Unsupervised Ensemble Clustering for Fake News Detection in Brazilian Portuguese using Contextual and Traditional Embeddings

Nathalia Raquel Bianchini[2], Jean Alex Firmino Pinhata[2], Alex Marino Gonçalves de Almeida[1], Rodrigo Capobianco Guido[2]

[1] Departamento de Análise e Desenv. de Sistemas – Faculdade de Tecnologia de Ourinhos
Av. Vitalina Marcusso, 1440, Campus Universitário – Ourinhos – SP – Brasil

alex.marino@fatecourinhos.edu.br

[2]Departament of Computer Science
São Paulo State University
Rua Cristóvão Colombo 2265, Jd Nazareth
15054-000, São José do Rio Preto - SP, Brazil

n.bianchini@unesp.br, jean.pinhata@gmail.com, guido@ieee.org

Abstract. *The automated detection of fake news in Brazilian Portuguese often relies on supervised learning, which may be constrained by the scarcity of labeled data. In this study, we evaluate an unsupervised ensemble clustering framework using the Fake.br Corpus and compare traditional baselines (TF-IDF, Doc2Vec, FastText) with contextual embeddings (BERTimbau). Our results suggest that BERTimbau tends to group deceptive narratives into clusters with higher internal consistency and to separate legitimate news into two thematic groupings. These results indicate that dense contextual vectors could have potential for unsupervised disinformation detection in this language.*

1. Introduction

The proliferation of fake news, defined as fabricated narratives, has been associated with efforts to discredit traditional media, undermine institutional trust, and influence public opinion. This poses a notable challenge in the information age, with potential to affect society by shaping political narratives, impacting public health communication, and posing risks to democratic processes [Colomina et al. 2021, Sparemberger and da Silva 2021].

The rise of social media platforms appears to have accelerated the dissemination of unverified information, often exceeding the capacity of traditional fact-checking agencies to respond. This imbalance may contribute to increased digital insecurity. According to [Newman et al. 2025], public concern about the authenticity of online content has risen to 56% globally, suggesting a growing crisis of confidence and underscoring the need for effective, scalable, and automated tools to address disinformation.

While significant progress has been made in automated detection, most existing studies rely heavily on supervised learning [Vieira et al. 2025, Trevilato 2025]. These models depend on massive, manually labeled datasets, which constitute a critical bottleneck during rapidly evolving disinformation crises where ground truth labels are scarce.

To circumvent this limitation, unsupervised methodologies offer a promising alternative by dynamically discovering latent thematic patterns [Paixão et al. 2020]. However, the efficacy of such algorithms inherently depends on how well the text is mathematically represented in the vector space [Jurafsky and Martin 2026]. Consequently, capturing the semantic complexity of fake news in Brazilian Portuguese requires a robust evaluation of different embedding architectures to ensure accurate and stable thematic segregation.

Motivated to address this gap in the literature, the present study designed a comprehensive experimental pipeline using the *Fake.br Corpus*[1] dataset [Monteiro et al. 2018]. For each extracted text representation, dimensionality reduction was applied, followed by optimal k tuning via the Elbow Method [Han et al. 2011] and Silhouette Score [Rousseeuw 1987] analysis. Subsequently, an ensemble clustering framework integrating six distinct algorithms was implemented. Finally, the quality and stability of the resulting clusters were assessed through a robust set of internal and external validation metrics.

The main contributions of this work are threefold, focusing on the following primary objectives:

- evaluating the extent to which contextual embeddings outperform traditional static text representations in capturing linguistic patterns and argumentative structures inherent to false narratives in Brazilian Portuguese;
- demonstrating the capacity of an unsupervised ensemble clustering methodology to achieve accurate and stable segregation of fake and authentic news without relying on ground truth labels; and
- identifying the underlying thematic structures that differentiate deceptive content from legitimate news when clustered in a high-dimensional semantic space.

Briefly outlining the findings of this study and addressing the first of the contributions aforementioned, our evaluation suggests that contextual embeddings may offer an advantage in distinguishing false information from legitimate news, when compared to traditional static models. In regard to our second objective, the ensemble methodology provided superior stability and accuracy compared to standalone clustering algorithms (e.g., Ward's method). The dynamic bidirectional vectors were extracted using *BERTimbau* [Souza et al. 2020], a pre-trained language model for Brazilian Portuguese based on the *BERT* architecture.

