Interpretable Features for Distinguishing Machine Generated News Articles

Ravi Teja Karumuri[1] and Huan Liu[1]

Arizona State University, Tempe AZ 85281, USA,
rkarumur@asu.edu, huanliu@asu.edu

Abstract. Rapid progress in large language models (LLMs) has made it trivial to generate near-human-quality news, flooding social media with deceptive machine-written articles. Because LLMs are opaque and prone to hallucination, we propose two syntactic features [7]: Bag-of-Relations (BoR), a histogram of dependency labels per article, and a TF-IDF–style weighting called Relation-Frequency Inverse-Document-Frequency (RF-IDF), over those labels to distinguish human versus machine text. Using only an off-the-shelf dependency parser, we extract these vectors for two benchmarks (NeuralNews, 64 K articles; Articles, 31 K) and train logistic-regression and random-forest classifiers. Our best BoR+RF model achieves an F_1 of 0.81 matching an n-gram baseline while highlighting key relations (e.g. `nn`, `punct`) that most separate real and fake news. SHAP analysis confirms these dependency labels capture the core stylistic footprints of machine-generated text. By combining competitive accuracy with directly interpretable grammatical cues, our approach offers a transparent, linguistically grounded defense against automated misinformation.

Keywords: misinformation detection, large language models, dependency parsing, stylometric analysis, Bag-of-Relations, interpretability, SHAP, explainable AI, Relation Frequency–Inverse Document Frequency

1 Introduction

The recent explosion of large language models (LLMs) has rendered automated news generation both trivial and widely accessible. This has lead to an inundation of social media platforms with seemingly innocuous but malignant machine-generated content masquerading as human-written content. Human moderation remains costly and slow, while LLMs' opaque,"black-box" behavior makes it difficult to diagnose or correct their well-documented tendency to hallucinate and fabricate information.

Detecting neural fake news is thus a pressing challenge. Prior work has shown that LLM sampling and decoding strategies leave subtle statistical artifacts in generated text [5]. We hypothesize [7] that these artifacts extend into the syntactic domain—specifically, that (1) dependency parse–tree structures encode

literary style and idiosyncrasies; (2) LLM decoding biases subtly alter those structures; and (3) by counting and weighting dependency relations, we can both distinguish and explain machine-versus-human news. Building on this intuition, we propose two features that capture important stylistic idiosyncrasies and can be explained through established grammatical framework:

- **Bag-of-Relations (BoR)**: a simple histogram of dependency labels (e.g. `nsubj`, `punct`, `conj`) aggregated across an article's sentences.
- **Relation Frequency–Inverse Document Frequency (RF-IDF)**: a TF-IDF–style weighting over the same relation vocabulary.

These features require only a fixed, off-the-shelf dependency parser [3] (trained on Universal Dependencies, with strong Unlabelled Attachment Score and Labelled Attachment Score [10]) and make no further model assumptions. We evaluate BoR and RF-IDF on two benchmarks: NeuralNews (64 K articles) and a new "Articles" corpus ($\approx$ 31 K), using both logistic regression and random forests. Our best BoR+RF model achieves an F_1 of 0.81, matching an n-gram baseline while revealing the most discriminative syntactic relations (e.g. `conj`, `punct`, `nsubj`). Finally, using SHAP we show that these parse-based features yield clear, human-readable explanations: the very dependency labels most biased in LLM output become our strongest indicators of neural fake news. By combining high accuracy with direct explainability, our work points toward more transparent, linguistically grounded defenses against the next generation of automated misinformation.

2 Related Work

We organize prior work into three broad categories: stylometric approaches, neural model–based detectors, and human-in-the-loop assistive tools.

2.1 Stylometric and Syntactic Methods

Early deception-detection research showed that shallow linguistic cues (n-grams, POS tags, readability scores) can reveal authorship and falsified content [9,11]. Moving beyond surface features, syntactic stylometry leverages deeper structure: Feng et al. [4] demonstrated that PCFG production rules and dependency-parse patterns improve deception classification on hotel reviews and essays. More recently, Schuster et al. [13] evaluated stylometric features specifically for machine-generated news, finding that syntactic idiosyncrasies degrade when both fake and real articles originate from neural generation pipelines.

