Cross-Linguistic Framing Analysis: Unveiling Political and Cultural Narratives in Spanish-Language News

Juan Cuadrado and **Elizabeth Martinez** and **Edwin Puertas** and **Juan Carlos Martínez-Santos**
Universidad Tecnológica de Bolívar
epuerta@utb.edu.co

Abstract

As the media landscape evolves, the study of news framing, key to shaping public opinion, has expanded. Our study presents a methodology for framing analysis in Spanish media using NLP to uncover the cultural and political factors shaping narratives. This method enhances linguistic inclusivity in framing studies and highlights Spanish media's role in public opinion, showing NLP's utility in diverse linguistic contexts and contributing to a more comprehensive global media analysis.

1 Introduction

In the contemporary media landscape, the influence of how news is framed—selectively presenting certain aspects of a story to shape public opinion is more pronounced than ever. This process, known as framing, is critical because it not only highlights specific angles of news stories but also significantly influences public attitudes and behaviors by shaping the discourse around key issues (Hassan et al., 2020; Yang et al., 2017; Puertas et al., 2018).

Despite extensive research on media framing in English, Spanish-language media dynamics are less studied, despite their global impact. This gap in research highlights a critical lack of understanding about the diversity of media manipulation and framing techniques across linguistic and cultural contexts (Camacho et al., 2022).

Our study introduces a methodology tailored for the analysis of framing in Spanish-language news using advanced Natural Language Processing (NLP) techniques. This approach is influenced by recent advancements in the field, particularly the insights gained from SemEval 2023 (Piskorski et al., 2023). By focusing on Spanish, we aim to enhance linguistic inclusivity in framing studies and better understand the unique cultural and political factors that influence media narratives (Kim et al., 2019; Solves et al., 2019).

Our methodology provides insights into the strategies used to frame news stories in Spanish and evaluates the effectiveness of these strategies in shaping public opinion. The results of our study reveal significant variability in how different framing categories are applied, underscoring the complexity of adapting analytical tools to different linguistic environments (Alhindi et al., 2020).

This research paves the way for future cross-linguistic studies. It enriches our comprehension of how media narratives are crafted and perceived across different cultures, reinforcing the importance of considering linguistic diversity in media studies(Cuadrado et al., 2023a; Puertas and Martinez-Santos, 2021).

2 Methodology

Our methodological framework begins with preprocessing using NLTK for text normalization and continues with web scraping for feature extraction, focusing on category-specific Spanish lexicons. We apply SMOTE for class balance and StratifiedShuffleSplit for robust data partitioning. Classifier selection is refined through a voting mechanism, leading to the evaluation phase where cross-validation ensures the accuracy and generalizability of our model. This concise yet comprehensive approach, outlined in Figure 1, provides a systematic pathway to analyze framing within Spanish media narratives (Puertas et al., 2021).

2.1 Preprocessing

Preprocessing with NLTK is key to refining the dataset for precise modeling. Text is made uniform by lowercasing and clearing non-essential characters, ensuring only pertinent information is retained for analysis. Tokenizing the text into discrete words and removing stop words pares down the data to its most informative elements. Lemmatization further streamlines the variability of words, setting a solid

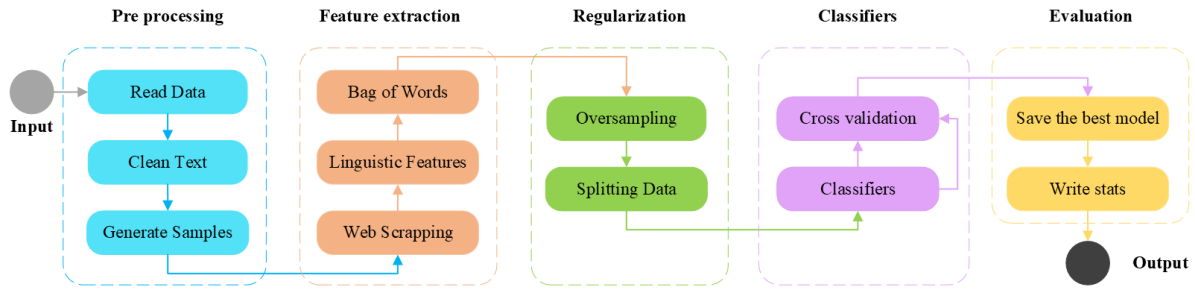

Figure 1: Methodology Pipeline

foundation for accurate feature analysis (Moreno-Sandoval et al., 2019; Puertas et al., 2019).

2.2 Feature Extraction

We customize feature extraction for the Spanish context. Initially, we collected data via web scraping, drawing from resources aligned with our categories, such as Wikipedia, using the Beautifulsoup library to compile relevant texts (Patel and Patel, 2020).

