# Reliable News or Propagandist News? A Neurosymbolic Model Using Genre, Topic, and Persuasion Techniques to Improve Robustness in Classification

**Géraud Faye**[1,2], **Benjamin Icard**[3], **Morgane Casanova**[4], **Guillaume Gadek**[1], **Guillaume Gravier**[4], **Wassila Ouerdane**[2], **Céline Hudelot**[2], **Sylvain Gatepaille**[1], **Paul Égré**[5]

[1]Airbus Defence and Space, France
[2]Université Paris-Saclay, CentraleSupélec, MICS, France
[3]LIP6, Sorbonne Université, CNRS, France
[4]Université de Rennes, CNRS, Inria, IRISA, France
[5]IRL Crossing, CNRS, Australia

**Abstract**

Among news disorders, propagandist news are particularly insidious, because they tend to mix oriented messages with factual reports intended to look like reliable news. To detect propaganda, extant approaches based on Language Models such as BERT are promising but often overfit their training datasets, due to biases in data collection. To enhance classification robustness and improve generalization to new sources, we propose a neurosymbolic approach combining non-contextual text embeddings (fastText) with symbolic conceptual features such as genre, topic, and persuasion techniques. Results show improvements over equivalent text-only methods, and ablation studies as well as explainability analyses confirm the benefits of the added features.

**Keywords:** Information disorder, Fake news, Propaganda, Classification, Topic modeling, Hybrid method, Neurosymbolic model, Ablation, Robustness

## 1. Introduction

Recent years have seen a sharp increase in online news manipulation, driven by renewed international tensions, as documented in Europe by various intelligence offices (VIGINUM[1] in France, ZEAM[2] in Germany) and monitoring organizations (viz. EU DisinfoLab[3]). Such manipulation of information, which we may refer to as "news disorder" (adapting the terminology of Wardle and Derakhshan 2017), is often orchestrated through press-like websites that mimic journalistic conventions and disseminate targeted narratives to shape opinion (aka. "pseudo-news", see Faye et al. 2024). This is concerning since such content is widely shared on social media, quickly reaching large audiences.[4]

Automated detection of news disorders has advanced in various directions, including media bias (Hamborg et al., 2019), misinformation via automated fact verification (Thorne et al., 2018), fake news (Zhou and Zafarani, 2020; Hu et al., 2025), and rumors (Shu et al., 2020). However, these methods often transfer poorly across other types of news disorder. In particular, they tend to break down on influence operations and propaganda, which hinge on context-dependent framing of genres and topics, and on several specific persuasion techniques (Da San Martino et al., 2019b, 2020b).

In this paper, we focus on the identification of propagandist news, which are particularly insidious because they tend to smuggle in oriented messages with factual reports intended to look like reliable news. For this detection task, LLMs show high performance on specific datasets, but suggestive of a potential overfit. To increase robustness in classification, we argue that hybrid approaches, which have been effective for other news disorders (e.g., Baly et al., 2018; Thorne et al., 2018; Ruchansky et al., 2017; Ma et al., 2017), can improve the detection of influence operations and propaganda in news. Drawing on existing corpora of transparent news versus propagandist news (Faye et al., 2024), we present a neurosymbolic detection model that combines static vector embeddings (fastText) with additional features that include genre, topic, and

[1] https://www.sgdsn.gouv.fr/notre-organisation/composantes/service-de-vigilance-et-protection-contre-les-ingerences-numeriques?

[2] https://www.bmi.bund.de/SharedDocs/schwerpunkte/EN/disinformation-election/zeam-artikel-en.html?

[3] https://www.disinfo.eu

[4] In September 2025, Pew Research Center estimates that 53% of U.S. adults used social media as a news source: https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/.



*Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026*, pages 55–65
12 May 2026. 

persuasion techniques. We provide evidence, using biased splits of the training, validation, and test tests, that the incorporation of these features improves performance over text-only methods.

Section 2 reviews related work on fake news detection and on the challenges posed by propaganda. Section 3 introduces two existing datasets that include reliable articles and then propaganda articles, to identify cross-corpus differences and features characteristic of propaganda. Building on these observations, Section 5 introduces a neurosymbolic model that merges dense text embeddings with interpretable conceptual features automatically extracted from the texts. Section 6 compares the performance of our hybrid method with that of a text-only benchmark; it evaluates its robustness relative to different partitions of the training/valid/test tests, and it uses ablation studies and explainability analyses to validate the method. Section 7 discusses the findings and outlines directions for future work.

