Multi-Label Fake News Detection with NLP using OBSINFOX [2] dataset

Victoire Ahyerre
victoire.ahyerre@ensae.fr
code realised with Etienne Selles

1 Introduction

The proliferation of fake news has emerged as a critical challenge in the digital age. While the automatic detection of fake news has been the focus of much research, existing models often reduce the task to a binary classification problem, labeling news articles as either "fake" or "not fake." This simplification overlooks the inherent complexity of fake news, which can involve elements such as factual inaccuracies, stylistic exaggeration or omission of sources.

The problem adressed in this work is the following : Can fake news be more effectively detected when modeled as part of a multi-label learning task rather than a standalone binary label? Most traditional approaches fail to account for the diversity of cues that characterize fake news. Fake news is rarely defined by falsity alone; it often exhibits subjective language, unverified information, and rhetorical devices like insinuation or offbeat titles.

To explore this question, we base cles annotated by eight experts across 11 complementary binary labels. These labels account both for factual and stylistic dimensions, including Fake News, False Information, Opinion, Subjective, and others. By jointly learning these labels, we hypothesize that the model can better distinguish fake news from legitimate information.

2 State of the Art

2.1 Fake News Datasets

The automatic detection of fake news has been extensively explored in recent years, often framed as a binary classification problem. Several benchmark datasets such as LIAR [7] have been used to train models that classify news articles as fake or real. However, these datasets largely ignore style and context, instead providing labels such as *true*, *half-true*, or *false*. Moreover, they are limited to the English language, restricting their applicability to french fake news.

To address these limitations, Icard et al. introduced OBSINFOX in 2024 [2], a French-language dataset composed of 100 news articles from 17 websites flagged as unreliable by fact-checking sources such as Conspiracy Watch and NewsGuard. Each article is annotated with 11 binary labels capturing both factual aspects (e.g., *False Information*, *Sources Cited*) and stylistic dimensions (e.g., *Exaggeration*, *Offbeat Title*, *Subjective*).

The dataset was annotated by eight raters, enabling detailed analysis of inter-annotator agreement and label correlation. Notably, the authors found strong associations between the *Fake News* label and others such as *Exaggeration*, and *False Information*, which supports the use of a multi-label learning framework for improved detection performance.

2.2 Models

Many existing approaches for the detection of fake news have been presented using traditional machine learning models ([1], [5], [6]). BERT-based deep learning approach (FakeBERT, [3] are also emerging.

In the context of French NLP, a prominent model is CamemBERT [4], a RoBERTa-based transformer pretrained on the OSCAR corpus—a large, multilingual web crawl filtered for quality. CamemBERT has demonstrated state-of-the-art results across a variety of French-language tasks.

Thanks to its contextual embeddings and strong transfer capabilities, CamemBERT is well suited for low-resource classification tasks, including those requiring multi-label or multi-task learning. Its architecture allows it to capture dependencies among labels, such as those in OBSINFOX.

3 Data

The dataset used in this project is based on the metadata provided in the OBSINFOX corpus [2], which contains URLs and annotations for 100 French news articles from 17 sources flagged as unreliable.

3.1 Label analysis

To better understand the distribution of labels across the dataset, we computed the mean annotation score for each label. Labels and their signification are reported in Table 1.

Label	Signification
Fake News	The article describes at least a false or exaggerated fact.
Places, Dates, People	The article mentions at least one place, date, or person.
Facts	The article reports at least one fact, i.e., a state of affairs or event, which may be true or false.
Opinions	The article expresses at least one opinion.
Subjective	The article contains more opinions than facts.
Reported Information	The information is reported by another person or source and is not directly endorsed.
Sources Cited	The article cites at least one source for at least one fact.
False Information	The article contains at least one false fact.
Insinuation	The article suggests a certain reading of a fact, without saying so explicitly.
Exaggeration	The article describes a real fact with exaggeration.
Offbeat Title	The article has a misleading headline not accurately reflecting the content.

Table 1: Labels and their signification

To investigate relationships between labels, we computed the Pearson correlation matrix over the mean label scores. As seen in Figure 1, *Fake News* shows strong positive correlations with *Subjective*, *Insinuation*, *Exaggeration*, and *False Information*, supporting the multi-label hypothesis that these auxiliary signals are predictive of misinformation. It also presented a small negative correlation with *Sources Cited*.

