might receive larger programs, either because their political systems are similar to those of the G5 countries or due to concerns among G5 policymakers about the risk of democratic regression.

4. Methodology

4.1. NLP Tool Development

NLP is a subfield of artificial intelligence, which aims to make human language understandable by computers. NLP starts with the formation of a corpus, which states the rules and associations between words in text. The process involves parsing text into a syntax tree, removing stopwords (i.e., words that do not add value to meaning),stemming the words and extracting the meaning of the text by utilizing one of the many information retrieval/ML algorithms used in NLP.

While some NLP tools have been developed for general language and specialized domains, such as the analysis of Federal Reserve meeting minutes (Zadeh and Zollmann, 2009), applying NLP to automatically analyze IMF Executive Board meeting minutes requires the integration of domain knowledge into the construction of the corpus. With the support of over 80 students who have worked for this project since 2020, we first manually annotated EB meeting minutes to build a specialized corpus, on which an NLP model was trained. This model was then utilized to analyze the entire set of meeting minutes to extract the sentiments of EB members regarding program design.

Consequently, a new dataset was created by merging the sentiment analysis data with IMF loans and conditionality data from Kentikelenis et al. (2016). The sentiment analysis data include our primary EB discontent variables, which measures the number of negative sentences made by Executive Directors during EB meetings.

Apart from creating the Executive Board discontent variable for statistically testing my hypothesis, the NLP tool developed in this research also contributes to the IO literature with its high accuracy. The tool can be easily fine-tuned and applied to other IOs, such as the World Bank.

Statistical Models

To analyze the size of IMF loans, we first estimate the following ordinary least squares (OLS) model with standard errors clustered by borrowing country j:

$$\begin{aligned}\text{Loan/Quota}_{j,t} = \beta_0 &+ \beta_1 \text{EBD}_{j,t-1} + \beta_2 \text{GDP Growth}_{j,t-1} \\ &+ \beta_3(\text{EBD}_{j,t-1} \times \text{GDP Growth}_{j,t-1}) + \gamma \mathbf{X}_{j,t-1} \\ &+ \eta_j + \theta_t + \varepsilon_{j,t}\end{aligned}$$

where EBD_j represents the EB sentiment, $\mathbf{X_j}$ is a vector of control variables, η_j captures borrowing country fixed effects, and θ_t captures year fixed effects.

Country fixed effects are included to account for any time-invariant characteristics that are unique to each recipient country, while year fixed effects control for annual variations and unobserved global factors. To address the sequential nature of the causal relationship and reduce the risk of simultaneity bias, EBD and $GDP\ Growth$ are lagged by one year.

Given that the conditions are count variables, we utilize negative binomial regression to estimate policy conditions, with standard errors clustered by borrowing country j. The control variables used in these models are consistent with those in our loan size models, except for Polity2. In the conditionality models, we anticipate that the coefficient on EBD will be positive because higher levels of Executive Board discontent are likely to lead to an increase in the number of conditions attached to IMF programs. This is based on the assumption that when there is significant discontent or disagreement among the Executive Directors regarding the policy direction or economic management of a borrowing country, the IMF may respond by imposing more stringent conditions to mitigate perceived risks.

4.2. Machine Learning Models

Recent advances in artificial intelligence and high-performance digital data processing architectures have enabled researchers across various fields to efficiently extract valuable insights from diverse data sources in an automated or semi-automated manner. Machine Learning, which not only facilitates the extraction of meaningful patterns but also enables accurate forecasting of future events based on historical data and current conditions, is becoming an indispensable tool for economists at international organizations. ML techniques are increasingly used to make estimations about macroeconomic variables in different countries. Recent studies have shown that ML techniques provide more accurate predictions of financial crises (Ahn et al., 2011; Kim et al., 2009), currency crises (Lin et al., 2008), stock market volatility (Son et al., 2009), and economic variables (Jung et al., 2018; Tiffin, 2016) than traditional statistical models, which often impose stricter assumptions on the relationships between dependent and independent variables. While traditional statistical models excel in explaining the relationships between variables in historical data, ML models offer superior predictive power, making them a powerful tool for forecasting future outcomes based on past data.

This study employs a machine learning approach to understand the design of IMF programs by exploring how various economic and political variables interact to shape program outcomes, particularly in terms of conditionality. Unlike traditional statistical methods, which take a hypothesis-driven approach to identify correlations between independent and dependent variables, ML models adopt a data-driven approach, allowing for the construction of models with high predictive power. The data-driven nature of ML enables the discovery of latent structures within large datasets—patterns that might

go unnoticed by human analysts—and facilitates the integration of numerous independent variables into the predictive process. The ML model itself reveals the relative importance of these variables in determining the predicted outcomes.

How can ML approaches be more effective in capturing the complexities of IMF lending practices than traditional statistical models like OLS or Negative Binomial regression? Firstly. ML models, such as Random Forest and Gradient Boosting, are capable of capturing non-linear relationships and complex interactions between variables that traditional regression models might miss. This flexibility allows them to better fit the data, especially when relationships are not strictly linear or when there are multiple interacting factors influencing the outcome.

Secondly, ML models can handle a larger number of predictors and potential interactions without the same risk of overfitting that would typically plague traditional statistical models. This capability allows for the inclusion of more nuanced variables that might influence IMF lending decisions, which traditional models might exclude due to concerns about multicollinearity or overfitting.

Thirdly, ML models provide mechanisms, such as feature importance rankings, that help identify which variables have the most predictive power. This can lead to a better understanding of the key drivers of IMF program design and conditionality.

Last but not least, unlike OLS or Negative Binomial regression, ML models are generally more robust to issues like overdispersion and heteroskedasticity. This robustness can lead to more reliable and accurate predictions, even when the assumptions underlying traditional statistical models are not fully met.

In this research, I employ supervised machine learning techniques, where each sample in the training dataset is labeled to indicate whether a specific conditionality category was present in a particular IMF program. The task is to predict the class label (dependent variable) for each data point, where each data point consists of a set of feature values (i.e., a feature vector), representing the independent variables in IMF program design as discussed earlier. The ML model is trained using the prepared training set, enabling it to learn the relationships between the independent and dependent variables and how variations in the independent variables collectively influence program outcomes. The model's accuracy is then evaluated by allowing it to predict the class labels of the samples in the test dataset and comparing

the predicted labels with the actual class labels. The model parameters are iteratively refined by repeating the training process until the best possible prediction results are achieved.

5. Empirical Analysis

5.1. Results from NLP Analysis

As explained in the previous section, an NLP tool for automatic extraction of Executive Board views on IMF program design has been developed for this research, which is able to process the Executive Board meeting minutes. The output of the tool is the values of the specific Executive Board-related variable defined in the theoretical model (Executive Board discontent with program design) for the different IMF programs analyzed. I designed the NLP tool specific to the Executive Board meeting minute domain, while also being responsible for the implementation of the software. Figure 1 includes sample sentences from the meeting minutes data and their corresponding sentiments (1 for positive, 0 for neutral and -1 for negative sentiments).