## 2. Related work

The detection of information manipulation and news disorders is a broad domain that involves identifying different types of problematic content. The broadest category for misinformation is that of *fake news*, often defined as false or biased information, produced with an agenda or by negligence (Baptista and Gradim, 2022). Other varieties of news disorder include *rumors*, typically news reporting information that is hardly verifiable when published, intending to capture public attention. A specific variety of fake news is *propaganda*, namely partisan information seeking to set a narrative in order to debunk an enemy and glorify a state or organization.

Fake news detection (Hu et al., 2022; Zhou and Zafarani, 2020) is a popular area of research. It can be detected using transformer-based models (Pelrine et al., 2021), or using linguistic features and web markup features (Castelo et al., 2019), or combining linguistic and knowledge features (Seddari et al., 2022; Guelorget et al., 2021).

Fake news can also be easily detected using social media propagation patterns (Silva et al., 2021a; Davoudi et al., 2022), reaching in some cases more than 99% accuracy on four-day propagation data. However, in the current context, content-based fake news detection is more relevant, as a four-day delay is too long for effective detection.

Other approaches rely on user reports (Tschiatschek et al., 2018) and are designed to remain effective even when the majority of users engage in malicious reporting behavior. Various datasets exist for this subtask, often annotated by journalists, such as PolitiFact (Shu et al., 2020) and Horne2017 (**?**) on US politics, CoAID (Cui and Lee, 2020) on Covid-19, as well as LIAR (Wang, 2017), MultiFC (Augenstein et al., 2019) or MuMIN (Nielsen and McConville, 2022) on diverse topics. In the case of rumors, related datasets for this task are Fakenewsnet (Shu et al., 2020), relying on labels produced by PolitiFact and GossipCop, and PHEME (Kochkina et al., 2018).

A method put forward to analyze the news and to track whether they count as reliable or fake is *stance classification* (Riedel et al., 2017). In one version of the task, the goal is to assign a claim-evidence pair to one of three categories: the evidence either supports the claim, contradicts it, or fails to provide sufficient evidence. The primary dataset for this task is FEVER (Thorne et al., 2018), containing more than 300,000 facts. This task could help detect misinformation based on the content to be checked and a small collection of related evidence, making this task also close to fact-checking. The method has also been used to detect propaganda (Hanley, 2025).

Propaganda differs from other forms of news disorder by explicitly mimicking news articles and relying on frames and persuasion techniques (Da San Martino et al., 2020b). Barrón-Cedeño et al., 2019 computes a propaganda score, defined as the estimated likelihood that an article contains propagandistic mechanisms, using engineered stylistic and lexical features such as readability, lexical richness, and TF-IDF n-grams, whereas Da San Martino et al., 2019b introduces a fragment-level annotation scheme in which expert annotators mark the exact spans in news articles that realize propaganda and label each marked span with one of 18 propaganda persuasion techniques. This formulation, combining span identification with technique classification, is benchmarked in NLP4IF-2019 (Da San Martino et al., 2019a) and SemEval-2020 Task 11 (Da San Martino et al., 2020a).

Faye et al. (2024) compare human annotations with model predictions for multi-label propaganda analysis of press articles and systematically evaluates which stylistic cues explain performance. In particular, they show that feature sets targeting vagueness and subjectivity, together with syntactic and lexical cues, can achieve performance comparable to RoBERTa, while making the textual correlates of predictions more explicit.

Classical content-only approaches for news disorder detection often overfit to dataset or domain artifacts and therefore generalize poorly across outlets, genres, topics, and types of news disorder (Suprem et al., 2022; Silva et al., 2021b; Pan et al., 2023; Krieger et al., 2022). To improve robustness, neurosymbolic hybrid approaches pair neural representations of news articles with structured, complementary signals such as external evidence, source

metadata, lexicon and stylistic indicators, or multimodal cues. This strategy has been effective for misinformation through evidence-based verification and debunking (Thorne et al., 2018; Popat et al., 2018), for fake news by incorporating social-context and multimodal signals (Ruchansky et al., 2017; Wang et al., 2018), and for media bias by combining article text with source-level features and distant supervision (Baly et al., 2018; Spinde et al., 2021).

