What Makes Fake News? A Comparison of Lexical, Stylistic, and Semantic Models

Josephine Mayans
ENSAE
josephine.mayans@ensae.fr

Kenza Miousset
ENSAE
kenza.miousset@ensae.fr

Antonin Henriet
ENSAE
antonin.henriet@ensae.fr

Abstract

Fake news, that is false or misleading information disguising as legitimate news, poses a serious threat by influencing public opinion and spreading misinformation. In this context, fake news detection has become an urgent task for mitigating the societal harms caused by misinformation, especially in political contexts. This study investigates which textual characteristics are most predictive of fake news by comparing three different types of text representations: world-level features (TF-IDF), linguistic stylistic features (such as punctuation and sentence structure), and deep contextual embeddings (BERT). Our dataset contains approximately 21,500 fake and 21,000 true news articles which allows us to evaluate and compare models in terms of accuracy and precision. All models tested attain high accuracy within the dataset (over 90%), with BERT slightly outperforming the two other models. Our findings suggest that writing style alone can be highly predictive, highlighting the importance of both lexical and linguistic signals in fake news detection.

1 Introduction (*written by Joséphine*)

Fake news has a very detrimental impact on society, from influencing elections to spreading dangerous health misinformation, as we've seen during the Covid-19 crisis. Fake news is usually defined as news that are intentionally and verifiably false, spread to mislead readers, often with the goal of political gain or ad revenue. Considering the threat fake news represents, developing robust automated fake news detection systems has become a crucial task within the field of NLP.

Researchers have developed many automated detection techniques, most of which rely on machine learning models trained on the textual content of news articles. Traditional evaluations show that these models can achieve very high classification accuracy when tested on unseen portions of the same dataset used for training. However, it remains an open question to know what fundamentally characterizes a fake news article. Is it mainly a matter of writing style, such as a catchy tone, simplified language and short sentences? Is it the use of emotionally charged keywords and controversial topics? Or could it be deeper contextual and semantic patterns that distinguish fake news from legitimate reporting? Understanding which textual features are most predictive is essential for building more robust detection models.

This study focuses on fake news detection within political news using a dataset composed of fake and true news articles. Our objective is to evaluate and compare the performance of models trained with three distinct types of textual representations: word-level features (TF-IDF), linguistic stylistic

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

features, and deep contextual embeddings (BERT). We investigate whether models relying on writing style indicators can achieve comparable accuracy to those based on word occurrence patterns, and how they perform relative to models leveraging semantic understanding through deep learning. In short, we try to better understand which textual characteristics are most predictive of fake news, by contrasting signals captured by each of the three approaches: keyword frequency for TF-IDF, writing patterns for stylistic features, and contextual meaning for BERT.

2 Brief State-of-the-Art *(written by Antonin)*

Fake news detection has emerged as a critical application of natural language processing techniques. The dominant methodological paradigm follows a three-stage pipeline: text pre-processing, feature engineering, and supervised classification, typically implemented using Support Vector Machines (SVMs) or neural networks. Two primary families of features underpin current systems, linguistic clue-based and word-level analyses. The techniques used until now to detect fake news from a NLP perspective primarly stem from these two families. It is therefore necessary to evaluate said techniques, and comparing them on a fixed dataset can help better understand the structure of fake news.

The first category comprises word-level representations, which encode textual information into numerical formats. Basic techniques such as Bag-of-Words (BoW) rely on just the number of words to represent documents, but more elaborate ones exist, such as Term Frequency-Inverse Document Frequency (TF-IDF), which focuses on particularly distinctive terms in a document. More sophisticated embedding frameworks, including Word2Vec, produce dense vector embeddings that capture the semantic relationships between words. These three types of techniques have all been used by Thota et al. (2018), who consider the relationship between title and content of articles, and find that TF-IDF is the best approach for their particular framework. Poddar et al. (2019) conclude the same, achieving a 92,8 % accuracy in fake news detection using TF-IDF and a SVM.