Aligning with our third contribution, these embeddings underwent consensus clustering, which revealed a nuanced tripartite structure within the corpus. In our experiments, the unsupervised framework grouped deceptive information into a single, highly homogeneous cluster reaching 97.46% purity, while legitimate news was separated into two distinct thematic domains.

The paper is organized as follows: Section 2 reviews recent studies on the development of Brazilian Portuguese corpora and others that endeavored to detect fake news in the language. In Section 3, the proposed methodology—encompassing dataset selection, data preparation, embedding extraction, and the ensemble clustering routine—is detailed. The experimental results and a comparative analysis of the representation models are presented in Section 4, and Section 5 concludes the paper.

[1] https://github.com/roneysco/Fake.br-Corpus

2. Related Work

The detection of disinformation dissemination is an active research field, with most studies on Brazilian Portuguese relying on supervised learning paradigms to establish reference benchmarks. [Monteiro et al. 2018] pioneered this field by introducing the *Fake.br Corpus*. By applying classical models such as *Support Vector Machines* (*SVM*) and *Multilayer Perceptron* (*MLP*), they demonstrated that the lexical and emotional signatures of fake news are identifiable when ground-truth labels are provided, achieving 89% accuracy and setting the baseline for automatic detection.

Building upon this supervised foundation, subsequent research has focused on optimizing text representation architectures and addressing data scalability. [Vieira et al. 2025] investigated the impact of different embeddings on multiclass classification, finding that while classical methods like *Random Forest* remain competitive, *Transformer*-based models, specifically *BERTimbau*, significantly improve performance by capturing contextual depth.

In parallel, [Trevilato 2025] analyzed the influence of incremental data growth on model robustness. The author concluded that integrating multiple datasets with *TF-IDF* representations and SVC classifiers yields highly stable results, with F1-scores exceeding 0.96 in high-data regimes. Despite these algorithmic advancements, a common limitation across these studies is their reliance on large-scale manual annotations—a resource often scarce in real-time disinformation crises.

To circumvent reliance on labeled data, unsupervised methodologies have emerged as a necessary alternative for capturing dynamic thematic patterns. In contrast to the predominantly supervised landscape, [Paixão et al. 2020] explored whether latent thematic structures could differentiate fake news from real information without prior labeling. Using *Latent Dirichlet Allocation* (*LDA*) for topic modeling, they successfully identified dominant disinformation themes, demonstrating that unsupervised discovery is a viable approach for detecting hidden patterns in Brazilian Portuguese discourse.

3. Methodology

Our approach follows the Knowledge Discovery in Databases (KDD) process, "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [Fayyad et al. 1996]. As shown in Figure 1, the pipeline comprises eight stages: (i) data acquisition; (ii) preprocessing; (iii) feature extraction with traditional and contextual embeddings; (iv) dimensionality reduction; (v) k-optimal tuning; (vi) ensemble clustering; (vii) performance evaluation; and (viii) knowledge discovery.

These refine the five canonical KDD steps to the needs of unsupervised text mining: data acquisition maps to *selection*; preprocessing is retained; feature extraction, dimensionality reduction, and k-optimal tuning jointly form *transformation*; ensemble clustering is the *data mining* stage; and performance evaluation with knowledge discovery realize *interpretation/evaluation*.

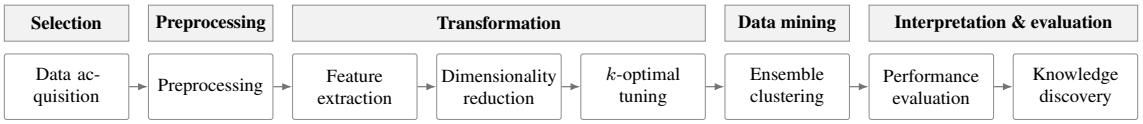


Figure 1. Methodological pipeline for unsupervised fake news detection: eight operational stages (bottom) grouped under the five canonical KDD phases (top).

3.1. Data Acquisition

We use the *Fake.br Corpus* [Monteiro et al. 2018], the first dataset designed for fake news detection in Brazilian Portuguese and the field benchmark. It relies on news alignment—every fake story has a true counterpart on the same topic or event—preventing thematic dominance in the vector space and forcing the framework to capture the structures that distinguish deceptive from legitimate content.

Table 1. Text distribution per category in the *Fake.br Corpus* dataset.