In the wake of these approaches, here we present a neurosymbolic model for propaganda detection grounded in stylistic features and observable symbolic cues, motivated by differences reported across previously published news corpora. We first conduct a comparative analysis of propaganda and mainstream news, then leverage the observed contrasts to design and evaluate our feature-based model using different organizations of our dataset to aim for more robustness.

## 3. Datasets used

### 3.1. Two corpora

In the rest of this paper, we exploit two datasets recently presented in Faye et al. (2024), which respectively consist of a corpus of propagandist pseudo-news articles (PPN) and a corpus of reliable articles from the mainstream press (MAINSTREAM).

- PPN[5] (Faye et al., 2024), for Propagandist Pseudo-News, is a collection of 12,427 articles from sources identified as propaganda outlets by the expert organizations NewsGuard and VIGINUM.[6] The five sources were created after the Russian invasion of Ukraine in February 2022 and contain propagandist news in 9 different languages (Arabic, Chinese, English, French, German, Italian, Russian, Spanish and Ukrainian).

- MAINSTREAM is a corpus of French and English articles, this time of regular news coming from established newspapers, and used as a control for the analysis of the PPN corpus. The MAINSTREAM articles were selected based on publication dates and on keywords related to the Ukraine conflict. MAINSTREAM consists of 1,004 English articles and 1,367 French articles from 11 and 5 sources, respectively.

While both datasets were introduced in Faye et al. (2024), only a small portion was analyzed in the context of an annotation experiment (100 in French split across both sources). Here, we analyze the whole corpus in full and using different methods.[7]

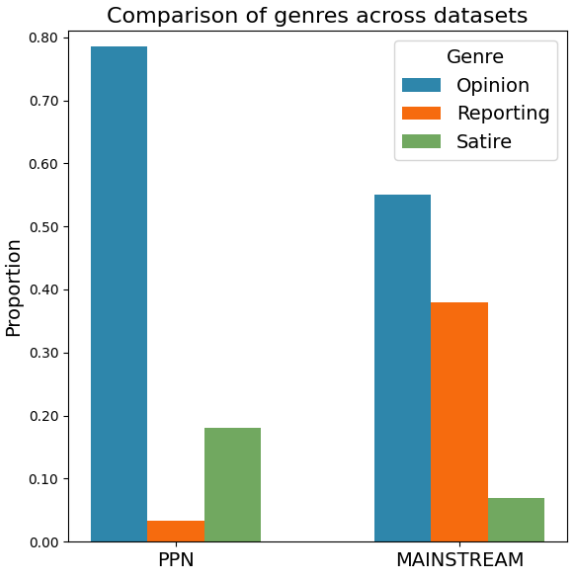


Figure 1: News genre distribution on the two corpora.

### 3.2. Review of published observations

Before presenting new analyses, we briefly summarize the main results of previous analyzes.

Firstly, a French subset of PPN and MAINSTREAM was manually annotated using 11 labels, adapted from a previous annotation experiment.[8] The labels included "vague", "subjective", "exaggeration", "pejorative", "descriptive", "propaganda", "satirical", "dishonest title", "adequate sources", "false information", and "fake news".[9]

Mainstream articles were found to be generally descriptive (close to 90%) and to adequately cite their sources (above 80%). They received low scores for "subjective" (20%) and "vague" (10%), and hardly any of the other labels, including no ascription of "false information" (0%). By contrast, propaganda articles were labeled as descriptive for only 60%, and they were overwhelmingly labeled

[5] https://github.com/hybrinfox/ppn

[6] https://www.sgdsn.gouv.fr/publications/maj-19062023-rrn-une-campagne-numerique-de-manipulation-de-linformation-complexe-et

[7] Because of copyright issues, the text is not redistributed, but direct links to web pages are provided with the corresponding annotations.

[8] See the OBSINFOX dataset, https://github.com/obs-info/obsinfox and its analysis in Icard et al. (2024). The focus on French was to take advantage of the native competence of the annotators.

[9] The difference between "fake news" and "false information" was that in order to be ascribed the former label, an article had to contain "at least one false information". The definition of "fake news" was deliberately left up to the annotators in the experiment, see (Faye et al., 2024) for details.

as manifesting subjectivity (70%), with up to 30% of "false information" and over 40% of "fake news".

