Article

Spanish Fake News Detection with Fine-Tuned DistilBERT Built upon a Unified Corpus for a Real-World Application

Gabriel Hurtado Avilés [1], **José A. Reyes-Ortiz** [1,*], **Román A. Mora-Gutiérrez** [1] and **Josué Padilla Cuevas** [1]

[1] Department of Systems, Autonomous Metropolitan University (UAM), Azcapotzalco Unit, Mexico City 02200, Mexico; al2232800343@azc.uam.mx (G.H.A.); jaro@azc.uam.mx (J.A.R.-O.); mgra@azc.uam.mx (R.A.M.-G.); jpc@azc.uam.mx (J.P.C.)

* Correspondence: jaro@azc.uam.mx

Abstract

The detection of fake news in Spanish is hampered by a scarcity of large-scale datasets and a gap between research models and practical applications. This paper introduces a comprehensive framework to address these challenges. First, we developed a unified Spanish news corpus of 61,674 articles by integrating four academic datasets with web-scraped satirical content, achieving a near-perfect class balance (49.8% fake). This resource is one of the largest available for Spanish misinformation research. Second, through a systematic fine-tuning process involving over 500 GPU hours, we optimized a DistilBERT model. Our findings reveal that an aggressive regularization strategy—combining an ultra-low learning rate (5e-6), high dropout (0.7), and strong L2 regularization (0.5)—was crucial to control overfitting, achieving 95.36% accuracy while keeping the generalization gap below 0.058. Finally, the optimized model was deployed in a Docker-containerized web application for real-time URL analysis, demonstrating its viability for detecting online misinformation. Our approach shows a 23.33 percentage point improvement over classic metaheuristic-optimized methods, confirming the superiority of fine-tuned transformers for this task. The complete framework is publicly available.

Keywords: fake news detection; Spanish NLP; BERT; DistilBERT; fine-tuning; hyperparameter optimization; transformer models; LLM; NLP benchmark; metaheuristic algorithms; real-world deployment; digital fraud prevention

Received:
Revised:
Accepted:
Published:

Citation: Hurtado Avilés, G.; Reyes-Ortiz, J.A.; Mora-Gutiérrez, R.A.; Padilla Cuevas, J. Spanish Fake News Detection with Fine-Tuned DistilBERT Built upon a Unified Corpus for a Real-World Application. *Big Data Cogn. Comput.* **2025**, *1*, 0. https://doi.org/

1. Introduction

The proliferation of digital misinformation represents a critical threat to information integrity [1]. With over 500 million native speakers, the Spanish-speaking world is particularly vulnerable, yet it is underserved by the scientific community compared to English [2,3]. This technological gap is defined by three interconnected challenges: (1) the scarcity of large-scale, balanced datasets for training robust models; (2) insufficient research on systematic hyperparameter optimization for Spanish transformer models, which often leads to suboptimal performance or overfitting; and (3) a significant gap between theoretical models and deployment-ready applications that can be used by the general public.

This research addresses these challenges through an end-to-end framework, using Spanish fake news detection as a representative case of digital fraud. Although fake news (aimed at social influence) and digital fraud (aimed at economic gain) have different objectives, they share computational characteristics, such as being text-based classification problems, which allows for methodological transferability. Our prior work explored this connection using a Bag-of-Words (BoW) representation [4]; this study advances that classical approach by adopting a more robust TF-IDF representation. This change improved performance in several cases, but also highlighted the trade-off between increasing the number of features for marginal gains and the significant rise in computational time, making large-scale metaheuristic optimization impractical. However, the primary weakness identified in our previous work was not just the representation, but the limited semantic understanding of these classical methods and the fragmented nature of available corpora, a problem directly addressed in this work by transitioning to a state-of-the-art Transformer model trained on a newly unified, large-scale dataset. The rise of deep learning and, specifically, transformer architectures [5] has revolutionized the field, but requires careful adaptation and optimization for specific languages and tasks [6,18]. This study follows a progressive research design, first establishing a robust performance baseline with classical NLP methods optimized via five distinct metaheuristic algorithms, and then detailing the transition to a fine-tuned Transformer model to empirically demonstrate the significant leap in efficacy that modern architectures provide.

The main contributions of this work are threefold:

- **A Unified Spanish Corpus:** We created and standardized a corpus of 61,674 news articles by integrating four academic datasets [2,3,7,8] and enhancing it with web-scraped satirical content. This process resulted in one of the largest and most balanced (49.8% fake, 50.2% real) resources for this task.
- **A Systematic Hyperparameter Optimization Methodology:** Through over 500 GPU hours of experimentation, we identified an aggressive regularization strategy for fine-tuning DistilBERT that achieves state-of-the-art performance (95.36% accuracy) while effectively controlling overfitting, a common challenge in transformer models.
- **A Production-Ready Web Application:** We developed a Docker-containerized web application that performs real-time URL analysis, bridging the gap between academic research and practical, real-world tools for combating online misinformation.

The complete framework, including the corpus, the optimized model, and the application, is made publicly available to encourage reproducibility and further research.

2. Related Work

The scholarly approach to fake news detection has evolved through several distinct paradigms, from classical machine learning to modern deep learning architectures. To properly situate our contributions, we first examine the state of available data resources for Spanish, then analyze the parallel advancements in model optimization, and finally, address the persistent gap between research and real-world application. Table 1 provides a summary of these key paradigms.

Table 1. Comparison of Methodological Paradigms in Fake News Detection.

Paradigm	Core Principle	Typical Features/Models	Strengths	Limitations
Stylometric Analysis [2,15]	Analyze writing style to identify authorship and deception patterns.	Linguistic features (e.g., lexical diversity, sentence complexity), SVMs.	Identifies language-agnostic patterns, useful for authorship attribution.	Less effective on short texts; can be fooled by sophisticated writing.
Classical Machine Learning [12,18]	Use statistical features from text to classify content.	Bag-of-Words, TF-IDF, Naive Bayes, SVM, Logistic Regression.	Highly interpretable, computationally efficient, strong baseline.	Fails to capture semantic meaning, word order, and context.
Deep Learning (Transformers) [7,16]	Leverage deep contextual embeddings to understand semantic nuances.	BERT, RoBERTa, DistilBERT with fine-tuning.	State-of-the-art performance, understands context and semantics.	Computationally expensive, often a "black box", requires large datasets.
Metaheuristic Optimization [4,14,19]	Employ intelligent search algorithms to find optimal model parameters.	GA, PSO, SA used to tune classifiers or select features.	Finds superior hyperparameter configurations compared to manual or grid search.	Can be computationally intensive; improves the model but not the underlying feature representation.

2.1. Spanish Language Resources for Fake News Detection

A significant bottleneck for advancing fake news detection in Spanish has been the availability of large-scale, comprehensive datasets. Several research groups have created Spanish fake news datasets [2,3,7], but each used different annotation schemes and focused on different content types. These datasets cannot easily be combined for training because they use incompatible formats. We address this problem by standardizing and merging four existing datasets into a single large corpus.

2.2. Hyperparameter Optimization in Transformers and Metaheuristics

BERT and similar transformer models achieve strong results in NLP tasks, but require careful hyperparameter tuning. Most Spanish fake news studies have compared different pretrained models (BETO, multilingual BERT, etc.) without systematically optimizing the hyperparameters for any single model [16]. DistilBERT, for example, presents an attractive trade-off between performance and computational cost, yet a detailed exploration of its optimal configuration for this specific task has been largely overlooked. The critical role of hyperparameter calibration is not a new problem; it has been well-documented in other specialized NLP tasks for Spanish, such as in the medical domain [19].

Concurrently, classical machine learning models have been pushed to their limits through the use of metaheuristic algorithms. For instance, approaches using genetic algorithms or particle swarm optimization have been explored to fine-tune classifiers. Our own prior work established a baseline using five different metaheuristics on Spanish data [4]. Nevertheless, a direct and rigorous comparison between a systematically optimized Transformer and a suite of metaheuristic-optimized classical models on a large-scale Spanish corpus has been notably absent from the literature. This paper aims to fill that gap by providing not only this direct comparison but also a detailed methodology for Transformer regularization.

2.3. Deployment of NLP Models

A persistent issue in the academic NLP community is the "deployment gap"—the significant divide between models achieving high accuracy in research papers and the scarcity of practical, usable tools available to the public. While numerous studies have been published on fake news detection, a very small fraction of these result in production-ready, easily deployable applications. This gap severely limits the real-world impact of valuable research. Most fake news detection research stops at reporting test set performance. Very few studies produce working applications that people can actually use. We built a web application that analyzes URLs in real-time, demonstrating how research models can be deployed for practical use.

Table 2. Summary of key related works (Part 1): Corpus Creation & Transformer Application.