Category	Quantity	Percentage
Politics	4,180	58.0%
TV & Celebrities	1,544	21.4%
Society and Daily News	1,276	17.7%
Science and Technology	122	1.5%
Economy	44	0.7%
Religion	44	0.7%
Total	**7,200**	**100.0%**

As Table 1 shows, the corpus holds 7,200 plain-text items collected from January 2016 to January 2018, split evenly into 3,600 true and 3,600 fake news.

3.2. Preprocessing

To keep the setting fully unsupervised, the *Fake.br Corpus* ground-truth labels were withheld during preprocessing and used only later, in *knowledge discovery*, to validate the clusters with external metrics (Section 3.6); clusters are formed solely from the documents' latent features.

Preprocessing was tailored per model. All methods start from the standardized, pre-cleaned corpus released by the original authors [Monteiro et al. 2018]. For the lexical and statistical models (*TF-IDF*, *Doc2Vec*, *FastText*) we added a routine using the spaCy library (version 3.8.9)[2] with the `pt_core_news_sm` model: (a) tokenization; (b) normalization (lowercasing, accent and punctuation removal); (c) stopword removal of low-semantic high-frequency terms (e.g., pronouns, prepositions); and (d) lemmatization via Part-of-Speech (PoS) tagging, reducing each word to its canonical lemma. For *BERTimbau*, we kept the authors' structured input to leverage its bidirectional attention and preserve the syntactic nuances crucial for deception detection.

[2]https://spacy.io

3.3. Feature Extraction

An embedding is the vector representation of a word—or, in some cases, a sentence—capturing semantic and some grammatical properties, and treating that unit as a single concept [Das et al. 2024]. We examine two parallel approaches, denoted in Section 3 as traditional and contextual embeddings. The traditional approach covers lexical-statistical methods, which rely on term frequency but miss semantic proximity, and static representations, which capture semantics but assign a fixed vector to each word regardless of context. The contextual approach uses pre-trained neural architectures that consider the whole sentence, so the same word takes different values depending on its surroundings. Their dimensions and setups are detailed below.

3.3.1. Traditional Embeddings

Term Frequency-Inverse Document Frequency (*TF-IDF*) is a statistical measure of a term's relevance within a document relative to a corpus. Using the Gensim library (version 4.4.0)[3]—our primary framework for non-*Transformer* representations—we built a Bag-of-Words (BoW) corpus after tokenization, computed the *TF-IDF* weights, discarded terms occurring in fewer than five documents to reduce noise, and mapped the sparse representation into a dense $26,500$-dimensional matrix.

Doc2Vec, or *Paragraph Vectors*, is an unsupervised extension of *Word2Vec* [Le and Mikolov 2014]. We implemented the *Distributed Memory* (*PV-DM*) architecture, which predicts a paragraph's terms from two vector types—a per-instance paragraph vector and several shared, fixed word vectors—so each paragraph vector captures its sentence's specific context. Vectors were trained with Stochastic Gradient Descent and backpropagation of error [Le and Mikolov 2014], converging after 60 epochs, over a grid search $d \in \{50, 100, 200\}$; a frequency threshold filtered out rare tokens to focus on the corpus's most representative vocabulary.

FastText is a subword-based representation [Bojanowski et al. 2017] that decomposes words into character n-grams, letting it embed Out-Of-Vocabulary (OOV) words by combining their constituent n-gram vectors. This is reported as beneficial for morphologically rich languages such as Brazilian Portuguese, offering robustness to grammatical errors, slang, and neologisms common in web publications while capturing broader context and nuance. We used the *Skip-Gram* architecture with a 5-word context window, a subword range of 3–6 characters, and mean pooling (arithmetic average of word embeddings) for document-level vectors; as with *Doc2Vec*, we evaluated $d \in \{50, 100, 200\}$, training for 30 epochs.

3.3.2. Contextual Embeddings

The *Bidirectional Encoder Representations from Transformers* (*BERT*) is an encoder-only model based on the *Transformer* architecture that derives each word's semantic value from its left- and right-context [Shah et al. 2025], yielding dynamic representations that surpass static models, particularly for polysemous terms. For Brazilian Portuguese

[3]https://radimrehurek.com/gensim/

we employed *BERTimbau* (`neuralmind/bert-base-portuguese-cased`) via the SentenceTransformers framework [Reimers and Gurevych 2019]; its large-scale Portuguese pre-training suits the subtle semantic ambiguities central to identifying deceptive patterns [Souza et al. 2020]. The model extracted dense 768-dimensional vectors, condensed by mean pooling as in *FastText*, with a batch size of 64 for efficiency.