The second category focuses on linguistic and stylistic cues, which are manually developed features that define the form and structure of the text rather than its semantic content. The average length of words or sentences, the distribution of parts of speech and sentimental cues are used to capture the author's style. Fernandez and Devaraj (2019) use this focus to achieve an accuracy as high as 94 % in fake-news detection, noting for instance that titles of fake news articles contain on average more words than true information, whereas they are overall shorter. By capturing the stylistic dimensions of writing, linguistic cues aim to provide models with signals that can be better generalised across different content areas. A combination of the word-centric and linguistic cue approach is implemented by Gautam and Jerripothula (2020), which also provides high accuracy scores.

Further techniques include contextual token-level embedding techniques as BERT (Bidirectional Encoder Representations from Transformers). BERT represents the state of the art in fake news detection due to its ability to generate rich contextual embeddings, outperforming models based on Word2Vec or TF-IDF. However, as highlighted by Hoy and Koulouri (2025), its performance drops significantly in cross-domain settings due to overfitting to lexical cues. Other studies, such as Poddar et al. (2019), confirm this limitation, emphasizing the need for more generalisable approaches.

Constrasting these three approaches should allow us to better determine what the most proeminent characteristics of fake news are, and what are the best methods to detect and mitigate the effects of fake news.

3 Experiment *(written by Joséphine)*

Despite the abundance of research on fake news detection, it remains unclear which aspects of a news article's text contribute to a model's ability to identify misinformation. Previous work did show that high accuracy can be obtained using models trained on raw text, but this might amalgamate distinct sources of information such as lexical content, writing style or semantic meaning. In order to build more interpretable and generalizable detection system, we try to disentangle these different textual signals.

We propose to evaluate three families of models that each rely on a different type of textual representation:

- **TF-IDF + Logistic Regression**: emphasizes keyword frequency and lexical patterns.
- **Stylistic Features + Random Forest**: captures writing style indicators (punctuation, sentence structure, vocabulary richness. . .).
- **BERT + Fine-tuning**: provides deep contextual language understanding via Transformer-based embeddings.

By holding the dataset and evaluation protocol constant, we assess how different text representations influence predictive performance, which allows us to ask: What kind of information is each model using to detect fake news? For example, does writing style alone contain enough signal to distinguish fake from real news? Or is contextual understanding necessary for more robust classification?

4 Dataset Description (*written by Kenza*)

We use the ISOT Fake News Dataset, developed by the Information Security and Object Technology (ISOT) Research Lab at the University of Victoria.

The dataset comprises political news articles, from 2015 to 2018, divided into two categories: fake news articles collected from unreliable websites, and real news articles sourced from Reuters. The original dataset is provided in two separate CSV files, Fake.csv and True.csv. We merged these files into a single dataframe and assigned a binary label: 1 for fake news and 0 for real news.

To improve data quality, we performed a filtering step by removing articles with fewer than 50 words. This allowed us to eliminate empty or incomplete entries that could negatively impact model performance. After this filtering, the final dataset contains 42,677 articles, of which 21,586 are labeled as fake and 21,091 as real.

We then constructed a new column by concatenating the article's title and body, as titles often contain informative cues about the article's intent, especially in fake news where sensational or misleading headlines are common.

For each article, we also computed basic statistics such as article length, which revealed differences across the two classes. On average, fake news articles tend to be longer (mean: 459 words, median: 381), while real news articles are slightly shorter (mean: 391 words, median: 364). The longest article in the dataset contains 8,135 words, and the shortest valid article, after filtering, contains exactly 50 words. Overall, the average article length across the dataset is 425 words.

5 Methodology (*written by Joséphine*)

This section details the preprocessing steps, feature extraction pipelines, and model training procedures used for the three distinct modeling approaches.

