Quantifying Formality in Economic Texts: A Novel Adaptation of the Formality Formula for Economic Data Analysis

HARSH MEHTA, SANTOSH KUMAR BHARTI, NISHANT DOSHI
Computer Engineering,
Pandit Deendayal Energy University,
Raisan, Gandhinagar,
INDIA

Abstract: - Automatic text summarization and formality analysis have been extensively studied in linguistics and natural language processing (NLP). However, their application to economic data remains underexplored. Economic texts, such as policy documents, financial reports, and news articles, often exhibit varying levels of formality that influence decision-making and communication. This study introduces an innovative adaptation of a formality formula tailored specifically for economic data. Our modified formula incorporates numerical values, domain-specific keywords, and weighted grammatical features to quantify the formality of economic texts. By applying this formula to diverse datasets, including central bank policy statements and social media discussions, we demonstrate its effectiveness in identifying formal and informal tones. Statistical validation reveals that our approach achieves significant improvements in distinguishing formal texts from informal ones, with applications in sentiment analysis, readability assessment, and document classification. The findings underscore the importance of adapting linguistic tools to specialized domains like economics, paving the way for more nuanced text analysis in financial and policy contexts.

Key-Words: - Text Formality Analysis, Economic Discourse, Linguistic Quantification, Financial Communication, Domain-Specific Formality, Monetary Policy Language, Computational Linguistics, Economic Text Classification, Formality Formula Adaptation, Natural Language Processing in Economics.

Received: September 23, 2024. Revised: March 19, 2025. Accepted: April 6, 2025. Published: May 2, 2025.

1 Introduction

The Textual data contains very important information related to financial data, whether it could be the Bank reports or the informal discussions in social media. Interpreting and get useful information from the financial textual data is crucial. Bank reports may signal authority and a precise tone to convey the financial details, whereas other discussion arena like social media may contain financial details in informal text. However, the tone and formality of these texts can vary significantly, influencing their credibility, clarity, and impact.

In linguistics, formality has been explored thoroughly with the English language. Formality research is more focused on using grammatical structure and word frequencies, [1]. We have curated a formality formula for the economic reports or the financial analyses. As general-purpose formality formula fails to address domain-specific data. Our formality formula incorporates features unique to economic texts, such as numerical values, domain-specific keywords, and weighted grammatical tags.

We have considered the base work done by [1], who introduced a formality score based on part-of-speech frequencies.

We extend this framework by integrating: Numeric values, domain-specific keywords, and weighted features. Numeric values are important for the economic data as they contain GDP growth and inflation percentages. Domain-specific keywords like ”monetary policy,” ”fiscal stimulus,” and ”inflation” do have more weight in financial data. And weighted features give more importance to nouns, adjectives, and formal sentence structures. We aim to achieve quantifying economic text, identify the pattern in tone and professionalism across different types of documents, and provide actionable insights for policymakers, analysts, and researchers with our custom-made formality formula.

This paper is structured as follows: Section 2

reviews related work on formality analysis and its applications. Section 3 details the methodology, including the adapted formula and preprocessing steps. Section 4 presents experimental results and discusses their implications. Finally, the Discussion and Implications have been described in Section 5, and Section 6 concludes with future research directions.