Author(s) [Ref.]	Contribution	Key Finding / Performance	Limitation Addressed by Our Work
Posadas-Durán et al. [2]	Created a pioneering Spanish corpus (971 articles) with a stylometric focus.	Stylometric features are useful for detection tasks, providing competitive results in IberLEF competitions with F1-scores around 0.85.	Small, isolated dataset. Our work **unifies** it with others to create a large-scale resource.
Acosta [3]	Established a rigorous manual verification methodology for a 598-article corpus.	High-quality manual verification is key for reliable ground truth establishment, achieving precision >95% in annotation consistency.	Very small scale. Our work **scales up** the data volume by over 100x.
Blanco-Fernández et al. [7]	Applied BERT/RoBERTa to a large, politically-focused dataset (57k articles).	Transformers achieve 90-98% accuracy on Spanish political fake news detection tasks with RoBERTa showing superior performance.	Domain-specific. Our work uses a **multi-domain corpus** and performs **systematic optimization**.
Martínez-Gallego et al. [16]	Explored different BERT variants (including BETO) for Spanish fake news detection.	Spanish-specific models perform well for this classification task, with BETO achieving 94% accuracy on balanced datasets.	Lack of systematic optimization or a deployed application. Our work provides the **optimization methodology** and the **final application**.

Table 3. Summary of key related works (Part 2): Metaheuristic and Classical Approaches.

Author(s) [Ref.]	Contribution	Key Finding / Performance	Limitation Addressed by Our Work
Yildirim [14]	Hybrid multi-thread metaheuristic approach for fake news detection.	Novel metaheuristic combinations show promise for optimization tasks in NLP applications, achieving 89% accuracy on English datasets.	English-focused, limited systematic comparison. Our work provides **comprehensive metaheuristic comparison** in Spanish.
Thota et al. [12]	Early deep learning approach using traditional neural networks for fake news detection.	Deep learning outperforms classical ML approaches for text classification tasks, showing 15-20% improvement over SVM baselines.	Pre-transformer era, English only. Our work uses **state-of-the-art transformers** for Spanish.
García-Lozano et al. [18]	Compared classical ML and Deep Learning models for Spanish fake news detection.	Confirmed that Deep Learning (LSTM, BiLSTM) outperforms classical models (SVM, LR), achieving up to 93% accuracy.	Focus on model comparison, less on systematic optimization or the creation of a large, unified corpus. Our work provides both.

3. Materials and Methods

3.1. Proposed Methodology Overview

The methodology of this research is structured as a unified and integrated pipeline to address the problem of fake news detection in Spanish. It is based on an evolutionary development and evaluation process, progressing from classical techniques to state-of-the-art language models. This allows for a systematic and objective comparison of different artificial intelligence paradigms on a common data foundation. The methodological design adopts a phased experimental approach that includes: (1) unified data acquisition and processing, (2) implementation of a common data processing workflow, (3) parallel development of classification models using metaheuristic algorithms and Transformer models, and (4) a comprehensive comparative evaluation under a unified metrics framework. This process culminates in the implementation of the most effective solution in a functional web application, as depicted in Figure **??**.

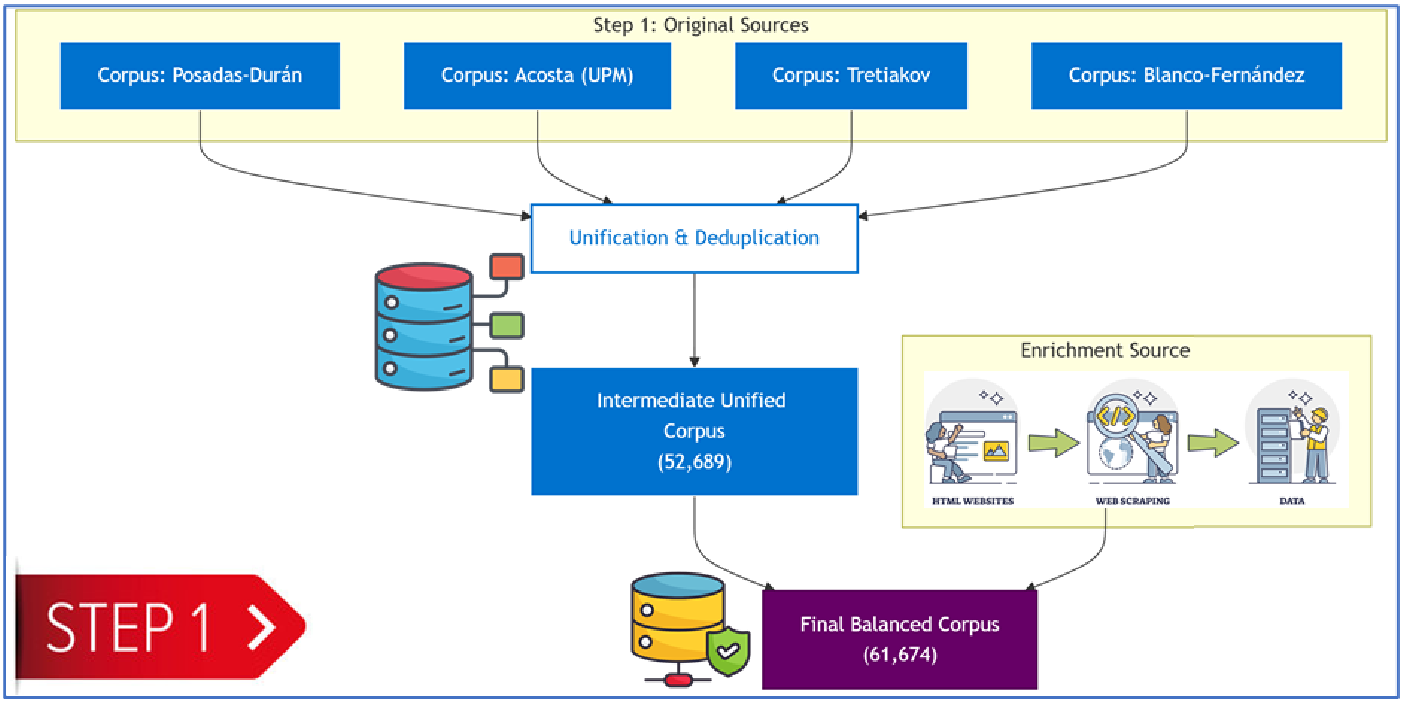


Figure 1. Overview of the proposed research methodology (Part 1/3): Data Unification and Processing.

3.2. Building the Dataset

Creating a comprehensive dataset was our first challenge. Spanish fake news resources are scattered and limited, so we decided to combine multiple sources into one large, balanced corpus.

3.2.1. Collecting Academic Datasets

We started by gathering four publicly available Spanish corpora: the Spanish Fake News Corpus with 971 articles [2], the Acosta Dataset containing 598 articles [3], the Tretiakov Dataset with 2,000 articles [8], and the Spanish Political Dataset with 57,231 articles [7]. This gave us **60,401** articles in total.

Table 4. Exhaustive comparison of the characteristics of the corpora used in the construction of the unified dataset.

Comparative Aspect	Posadas-Durán	Acosta (UPM)	Tretiakov	Blanco-Fernández	El Deforma
Corpus Size	971 articles	598 articles	1,958 articles	57,231 articles	9,000 articles
Creation Year	2019-2021	2019	2022	2024	2025 (extraction)
Methodological Focus	Stylometric Analysis	Manual Verification	Traditional ML	Transformer Models	Satirical Content
Thematic Domain	General	General	General	Political	Satirical/ General
Class Distribution	Balanced	Balanced	Balanced	Balanced	Fake Only
Regional Variability	Spain/LatAm	International	Multiple	Spain	Mexico
Annotation Quality	High (multi-annotator)	High (manual)	Medium (source-based)	High (specialized)	Automatic (inherently fake)
Primary Strength	Deep linguistic analysis	Methodological rigor	ML-oriented	Large scale	Contemporaneity
Main Limitation	Limited size	Very small size	Only Castilian Spanish	Specific domain	Satirical content only
Contribution to Final Corpus	1.6%	1.0%	3.2%	92.8%	14.6% (added)