3.4. Dimensionality Reduction and K-Optimal Tuning

The extracted features are high-dimensional—from 768 (*BERTimbau*) to $26,500$ (*TF-IDF*)—raising the "curse of dimensionality" [Bellman 1961]. We therefore applied the *Uniform Manifold Approximation and Projection* (*UMAP*) library (version 0.5.9.post2)[4] to project the vectors, chosen for preserving both local and global topology, essential for identifying semantic clusters in text [McInnes et al. 2020]. The space was set to 10 components—a trade-off between efficiency and information preservation that still yields stable distance calculations—with a fixed seed (`random_state = 42`) for reproducibility.

Parameter tuning used a two-fold Silhouette Score analysis: first on the raw embeddings to verify the data's inherent structure, then via a K-Elbow Visualizer on the 10-dimensional *UMAP* space over $k \in [2, 30]$. Reduction increased cluster cohesion for *BERTimbau*, with the maximum Silhouette Score rising from 0.100 in the original space to 0.394 in the reduced manifold. Although the visualizer peaked at $k = 3$—possibly a third distinct cluster—the high $k = 2$ scores indicate the 10-dimensional space adequately captures the binary fake-vs-true distinction present in the *Fake.br Corpus*.

3.5. Ensemble Clustering

To increase stability and robustness, the framework applies consensus clustering over the unlabeled corpus, integrating multiple paradigms to mitigate individual biases and capture complex topology in the 10-dimensional *UMAP* manifold. The ensemble comprises six algorithms—*K-Means*, *K-Medoids*, *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (*HDBSCAN*), *Bayesian Gaussian Mixture* (*GMM*), *Agglomerative Clustering* (Ward), and *Fuzzy C-Means*—implemented with the Scikit-learn library (version 1.7.2).[5] Each algorithm is run independently to label every document, and these labels build a co-association matrix C of size $N \times N$ ($N = 7,200$), where each entry C_{ij} is the proportion of algorithms assigning documents (i, j) to the same cluster, a similarity measure. Finally, from the distance matrix $D = 1 - C$, a last *Agglomerative Clustering* solver with average linkage and the target number of clusters k derives the definitive ensemble labels, yielding a more reliable segmentation of the *Fake.br Corpus* than any individual baseline.

3.6. Performance Evaluation

We evaluate the ensemble and individual clusters with internal and external validation metrics. The internal Silhouette Score, based on the data's intrinsic attributes, measures cluster cohesion and separation within the 10-dimensional *UMAP* manifold:

$$\text{Sil.} = \frac{b - a}{\max(a, b)} \tag{1}$$

[4]https://umap-learn.readthedocs.io/en/latest/

[5]https://scikit-learn.org/stable/

where a is the mean intra-cluster distance, b is the mean nearest-cluster distance, and $\max(a, b)$ is the maximum value between a and b. The external metrics—Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Homogeneity, and V-Measure—quantify agreement between the predicted clusters and the corpus ground-truth labels.

4. Results and Discussion

The results obtained are presented and discussed in this section. First, the performance and robustness of the consensus mechanism are analyzed using a comparative approach, considering the experimental values for ARI, NMI, Homogeneity, V-Measure, and Silhouette Score across different embeddings. Thus, given the superior performance metrics achieved, the *BERTimbau* model performance is further examined through graphical visualizations and word clouds to unveil and interpret how the model distinguishes between deceptive and legitimate content in an entirely unsupervised manner.

4.1. Comparative Analysis of Text Representations

Based on the experiments conducted and detailed in Sections 3.5 and 3.6, performance results were obtained for each text representation across its respective dimensionality settings and clustering algorithms. The findings are organized in Table 2, where the best results are highlighted in bold.

Table 2. Complete performance comparison of the Consensus mechanism across all embeddings and dimensions.

