Political Ideology Classification in Tweets

A Natural Language Processing Deep Learning Approach

José Walter Hernández Pérez

AI Engineer

Abstract

This report details the end-to-end development of a text classification pipeline designed to identify political ideologies in Twitter data. The system analyzes semantic patterns to categorize user stances. The project covers data curation, preprocessing techniques, model architecture, and performance evaluation metrics.

Contents

Chapter 1

Introduction

Political ideology classification in social media has become a cornerstone of computational social science and public opinion monitoring. In an era where digital platforms serve as the primary arena for political discourse, the ability to automatically categorize user stances—specifically within the nuanced spectrum of the Spanish political landscape—presents significant challenges due to the informal, ambiguous, and highly polarized nature of the language used.

This project addresses the problem of multi-class ideology classification (Left, Moderate Left, Right, Moderate Right) using a dataset of Spanish tweets. Unlike traditional approaches that rely heavily on metadata, this work rigorously evaluates the efficacy of pure Natural Language Processing (NLP) techniques, ensuring robust performance even in privacy-preserving scenarios where user demographics are unavailable.

1.1 Project Objectives

The primary goals of this technical report are:

1. **Benchmarking Architectures:** To evaluate and compare three distinct NLP paradigms: Domain-Adapted Transformers (PolitiBETO), Classical Machine Learning with Feature Engineering (Logistic N-grams), and Hybrid Transformer Ensembles (DualBERT).
2. **Bias Mitigation:** To identify and rectify critical data leakage issues inherent in metadata-rich datasets, ensuring the model learns semantic patterns rather than memorizing user identities.
3. **Efficiency vs. Performance Analysis:** To determine the optimal trade-off between computational cost and classification accuracy for deployment in resource-constrained environments.

1.2 Report Structure

This document is organized as follows: Chapter 2 reviews the state-of-the-art in political profiling. Chapter 3 presents a critical audit of the dataset, revealing significant biases. Chapter 4 details the engineered pipelines. Chapter 5 discusses the experimental results, and Chapter 7 summarizes the key findings and engineering recommendations.

Chapter 2

State of the Art

2.1 Political Profiling in Social Media

The automatic profiling of political actors has evolved from simple keyword matching to complex contextual embedding models. A pivotal benchmark in this domain is the *PoliticEs 2022* shared task García-Díaz et al. (2022), which established a standardized corpus for profiling Spanish politicians based on gender, profession, and ideology.

The landscape of solutions can be categorized into three dominant approaches:

2.1.1 Transformer-Based Models

The current state-of-the-art is dominated by Large Language Models (LLMs). The *LosCalls* team (Santamaría Carrasco and Cuervo, 2022) achieved the top performance (Macro-F1=0.902) by ensembling BETO and RoBERTa-MarIA. Their success highlights the superiority of pre-trained models in capturing the subtle semantic shifts characteristic of political irony and dog-whistles.

2.1.2 Domain Adaptation Strategies

General-purpose models often underperform in specialized domains. *NLP-CIMAT* introduced **PolitiBETO** (NLP-CIMAT-GTO, 2022), a BERT variant further pre-trained on a massive corpus of political tweets. Achieving an F1-macro of 0.896, this approach demonstrates that domain-adaptive pre-training (DAPT) is a crucial step for maximizing performance in niche verticals.

2.1.3 Feature Engineering Classical ML

Despite the deep learning hype, classical approaches remain competitive. Mosquera (2022) achieved remarkable results (F1=0.889) using N-grams and logistic regression. This highlights a critical insight for production systems: traditional models offer comparable accuracy with a fraction of the inference latency and carbon footprint.

2.2 Critical Analysis of Existing Approaches

While previous works achieved high metrics, they suffer from two major limitations that this project aims to address:

- **Over-reliance on Metadata:** Many top-performing systems (HITZ-IXA, 2023) exploit demographic features (gender, profession). While effective in closed-set competitions, these models fail in real-world "cold-start" scenarios where user profiles are incomplete.
- **Adversarial Fragility:** As noted by Mosquera (2022), models lacking robust regularization are vulnerable to simple adversarial attacks (e.g., typos), a vulnerability often overlooked in pure Transformer implementations.

