An Empirical Study on the Generalizability of Fake News Detection Models

Julia Lu
ENSAE Paris
julia.lu@ensae.fr

Abstract

This report investigates the generalizability of fake news detection models across datasets, inspired by the work of Hoy and Koulouri (2022). We examine how well models trained on one dataset perform when tested on another, using a range of text representations including Bag of Words, TF-IDF, Word2Vec, BERT, and engineered linguistic features. While models achieve high performance when evaluated on the same dataset they were trained on, their effectiveness drops significantly when applied to unseen datasets. These results confirm the challenges of cross-domain generalization in fake news detection and align with previous findings that emphasize the need for more robust and transferable approaches.

1 Introduction

In recent years, the proliferation of fake news has raised serious concerns, as it has been massively used to manipulate public opinion, influence elections, or spread false health-related narratives. In response, numerous machine learning models have been developed to automatically detect fake news. However, many of these models achieve strong performance only within the specific datasets on which they were trained and tested, but struggle to generalize effectively to new, unseen datasets, even those within the same domain. This limits their usefulness in real-world situations where news varies in topic. Hoy and Koulouri (2022) examine this issue of generalization by evaluating the performance of various fake news detection models and feature extraction methods across different datasets. Their study highlights the limitation of current approaches and explores possible reasons behind such poor generalizability.

In this report, we replicate the analysis conducted by Hoy and Koulouri (2022). In particular, we compare the performance of different feature extraction techniques combined with different models across two fake news datasets to evaluate how well these models generalize.

2 State of the art

Fake news detection has emerged as a crucial application of Natural Language Processing (NLP), relying on effective feature extraction and robust classification models. While various approaches have been proposed, a major challenge remains: ensuring that these models generalise well beyond the specific datasets they are trained on.

Two major categories of features are commonly used in fake news detection:

- **Word-level representations** aim to encode words into numerical vectors with varying degrees of complexity. Simple approaches like Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF) represent text based on word counts and frequency-adjusted scores, respectively. However, they ignore syntax and semantics. More sophisticated approaches, such as word embeddings (e.g., Word2Vec) or contextual models like BERT, incorporate semantic relationships between words by representing them as dense vectors in continuous space (Prachi et al., 2022).

- **Linguistic cues** are derived through analysis of the corpus text. These cues may include statistical features such as average word length, sentence complexity, part-of-speech (POS) tag distributions, and the frequency of specific elements like quotations, pronouns, verbs, and named entities. Linguistic features can offer insights into the author's writing style and emotional tone, enhancing the analysis of text data (Gravanis et al., 2019).

A wide range of machine learning algorithms has been applied to the task of fake news detection. Traditional classifiers such as Logistic Regression, and Support Vector Machines (SVM) are widely used for their simplicity and interpretability. Ensemble methods, like Random Forest, Gradient Boosting, and AdaBoost, often achieve higher performance by combining the predictions of multiple base learners. More recently, deep learning architectures have been explored for their ability to model complex patterns in textual data, though they often require larger datasets and more computational resources (Sastrawan et al., 2022).

Most fake news detection models are evaluated using either holdout testing or cross-validation on a partitioned subset of the dataset used for training. While many of these models report strong performance, often achieving around 80% accuracy on average, their generalisability remains a major concern. Hoy and Koulouri (2022) focused on four political news datasets to assess generalisability within a single domain and found significant accuracy drops when testing across datasets, especially in word-representation models. While models using linguistic cues also struggled with generalizability, they performed more consistently across datasets, suggesting lower sensitivity to dataset biases.

3 Experimental design

3.1 Datasets

The datasets used in our experiments are the **ISOT Fake News dataset**[1] and the **Fake or Real (FoR) News dataset**[2]. They are both publicly available on Kaggle.

- The ISOT dataset contains 44,898 news articles labeled as either real or fake. The real news articles were sourced from different legitimate news sites, while the fake news articles were collected from unreliable sources known for spreading misinformation. The dataset is balanced, with 23,481 fake and 21,417 real articles, and focuses predominantly on political content.
- The FoR dataset includes 6,335 news articles, with 3171 articles labeled as real and 3164 labeled as fake. The topic of the articles in FoR is similar to ISOT (political news).

3.2 Text preprocessing

Text preprocessing is essential for cleaning noisy data and improving model performance. In this experiment, different levels of preprocessing were applied depending on the feature extraction method.