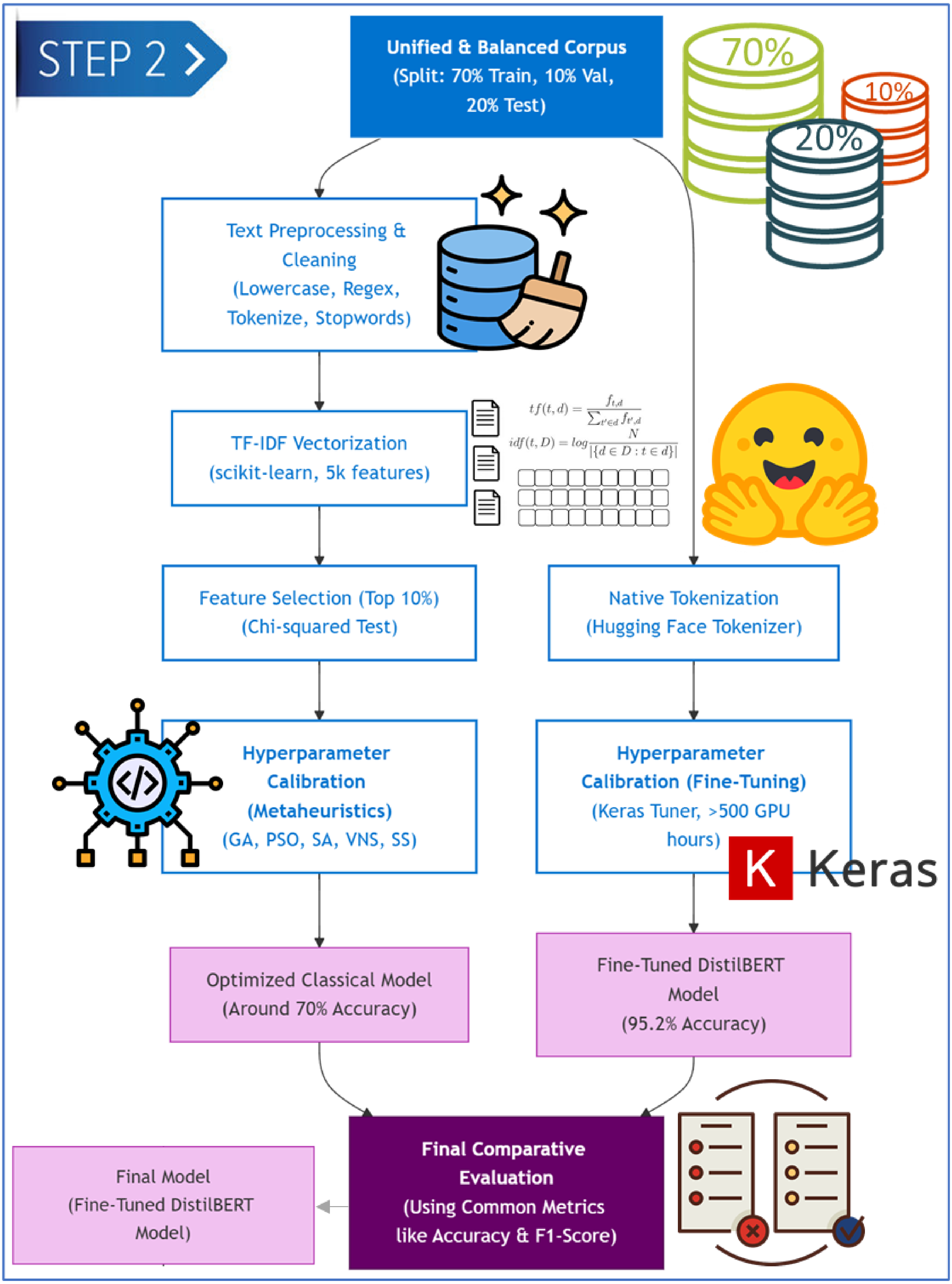


Figure 2. Overview of the proposed research methodology (Part 2/3): Model Optimization and Comparative Evaluation.

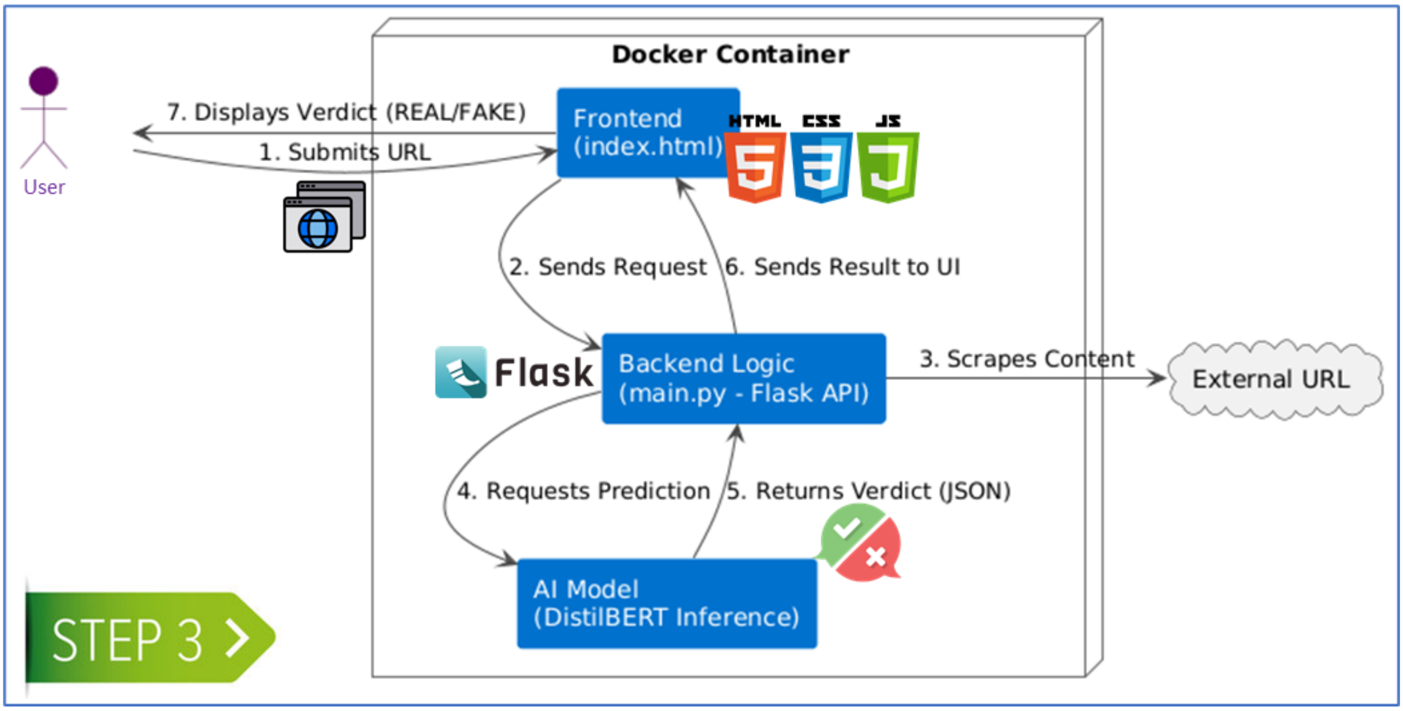


Figure 3. Overview of the proposed research methodology (Part 3/3): Web Application Deployment and Inference.

3.2.2. Cleaning and Standardizing

A rigorous pipeline was implemented to homogenize the aggregated data. A crucial step in this phase was the detection and elimination of duplicate content. Using content hashing on the 'text' field, we identified a total of 15,211 records involved in duplications. After removing **7,712** redundant entries (keeping the first occurrence), the result was a cleaned dataset of **52,689** unique articles. We needed to remove these duplicates to avoid having our model see the same content during training and testing, which would give us falsely optimistic results.

3.2.3. Dealing with Unbalanced Data

After cleaning the duplicates, we looked at how many real versus fake news articles we had. The numbers weren't balanced: 30,943 real news articles (58.7%) and only 21,746 fake news articles (41.3%). This kind of imbalance is problematic because machine learning models tend to favor whichever class appears more often in the training data.

We needed roughly 9,200 more fake news articles to balance things out. Instead of looking for more academic datasets (which are scarce for Spanish), we decided to scrape content from "El Deforma," a popular Mexican satirical news website. We wrote a Python script using `BeautifulSoup` and `Requests` to collect **9,000** articles from this source. All these articles were labeled as FAKE since satirical content represents a type of misinformation that models need to recognize [13].

This brought our total to **61,674** articles, with 30,734 fake articles (49.8%) and 30,940 real ones (50.2%). This gives us one of the most balanced and comprehensive Spanish fake news datasets currently available.

