

Separating Human from Machine: Deep Linguistic Analysis and Automatic Classification of Fake News in Brazilian Portuguese

Pedro Lucas Castro de Andrade^[0009–0001–3911–1870],
Renato Moraes Silva^[0000–0001–6687–8981], and
Thiago Alexandre Salgueiro Pardo^[0000–0003–2111–1319]

Interinstitutional Center for Computational Linguistics (NILC),
Institute of Mathematics and Computer Science, University of São Paulo,
São Carlos, SP, Brazil
`pedroandrade@usp.br`, `{renatoms,taspardo}@icmc.usp.br`

Abstract. Fake news are a growing challenge for society, further intensified by the proliferation of synthetic content generated by Large Language Models (LLMs). In underrepresented languages such as Portuguese, the distinction between human-written and synthetic fake news remains insufficiently explored, particularly for Brazilian Portuguese, hindering accountability on the issue. This work contributes to this research line by performing a deep characterization of two Brazilian Portuguese corpora containing both human-written and LLM-generated fake news, exploring varied linguistic features (building over previous studies). We then evaluate statistical and neural classifiers, including multilingual and Portuguese-specialized models, showing that the texts contain discriminative features and that neural approaches achieve F1 scores above 99% for the task of identifying human-written and synthetic fake news.

Keywords: Natural Language Processing · Large Language Models · Fake News Authorship · Brazilian Portuguese

1 Introduction

The dissemination of fake news is one of the most pressing challenges for society, with significant implications for democratic processes, public health, and social cohesion. This challenge has been dramatically amplified by the Large Language Models (LLMs), which facilitates the generation of false content at scale with professional writing quality that closely mimics human-authored texts [1, 2]. While considerable progress has been made in both fake news detection and machine-generated text identification, these two research areas have historically evolved as distinct lines of inquiry, with limited investigation of their intersection.

The proliferation of generative AI has introduced a new dimension to the fake news problem: the need to distinguish between human-authored and machine-generated false content. This distinction has become increasingly important for several reasons. First, understanding the source of false information is crucial

for developing effective accountability mechanisms and tracking disinformation campaigns [3]. Second, different detection strategies may be required for each type of content, as human-written and machine-generated texts exhibit distinct linguistic characteristics that can be exploited by classification systems [4, 5]. Third, the ability to identify machine-generated fake news enables more targeted interventions, such as focusing moderation efforts on automated disinformation networks rather than individual human actors.

For underrepresented languages, this problem becomes more critical. For instance, Brazilian Portuguese, with over 200 million speakers, has received considerably less attention in Natural Language Processing (NLP) research compared to languages such as English. This disparity is particularly evident in the context of fake news and machine-generated text detection, where available resources and datasets remain scarce [6]. There are also important cultural and linguistic factors that make language-specific investigations essential: fake news narratives, stylistic conventions, and discourse patterns are deeply embedded in cultural contexts and may not transfer directly from one language to another [6].

Silva et al. [7] recently took a step toward changing this scenario. They created extended versions of two well-known Brazilian Portuguese fake news corpora, Fake.br [8,9] and FakeTrueBR [10], by generating synthetic fake news with LLMs. To the best of our knowledge, these are the only available paired corpora for Brazilian Portuguese containing both human-written and LLM-generated fake news. Following their release, Andrade et al. [11] conducted the lexical and syntactic characterization of these corpora. While their study reveals differences between human and machine-generated fake news, it is limited to basic metrics and does not explore deeper psycholinguistic or semantic-discursive properties. Furthermore, they do not evaluate whether these features can actually be used to automatically classify the source of the text as human-written or synthetic.

