



Fake News Detection and Dataset Analysis in Ecuador's 2025 Elections: A Comparative Study of Machine Learning Models with X Social Media Platform Data

Milagros Delgado¹, Britney Robalino Ramírez¹, Brithany Suárez¹,
Aisha Gómez², Diego Almeida-Galarraga³, Paulina Vizcaíno-Imacaña⁴,
and Andrés Tirado-Espín^{*2}

¹ School of Physical Sciences and Nanotechnology, Yachay Tech University, Hacienda San José s/n, San Miguel de Urcuquí 100119, Ecuador.

² School of Mathematical and Computational Sciences, Yachay Tech University, Hacienda San José s/n, San Miguel de Urcuquí 100119, Ecuador.

³ School of Biological Sciences and Engineering, Yachay Tech University, San Miguel de Urcuquí 100119, Ecuador.

⁴ Facultad de Ingenierías Digitales y Tecnologías Emergentes, Universidad Internacional del Ecuador (UIDE), Quito 170411, Ecuador.

ctirado@yachaytech.edu.ec

Abstract. In a fragmented and polarized society, recognizing fake content is essential for safeguarding democratic processes. This problem was particularly relevant in the context of Ecuador's 2025 Presidential Elections, where misinformation can significantly influence public opinion. This study evaluates the effectiveness of large language models (LLMs) in detecting fake news related to this electoral process on X (formerly Twitter). A dataset of 600 news samples was compiled. Three traditional machine learning models and five pre-trained Spanish-language LLMs were fine-tuned and compared. The analysis revealed that 'Daniel Noboa' was the most frequent term in both real and fake news, reflecting a strong social media presence during the campaign. Among the evaluated models, RoBERTuito achieved the best performance, reaching 90% accuracy. The results confirm the potential of transformer-based models for fake news detection in Spanish-speaking political contexts and highlight their contribution to strengthening public discourse and promoting electoral transparency.

Keywords: Social media, medical pseudoscience, digital misinformation, public health, media literacy.

1 Introduction

Social media platforms have become key channels for maintaining an informed society, but they also facilitate the spread of false information. MIT research reveals false news is 70% more likely to be retweeted than truthful content and reaches people six times faster [1]. In Ecuador, fake news intensifies during electoral periods, with content often created to oppose specific figures or issues [2], significantly affecting election outcomes [3].

According to recent data, around 70% of Latin Americans do not know how to detect fake news [5], and a survey indicates that almost half of the population admits that they occasionally question what they read or do not question it at all [6]. A different study showed that only 26% of organizations have adopted formal guidelines to address it, and only 7.2% have implemented technologies specifically designed to identify disinformation. These challenges reflect two threats to democratic systems in the region: widespread manipulation through disinformation and distrust in political leadership, which contribute to the rise of populist figures in Latin America [5].

Ecuador is particularly vulnerable, ranking second in Latin America for the least discussion of fake news, with 78% of citizens struggling to identify false information [7],[8]. To combat this, Ecuador has developed initiatives such as Ecuador Chequea and the Ecuador Verifica coalition to promote fact-checking and digital literacy [9]. To complement these journalistic efforts, recent studies have begun to explore the use of machine learning and natural language processing to automate the detection of fake news; however, in Spanish-speaking contexts, this remains a technical and linguistic challenge.

Therefore, this study aims to evaluate the reliability of large language models (LLMs) and three traditional machine learning models in detecting fake news in the 2025 Ecuadorian Election. The goal of this research is to contribute to the development of more accurate and efficient real-time fake news detection methods within Spanish-speaking contexts.

2 State of Art

Large Language Models (LLMs) are advanced neural networks containing billions of parameters. They have the ability to learn from large amounts of unlabeled text using self-supervised learning to capture complex linguistic patterns and semantic relationships [10]. The LLMs development process generally follows two main stages: pre-training and transfer learning. During the first step, the model learns a generalized representation of language through unsupervised objectives, while in the transfer learning phase, this representation is fine-tuned for specific downstream tasks using labeled data [11]. This allows LLMs to perform well across a wide range of applications with minimal task-specific training. Moreover, instead of building models from scratch, these pre-trained models can serve as a flexible base for efficiently developing new models for various tasks [12].