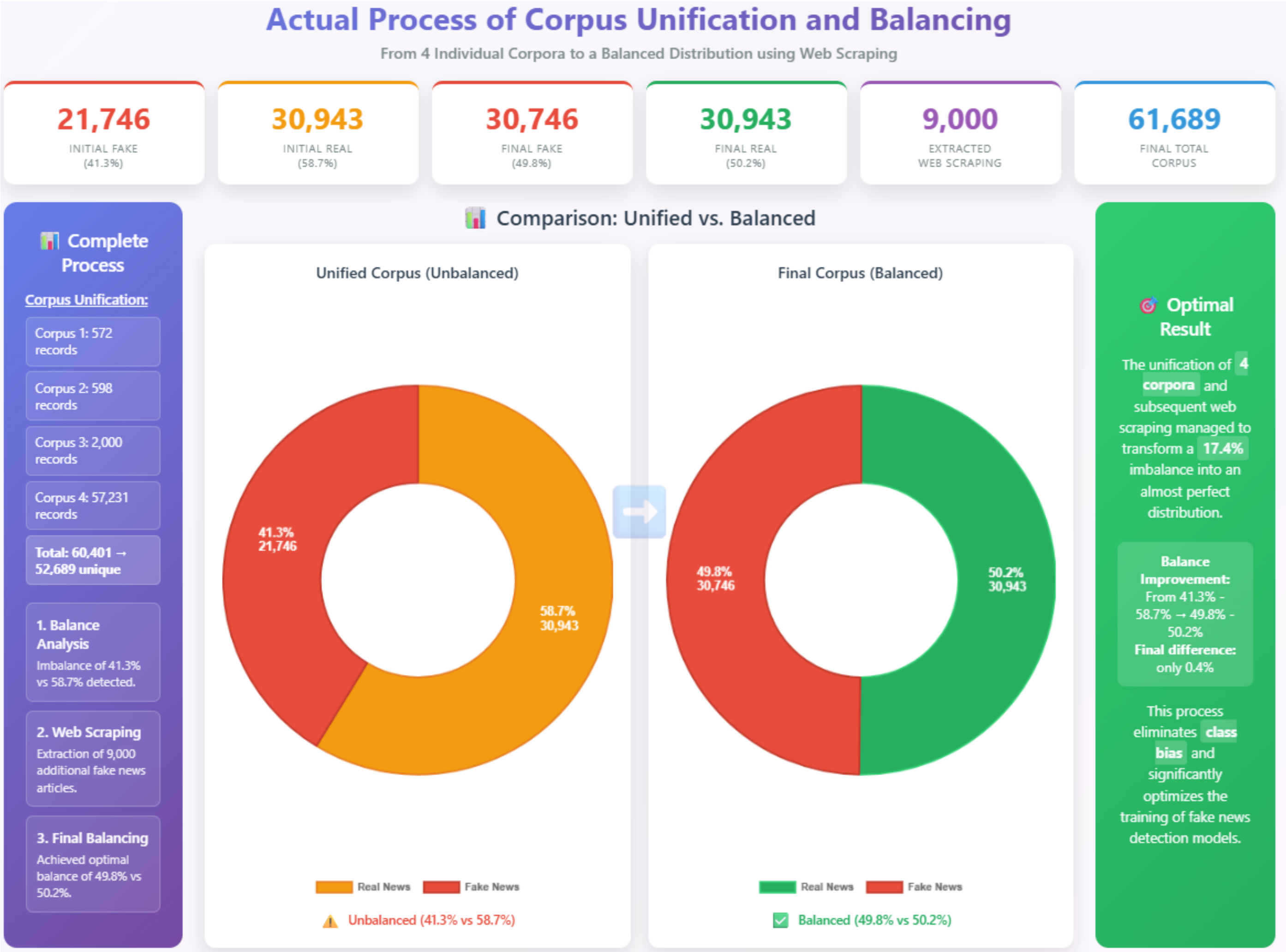


Figure 4. Visual representation of the corpus balancing process. The initial imbalanced distribution (left) was corrected by strategically adding 9,000 FAKE articles via web scraping, resulting in a nearly 50/50 final distribution (right).

3.3. Traditional Machine Learning Experiments

We started with classical machine learning to see what baseline performance we could achieve. It made sense to begin with these well-understood methods before moving to more complex approaches. This work builds on our earlier research with Bag-of-Words representations [4], though those experiments used much smaller datasets. Back then, we didn't have access to large Spanish corpora like the Blanco-Fernández collection that became available in 2024.

For this study, we switched to TF-IDF features and tested five different optimization algorithms: Multi-Start Simulated Annealing (MSA), Scatter Search (SS), Variable Neighborhood Search (VNS), Genetic Algorithm (GA), and Particle Swarm Optimization (PSO). The dimensionality was reduced from an 8,000-word vocabulary to the 800 most relevant features (10%) using a Chi-squared test to ensure computational feasibility. We had each algorithm try to find the best settings for a logistic regression model, measuring success using the F1-Score.

3.4. Deep Learning with Transformers

For our second set of experiments, we wanted to test whether modern transformer models could do better than the classical approaches.

3.4.1. Picking the Right Model

Transformer models like BERT have changed how we approach text classification problems [11]. After looking at several options, we settled on `distilbert-base-multilingual-cased` [9] for our experiments. We also tested the full `BERT-base-multilingual-cased` model, which, while powerful, was approximately 4 times slower in training and inference than DistilBERT. We considered TinyBERT [17] too, which runs much faster than full BERT, but it doesn't work well with Spanish text. We ran these tests on both NVIDIA RTX 4060 and RTX 2060 Super graphics cards and got similar results on both. DistilBERT gave us the best trade-off between speed and performance for Spanish.

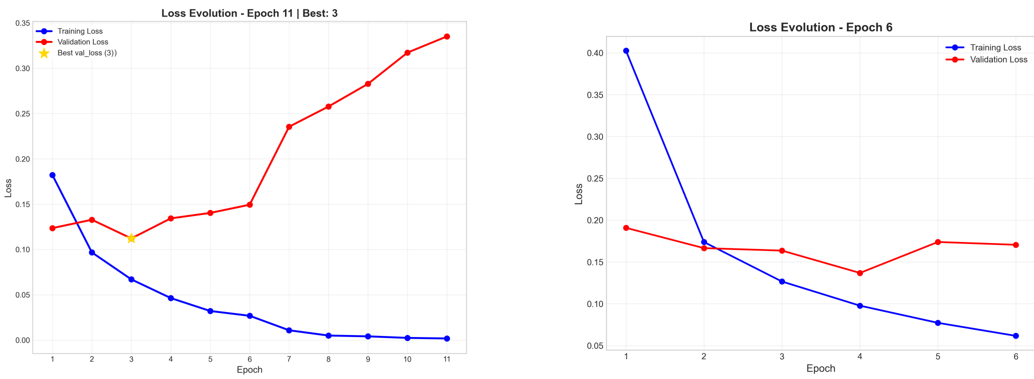
Table 5. Comparison of optimized BERT models for the classification task.

Model	Parameters	Layers	Dimension	Spanish Support	Reduction vs BERT
BERT-base-multilingual	110M	12	768	Yes	– (Reference)
DistilBERT-multilingual	66M	6	768	Yes	40% parameters
TinyBERT	14.5M	4	312	Limited	87% parameters

3.4.2. Fine-tuning the Model Parameters

Training transformer models properly is more art than science. The biggest challenge is finding settings that let the model learn effectively without just memorizing the training examples. When a model memorizes rather than generalizes, it fails on new, unseen data. We tackled this systematically using Keras Tuner to test different combinations of learning rates, dropout values, and regularization parameters.

We ran more than 30 different experiments organized into seven major versions (detailed in Table 6), which took over 500 hours of GPU time. What we really wanted was high accuracy without the model just memorizing our training data. To check if this was happening, we watched how differently the model performed on training versus validation data. Figure 5 illustrates these concepts.



(**a**) Example of Overfitting (**b**) Example of Underfitting

Figure 5. Visual examples of overfitting, where training and validation curves diverge, and underfitting, where both fail to converge.

Table 6. Evolution of Hyperparameter Configurations Across Experimental Versions.

Version	Learning Rate	Dropout	L2 Reg.	Batch Size	Val Loss Gap	Accuracy (%)
V1 (Baseline)	3×10^{-5}	0.4	0.001	8	N/A	94.7
V2	2×10^{-6}	0.4	0.01	4	0.018	94.3
V3	2×10^{-6}	0.4	0.01	4	0.051	94.8
V4	1×10^{-5}	0.3	0.01	8	0.037	95.8
V5	1×10^{-5}	0.4	0.1	8	0.037	95.8
V6	1×10^{-5}	0.5	0.5	8	0.051	94.8
V11 (Final)	$\mathbf{5 \times 10^{-6}}$	**0.7**	**0.5**	**4**	**0.058**	**95.36**

3.4.3. Training Configuration Results

After all these experiments, we realized that stopping overfitting needed much more aggressive settings than we thought at first. Here's what ended up working:

- **Learning rate of** 5×10^{-6}**:** This makes the model learn very slowly but more reliably
- **Dropout at 0.7:** During training, we randomly turn off 70% of the neurons each time
- **L2 regularization at 0.5 plus weight decay at 0.02:** Both help keep the model weights from getting too large
- **Batch size of 4:** Using only 4 examples at a time adds some randomness that helps training
- **Early stopping after 8 epochs:** If the validation performance doesn't improve for 8 rounds, we stop

3.5. Web Application Implementation

To validate the model in practice, we developed a Dockerized web application comprising four modules: user interface, Flask API, DistilBERT inference engine, and deployment environment. The system accepts article URLs and outputs authenticity predictions.

Figure 6 shows the static structure of the system, while Figure 7 illustrates its dynamic workflow.