Figure 1: Sample Data from EB Meeting Minutes

Country	Year	Sentence	Sentiment	Author
Bulgaria	2004	Mr. Mozhin and Mr. Lissovolik submitted the following statement:	0	Mr Lissovolik;Mr Mozhin
Bulgaria	2004	We thank the staff for a comprehensive and lucid paper.	0	Mr Lissovolik;Mr Mozhin
Bulgaria	2004	Bulgaria is making important progress in consolidating the achievements of its economic reforms, which in recent periods have been duly rewarded by upgrades of credit rating agencies.	1	Mr Lissovolik;Mr Mozhin
Bulgaria	2004	At the same time the staff report rightly notes that on some of the fronts vulnerabilities have increased, most notably this concerns the high level of the external gap.	-1	Mr Lissovolik;Mr Mozhin
Bulgaria	2004	On the whole, however, we were pleased to learn from the statement by the staff representative on Bulgaria that the prior actions for the stand-by request have been met.	1	Mr Lissovolik;Mr Mozhin
El Salvador	1990	Ms. Powell made the following statement:	0	Ms Powell
El Salvador	1990	The staff notes that in the last five years, economic growth in El Salvador averaged around 1.5 percent a year, and the external position deteriorated–a result of armed conflict, adverse external developments, and political uncertainty.	-1	Ms Powell
El Salvador	1990	The new authorities in El Salvador should, therefore, be congratulated for the initiative they are taking to reduce imbalances and secure stronger economic growth in that country.	1	Ms Powell
El Salvador	1990	By their recent actions, which included a tightening of monetary policy and the unification of the exchange rate, the authorities have shown a strong commitment to this process, and the program they have outlined for 1990 and 1991 deserves our support.	1	Ms Powell
El Salvador	1990	It is clear that the ability of the authorities to successfully implement this program will depend critically on a resolution of the military conflict.	0	Ms Powell

Sentiment analysis can be considered as a text classification problem where the classes are degrees of positive and negative sentiments. I developed machine learning based sentiment analysis models to classify each sentence in the EB meeting minutes as having a positive, negative or neutral sentiment, which demonstrates how satisfied the EB board members with the program discussed in that meeting. The development of the sentiment

analysis models involved (1) development of a model that effectively combines context and sentiment information when building a text representation for improving classification accuracy (2) development of a feature ensemble model for sentiment analysis.

Sentiment analysis, a technique for automatically identifying opinions or emotions, plays a crucial role in the decision-making process. It provides insight into people's views on entities such as products, services, or organizations. In recent years, it has gained significant interest for its successful applications in a variety of fields, including product evaluation, finance, election forecasting, critical event identification, and recommendation systems. There are two main approaches to sentiment analysis: lexicon-based and machine learning-based. In the lexicon-based method, a dictionary of words and their sentiment scores is used to calculate the sentiment of a text by aggregating the scores of all words present. In the machine learning-based approach, features representing the text are extracted and then used to train a machine learning model to determine the sentiment of the text.

After the formation of the EB meeting minutes dataset, I first ran well-known ML algorithms on the dataset to evaluate their performance in order to get an understanding of the baseline accuracy results of existing algorithms. Figure 2 summarizes the results of the tested algorithms and feature extraction methods.

Figure 2: Accuracy Results of Different Feature Extraction Methods for Sentiment Analysis

Method	Number of features	LR accuracy	SVM accuracy	RF accuracy
BoW	17,527	0.852	0.851	0.822
N-gram	33,770	0.827	0.826	0.755
TF-IDF	17,527	0.853	0.854	0.816

The highest accuracy, 85.4 percent, was achieved by the Term Frequency-Inverse Document Frequency (TF-IDF) method using Support Vector Machines (SVM) as the classifier. TF-IDF was followed by Bag-of-Words (BoW) and N-gram in terms of accuracy, with N-gram extracting more features (nearly double) than BoW and TF-IDF, which increased its processing time. Among machine learning techniques, SVM achieved the highest accuracy,

but Logistic Regression (LR) performed better in two out of three methods, while Random Forest (RF) showed the worst results for all methods.

In sentiment analysis, word embeddings have been shown to be highly effective in recent studies, among various techniques for text modeling and feature extraction. Word embeddings convert words into vectors of a specific length with real numbers, where each dimension of the vector represents a feature of the word. To achieve efficient sentiment classification, it's crucial to obtain a low-dimensional and non-sparse word vector representation.

There are pre-trained word embedding models available, such as those based on large datasets like Wikipedia, commonly used in NLP tasks. However, these pre-trained word embeddings may not be suitable for the specific context of a domain-specific sentiment analysis task. Deep learning models have the ability to capture context information but require large amounts of data and computational power for training. Additionally, for sentiment analysis, word embeddings models like GloVe lack sentiment information for the words and their context. Words with contrasting sentiments, such as "happy" and "sad" or "good" and "bad", may have similar semantics and thus similar vector representations.

In an effort to address the shortcomings of existing word embeddings for the sentiment analysis on EB meeting minutes, in this project I developed a model that effectively combines context and sentiment information when building a text representation. The proposed model refines word vectors with immediate context information without relying on any specific domain, and can be employed with any pre-trained word embedding models. It further integrates sentiment scores obtained using a lexicon-based method when building the final word vectors of sentences to be classified. Additional details on our NLP tool, the accuracy results of RoBERTa-large and FinBERT, and the fine-tuned BERT, can be found in the Appendix of the paper.

5.2. Statistical Findings

The final dataset used to analyze the relation between the IMF's Executive Board sentiment and number of conditions and loan amounts covers, in 112 countries from 1976 to 2017 for the main independent and dependent variables.

5.3. OLS Regressions Results

In our empirical analysis, we estimated the relationship between our main independent variables, EB discontent and GDP growth, and the loan

amount/IMF quota using an OLS regression with robust standard errors. Our models include lagged values of EB discontent and GDP growth, their interaction term, and a set of control variables, including the natural logarithm of population, a binary variable indicating whether a country is an ally of major IMF shareholders, and the polity score of the borrowing country. We also include country and year fixed effects to account for unobserved heterogeneity across countries and over time. The regression results are presented in Table 1. Overall, the models explain a substantial proportion of the variance in the loan amount/quota, as indicated by their R-squared values.

Model 1 serves as a baseline model, including only the lagged values of EBD and GDP growth. The results show that neither EBD (Coefficient = 0.0009317, p = 0.466) nor GDP growth (Coefficient = -0.012374, p = 0.159) have significant effects on the IMF quota percentage when considered in isolation.

Models 2 through 4 introduces additional controls, including G5 ally, polity2 and log of population. The results show that being an ally of the G5 is positively and significantly associated with a larger IMF loan size (Coefficient = 3.962, $p < 0.05$).This finding is consistent with the literature suggesting that geopolitical considerations play a role in IMF lending decisions.