2 Literature Survey

There has been a thorough study of formal analysis in linguistics and natural language processing(NLP). [1] introduced a formality score based on part-of-speech frequencies, which has been widely adopted for analyzing text tone. The formula is made for a common purpose, and it does not target domain-specific text like financial documents, where a lot of numeric data and some special economy-related terminologies are present. However, the research has not stopped making domain-specific formulas. For legal text, Researchers have adapted linguistic tools to analyze legal documents, focusing on formal sentence structures and technical jargon, [2]. For medical reports, domain-specific features have also been considered. A recent study introduced a multi-method framework including Medical-Enriched Deep Learning (MEDEL) to analyze the helpfulness of caregiver-generated content, showing that less formal medical language improves user perception in senior care communities, [3]. This highlights the nuanced role of formality in domain-specific communication, suggesting that informality can enhance perceived utility in certain contexts. On the other hand, research focusing on informal digital text, especially from social media, indicates that formality-based approaches often overlook the presence of emojis, sarcasm, and domain-specific cues, which are essential in sentiment detection tasks. These insights reinforce the importance of adapting sentiment and formality tools to the structure and style of informal language, [4]. Still, the economic field remained untouched. Like policy documents, financial reports contain unique features like numeric values, a lot of domain-specific keywords, and a very structured format that requires a custom format. Our article aims to provide a formula for adapting these features to the economic text. Recent studies have also explored formality in financial communication. [5] this study presents a methodology for assessing the clarity of central bank communications, specifically analyzing the European Central Bank's (ECB) written materials. The authors find that clearer communication correlates with better public understanding, emphasizing the importance of readability in conveying economic policies effectively. [6] demonstrated that the tone of media articles around earnings announcements provides incremental information useful for price discovery, emphasizing the practical relevance of formality in financial disclosures. [7] analyzed the content of earnings calls, revealing how variations in corporate messaging affect firms' financial performance, particularly during periods of economic uncertainty. Recent studies have also applied quantitative methods to economic phenomena. For instance, nonlinear autoregressive models have been used to analyze the impact of foreign trade on economic growth in Sudan [8], while the relationship between tourism and economic growth in Greece has been explored using autoregressive distributed lag approaches, [9]. More recently, researchers applied machine learning and deep learning techniques to classify economic text into positive, negative, and neutral sentiment categories. Despite modest accuracy across models, this work demonstrates the increasing interest in deriving public opinion insights from economic discourse using computational methods, [10]. These applications demonstrate the growing interest in quantitative methods for economic discourse analysis, aligning with our goal of adapting linguistic tools to specialized economic contexts.

3 Methodology

3.1 Formality Formula

[1] gave the original formula, which is given in Equation 1.

$$F = (nounfrequency + adjectivefrequency + prepositionfrequency + articlefrequency - pronounfrequency - verbfrequency - adverbfrequency - interjectionfrequency + 100)/2 \quad (1)$$

After considering the numeric values, domain-specific keywords, and weightage features, we have modified the formula, which is given in Equation 2.

$$F = (w1 * nounfrequency + w2 * numericfrequency + w3 * keywordfrequency - w4 * pronounfrequency - w5 * verbfrequency)/2 \quad (2)$$

Here, numeric frequency contains the Number/Digit counts of the sentence. Keyword frequency contains domain-specific words like "GDP", "inflation", etc. Weight values(w1,w2..) are experimentally determined to reflect feature importance.

3.2 Preprocessing Steps

We follow the following preprocessing steps: Tokenization: Split text into words and phrases. Part-of-Speech Tagging: Identify nouns, verbs, adjectives, etc., using NLP libraries like SpaCy or NLTK. Numeric Extraction: Use regular expressions to extract numerical values. Keyword Matching: Match words against a predefined list of economic keywords. Stopword Removal: Remove common stopwords unless they contribute to formality. Based on domain expertise, we assign weights as follows:w1=1.0 (nouns),w2=2.0 (numerical values),w3=1.5 (keywords),w4=1.0 (pronouns),w5=0.5 (verbs).

3.3 Economic Keyword Dictionary

To identify domain-specific terminology, we compiled a dictionary of economic keywords through expert consultation and corpus analysis. The dictionary includes:

Macroeconomic terms: GDP, inflation, unemployment, recession, fiscal policy Financial terms: bond yields, interest rates, equity, volatility, liquidity Institutional terms: central bank, Federal Reserve, treasury, IMF, World Bank Policy-related terms: monetary policy, quantitative easing, austerity, stimulus.

Each keyword was assigned a formality weight based on its prevalence in formal economic documents versus informal discussions.

3.4 Numerical Pattern Recognition

We developed specialized regular expressions to capture economic numerical patterns:

Percentages (e.g. "2. 5%", "inflation of 7 percent") Monetary values (e.g. "$10 million", "€50 billion") Indices (e.g., "S&P 500 at 4,200") Statistical measures (e.g., "p-value < 0.05", "r^2 = 0.78")

These patterns were given higher weights in the formality calculation as they tend to appear more frequently in formal economic analyses.