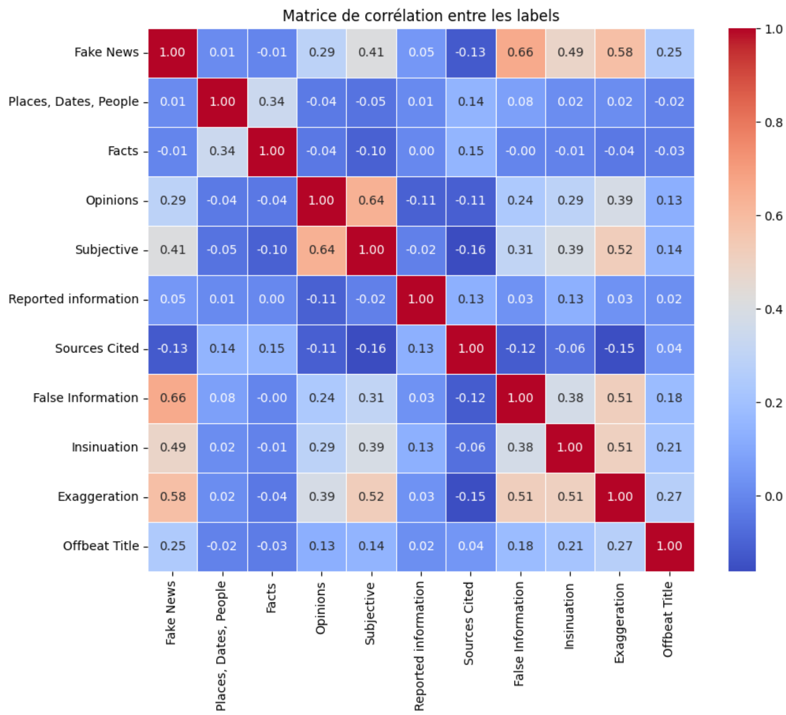

Figure 1: Correlation matrix between annotated labels.

As shown in Figure 2, elements such as *Reported Information* and *Offbeat title* are less frequently present across articles, while *Facts* and *Places, Dates, People* appear almost every time. Inter-annotator agreement is generally higher for more concrete labels (e.g., *Places, Dates, People*, *Offbeat title*) and lower for more subjective ones (e.g., *Fake News*, *Insinuation*).

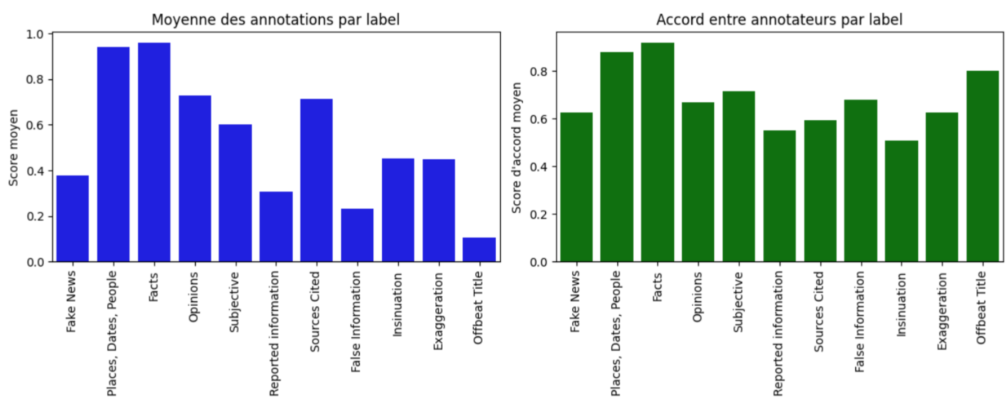

Figure 2: Left: Average annotation scores per label. Right: Inter-annotator agreement per label.

3.2 Data collection : articles' texts

Since the article texts themselves are not included in the dataset, we implemented a web scraping pipeline using `requests` and `BeautifulSoup` to extract the raw content from the URLs. In total, we successfully retrieved the full text for 92 articles. The last 8 articles were left out of our study.

3.3 Data preprocessing

Each article in the dataset is annotated across 11 labels, including both factual (*False Information*, *Sources Cited*) and stylistic (*Exaggeration*, *Subjective*, *Offbeat Title*) cues. Annotations were provided by 8 raters. We computed the mean score for each label across annotators to create soft-label targets in the range $[0, 1]$. The labels were kept soft during the training. The ground truth of the label, used to calculate the accuracy, is considered to be the binarized value of the label using a threshold of 0.5.

The final dataset consists of 92 articles, each annotated with 11 mean-label values and ready for input into a multi-label classification model.

4 Model

4.1 Architecture and Method

We used the `camembert-base` model as a feature extractor with pretrained weights and trained its classifier only (last layer). We used the `BCEWithLogitsLoss` loss function, which is well suited for multi-label learning.

The classifier is fine-tuned using the annotated OBSINFOX dataset. During training, we keep the soft label values (averaged across 8 annotators) in the $[0, 1]$ range as targets, while during evaluation, we binarize predictions using a threshold of 0.5.

4.2 Preprocessing and Tokenization

Articles are tokenized using the CamemBERT tokenizer. We split the dataset into training and testing subsets (80/20 ratio).

4.3 Training Procedure

We train the model for 30 epochs using a batch size of 16 and a learning rate of 5×10^{-5}. Optimization is performed using AdamW. During each epoch, we compute the average training loss to monitor convergence.