Repres.	Dim.	k	ARI	NMI	Homogeneity	V-Measure	Sil.
TF-IDF	26,500d	2	0.0029	0.0022	0.0022	0.0022	0.3453
Doc2Vec	50d	2	0.0029	0.0022	0.0022	0.0022	0.3453
Doc2Vec	100d	2	0.0014	0.0011	0.0011	0.0011	0.3108
Doc2Vec	200d	2	0.0249	0.0198	0.0193	0.0198	0.2935
FastText	50d	2	0.0001	0.0002	0.0002	0.0002	0.4253
FastText	100d	2	0.0009	0.0008	0.0008	0.0008	0.4338
FastText	200d	2	0.0021	0.0016	0.0016	0.0016	**0.4375**
BERTimbau	**768d**	**3**	**0.3534**	**0.3446**	**0.4446**	**0.3446**	0.3932

The experimental data indicate comparatively stronger performance for contextual embeddings. The representative bidirectional model, *BERTimbau*, obtained the highest performance in the experiments with $k = 3$, reaching an ARI of 0.3534 and an NMI of 0.3446. Despite its lower spatial density, as reflected in a Silhouette Score of 0.3932, it was the only model in this study that aligned clusters with ground truth labels to a notable degree, suggesting its potential to capture complex linguistic patterns, semantic subtleties, and ambiguities present in Brazilian Portuguese.

Among the results obtained, it is noteworthy that *BERTimbau* was the only text representation model in this study to support a three-cluster ($k = 3$) configuration, as shown in Table 2. While traditional static representations converged to a binary ($k = 2$) setup with lower external metrics, the contextual embeddings generated by *BERTimbau*

enabled the ensemble to identify a more nuanced subdivision of the corpus, which is further examined in Section 4.2.

This outcome indicates that high-dimensional latent features captured by *Transformer* architectures may facilitate the identification of subtle structural attributes present in fake news that simpler models might not capture.

In contrast to these contextual results, the traditional static models performed poorly. The external *TF-IDF* results suggest that term frequency is highly similar across the corpus, making it insufficient to distinguish between legitimate and deceptive content based on keyword recurrence alone. Alternatively, *Doc2Vec* was evaluated across three different dimensions; it demonstrated a noticeable improvement as dimensionality increased, with its NMI rising from 0.0022 at 50 dimensions to 0.0249 at 200 dimensions. This trend indicates that, while still underperforming contextual models, *Doc2Vec* benefits significantly from higher-dimensional vector representations.

In particular, *FastText* configured for 50 dimensions displayed the least satisfactory external results, with an ARI of 0.0001. Conversely, the 200-dimensional configuration achieved the highest Silhouette Score of 0.4375. However, its near-zero ARI suggests that although the clusters are spatially cohesive, *FastText* fails to capture the actual stylistic features that characterize misleading information.

4.2. Linguistic Patterns and Structural Analysis of the *BERTimbau* Ensemble

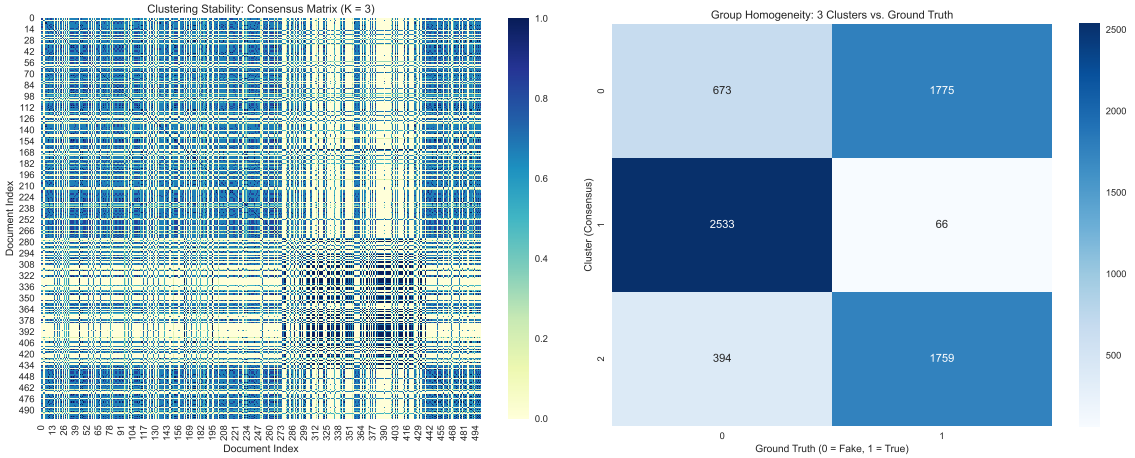
The comparatively stronger performance of the transformer architecture is also reflected in the individual algorithm results obtained using *BERTimbau*, as detailed in Table 3, with the best metrics highlighted in bold. Notably, the Ward clustering algorithm achieved the highest external metrics (ARI of 0.4167, NMI of 0.3785, and V-Measure of 0.3785), which may be attributable to its approach of minimizing intra-cluster variance and its compatibility with dense, high-dimensional model embeddings.