Our study addresses this gap by providing a comprehensive characterization across three linguistic levels: lexical, syntactic, and semantic-discursive. Specifically, we employ LIWC (Linguistic Inquiry and Word Count) [12,13] to examine psychological and cognitive processes reflected in the texts, and NILC-Metrix [14] to extract semantic-discursive features related to cohesion, coherence, and textual complexity. Additionally, we analyze enhanced dependency relations to capture specific patterns characteristic of each group. Beyond characterization, we evaluate the discriminative potential of these linguistic features by implementing and assessing traditional machine learning classifiers, addressing a fundamental question: to what extent can linguistic attributes alone distinguish between human-written and machine-generated fake news in Brazilian Portuguese?

Answering this question is crucial for two main reasons. First, it could provide clear, linguistically based explanations for why a text is flagged as machine-generated, rather than relying only on “black-box” deep learning models. This transparency could provide clear and verifiable evidence to support investigations into the creation and dissemination of AI-generated fake news. Second, if proven effective, these attributes could enable the development of lighter and cheaper

detection systems that look for stable writing styles instead of using expensive and heavy neural networks.

Our main contributions are threefold: (1) the first deep linguistic profile of machine-generated Portuguese fake news, uncovering psychological, cognitive, and discursive patterns that distinguish them from human-written texts; (2) a robust classification benchmark that systematically evaluates how these linguistic features perform across traditional and neural models, proving their high discriminative potential; and (3) an integrated profile of complementary linguistic levels, offering comprehensive insights for both theoretical understanding and practical application in fake news detection systems.

The remainder of this paper is organized as follows: Section 2 briefly reviews the main related work; Section 3 describes our methodology; Section 4 presents and discusses our results; and Section 5 brings final remarks.

2 Related Work

Fake news detection has been addressed through various computational approaches, including linguistic analysis [15, 16], source credibility assessment [17], and social network context [18]. For Brazilian Portuguese, two publicly available corpora are Fake.br [8, 9] and FakeTrueBR [10], though existing work has focused primarily on distinguishing human-written real from fake news, leaving machine-generated content largely unexplored until recently [6].

Some studies have characterized linguistic differences between human and AI-generated texts without considering fake news. Sardinha [5] found such differences through multidimensional analysis in Portuguese, while Zhou et al. [4] observed that AI texts are more formal and analytical, and Muñoz-Ortiz et al. [19] reported greater syntactic variety in human texts. These studies have mainly focused on high-resource languages and general domains.

Recent initiatives have addressed the intersection of fake news and machine-generated content in Portuguese. Silva et al. [7] extended both corpora by generating synthetic fake news using Sabiá-3 [20, 21], a Portuguese-specialized LLM. Andrade et al. [11] then provided a lexical and syntactic characterization of these corpora, revealing that machine texts used longer words and sentences, favored formal expressions, and showed lower stylistic variability than their human-written counterparts.

Rodrigues et al. [22] compared human- and AI-generated content in European Portuguese (GPT-4o, Mistral Large, and Llama 3.3 70B), building a dataset upon the fake-news-pt-eu corpus [23] and finding that LLM texts are more formal, positive, and structured while human texts show greater variability and informal markers. They also show that a fake news detector trained on human content degrades significantly when applied to LLM-generated misinformation. Unlike their work, which covers both factual and fake news in European Portuguese without paired corpora or dedicated classifiers, we focus exclusively on Brazilian Portuguese fake news using paired corpora and a systematic classification benchmark.

While these analyses established baseline differences at lexical and syntactic levels, semantic-discursive characteristics, including psycholinguistic dimensions, text cohesion and coherence, remained unexplored. Additionally, the discriminative potential of linguistic features for classification has not been systematically evaluated. The present work addresses these gaps.

3 Methodology

3.1 Corpora

This work uses a unified dataset composed of fake news articles from the Fake.br and FakeTrueBR corpora, together with machine-generated counterparts created by Silva et al. [7] using Sabiá-3 [20, 21], totaling 10,782 documents. The dataset comprises 5,391 human-written fake news articles (3,600 from Fake.br and 1,791 from FakeTrueBR) paired with 5,391 machine-generated versions sharing the same topics, enabling direct stylistic comparison. To the best of our knowledge, these are the only available corpora containing paired human-written and LLM-generated fake news for Brazilian Portuguese.

