Generalization of Fake News Detection Models

Chappuis Maxime, Deroo Marc
ENSAE
maxime.chappuis@ensae.fr, marc.deroo@ensae.fr

Abstract

Fake news detection is a key challenge in Natural Language Processing (NLP), especially with the rapid spread of misinformation. This paper replicates and analyzes the findings of [1] by evaluating the generalization of six traditional machine learning models (Logistic Regression, SVM, Random Forest, Gradient Boosting, AdaBoost, and a Neural Network) across different preprocessing methods and datasets. Five preprocessing techniques are compared, including full text normalization for Bag-of-Words (BoW) and TF-IDF, as well as lighter preprocessing for Word2Vec and BERT. Results show that while models achieve near-perfect accuracy on the same dataset, their performance drops significantly when tested on the external dataset, highlighting challenges in generalization.

1 Introduction

The rapid spread of fake news has significant societal and political consequences. With growing reliance on digital media, fake news detection has become a critical task in Natural Language Processing (NLP), aiming to curb misinformation. Various models have been proposed, ranging from traditional machine learning approaches to deep learning-based methods.

A key issue in fake news detection is the **lack of generalization** of classification models. While many models perform well when trained and tested on the same dataset, they often fail to maintain similar performance when tested on unseen data from different domains [1]. This raises concerns about the models' real-world applicability.

Objective: This paper replicates the findings of [1] by systematically evaluating the generalization ability of fake news detection models. Specifically, we:

- Implement six traditional machine learning models (Logistic Regression, SVM, Random Forest, Gradient Boosting, AdaBoost, and a Neural Network).
- Compare five preprocessing techniques: full text normalization for Bag-of-Words (BoW) and TF-IDF, as well as lighter preprocessing for Word2Vec, BERT and Linguistic Cues.
- Train models on a specific dataset and test them on an external dataset.
- Analyze whether models using linguistic features (BoW, TF-IDF) generalize better than those based on word embeddings (Word2Vec, BERT).

This study compares model performance across datasets and preprocessing methods to provide insights into the factors that affect generalization in fake news detection.

2 Brief State-of-the-Art

Fake news detection in NLP has led to a variety of classification approaches, broadly divided into two categories: traditional machine learning models and feature extraction methods.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2.1 Traditional Machine Learning Approaches

Early fake news detection methods focused on manually engineered linguistic features, such as TF-IDF, bag-of-words (BoW), and n-grams. These features were used with classifiers like **Support Vector Machines (SVM)**, **Random Forest (RF)**, **Logistic Regression**, **Gradient Boosting**, and **AdaBoost** [3]. While these methods are interpretable and computationally efficient, they often fail to capture deeper semantic meaning and may overfit dataset-specific patterns.

The study by Hoy and Koulouri [1] highlights that while traditional models achieve high accuracy on the ISOT dataset, their performance drops significantly when applied to an external dataset, indicating a lack of robustness to domain shifts.

2.2 Challenges in Model Generalization

A major challenge in fake news detection is **domain dependence**, where models trained on one dataset often show significant performance drops when tested on another. This suggests that they rely on dataset-specific patterns rather than universal linguistic features [1], raising concerns about the real-world applicability of these models.

To improve generalization, previous research has explored various preprocessing strategies and feature representations. Some models use statistical text representations, while others leverage word embeddings to capture semantic relationships between words.

2.3 Analyzing the Impact of Linguistic vs. Word Embedding Representations

The choice of text representation is crucial for model generalization. Fake news classification models rely on two main feature types: **linguistic-based features** and **word embeddings**. Understanding their differences is vital for designing models that perform well across different datasets.

Linguistic-Based Representations (TF-IDF, Bag-of-Words): Traditional models, such as SVM, Random Forest, and Gradient Boosting, rely on statistical representations like:

- **Bag-of-Words (BoW)**: Represents text as a set of individual words, ignoring word order.
- **TF-IDF (Term Frequency - Inverse Document Frequency)**: Weighs words based on their importance across documents, reducing the impact of common words.

These methods are efficient but may struggle to capture deeper semantic relationships.

Word Embedding-Based Representations (Word2Vec, BERT): Recent approaches use embeddings to capture semantic meaning:

- **Word2Vec**: Represents words as dense vectors based on co-occurrence patterns.
- **BERT**: Generates word representations based on surrounding context, improving understanding.