Table 3. Clustering performance results using BERTimbau embeddings ($k = 3$) after UMAP reduction.

Algorithm	ARI	NMI	Homogeneity	V-Measure	Sil.
K-Means	0.3573	0.3461	0.4465	0.3461	0.3935
Ward	**0.4167**	**0.3785**	0.4836	**0.3785**	0.3633
HDBSCAN	0.0405	0.1535	0.3218	0.1535	-0.1189
GMM	0.0961	0.2242	**0.5423**	0.2242	0.2419
K-Medoids	0.3579	0.3499	0.4515	0.3499	**0.3938**
Fuzzy C-Means	0.3417	0.3296	0.4255	0.3296	0.3908
Consensus	0.3534	0.3446	0.4446	0.3446	0.3932

Although the Consensus clustering performance (ARI of 0.3534) was not the highest absolute result, it appeared to be the most robust outcome in this experimental context. The consensus mechanism balanced predictions across all the models tested, reducing the influence of individual algorithmic variability—such as the lower performance of *HDBSCAN*—and producing stable clusters, as illustrated by the blue blocks in the main diagonal of Figure 2a. This observation suggests that, for $k = 3$, the different clustering

algorithms generally agreed on the partitioning of most data points. The observed purity of these segmented groups is further supported by the contingency matrix in Figure 2b, which compares the consensus clusters with the binary ground truth.



(a) Clustering stability: consensus matrix **(b) Group homogeneity vs. Ground Truth**

Figure 2. Validation of the BERTimbau Ensemble ($k = 3$): (a) consensus stability and (b) cluster purity against the Fake.br Corpus labels.

As discussed in Section 4.1, the *BERTimbau* ensemble identified a tripartite structure within the corpus. While the ground-truth labels from the *Fake.br Corpus* are binary, the model separated legitimate news into two clusters, suggesting the presence of distinct writing styles or themes within the authentic news in the dataset that are sufficiently different to be separated through unsupervised clustering.



(a) Cluster 0 (True): 72.51% purity **(b) Cluster 2 (True): 81.70% purity**



(c) Cluster 1 (Fake): 97.46% purity

Figure 3. WordCloud qualitative analysis of the three identified clusters ($k = 3$) using BERTimbau embeddings.

Thus, the legitimate texts were segmented into two apparent thematic clusters. Cluster 0, comprising 2,488 documents, achieved a purity of 72.51% and focused on social and international narratives (e.g., 'Coreia Norte', 'Estados Unidos', 'Trump'), as illustrated in Figure 3a. Meanwhile, Cluster 2 grouped 2,153 articles with an 81.70% purity (Figure 3b) and was characterized by institutional and legal Brazilian discourse (e.g., 'Tribunal Federal', 'Ministerio Publico', 'Lava Jato'). Although political in nature, its vocabulary consists of more legal jargon than the misleading texts.

In contrast, Cluster 1 consisted predominantly of deceptive content (97.46%), containing 2,599 instances, of which 2,533 were labeled as fake news. As visualized in Figure 3c, this group was mainly associated with proper nouns and political rhetoric. This observation may reflect trends in real-world disinformation patterns, where false narratives often reference political attacks and authority figures.

5. Conclusion

In this study, we investigated the problem of disinformation detection in Brazilian Portuguese using an entirely unsupervised approach, comparing the performance of traditional static feature extraction with that of dynamic contextual embeddings. The proposed methodological approach, guided by the KDD process, integrated dimensionality reduction and an ensemble clustering mechanism. The *BERTimbau* results indicated comparatively stronger performance for bidirectional *Transformer* architectures relative to traditional and statistical baselines. These findings suggest that unsupervised frameworks incorporating dense contextual embeddings may facilitate the identification of structural and argumentative characteristics associated with disinformation, even without prior labeling.