- For Bag-of-Words and TF-IDF models, the text was fully cleaned. This included converting to lowercase, lemmatization, and removing punctuation, URLs, Twitter handles, extra whitespace, and stop words. These steps help reduce redundancy and standardize the input for more effective pattern recognition.
- In contrast, Word2Vec and BERT require richer contextual information, so only light cleaning was performed—lowercasing, spell checking, and removing URLs and Twitter handles—to preserve syntax and semantics.
- No preprocessing was applied when extracting linguistic cues, as these features depend on the original structure and content of the text.

[1] https://www.kaggle.com/datasets/csmalarkodi/isot-fake-news-dataset

[2] https://www.kaggle.com/datasets/jillanisofttech/fake-or-real-news

3.3 Feature extraction

Various methods were tested to convert textual data into numerical representations : Bag-of-Words (BoW), TF-IDF, word embeddings (Word2Vec and BERT), and linguistic cues. These feature sets were chosen to represent both traditional and contextualised approaches to text representation. BoW and TF-IDF focus on word frequency and importance, while word embeddings capture semantic relationships between words. Linguistic cues provide structural and stylistic information, such as part-of-speech patterns, sentence complexity, and sentiment indicators. Using a variety of feature types allows for a comparative analysis of how well each supports fake news detection and generalises across datasets.

3.4 Model training

For each feature set, a range of machine learning models was trained to evaluate performance and generalisability. The models included Logistic Regression, Random Forest, Gradient Boosting, AdaBoost, and a Neural Network—all commonly used in fake news detection.

Models were trained using standard scikit-learn implementations. Each model was first evaluated using within-dataset validation, where training and testing were conducted on the same dataset via stratified 5-fold cross-validation. To assess generalisability, cross-dataset evaluation was then performed by training on one dataset (e.g., ISOT) and testing on the other (e.g., Fake or Real), and vice versa.

This training strategy enables both a comparison of model performance across feature sets and an analysis of how well each approach generalises to unseen data from a different source.

3.5 Evaluation metrics

To assess model performance, several standard classification metrics were used : accuracy, precision, recall, and F1-score. These metrics provide a balanced view of each model's effectiveness, especially in the context of binary classification tasks like fake news detection.

- Accuracy measures the overall proportion of correctly classified instances.
- Precision evaluates the proportion of true positives among all predicted positives, reflecting how well the model avoids false alarms.
- Recall assesses the proportion of true positives among all actual positives, indicating the model's ability to detect fake news.
- F1-score is the harmonic mean of precision and recall, it represents both precision and recall in one metric.

4 Data analysis

Before evaluating model performance, a preliminary analysis was conducted to better understand the characteristics of each dataset. This includes examining the distributions of article and title lengths, as well as visualizing the most frequent words in fake and real news content using word clouds. These insights provide context for potential linguistic differences and biases in the datasets, which may influence model performance and generalisability.

4.1 ISOT dataset

Figure 1 presents the distribution of article lengths (left) and title lengths (right) in number of words for fake and real news in the ISOT dataset. Fake news articles tend to be slightly shorter on average than real ones, especially in headline length, which may reflect their often less detailed or more sensational nature.

Figure 2 shows word clouds generated from the content of fake (left) and real (right) news articles. Common words in both include *"donald trump"*, and *"white house"*, which indicates a strong focus on american politics across the dataset. However, subtle differences in vocabulary are apparent. The fake news cloud features words like *"image"*, *"featured"*, and *"twitter"*, possibly hinting at

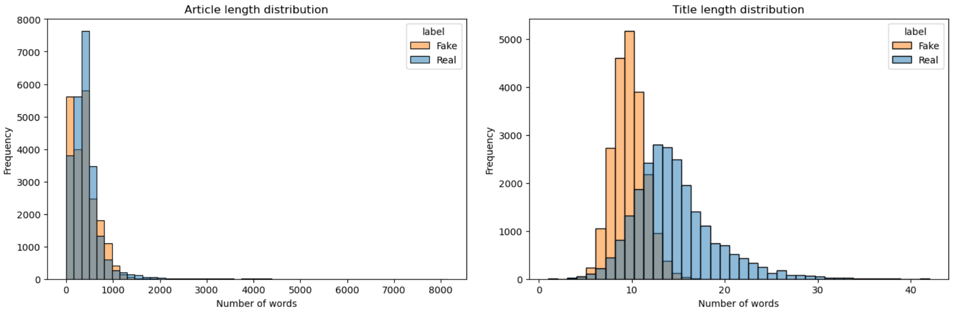


Figure 1: Article length and title length distribution for ISOT dataset

sensationalist or visually-driven content. The real news cloud, on the other hand, includes terms like *"united states"*, *"washington"*, and *"government"*, suggesting a more formal and institution-focused tone.