The strategies used for the literature review were the search applied in academic databases such as Google Scholar, Scopus, IEEE, and the use of literature reviewers such as LitMaps that helped visualize and connect information according to key search terms such as 'Fake news detection', 'Machine Learning (ML)', 'Large Language Models (LLM)', 'Spanish speaking contexts', 'Ecuadorian elections'. The selected articles ranged from 2018 to 2025 with a focus on experimental studies or systematic reviews. Finally, these findings were classified according to the type of model used, its precision metrics, accuracy metrics, and the dataset used, to identify trends, strengths, and challenges in the field, which are described in Table 1.

Transformer-based models consistently achieve high accuracy (96-98%) across languages and datasets [23],[25],[26], with hybrid approaches showing particular promise for resource-limited contexts [24],[27]. Data quality remains critical regardless of model choice: using validated labels [23], fact-checking sources [27], or data augmentation [26] significantly improves generalization and accuracy.

Table 1. Comparison of different machine learning methods for fake news detection in Spanish languages.

Study	Methods	Results	Strengths	Weaknesses
Mafla N. et al. (2022) [4].	TF-IDF, SVM, Naïve Bayes, Random Forest.	SVM achieved the best results, improving performance by 15%.	Good for structured text classification.	Poor performance on unstructured or satirical news.
Park, M. and Chai, S. (2023) [13].	LR, NNET, CART, SVM, and RF.	CART, NNET achieved 94.6% accuracy and LR, SVM 91.3%.	Feature selection (XGBoost), focusing on relevant attributes.	Small Dataset (402 tweets) limits the robustness of findings.
Jehad, R. and Yousif, S. (2020) [14].	Random Forest and Decision Tree.	Random Forest 84.97% and Decision Tree 89.11% accuracy.	Larger dataset (20,761 samples), content-centric (TF-IDF vectorization).	Lower accuracy (complex content) and limited feature diversity.
Verma, K. et al. (2021) [15].	Hybrid model WELFake (linguistic features + word embeddings).	Achieving an accuracy of 96.73% using SVM.	Outperform BERT and CNN models.	Barriers in applicability to nontextual or multilingual content.
Toapanta M. et al. (2024) [16].	BETO, MarIA, RoBERTuito, BERTuit, BERTin.	MarIA outperformed other models.	Strong for Spanish-specific text classification.	Requires large, labeled datasets.
Moreno P. et al. (2023) [17].	LSTM, MLP, CNN.	LSTM achieved the best F1 score (0.746).	Works well for sequential data.	High computational cost.
Sánchez A. et al. (2024) [18].	Logistic Regression, Naïve Bayes, SVM, LSTM, BETO.	BETO achieved the highest accuracy (0.98).	Transformer-based models excel in feature learning.	Susceptible to bias from training data.

Queiroz H. et al. (2020)[19].	k-NN, SVM, Random Forest, XGB.	Random Forest achieved 85.3% accuracy.	Language-independent features improve generalization.	Lower performance on highly context-dependent news.
Martínez K. et al. (2021) [20].	BERT, BETO, LSTM, CNN.	BETO + LSTM achieved 80% accuracy.	Pretrained embeddings enhance performance.	Error propagation in translation strategies.
Sultana, A. et al. (2023) [21].	Logistic Regression (LR) and SVM.	LR achieved 95.12% accuracy, and SVM 91.68%.	Using 4 features makes the model fast and easy to interpret.	Small dataset limits the robustness of findings.
Oluwabunmi, A. et al. (2024) [22].	Decision Tree (DT), Random Forest (RF), LR.	DT had 99.64%, RF had 99.23%, and LR had 98.80% accuracy.	Large-scale dataset and its features.	Kaggle datasets may not fully represent real-world scenarios.
Raza, S. et al. (2025) [23].	BERT models (RoBERTa, BERT-based).	Achieving a classification accuracy of 89-90%.	Outperform LLMs like Llama2-7B.	Poor data quality affects model performance.