4 Experiment

4.1 Dataset

We evaluate our approach on three datasets:

Central Bank Policy Statements: Formal documents issued by central banks. Financial Reports: Annual reports of publicly traded companies. Social Media Discussions: Informal posts about economic topics (e.g., Reddit threads).

4.1.1 Dataset Statistics

Corpus statistics are shown in Table 1 (Appendix).

4.2 Evaluation Metrics

We have evaluated using the following evaluation metrics:

Formality Scores are calculated using the adapted formula. Statistical Significance is calculated using paired t-tests to compare scores across document types. Paired t-tests provide higher sensitivity for detecting changes in dependent samples. Hence, we have zeroed in on paired t-text. All the Case Studies Sentiment and readability analysis have been done to validate findings.

4.3 Results

The following results we have received:

4.3.1 Formality Score Distribution

For the Central Bank Statements, the Average formality score is 8.5 (highly formal). High frequency of numerical values and keywords like "monetary policy." The standard deviation is 0.7, indicating consistency in a formal tone. For the Financial Reports, the Average formality score is 7.2 (moderately formal). Balanced use of nouns, verbs, and domain-specific terms. The standard deviation is 1.3, showing more variation than central bank documents. Where in Social Media Discussions, the Average formality score is 3.0 (informal). Low presence of numerical values and high use of pronouns. Standard deviation is 2.1, reflecting high variability in formality.

4.3.2 Comparative Analysis

Our adapted formula effectively distinguishes formal from informal texts. For example, A 15% increase in formality scores for central bank statements compared to financial reports. A 60% decrease in scores for social media discussions, highlighting their informal nature. These results demonstrate the utility of our approach in quantifying formality in economic

texts, with applications in sentiment analysis, readability assessment, and document classification.

4.3.3 Feature Importance Analysis

We conducted feature ablation studies to determine the relative importance of different components in our formula. This analysis confirms that numerical values and domain-specific keywords are the most critical features for accurately measuring formality in economic texts. Table 2 (Appendix) shows the feature importance analysis.

4.4 Case Studies

4.4.1 Federal Reserve Policy Statements

We analyzed 50 Federal Reserve policy statements from 2018-2023, finding is like the Average formality score is 8.7, a High correlation ($r=0.82$) between formality and market volatility following releases, and Significant increase in formality during economic crises (9.2 during COVID-19 vs. 8.3 pre-pandemic).

4.4.2 Earnings Call Transcripts

Analysis of 100 earnings call transcripts revealed which is Prepared statements have, average formality score of 7.6. Q&A sections, average formality score of 5.8. Higher formality correlated with positive market reactions ($p<0.01$).

5 Discussion and Implications

5.1 Theoretical Contributions

Our research extends the traditional formality formula in several important ways:

Domain adaptation for economic texts, Integration of numerical pattern recognition, Weighted approach to feature importance, Validation across diverse economic document types.

These contributions address a significant gap in the literature regarding the specialized nature of economic communication.

5.2 Practical Applications

The adapted formality formula enables several practical applications:

Automated classification of economic document types, Prediction of market reactions to formal communications, Assessment of communication clarity for diverse audiences, and Tracking of formality changes over time in institutional communications.

5.3 Limitations and Challenges

Despite its effectiveness, our approach has several limitations:

Language dependency (currently optimized for English), Sensitivity to domain shifts (e.g., emerging economic terminology), Computational complexity of processing large financial documents, Potential biases in the keyword dictionary.

6 Conclusion

This study introduces a novel adaptation of the formality formula for economic text analysis. By incorporating numerical values, domain-specific keywords, and weighted grammatical features, our approach effectively quantifies the formality of various economic documents. The significant differences in formality scores across central bank statements, financial reports, and social media discussions validate the utility of our method. Future research directions include:

Extending the approach to multilingual economic texts. Developing dynamic keyword dictionaries that adapt to emerging terminology. Exploring the relationship between formality and economic sentiment. Creating user-friendly tools for policymakers and financial analysts. Investigating temporal patterns in formality across economic cycles.