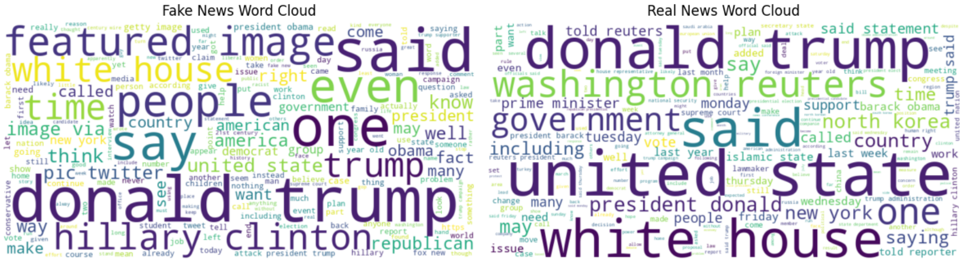


Figure 2: Word Cloud for ISOT dataset

4.2 Fake or Real dataset

Figure 3 shows the distribution of article lengths (left) and title lengths (right) for fake and real news articles in the Fake or Real dataset. In contrast to the ISOT dataset, real news articles tend to be shorter in length than fake ones, with a noticeably sharper peak in the lower word count range. Title lengths, however, show very similar distributions across both fake and real news, with most headlines falling between 5 and 15 words.

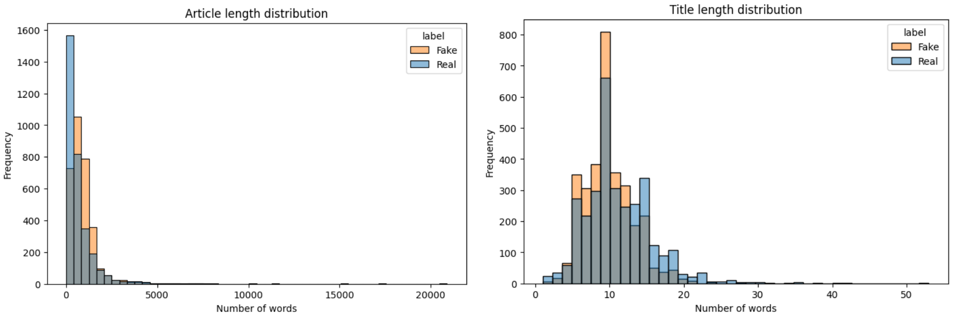
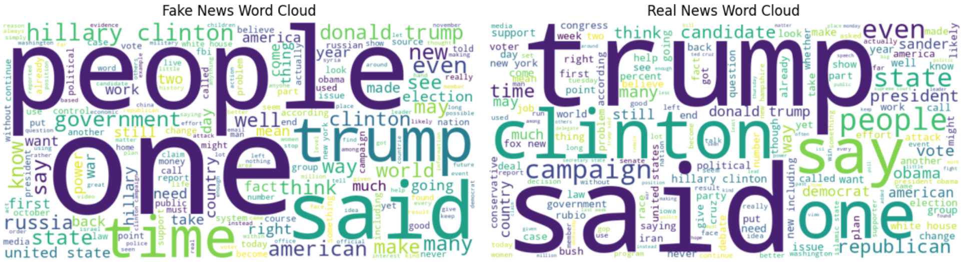


Figure 3: Article length and title length distribution for Fake or Real dataset

Figure 4 presents word clouds generated from fake (left) and real (right) news articles. Both classes prominently feature key political figures like *"trump"* and *"clinton"*, as well as frequent reporting terms such as *"said"* and *"one"*.


Fake News Word Cloud

Real News Word Cloud

Figure 4: Word Cloud for Fake or Real dataset

5 Results

This section presents the performance of the different fake news detection models across the two datasets used in the study. The focus is on both within-dataset accuracy and cross-dataset generalisability.

5.1 Within-dataset evaluation

5.1.1 ISOT dataset

Models trained and evaluated on the ISOT dataset using stratified 5-fold cross-validation (SCV) showed strong within-dataset performance. The table 1 below summarises the performance metrics—accuracy, precision, recall, and F1-score—across different combinations of features and models. Results are averaged across folds for each setup.