Chapter 3

Data Analysis Bias Detection

3.1 Class Distribution

The dataset exhibits a significant class imbalance, a common characteristic in social media monitoring. As shown in Table 3.1, moderate ideologies constitute the majority (62.6%), while extreme positions are underrepresented, particularly the 'Right' class (13.03%). This distribution necessitates the use of Macro-F1 as the primary evaluation metric to prevent majority-class bias.

Class	Count	Percentage
Moderate Left	9,158	32.63%
Moderate Right	8,412	29.97%
Left	6,839	24.37%
Right	3,656	13.03%

Table 3.1: Target Class Distribution

3.2 The Metadata Illusion: Detecting Data Leakage

Initial exploratory data analysis revealed anomalously high correlations between metadata features and the target variable. Specifically, the *Ideology Binary* feature showed a correlation of 0.87 with the multiclass target.

To investigate potential **data leakage**, we conducted a series of ablation studies:

1. **Baseline with Metadata:** A model using gender, profession, and binary ideology achieved near-perfect accuracy, raising immediate red flags for overfitting.
2. **Feature Ablation:** Removing the *Ideology Binary* feature did not significantly drop performance, suggesting the leakage was systemic.

3.2.1 Root Cause Analysis

A deep dive into the user IDs revealed a critical flaw: the corpus contains only **313 unique users** across 28,000 tweets. Since the train/test split was likely random rather than user-stratified, the model was essentially memorizing user identities (e.g., "User_10 always tweets Right") rather than learning linguistic patterns.

Validation Experiments:

- **User ID Removal:** Removing explicit User IDs caused performance to drop significantly, confirming reliance on identity.
- **Anonymization Test:** Replacing User IDs with unseen tokens caused model failure.

3.3 Strategic Decision

Based on these findings, we decided to **discard all metadata** for the final system architecture. The resulting models rely exclusively on textual content, ensuring they solve the actual NLP task rather than exploiting dataset artifacts. This makes the developed solution robust and applicable to any unseen user.

Chapter 4

System Architecture Methodology

To address the classification task rigorously, we engineered three distinct pipelines, each representing a different paradigm in modern NLP.

4.1 Architecture 1: Domain-Adapted Transformers (PolitiBETO)

This approach leverages transfer learning from a domain-specific base model. We utilized `nlp-cimat/politibeto`, a BERT-based model pre-trained on 5 million political tweets.

4.1.1 Implementation Details

We developed two iterations of this architecture:

- **v1 (Lightweight Fine-Tuning):** To mitigate catastrophic forgetting, we froze the initial layers of the encoder and only fine-tuned the last 4 layers and the classification head.
- **v2 (Seed Ensemble):** To reduce variance and improve generalization, we implemented a multi-seed ensemble strategy. Three separate instances were trained with seeds $\{42, 56, 89\}$ using decoupled weight decay optimization. The final prediction is obtained via soft voting (averaging logits).

4.2 Architecture 2: Feature Engineering (Logistic N-grams)

Prioritizing interpretability and speed, this pipeline relies on explicit feature extraction.

4.2.1 Feature Extraction

- **Lexical Features:** TF-IDF vectorization of word n-grams (1-4) and character n-grams (3-5) to capture morphological patterns.
- **Stylometric Features (v2):** We engineered a set of "text complexity" features including Flesch-Kincaid readability scores, punctuation density, and emoji usage frequency.

4.2.2 Classifier

The features are fed into a **Voting Classifier** composed of Logistic Regression (L2 penalty), Linear SVM, and Random Forest. This ensemble approach balances the bias-variance trade-off inherent in linear vs. non-linear models.

4.3 Architecture 3: Hybrid DualBERT (BETO + MarIA)

This experimental architecture attempts to combine the strengths of two leading Spanish language models: BETO (trained on general Spanish) and MarIA (RoBERTa-based).

4.3.1 Low-Rank Adaptation (LoRA)

Due to the high computational cost of fine-tuning two large transformers simultaneously (220M parameters), we implemented **LoRA (Low-Rank Adaptation)**. By injecting trainable rank-decomposition matrices ($r = 8$) into the attention layers while freezing the pre-trained weights, we reduced the trainable parameter count by 98%, enabling training on consumer-grade hardware.

Chapter 5

Experimental Results

5.1 Benchmarking Overview

Table 5.1 summarizes the performance of all developed architectures.