2.2 Neural Model–Based Detection

The advent of large pretrained LLMs has inspired adversarial benchmarks in which a generator and discriminator co-evolve. [15] introduced GROVER, a GPT-2–style generator conditioned on headline, author, and date fields, paired

with BERT[2]- and GPT-based [12] discriminators; GROVER's discriminator achieves near-perfect accuracy on its own generated outputs. Ippolito et al. [6] studied the effect of decoding algorithms (top-k, nucleus sampling) on detection performance, showing that discriminators trained on one sampling regime do not generalize to others. Bakhtin et al. [1] explored Energy-Based Models (EBMs) for text discrimination, generating negative samples via pretrained generators and training EBMs to distinguish them from human text.

2.3 Assistive Visualization and Human–Machine Collaboration

Complementing fully automated detectors, several works provide interpretable insights to human raters. The Giant Language model TRacer (GLTR) by Gehrmann et al. [5] color-codes words by their rank in the LLM's token distribution, enabling users to spot overly "head-focused" sampling. In user studies, GLTR improved human detection accuracy from roughly 52% to over 70%, demonstrating the value of statistical visualizations in augmenting human judgment.

2.4 Our Contribution

While prior approaches excel in accuracy, they often trade away interpretability (deep neural classifiers) or rely on shallow stylometric proxies. In contrast, our work bridges these paradigms by extracting *dependency-parse*–based features that (1) capture the *syntactic footprints* of LLM decoding, (2) retain interpretable semantics at the relation level, and (3) achieve competitive performance against both n-gram and neural baselines. [7]

3 Methodology

3.1 Problem Formulation

The domain of misinformation detection is extensive and evolving, encompassing a vast number of definitions as to what constitutes misinformation. Therefore, we confine our research to defining misinformation as any news article generated by a Large Language Model based on a prompt that constitutes the article title, authors, published date, etc. The phenomenon of misinformation arises from the fact that LLMs tend to hallucinate information, leading to the unintentional generation of falsified information.

The findings of [15] emphasize how LLMs using generated context during test time, unlike ground truth context during train time, lead to exposure bias. Furthermore, the utilization of decoding strategies during text generation imprints the text with statistical idiosyncrasies. In our current study, we addressed the misinformation detection problem through the lens of stylometric features, building on these findings, and formulated three fundamental hypotheses: 1.) The structure of dependency parse trees captures stylistic information due to subjective ordering and choice of words. 2.) Decoding strategies affect the dependency

parse tree structure due to the sampling space of words while generating text. 3.) The inherent meaning attached to the dependency relationships can explain the statistical features built from dependency-parse tree structures.

3.2 Datasets

NeuralNews We evaluate our features on the **NeuralNews** corpus, originally introduced by Tan et al. (2020) [14] as a large-scale benchmark for detecting machine-generated news articles. NeuralNews consists of 128 K article–caption pairs derived from the GoodNews dataset, which itself is drawn from New York Times content. For each real article (with its associated metadata: headline, author(s), publication date), Zellers et al.'s [15] GROVER model was conditioned on these fields to generate a matching fake article, yielding 32 K samples in each of four categories (real/gold vs. generated captions and articles).

Class	**#Train**	**#Validation**	**#Test**	**Total**
Fake	22309	2885	6806	32000
Real	22491	2875	6634	32000
Total	44800	5760	13440	64000

Table 1: Distribution of Fake and Real samples in NeuralNews dataset.

To tailor NeuralNews to pure text-based fake-news detection, we omit all image captions and restrict our focus to the article bodies and metadata. This transforms the corpus into 64 K real–fake article pairs. We then split this refined dataset into 70% train, 10% validation, and 20% test subsets. By leveraging NeuralNews sourced from professionally edited journalism and contrasting it with GROVER's adversarially generated outputs, we obtain a challenging, high-quality benchmark that reflects both linguistic diversity and real-world domain shifts.

Articles Our **Articles** dataset complements NeuralNews by incorporating more recent events and varied outlets. We curated nearly 15K human-written articles from sources such as the BBC, Al-Jazeera, and The New York Times across three broad topics including *COVID-19*, *Climate Change*, and *Military Ground Vehicles*. We further generated matching fake versions through the same GROVER pipeline utilized in generating NeuralNews dataset. The resulting 30,873 sample corpus (split 70%/10%/20% for train/validation/test) enables evaluation on contemporary content not covered in NeuralNews.