5.1 Data Preprocessing

For all models, we used the same dataset composed of political news articles labeled as either fake or true. Since the title often contains important information, such as misleading clickbait phrasing in fake news, we merged the title and full text of each article. We also performed light preprocessing such as lowercasing for TF-IDF and BERT. More aggressive preprocessing (e.g., lemmatization, punctuation removal, stop-words/URL/Twitter handles removal) was applied only for TF-IDF, because such tokens mays carry contextual significance. After cleaning, the dataset is randomly split into training and testing sets (80/20 split) stratified by class label. This yielded 34,000 training articles and 8,500 testing articles, with roughly equal fake/real proportions in each split.

5.2 TF-IDF + Logistic Regression

The first approach, TF-IDF, relied on the frequency of keywords. We cleaned the text by removing URLs, digits, and punctuation, converting all text to lowercase, and applying lemmatization. We then computed TF-IDF vectors using Scikit-learn's TfidfVectorizer limited to the top 10,000 features. These features were used to train a logistic regression model with L2 regularization. The model was evaluated on the test set using standard classification metrics.

For TF-IDF we also looked at a ranked list of keywords that contributed to the classification for each dataset, that is we focused on the frequency of key words used by the model trained on the dataset, and whether they contributed to a “True News” classification or a “Fake News” classification.

5.3 Stylistic Features + Random Forest

The second approach relied on writing style. We manually generated a set of several linguistic features. We began with a basic set of 7 features and progressively expanded it to the 11 most relevant stylistic features based on prior research and exploratory analyses. The final list includes: average word length, average sentence length, total punctuation count, ratio of uppercase letters to total letters, number of exclamation marks, number of question marks, lexical diversity (ratio of unique words), stop word ratio (“filler words” characteristic of fake news), number of all-uppercase words, average number of characters per sentence, Flesch Reading Ease score (a readability metric that estimates how easy a text is to understand).

We trained a RandomForestClassifier with 100 estimators on the resulting feature vectors. In addition to standard classification metrics, we examined the feature importance scores learned by the model to better understand which aspects of writing style were most discriminative between fake and true articles.

5.4 BERT + Fine-Tuning

The third model captured deep contextual meaning. Unlike TF-IDF, BERT embeddings are sensitive to both word meaning and position in a sentence. BERT is pretrained using two tasks: Masked Language Modeling (MLM), which enables it to model word relationships, and Next Sentence Prediction (NSP), which helps it understand inter-sentence dependencies.

We used the bert-base-uncased model and fine-tuned it for binary classification using a PyTorch-based training loop (without Hugging Face’s Trainer class for transparency and control). Text was tokenized using BertTokenizer and padded/truncated to a maximum of 512 tokens. We wrapped the encoded inputs and labels in a custom Dataset class and loaded them into PyTorch DataLoaders with batch size 8. We trained the model on GPU for 3 epochs using the AdamW optimizer and a learning rate of 2e-5. During training, we monitored loss per epoch.

5.5 Comparison

In order to compare the three models, we evaluated each model, after training, on the test set by computing accuracy, precision, recall, F1 score, and a confusion matrix. This method allows us to compare shallow models that rely on surface-level or stylistic signals with a deep model that captures rich semantic structure.

6 Results and Analysis (*written by Kenza*)

This section presents and compares the performance of the seven models trained to detect fake news in political news articles. We evaluate three main modeling strategies: one based on word-level frequency (TF-IDF + Logistic Regression), a series of models using stylistic features (Random Forest with incremental feature additions), and a contextual deep learning model (BERT fine-tuned on our dataset).

6.1 Overall Performance Comparison

To evaluate model performance, we report four standard classification metrics: accuracy, precision, recall, and F1-score.

- **Accuracy** measures the proportion of correctly classified articles overall.
- **Precision** is the proportion of articles predicted as fake that are truly fake, which reflects the model’s reliability when flagging fake news.
- **Recall** is the proportion of actual fake news articles that are correctly identified, indicating the model’s ability to catch misinformation.

- **F1-score** is the harmonic mean of precision and recall, providing a balanced metric especially useful when the cost of false positives and false negatives differs.

