Fake news detection

Catherine Berleur
ENSAE
ML for NLP

Abstract

This study addresses the task of fake news detection using machine learning approaches on textual data. We explore both traditional models—including Logistic Regression, Naive Bayes, and Support Vector Machines—and a deep learning approach based on BERT. Using the ISOT Fake News dataset, our findings show that classical models perform strongly, achieving F1-scores above 0.94. Furthermore, we successfully fine-tuned a BERT model, which outperformed all classical approaches with near-perfect classification results (F1-score = 1.00). While the BERT model demonstrated excellent accuracy, it required significantly more computational resources and training time. This highlights the trade-off between model complexity and practical deployment constraints in resource-limited environments.

1 Research question and state of the art

The proliferation of false information, or fake news, represents a major challenge in today's digital society. With the rise of social networks, the rapid dissemination of misleading content has significant repercussions on public perception, trust in institutions and social cohesion [**Wahab2025**]. In their study, Hoy et al (2022) [**Hoy2022**] explore various machine learning techniques for detecting fake news, focusing on the use of methods such as logistic regression, random forests and support vector machines. Their research highlights the varying effectiveness of these approaches, depending on the characteristics of the data and the pre-processing techniques used.

In the same vein, other work, such as that by Shu et al. (2017) [**Shu2017**], highlights the importance of integrating contextual information, such as the source of the information and distribution patterns on social networks, to improve the detection of fake news. They propose a multidimensional approach combining textual content, metadata and social interactions.

Some researchers, such as Zhou and Zafarani (2020) [**Zhou2020**], propose a typology of fake news detection methods into three main categories: content-based, context-based and user-driven approaches. They too point out that, although content-based methods are widely used, they have limitations when applied to a variety of contexts without taking external factors into account.

The emergence of pre-trained language models, such as BERT [**Devlin2019**], has opened up new perspectives for the detection of fake news. These models, capable of understanding the bidirectional context of words, have demonstrated superior performance in a variety of natural language processing tasks. However, their application to the detection of fake news requires specific adaptation to the datasets concerned.

These felt advances and the state of the art lead us to ask, to what extent does the integration of pre-trained language models, such as BERT, improve fake news detection over traditional machine learning methods, taking into account performance, robustness and generalizability on varied datasets?

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

2 Empirical approach to evaluate the model

In order to evaluate the performance of the models on the false information detection task, we have designed an experimental approach based on two axes: a comparative basis based on classical models, and a fine-tuning extension of a BERT-type pre-trained language model.

The aim of our empirical protocol is to answer the following question: to what extent do pre-trained language models on large masses of textual data (such as BERT) outperform traditional supervised classification approaches on the fake news detection task? Is this improvement really significant, given the higher complexity and computational power required for the BERT model?

This question follows on from the findings established in the state of the art, notably by Wahab et al. (2025) [**Wahab2025**], who show that BERT's effectiveness on semantic classification tasks remains superior to that of conventional models, but that this is highly dependent on preprocessing, domain and data size.

We propose the following steps to carry out the evaluation:

1. **Data preparation** : The ISOT dataset, consisting of two separate files containing articles classified as "fake" or "true", was cleaned, merged, labeled and linguistically pre-processed (removal of punctuation, lemmatization, removal of stopwords, etc.). A final balanced dataset was constructed, and a .csv save was made for reproducibility.
2. **Evaluation using traditional models** : We implemented a classical text processing pipeline with TF-IDF vectorization, then trained three supervised models: a logistic regression, a Naive Bayes classifier, and a linear SVM. The performance of each model was measured by cross-validation and $classification_report$ from `scikit-learn` on an independent test set.
3. **Fine-tuning the BERT model**: In a second phase, we used the pre-trained `bert-base-uncased` model, which we fine-tuned on our corpus. Training was carried out without the use of the Trainer module, to ensure compatibility with environments where the latest versions of Transformers are not available. Evaluation was performed using standard metrics (accuracy, precision, recall, F1-score), calculated from the model's predictions on the validation set.
4. **Comparison and analysis** : The final step is to compare the performance of conventional models with that of the fine-tuned BERT model, in order to judge the empirical contribution of transfer learning in the context of automated false information detection.

For the purposes of comparison, we use the usual evaluation indicators for binary classification:

1. Accuracy: proportion of correct predictions
2. Precision: ability to avoid false positives
3. Recall: ability to detect all true positives
4. F1-score: harmonic mean between precision and recall

All these scores will be compared between models, with a particular focus on the "fake" class, often under-represented in real-life contexts.