3.3 Feature Engineering: Bag-of-Relations and Relation Frequency Inverse Document Frequency

Inspired by Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), we propose two novel features called Bag-of-Relations (BoR)

Class	#Train	#Validation	#Test	Total
Fake	10106	1013	4320	15439
Real	10090	1007	4337	15434
Total	20196	2020	8657	30873

Table 2: Distribution of Fake and Real samples in Articles dataset.

and Relation Frequency-Inverse Document Frequency (RF-IDF),[7] which leverage the dependency relationships between words to capture stylistic idiosyncrasies and build explainable features that can aid in understanding the authenticity of news articles.

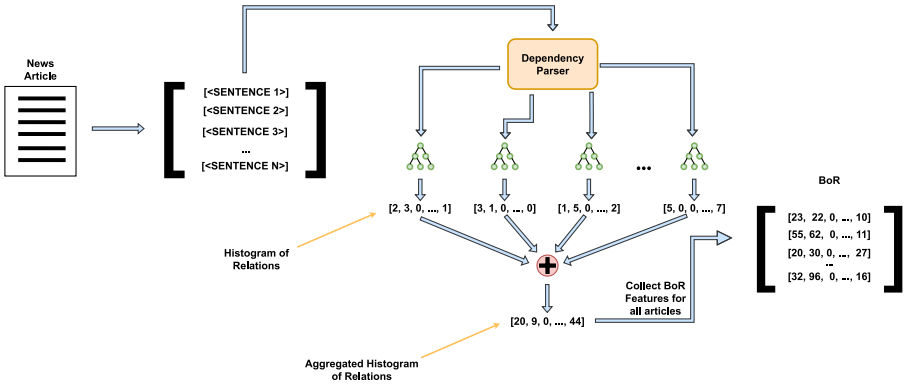


Fig. 1: Algorithm for Bag-of-Relations Features

Similar to the BoW feature, we construct the BoR features by counting the occurrences of dependency relationships extracted from the dependency parse trees of sentences. To build these features, we tokenize each news article into sentences and pass it through a dependency parser, obtaining dependency relations for each sentence. We then aggregate these relations across all sentences within an article, resulting in a vector representation where each dimension corresponds to the count of a specific dependency relation. Since the dependency standards, such as Universal Dependencies and Stanford Dependencies, contain a fixed set of grammatical dependency relations, we transform the text data into fixed-length, interpretable, and style-oriented vectors that highlight linguistic patterns characteristic of either human-written or machine-generated content. We outline this process in figure 1

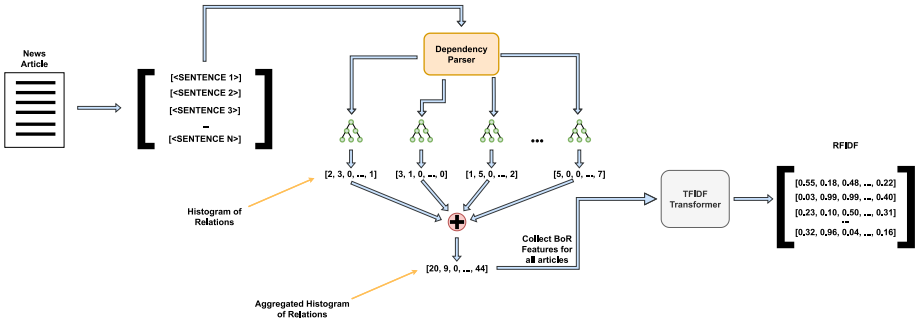


Fig. 2: Algorithm for Relation-Frequency Inverse-Document-Frequency Features

The RFIDF features are an extension to the BoR features and similar to the TFIDF feature, we compute the product of relation frequency and inverse document frequency to encode the peculiarities of dependency relationships between machine-generated news articles and human-written news articles. We outline this process in figure 2

3.4 Models and Interpretability

To build the proposed features, we use a pretrained dependency parsing (DP) [3] model that is trained to recognize dependency relations from the Universal Dependency framework and performs well on the Unlabelled Attachment Score (UAS) and Labelled Attachment Score (LAS) metrics [10]. We also fix the DP model across experiments to keep our results consistent on different datasets and eliminate noise due to the DP model biases.

One of the core motivations of our research is to understand what distinguishes a human-written news article from a machine-generated news article. Consequently, we choose models like Logistic Regression and Random Forests that inherently allow interpretability of features by design. Once we train the models, we analyze the model weights and the feature importance scores to identify essential dependency relations that contribute most towards detecting a misinformed news article. Additionally, we leverage SHAP values to quantify and interpret each dependency relation's contribution to model predictions.