These models capture deeper semantic relationships but may be more susceptible to domain shifts and dataset-specific cues.

Generalization Issue: Linguistic-based models tend to generalize better, while word embeddings like Word2Vec and BERT often overfit to dataset-specific characteristics.

Key Question: How do different feature representations affect the ability of fake news detection models to generalize across datasets?

3 Datasets

3.1 ISOT Fake News Dataset

The ISOT Fake News dataset [2] is a widely-used benchmark for fake news detection, containing **44,898** articles, with **23,481** fake and **21,417** real articles. The dataset spans multiple domains, such as politics and world news, and includes articles from both unreliable and reputable sources. The average article length is **405 words** with a standard deviation of **351 words**.

For our experiments, the dataset is split into training and testing subsets using an **80-20** stratified split to preserve class distribution.

3.2 Fake or Real News Dataset (FoR)

To assess generalization, we use the **Fake or Real News** dataset (FoR) from Kaggle. This dataset, containing **6,296** articles (**3,171 real** and **3,125 fake**), serves as an independent evaluation set, allowing us to assess model performance on unseen data from a different distribution.

By comparing model performance on both ISOT and FoR, we aim to identify potential overfitting to ISOT and assess the robustness of different feature extraction methods (TF-IDF, Word2Vec, BERT) across datasets with varied linguistic structures.

4 Experiment Proposal

The goal of this experiment is to evaluate the effectiveness of various machine learning models in detecting fake news, with a focus on their ability to generalize across different datasets. We will compare six models, including traditional machine learning methods such as Logistic Regression, Support Vector Machines, Random Forest, Gradient Boosting, AdaBoost, and a Neural Network. These models represent a range from simpler, interpretable algorithms to more complex ones capable of capturing non-linear relationships.

The comparison between these models is crucial to understand how simpler models perform in relation to more advanced ones, and whether the added complexity of neural networks leads to better performance in fake news detection. Additionally, the models will be evaluated based on multiple metrics: accuracy, precision, recall, F1-score, and AUC-ROC, which will allow us to assess their classification performance.

5 Data Retrieval, Formatting, and Statistical Description

In this study, we utilize two datasets: ISOT and FakeOrReal (FoR), both containing news articles aimed at detecting misinformation. The ISOT dataset contains 44,898 articles, while the FakeOrReal dataset (FoR) contains 6,335 articles. Both datasets are categorized into fake news (label 0) and real news (label 1).

5.1 Dataset Structure

The datasets contain the following columns:

- `title`: The title of the news article.
- `text`: The main content of the news article.
- `label`: The label indicating whether the article is fake (0) or real (1).

5.2 Descriptive Statistics

Class Distribution:

The class distributions for the 2 datasets are quite similar with nearly half of negative and half of positive label. The number of label are as follows:

ISOT Dataset

Label	Count
0	$23,481$
1	$21,417$

FoR Dataset

Label	Count
$FAKE$	$3,164$
$REAL$	$3,171$

Table 1: Class Distribution Comparison between ISOT and FoR Datasets

Text Length Statistics:

The length of the articles in both datasets varies significantly. Indeed, we see that the average number of words in FoR Dataset is 776 whereas it is 405 for ISOT (nearly 2 times more). It will be interesting to see if the models tend to generalize on articles that don't have the same length. Below are the descriptive statistics for text lengths:

ISOT Dataset

Statistic	Value
Count	44,898
Mean	405.28
Std Dev	351.27
Min	0
25% Quantile	203
50% Quantile (Median)	362
75% Quantile	513
Max	8,135

FoR Dataset

Statistic	Value
Count	6,335
Mean	776.30
Std Dev	854.33
Min	0
25% Quantile	289
50% Quantile (Median)	597
75% Quantile	1,024
Max	20,891

Table 2: Text Length Statistics Comparison between ISOT and FoR Datasets

Missing Data:

Both datasets have no missing values across any of the columns.