Together, these metrics allow us to assess not only how often the model is right, but also how trustworthy and thorough it is in detecting fake news. The table below summarizes the performance of each model on the test set.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
TF-IDF + Logistic Regression	0.99	0.99	0.98	0.99
Stylistic Features (Base)	0.90	0.90	0.91	0.91
+ Stopword Ratio	0.94	0.93	0.96	0.94
+ ALL CAPS Words	0.94	0.93	0.96	0.95
+ Avg Characters per Sentence	0.94	0.93	0.96	0.94
+ Readability Score (Final model)	0.95	0.93	0.96	0.95
BERT	1.00	0.999	1.000	1.000

Table 1: Performance of all models on the test set

The BERT model achieves near-perfect results. This performance confirms the power of deep contextual language models in capturing subtle semantic nuances and complex dependencies across sentences. However, such impressive results should be interpreted with caution: the extremely high scores may indicate overfitting to the specific dataset used, especially considering its relatively homogeneous content.

TF-IDF also performed well with an F1-score of 0.99. This suggests that, in this dataset, the frequency and distribution of certain keywords are highly predictive of whether an article is fake or real. In particular, TF-IDF is able to identify discriminative terms that are overrepresented in one class relative to the other, allowing even a simple linear classifier to achieve excellent results. The model's success also reinforces the idea that lexical signals alone, when appropriately weighted, can offer a powerful signal for classification.

The final stylistic model (Random Forest with 11 handcrafted features) achieved a very competitive F1-score of 0.95. Fake news are not only different in content but often in form.

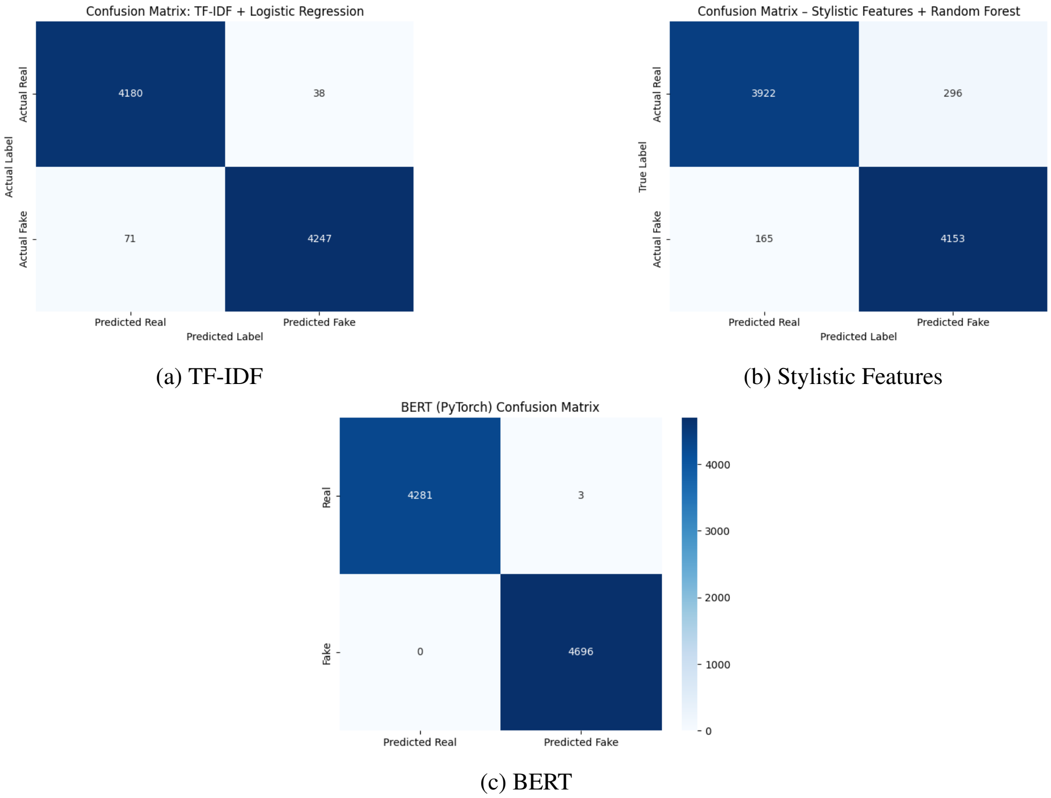
6.2 Confusion Matrices

The confusion matrices of the three main models provide valuable insight into the types of errors each approach tends to make.