To test the robustness of these findings, Models 5 through 7 include interaction terms and covariates that could potentially confound the relationship between the key independent variables and the loan size/quota. Specifically, Model 5 adds the interaction term between EBD and GDP growth, which is significant (Coefficient = -0.0004275, p = 0.036). This indicates that the effect of EB discontent on IMF loan size is moderated by the economic performance of the borrowing country. Model 6 includes both the interaction term and the G5 ally variable, confirming the significant effects of these variables on IMF loan size. The interaction term remains significant (Coefficient = -0.0004452, p = 0.022), demonstrating that the political and economic conditions within the borrowing countries are crucial in determining the size of the loans. Model 7 is the most comprehensive, combining all variables and interaction terms. This model confirms the significant effects of population size (Coefficient = -5.633, $p < 0.01$), G5 ally status (Coefficient = 3.822, $p < 0.05$), and the interaction between GDP growth and EBD (Coefficient = -0.0003536, $p < 0.1$). The results highlight the importance of both domestic economic indicators and international political alliances in determining IMF loan size.

Country and year fixed effects were included in the model to control for

Table 1: OLS Regression Results

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
EBD	0.0009317	0.0000781	-0.0001091	-0.0000948	0.0016	0.0007681	0.0003801
	(0.0012779)	(0.0011519)	(0.0011219)	(0.0011343)	(0.0013122)	(0.0011778)	(0.0012018)
GDP growth	-0.012374	-0.0149505*	-0.0272506**	-0.0254585**	0.0287826	0.0278941	0.0107363
	(0.0087553)	(0.0081904)	(0.0122146)	(0.0120866)	(0.0195599)	(0.01828)	(0.0225004)
EBD X GDP growth					-0.0004275**	-0.0004452**	-0.0003536
					(0.000203)	(0.000193)	(0.0002308)
ln_population	-4.211516***	-5.86716***	-5.589558***		-4.241117***	-5.633472***	
	(1.255579)	(1.406804)	(1.413543)		(1.242561)	(1.417304)	
G5 ally			3.961849**	3.970245**		3.822201**	
			(1.897401)	(1.849661)		(1.845049)	
Polity2				-0.0422145**		-0.0397415**	
				(0.0184556)		(0.0184175)	
Observations	477	477	358	343	477	477	343
R-squared	0.7679	0.7945	0.8329	0.8160	0.7723	0.7992	0.8197

Notes: Note: These estimates are from ordinary least squares (OLS) regression. The dependent variable is the loan size divided by the borrowing country IMF quota in Year t. All models include country and year fixed effects. Robust standard errors in parentheses. ***, **, and * indicate statistical significance levels of 0.1, 1, and 5%, respectively.

time-invariant country characteristics and global trends that could influence IMF quotas. The results show that several country fixed effects are statistically significant, indicating that some countries have systematically higher or lower quotas than others, even after controlling for the variables included in the model. Year fixed effects capture the impact of global economic conditions on IMF quotas.

5.4. Negative Binomial Regression Results

In addition to the OLS regressions, we employed negative binomial regressions to further analyze the relationship between our key independent variables—Executive Board discontent and GDP growth—and the number of conditions attached to IMF loans. The NB regression is appropriate given the count nature of our dependent variable, which measures the total number of conditions in IMF programs. This model accounts for overdispersion in the count data, which was confirmed by the likelihood-ratio test comparing the NB model to a Poisson model, indicating that the variance of number of conditions exceeds its mean.

The baseline NB model, presented in Table 2 (Model 1), includes lagged values of EB discontent and GDP growth as the IVs. The results show a statistically significant positive effect of EB discontent on the number of conditions (Coefficient = 0.0241, $p < 0.01$), suggesting that higher levels of EB discontent are associated with an increased number of conditions in IMF programs. This finding is in line with our hypothesis that increased dissatisfaction within the EB leads to stricter loan conditions. On the other hand, GDP growth does not have a statistically significant effect on the number of conditions (Coefficient = -0.0664, $p = 0.478$), indicating that economic performance alone does not directly impact the stringency of IMF loan terms.

In Models 2 through 4, we introduce additional control variables, including the natural logarithm of population and an indicator for whether a country is an ally of the G5. These models aim to control for domestic and geopolitical factors that might influence the number of conditions attached to IMF loans. The inclusion of population reveals a statistically significant positive effect on the number of conditions (Model 2: Coefficient = 1.2113, $p < 0.05$).

Model 3 further includes the G5 ally variable, which captures geopolitical considerations. The coefficient for ally is negative and marginally significant

(Coefficient = -4.0447, p = 0.084), indicating that countries allied with major IMF shareholders might face fewer conditions. This finding aligns with existing literature on the role of geopolitical factors in IMF decision-making, where allied countries may receive more favorable terms.

To explore the potential interaction between economic performance and EB discontent, Models 4 and 5 introduce an interaction term between EBD and Gdp growth (both lagged). The interaction term is negative but not statistically significant in Model 4 (Coefficient = -0.0023, p = 0.199), suggesting that the effect of EB discontent on the number of conditions does not significantly vary with the economic performance of the borrowing country.

Model 5 presents the most comprehensive analysis, incorporating all control variables and interaction terms. In this model, the interaction term between EB discontent and GDP growth remains negative and marginally significant (Coefficient = -0.0026, p = 0.156), indicating a potential moderating effect of economic growth on the relationship between EB discontent and loan conditions. Additionally, the negative coefficient for ally (Coefficient = -5.6670, $p < 0.05$) persists, reinforcing the notion that geopolitical considerations influence IMF conditionality.

To ensure the robustness of our findings, we conducted several sensitivity analyses, including alternative model specifications and the inclusion of other potential confounders. The results remain consistent across different model specifications, suggesting that our findings are robust to changes in model assumptions. To better understand how EBD and economic performance influence specific types of IMF conditions, I also conducted negative binomial regressions for different categories of conditions: financial-related, fiscal policy-related, and labor-related.

The results for financial-related conditions show a positive and significant association with lagged EBD, suggesting that higher discontent among Executive Board members tends to result in more stringent financial conditions. Similarly, for fiscal policy-related conditions, EB discontent is also positively associated, reinforcing the idea that greater discontent leads to tighter fiscal requirements.

For labor-related conditions, while there is a positive association with EB discontent, the relationship is not consistently significant across all models. This suggests that the impact of EB discontent on labor-related conditions is less robust compared to financial and fiscal conditions.

These findings indicate that the effects of EB discontent and economic performance differ by condition type, with more consistent impacts observed

for financial and fiscal policy conditions. This highlights the importance of considering specific types of conditions to fully understand the nuances in how IMF programs are shaped by internal dynamics and economic factors.

Overall, the results from the Negative Binomial regression models underscore the significant and positive relationship between EB discontent and the number of conditions, coupled with the impact of geopolitical alliances, highlighting the interplay between economic and political factors in shaping IMF lending practices.

5.5. Testing Hypotheses with ML Models

In this section, the datasets and the pre-processing, parameter settings and the results of the applied machine learning algorithms are explained in detail, respectively. The machine learning models follow the statistical models reported in the Statistical Analysis section with respect to the reasoning, model building and the dataset.

5.5.1. Preliminaries

We use support vector machines (SVM) with linear kernel, random forest (rf), k-nearest neighbors (knn), Adaboost, Naive Bayes, Latent Dirichlet Allocation (lda), and Gradient Boosting with their default parameter settings from the Python scikit-learn library, to build various models in line with the statistical models described in the Statistical Section Analysis.