3 Research Design and Methodology

3.1 Dataset Acquisition and Preprocessing

A dataset of 600 samples equally divided between real and fake news was manually compiled from three different sources relevant to Ecuador’s 2025 presidential elections, in the time interval consisting of March 2024 and April 2025. These sources included:

- Fact-checking organizations: Posts verified as true or false by the accounts Ecuador Chequea and Ecuador Verifica on platform X.
- Official repositories: Fake news archives maintained by the Consejo Nacional Electoral (CNE) of Ecuador.
- Verified news outlets: Posts from legitimate, recognized media organizations, such as the official X account of Teleamazonas.

To ensure clean, standardized input for model fine-tuning, the collected dataset underwent multiple preprocessing steps. These included URL removal, HTML tag stripping, emoji filtering, and punctuation removal.

3.2 Models Selection and Fine-tuning

We selected five publicly available pre-trained LLMs specialized in Spanish or multi-lingual contexts and three traditional MLMs. The large language models (LLMs) include BETO, a Spanish BERT model trained on a large Spanish corpus [28]; RoBERTuito, a RoBERTa-based model fine-tuned for Spanish Twitter content [29]; MARIA/RigoBERTa, Spanish transformer models trained on mixed-domain data including news and social media [30]; BERTuit, a Spanish BERT variant tailored to Twitter and short texts [31]; and BERTin, a RoBERTabased model pre-trained on massive

Spanish corpora [32]. In addition, we employed three traditional machine learning models: a logistic regression model, a neural network with one hidden layer and ReLU activation, and a neural network with two hidden layers and dropout for regularization. For the fine-tuning of the large language models, we used the HuggingFace Transformers library [33].

3.3 Performance Evaluation Metrics

To evaluate the classification performance of each model, we used the following standard metrics: accuracy, precision (positive predictive value), recall (sensitivity), and F1 Score (harmonic mean).

3.4 Availability of Materials

All code, preprocessed datasets, and training configurations used in this study are publicly available at: [https://github.com/britssc/ecuadorian fake news detection](https://github.com/britssc/ecuadorian_fake_news_detection).

4 Results

4.1 Dataset Analysis

The 20 most frequent bigrams found in fake and real news can reveal distinct linguistic and thematic patterns between the two classes (see Fig. 1). In real news, bigrams such as “segunda vuelta” (second round), “las elecciones” (the elections), and “13 abril” (April 13), these suggest that are centers around high-impact electoral events. In contrast, fake news exhibits bigrams that are not informative, but rather function-word combinations or name tokens with low semantic content. Notably, the most frequent bigrams in real news appear less in fake news. This suggests a broader lexical diversity in journalism.

These lexical differences reinforce the hypothesis that fake and real news differ not only in factual accuracy but also in linguistic style and structural composition. Such distinctions can be leveraged by automated detection systems to improve the classification of fake news.

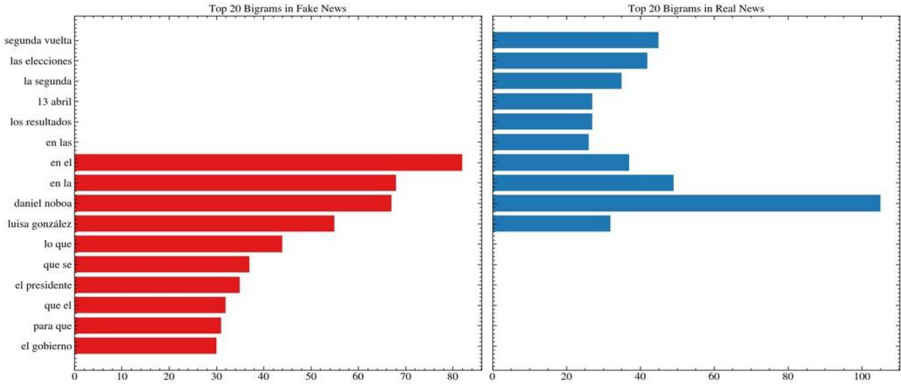


Fig. 1. Top 20 most frequent bigrams in fake news (left, in red) and real news (right, in blue).