4 Experiments and Results

We outline the experiments conducted with the proposed features, Bag-of-Words and Relation Frequency Inverse Document Frequency, and evaluate their effectiveness in distinguishing and explaining the differences between machine-generated and human-written text. Furthermore, we conduct our experiments on the Neural News and Articles dataset, with each containing roughly 64,000

and 30,000 samples, respectively. Both datasets contain almost an even split of human-written and machine-generated news articles. Next, we partition the datasets into training, validation, and testing subsets with a proportional split of 70%, 10%, and 20%, respectively, and evaluate our models using standard classification metrics such as precision, precision, recall, and F-1 scores.

4.1 Experiment 1: Analyzing features in lower dimensions

We investigate the validity of the proposed features and check their discriminatory nature by transforming the features to lower dimensions using Principal Component Analysis (PCA).

(a) PCA Plot of NeuralNews Dataset with Top-3 Principal Components

(b) PCA Plot of Articles Dataset with Top-3 Principal Components

Fig. 3: PCA Plots for NeuralNews and Articles Datasets.

Upon plotting the features in three-dimensional space as seen in figure 3, we clearly observe the distinction between machine-generated and human-written news articles. This confirms our intuition that the features built on dependency parse trees capture stylistic differences between human-written and machine-generated news samples. We also observe significant variance in the human-written news articles compared to the machine-generated news articles, which highlights biases due to the decoding strategies employed during text generation.

4.2 Experiment 2: Classification using BoR and RFIDF features.

We applied our BoR and RFIDF features to train the Logistic Regression and Random Forest models to observe significant differences in the stylistic cues between machine-generated news articles from human-written news articles. We systematically explore each classifiers' hyperparameter configuration space such as varying regularization parameters, penalty scores for logistic regression model and maximum tree depth for random forest models, etc. We performed a grid search on each model-feature combination evaluating models on validation set and reported the final performance on the test set. We also estimated baseline

performance of these models on the n-gram features which we provided for comparison.

Logistic Regression		**Random Forest**	
Parameter	**Values**	**Parameter**	**Values**
penalty	none, l1, l2, elasticnet	n_estimators	10–100 (step 10)
solver	newton-cg, lbfgs, liblinear	criterion	Gini, Entropy
C	1e–5, 1e–4, 1e–3, 1e–2, 1	max_depth	2, 5, 10, 20, 50
		min_samples_leaf	1, 5, 10

Table 3: Hyperparameter search spaces for Logistic Regression (left) and Random Forest (right).

Our results indicate the superior performance of the RF-BoR combination on Articles dataset but falls short in comparison to the Neural News baseline. One contributing factor to the RF–BoR model's inability to outperform the n-gram baseline on NeuralNews is the extreme stylistic consistency of the genuine articles. All "real" NeuralNews samples originate from a single editorial source: the New York Times via the GoodNews corpus, leading to a very uniform dependency-parsing patterns.

Data/model	**LR-BoR**	**LR-RFIDF**	**RF-BoR**	**RF-RFIDF**	**Baseline**
NeuralNews	0.79	0.78	0.79	0.78	**0.92**
Articles	0.77	0.77	**0.81**	0.80	0.79

Table 4: F-1 scores of all the (model, feature, dataset) combinations.

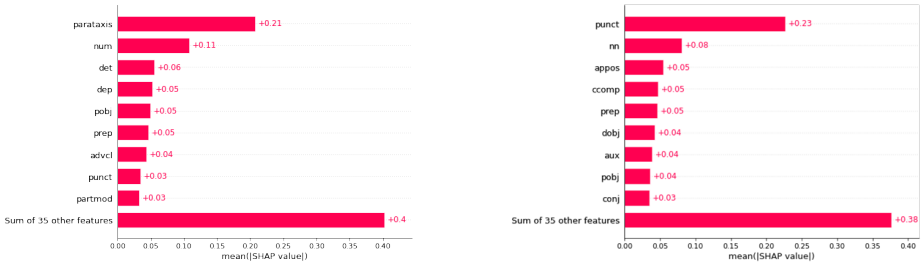
The GROVER model was also trained on the NY-Times articles, and hence its synthetic outputs reproduce similar motifs in generated news articles. As a result, our Bag-of-Relations counts exhibit little variance between real and generated articles, which limits their discriminative power relative to lexically rich n-gram representations. These experiments serve the primary objectives of this research which is explainability. Each BoR and RFIDF feature correspond to semantically verifiable grammatical rules and relations. In practice, experts can see how a particular dependency relationship (say `punct`) scores much higher in fake articles compared to their real counterparts.