Table 1: Stratifed 5-Fold cross validation results on the ISOT dataset

Feature	Model	Acc	Pre	Rec	F1
BoW	AdaBoost	1.00	1.00	1.00	1.00
	Gradient Boosting	1.00	1.00	1.00	1.00
	Logistic Regression	1.00	1.00	1.00	1.00
	Neural Network	0.99	0.99	0.99	0.99
	Random Forest	1.00	1.00	1.00	1.00
TF-IDF	AdaBoost	0.99	0.99	0.99	0.99
	Gradient Boosting	1.00	1.00	1.00	1.00
	Logistic Regression	0.99	0.99	0.99	0.99
	Neural Network	0.99	0.99	0.99	0.99
	Random Forest	1.00	1.00	1.00	1.00
Word2Vec	AdaBoost	0.93	0.93	0.93	0.93
	Gradient Boosting	0.95	0.95	0.95	0.95
	Logistic Regression	0.96	0.96	0.96	0.96
	Neural Network	0.99	0.99	0.99	0.99
	Random Forest	0.96	0.96	0.96	0.96
BERT	AdaBoost	0.95	0.95	0.95	0.95
	Gradient Boosting	0.97	0.97	0.97	0.97
	Logistic Regression	0.99	0.99	0.99	0.99
	Neural Network	1.00	1.00	1.00	1.00
	Random Forest	0.97	0.97	0.97	0.97
Linguistic Cues	AdaBoost	0.95	0.95	0.95	0.95
	Gradient Boosting	0.96	0.96	0.96	0.96
	Logistic Regression	0.92	0.92	0.92	0.92
	Neural Network	0.96	0.96	0.96	0.96
	Random Forest	0.97	0.97	0.97	0.97

Across all feature representations, the best performances were obtained using Bag of Words and TF-IDF vectors, with both features achieving near perfect scores across all metrics. Models using Word2Vec and BERT embeddings also demonstrated high performance, especially with the neural network. Linguistic cues, while slightly less effective than word-based representations, still yielded strong performance across all models, with accuracy ranging from 92% to 97%.

5.1.2 Fake or Real dataset

Similarly, models were trained and evaluated on the Fake or Real dataset using stratified 5-fold cross-validation (SCV). The table 2 below summarises the performance metrics across different combinations of features and models.

Table 2: Stratifed 5-Fold cross validation results on the Fake or Real dataset

Feature	Model	Acc	Pre	Rec	F1
BoW	AdaBoost	0.86	0.86	0.86	0.86
	Gradient Boosting	0.89	0.89	0.89	0.89
	Logistic Regression	0.92	0.92	0.92	0.92
	Neural Network	0.92	0.92	0.92	0.92
	Random Forest	0.91	0.91	0.91	0.91
TF-IDF	AdaBoost	0.86	0.86	0.86	0.86
	Gradient Boosting	0.90	0.90	0.90	0.90
	Logistic Regression	0.92	0.92	0.92	0.92
	Neural Network	0.93	0.93	0.93	0.93
	Random Forest	0.91	0.91	0.91	0.91
Word2Vec	AdaBoost	0.82	0.82	0.82	0.82
	Gradient Boosting	0.88	0.88	0.88	0.88
	Logistic Regression	0.85	0.85	0.85	0.85
	Neural Network	0.91	0.91	0.91	0.91
	Random Forest	0.87	0.87	0.87	0.87
BERT	AdaBoost	0.82	0.82	0.82	0.82
	Gradient Boosting	0.85	0.85	0.85	0.85
	Logistic Regression	0.89	0.89	0.89	0.89
	Neural Network	0.91	0.91	0.91	0.91
	Random Forest	0.84	0.84	0.84	0.84
Linguistic Cues	AdaBoost	0.81	0.81	0.81	0.81
	Gradient Boosting	0.84	0.84	0.84	0.84
	Logistic Regression	0.80	0.80	0.80	0.80
	Neural Network	0.78	0.81	0.78	0.78
	Random Forest	0.85	0.85	0.85	0.85

Performance across all models and features was lower on the Fake or Real dataset than on the ISOT dataset, though still very good with models often achieving more than 80% accuracy. Among the tested representations, TF-IDF and Bag of Words again yielded the strongest results.

Overall, all models across all feature extraction approaches performed well when trained and tested on the same dataset.

5.2 Cross-dataset evaluation

Next, to assess the generalizability of fake news detection models, we conducted cross-dataset experiments by training models on one dataset and evaluating them on a different one. This setup aims to simulate real-world deployment scenarios, where a model may encounter fake news articles from a distribution different from its training data.

