

# VerbaNex AI at DIPROMATS 2024: Enhancing Propaganda Detection in Diplomatic Tweets with Fine-Tuned BERT and Integrated NLP Techniques

Jose Cuadrado<sup>2</sup>, Elizabeth Martinez<sup>1</sup>, Juan Cuadrado<sup>1</sup>, Juan Carlos Martinez-Santos<sup>1</sup> and Edwin Puertas<sup>1,\*,‡</sup>

<sup>1</sup>Universidad Tecnologica de Bolivar, School of Engineering, Cartagena de Indias 130010, Colombia.

<sup>2</sup>Universidad Industrial de Santander, School of Engineering, Bucaramanga 680002, Colombia.

## Abstract

This paper outlines the methodology used by VerbaNexAI for the DIPROMATS 2024 competition, part of the Iberian Languages Evaluation Forum (IberLEF). The challenge involves detecting propaganda and strategic narratives in tweets from diplomats of the USA, Europe, Russia, and China, in English and Spanish. Task 1, focuses on the identification and characterization of propaganda, we implemented four methodologies. Our pre-processing steps include text normalization, removal of URLs, retweets, user mentions, stop words, and lemmatization. Feature extraction was performed using TF-IDF vectorization, transformer fine-tuning, combined feature extraction from TF-IDF and transformers, and a hashtag-specific feature extraction. Regularization techniques such as class balancing and k-fold cross-validation were applied to ensure robust model performance. Various classifiers, including Random Forest, Support Vector Classifier, Naive Bayes, and Logistic Regression, were evaluated to determine the most effective models. Our approach aims to enhance the detection of propaganda in diplomatic tweets, contributing to a broader understanding of how propaganda operates in social media.

## Keywords

DIPROMATS 2024, Propaganda Detection, NLP, TF-IDF, Transformers, Diplomatic Tweets

## 1. Introduction

Propaganda has been a powerful tool for shaping public opinion and influencing political beliefs throughout history. In the digital age, social media platforms have amplified the reach and impact of propaganda, making it a critical area of study [1]. The DIPROMATS 2024 competition [2], part of the Iberian Languages Evaluation Forum (IberLEF)[3], aims to address this issue by challenging participants to develop systems capable of detecting and characterizing propaganda and strategic narratives in tweets written by diplomats from the USA, Europe, Russia, and China, in both English and Spanish.

---

*IberLEF 2024, September 2024, Valladolid, Spain*


\*Corresponding author.

✉ Jose2200485@correo.uis.edu.co (J. Cuadrado); eayala@utb.edu.co (E. Martinez); jflechas@utb.edu.co (J. Cuadrado); jcmartinezs@utb.edu.co (J. C. Martinez-Santos); epuerta@utb.edu.co (E. Puertas)

🆔 0009-0009-8436-8083 (J. Cuadrado); 0000-0001-6592-347X (E. Martinez); 0000-0002-8226-1372 (J. Cuadrado); 0000-0003-2755-0718 (J. C. Martinez-Santos); 0000-0002-0758-1851 (E. Puertas)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

The importance of this problem lies in the subtle and pervasive nature of propaganda in social media. Unlike disinformation, which often involves outright falsehoods, propaganda can be more insidious, blending plausible suggestions, half-truths, and manipulative assertions to influence emotions and prejudices [4]. This makes it difficult to detect and counteract, particularly when used by influential figures such as diplomats and government officials. Understanding and identifying these techniques are crucial for safeguarding democratic processes and promoting informed public discourse.

Previous research in natural language processing (NLP) has made significant strides in detecting various forms of disinformation and malicious content. Studies have explored machine learning models, including transformer-based architectures, for text classification tasks [5, 6]. However, propaganda detection, specifically in diplomatic communication, presents unique challenges due to the nuanced language. Our study builds on this existing body of work by focusing on a more targeted application of these techniques to identify and analyze propaganda.

The theoretical framework for this study is grounded in the principles of communication theory and persuasion. Propaganda techniques often bypass rational thought processes, appealing instead to emotions and deeply held beliefs [7]. By systematically analyzing the language and narratives used in diplomatic tweets, we can gain insights into the strategies employed to shape public perception and influence geopolitical dynamics. This study has significant theoretical implications for understanding propaganda and practical implications for developing tools to detect and mitigate its effects.