4.4 Labels subsets

We also experimented with different sets of target labels to compare single-label, subset multi-label, and full multi-label configurations:

- Training on `Fake News` only (baseline).
- Training on `Fake News, Opinions, Subjective, Sources Cited` to focus on labels most relevant to detect misinformation according to the article [2].
- Training on all 11 available labels to exploit the full richness of the annotation scheme.

4.5 Evaluation

At test time, we evaluate the model using per-label accuracy, computed by thresholding the sigmoid outputs at 0.5 and comparing them to the binarized ground truth. We also compute the test loss over the entire dataset and visualize the training loss over epochs to ensure proper convergence.

5 Results

5.1 Single-label Classification: *Fake News* Only

In the first experiment, we trained the model to predict only the *Fake News* label.The model achieved an accuracy of **78.95%** on this label, with a test loss of 0.5480. This baseline demonstrates the capacity of CamemBERT to capture patterns indicative of misinformation in French news content.

5.2 Multi-label Classification with 4 Labels

We then extended the task to a multi-label setup with four key indicators: *Fake News*, *Opinions*, *Subjective*, and *Sources Cited*. The hypothesis was that incorporating these complementary dimensions would help the model better contextualize the fake news prediction.

The model achieved higher accuracy on the *Fake News* label (84.21%) and also performed well on *Opinions* (89.47%), *Subjective* (84.21%), and *Sources Cited* (68.42%), suggesting that jointly learning these dimensions improves predictive power.

5.3 Full Multi-label Classification (11 Labels)

Finally, we trained the model on all 11 labels simultaneously. This configuration yielded strong overall results : *Fake News*, *Opinions*, *Subjective*, and *Reported Information* all reached the same accuracy of 84.21%. Other labels such as *Facts* (94.74%), *Exaggeration* (89.47%), and *Offbeat Title* (89.47%) were also highly predictable, while *Places, Dates, People* reached perfect classification (100%).

The only relatively lower-performing labels were *Sources Cited* (68.42%) and *Insinuation* (73.68%).

Interestingly, several labels (e.g., Fake News, Opinions, Subjective) yielded identical accuracy scores across different setups. This could be due to label correlation. These labels might consistently co-occur in the dataset, leading to similar decision boundaries during training.

5.4 Discussion

The results are summarized in Table 2. The results demonstrate a clear benefit to incorporating additional, related labels during training. While the single-label model predicting only "Fake News" achieved a respectable accuracy of 78.95%, this performance improved to 84.21% when training jointly on three related labels (Opinions, Subjective, Sources Cited). This trend continued in the full multi-label setup, where all 11 labels were used: although the "Fake News" accuracy remained at 84.21%, the test loss dropped significantly from 0.5642 to 0.4911, suggesting a more confident and calibrated model.

This indicates that multi-task learning helps the model learn more generalizable and robust features by leveraging correlations between labels. Training with additional labels likely acts as a regularizer, reducing overfitting on the Fake News task alone and improving overall model generalization.

Overall, the results support the idea that a multi-label learning approach improves performance on the primary task (Fake News detection) and yields more nuanced predictions for related linguistic or stylistic attributes.

Label Configuration	**Test Loss**	**Fake News Accuracy**
Only *Fake News*	0.5480	0.7895
Fake News + 3 Related Labels	0.5642	0.8421
All 11 Labels	0.4911	0.8421

Table 2: Test loss and Fake News accuracy across different label configurations.

6 Conclusion

In this study, we explored fake news detection in French using a multi-label formulation based on CamemBERT. By gradually incorporating additional labels related to factuality, subjectivity, and stylistic cues, we demonstrated that the model not only maintained high accuracy on the primary *Fake News* label but also gained predictive power on complementary dimensions. Our findings suggest that misinformation is best approached as a multi-faceted phenomenon, and that multi-label learning provides a more robust and explainable framework than single-label classification. Our study remains limited due to the small size of the dataset but proved the interest of the OBSINFOX dataset.

References

[1] Saad S Ahmed H, Traore I. Detection of online fake news using n-gram analysis and machine learning techniques. *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*, 2017.

[2] Thomas Icard, Richard Dufour, and Yannick Esteve. Obsinfox: A french multi-label dataset for analyzing stylistic and factual cues in fake news. *Proceedings of the 14th Language Resources and Evaluation Conference (LREC)*, 2024.

[3] Goswami A. Narang P. Kaliyar, R.K. Proceedings of the twelfth acm international conference on web search and data mining. *Multimed Tools Appl*, 2021.

[4] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suarez, Yoann Dupont, Laurent Romary, Eric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: A tasty french language model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[5] Vigneswara Ilavarasan P Reema A, Kar AK. Detection of spammers in twitter marketing: a hybrid approach using social media analytics and bio inspired computing. *Information Systems Frontiers*, 2018.

[6] Liu H Shu K, Wang S. Beyond news contents: The role of social context for fake news detection. *Proceedings of the twelfth ACM international conference on web search and data mining*, 2019.

[7] William Yang Wang. “Liar, Liar Pants on Fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.