Daniel Noboa is mentioned substantially more often than Luisa González in both true (blue) and false news (see Fig. 2). Approximately one third of Noboa's mentions occurring in false news, indicating greater media exposure and substantial susceptibility to misinformation. In contrast, Luisa González shows fewer mentions with a more balanced true–false distribution. This disparity suggests uneven media visibility among candidates, which may influence public perception during the electoral process.

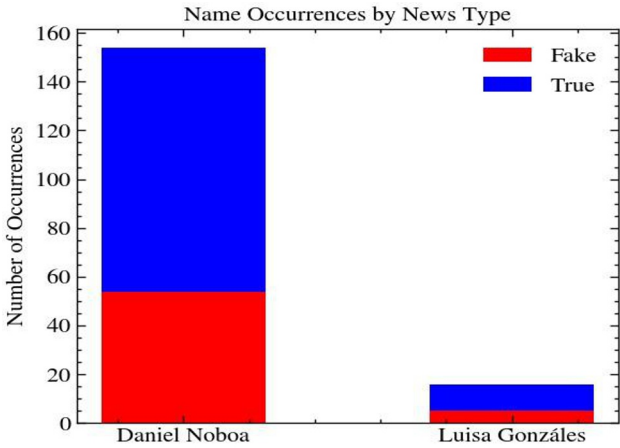


Fig. 2. Frequency of mentions of presidential candidates Daniel Noboa and Luisa González in the dataset.

4.2 Model Performance Comparison

We obtained a comparative evaluation of eight models based on key performance metrics, as described in Table 2.

Table 2. Models comparison in terms of Accuracy, Weighted Precision (WP), Weighted Recall (WR), F1-score, and class-specific F1-scores for class 0 (Fake) and class 1 (Real).

Model	Accuracy	WP	WR	Macro F1	Class 0 F1	Class 1 F1
BETO	0.91	0.92	0.91	0.91	0.9	0.92
RoBERTuito	0.93	0.93	0.93	0.92	0.92	0.93
BERTin	0.85	0.86	0.85	0.85	0.83	0.87
RoBERTa	0.87	0.87	0.87	0.87	0.85	0.89
Neural Network 1	0.84	0.84	0.84	0.84	0.81	0.86
Logistic Regression	0.85	0.85	0.85	0.85	0.83	0.87
RoBERTa / MarIA	0.81	0.81	0.81	0.8	0.77	0.84
Neural Network 2	0.77	0.81	0.77	0.76	0.78	0.74

RoBERTuito achieved the best performance across all metrics (accuracy, precision, and recall of 0.93; macro-F1 of 0.92), followed closely by BETO (macro-F1 of 0.91). In contrast, Neural Network 2 and RoBERTa/MarIA obtained the lowest scores, while classical models showed moderate performance. Overall, Table 2 confirms that Spanish-specific transformer models, particularly RoBERTuito and BETO, outperform traditional approaches in fake news detection.

5 Discussion

One of the main challenges was building the dataset due to the scarcity of publicly available resources in Spanish, especially regarding disinformation in Ecuador. Unlike English, Spanish has limited resources for artificial intelligence datasets, leading to the manual collection and review of news from entities such as Ecuador Chequea. Although contextual relevance was ensured, it was a laborious process and difficult to scale. This lack of public datasets delays and limits academic research. Consequently, it is necessary to promote collaborative work for the development of open and properly labeled datasets.

Furthermore, the analysis of the dataset revealed some interesting patterns. Fake news tends to show fewer bigrams related to specific events or election dates, and more bigrams related to candidates’ names. This can make it difficult to verify information with official sources and could be part of an emotional manipulation or negative association strategy. One notable finding was Daniel Noboa’s strong presence on social media, suggesting a well-established digital campaign that may have influenced his electoral success. This data provides a basis for analysis; however, more in-depth and focused research is needed to reach definitive conclusions.