The objective is to develop and evaluate methodologies for detecting propaganda in diplomatic tweets. We hypothesize traditional NLP techniques, such as TF-IDF vectorization, and advanced transformer models, will effectively identify and characterize propagandistic content. Specifically, we aim to develop a robust pre-processing pipeline to clean and standardize tweet data, implement and compare multiple feature extraction techniques, and evaluate the performance of various classifiers in detecting propaganda.

Experimental results during the training phase demonstrated promising performance across different methodologies. For instance, using the Linear SVC model, we achieved an accuracy of 91% with a precision of 92%, recall of 91%, and an F1-score of 91% for our first methodology. The fine-tuned BERT transformer model yielded an accuracy of 84%, with consistent precision, recall, and F1-scores. The combined feature extraction methodology with Linear SVC resulted in a 90% accuracy and an F1-score of 90%. Finally, the hashtag-specific feature extraction approach, also using Linear SVC, achieved an accuracy of 88%, and an F1-score of 88% [8, 9].

In the competition evaluation, our approach achieved varying levels of success. One of our methods ranked 24th in Spanish Task 1 based on the Information Contrast Model (ICM) score. It performed significantly better in F1-macro-F, achieving a score of 0.7279, indicating its effectiveness at distinguishing between propaganda and non-propaganda tweets. These results highlight the effectiveness of our approach and provide a solid foundation for further refinement and application to real-world data.

## 2. Related Work

Detection of propaganda in textual content has become an increasingly important area of research, especially in social media and digital communication. Propaganda, which can be defined as biased or misleading information used to promote a particular political cause or point of view, poses significant challenges to detection due to its often subtle and nuanced nature.

Early research in this field primarily focused on traditional machine-learning techniques. For example, Barrón-Cedeño et al. [10] developed Propopy, a system for organizing news based on their propagandistic content. Their approach utilized a combination of linguistic features and machine learning algorithms, highlighting the importance of feature selection in improving model performance.

With the advent of deep learning, more sophisticated models have been employed to tackle the problem of propaganda detection. Da San Martino et al. [11] proposed a fine-tuned BERT model for detecting propaganda in news articles. This study demonstrated that transformer-based models outperform traditional approaches by capturing the nuanced language patterns associated with propaganda.

Glazkova et al. [12] extended this work by applying various deep learning models, including BERT and GPT-2, to identify propaganda techniques. Their comprehensive study underscored the effectiveness of contextual embeddings in improving classification accuracy, indicating the potential of these models in handling complex language tasks.

In the realm of social media, Zubiaga et al. [13] explored the use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for identifying propaganda in tweets. Their findings indicated that combining textual features with user metadata can enhance detection performance, providing a robust approach to handling the diverse and noisy nature of social media data.

Another significant contribution to the field is the work by Rashkin et al. [14], who developed a system to detect different types of misleading content, including propaganda, in news articles. They utilized a combination of linguistic and context-based features, demonstrating the utility of combining multiple types of information to improve detection accuracy.

Moreover, Guerini et al. [15] introduced a dataset specifically designed for analysis of propaganda techniques in news articles. This dataset has been widely adopted in subsequent research and has facilitated the development of more targeted detection models. Their work also provided a detailed taxonomy of propaganda techniques, which has been instrumental in advancing the understanding and identification of these techniques in text.

Despite these advances, the detection of propaganda in diplomatic communication remains relatively unexplored. Diplomatic tweets often employ subtle and sophisticated language, making it challenging to identify propaganda without advanced NLP techniques. Our study builds on this body of work by focusing on the specific context of diplomatic communication and employing a combination of traditional and transformer-based methods to detect propaganda in tweets.

By leveraging the strengths of both traditional and advanced NLP techniques, we aim to develop a comprehensive approach to propaganda detection that can handle the unique challenges posed by diplomatic communication.