5.3 Word Cloud

A word cloud was generated to visually highlight the most frequent words in the articles, providing insight into the content of fake versus real news. We can see that words like Trump, Clinton or the White House appear for the 2 dataset in Fake News and True News. Below are the word clouds for both datasets:

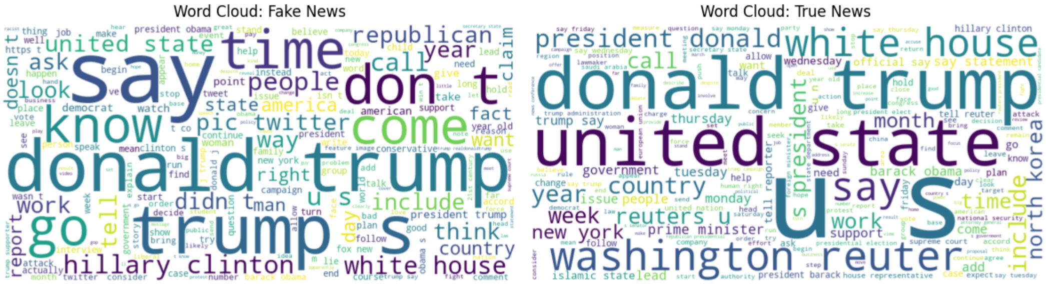

Figure 1: Word Cloud for ISOT Dataset

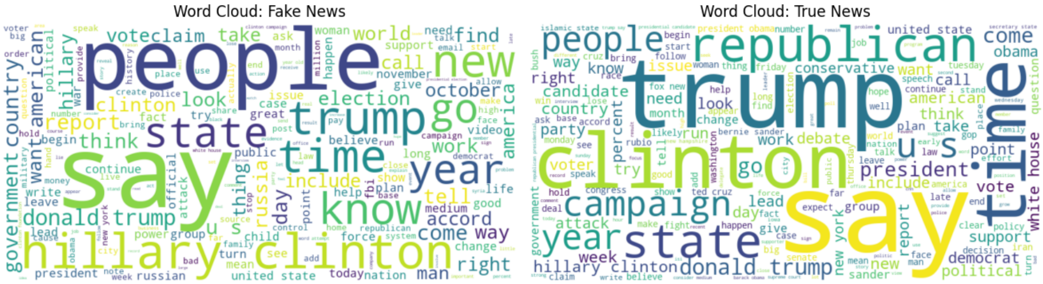

Figure 2: Word Cloud for FoR Dataset

6 Experiments with the proposed model

6.1 Experimental Setup

To evaluate the generalizability of fake news detection models, we implement the six classification models used in [1]: Logistic Regression, Support Vector Machines (SVM), Random Forest, Gradient Boosting, AdaBoost, and a Neural Network (MLPClassifier). For the first step we train and test on ISOT, the we train and test on FoR Dataset. The second step has been to use cross Datasets to test the generality on another dataset. Indeed, we train on ISOT and then test on FoR Dataset and then we train FoR Dataset and test on ISOT.

6.1.1 Preprocessing Methods

As outlined in [1], we apply different levels of preprocessing based on the type of feature extraction method used:

- **Bag-of-Words (BoW) and TF-IDF**: Full preprocessing, including lowercasing, lemmatization, stopword removal, punctuation removal, and elimination of extra whitespace.
- **Word2Vec, BERT and Linguistic Cues**: Light preprocessing, including lowercasing, removal of URLs and Twitter handles, and spell-checking.

This ensures that models relying on explicit feature representations (BoW, TF-IDF) do not introduce noise, while embedding-based models (Word2Vec, BERT) retain as much contextual information as possible.

6.1.2 Evaluation Metrics

Following [1], we evaluate each model using the following metrics:

- **Accuracy**: The proportion of correctly classified samples.
- **Precision**: The ratio of correctly predicted positive samples to the total predicted positive samples.
- **Recall**: The ratio of correctly predicted positive samples to the total actual positive samples.
- **F1-score**: The harmonic mean of precision and recall.

6.2 Results on ISOT Dataset

Table 3: Comparison of ISOT SCV results from [1] and our experiment.