Secondly, Large Language Models fine-tuned for the task, such as RoBERTa-base (Liu et al., 2019), were trained and tested on the whole English corpus. The LLMs were found to distinguish with very high performance between propaganda and regular articles (99.7% of test accuracy). Other models, such as TF-IDF (Sparck Jones, 1972), gave a similar high performance (98.5%), and also evidenced lexical differences between the two English corpora.

In both cases, however, this very high performance raises suspicion, since it can be a sign that the methods won't transfer to new datasets. While we do not see direct evidence of overfitting in the comparison between our training, validation, and test tests, such high accuracy points to the risk of a lack of robustness regarding other datasets. Furthermore, it is also possible that articles from the PPN dataset are partially written by AI, adding further biases to the model, leading to more overfitting on machine-written misinformation (Su et al., 2023).

## 4. Genres, topics, and persuasion techniques

To get a more general comparison of the two datasets, here we use distinct analytics of *genre*, *topic*, and *persuasion-techniques* distributions across the two corpora, following the distinctions proposed for the SemEval-2023 Task 3, and using the public APIs of the news classifier GATE Cloud.[10]

**Genres.** A three-fold genre distribution into Reporting articles / Opinion articles / Satire-like articles, is shown in Figure 1 for the two corpora.

The comparison reveals significant differences between the datasets: MAINSTREAM is characterized by a larger proportion of Reporting articles, more than six times the proportion in PPN. MAINSTREAM also shows a lower proportion of Satire. While these articles do not necessarily display humorous content characteristic of satire (see Icard et al. 2024, where the label "satirical", defined as intending to produce laughter, was applied less than 5% even for propagandist articles), this finding supports previous observations that fake news are stylistically closer to satire in style than regular news (Horne and Adali, 2017).

Finally, while Opinion is significantly represented across the two corpora, including the MAINSTREAM one, the class represents more than three quarters of the propagandist corpus PPN, confirming the link between propaganda and persuasion, and the tendency of propaganda to blur the frontier between factual reports and opinion pieces.

**Topics.** We used the same suite of annotating tools to get the topic distribution of the articles. A division along nine topics is shown in Figure 2, showing the two corpora to have relatively similar distributions.[11] This suggests that the two datasets can meaningfully be compared in terms of stylistic features, since they broadly have the same coverage.

**Persuasion techniques.** Finally, we used a third distributional analysis, this time relative to the set of persuasion techniques defined in Piskorski et al. (2023), again using the Cloud multilingual persuasion technique classifier. The distribution of persuasion techniques by articles is shown in Figure 3.

The plots indicate that propaganda articles from the PPN corpus tend to use more of these persuasion techniques, which is coherent with the fact that more than 90% of that corpus is identified as Opinion or Satire. In particular, we see a more prevalent use of *loaded language*, *repetition*, *exaggeration-minimization*, and *appeal to prejudice* in the propagandist corpus.

Overall, these analyses show us that in terms of genres as well as persuasion techniques, propagandist articles are more easily recognizable than other kinds of articles. This fact, combined with the additional characteristics explored in Faye et al. (2024), suggests that we may enhance the detection of propaganda by taking into account genres, persuasion techniques, and other stylistic features. To address this, the next section introduces a hybrid approach that integrates neural and symbolic representations by combining text-based features with concepts extracted from the content.

## 5. Neurosymbolic approach for propaganda detection

For the remaining of the paper, we focus on the English part of the corpus (3219 articles for PPN and 1004 articles for MAINSTREAM), as the models used perform better on this language, and will allow for better quantitative evaluation of our approach. The imbalance between the classes is managed by using the WeightedRandomSampler of PyTorch to expose the model to a balanced amount of classes during training.

To enhance robustness, here we embed text using neurosymbolic methods, and we add conceptual information to the articles to enhance classi-

[10] https://cloud.gate.ac.uk/shopfront#tagged=Misinformation

[11] A similar distribution was also found in the OBSINFOX dataset mentioned in fn. 8.