TF-IDF performs very well overall, misclassifying only 38 real news articles as fake and 71 fake news articles as real, resulting in a low overall error rate.

The final version of the stylistic features model, while still accurate, commits slightly more errors, particularly in the form of false positives : real news articles that are incorrectly labeled as fake. This suggests that stylistic cues alone may not be sufficient to fully distinguish between the two categories.

BERT achieves near-perfect classification on the test set, with only three fake-negatives.



(a) TF-IDF

(b) Stylistic Features

(c) BERT

Figure 1: Confusion matrices for the three classification models.

6.3 TF-IDF Keyword Analysis

We analyzed the top features learned by the logistic regression classifier trained on TF-IDF vectors. These features are the words most strongly associated with each class.

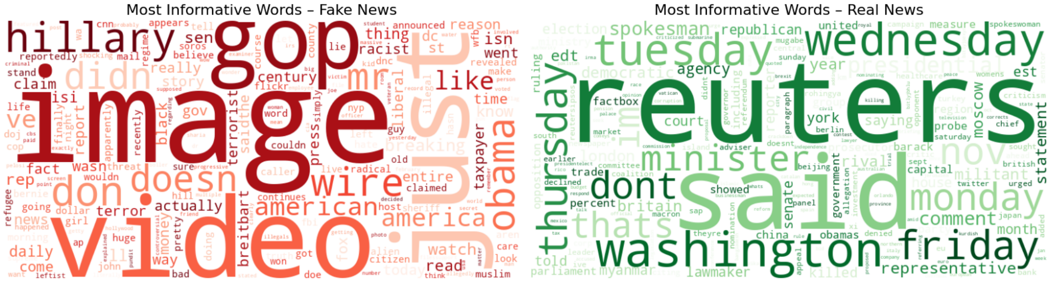
Fake Word	Coef	Real Word	Coef
image	8.57	reuters	-24.33
video	7.90	said	-18.70
just	7.20	washington	-6.07
gop	6.09	wednesday	-5.24
hillary	5.57	tuesday	-4.95
dont	5.49	thursday	-4.43
wire	4.93	friday	-4.35
obama	4.83	monday	-3.84
mr	4.67	nov	-3.82
didnt	4.52	dont	-3.76
doesnt	4.40	im	-3.58
american	4.35	minister	-3.49
ve	4.31	thats	-3.31
like	4.20	presidential	-3.30
america	4.16	spokesman	-3.25

Table 2: Most Informative Words by Logistic Regression Coefficients

The coefficients shown in the table correspond to the weights learned by the logistic regression classifier trained on TF-IDF vectors. In this setting, each word is treated as a separate feature, and the model learns a weight for each word during training. These coefficients represent the strength and direction of the association between the presence of a word and the predicted class (fake or real).

A positive coefficient means that the presence of the corresponding word increases the probability that the article will be classified as fake news. The larger the value, the stronger the contribution to a "fake" classification. Conversely, a negative coefficient indicates that the word contributes to the classification of an article as real news. The most negative values thus represent words strongly associated with true reporting.

A word cloud visualization further illustrates the lexical signals selected by the model.


Most Informative Words – Fake News

Most Informative Words – Real News

Figure 2: Word clouds: Fake news versus Real news

The words in red are those with the most positive coefficients in the logistic regression model, meaning that their presence increases the probability that the article will be classified as *fake news*. The words in green have the most negative coefficients, and are therefore characteristic of the *real news* class.