### 3. Data

For the DIPROMATS 2024 competition, we utilized the dataset provided by the organizers, focusing specifically on the identification and characterization of propaganda techniques in tweets for Task 1. This dataset includes tweets in both English and Spanish, written by diplomats and authorities from the USA, Europe, Russia, and China. The dataset is composed of tweets collected from official diplomatic accounts, including government officials, embassies, ambassadors, consuls, and various diplomatic missions. This diverse collection ensures a comprehensive representation of diplomatic communication across different geopolitical contexts. The dataset comprises raw textual content of the tweets and binary labels indicating the presence (1) or absence (0) of propaganda techniques. The dataset’s bilingual nature (English and Spanish) and its coverage of tweets from diplomats representing four major geopolitical entities add significant complexity and value. This diversity necessitates the use of advanced natural language processing techniques capable of handling different languages and communication styles. The dataset provides a unique opportunity to study the variations in propaganda techniques across different cultures and geopolitical contexts.

### 4. Methodology

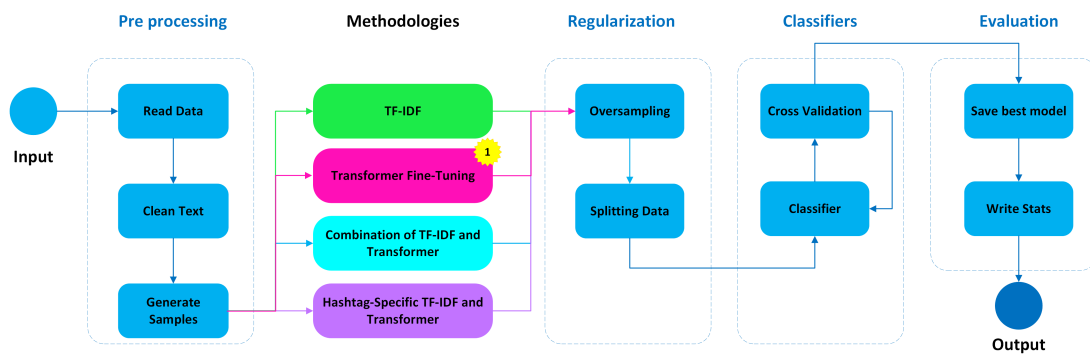


Figure 1: System Pipeline.

In this section, we detail the methodology used to detect propaganda in diplomatic tweets as seen in Figure 1. Our approach involves several key steps, including pre-processing the tweet data, extracting relevant features, employing various machine learning models, and evaluating their performance. We developed and tested four distinct methodologies, each leveraging different combinations of traditional and advanced NLP techniques to identify propaganda in a complex and nuanced textual environment.

#### 4.1. Pre-Processing

The first step in our pre-processing pipeline is to convert all text to lowercase. This standardizes the text and ensures consistency in our analysis. We then remove URLs, retweets, and user

mentions from the tweets as they often contain noise and do not contribute to the identification of propaganda techniques. Common stop words, such as "the", "and", "is", etc., are removed from the text to focus on the meaningful content of the tweets. Additionally, lemmatization is applied to standardize words to their base or root form, reducing inflectional forms to a common base form.

## **4.2. Feature Extraction**

Feature extraction enables the transformation of raw text into meaningful representations that machine learning models can process. We utilized both traditional and advanced NLP techniques across four different methodologies to capture the relevant features from the pre-processed text.

### **4.2.1. Methodology 1: TF-IDF**

In the first methodology, we employed Term Frequency-Inverse Document Frequency (TF-IDF) to extract features from the pre-processed text data [16]. This technique transforms raw text into a numerical representation that can be used as input for machine learning models. TF-IDF captures the relevance of each word in the context of the document and the overall corpus, providing a solid foundation for subsequent classification tasks.

### **4.2.2. Methodology 2: Transformer Fine-Tuning**

For the second methodology, we performed fine-tuning of the BERT (bert-base-cased) transformer model specifically for the task of text classification [17]. This approach leverages the contextual embeddings generated by BERT, which have been shown to significantly improve the performance of classification tasks by capturing the relationships between words in a sentence.

### **4.2.3. Methodology 3: Combination of TF-IDF and Transformer**

The third methodology involves a combination of feature extraction techniques. We implemented both TF-IDF and a transformer model (cardiffnlp/xlm-roberta-base-sentiment-multilingual) and combined the features extracted by these two techniques [18]. This approach captures both local information from TF-IDF and contextual information from the transformer model, providing a comprehensive representation of the text.