Our findings underscore the importance of domain-specific adaptations in linguistic analysis, particularly for specialized fields like economics, where communication style can have far-reaching implications for markets, policy, and public understanding.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the authors used Grammarly and ChatGPT to improve the grammar and overall quality of the text. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References:

[1] Heylighen, F., Dewaele, J. M. (1999). Formality of language: Definition, measurement, and behavioral determinants. *Internal Report, Center "Leo Apostel",* Free University of Brussels.

[2] Brown, J. D., Rodgers, T. S. (2002). *Doing*

second language research. Oxford University Press.

[3] Xie, J., Zhang, B., Brown, S., & Zeng, D. (2021). Write like a pro or an amateur? Effect of medical language formality. *ACM Transactions on Management Information Systems*, *12*(3), 1-25. https://doi.org/10.1145/3458752.

[4] Alam, M. S., Mrida, M. S. H., & Rahman, M. A. (2025). Sentiment analysis in social media: How data science impacts public opinion knowledge integrates natural language processing (NLP) with artificial intelligence (AI). *American Journal of Scholarly Research and Innovation,* 4(1), 63–100. https://doi.org/10.63125/r3sq6p80.

[5] Bulíř, A., Čihák, M., & Šmídková, K. (2013). Writing Clearly: The ECB's Monetary Policy Communication. *German Economic Review*, 14(1), 50–72. https://doi.org/10.1111/j.1468-0475.2011.00562.x.

[6] Ardia, D., Bluteau, K., & Boudt, K. (2021). Media abnormal tone, earnings announcements, and the stock market. *arXiv preprint arXiv:2110.10800*. https://doi.org/10.48550/arXiv.2110.10800.

[7] Meursault, V., & Kogan, S. (2022). Disentangling the content of earnings calls: How corporate messaging affects firms' financial performance. *Federal Reserve Bank of Philadelphia.*

[8] Abdulrahman, B. M. A., Ibrahim, A. G. A., Dawalbait, H. A. A. (2024). Using Nonlinear Autoregressive Models to Investigate the Impact of Foreign Trade on Economic Growth in Sudan. *WSEAS Transactions on Business and Economics,* Vol.21, pp.719-725. https://doi.org/10.37394/23207.2024.21.60.

[9] Mavrommati, A., Kazanas, T. H. A. N. A., S. S. I. S., Pliakoura, A. L. E. X. A. N. D., R. A., Kalogiannidis, S. T. A. V. R. O. S., Chatzitheodoridis, F. O. T. I. O. S. (2024). An empirical study on tourism and economic growth in Greece: an autoregressive distributed lag boundary test approach. *WSEAS Transactions on Business and Economics,* vol.21, pp.588-602. https://doi.org/10.37394/23207.2024.21.49.

[10] Ojo, O.E., Gelbukh, A., Calvo, H., Adebanji, O.O., & Sidorov, G. (2020). *Sentiment Detection in Economics Texts*. In L. Martínez-Villaseñor, O. Herrera-Alcántara, H. Ponce, & F.A. Castro-Espinoza (Eds.), *Advances in Computational Intelligence* (pp. 303–314). Springer, Cham. https://doi.org/10.1007/978-3-030-60887-3_24.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

- Harsh Mehta: Writing an original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.
- Santosh Kumar Bharti: Writing–review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.
- Nishant Doshi: Writing–review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

APPENDIX

Table 1. Dataset Statistics

Central Bank Statements	150	2500	2018-2023
Financial Reports	300	15000	2019-2023
Social Media Discussions	5000	200	2020-2023
Dataset Type	**No.of Documents**	**Average Word Count**	**Date Range**

Table 2. Feature Importance

Feature Removed	Average Decrease in Accuracy
Numerical values	42%
Domain-specific keywords	35%
Noun frequency	18%
Pronoun frequency	12%
Verb frequency	7%