In the case of fake news, significant terms include *video*, *image*, *GOP*, *obama*, *hillary*, *terror*, *racist*, *money*. Many of these words refer to controversial political figures or polarizing topics, while others suggest the presence of sensationalist or emotionally evocative content. Not all high-weight terms are ideological in nature. Words like *video*, *image*, *just*, or *watch* may point toward the structural features of fake news, particularly its tendency to rely on visual content and catchy phrasing to draw attention, typical of clickbait-style articles.

In contrast, the words most associated with real news articles include *reuters*, *said*, *reporter*, *official*, *ministry*, *statement*, *measure*, *representative* and days of the week. These terms are indicative of a more formal and source-based reporting style, characteristic of legitimate journalistic standards. Their presence suggests solid sources that lend credibility and objectivity to the article.

This contrast highlights how lexical choices can serve as strong signals for automated classifiers and provides an interpretable basis for understanding the predictions made by the TF-IDF + logistic regression model.

6.4 Stylistic Feature Importance

To better understand which stylistic cues are most predictive of fake news, we conducted an ablation analysis by progressively adding handcrafted features to the Random Forest model. We started with a set of 7 basic features, and then incrementally introduced four additional features: the proportion of stopwords, the number of all-uppercase words, average characters per sentence, and the Flesch Reading Ease score. The inclusion of these features led to modest but consistent improvements in accuracy and F1-score.

This graph shows the importance of the different stylistic variables in the Random Forest model trained on our final set of 11 features. The *stopword ratio* stands out clearly as the most predictive variable for distinguishing fake news from real news, with a much higher importance score than all the others. This suggests that fake news articles tend to contain a higher proportion of grammatical or "empty" words, which may reflect a looser or less informative writing style.

Next in importance are the number of questions and the uppercase ratio, two features often associated with emotional or emphatic writing, which is typical of sensationalist style. The presence of exclamation marks and shorter average sentence length also emerge as strong signals. Fake news

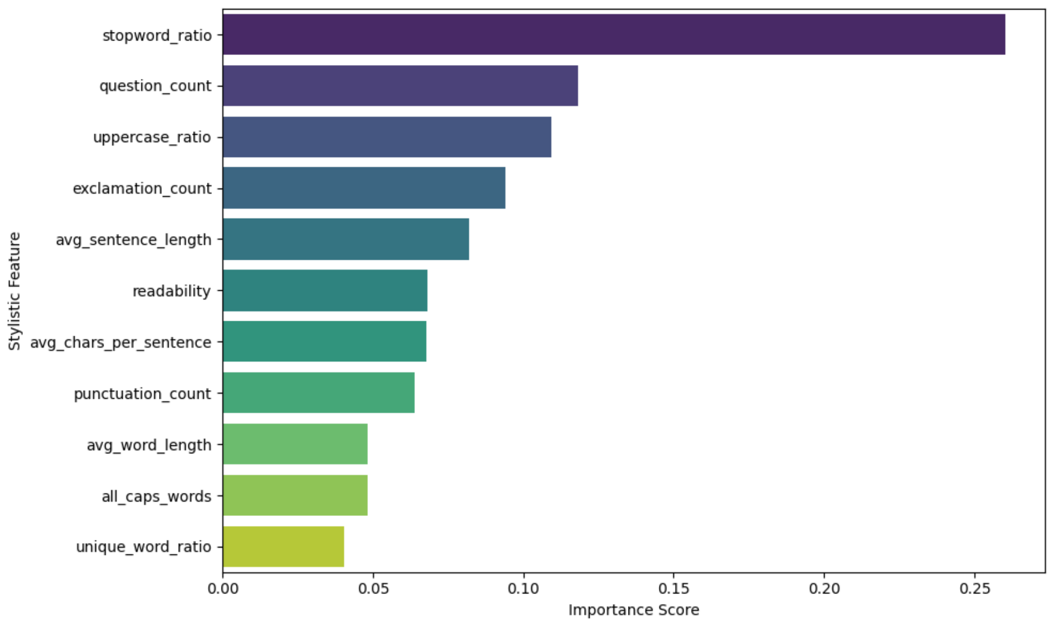


Figure 3: Feature Importances (Random Forest on Stylistic Features)

content thus seems to be more fragmented, more heavily punctuated, and more "aggressive" in its form.