### **4.2.4. Methodology 4: Hashtag-Specific TF-IDF and Transformer**

In the fourth methodology, we implemented a variation of the third approach. Here, the TF-IDF was applied exclusively to the hashtags within the tweets, creating a vocabulary based solely on these hashtags. These features were then combined with the features extracted from the transformer model. This combined approach aimed to leverage the specific semantic information conveyed by hashtags in addition to the contextual information from the transformer.

### 4.3. Regularization

To address the class imbalance, we implemented various regularization techniques. We utilized the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic examples by interpolating between existing samples from the minority class [19]. Class-balancing methods were also applied to ensure an equitable distribution of instances across different categories. Combining data augmentation with regularization techniques, we effectively addressed class imbalance issues, significantly enhancing the model's ability to predict propaganda accurately.

### 4.4. Cross-Validation

We employed k-fold cross-validation to assess the classifiers used. This method divides the dataset into k equally sized subsets, or folds. The model is trained and evaluated k times, with each fold serving as the test set once, while the remaining k-1 folds are used for training [20]. This evaluation method provides a reliable estimate of the model's performance on unseen data and helps identify the best-performing models.

### 4.5. Classification

We experimented with various classifiers, including Random Forest, Support Vector Classifier (SVC), Naive Bayes Classifier, and Logistic Regression, to identify the most suitable model for detecting propaganda techniques. To evaluate the performance of our models in detecting propaganda techniques in diplomatic tweets, we conducted comprehensive experiments using various classifiers. We employed standard metrics such as accuracy, precision, recall, and F1-score to assess the effectiveness of each classifier in classifying tweets as propaganda or non-propaganda.

## 5. Evaluation

### 5.1. Model Testing

In the testing section, the experiments conducted to assess the performance of models are described. Standard metrics such as precision, recall, F1-score, and accuracy are used to evaluate the effectiveness of each classifier in classifying tweets as propaganda or non-propaganda.

**Table 1**  
Performance of Different Methodologies

Approach	Classifier	Accuracy	Precision	Recall	F1-score
TFiDF	Support Vector Classifier	0.91	0.92	0.91	0.91
Bert Fine-Tuning	Linear Classifier	0.84	0.85	0.84	0.84
Combinations of Features	Support Vector Classifier	0.9	0.91	0.9	0.9
Combinations of Features Hashtags	Support Vector Classifier	0.88	0.89	0.88	0.88

The TFiDF achieved the highest performance among the approaches, with an accuracy of 0.91 and an F1 score of 0.91. This indicates that TFiDF was the most effective model for distinguishing between propaganda and non-propaganda tweets during training.

## 5.2. Competition Evaluation

The results revealed that VerbaNex AI ranked 24th, 28th, 29th, and 30th out of the participating teams for Spanish. The detailed performance metrics provided by the competition organizers are summarized in Table 2. These metrics include the ICM score, F1-True, F1-False, and F1-macro-F scores.

**Table 2**  
Competition Results for VerbaNex AI

Approach	Position	ICM	F1-True	F1-False	F1-macro-F
Bert Fine-Tuning	24	-0.0012	0.5301	0.9258	0.7279
TFiDF	28	-0.3662	0.3230	0.8636	0.5933
Combinations of Features	29	-0.3686	0.3333	0.8607	0.5970
Combinations of Features Hashtags	30	-0.3715	0.3058	0.8657	0.5858

The analysis of the results suggests that while our approaches showed promising performance during the training phase, they did not generalize well to the test data used in the competition. Our approach ranked 24th based on the ICM score [21], performed significantly better in F1-macro-F and other F1 metrics. This indicates our approach did not align well with the competition’s main evaluation metric. Future improvements should focus on better understanding and optimizing the ICM metric to achieve a higher overall ranking.

## 6. Conclusion

This study presents the methodology and results of VerbaNex AI’s participation in the DIPROMATS 2024 competition. Our approaches involved traditional NLP techniques, such as TF-IDF vectorization, and advanced transformer models, to detect propaganda in diplomatic tweets. Despite achieving promising results during the training phase, our performance in the competition indicated the need for further refinement.