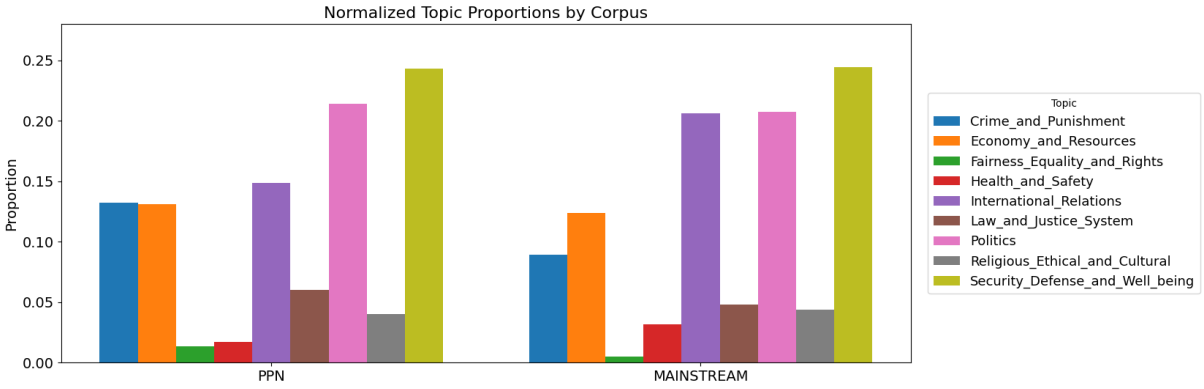


Figure 2: Topic distribution on the two corpora.

fication. Text embeddings are an efficient way of encoding texts and classifying them in downstream tasks. However, there is little understanding of the information they contain. To create a more robust model, we propose combining simple statistical text embeddings with the features observed in the previous section. Thus, texts are encoded using pre-trained fastText embeddings (Bojanowski et al., 2017). While these are non-contextual embeddings, they provide stable and parcimonious lexical representations for large amounts of data.

### 5.1. Proposed architecture

For each article, the process creates a 300-dimensional vector representing the text's distributional lexical characteristics. In addition to this vector, information about the genre, topic, and persuasion techniques contained in the articles is added.

- Genre information is encoded using one-hot encoding (OHE), creating a vector containing only zeros with the exception of the encoded feature, which contains a one. In this case, it adds 3 specific dimensions (for Reporting, Opinion, and Satire).
- Topic information is also one-hot encoded, adding 9 dimensions (for the topics displayed in Figure 2).
- Information about persuasion techniques is added into a vector counting how many persuasion techniques of each type are contained in the article. Fine-grained persuasion techniques represent 23 dimensions. However, these techniques can be grouped into coarser-grained groups, resulting in only 6 dimensions (Piskorski et al., 2023).

In total, a 35-dimensional vector (=3+9+23) for fine-grained persuasion techniques is added to the 300-dimensional fastText embedding. This vector goes through a two-layer perceptron containing a dense layer (335 dimensions to 335 dimensions) with ReLU activation function, and a dense layer (335 dimensions to 1 dimension) with sigmoid activation function to get a propaganda estimation score between 0 and 1. The text is then classified relative to a threshold of 0.5. A global view of the proposed architecture is presented in Figure 4. When all persuasion techniques are included, we call the resulting method the *Hybrid Method*. A scaled-down method is obtained by using an 18-dimensional (=3+9+6) vector incorporating only coarse-grained persuasion techniques: we call the corresponding method *Hybrid Lite*.

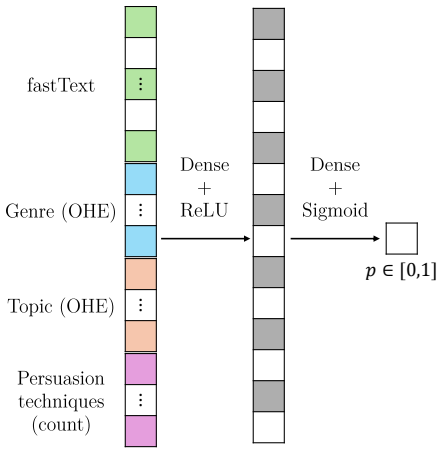


Figure 4: Our hybrid architecture, combining neural features and extracted concepts.

This architecture is voluntarily made simple in order to perform explainability analyses, which are hardly possible with Large Language Models. For this binary classification class, we use a single neuron as an output as we frame the task as we oppose directly reliable news to propagandist news.