Preprocessing included basic character and whitespace normalization and line break correction. All linguistic characterization analyses were performed on the full, untruncated texts.

3.2 Linguistic Characterization Framework

Lexical and Syntactic Features: Building upon the lexical and syntactic characterization of Andrade et al. [11], which analyzed vocabulary patterns through SAGE modeling [24], syllable complexity, and the distribution of part-of-speech (POS) tags and dependency relations (DEPREL) following Universal Dependencies (UD) [25], this work extends the analysis to deeper linguistic levels, as follows.

Enhanced Syntactic Rules: To capture syntactic-semantic patterns, we extended UD annotations with Enhanced UD (EUD) [26, 27]. Texts were automatically parsed and enriched with EUD (using the state of the art systems of [28] and [29]), capturing phenomena like subject/object propagation and lexicalized prepositional relations. We extracted production rules in the form used in [11], as the following template illustrates (actual examples will be presented later): `POS_of_the_head(POS_of_the_1st_dep/DEPREL, *, POS_of_the_2nd_dep/DEPREL, etc.)`, where `*` marks the head position and dependents are ordered linearly, yielding 32,819 unique rule types. Rules were vectorized via TF-IDF (Term Frequency-Inverse Document Frequency) and analyzed using Mann-Whitney U tests and Cohen’s *d* to identify discriminative patterns.

Psycholinguistic Analysis with LIWC: We employed LIWC [12] lexicon in its Brazilian Portuguese adaptation [13] to analyze psycholinguistic patterns. This resource contains 11,823 exact-match words and 2,634 prefix entries (marked with *) distributed across 73 categories, which span multiple dimensions: linguistic processes (function words, pronouns, verbs, adjectives), psychological processes (affective, social, cognitive, perceptual), temporal focus (past, present, future), personal concerns (work, leisure, home), and informal language markers. For each document, we computed the percentage of tokens assigned to each category following standard LIWC scoring. Statistical comparison employed Mann-Whitney U tests (non-parametric, given non-Gaussian distributions) with Cohen’s d effect sizes, identifying categories with medium-to-large effects ($|d| \geq 0.5$) as particularly discriminative.

Semantic-Discursive Analysis with NILC-Metrix: We employed NILC-Metrix [14], a Brazilian Portuguese text analysis tool, extracting metrics covering readability, lexical diversity, cohesion, and psycholinguistic properties. We selected 56 semantic-discursive metrics across 10 categories (descriptive complexity, lexical diversity, readability, psycholinguistics, referential cohesion, semantic cohesion via LSA, connectives, and lexical semantics), avoiding redundancy with lexical-syntactic analyses [11]. Statistical analysis employed non-parametric methods: Mann-Whitney U tests and Cliff’s δ as effect size. We defined “strong discriminative signals” as metrics achieving $|\delta| \geq 0.330$ and Benjamini-Hochberg FDR-corrected $q < 0.05$, yielding 56 metrics meeting both criteria.

While some overlap exists across dimensions, for instance, lexical information appears in both syntactic and semantic-discursive analyses, and perceptual verbs (LIWC) share conceptual ground with sensorial vocabulary (NILC-Metrix), these dimensions are treated as intentionally complementary, capturing the same phenomena from distinct theoretical perspectives.

3.3 Classification Experiments

We constructed a comprehensive feature set integrating attributes from all linguistic levels: (i) NILC-Metrix metrics (semantic-discursive), (ii) EUD rules as TF-IDF vectors (syntactic-semantic), and (iii) lexical features combining LIWC categories, part-of-speech ratios, lexical frequency, and syllable complexity measures. Features were merged by document identifier, with paired human-machine texts maintaining alignment. The EUD rules, initially represented as strings, were vectorized via TF-IDF with L2 normalization to account for length differences.