Regarding model performance, in our experiment, RoBERTuito achieved the best performance ($F1 = 0.92$), closely followed by BETO, which is consistent with the findings of Vélez-Bermello [9], where models such as MarIA and BERTin also achieved outstanding metrics ($F1 > 0.95$). RoBERTuito’s success can be attributed to its pre-training on social media texts. Otherwise, the classical models such as Logistic regression ($F1 = 0.85$), demonstrated solid performance, outperforming deep neural networks

such as Neural Network 1 ($F1 = 0.84$). This result reflects the usefulness of traditional models; however, our results do not compare with the metrics obtained by the English models [22], which achieved even higher metrics ($F1 = 0.99$).

In our case, the limited performance of some architectures, such as Neural Networks, can be explained by the relatively small size of the dataset, suggesting overfitting and low generalization. Furthermore, the superior results reported in [22] could be related to the fact that their models were trained on an open-source dataset available on Kaggle with over 40,000 labeled instances. This contrast emphasizes the crucial importance of having large balanced datasets for the effective development of robust models. To illustrate this difference more clearly, Table 3 compares the best-performing models from our experiment with some relevant related studies.

Table 3. Comparison of the best-performing models from our work and related studies.

Model	Study	Accuracy	F1 Score	Dataset (Size / Language)
MarIA	Vélez-Bermello [13]	0.96	0.96	1600 / Spanish
Decision Tree	Ayankemi et al. [28]	1	1	40000 / English
RoBERTuito	Our experiment	0.93	0.92	600 / Spanish

It is important to emphasize that a consistent pattern emerged, models tended to underperform in identifying fake news compared to real news, demonstrating greater accuracy and precision in classifying the latter. This performance disparity highlights the inherent difficulty of detecting disinformation, which, as evidenced by the dataset analysis, lacks clear and well-defined patterns, such as specific dates or distinctive keywords not also found in legitimate news.

Added to this difficulty is the nature of advanced models like Transformers. While they offer better results, their “black box” nature complicates the interpretation of their decisions. Furthermore, these models require considerably more computational resources, training time, and storage capacity compared to other models. This can limit their practical application in resource-constrained environments or in contexts where efficiency and interpretability are prioritized.

6 Conclusion

Overall, this study examined the challenges of detecting fake news in Spanish, especially in politically specific contexts such as the 2025 Ecuadorian elections. The scarcity of large, publicly available Spanish-language datasets poses a significant obstacle, as it necessitates time-consuming and resource-intensive manual data collection processes. Despite these limitations, models based on transformers pre-trained with Spanish text, particularly RoBERTuito, demonstrated promising performance. However, the adaptable nature and variability of disinformation, along with the computational cost and interpretability challenges of these models, present obstacles to their practical application in resource-constrained environments.

On the other hand, the analysis of the dataset revealed patterns such as the avoidance of explicit references and the strategic repetition of candidate names in fake news. A possible relationship between social media visibility and electoral success was observed; however, further research is needed to draw conclusive results. Future work should prioritize the development of larger open datasets in Spanish, and the application of techniques to improve the explainability of models and their efficiency.

Acknowledgements

The authors thanks to Yachay Tech University for the financial support given to the project PII26-15 through the Internal Call for Research Project Funding 2026.