Other structure-related indicators, such as readability (measured with the Flesch Reading Ease score), average number of characters per sentence, and average word length, play a more modest but still noticeable role. Finally, lexical diversity (the ratio of unique words) turns out to be the least informative feature. This may be due to the fact that fake and real news articles share a relatively similar vocabulary but differ in how they use it.

6.5 Trade-offs Between Lexical, Stylistic, and Semantic Models

While all three models achieve strong performance within the dataset, they differ significantly in terms of interpretability. The TF-IDF model offers a transparent representation of the input, where individual word coefficients can be examined to understand what drives classification. Similarly, the stylistic model provides feature importance scores that reveal which linguistic markers contribute most to detecting fake news. These interpretable models are appealing in contexts where model transparency is critical, such as journalistic verification.

In contrast, the BERT model captures deep contextual and semantic information, which enables higher predictive performance but at the cost of interpretability. Its decisions are encoded in abstract representations that are difficult to trace back to specific linguistic patterns.

This trade-off between accuracy and transparency highlights an important practical consideration: while deep learning models like BERT are powerful, simpler models may be preferable when the reasoning process behind a prediction must be understood or explained.

7 Conclusion (*written by Antonin*)

Our study explores which types of textual features are most effective at detecting fake news by comparing three distinct approaches: lexical features (TF-IDF), linguistic cues, and deep contextual embeddings (BERT). We trained and evaluated models based on each representation and analyzed their relative performance and interpretability, using a dataset of political news articles.

Our results show that, while BERT outperforms other models in terms of classification metrics (accuracy, precision, recall and F1-score), other simpler models, relying on keyword frequency or stylistic patterns, also achieve strong results. Surprisingly, the stylistic model, based on handcrafted features (punctuation, capitalization, sentence structure…) demonstrated that writing style alone carries great predictive power. In particular, it showed that stop-word usage and the presence of all-caps words contribute a lot to fake news classification. Overall, our study shows that fake news is not only about what is said, but also about how it is written.

Beyond accuracy, it is also important to emphasize the importance of model transparency. Indeed, while deep learning models capture richer semantic information, they operate as black boxes, and they don't tell us much about what makes fake news. Interpretable models like logistic regression and random forests allow researchers to understand the signals driving classification. Explainability is essential when it comes to fake news detection, because it helps inform users about the linguistic and stylistic cues commonly associated with misinformation. By identifying these features, we can support critical media literacy, and help readers recognize potential fake news themselves.

While our models achieve high performance on the ISOT dataset, it remains unclear whether these results generalize to news from other domains or sources. Future work could evaluate the generalizability of these models on other datasets and investigate hybrid approaches combining stylistic, lexical and semantic features.

References

Fernandez, A. C. T. and Devaraj, M. (2019). Computing the linguistic-based cues of fake news in the philippines towards its detection. In *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics (WIMS2019)*, pages 1–9, New York, NY, USA. ACM Press.

Gautam, A. and Jerripothula, K. R. (2020). Sgg: Spinbot, grammarly and glove based fake news detection. In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, pages 174–182. IEEE.

Hoy, N. and Koulouri, T. (2025). An exploration of features to improve the generalisability of fake news detection models. Technical report, Brunel University London.

Poddar, K., D., G. B. A., and Umadevi, K. S. (2019). Comparison of various machine learning models for accurate detection of fake news. In *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, volume 1, pages 1–5. IEEE.

Thota, A., Tilak, P., Ahluwalia, S., Lohia, N., Ahluwalia, S., and Lohia, N. (2018). Fake news detection: A deep learning approach. *SMU Data Science Review*, 1(3):10.