To eliminate length bias, we applied paired truncation before classification only: both texts in each pair were truncated to the length of the shorter one. While truncation may influence features dependent on full discourse structure, its impact is mitigated for the majority of features employed, which are normalized by text length; non-normalized features were excluded via the length-bias filtering criterion applied in Condition 3 (presented below).

We evaluated three experimental conditions to assess the discriminative potential of linguistic features:

- *Condition 1: Neural Models.* BERTimbau (Portuguese-specific) [30] and Multilingual BERT [31] were tested with and without fine-tuning for binary classification. This condition comprises direct classification over the [CLS] embeddings, without using the explicit linguistic features that were previously extracted.
- *Condition 2: Statistical Classifiers without Linguistic Features.* Bag-of-words (BoW) and TF-IDF representations combined with six classifiers: Support Vector Machines, Random Forest, Gaussian Naive Bayes, Logistic Regression, Multi-Layer Perceptron, and XGBoost. This establishes lexical baselines without explicit linguistic annotation.
- *Condition 3: Statistical Classifiers with Linguistic Features.* TF-IDF and BoW representations concatenated with a sanitized set of 780 linguistic features, trained with the same classifiers. Feature sanitization applied a filtering step to remove length-biased metrics, removing a total of 90 absolute-count features and metrics with documented length dependencies. This filtering step ensures that classifiers learn genuine stylistic patterns rather than residual length correlations.

Comparing Conditions 2 and 3 quantifies the contribution of linguistic features over raw lexical distributions. Comparing Conditions 1 and 3 measures the gap between contextual neural embeddings and interpretable feature-based approaches.

Experimental Setup: Following Silva et al. [7], all experiments employed a holdout validation strategy with an 80% training set and a 20% testing set, shared across all conditions to ensure comparability, while paired documents were never split across sets to prevent topic leakage. To eliminate length bias as a discriminative factor, we applied paired truncation before classification: for each document pair (human and LLM versions of the same article), both texts were truncated to the length of the shorter document. This strategy prevents classifiers from learning to distinguish classes based on document length rather than linguistic patterns, as LLM-generated texts tend to be systematically shorter than their human counterparts. Hyperparameter optimization used GridSearchCV with 5-fold stratified cross-validation on the training set, maximizing the weighted F1 score.

4 Results

4.1 Linguistic Characterization Results

Enhanced Syntactic Rules Analysis: EUD analysis identified 32,819 unique production rules. Table 1 presents the top five most discriminative rules for each group, with illustrative examples shown in Figure 1.

Table 1. Top discriminative EUD production rules per group.

Rule Pattern	Diff	<i>p</i>	<i>d</i>
<i>Higher in Human Texts</i>			
NOUN(*, NOUN/nmod:de)	+0.0111	<0.001	0.544
NOUN(*, ADV/nmod:de)	+0.0090	<0.001	0.270
NOUN(*, PRON/nmod:de)	+0.0076	<0.001	0.205
VERB(PRON/nsubj, *, SYM/obl:a)	+0.0065	<0.001	0.254
VERB(*, VERB/conj:e, VERB/conj:e)	+0.0053	<0.001	0.188
<i>Higher in Machine Texts</i>			
VERB(*, VERB/conj:mas)	-0.0221	<0.001	0.632
NOUN(*, PROPN/nmod:em)	-0.0196	<0.001	0.546
NOUN(*, NOUN/nmod:sobre)	-0.0193	<0.001	0.534
ADJ(*, ADJ/conj:e)	-0.0177	<0.001	0.512
NOUN(*, NOUN/nmod:de, NOUN/conj:e)	-0.0169	<0.001	0.484

Note: Portuguese words appearing as dependency label lexicalizations: prepositions *de* (of/from), *em* (in/at), *a* (to/at), *sobre* (about); conjunctions *e* (and), *mas* (but).

Examples of sentences that use the discriminative rules (corresponding passages in bold):

Human – NOUN(*, NOUN/nmod:de):

“Do jeito que ele falou, parece que o suposto **medo de venda** aos chineses é pretexto para evitar a privatização.”