References

1. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. *Science* **359**, 1146–1151 (2018). <https://doi.org/10.1126/science.aap9559>
2. Vite, M.A.B., Albuja, P.E.H.: Comportamientos generacionales y contenidos informativos que construyen el mundo de las Fake News en Ecuador. *Revista Scientific* **5**, 208–229 (2020). <https://doi.org/10.29394/scientific.issn.2542-2987.2020.5.16.11.208-229>
3. Olan, F., Jayawickrama, U., Arakpogun, E.O., Suklan, J., Liu, S.: Fake news on Social Media: The Impact on Society. *Information Systems Frontiers* **26**, 443–458 (2022). <https://doi.org/10.1007/s10796-022-10242-z>
4. Mafla, N., Flores, M., Castillo, S., Andrade, R.: Automatic detection of fake news in Spanish: Ecuadorian political satire. *Revista Politécnica* **50**, 7–16 (2022). <https://doi.org/10.33333/rp.vol50n3.01>
5. Hernández, S., Herrera, M.: Fake News and Democracy in Latin America. *Politeja* **19**, 85–101 (2023). <https://doi.org/10.12797/Politeja.19.2022.81.04>
6. Athaydes, A.S., Álvarez Nobell, A., Sadi, G.: Fake news y relaciones públicas en América Latina. In: *Mediaciones de la Comunicación* **15** (2020)
7. Moreno, A., Molleda, J.C., Álvarez Nobell, A., Herrera, M., Athaydes, A., Suárez, A.M.: Comunicación estratégica y sus retos: fake news, confianza, información para la toma de decisiones, liderazgo y compromiso laboral. Resultados de una encuesta en 19 países. EUPRERA/DIRCOM, Bruselas–Madrid (2019)
8. Castillo, E.D.: Las Fake News: Descripción e impacto de la desinformación en medios digitales ecuatorianos. *MLS Communication Journal* **1**, 7–24 (2021)
9. Vélez-Bermello, G.L.: Fact-checking: campaña política de elecciones seccionales en Ecuador. Caso Ecuador Verifica. *Comhumanitas* **15**, 214–232 (2024). <https://doi.org/10.31207/rch.v15i2.458>
10. Raiaan, M.A.K., et al.: A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access* **12**, 26839–26874 (2024). <https://doi.org/10.1109/ACCESS.2024.3365742>
11. Patil, R., Gudivada, V.: A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). *Applied Sciences* **14**, 2074 (2024). <https://doi.org/10.3390/app14052074>
12. Myers, D., Mohawesh, R., Chellaboina, V.I., Sathvik, A.L., Venkatesh, P., Ho, Y.H., Henshaw, H., Alhawawreh, M., Berdik, D., Jararweh, Y.: Foundation and Large Language Models: Fundamentals, Challenges, Opportunities, and Social Impacts. *Cluster Computing* **27**, 1–26 (2024). <https://doi.org/10.1007/s10586-023-04203-7>