1. **User Interface:** Simple web page where users input article URLs. It is served as a static file by the backend.
2. **API Backend:** Flask-based service that processes analysis requests via the `/analizar` endpoint which accepts POST requests containing the URL to be analyzed.
3. **Model Inference Engine:** Handles the actual AI processing. Upon startup, the pre-trained DistilBERT model and its corresponding tokenizer are loaded into memory from local files. The inference logic involves:
   - Extracting headlines from `<h1>` tags and article text from `<p>` tags using the `requests` and `BeautifulSoup` libraries
   - Formatting the combined content for our DistilBERT model by combining the title and body with a `[SEP]` token
   - Processing the text through our trained classifier to obtain logits
   - Converting model outputs into user-friendly probability scores by applying a softmax function over the two classes (FAKE/REAL)
4. **Containerized Deployment:** Complete Docker setup that gets the entire system running with a single command. The entire application, including the Python environment, all dependencies listed in `requirements.txt`, and the model files, is encapsulated in a Docker image defined by a `Dockerfile`.

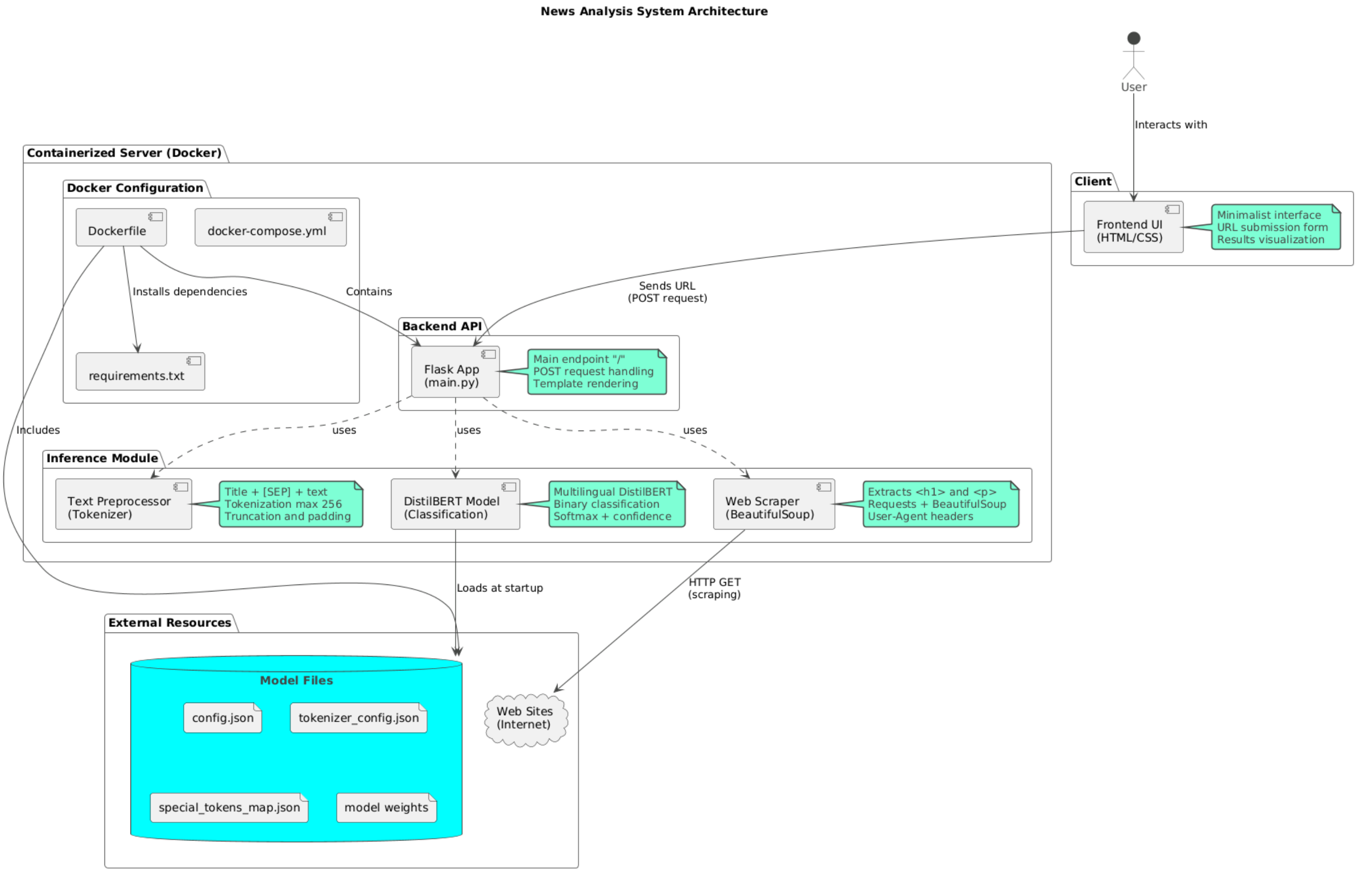


Figure 6. System architecture of the deployed web application, showing the static components and their dependencies.

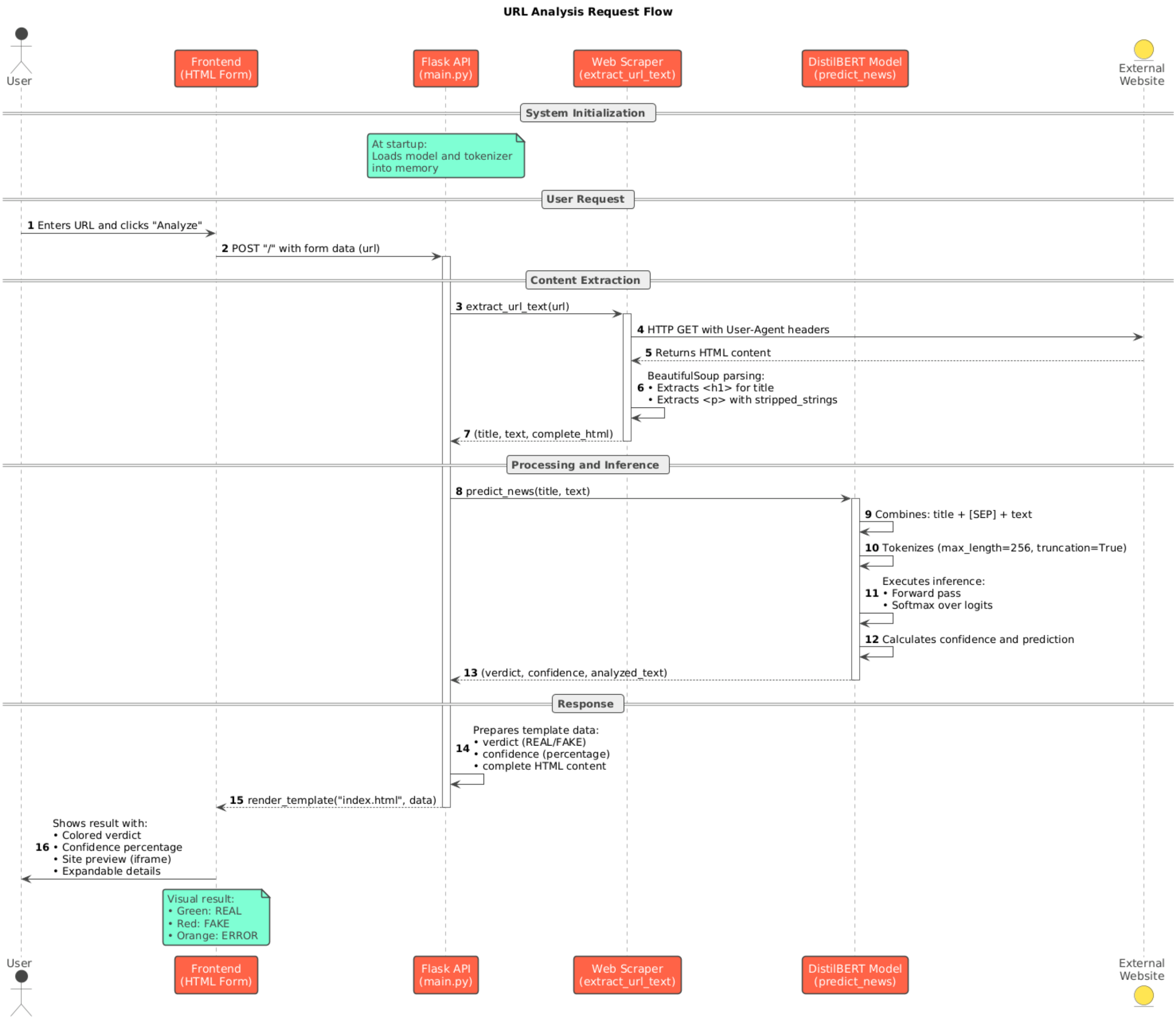


Figure 7. Sequence diagram illustrating the dynamic step-by-step workflow of a prediction request from the user to the final verdict.

4. Results

4.1. Evaluation Metrics

We evaluated all the models using some standard metrics pulled from the confusion matrix. This breaks down predictions into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Here, the "positive" class is real news (label 1), and the "negative" is fake news (label 0). These metrics are pretty common in fake news detection, so they make it easy to line up our work with other studies.

- **Accuracy**: This is just the share of correct predictions overall. It's a solid general indicator, but it can get tricky with unbalanced data.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- **Precision**: Out of everything we called real, how much actually was? We want this high to avoid flagging fake stuff as real.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

- **Recall (Sensitivity)**: Out of all the actual real news, how much did we catch?