Table 3 shows the average accuracy of models trained on one dataset, as well as the average accuracy when testing these models on the remaining dataset. It demonstrates how well models generalise depending on the dataset on which they are trained.

Dataset	Baseline Avg Acc	Cross-dataset Avg Acc
ISOT	0.97	0.53
Fake or Real	0.87	0.56

Table 3: Baseline and Cross-dataset performance comparison

The comparison reveals a consistent drop in performance, regardless of the dataset used for training. Compared to the within-dataset performance, the accuracy of models trained on ISOT and tested on Fake or Real dropped by 44%, while the accuracy of models trained on Fake or Real and tested on ISOT dropped by 32% on average. This highlights a key limitation of many current fake news detection approaches : poor generalizability.

5.3 Generalizability by model type

Further analysis was conducted to assess whether any of the machine learning models demonstrated superior generalizability across a different dataset. Figure 5 shows that model selection does not notably influence generalisability, as all models achieve poor cross-dataset accuracy (ranging between 50% and 60%).

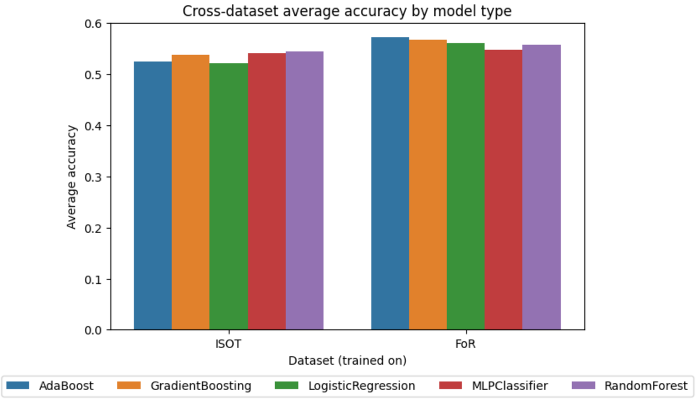


Figure 5: Cross-dataset performance by model

5.4 Generalizability by feature type

We finally examined whether any of the feature extraction methods generalize better than the others across a different dataset. Figure 6 shows that regardless of the feature method used, models fail to generalize effectively, although models using Word2Vec or Linguistic Cues appear to suffer less.

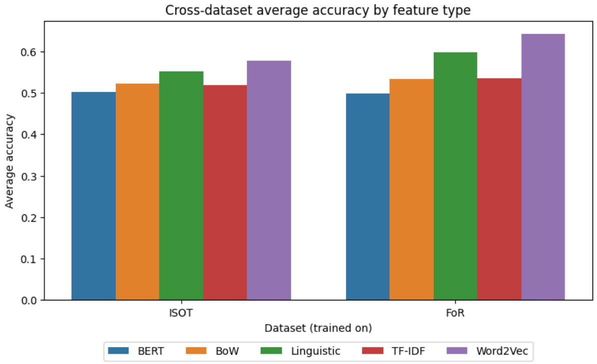


Figure 6: Cross-dataset performance by feature type

6 Conclusion

Our experiments confirm that fake news detection models tend to perform well on the datasets they are trained on, but their performance deteriorates notably when tested on different datasets. This lack of generalizability raises concerns about the deployment of such models in real-world scenarios where data distributions vary. Despite testing several types of feature representations—including Bag of Words, TF-IDF, Word2Vec, BERT, and linguistic features—no approach demonstrated strong cross-dataset robustness. These findings support those of Hoy and Koulouri (2022), highlighting the ongoing need for strategies that enhance the generalizability of fake news detection systems.

Code

The code used to carry out this work is publicly available at : `https://github.com/lu-julia/fake-news-detection`

References

Gravanis, G., Vakali, A., Diamantaras, K., and Karadais, P. (2019). Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, 128:201–213.

Hoy, N. and Koulouri, T. (2022). Exploring the generalisability of fake news detection models. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5731–5740.

Prachi, N. N., Habibullah, M., Rafi, E. H., Alam, E., and Khan, R. (2022). Detection of fake news using machine learning and natural language processing algorithms. *Journal of Advances in Information Technology*, 13.

Sastrawan, I. K., Bayupati, I., and Arsa, D. M. S. (2022). Detection of fake news using deep learning cnn–rnn based methods. *ICT Express*, 8(3):396–408.