In order to observe the effects of the various independent variables on model performance, different combinations of the independent variables were included as features in the models. Here, the dependent variable is the number of IMF conditions, whereas the independent variables are the lagged values of EB Discontent measure (ebd), GDP growth (gdp-gr), GDP per capita (gdp-pc) and the vector of controls including population (pop), and G5-allies. The dataset is identical to the statistical models described in the previous section.

5.5.2. Preprocessing

The pre-processing needs to be done with respect to the machine learning models. The number of IMF conditions is a discrete target variable. After removing the missing observations, we are left with a total of 417 observations. The target variable has 60 unique values that range between 1 and 105 with a median value of 17. I created three categories for the target variable as follows: low: 0-17, medium: 18-40, high: >40. In order to get an equal

Table 2: NB Regressions for Conditions

Variable	(1)	(2)	(3)	(4)	(5)
EBD (lagged)	0.0241***	0.0196**	0.0239**	0.0274***	0.0254***
	(0.0076)	(0.0078)	(0.0096)	(0.0099)	(0.0079)
GDP growth (lagged)	-0.0664	-0.0684	-0.0961	0.1377	0.0221
	(0.0936)	(0.0931)	(0.1226)	(0.2211)	(0.1708)
EBD X GDP growth				-0.0023	-0.0009
				(0.0018)	(0.0014)
ln_population		1.2113**	0.9654	0.9126	
		(0.5518)	(0.7405)	(0.7373)	
G5 ally			-4.0447*	-4.2468*	
			(2.3373)	(2.3256)	
Observations	503	503	358	358	503
R-squared (overall)	0.0542	0.0624	0.0817	0.0873	0.0572

Notes: These estimates are from negative binomial regression models. The dependent variable is the total number of IMF program conditions assigned to borrowing country j in Year t. EBD and GDP growth are lagged by one year to reflect the conditions of borrowing country j in Year $t-1$. Standard errors are clustered by borrowing country and are shown in parentheses. ***, **, and * indicate statistical significance levels of 0.1, 1, and 5 percent, respectively.

number of samples from each category in ML model training, so that the results would not be biased towards any particular category, I oversampled the data, resulting in 200 samples from each category in the classification task.

5.5.3. Parameter Settings

The repeated k-fold cross-validation procedure is an established standard method to measure the performance of a machine learning algorithm. Repeating the cross-validation procedure multiple times and stating the mean result across all folds from all runs decreases the bias in the measure of the performance of a machine learning model.

We use repeated k-fold cross validation where the number of folds is 5 and the number of repeats is 5.

5.5.4. Results

The accuracy results achieved in the classification of the total number of conditionality with the different machine learning models tested are listed in Table 3. Here the leftmost column lists the name of the machine learning algorithm used for classification in each row, and the remaining column titles state the list of independent variables used as features in ML model training.

Table 3: ML Models' Accuracy Results

Algorithm	ebd, gdp-pc	gdp-gr, gdp-pc	ebd, gdp-gr, gdp-pc	ebd, gdp-gr, gdp-pc, pop	ebd, gdp-gr, gdp-pc, G5-allies	ebd, gdp-gr, gdp-pc, pop, G5-allies
rf	0.817	0.635	0.821	0.829	0.883	0.893
svm	0.429	0.377	0.438	0.500	0.511	0.556
knn	0.636	0.524	0.663	0.678	0.723	0.722
Adaboost	0.546	0.518	0.574	0.613	0.615	0.653
Naive Bayes	0.376	0.378	0.370	0.444	0.421	0.527
lda	0.402	0.382	0.418	0.450	0.515	0.529
Gradient Boosting	0.801	0.628	0.821	0.844	0.886	0.896

5.5.5. Discussion

As seen in Table 3, the ML models with the highest classification accuracy are Random Forest and Gradient Boosting, both of which are ensemble models. The accuracy achieved by these models when all of the mentioned independent variables are included as features is close to 90%, which is quite significant. This proves that ML is an important tool for analysis and prediction of IMF program design.

When we compare the effects of including different independent variables in model training, we observe that the EB discontent variable has a significant effect on accuracy, resulting in a difference of about 18% between the cases where it is included and left out (columns 3 and 4 in the table). We also observe that among the control variables, G5-allies has a significant effect on accuracy, achieving an increase of at least 6% over the case where it is not included in training, while population has a negligible effect.

6. Conclusion

This paper seeks to expand our understanding of IMF lending practices by focusing on a previously underexplored factor: the sentiments and internal dynamics within the IMF's Executive Board. By introducing a novel variable, EB discontent, derived from extensive analysis of IMF Executive Board meeting minutes using natural language processing techniques, this study provides new insights into how EB discontent influence the design and stringency of IMF programs.

While the OLS regression models reveal that while EB discontent alone does not significantly affect the size of IMF loans, the interaction between EB discontent and GDP growth is significant, indicating that the economic context of the borrowing country can influence the impact of Board sentiment on loan size. Negative Binomial regression models, on the other hand, demonstrate that higher levels of EB discontent are associated with a greater number of conditions attached to IMF loans. This suggests that internal disagreements within the Board may lead to stricter loan terms, particularly for labor, financial, and fiscal policy-related conditions. While the OLS and Negative Binomial regression models have provided valuable insights into the factors associated with IMF loan size and conditionality, their relatively low

R-squared values indicate that these models do not capture the full complexity of IMF lending decisions, however.

In contrast, the machine learning models, particularly Random Forest and Gradient Boosting, demonstrated high accuracy in predicting the number of conditions in IMF programs, achieving close to 90 percent accuracy in some cases. The inclusion of EB discontent as a predictor significantly improved model accuracy, further validating the relevance of internal Board dynamics in shaping IMF lending decisions. This suggests that ML approaches may be more effective in capturing the patterns and interactions that characterize IMF lending practices. The superior performance of ML models likely stems from their ability to handle non-linear relationships, manage high-dimensional data, and identify important features without being constrained by the assumptions required for traditional statistical models.

The findings of this study highlight the potential of machine learning techniques to enhance our understanding of IMF lending practices and offer robust tools for predicting program outcomes. Future research should continue to explore the use of machine learning in IPE, particularly in contexts where complex, non-linear relationships are likely to play a crucial role.

Overall, this research highlights the importance of considering a more holistic set of factors—spanning internal organizational dynamics, economic performance, and geopolitical alliances—when analyzing IMF lending practices. By integrating a novel variable capturing EB sentiment, this study not only fills a critical gap in the literature but also offers a robust framework for future research in IPE. The combination of traditional econometric methods with advanced machine learning techniques opens new avenues for understanding and predicting the complex decision-making processes within international financial institutions like the IMF.

Future studies could build on this framework by exploring additional internal and external factors that influence IMF lending, employing more granular data, and further refining the methodologies used to capture and analyze EB sentiment. Such efforts will deepen our understanding of how international financial governance is shaped by a complex interplay of political, economic, and institutional forces.