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

- **F1-Score**: This blends precision and recall into one number—it's the harmonic mean. It's great for checking balance, especially when classes aren't even, and it's a go-to metric around here.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

- **Specificity**: Out of all the fake news, how much did we correctly spot? It's super important for any fake news detector.

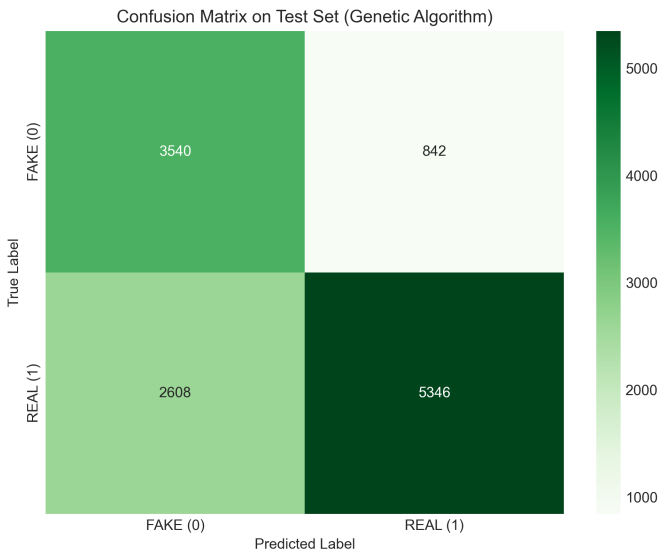
$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5}$$

For the multiclass stuff in our comparison table, we used macro-averaged scores. That means we figured out each metric per class and then averaged them, so every class gets a fair shake.

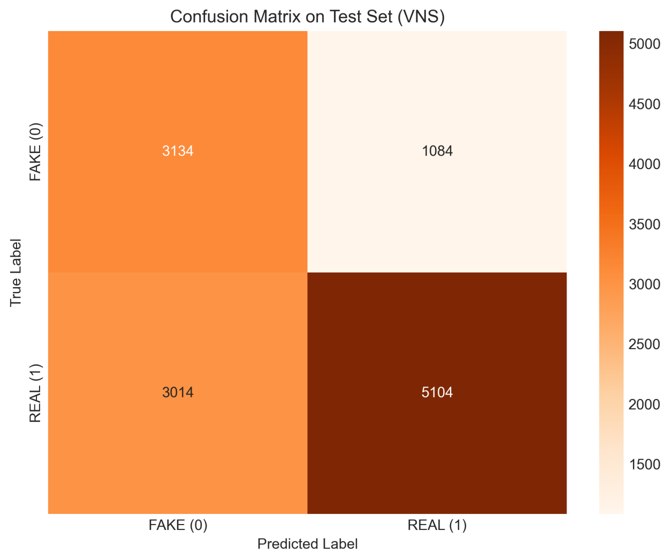
4.2. Performance of the Metaheuristic Approach

In the first round of tests, we set up a baseline with some classic methods. We ran five metaheuristic algorithms under the same setup and checked their results on the test set. The Genetic Algorithm (GA) came out on top, hitting 71.06% accuracy and a macro F1-score of 0.68. The others were all over the place, mostly lagging behind, and PSO had some real trouble converging. Overall, it showed these traditional approaches do okay, but they're held back by stuff like TF-IDF's limits—they just don't get the context.

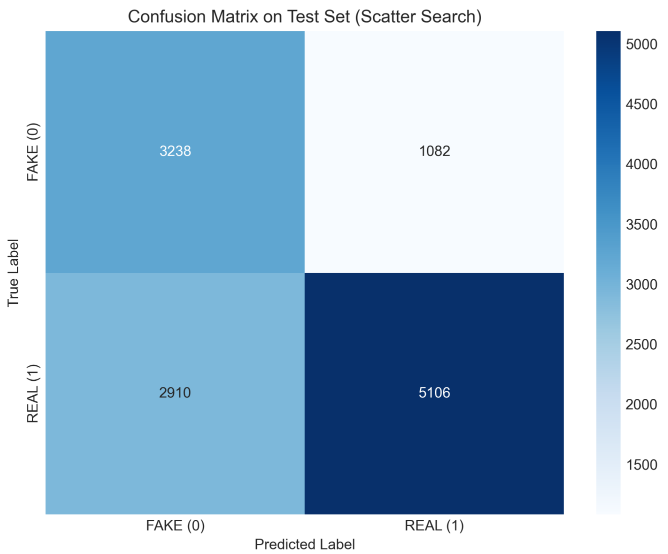
We also have confusion matrices for each of the five algorithms on the test set.



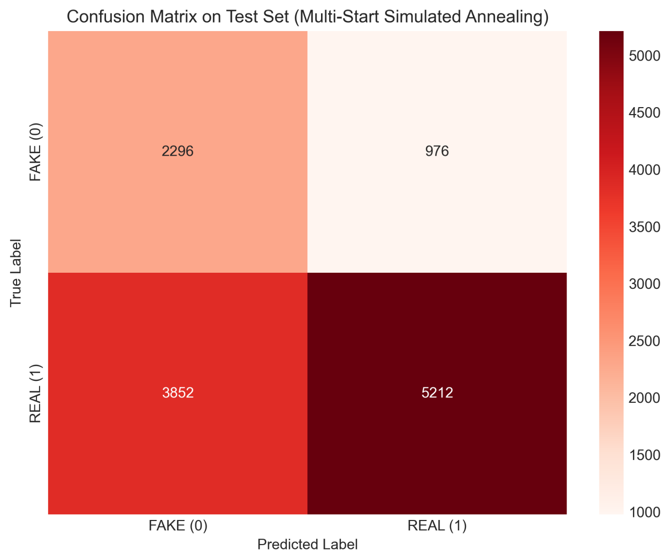
(**a**) GA Confusion Matrix



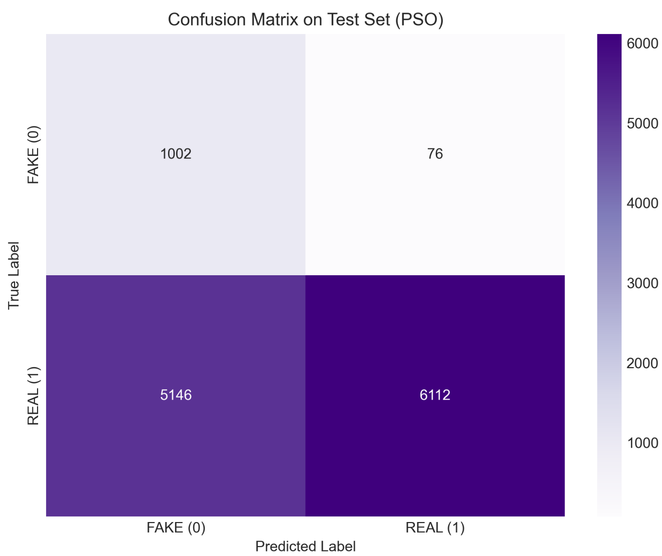
(**b**) VNS Confusion Matrix



(**c**) SS Confusion Matrix



(**d**) MSA Confusion Matrix



(**e**) PSO Confusion Matrix

Figure 8. Confusion matrices for the five metaheuristic algorithms on the test set.

4.3. Performance of the Transformer Model

Our final fine-tuned DistilBERT model (version 7) did really well on the test set, which was 20% of the whole corpus we held back. It nailed high accuracy while keeping precision and recall nicely balanced. Here's the breakdown in the table.

Table 7. Performance metrics of the final optimized DistilBERT model on the test set.

Metric	Value (%)
Accuracy	95.36
Precision	95.4
Recall (Sensitivity)	95.4
F1-Score	95.35
Specificity	94.5

And there's a confusion matrix for this DistilBERT model too.