Table 5.1: Model Performance Comparison (Metrics in %)

Model Architecture	Macro-F1	Accuracy	Inference Speed	Resource Usage
PolitiBETO v2 (Ensemble)	**63.35**	**65.00**	Medium	High (GPU)
PolitiBETO v1	61.78	63.00	High	High (GPU)
Logistic N-gram v2	59.75	62.29	**Very High**	**Low (CPU)**
Logistic N-gram v1	53.00	56.20	Very High	Low (CPU)
DualBERT v2 (LoRA)	37.64	44.67	Low	Very High (GPU)

5.2 Key Findings

5.2.1 The Dominance of Domain Adaptation

PolitiBETO v2 emerged as the superior model with a Macro-F1 of 63.35%. This validates the hypothesis that pre-training on domain-specific data (political tweets) is more impactful than increasing model complexity. The model showed particular strength in classifying "Moderate" ideologies (F1 68%), likely due to the higher density of training data in these regions.

5.2.2 Efficiency vs. Accuracy Trade-off

A crucial finding for production deployment is the performance of **Logistic N-gram v2**. Despite being orders of magnitude faster and cheaper to run than Transformers, it achieved 59.75% F1, lagging only 3.6 points behind the state-of-the-art deep learning model. This suggests that for real-time applications where latency is critical, well-engineered classical models remain a viable, high-ROI alternative.

5.2.3 The Failure of Complexity (DualBERT)

The DualBERT architecture underperformed significantly (37.64%). The analysis suggests that simply concatenating embeddings from two disparate models without sufficient fine-tuning data leads to representation dilution. Furthermore, the reliance on LoRA, while computationally efficient, may have restricted the model's capacity to align the two latent spaces effectively.

Chapter 6

Discussion Data Integrity Audit

6.1 Test Set Contamination Analysis

Following the bias detection in the training phase (Chapter 3), we conducted a forensic audit of the provided test set. The findings confirmed a severe **Data Leakage** scenario:

- **User Overlap:** 100% of the 4,678 users in the test set were present in the training set.
- **Metadata Consistency:** Every user maintained the exact same gender/profession profile.

6.2 Implications for Model Evaluation

These findings imply that the "Test Set" is not a true measure of generalization to unseen data, but rather a test of the model's ability to recognize known entities. If we consider a metadata-based Random Forest as the "Ground Truth" (since it achieves 100% by memorizing users), our text-based models are effectively being evaluated on how well their linguistic patterns correlate with user identity. The fact that PolitiBETO achieves 63% indicates that while ideology strongly influences language, there is significant variance in how individuals express their stance tweet-by-tweet.

Chapter 7

Conclusion Future Work

This project successfully developed and benchmarked robust NLP pipelines for political ideology classification. Our rigorous audit of the dataset exposed critical flaws in metadata reliance, steering the project towards a pure-text approach that guarantees privacy and generalization.

Key takeaways for engineering decision-making:

- **Best Performance:** PolitiBETO v2 (63.35% F1) is recommended for offline analytics where accuracy is paramount.
- **Best Efficiency:** Logistic N-gram v2 (59.75% F1) is the optimal choice for real-time stream processing, offering 95% of the performance at a fraction of the cost.
- **Data Integrity:** Future iterations must enforce strict user-stratified splitting to prevent the "identity memorization" observed in this dataset.

Future work should focus on **Few-Shot Learning** techniques to handle the underrepresented "Right" wing class and exploring **Adapter-based** methods to combine the efficiency of feature engineering with the semantic power of Transformers.

Bibliography

García-Díaz, J. A., Jiménez-Zafra, S. M., Martín Valdivia, M.-T. et al. (2022), 'Overview of politices 2022: Spanish author profiling for political ideology', *Procesamiento del Lenguaje Natural* **69**, 265–272.

HITZ-IXA, E. (2023), 'Document and sentence level representations for demographics and ideology', CEUR Workshop Proceedings. PLACEHOLDER - Incluir URL.

Mosquera, A. (2022), Towards robust spanish author profiling and lessons from adversarial attacks, Technical report, Universidad. PLACEHOLDER - Completar.

NLP-CIMAT-GTO, E. (2022), Politibeto, a domain-adapted transformer for multi-class political author profiling, *in* 'IberLEF Workshop'. PLACEHOLDER - Añadir detalles.

Santamaría Carrasco, A. and Cuervo, A. (2022), 'Loscalls at politices 2022: Political author profiling using beto and maria', *CEUR Workshop Proceedings* . PLACEHOLDER - Actualizar datos.