Acknowledgements

I would like to express my deep appreciation to over 80 students who annotated more than 50 thousand pages for this research since 2020. Simay Ayd-

eniz, Yusuf Hayta, Mariam Manaullah, Ozlem Ozmen, and Aleyna Temelkaya's assistance in data collection is gratefully acknowledged. Special thanks go to Daniele Aglio, Gulce Elif Atabey, Celikhan Baylan, Ayca Deniz, Eylem Guner, and Mustafa Yahsi for their research assistance. I also would like to thank Pelin Angin Ulkuer, Hakan Emekci, and Gokhan Gunes for their valuable feedback throughout the project.

Funding

This research was funded by H2020 Marie Skłodowska-Curie Actions (H2020-MSCA-IF2019) grant number 896716, and the 2232 International Fellowship for Outstanding Researchers Program of TÜBİTAK (Project No: 118C309). The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

Ahn, K.S., Oh, Y.H., Kim, J.T., Kim, M.J., 2011. A comparative study of financial crisis prediction using ml models. Journal of the Korean Institute of Information Scientists and Engineers 38, 2966–2972.

Angin, M., Shehaj, A., Shin, A.J., 2023. Imf: International migration fund. International Interactions 49, 86–113. doi:10.1080/03050629.2023.2172002.

Barro, R.J., Lee, J.W., 2003. Imf programs: Who is chosen and what are the effects? Journal of Monetary Economics 52, 1245–1269.

Bird, G., 2001. Imf programs: Do they work? can they be made to work better? World Development 29, 1849–1865.

Breen, M., 2014. Imf conditionality and the economic exposure of its shareholders. European Journal of International Relations 20, 416–436. doi:10.1177/1354066112448257.

Chwieroth, J.M., 2010. Capital Ideas: The IMF and the Rise of Financial Liberalization. Princeton University Press, Princeton, NJ.

Copelovitch, M.S., 2010a. The International Monetary Fund in the Global Economy: Banks, Bonds, and Bailouts. Cambridge University Press.

Copelovitch, M.S., 2010b. Master or servant? common agency and the political economy of imf lending. International Studies Quarterly 54, 49–77.

Dreher, A., 2004. A public choice perspective of imf and world bank lending and conditionality. Public Choice 119, 445–464.

Dreher, A., Sturm, J.E., Vreeland, J.R., 2015. Politics and imf conditionality. Journal of Conflict Resolution 59, 120–148.

Edwards, S., 2003. Debt relief and the current account: An analysis of the hipc initiative. World Economy 26, 513–531.

Gutner, T., 2005. World bank environmental reform: Revisiting lessons from agency theory. International Organization 59, 773–783.

Helleiner, E., 1987. The imf and africa in the 1980s. World Development 15, 681–687.

James, H., 1996. International Monetary Cooperation Since Bretton Woods. Oxford University Press, Oxford.

Jung, H.S., Patnam, M., Ter-Martirosyan, A., 2018. Ml techniques in economic forecasting: A comparative study. IMF Working Papers 18, 1–23.

Kang, S., 2007. Agree to reform? the political economy of conditionality variation in international monetary fund lending, 1983-1997. European Journal of Political Research 46, 685–720.

Kentikelenis, A.E., Stubbs, T.H., King, L.P., 2016. Imf conditionality and development policy space, 1985–2014. Review of International Political Economy 23, 543–582. doi:10.1080/09692290.2016.1174953.

Killick, T., 1995. IMF Programmes in Developing Countries: Design and Impact. Routledge, London.

Kim, H.D., Lee, J.H., Oh, Y.H., Kim, J.T., 2009. Prediction of financial crises using ml techniques. Expert Systems with Applications 36, 260–265.

Lang, V., 2021. The economics of the democratic deficit: The effect of imf programs on inequality. The Review of International Organizations 16, 599–623. URL: https://doi.org/10.1007/s11558-020-09405-x, doi:10.1007/s11558-020-09405-x.

Lin, C.Y., Khan, M.J., Chang, C.L., Wang, K.H., 2008. Forecasting currency crises using ml techniques. Neurocomputing 71, 1098–1104.

Martin, L.L., 1991. Coercive Cooperation: Explaining Multilateral Economic Sanctions. Princeton University Press, Princeton, NJ.

Metinsoy, S., 2024. Who adjusts? exchange rate regimes and finance versus labor under imf programs. The Review of International Organizations doi:10.1007/s11558-024-09540-9.

Momani, B., 2007. Imf staff: Missing link in fund reform proposals. Review of International Organizations 2, 39–57.

Nelson, S.C., 2017. The Currency of Confidence: How Economic Beliefs Shape the IMF's Relationship with Its Borrowers. Cornell University Press, Ithaca, NY.

Rickard, S.J., Caraway, T.L., 2014. International negotiations in the shadow of national elections. International Organization 68, 701–720.

Son, S.R., Oh, Y.H., Kim, J.T., Kim, M.J., 2009. Predicting stock market volatility using ml models. Journal of Computational Finance 39, 4951–4957.

Stiglitz, J.E., 2003. Globalization and Its Discontents. W.W. Norton & Company, New York.

Stone, R.W., 2002. Lending Credibility: The International Monetary Fund and the Post-Communist Transition. Princeton University Press, Princeton, NJ.

Stone, R.W., 2008. The scope of imf conditionality. International Organization 62, 589–620.

Stone, R.W., 2011. Controlling Institutions: International Organizations and the Global Economy. Cambridge University Press, Cambridge.

Taylor, L., 1987. Varieties of Stabilization Experience: Towards Sensible Macroeconomics in the Third World. Clarendon Press, Oxford.

Thacker, S.C., 1999. The high politics of imf lending. World Politics 52, 38–75.

Tiffin, A., 2016. Ml in macroeconomic forecasting. IMF Working Papers 16, 1–32.

Vaubel, R., 1986. A public choice approach to international organization. Public Choice 51, 39–57.

Vreeland, J.R., 2003. The IMF and economic development. Cambridge University Press.

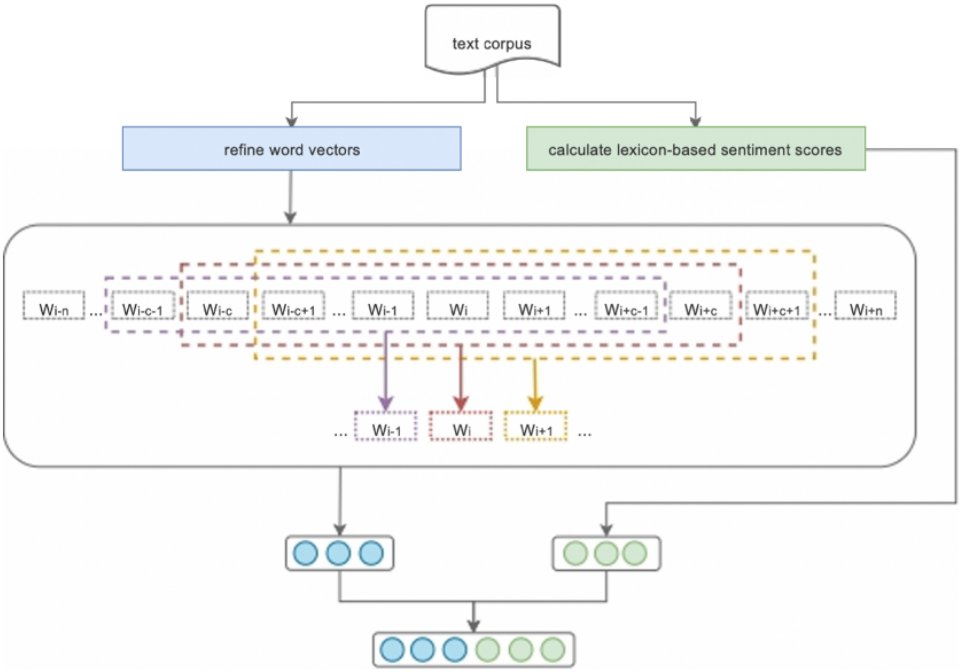
Woods, N., 2006. The globalizers: the IMF, the World Bank, and their borrowers. Cornell University Press.