Machine – VERB(*, VERB/conj:mas):

“Os advogados de Frederico Pacheco e Mendherson Lima também expressaram satisfação, enfatizando que a justiça **tarda, mas não falha.**”

Fig. 1. Illustrative examples of sentences containing the most discriminative EUD rules for each group.

Human-written texts favor constructions of the form *de*-nominal modification and pronominal subjects. The pattern of triple verb coordination (*e*-coordination) suggests informal, additive listing style. Machine-generated texts show systematic preference for adversative coordination (*mas*, $d = 0.632$), indicating argumentative structure with explicit contrast marking. Higher rates of adjective and noun coordination (*e*) suggests formal, information-dense nominal style.

Psycholinguistic Patterns (LIWC Results): LIWC analysis revealed systematic psycholinguistic differences. Human texts exhibited stronger social and temporal grounding, with the most discriminative categories being *past focus* and *informal language* (both $d = 0.595$), *hear* (perceptual verbs, $d = 0.552$), *social processes* ($d = 0.517$), and *male references* ($d = 0.456$). Machine texts showed elevated cognitive and affective processing, with highest discriminative

power in *positive emotions* ($d = 0.472$), *affect* ($d = 0.460$), *anxiety* ($d = 0.449$), *insight* (reasoning verbs, $d = 0.435$), and *achievement* ($d = 0.342$).

Semantic-Discursive Features (NILC-Matrix Results): From 56 semantic-discursive metrics with strong discriminative power ($|\delta| \geq 0.330$), human texts revealed accessible, sensorial vocabulary with the most discriminative features being *punctuation diversity* (ratio of distinct punctuation marks used in a text, $\delta = +0.725$), *age of acquisition: early-acquisition ratio* (proportion of words learned by children before age 2.5 years, $\delta = +0.667$), *imageability 4-5.5 range* (proportion of words rated 4-5.5 on a 1-7 scale of mental imagery evocability, $\delta = +0.657$), *imageability mean* (average imageability score of all words in text, $\delta = +0.626$), and *concreteness mean* (average concreteness rating of all words, concrete vs. abstract, $\delta = +0.583$). These patterns indicate that human texts employ concrete, imageable vocabulary acquired early in childhood, with greater punctuation variety reflecting colloquial style. Machine texts showed sophisticated, abstract vocabulary with highest discriminative power in *Honoré statistic*¹ (lexical richness, $\delta = -0.798$), *age of acquisition mean* (average age at which words were learned by children, late-acquired words, $\delta = -0.747$), *LSA span mean* (Latent Semantic Analysis: average semantic overlap between adjacent sentences smooth semantic progression, $\delta = -0.699$), *age of acquisition: late-acquisition ratio* (proportion of words learned by children after age 5.5 years, $\delta = -0.686$), and *LSA givenness mean* (LSA-based measure of information status: proportion of given content between sentences, $\delta = -0.685$). These reflect formal, abstract style with continuous, uniform semantic progression where sentences appear to share substantial overlap. Higher positive words and systematic use of connectives suggest formal argumentative structure.

Integrated Linguistic Profile: Integrating findings across all linguistic levels suggests distinct profiles: human texts tend to employ accessible, concrete vocabulary with informal markers (netspeak, fillers) and irregular semantic transitions reflecting thematic variability. Machine texts tend to employ more abstract, sophisticated vocabulary with formal structures (adversative coordination, complex genitive chains) and smooth semantic progression with continuous overlap. The cross-dimensional consistency — accessible versus sophisticated (lexical), informal versus formal (psycholinguistic), simple versus complex (syntactic), variable versus smooth (discursive) — shows the discriminative potential of interpretable linguistic features.

4.2 Classification Results

We evaluated the three experimental conditions for binary authorship attribution (human vs LLM-generated fake news) listed before. The results are synthesized in what follows.