4.3 Experiment 3: Interpretability analysis

Upon examining the feature importance values assigned to each dependency relationship, we observe that eight relations namely `punct`, `det`, `nn`, `prep`, `pobj`,

`nsubj`, `dobj`, and, `parataxis` significantly contribute to determining machine-generated and human-written text. These eight attributes encode core clause structure and parenthetical asides that are characteristic to the text. The divergence of their frequencies in machine-generated vs human-written text gives us an insight into unique syntactic fingerprints of the different sources.

- **NeuralNews RF–BoR (Top-10):** `punct`, `nn`, `prep`, `pobj`, `nsubj`, `det`, `dobj`, `appos`, `ccomp`
- **Articles RF–BoR (Top-10):** `parataxis`, `num`, `prep`, `det`, `punct`, `pobj`, `dobj`, `dep`, `nsubj`, `nn`



(a) RF–BoR (Articles) (b) RF–BoR (NeuralNews)

Fig. 4: SHAP values for the RF–BoR model on (a) the Articles dataset and (b) the NeuralNews dataset.

Further investigation with the SHAP[8] values does infact corroborate earlier finding on the important dependency parsing relationships that capture stylistic differences between machine-generated and human-written text. Since each feature corresponds to a well-defined grammatical function, we gain a transparent and linguistically grounded explanation of where LLMs diverge from human prose. For example, the pronounced parataxis spike in Articles suggests that models overuse parenthetical constructions, while the `punct` elevation in NeuralNews reflects different comma-and-period placement patterns. These insights point directly to specific stylistic quirks rather than abstract embeddings, making our BoR features both powerful and readily interpretable.

5 Conclusion and Future Work

We have shown that two simple, dependency-parsing-based features: Bag-of-Relations and Relation-Frequency Inverse-Document-Frequency, robustly capture the stylistic footprints that distinguish machine-generated from human-written news. Across two large benchmarks and both logistic-regression and

random-forest classifiers, our BoR+RF model matched n-gram baselines (F_1 = 0.81) while pointing directly to interpretable syntactic cues (e.g. nn, punct, parataxis) as the key discriminators. Although our current study fixes a single off-the-shelf parser and "shallow" classifiers to maximize transparency, it will be valuable to explore whether these dependency-based signals hold (or even strengthen) under more complex architectures such as graph-neural-network encoders or the latest LLM-driven discriminators and on newer, more varied corpora of automatically generated text.

References

1. Bakhtin, A., Gross, S., Ott, M., Deng, Y., Ranzato, M., Szlam, A.: Real or fake? learning to discriminate machine from human generated text. arXiv preprint arXiv:1906.03351 (2019)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Dozat, T., Manning, C.D.: Deep biaffine attention for neural dependency parsing. arXiv preprint arXiv:1611.01734 (2016)
4. Feng, S., Banerjee, R., Choi, Y.: Syntactic stylometry for deception detection. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 171–175 (2012)
5. Gehrmann, S., Strobelt, H., Rush, A.M.: Gltr: Statistical detection and visualization of generated text. arXiv preprint arXiv:1906.04043 (2019)
6. Ippolito, D., Duckworth, D., Callison-Burch, C., Eck, D.: Automatic detection of generated text is easiest when humans are fooled. arXiv preprint arXiv:1911.00650 (2019)
7. Karumuri, R.T.: Interpretable features for distinguishing machine generated news articles. Master's thesis, Arizona State University (2022)
8. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions (2017), https://arxiv.org/abs/1705.07874
9. Mihalcea, R., Strapparava, C.: The lie detector: Explorations in the automatic recognition of deceptive language. In: Proceedings of the ACL-IJCNLP 2009 conference short papers. pp. 309–312 (2009)
10. Nivre, J., Fang, C.T.: Universal dependency evaluation. In: Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017). pp. 86–95 (2017)
11. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. arXiv preprint arXiv:1107.4557 (2011)
12. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog 1(8), 9 (2019)
13. Schuster, T., Schuster, R., Shah, D.J., Barzilay, R.: The limitations of stylometry for detecting machine-generated fake news. Computational Linguistics 46(2), 499–510 (2020)
14. Tan, R., Plummer, B.A., Saenko, K.: Detecting cross-modal inconsistency to defend against neural fake news. arXiv preprint arXiv:2009.07698 (2020)
15. Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., Choi, Y.: Defending against neural fake news. Neurips (2020)