Zadeh, R.B., Zollmann, A., 2009. Predicting market-volatility from federal reserve board meeting minutes nlp for finance. URL: https://stanford.edu/ rezab/papers/finlpreport.pdf. accessed: 2024-08-30.

Appendix A. NLP Tool

My proposed model for word embeddings refines word vectors obtained using effective pre-trained models in two steps: contextual refinement and valence addition. Figure A.3 shows a graphical representation of how the model works.

Figure A.3: Proposed model for sentiment and context-refined word embeddings



The sentiment analysis task relies heavily on the context of words, which is defined by the words around them. The proposed model seeks to incorporate this context information by utilizing the word vectors of the preceding and following words. To accomplish this, the vector values for each token are updated by taking the average of the vectors for c neighboring words. This approach not only adds context information but also addresses the issue of unknown words, as these words can be represented through the vectors of their surrounding words.

Figure A.4 demonstrates an example of context refinement for a sentence. The table displays the pre-trained GloVe word embeddings and their refined versions based on context. The sample sentence, “This is a truly truly bad

Figure A.4: Sample Context Refinement in Sentiment Analysis

word	plain word embeddings	context-refined word embeddings
this	[-0.204, 0.164, 0.042, -0.137, -0.298, ...]	[-0.225, 0.163, 0.065, -0.229, -0.202, ...]
is	[-0.175, 0.230, 0.249, -0.205, -0.123, ...]	[-0.102, 0.131, 0.100, -0.264, -0.056, ...]
a	[-0.297, 0.094, -0.097, -0.344, -0.185, ...]	[-0.028, 0.112, 0.121, -0.285, 0.032, ...]
truly	[0.267, 0.035, 0.206, -0.369, 0.383, ...]	[0.074, 0.053, 0.097, -0.260, 0.062, ...]
truly	[0.267, 0.035, 0.206, -0.369, 0.383, ...]	[0.082, -0.017, 0.048, -0.221, 0.113, ...]
bad	[0.309, -0.127, -0.078, -0.011, -0.146, ...]	[0.176, -0.045, 0.085, -0.190, 0.188, ...]
movie	[-0.138, -0.122, 0.005, -0.010, 0.131, ...]	[0.146, -0.071, 0.044, -0.130, 0.123, ...]

movie.", is taken from the Stanford Sentiment Treebank dataset and the neighbor radius c is set at 2 for this example. The refined embedding for the first occurrence of the word "truly" is calculated by averaging the pre-trained vectors of "is", "a", "truly", "truly", and "bad". For example, the first dimension of the refined vector for "truly" is calculated as follows: ((-0.175)+(-0.297)+(0.267)+(0.267)+(0.309)) / 5 = 0.074. If there are fewer words than c in the vicinity, the maximum number of available neighbors are used in the calculation. For instance, the refined embedding for the word "this" is determined by averaging the vectors of "this", "is", and "a". Similarly, the refined embedding for "is" is computed by taking the average of the vectors for "this", "is", "a", and "truly". In the original GloVe model, each word has a unique vector, regardless of its context, e.g. the word "truly" has the same vector in both instances in the sample sentence. However, after context refinement, the two occurrences of "truly" have different vectors as context information is incorporated based on the word's position in the sentence and its neighboring words.

Additionally, I believe that incorporating sentiment information in addition to contextual information can enhance prediction accuracy. To this end, the second step of the model involves incorporating sentiment predictions from VADER. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a sentiment analysis tool based on a lexicon, accessible at https://github.com/cjhutto/vaderSentiment. It computes four ratios (positive, neutral, compound, and negative) for a given text. An example result would be: 'positive': 0.706, 'compound': 0.9469, 'neutral': 0.294, 'negative': 0.0. The compound score is computed in three steps: by first summing up the valence scores of all words in the text, adjusting the result with rules such as the presence of a booster word, and finally normalizing the score

between -1 and 1, indicating strongly negative and strongly positive, respectively. Sentences are classified as positive if the compound score is greater than or equal to 0.05, negative if the score is less than or equal to -0.05, and neutral if the score falls between -0.05 and 0.05.

Finally, the proposed domain-specific word embeddings are created by combining the contextually refined word vectors with sentiment predictions. This combination is achieved through concatenation. For machine learning algorithms like Logistic Regression, the final word embeddings are obtained by concatenating the context-refined word embeddings and the lexicon-based sentiment scores. Model training is then carried out using these refined word embeddings. For deep learning models such as Convolutional Neural Networks (CNN), the concatenation operation can occur at different points in the network, depending on the architecture. In the experiments with CNN, the two inputs were concatenated at the start of the model and all neural network layers were performed on the refined word embeddings.

While deep learning models can intrinsically capture the context information, they require a high amount of data and computational power to train. On the other hand, BERT is a context-sensitive language model which has become popular lately with its high performance on various NLP tasks. It provides pre-trained deep learning models that can be used in many tasks, including sentiment classification. Therefore, I tried to improve the performance of BERT on the sentiment analysis task.

In Table A.4, I present the accuracy results achieved by two pre-trained BERT models, namely RoBERTa-large and FinBERT, and the fine-tuned BERT model built by utilizing the annotated data from the EB meeting minutes. Among pre-trained models, RoBERTa-large achieves the maximum accuracy of 81.9%. When we fine-tune the BERT model with randomly selected 80% of our data, we achieve 97.2% accuracy in the remaining test data. Fine-tuning increases the performance significantly; however, the execution time of it is much more than applying the pre-trained models.

Table A.4: Accuracy results of BERT

Model	Accuracy	Execution time (min)
RoBERTa-large	0.819	81
FinBERT	0.609	24
Fine-tuned BERT	0.972	937

Experiment results show that highly accurate results are only achieved when the training data includes data annotated specifically for the domain in question in addition to the data of more generic pre-trained models. This is most likely due to the usage of technical terminology specific to the domain (IMF in this case) and the different usage of language to maintain political correctness.