Feature	**Model**	**Results from [1]**				**Our Results**			
		Acc	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
Count	AdaBoost	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00
	Gradient Boosting	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99
	Logistic Regression	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00
	Neural Network	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	Random Forest	0.98	0.98	0.97	0.98	1.00	1.00	1.00	1.00
	SVM	0.90	0.91	0.91	0.90	1.00	1.00	1.00	1.00
TF-IDF	AdaBoost	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00
	Gradient Boosting	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00
	Logistic Regression	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99
	Neural Network	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	Random Forest	0.97	0.97	0.97	0.97	1.00	1.00	1.00	1.00
	SVM	0.96	0.96	0.96	0.96	0.99	0.99	0.99	0.99
Word2Vec	AdaBoost	0.94	0.94	0.93	0.93	0.94	0.94	0.94	0.94
	Gradient Boosting	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
	Logistic Regression	0.96	0.96	0.96	0.96	0.95	0.95	0.95	0.95
	Neural Network	0.97	0.95	0.96	0.95	0.99	0.99	0.99	0.99
	Random Forest	0.94	0.94	0.94	0.94	0.95	0.95	0.95	0.95
	SVM	0.92	0.92	0.92	0.92	0.97	0.97	0.97	0.97
BERT	AdaBoost	0.96	0.96	0.96	0.96	0.94	0.94	0.94	0.94
	Gradient Boosting	0.97	0.97	0.97	0.97	0.96	0.96	0.96	0.96
	Logistic Regression	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	Neural Network	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99
	Random Forest	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
	SVM	0.95	0.95	0.95	0.95	0.98	0.98	0.98	0.98
Linguistic Cues	AdaBoost	0.95	0.95	0.94	0.94	0.95	0.95	0.95	0.95
	Gradient Boosting	0.94	0.94	0.94	0.94	0.96	0.96	0.96	0.96
	Logistic Regression	0.90	0.90	0.90	0.90	0.91	0.91	0.91	0.91
	Neural Network	0.89	0.84	0.88	0.86	0.94	0.94	0.94	0.94
	Random Forest	0.94	0.94	0.94	0.94	0.97	0.97	0.97	0.97
	SVM	0.51	0.70	0.53	0.39	–	–	–	–

Note: The combination of Linguistic Cues with SVM was excluded from the experiments due to excessive computational time, which made training infeasible on our available hardware.

6.3 Results on Fake or Real Dataset

Table 4: Comparison of FoR SCV results from [1] and our results on Fake or Real News (FoR).

Feature	Model	FoR Results (Hoy et al. 2022)				Our Results on FoR			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
Count	AdaBoost	0.86	0.87	0.86	0.86	0.88	0.88	0.88	0.88
	Gradient Boosting	0.89	0.89	0.89	0.89	0.90	0.90	0.90	0.90
	Logistic Regression	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
	Neural Network	0.93	0.94	0.93	0.94	0.92	0.92	0.92	0.92
	Random Forest	0.84	0.85	0.84	0.84	0.91	0.91	0.91	0.91
	SVM	0.85	0.87	0.85	0.85	0.88	0.88	0.88	0.88
TF-IDF	AdaBoost	0.86	0.86	0.86	0.86	0.87	0.87	0.87	0.87
	Gradient Boosting	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
	Logistic Regression	0.91	0.92	0.91	0.91	0.91	0.91	0.91	0.91
	Neural Network	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	Random Forest	0.83	0.84	0.83	0.83	0.92	0.92	0.92	0.92
	SVM	0.90	0.91	0.90	0.90	0.92	0.92	0.92	0.92
Word2Vec	AdaBoost	0.85	0.85	0.85	0.85	0.84	0.84	0.84	0.84
	Gradient Boosting	0.89	0.89	0.89	0.89	0.87	0.87	0.87	0.87
	Logistic Regression	0.87	0.87	0.87	0.87	0.85	0.85	0.85	0.85
	Neural Network	0.86	0.80	0.85	0.82	0.91	0.91	0.91	0.91
	Random Forest	0.84	0.85	0.84	0.84	0.86	0.86	0.86	0.86
	SVM	0.89	0.89	0.89	0.89	0.87	0.87	0.87	0.87
BERT	AdaBoost	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86
	Gradient Boosting	0.89	0.89	0.89	0.89	0.88	0.88	0.88	0.88
	Logistic Regression	0.91	0.91	0.91	0.91	0.90	0.90	0.90	0.90
	Neural Network	0.90	0.88	0.89	0.88	0.92	0.92	0.92	0.92
	Random Forest	0.83	0.84	0.83	0.83	0.88	0.88	0.88	0.88
	SVM	0.88	0.88	0.88	0.88	0.91	0.91	0.91	0.91
Linguistic Cues	AdaBoost	0.82	0.82	0.82	0.82	0.81	0.81	0.81	0.81
	Gradient Boosting	0.84	0.84	0.84	0.84	0.83	0.83	0.83	0.83
	Logistic Regression	0.81	0.81	0.81	0.81	0.79	0.79	0.79	0.79
	Neural Network	0.79	0.72	0.71	0.71	0.73	0.79	0.72	0.71
	Random Forest	0.83	0.83	0.83	0.83	0.84	0.84	0.84	0.84
	SVM	0.54	0.72	0.54	0.42	–	–	–	–