13. Park, M., Chai, S.: Constructing a User-Centered Fake News Detection Model by Using Classification Algorithms in Machine Learning Techniques. *IEEE Access* **11**, 71517–71527 (2023). <https://doi.org/10.1109/ACCESS.2023.3294613>
14. Jehad, R., Yousif, S.A.: Fake news classification using random forest and decision tree (J48). *Al-Nahrain Journal of Science* **23**, 49–55 (2020). <https://doi.org/10.22401/ANJS.23.4.09>
15. Verma, P.K., Agrawal, P., Amorim, I., Prodan, R.: WELFake: Word Embedding Over Linguistic Features for Fake News Detection. *IEEE Transactions on Computational Social Systems* **8**, 881–893 (2021). <https://doi.org/10.1109/TCSS.2021.3068519>
16. Bernabe, M.T., Garcia-Cumbreras, M.A., Urena-Lopez, L.A.: Fake News Detection and Fact Checking in X posts from Ecuador Chequea and Ecuador Verifica using Spanish Language Models. *Revista Tecnológica - ESPOL* **36**, 158–173 (2024). <https://doi.org/10.37815/rte.v36n2.1219>
17. Moreno-Vallejo, P.X., Bastidas-Guacho, G.K., Moreno-Costales, P.R., Chariguaman-Cuji, J.J.: Fake News Classification Web Service for Spanish News by using Artificial Neural Networks. *International Journal of Advanced Computer Science and Applications* **14** (2023). <https://doi.org/10.14569/ijacsa.2023.0140334>
18. Gálvez, A.M.S., Albores, F.J., González, R., Conde, S.G., Gálvez, S.S.: A study on the Detection of Fake News in Spanish. *International Journal of Combinatorial Optimization Problems and Informatics* **15**, 85–94 (2024). <https://doi.org/10.61467/2007.1558.2024.v15i2.467>
19. Abonizio, H.Q., De Moraes, J.I., Tavares, G.M., Barbon, S.J.: Language-Independent fake news detection: English, Portuguese, and Spanish mutual features. *Future Internet* **12**, 87 (2020). <https://doi.org/10.3390/fi12050087>
20. Martínez-Gallego, K., Alvarez Ortiz, A.M., Arias-Londoño, J.D.: Fake news detection in Spanish using deep learning techniques. *arXiv* (2021). <https://doi.org/10.48550/arXiv.2110.06461>
21. Sultana, A., Islam, M., Hasan, M., Ahmed, F.: Fake News Detection Using Machine Learning Techniques. In: *Proceedings of SERA 2023*, pp. 98–103 (2023). <https://doi.org/10.1109/SERA57763.2023.10197712>
22. A., O., R., O., A., A.: Fake News Detection System Using Logistic Regression, Decision Tree and Random Forest. *British Journal of Computer, Networking and Information Technology* **7**, 115–121 (2024). <https://doi.org/10.52589/BJCNIT-IOYRPY7G>
23. Raza, S., Paulen-Patterson, D., Ding, C.: Fake news detection: comparative evaluation of BERT-like models and large language models with generative AI-annotated data. *Knowledge and Information Systems* (2025). <https://doi.org/10.1007/s10115-024-02321-1>
24. Abdelhakim Othman, N., Elzanfaly, D.S., Elhawary, M.M.M.: Arabic Fake News Detection Using Deep Learning. *IEEE Access* **12**, 122363–122376 (2024). <https://doi.org/10.1109/ACCESS.2024.3451128>
25. Roumeliotis, K.I., Tselikas, N.D., Nasiopoulos, D.K.: Fake News Detection and Classification: A Comparative Study of Convolutional Neural Networks, Large Language Models, and Natural Language Processing Models. *Future Internet* **17** (2025). <https://doi.org/10.3390/fi17010028>
26. Oad, A., Hamza Farooq, M., Zafar, A., Ayesha Akram, B., Zhou, R., Dong, F.: Fake News Classification Methodology With Enhanced BERT. *IEEE Access* **12**, 164491–164502 (2024). <https://doi.org/10.1109/ACCESS.2024.3491376>
27. LekshmiAmmal, H.R., Madasamy, A.K.: A reasoning based explainable multimodal fake news detection for low resource language using large language models and transformers. *Journal of Big Data* **12** (2025). <https://doi.org/10.1186/s40537-025-01093-x>

28. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish PreTrained BERT Model and Evaluation Data. In: Proceedings of PML4DC at ICLR 2020 (2020)
29. Pérez, J.M., Furman, D.A., Alonso Alemany, L., Luque, F.M.: RoBERTuito: a pre-trained language model for social media text in Spanish. In: Proceedings of the Language Resources and Evaluation Conference (LREC), Marseille, France, pp. 7235–7243 (2022)
30. Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., SilveiraOcampo, J., Carrino, C.P., Gonzalez-Agirre, A., Armentano-Oller, C., RodriguezPenagos, C., Villegas, M.: MARIA: Spanish Language Models. arXiv (2021). <https://doi.org/10.26342/2022-68-3>
31. Huertas-Tato, J., Martín, A., Camacho, D.: BERTuit: Understanding Spanish language in Twitter with transformers. *Expert Systems* **40** (2023). <https://doi.org/10.1111/exsy.13404>
32. la Rosa, E.G., Ponferrada, E.G., Romero, M., Villegas, P., González de Prado Salas, P., Grandury, M.J.D.: BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling. *Procesamiento del Lenguaje Natural* **68**, 13–23 (2022)
33. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Transformers: State-of-the-Art Natural Language Processing. In: Liu, Q., Schlangen, D. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020). <https://doi.org/10.18653/v1/2020.emnlp-demos.6>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