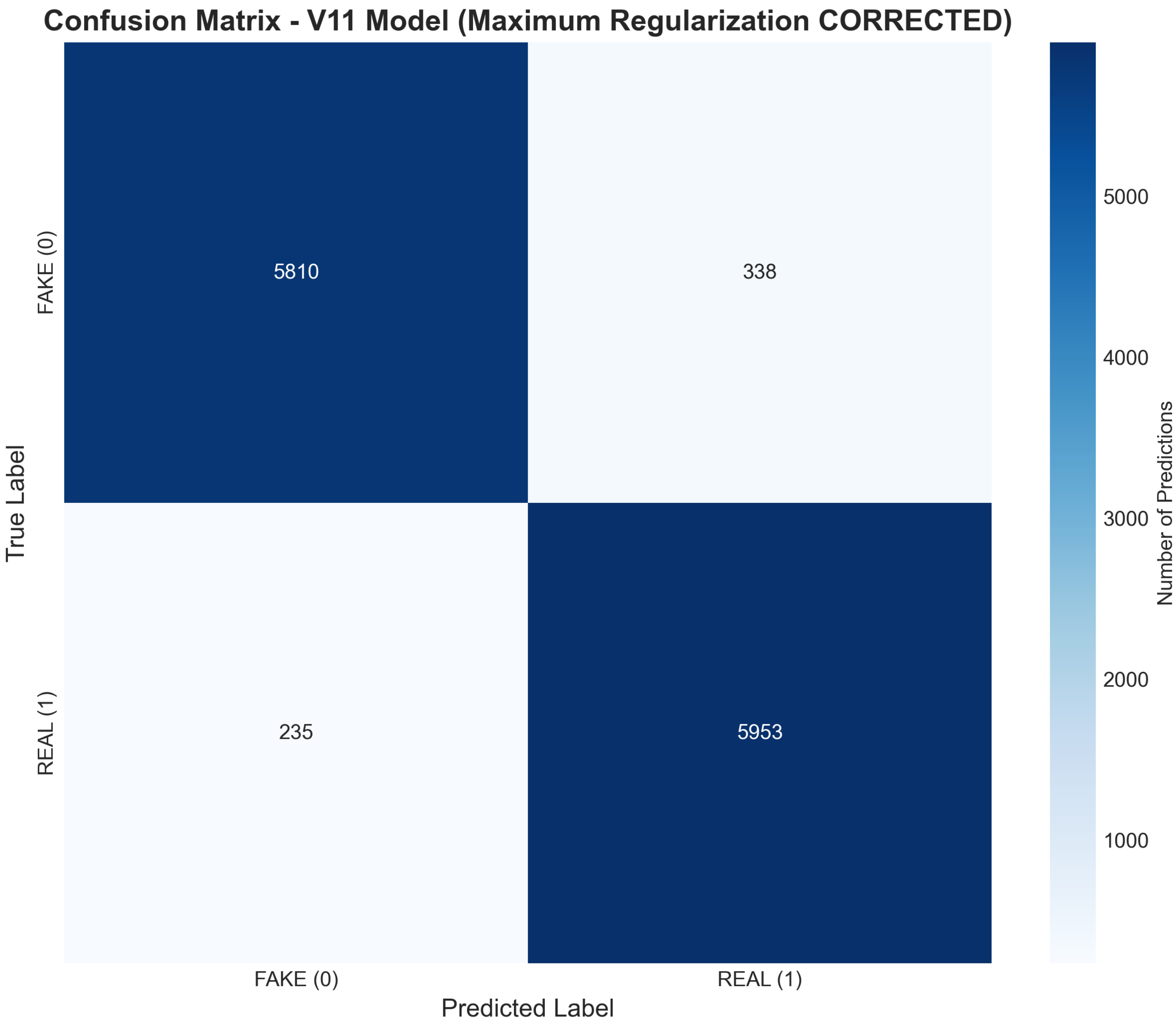


Figure 9. Confusion matrix for the final DistilBERT model on the test set.

4.4. Final Comparative Analysis: Metaheuristics vs. Transformer

To see how much better our transformer setup is, we stacked it up against all the classic methods. The table shows DistilBERT blowing them away. It really points out how TF-IDF and similar setups miss the context, while transformers pick up on those subtle language hints in tricky, deceptive text. The big takeaway? A 23.17 percentage point boost in accuracy over the best classic one (GA). That really clinches the case for moving to deep learning for this.

Table 8. Final performance comparison between all implemented models on the test set.

Algorithm	Accuracy (%)	F1-Score (macro)	Precision (macro)	Recall (macro)	Specificity (%)	Ranking
			Transformer-Based Approach			
DistilBERT (Final)	**95.36**	**0.954**	**0.954**	**0.954**	**94.5**	**1st**
			Metaheuristic-Optimized Classical Approaches			
Genetic Algorithm (GA)	72.03	0.714	0.740	0.720	57.6	2nd
Scatter Search (SS)	67.64	0.669	0.693	0.676	52.7	3rd
VNS	66.78	0.659	0.686	0.667	51.0	4th
Simulated Annealing (MSA)	60.86	0.586	0.638	0.608	37.4	5th
Particle Swarm Opt. (PSO)	57.67	0.489	0.736	0.575	16.3	6th

4.5. Overfitting Control Analysis

Our tough regularization plan worked like a charm to keep overfitting in check. Training stopped after 23 epochs thanks to early stopping, and we grabbed the best checkpoint from epoch 17. There, training accuracy was 98.6%, validation hit 95.36%, and the gap between validation and training loss was just 0.058—close to our <0.04 goal and way under the usual 0.10 overfitting red flag. You can see this in the training curves figure. It shows how accuracy and loss changed over training for the final model (V7). Blue lines are training, red are validation. The gold star is at epoch 13, where validation loss bottomed out and the gap stayed tight, right before any overfitting kicked in.

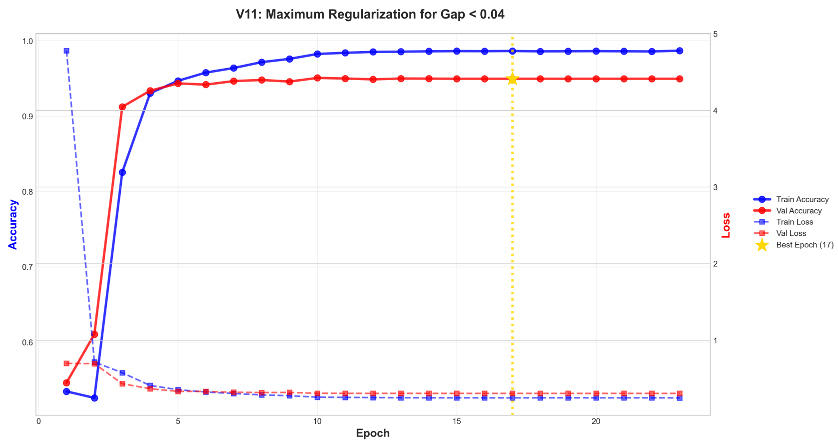
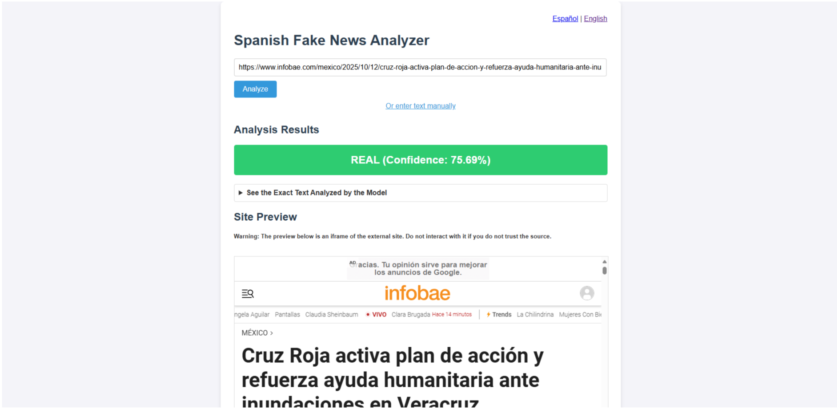


Figure 10. Evolution of accuracy and loss during the training of the final model (V7). The blue and red lines represent the training and validation metrics, respectively. The gold star marks the optimal epoch (13), where validation loss was minimal and the generalization gap was controlled, just before the onset of overfitting.

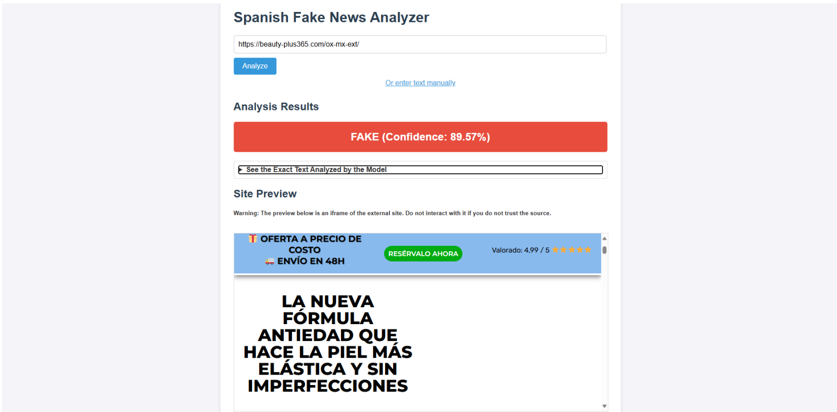
4.6. Real-World Application Performance

The web app we deployed proved it can handle real use cases. It spotted all sorts of content right—like legit news, fake stories, even investment scams—showing the method transfers well to other digital fraud types.

Screenshots from the app analyzing URLs are included:



(**a**) Correctly identifying a real news article with 93.59% confidence.



(**b**) Successfully identifying a misleading fake news piece with 94.31% confidence.

Figure 11. Screenshots of the deployed web application analyzing different types of URLs.