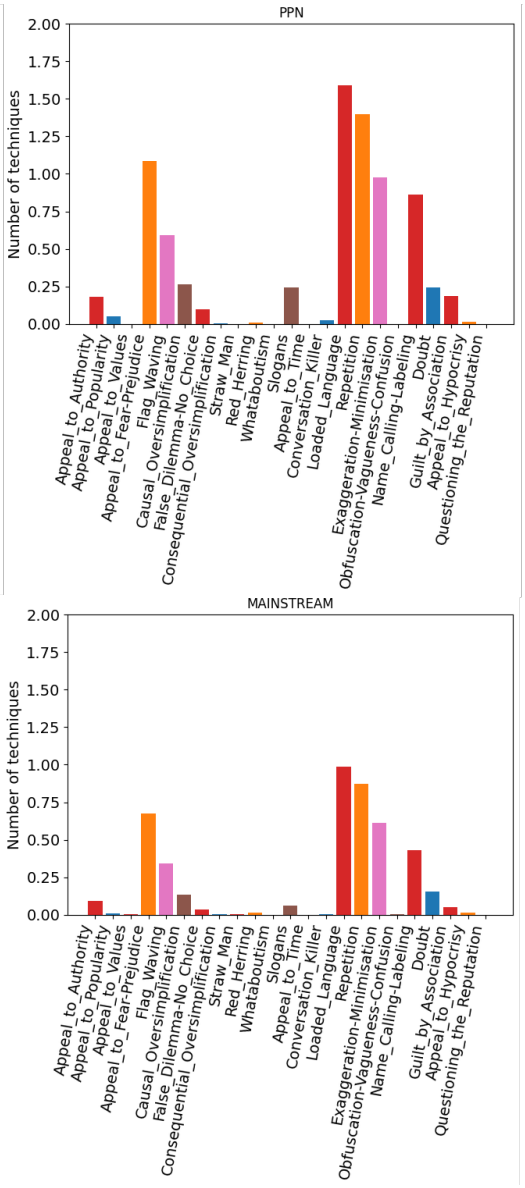


Figure 3: Mean number of persuasion techniques by article, on the two corpora.

### 5.2. Evaluation methodology

In order to prevent overfitting, we designed a new evaluation methodology to ensure that the proposed approach could generalize better to new events and sources.

The general idea is to split the data into train/valid/test sets depending on different criteria explained below. The model is trained on the train set and evaluated on the validation set after each epoch. Early stopping with patience 20 is used to monitor the F1-score on the validation set. The F1-score is the harmonic mean between precision (true positive/true positives+false positives) and recall (true positives/true positives+false negatives). In what follows, propaganda articles are the positive class to detect.

The models are trained with a cross-entropy loss and the Adam optimizer (Kingma and Ba, 2015) with a learning rate of $10^{-4}$ for a maximum of 300 epochs. If no improvement in the validation F1-score is observed after 20 epochs, the best model is restored and evaluated on the test set, giving the scores reported in the results tables. We ran the experiment five times with different seeds and report the average results over these 5 runs.

The way the train/valid/test sets are defined can help measure robustness, namely how effective the learned features are in new contexts. Toward that goal, we define four types of split for our experiments (again on the English part of each corpus):

- **Random**: The articles are randomly sampled to produce 80%-10%-10% random sets.
- **Sources**: The split sets only contain articles from specific sources. The sources were chosen to have an approximately 80%-10%-10% split distribution. The split of sources is given in Table 1 (top).
- **Political**: Mainstream articles are split according to their political leaning annotation by MediaBiasFactCheck.[12] As a majority of articles come from left-leaning sources, we use them as training sources, and randomly split right-leaning sources between validation and test. Propaganda articles are randomly chosen for each set following an 80%-10%-10% distribution (see Table 1 middle).
- **Credibility**: Similarly to political leaning, MediaBiasFactCheck proposes credibility ratings of sources based on freedom of the press, articles' factuality, ownership, and previous fact-checks. In this sense, all propaganda articles come from low credibility sources. As a large majority of regular articles come from high-credibility sources, we use these sources for training and low-credibility sources for validation and test. Propaganda articles are randomly chosen for each set following an 80%-10%-10% distribution (see Table 1, bottom).