Adapting the method for Spanish (Cuadrado et al., 2023a), we cleaned the text by removing extraneous characters and standardized words(Cuadrado et al., 2023b). We then broke down the text into tokens, a step divergent from the NLTK library's (Bird, 2006) approach for English, to better suit the Spanish lexicon. Leveraging the 'Weirdness' concept, we compared the token frequencies with Google's Spanish unigrams to curate a lexicon unique to each framing category (Geyken and Lemnitzer, 2012; Brooke et al., 2009).

The final lexicon, representing the most salient terms within each contextual category, was incorporated into a Bag-of-Words model. These representations formed the foundation for our model's input features, essential for the subsequent voting-based classification system (Martinez et al., 2023a,b), thus ensuring a precise analysis of framing within Spanish media narratives.

2.3 Regularization

SMOTE equilibrates our dataset, crucial for unbiased model predictions. The StratifiedShuffleSplit method ensures a representative division of data, aiding in model validation and promoting its applicability to broader data sets (Chawla et al., 2002; Mahesh et al., 2023).

2.4 Classifiers

The voting mechanism in our classifiers enhances accuracy by amalgamating the strengths of individual models. By evaluating classifiers through the Python lazy classifier tool, we identified the top three for each category by their performance. In action, these classifiers vote on classifications, with the majority decision shaping the final outcome. This approach not only bolsters the model's robustness and precision but also has been confirmed through performance tests. Utilizing this ensemble method, our classification process is precisely calibrated for distinct categories, ensuring broad applicability and reliability.

2.5 Evaluation

In the evaluation stage, we conduct a thorough appraisal to select the best-performing model. This stage is essential for confirming the model's efficacy in multi-label classification, ensuring it can deftly navigate the complexities of various framing categories.

3 Experimental Setup

The "Framing Across Borders" corpus (Cuadrado et al., 2024) serves as the foundation of our experimental setup, incorporating 140 Spanish-language articles from "El Universal" and "El Tiempo." Articles were selected to cover a broad spectrum of socio-political topics and tones, with neutral (45.85%), negative (36.68%), and positive (17.47%) perspectives represented. The dataset predominantly features frames such as 'Public opinion' (57.89%), 'Political' (42.98%), and 'Security and defense' (27.63%), highlighting the salience of governance and societal issues in news narratives.

For our model evaluation, we partitioned these articles into training (80%), development (10%), and testing (10%) sets. This distribution was strategically chosen to foster effective learning and validation, facilitating a thorough assessment of the model's predictive performance.

Figure 2: Confusion Matrix Summary for Traning and Test Results.

	Precision	Recall	F1 Score
micro avg	0.31	0.33	0.32
macro avg	0.21	0.31	0.20
weighted avg	0.41	0.33	0.33
samples avg	0.28	0.34	0.29

Table 1: Metrics Summary.

4 Results

The overall performance metrics of our model, presented in the Table 1, show precision, recall, and F1 scores that might initially suggest modest effectiveness. Specifically, the micro average scores are 0.31 for precision, 0.33 for recall, and 0.32 for F1; macro averages are even lower at 0.21 for precision, 0.31 for recall, and 0.20 for F1. While these metrics appear low, they do not necessarily indicate poor model performance. Instead, they reflect challenges in accurately detecting certain framing categories, which are not uniformly easy to identify.

These challenges are further elucidated by the confusion matrices shown in Figure 2, which detail the model's performance across various framing categories. For instance, while there are categories where the model over-predicts false positives, such as Legality, Constitutionality, and Jurisprudence and External Regulation and Reputation, it performs well in accurately identifying positives in other categories like Cultural Identity, Public Opinion, and Economic.

The detailed examination of individual category performance within the confusion matrices helps to affirm that the model is effectively identifying frames in several key areas, despite the lower average scores. This nuanced understanding is crucial for directing future refinements and ensuring a balanced evaluation of the model's capabilities.

5 Conclusions

This study has demonstrated a gradual yet consistent enhancement of the model's ability to classify framing in Spanish-language news. While the model performed well in distinguishing several frames, it became evident that achieving uniformly high levels of accuracy across all categories remains a challenge.

6 Limitations and Future Work

The limitations of our study primarily revolve around the difficulties in achieving consistent accuracy across all framing categories, highlighting the need for further refinement of the model. Future work will focus on ensuring that every category of framing is recognized with similar precision. This will likely involve expanding the training data and refining the algorithm to better understand and categorize the nuanced language of news frames. The goal is to develop a balanced and reliable tool for framing analysis, ultimately strengthening our understanding of how media narratives shape and reflect public discourse.

Acknowledgments

To the master's degree scholarship program in engineering at the Universidad Tecnológica de Bolivar (UTB) in Cartagena Colombia and to the contributors who made the "Framing Across Borders: News Media Corpus Framing in Mexico and Colombia".

References

Tariq Alhindi, Smaranda Muresan, and Daniel Preotiuc-Pietro. 2020. Fact vs. opinion: the role of argumentation features in news classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6139–6149, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Julian Brooke, Milan Tofiloski, and Maite Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. In *Proceedings of the international conference RANLP-2009*, pages 50–54.