In order to improve the performance of pre-trained BERT models I proposed a sentiment-aware unsupervised method that utilizes the confidence scores provided by pre-trained BERT models and propagates the sentiment information through similarity information.

Research on sentiment analysis is rapidly evolving, especially in terms of supervised learning algorithms. Various approaches have been proposed to build sentiment classification models, including feature-based or neural-network-based models. Many studies improve the performance of their approach by incorporating internal or external knowledge into existing models such as sentiment-aware BERT or knowledge-enabled BERT. However, unsupervised learning algorithms could be more practical in many settings, especially when there is a lack of experts to perform the labeling. Unsupervised learning studies mostly focus on statistical approaches that utilize lexicon-sentiment pairs. Some of the well-known lexicon-based sentiment analysis tools are SentiStrength, SentiWordNet, and VADER. Other than the lexicon-based approaches, BERT provides an off-the-shelf language model that does not necessarily require labeled data, which makes it appealing for unsupervised NLP tasks. Moreover, when training is not possible, self-training may boost pre-training. There exist many studies that improve unsupervised learning performance via self-training. Although they perform better than the lexicon-based approaches, they execute slower due to the nature of the training process. There also exist studies that utilize co-occurrence information to improve the performance of the classification task. In this project, I employed a similar strategy by propagating sentiment information retrieved from a pre-trained BERT model from the most confident estimate to the least in a recursive way.

The approach consists of two stages, as presented in Figure A.5. The first stage involves the utilization of a pre-trained BERT model fine-tuned for the sentiment classification task to predict all sentences as positive or negative. Then a threshold value is determined based on a predetermined percentile. The model assumes that all instances having a confidence score larger than this threshold are correctly identified by the pre-trained model and their

predicted values are assigned as their final predictions. This assignment concludes the first stage.

In the second stage of the algorithm, the information within the sentences having final predictions is propagated onto the sentences that are yet to be finalized. For this purpose, the k-Nearest Neighbours (k-NN) algorithm is used to assign the final predictions of the non-finalised instances. In the implementation, the pre-trained BERT model's prediction is considered as the first neighbor regardless of its confidence score. The remaining k-1 neighbors are selected among the finalized instances. Additionally, at the beginning of the second stage, sentences are sorted based on their confidence scores in descending order. The reason for sorting is to process the sentences with higher confidence scores first. Accordingly, those instances may contribute to the prediction of the sentences with lower confidence scores.

In order to apply the k-NN algorithm to propagate the information to the similar sentences, we need to be able to calculate the similarities between them. For this purpose, I leverage two measurements. BERT does not map sentences to a vector space considering the common similarity measures, such as cosine similarity. In other words, the cosine similarity of two sentences does not imply any meaningful information in terms of how similar they semantically are. To overcome this issue Sentence-BERT (or SBERT) was proposed, which fine-tunes BERT in this regard. In my proposed method, SBERT is utilized to calculate the similarities between sentences. The cosine similarity between the vector embeddings of the sentences constitutes the first metric.

SentiWordNet is used as the second metric to calculate the distances between the sentiments of the sentences. For this purpose, all the nouns, adjectives, and adverbs in the sentences are found and the rest are filtered out. Then, the words are lemmatized and queried over in WordNet. If the lemma exists in WordNet, the positive and negative sentiment scores of its most common definition from SentiWordNet are retrieved. The positive and negative scores of all lemmas are summed separately and concatenate into a vector as <positivescore, negativescore>. The cosine similarity between these vectors constitutes the second metric.

I present the accuracy results in Table A.5. This table also presents the ablation study that reflects the effects of including each step in the proposed method. From the table, we see that the lexicon-based SentiWordNet has the lowest accuracy, and it performs merely better than random guessing. The pre-trained BERT model substantially outperforms SentiWordNet. The pos-

itive effect of propagating the sentiment information from highly confident instances to the neighboring low confident instances is clearly seen on the third line, i.e., proposed method without initial sorting. However, it is outperformed by the proposed method, which sorts and processes the unlabeled sentences in descending order according to their confidence scores.

Table A.5: Accuracy Results of Proposed Confidence Propagation Algorithm

Model	Accuracy
SentiWordNet	0.637
Pre-trained BERT model	0.785
Proposed method without initial sorting	0.829
Proposed method	0.835

As the next step, I completed building the feature ensemble model. The model leverages context and valence information to improve sentiment classification performance. Context information has a critical importance for sentiment analysis as a word may have different meanings along with different sentiments in different contexts, e.g., an increase in inflation vs an increase in the gross domestic product (GDP). Fundamentally, the model creates an ensemble by combining various feature types. This process involves several steps as follows. First, I generate vector representations of the sentences using Bag-of-Words after I clean them with common preprocessing steps: removing punctuation marks, removing extra whitespaces, converting to lowercase, and removing stop words. With this representation, every unique word becomes a feature. To eliminate uninformative features, I employed a filter-based feature selection technique. Simply, I calculate the information gain (IG) values of all features and select the third quartile value as the threshold. Then, we remove all the features whose IG values are lower than this threshold. The filtered vectors constitute the first part of our feature pool. I also utilize these vectors in the second step of the algorithm, where I involve the context information of the words. At this point, a network from the features is built by calculating their co-occurrence relation within a specified window, also called their context range, as the contexts of words are built around their surrounding words. In this graph, nodes represent words, and edges represent the sentiment intensity between the two words. If two words appear together within the context range in a sentence, their relation increases or decreases by one unit regarding the sentence's sen-

timent, positive or negative, respectively. To provide a clear understanding of this procedure, I present the graph of a sample dataset in the Figure A.6. The sample dataset consists of the following four sentences:

- The reduction in the budgetary wage bill increased unemployment. (Negative)
- Consolidating into one basic wage also increased transparency. (Positive)
- Inflation has increased to more than 50 percent. (Negative)
- Reforms to increased fiscal transparency are welcome. (Positive)

To keep the graph simple, I set the context range as two. The nodes in the graph represent the words in the sentences after the data-cleaning process. The edges between the nodes carry the sentiment-oriented relation. The sign of the edge weights represents the polarity of the sentiment, while the magnitude represents how strong the relationship is. For example, the edge weight between the nodes increased and transparency indicates a strong relationship towards the positive sentiment. Other than that, the relationship between the nodes increased and wage is noteworthy with its weight of zero. The reason for this is that there exist two sentences with contradicting sentiments in our sample dataset that have both words within their context range. Accordingly, the relation between these nodes does not carry conclusive sentiment information for this domain. As seen from the sample figure, we obtain a domain-specific representation that contains sentiment-oriented relationship information.

We obtain encoded relationships of the words by projecting this graph into low-dimensional space with Node2Vec. Node2Vec is a recent technique for embedding graph-like data into machine learning models. It takes a graph as input, analyzes the existences and weights of edges between the vertices in the graph, and produces a fixed-length vector for each vertex of the graph as its output. Finally, we compute the sentence vector by averaging the vectors of all words in the sentence.