Note: The combination of Linguistic Cues with SVM was excluded from the experiments due to excessive computational time, which made training infeasible on our available hardware.

6.4 Results on Cross Datasets

In this experiment, we evaluate the performance of the fake news detection models on cross-dataset settings. Specifically, we train the models on the ISOT dataset and test them on the Fake or Real (FoR) dataset, and vice versa, as done in the original paper. This allows us to assess the generalization ability of the models across different distributions of news articles.

Table 5 presents the baseline performance of the models when trained and tested on the same dataset, as well as the performance when trained on one dataset and tested on the other (cross-dataset evaluation). As expected, the cross-dataset accuracy is significantly lower than the in-dataset performance, highlighting the challenge of generalizing to unseen datasets.

Dataset	Baseline Acc.	Cross-Dataset Avg. Acc.	Our Baseline Acc.	Our Cross Acc.
ISOT	0.95	0.52	0.97	0.59
Kaggle FoR	0.86	0.52	0.88	0.61

Table 5: Baseline and Cross-Dataset Performance Comparison

7 Analysis of Results and Conclusion

The results obtained in this study show that our models perform similarly to the ones reported in the article [1], with comparable orders of magnitude for classification accuracy. However, there are some differences in the results, which can be attributed to several factors. First, the article lacks specific details regarding the models used, making it difficult to exactly replicate their setup. For instance, in our study, we used the MLPClassifier for the neural network, which may differ from the architecture used in the original work. Similarly, for the Support Vector Machine (SVM), we opted for the SVC model with a linear kernel, whereas other variations like linearSVC could have been used in the article.

Due to the lack of detail on the specific configurations of the models, we chose what we believe to be the most basic versions of the models, with default hyperparameters, without any optimization. This choice of default settings likely contributed to the small differences observed between our results and those of the original paper. The models tested in this study, while following standard practices, may not have been fine-tuned to their optimal settings, which could explain the discrepancies.

In terms of performance, our findings align with the article's conclusions: the models perform well on the same dataset, but their performance drops significantly when evaluated on a new dataset, with accuracy approaching random chance. This confirms the generalization issue highlighted in the original study, where models that perform well on a familiar dataset struggle when applied to new, unseen data. This result underscores the challenge of developing fake news detection models that can generalize across different domains or sources.

In conclusion, while our results are consistent with the general trends observed in the original paper, the differences in specific model configurations and the performance on new datasets point to the importance of model optimization and careful attention to the fine details in machine learning pipelines for fake news detection.

Contributions

Although the project was conducted in close collaboration, with many tasks carried out jointly, we provide here a general overview of how the work was distributed between the two contributors:

- **Marc Deroo** was primarily responsible for the implementation and experimentation on the ISOT dataset.
- **Maxime Chappuis** focused on the implementation and experimentation on the FakeOrReal (FoR) dataset.
- The cross-evaluation experiments, where models trained on one dataset were tested on the other, were carried out jointly.
- Regarding the writing of the paper:
  - The *state-of-the-art* section was written by Maxime Chappuis.
  - The *data analysis and description* section was written by Marc Deroo.
  - The sections concerning the models used, the experimental setup, and the analysis of results were written collaboratively, based on the results obtained by each contributor in their respective notebooks.

We would like to emphasize that regular communication and mutual feedback played an essential role throughout the entire project.

References

[1] Nathaniel Hoy and Theodora Koulouri. Exploring the generalisability of fake news detection models. *Brunel University Research*, 2022.

[2] ISOT Research Team. Isot fake news dataset, 2020. Accessed: 2025-03-09.

[3] Xinyi Zhou and Reza Zafarani. A survey of fake news detection: Methods, evaluation, and challenges. *ACM Computing Surveys*, 53(5):1–40, 2020.