Luís Camacho, Georgios Douzas, and Fernando Bacao. 2022. Geometric smote for regression. *Expert Systems with Applications*, 193:116387.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Juan Cuadrado, Elizabeth Martinez, Juan Carlos Martinez-Santos, and Edwin Puertas. 2024. Framing Across Borders: News Media Corpus Framing in Mexico and Colombia. *Zenodo*.

Juan Cuadrado, Elizabeth Martinez, Anderson Morillo, Daniel Peña, Kevin Sossa, Juan Martinez-Santos, and Edwin Puertas. 2023a. UTB-NLP at SemEval-2023 task 3: Weirdness, lexical features for detecting categorical framings, and persuasion in online news. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1551–1557, Toronto, Canada. Association for Computational Linguistics.

Juan Cuadrado, Elizabeth Martinez, Anderson Morillo, Daniel Peña, Kevin Sossa, Juan Martinez-Santos, and Edwin Puertas. 2023b. Utb-nlp at semeval-2023 task 3: Weirdness, lexical features for detecting categorical framings, and persuasion in online news. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1551–1557.

Alexander Geyken and Lothar Lemnitzer. 2012. Using google books unigrams to improve the update of large monolingual reference dictionaries. In *Proceedings of the 15th EURALEX International Congress*, pages 362–366.

Isyaku Hassan, Mohd Nazri Latiff Azmi, and Akibu Mahmoud Abdullahi. 2020. Evaluating the spread of fake news and its detection. techniques on social networking sites. *Romanian Journal of Communication and Public Relations*, 22(1):111–125.

Donghwa Kim, Deokseong Seo, Suhyoun Cho, and Pilsung Kang. 2019. Multi-co-training for document classification using various document representations: Tf–idf, lda, and doc2vec. *Information Sciences*, 477:15–29.

TR Mahesh, Oana Geman, Martin Margala, Manisha Guduri, et al. 2023. The stratified k-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification. *Healthcare Analytics*, 4:100247.

Elizabeth Martinez, Juan Cuadrado, Juan Carlos Martinez-Santos, Daniel Peña, and Edwin Puertas. 2023a. Automated depression detection in text data: leveraging lexical features, phonesthemes embedding, and roberta transformer model. *In IberLEF (Working Notes). CEUR Workshop Proceedings*.

Elizabeth Martinez, Juan Cuadrado, Juan Carlos Martinez-Santos, and Edwin Puertas. 2023b. Detection of online sexism using lexical features and transformer. In *2023 IEEE Colombian Caribbean Conference (C3)*, pages 1–5. IEEE.

Luis Gabriel Moreno-Sandoval, Edwin Puertas, Flor Miriam Plaza-del Arco, Alexandra Pomares-Quimbaya, Jorge Andres Alvarado-Valencia, and L Alfonso. 2019. Celebrity profiling on twitter using sociolinguistic. *CLEF (Working Notes)*.

Jay M Patel and Jay M Patel. 2020. Web scraping in python using beautiful soup library. *Getting Structured Data from the Internet: Running Web Crawlers/Scrapers on a Big Data Production Scale*, pages 31–84.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Edwin Puertas, Jorge Andres Alvarado-Valencia, Luis Gabriel Moreno-Sandoval, and Alexandra Pomares-Quimbaya. 2018. An automatic approach to generate corpus in spanish. In *Advances in Computing: 13th Colombian Conference, CCC 2018, Cartagena, Colombia, September 26–28, 2018, Proceedings 13*, pages 150–161. Springer.

Edwin Puertas and Juan Carlos Martinez-Santos. 2021. Phonetic detection for hate speech spreaders on twitter notebook for pan at clef 2021. *CEUR Workshop Proceedings*.

Edwin Puertas, Luis Gabriel Moreno-Sandoval, Flor Miriam Plaza-del Arco, Jorge Andres Alvarado-Valencia, Alexandra Pomares-Quimbaya, and L Alfonso. 2019. Bots and gender profiling on twitter using sociolinguistic features. *CLEF (Working Notes)*, pages 1–8.

Edwin Puertas, Luis Gabriel Moreno-Sandoval, Javier Redondo, Jorge Andres Alvarado-Valencia, and Alexandra Pomares-Quimbaya. 2021. Detection of sociolinguistic features in digital social networks for the detection of communities. *Cognitive Computation*, 13:518–537.

Josep Solves, Athanasios Pappous, Inmaculada Rius, and Geoffery Zain Kohe. 2019. Framing the paralympic games: A mixed-methods analysis of spanish media coverage of the beijing 2008 and london 2012 paralympic games. *Communication & Sport*, 7(6):729–751.

Fan Yang, Arjun Mukherjee, and Eduard Dragut. 2017. Satirical news detection and analysis using attention mechanism and linguistic features. *arXiv preprint arXiv:1709.01189*.