¹ Lexical richness measure: $H = \frac{(\log N)^2}{\log(N) - \log V}$, where N =total words, V =vocabulary size.

Neural Baselines (Condition 1): BERTimbau fine-tuned for direct binary classification achieved the strongest result overall with F1 score of 99.95% (1 error in 2,158 samples), while BERT Multilingual achieved 99.77% (5 errors). Fine-tuned embeddings with classifiers yielded comparable performance: BERTimbau combined with SVM/LR/MLP/XGB at 99.81% (4 errors) and Multilingual combined with Random Forest at 99.81% (4 errors). Embeddings without fine-tuning showed larger gaps, with BERTimbau combined with SVM at 99.07% (20 errors) versus Multilingual combined with MLP at 97.22% (60 errors). The consistent superiority of BERTimbau over Multilingual BERT reinforces the importance of language-specific pretraining for Portuguese.

Statistical Baselines (Condition 2): TF-IDF representation combined with multilayer perceptron achieved a 99.35% F1 score (14 errors), outperforming Support Vector Machines (99.30%, 15 errors) and Logistic Regression (99.21%, 17 errors) with TF-IDF. BoW representation yielded lower but still robust performance: MLP at 99.17% (18 errors), SVM at 98.75% (27 errors), and Logistic Regression at 98.70% (28 errors). The consistently high performance (F1 > 98%) demonstrates that systematic vocabulary differences between human and LLM texts are detectable through distributional patterns alone, without explicit linguistic annotation.

Linguistic Features (Condition 3): Classification concatenating lexical representations (TF-IDF or BoW) with interpretable linguistic features showed robust performance across all classifiers. XGBoost achieved the best result on both pipelines with a F1 score of 99.86% (3 errors), followed by MLP with BoW at 99.49% (11 errors), SVM and Logistic Regression with BoW at 99.56% (12 errors each), and Random Forest at 98.52% (32 errors).

A notable finding is that BoW outperforms TF-IDF for linear classifiers and MLP when linguistic features are present, while XGBoost performs identically on both. This suggests that TF-IDF’s log-compression of term frequencies suppresses discriminative signal that linear classifiers exploit more effectively from raw counts when stabilized by explicit linguistic features. Tree-based models are invariant to this transformation since they split on thresholds rather than magnitudes.

XGBoost feature importance (gain-based) reveals the most discriminative signals. The dominant feature is `nilc_indicative_present_ratio` (proportion of verbs conjugated in the present indicative tense relative to all verbs in the text) with 25.4% of total importance — a linguistically interpretable finding: LLMs tend to use present indicative more uniformly (a “timeless” generative style). Structural features follow: `sentences_per_paragraph` (average number of sentences per paragraph, capturing paragraph density, 13.7%), `clauses_per_sentence` (average number of clauses per sentence, measuring syntactic complexity 13.7%), `named_entity_ratio_text` (ratio of named entity tokens to total tokens in the document, 12.6%), and `adjunct_per_clause` (average number

of adjunct modifiers per clause, 6.6%). One Enhanced UD rule also appear: `PROPN(ADP/case, DET/det, *)`².

Across all classifiers and both vectorizers, four features consistently appear in the top-15: `nilc_sentences_per_paragraph`, `syll_media_silabas_por_palavra` (average number of syllables per word, a proxy for lexical complexity), `ud_VERB(PUNCT/punct, *, NOUN/obj)` (syntactic rule where a verb governs a punctuation-delimited inciso and a nominal object, characteristic of parenthetical constructions), and `nilc_syllables_per_content_word` (average syllables per content word (nouns, verbs, adjectives, adverbs), excluding function words). These represent the most stable discriminative signals across methodologically diverse classifiers.

4.3 Discussion

Systematic differences across EUD, LIWC, and NILC-Metrix reveal that linguistic features are relevant discriminators for text origin. Moreover, the convergence of high performance across distinct classification methodologies indicates that detection in Brazilian Portuguese is currently tractable. However, some caution is needed when generalizing these conclusions, as our analysis examined only one LLM (Sabiá-3) using a single generation protocol (rewriting human-authored articles) over only two corpora (the only ones available with the required information, to the best of our knowledge). These observations may guide future work on the task, as presented in the next section.