Key areas for improvement include addressing data distribution issues, refining feature engineering techniques, and selecting more appropriate models. By focusing on these aspects, we aim to enhance the robustness and accuracy of our propaganda detection models.

Our research contributes to the broader effort of understanding and combating propaganda in social media, particularly in the context of diplomatic communication. We remain committed to advancing our methodologies and improving our models identify and mitigate the effects of propaganda.

## 7. Future Work

Future work will focus on addressing the limitations identified in this study. A key area of improvement is enhancing the feature extraction process to better capture the nuances of propaganda in diplomatic tweets. This may involve incorporating more sophisticated linguistic and contextual features.

Additionally, we plan to explore more advanced models and techniques, such as ensemble methods and deep learning architectures, to improve classification performance. Data augmentation and transfer learning could handle data distribution issues and enhance model generalization.

Collaboration with experts in political science and communication studies will be crucial to validate our model's predictions in real-world settings. This interdisciplinary approach will ensure that our methodologies align with the theoretical and practical needs of identifying and combating propaganda in digital communication.

## Acknowledgments

The authors would like to acknowledge the support provided by the master's degree scholarship program in engineering at the Universidad Tecnológica de Bolívar (UTB) in Cartagena, Colombia.

## References

- [1] C. Wardle, H. Derakhshan, Information disorder: Toward an interdisciplinary framework for research and policy making, Council of Europe report, DGI(2017)09 (2017).
- [2] P. Moral, J. Fraile, G. Marco, A. Peñas, J. Gonzalo, Overview of DIPROMATS 2024: Detection, characterization and tracking of propaganda in messages from diplomats and authorities of world powers, *Procesamiento del Lenguaje Natural* 73 (2024).
- [3] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024)*, co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.
- [4] R. Marlin, *Propaganda and the ethics of persuasion*, Broadview Press, 2013.
- [5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, 2019, pp. 4171–4186. URL: <https://aclanthology.org/N19-1423>. doi:10.18653/v1/N19-1423.
- [6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).
- [7] G. S. Jowett, V. O'Donnell, *Propaganda & persuasion*, Sage publications, 2018.



- [8] E. Puertas, L. G. Moreno-Sandoval, J. Redondo, J. A. Alvarado-Valencia, A. Pomares-Quimbaya, Detection of sociolinguistic features in digital social networks for the detection of communities, *Cognitive Computation* 13 (2021) 518–537.
- [9] E. Puertas, J. C. Martinez-Santos, Phonetic detection for hate speech spreaders on twitter, *CLEF* (2021).
- [10] A. Barrón-Cedeño, P. Nakov, G. Da San Martino, T. Elsayed, Proppy: Organizing the news based on their propagandistic content, *Information Processing & Management* 57 (2020) 102–150.
- [11] G. Da San Martino, A. Barrón-Cedeño, I. Jaradat, H. Mubarak, W. Zaghouni, P. Nakov, Fine-grained analysis of propaganda in news articles, *arXiv preprint arXiv:1910.02517* (2019).
- [12] A. Glazkova, M. Glazkov, M. Tikhonova, Fine-tuned bert and gpt-2 for identification of propaganda techniques in news, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 153–162.
- [13] A. Zubiaga, M. Liakata, R. Procter, G. W. S. Hoi, P. Tolmie, Detection and resolution of rumours in social media: A survey, *ACM Computing Surveys (CSUR)* 51 (2018) 1–36.
- [14] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (2017) 2931–2937.
- [15] M. Guerini, C. Strapparava, G. Moretti, C. Pedrinaci, S. Tonelli, Towards zero-shot frame semantic parsing for propaganda detection, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (2018) 184–189.
- [16] J. Ramos, Using tf-idf to determine word relevance in document queries, *Proceedings of the first instructional conference on machine learning* 242 (2003) 133–142.
- [17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).
- [18] E. Martinez, J. Cuadrado, J. C. Martinez-Santos, E. Puertas, Detection of online sexism using lexical features and transformer, in: *2023 IEEE Colombian Caribbean Conference (C3)*, IEEE, 2023, pp. 1–5.
- [19] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, *Journal of artificial intelligence research* 16 (2002) 321–357.
- [20] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, in: *Statistics surveys*, volume 4, 2010, pp. 40–79.
- [21] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 5809–5819.