For future research, it will be important to assess the framework's adaptability to contemporary and rapidly evolving disinformation narratives, as the present experimental evaluation was limited to a single benchmark dataset covering news from 2016 to 2018. Therefore, the detected linguistic patterns—in particular those associated with proper nouns of political figures from the period—reflect the specific temporal context of the corpus. These results may not directly generalize to the disinformation landscape of 2026 without a revision of the underlying corpus or the re-training of embeddings to capture current semantic shifts.

Additionally, we plan to extend the proposed unsupervised pipeline to multi-domain corpora and real-time streaming environments. Given the comparatively stronger performance of *BERTimbau* in this study, subsequent investigations could benchmark this architecture against other state-of-the-art contextual models pre-trained on Portuguese, such as *Albertina* or *RoBERTa* variants, to further explore approaches for automated identification of deceptive patterns.

Generative AI Usage Declaration

The authors declare that the generative AI model Gemini 3.1 Pro was utilized as a supportive tool in the preparation of this manuscript, exclusively for English language editing and grammatical revision. The authors hold full responsibility for the final content, including plagiarism verification, data accuracy, and all implemented corrections.

References

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, NJ.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146. DOI: https://doi.org/10.1162/tacl_a_00051.

Colomina, C., Margalef, H. S., Youngs, R., and Jones, K. (2021). The impact of disinformation on democratic processes and human rights in the world. Study, European Parliament. DOI: https://doi.org/10.2861/59161.

Das, K., Kamlish, and Abid, F. (2024). Advancements in Word Embeddings: A Comprehensive Survey and Analysis. *Proceedings of the Pakistan Academy of Sciences: A. Physical and Computational Sciences*, 61(3):227–245. DOI: https://doi.org/10.53560/PPASA(61-3)842.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3):37–54. DOI: https://doi.org/10.1609/aimag.v17i3.1230.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Waltham, MA, 3rd edition. DOI: https://doi.org/10.1016/C2009-0-61819-5.

Jurafsky, D. and Martin, J. H. (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. Online manuscript released January 6, 2026.

Le, Q. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196. PMLR. Available at: https://proceedings.mlr.press/v32/le14.html. Accessed: April 21, 2026.

McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. DOI: https://doi.org/10.48550/arXiv.1802.03426.

Monteiro, R. A., Santos, R. L. S., Pardo, T. A. S., de Almeida, T. A., Ruiz, E. E. S., and Vale, O. A. (2018). Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results. In *Computational Processing of the Portuguese Language*, pages 324–334, Cham. Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-99722-3_33.

Newman, N., Ross Arguedas, A., Robertson, C. T., Nielsen, R. K., and Fletcher, R. (2025). *Digital News Report 2025*. Reuters Institute for the Study of Journalism. Available at: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2025. Accessed: April 17, 2026.

Paixão, M., Lima, R., and Espinasse, B. (2020). Fake News Classification and Topic Modeling in Brazilian Portuguese. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 427–432. DOI: https://doi.org/10.1109/WIIAT50758.2020.00063.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/D19-1410.

Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20:53–65. DOI: https://doi.org/10.1016/0377-0427(87)90125-7.

Shah, A., Shah, H., Bafna, V., Khandor, C., and Nair, S. (2025). Validation and Extraction of Reliable Information Through Automated Scraping and Natural Language Inference. *Engineering Applications of Artificial Intelligence*, 147:110284. DOI: https://doi.org/10.1016/j.engappai.2025.110284.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-61377-8_28.

Sparemberger, R. and da Silva, A. C. E. S. (2021). O IMPACTO DAS FAKE NEWS NO PROCESSO ELEITORAL BRASILEIRO. *Revista Reflexão e Crítica do Direito*, 9(2):251–277. Available at: https://revistas.unaerp.br/rcd/article/view/2438. Accessed: March 29, 2026.

Trevilato, L. A. (2025). Análise Comparativa de Modelos de Representação de Texto e Métodos de Aprendizado de Máquina na Classificação de Notícias Falsas em Português. Undergraduate thesis, Universidade Federal de Uberlândia. Available at: https://repositorio.ufu.br/handle/123456789/45951. Accessed: March 29, 2026.

Vieira, C., Souza, J., and Cavalcanti, G. (2025). Detecção de Fake News em Português: Análise Comparativa entre Métodos de Representação em Português, Inglês e Multilíngues. In *Anais do XIV Brazilian Workshop on Social Network Analysis and Mining*, pages 187–199, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.5753/brasnam.2025.9062.