After incorporating the context information, I also include valence information in our model. For this purpose, I first execute a pre-trained BERT model (available at https://huggingface.co/siebert/sentiment-roberta-large-english) on our original sentences. Then, I convert the labels and confidence

scores provided by the model into features. In addition, I include the VADER score in our feature pool in order to enrich the feature ensemble model in terms of valence information. The addition of this score concludes the external computation required for the feature pool. As the final step, I combine the vectors in the feature pool by concatenating them and provide this combined vector as input for the machine learning techniques.

Figure A.7 gives the ablation study accuracy results of the IMF Executive Board Meeting Minutes dataset. More specifically, it contains the performance results after including each feature type of the proposed model one by one for varying Node2Vec feature vector sizes (64 and 128) and machine learning techniques (LR for Logistic Regression and SVM for Support Vector Machines). The included feature types are as follows. I begin with the Bag-of-Words results as the baseline method (BoW). Then, I present the results after applying feature selection with information gain (IG). The third row (Node2Vec) is for the feature vectors obtained from the context-aware graph-based representation utilizing Node2Vec. The following row gives the accuracy results after concatenating the feature vectors of IG and Node2Vec. In the fifth row, I include the BERT features into the feature pool. Finally, in the last row, I enhance the pool with the VADER features.

The maximum accuracy, 87.1%, is achieved by the proposed model with the Node2Vec feature vector size of 128 and the SVM classifier. It is clear from the table that extending the feature pool with new feature vectors increases the classification performance, regardless of the selected Node2Vec feature vector size or classifier.

Table A.6 shares the comparison results with off-the-shelf methods employed in the proposed model; BERT and VADER. The proposed model achieves higher accuracy than off-the-shelf methods.

Table A.6: Comparison results of BERT and VADER

Model	Accuracy
BERT	0.8131
VADER	0.6545
Proposed feature ensemble model	0.8714

Figure A.5: Proposed Confidence Propagation Algorithm for BERT

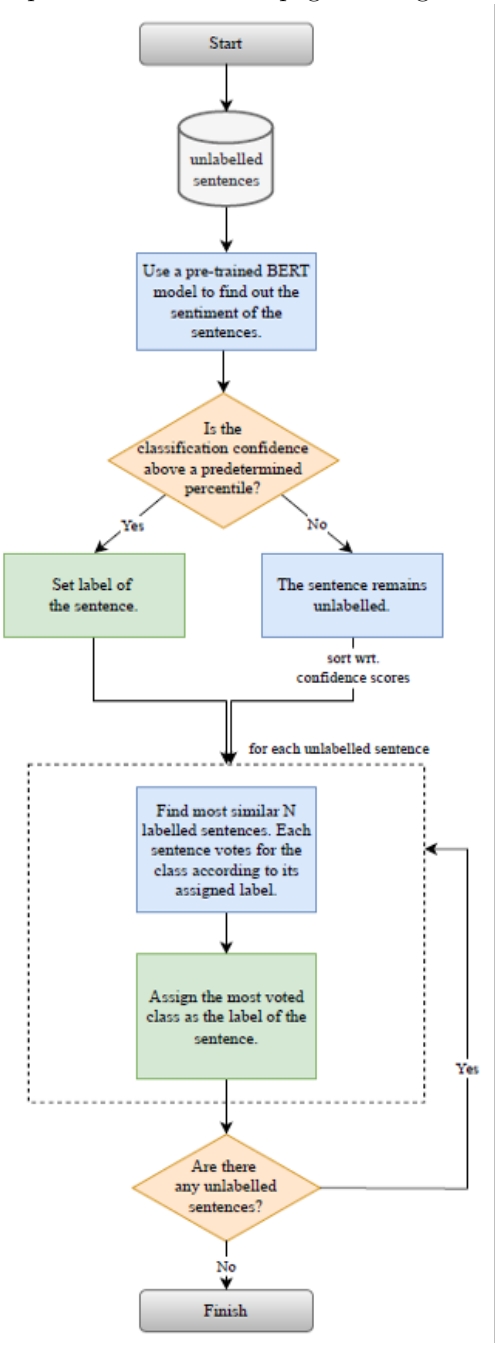


Figure A.6: Example Feature Network for Ensemble Model

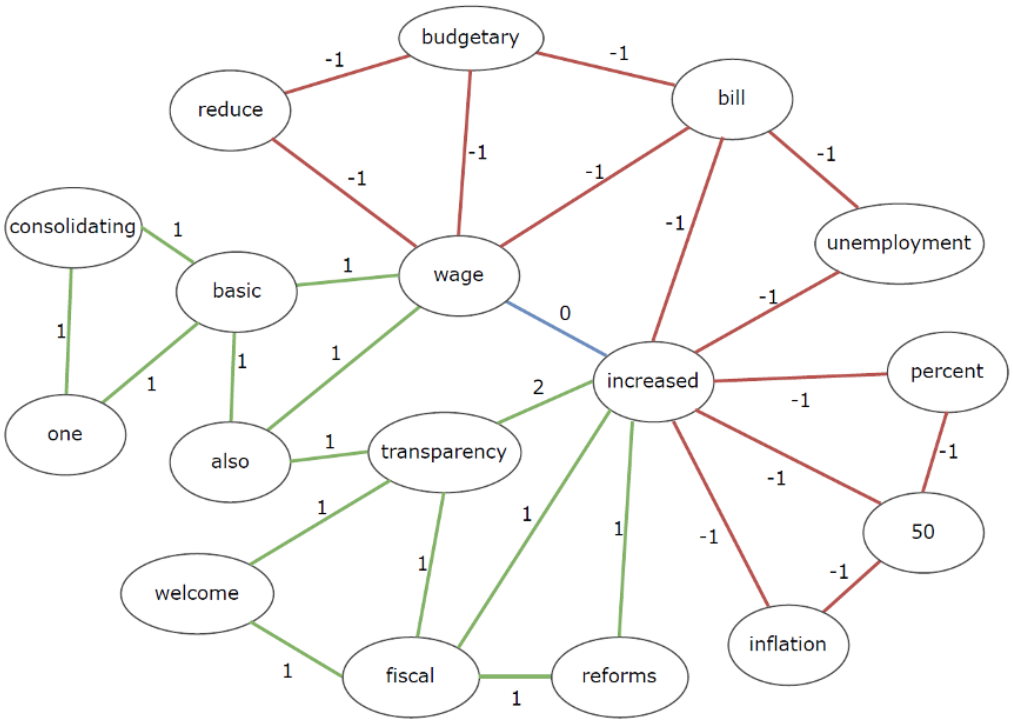


Figure A.7: Sentiment Analysis Accuracy Results on Executive Board Meeting Minutes Dataset

Method	**Node2Vec size**			
	64		**128**	
	LR	**SVM**	**LR**	**SVM**
BoW	0.8212	0.8212	0.8212	0.8212
IG	0.8331	0.8345	0.8331	0.8345
Node2Vec	0.8193	0.8188	0.8241	0.8232
IG + Node2Vec	0.8378	0.8391	0.8388	0.8393
IG + Node2Vec + BERT	0.8687	0.8690	0.8696	0.8698
IG + Node2Vec + BERT + VADER (Proposed model)	0.8696	0.8704	0.8704	0.8714