5 Conclusion

This work investigated the linguistic distinction between human-written and LLM-generated fake news in Brazilian Portuguese through comprehensive psycholinguistic, semantic-discursive, and syntactic analyses. Our main contributions are: systematic characterization across EUD (syntactic-semantic), LIWC (psycholinguistic), and NILC-Metrix (semantic-discursive) dimensions revealing coherent human-machine differences; and evaluation of three distinct classification approaches (neural, statistical, linguistic) demonstrating that discriminative patterns are robust across methodologically diverse approaches.

It is worth noting that the discriminative patterns identified in this work reflect the characteristics of current LLM outputs generated under a simple rewriting protocol, without any prompt engineering aimed at mimicking human writing style. As generative models evolve and more sophisticated prompting strategies are employed, these linguistic signals may become less reliable. This represents an inherent limitation of the field rather than of the specific methodology adopted here: detection systems will likely require continuous adaptation as generation techniques advance [1].

² Proper noun preceded by a case-marking preposition and determiner, e.g., *do Brasil, da Universidade*.

Future work includes (i) evaluating multiple LLMs to establish whether the observed patterns are LLM-specific or represent generalizable characteristics of synthetic text, and (ii) generating synthetic fake news using zero-shot prompting without reliance on real articles as source material, which would assess detection robustness against purely synthetic content generation rather than text rewriting.

Acknowledgments

This work was supported by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES), the São Paulo Research Foundation (FAPESP; grant no. 2024/17834-6) and the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq; grant no. 444933/2024-7). We also acknowledge the support of the Ministry of Science, Technology and Innovation, with resources of Law N. 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

References

1. Jinyan Su, Claire Cardie, and Preslav Nakov. Adapting fake news detection to the era of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1473–1490, Mexico City, Mexico, 2024. Association for Computational Linguistics.
2. Canyu Chen and Kai Shu. Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Magazine*, 45(3):354–368, 2024.
3. Jinyan Su, Terry Yue Zhuo, Javid Mansurov, Di Wang, and Preslav Nakov. Fake news detectors are biased against texts generated by large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10276–10288. Association for Computational Linguistics, 2023.
4. Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G. Parker, and Munmun De Choudhury. Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI ’23, pages 1–20, New York, NY, USA, 2023. ACM.
5. Tony Berber Sardinha. AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, 4(1):100083, 2024.
6. Renato Moraes Silva, Roney Lira de Sales Santos, and Thiago Alexandre Salgueiro Pardo. Detecção automática de notícias falsas. In Helena de Medeiros Caseli and Maria das Graças Volpe Nunes, editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 27, pages 635–653. BPLN, 3 edition, 2024.
7. Renato Moraes Silva, Hazem Amamou, Lucca Baptista Silva Ferraz, Fabio Kauê Araujo da Silva, and Anderson Raymundo Avila. Fake news detection in Portuguese under large language model-generated content. *Journal of the Brazilian Computer Society*, 31(1):1149–1166, 2025.
8. Rafael A. Monteiro, Roney L. S. Santos, Thiago A. S. Pardo, Tiago A. de Almeida, Evandro E. S. Ruiz, and Oto A. Vale. Contributions to the study of fake news