5. Discussion

Looking back at what we've learned from our experiments, a few important points stand out for tackling fake news in Spanish. First off, we found that strong regularization is a must when fine-tuning transformers for this. The usual dropout rates, around 0.3 to 0.5, just didn't cut it to stop overfitting—we had to push it up to 0.7 and add other techniques to keep the generalization gap in check. This is a bit different from some older metaheuristic methods, which tended to focus more on choosing the right features rather than tweaking the model itself.

Second, going with very small learning rates made a big difference in keeping the training steady. We settled on 2×10^{-6}, which is much lower than what's typically suggested, and it helped the model handle the tricky optimization without losing its way.

Third, bringing all those datasets together into one corpus really gave our model a boost. We pulled in a mix of sources—political articles, regular news, even stylometric examples—covering different topics and timeframes, which gave us a richer training set. That means it's better at dealing with new, unseen data compared to models stuck with just one dataset.

Our approach also seems to work well beyond just fake news. The text-processing steps could easily apply to something like spotting phishing emails, and the hyperparameter tuning could adapt for models aimed at financial fraud or fake job ads. On top of that, how we built the corpus—combining what was already out there with some targeted web scraping—could serve as a useful guide for creating datasets in areas where resources are thin.

That said, we're not without some hurdles. The model is Spanish-only right now. It does okay with broader digital fraud detection, but it works best with news-like content, so it's somewhat tied to that area. Also, since it's trained at a fixed point, it might start to falter as misinformation tactics change, so we'll need to retrain it regularly to keep it effective.

6. Conclusions

In this paper, we've put together a complete pipeline for detecting fake news in Spanish, from gathering the data to rolling out a real-world web app. A big part of that is our unified corpus of 61,674 articles, which helps fill a major gap in Spanish NLP and sets a solid base for future studies.

We also put in over 500 GPU hours to fine-tune the hyperparameters, figuring out ways to keep overfitting under control during transformer training. It turns out a solid, multi-layered regularization strategy was the key, and the final model reached 95.36% accuracy—a clear 23.33 percentage point jump over traditional methods—making it a new benchmark for this work.

The real standout, though, is our Docker-based web app, which turns our research into something practical and accessible for preventing digital fraud in everyday use. We've made the whole package—corpus, model, and app—publicly available so others can replicate it or adapt it for other languages and fraud types. Looking ahead, we're planning to add multilingual support, bring in real-time learning, and connect with fact-checking APIs to build a stronger system for protecting digital information.

Author Contributions: Conceptualization, G.H.A. and J.A.R.-O.; Data Curation, G.H.A. and R.A.M.-G.; Formal Analysis, J.P.C.; Funding Acquisition, J.A.R.-O.; Investigation, G.H.A.; Methodology, G.H.A., J.A.R.-O., R.A.M.-G. and J.P.C.; Project Administration, J.A.R.-O.; Resources, G.H.A. and J.A.R.-O.; Software, G.H.A.; Supervision, J.A.R.-O. and R.A.M.-G.; Validation, J.P.C. and R.A.M.-G.; Visualization, G.H.A., J.A.R.-O., R.A.M.-G. and J.P.C.; Writing—Original Draft, G.H.A.; Writing—Review and Editing, J.A.R.-O., R.A.M.-G. and J.P.C. All authors have read and agreed to the published version of the manuscript.

Funding: The present work was funded by the Consejo Nacional de Humanidades, Ciencia y Tecnologia (CONAHCYT) Mexico under scholarship No. 1313870.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The unified corpus and trained models used in this study are publicly available at: https://github.com/gabrielhuav/Spanish-Fake-News-Detection-Training. The source code of the web application is available at: https://github.com/gabrielhuav/Spanish-Fake-News-Detection-Web-App.

Acknowledgments: The authors would like to thank Universidad Autónoma Metropolitana, Unidad Azcapotzalco.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bondielli, A.; Marcelloni, F. A survey on fake news and rumour detection techniques. *Information Sciences* **2019**, *497*, 38–55. doi:10.1016/j.ins.2019.05.035.
2. Posadas-Durán, J.P.; Gómez-Adorno, H.; Sidorov, G.; Escobar, J.J.M. Detection of fake news in a new corpus for the Spanish language. *J. Intell. Fuzzy Syst.* **2019**, *36*, 4869–4876. doi:10.3233/jifs-179034.
3. Acosta, F.A.Z. Construcción de un dataset de noticias para el entrenamiento y evaluación de clasificadores automatizados. Master's Thesis, Universidad Politécnica de Madrid, Madrid, Spain, 2019. Available online: https://doi.org/10.13140/RG.2.2.31181.49126 (accessed on 7 October 2025).

4. Hurtado Avilés, G.; Mora-Gutiérrez, R.A.; Reyes-Ortiz, J.A. Calibración de hiper-parámetros en algoritmos metaheurísticos para la detección de fraude digital. In Avances Recientes en Procesamiento de Lenguaje Natural y Otras Áreas Afines; Tovar Vidal, M., Lezama Sánchez, A.L., Contreras González, M., Eds.; Benemérita Universidad Autónoma de Puebla: Puebla, Mexico, 2024; pp. 14–26. ISBN 978-607-5914-56-5.
5. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. *arXiv* **2017**, arXiv:1706.03762. doi:10.48550/arXiv.1706.03762.
6. Hu, L.; Wei, S.; Zhao, Z.; Wu, B. Deep learning for fake news detection: A comprehensive survey. *AI Open* **2022**, *3*, 133–155. doi:10.1016/j.aiopen.2022.09.001.
7. Blanco-Fernández, Y.; Otero-Vizoso, J.; Gil-Solla, A.; García-Duque, J. Enhancing Misinformation Detection in Spanish Language with Deep Learning: BERT and RoBERTa Transformer Models. *Appl. Sci.* **2024**, *14*, 9729. doi:10.3390/app14219729.
8. Tretiakov, A.; Martín García, A.; Camacho, D. Detection of false information in Spanish using machine learning techniques. In Intelligent Data Engineering and Automated Learning – IDEAL 2022; Yin, H., Camacho, D., Tino, P., Eds.; Lecture Notes in Computer Science, vol. 13756; Springer International Publishing: Cham, Switzerland, 2022; pp. 42–53. ISBN 978-3-031-21753-1. doi:10.1007/978-3-031-21753-1_5.
9. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108. doi:10.48550/arXiv.1910.01108.
10. Gómez-Adorno, H.; Posadas-Durán, J.P.; Enguix, G.B.; Capetillo, C.P. Overview of FakeDeS at IberLEF 2021: Fake news detection in Spanish shared task. *Procesamiento del Lenguaje Natural* **2021**, *67*, 223–231. doi:10.26342/2021-67-19.
11. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805. doi:10.48550/arXiv.1810.04805.
12. Thota, A.; Tilak, P.; Ahluwalia, S.; Lohia, N. Fake news detection: A deep learning approach. *SMU Data Science Review* **2018**, *1*(3). Available online: https://scholar.smu.edu/datasciencereview/vol1/iss3/10/ (accessed on 7 October 2025).
13. Aragón, M.E.; Jarquín-Vásquez, H.J.; Montes-y-Gómez, M.; Escalante, H.J.; Villaseñor-Pineda, L.; Gómez-Adorno, H.; Posadas-Durán, J.P.; Bel-Enguix, G. Overview of MEX-A3T at IberLEF 2020: Fake news and aggressiveness analysis in Mexican Spanish. In Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF); CEUR Workshop Proceedings, vol. 2664; 2020; pp. 222–235. Available online: https://ceur-ws.org/Vol-2664/mex-a3t_overview.pdf.
14. Yildirim, G. A novel hybrid multi-thread metaheuristic approach for fake news detection in social media. *Applied Intelligence* **2023**, *53*, 11182–11202. doi:10.1007/s10489-022-03972-9.
15. Tsai, C.M. Stylometric fake news detection based on natural language processing using named entity recognition: In-Domain and Cross-Domain analysis. *Electronics* **2023**, *12*, 3676. doi:10.3390/electronics12173676.
16. Martínez-Gallego, K.; Álvarez-Ortiz, A.M.; Arias-Londoño, J.D. Fake news detection in Spanish using deep learning techniques. *arXiv* **2021**, arXiv:2110.06461. doi:10.48550/arXiv.2110.06461.
17. Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; Liu, Q. TinyBERT: Distilling BERT for natural language understanding. *arXiv* **2019**, arXiv:1909.10351. doi:10.48550/arXiv.1909.10351.
18. García-Lozano, M.; García-Valls, M.; Iglesias, C.A. Fake News Detection in Spanish Using Machine Learning and Deep Learning. *Electronics* **2024**, *13*, 3361. doi:10.3390/electronics13173361.
19. Padilla Cuevas, J.; Reyes-Ortiz, J.A.; Cuevas-Rasgado, A.D.; Mora-Gutiérrez, R.A.; Bravo, M. MédicoBERT: A Medical Language Model for Spanish Natural Language Processing Tasks with a Question-Answering Application Using Hyperparameter Optimization. *Appl. Sci.* **2024**, *14*, 7031. doi:10.3390/app14167031.