3 Data

The empirical analysis begins with an in-depth exploration of the ISOT dataset, made up of 44,898 press articles, divided into 23,481 fake news articles and 21,417 true news articles. This relatively balanced distribution means that supervised learning can be envisaged without the need for resampling techniques to compensate for class imbalances.

Each article is described by several variables: a title, a text, a subject, a publication date and a binary label indicating the veracity of the article (0 for fake, 1 for true).

Analysis of article length, measured in words, reveals considerable heterogeneity. The average is 405 words per article, with a standard deviation of 351. An asymmetrical distribution can be observed:

while the majority of texts are under 600 words, some reach over 8,000 words. This wide variation calls for normalization or truncation when training models such as BERT, whose maximum sequence length is limited.

- **Min** : 0 word
- **Median** : 362 words
- **75th percentile** : 513 words
- **Maximum** : 8135 words

This exploration enables us to adjust the parameters of future models, taking into account sequence length constraints and the pre-processing required for efficient classification.

4 Main results

We evaluated three classical supervised classification models on TF-IDF representations of text: Logistic Regression, Naive Bayes Multinomial and Linear SVM. Performance was measured in terms of precision, recall, F1-score and overall accuracy. The results obtained are summarized in the following table:

Model	Mean accuracy	Mean Recall	Mean F1-score	Accuracy
Logistic Regression	0.987	0.987	0.987	0.987
Naive Bayes	0.931	0.931	0.931	0.931
Linear SVM	**0.994**	**0.994**	**0.994**	**0.994**

Table 1: Performance of conventional models on the ISOT dataset

The Linear SVM model clearly outperformed the others, achieving an accuracy of 99.4%, and an average precision, recall and F1-score of 0.994, indicating excellent generalization capability on the data tested. This result confirms the well-known effectiveness of SVMs on TF-IDF vectors in text classification tasks.

Logistic regression also performed very well, with an accuracy of 98.7%, making it an excellent compromise between performance and computation time. Its balanced results on both classes demonstrate the robustness of the model.

In contrast, the Naive Bayes model, although often used as a baseline for text classification tasks, performs less well here (accuracy 93.1%, average F1-score 0.931). This can be explained by the assumptions of strong independence between the variables on which it is based, which are ill-suited to the actual structure of natural language.

In a second phase of experimentation, we successfully fine-tuned a pre-trained BERT model (bert-base-uncased) on our cleaned ISOT dataset, using the PyTorch and Hugging Face Transformers libraries. The model was trained over two epochs with a batch size of 8, and evaluation was performed on 20% of the data reserved as a validation set.

The results obtained are remarkable: the model achieved nearly perfect performance with a precision, recall and F1-score of 1.00 for both the Fake and True classes. The confusion matrix confirms this, showing only 5 misclassified examples out of 8837. Specifically, only 4 false positives and 1 false negative were observed.

Compared to classical models tested earlier (Logistic Regression, Naive Bayes, and Linear SVM), which achieved F1-scores between 0.94 and 0.95, BERT clearly outperforms them in terms of accuracy and robustness.

This significant gain highlights the ability of transformer-based architectures to better capture the contextual and semantic richness of the text, especially in a binary classification task such as fake news detection. However, this performance comes at the cost of much higher computational demand. Training required over two hours using the Datalab Onyxia platform (SSPCloud) with GPU acceleration enabled.

While we did not compute the ROC and AUC metrics in time to include them in this version of the report, these would constitute valuable next steps to further investigate the model's behavior under threshold variation and support a more nuanced performance comparison.

5 Bibliography

[1] Devlin, J., Chang, M.-W., Lee, K.& Toutanova, K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, pp. 4171–4186. https://aclanthology.org/N19-1423/

[2] Hoy, N.& Koulouri, T. (2022) Exploring the Generalisability of Fake News Detection Models. In 2022 IEEE International Conference on Big Data (Big Data). https://doi.org/10.1109/BigData55660.2022.10020583

[3] Shu, K., Sliva, A., Wang, S., Tang, J.& Liu, H. (2017) Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter, 19(1), pp. 22–36. https://doi.org/10.1145/3137597.3137600

[4] Wahab, F., Khan, I.& Shankar, A. (2025) Fake News Detection Using Machine Learning Techniques. Journal of Computer Networks, Architecture and High Performance Computing, 7, pp. 440–461. https://doi.org/10.47709/cnahpc.v7i2.5717

[5] Zhou, X.& Zafarani, R. (2020) A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. ACM Computing Surveys (CSUR), 53(5), pp. 1–40. https://doi.org/10.1145/3395046