- in Portuguese: New corpus and automatic detection results. In *Computational Processing of the Portuguese Language*, volume 11122 of *Lecture Notes in Computer Science*, pages 324–334, Cham, 2018. Springer.
9. Renato M. Silva, Roney L. S. Santos, Tiago A. Almeida, and Thiago A. S. Pardo. Towards automatically filtering fake news in Portuguese. *Expert Systems with Applications*, 146:113199, 2020.
 10. Juan Pablo Chavarro, Jonata Tyska Carvalho, Tarlis Tortelli Portela, and Jonathan Cardoso Silva. FakeTrueBR: Um corpus brasileiro de notícias falsas. In *Anais da XVIII Escola Regional de Banco de Dados*, pages 108–117, Porto Alegre, 2023. SBC.
 11. Pedro Lucas Castro de Andrade, Renato Moraes Silva, and Thiago Alexandre Salgueiro Pardo. Caracterização lexical e sintática de notícias falsas em português produzidas por humanos e por máquinas. In Marlo Souza, Iria de Dios-Flores, Diana Santos, Larissa Freitas, Jackson Wilke da Cruz Souza, and Eugénio Ribeiro, editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 2*, pages 148–158, Salvador, Brazil, April 2026. Association for Computational Linguistics.
 12. James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of LIWC2015. *Technical Report*, 2015.
 13. Flavio Carvalho, Fabio Paschoal Junior, Eduardo Ogasawara, Lilian Ferrari, and Gustavo Guedes. Evaluation of the Brazilian Portuguese version of linguistic inquiry and word count 2015 (BP-LIWC2015). *Language Resources & Evaluation*, 58(1):203–222, 2024.
 14. Sidney Evaldo Leal, Magali Sanches Duran, and Sandra Maria Aluísio. NILC-Matrix: Assessing the complexity of written and spoken language in Brazilian Portuguese. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 668–677, Varna, Bulgaria, 2023.
 15. Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark, 2017. Association for Computational Linguistics.
 16. Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics.
 17. Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3364–3374. Association for Computational Linguistics, 2020.
 18. Fabrizio Barnabò, Federico Siciliano, Carlos Castillo, Stefano Leonardi, and Preslav Nakov. Deep learning for cross-lingual news stance detection. In *Proceedings of the ACM Web Conference 2022*, pages 1094–1103. ACM, 2022.
 19. Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review*, 57(10):265, 2024.
 20. Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. Sabiá: Portuguese large language models. In *Intelligent Systems*, volume 14197 of *Lecture Notes in Computer Science*, pages 226–240, Cham, 2023. Springer.

21. Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. Sabiá-3 technical report, 2024.
22. Flávia A. Rodrigues, Niclas F. Sturm, and Flávio L. Pinheiro. A linguistic comparison between human- and AI-generated content. *iScience*, 29:114976, 2026.
23. Ricardo Afonso and João Rosas. Development of a smartphone application and chrome extension to detect fake news in English and European Portuguese. *IEEE Latin America Transactions*, 22(4):294–303, 2024.
24. Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on Machine Learning*, ICML’11, pages 1041–1048, Madison, WI, USA, 2011. Omnipress.
25. Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal dependencies. *Computational Linguistics*, 47(2):255–308, 2021.
26. Magali Sanches Duran, Elvis A. de Souza, Maria das Graças Volpe Nunes, Adriana Silvina Pagano, and Thiago A. S. Pardo. Extending the enhanced Universal Dependencies – addressing subjects in pro-drop languages. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 143–152, Ljubljana, Slovenia, August 2025. Association for Computational Linguistics.
27. Adriana S. Pagano, Magali Sanches Duran, and Thiago Alexandre Salgueiro Pardo. Enhanced dependencies para o português brasileiro. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 453–462, Belo Horizonte, Brazil, 2023. SBC.
28. Lucelene Lopes and Thiago Alexandre Salgueiro Pardo. Towards portparser – a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024)*, pages 399–409, Santiago de Compostela, Galicia/Spain, 2024. Association for Computational Linguistics.
29. Elvis de Souza, Magali Duran, Maria das Graças Nunes, Gustavo Sampaio, Giovanna Belasco, and Thiago Pardo. Automatic annotation of enhanced universal dependencies for brazilian portuguese. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 217–226, Porto Alegre, RS, Brasil, 2024. SBC.
30. Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS*, pages 403–417, Rio Grande do Sul, Brazil, October 2020.
31. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics, 2019.