Each split is designed to evaluate one type of potential bias of our model based on sources, which may combine several types of unidentified biases, and then on political orientation, and credibility.

## 6. Results

This section is divided into three parts. The first provides the results of the proposed approach on the

[12] It is important to note that MediaBiasFactCheck seems to follow the American Overton window, so their political annotation may differ from what could be considered in other countries.

| Sources | Train | Validation | Test |
|---|---|---|---|
| MAINSTREAM | APNews<br>The Guardian | CNN USA Today Forbes Fox News<br>NBC News NYTimes Washington Post | CBSNews<br>Daily Mail |
| PPN | RRN | TribunalUkraine War on Fakes | |

| Political | Train | Validation | Test |
|---|---|---|---|
| MAINSTREAM | APNews CNN USA Today NBC News<br>NYTimes Washington Post CBSNews | Daily Mail Forbes Fox News | |
| PPN | Entire PPN corpus (English) | | |

| Credibility | Train | Validation | Test |
|---|---|---|---|
| MAINSTREAM | APNews CNN USA Today Forbes The Guardian<br>NYTimes Washington Post CBSNews | Daily Mail | Fox News |
| PPN | Entire PPN corpus (English) | | |

Table 1: News distribution in the **Sources**, **Political**, and **Credibility** splits.

different splits. In the second part, ablation studies are conducted to measure the benefit of adding conceptual embeddings to the textual embeddings. Finally, an explainability analysis highlights in which cases the proposed approach has more benefits than others.

### 6.1. Main results

Results for the different splits are shown in the first row of Table 2 (Hybrid), reporting Accuracy (Proportion Correct) and F1 score. Note that the test sets are different in each column. The **Random** column corresponds to classical evaluation. The **Sources** column corresponds to the system being confronted with new sources, the **Political** column to the system being confronted with new political ideas, and the **Credibility** column to the system being confronted to sources of different credibility from the training set.

The results obtained on the **Random** set are not high but are decent for such a small model. The system shows comparable performance for the **Sources** and **Credibility** splits, but has more difficulties dealing with new **Political** orientations. Compared to **Credibility**, **Political** does not include *Forbes* in the validation and test sets, but has it in the train set along with *The Guardian* and without *CBSNews*. These shifts suffice to lower performance, suggesting that the political orientation of training sets should be variegated to create more robust systems.

### 6.2. Ablation studies

The motivation for the proposed approach is to improve robustness in new scenarios by combining conceptual features with text embeddings to reduce overfitting. To evaluate the performance of our Hybrid method, we conducted ablation studies. To begin with, a model using only the fastText embeddings is trained and evaluated (Table 2, Text Only). Then, the features of the persuasion techniques are altered to take only into account the coarse persuasion categories (6 instead of 23, see Table 2, Hybrid Lite).

Several observations can be made:

- In nearly all cases, using fine-grained labels for persuasion techniques (Hybrid) improves performance over using coarse-grained labels (Hybrid Lite). One exception is the **Random** split, but the gain is large where the Hybrid Lite method struggles (**Sources**, **Credibility**), and the loss small otherwise.
- On average, the Hybrid method improves performance compared to the Text Only method (+26.89% accuracy and +41.85% F1-score average), even though it performs less well in **Random** and **Political**. Overall, however, the Hybrid method is the most robust across the four splits: it learns in all four cases, and it has the largest average performance with the least variance ($\mu_{F1} = 86.12, var_{F1} = 9.18$).
- By contrast, whereas the Text Only method outperforms the others in two splits (**Random**, **Political**, in the **Sources** and **Credibility** splits textual embeddings were not sufficient to learn to discriminate between propaganda and mainstream articles. This indicates that the Text Only method is not robust to perturbations of the training set, even though it does not overfit on political orientation.

In summary, even if the proposed approach does not perform best on traditional random splits, it shows better robustness and generalization than the equivalent text-only approach, which collapses in two cases. Another advantage of this approach is its simplicity, allowing for the application of explainability methods, to which we turn next.

### 6.3. Explainability analyses

To explain our hybrid model, we used SHAP (Lundberg and Lee, 2017). This game-theory-based ap-

| Method | Random | | Sources | | Political | | Credibility | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 |
| Hybrid | 78.08 | 87.5 | **79.45** | **88.37** | 69.86 | 81.66 | **79.45** | **86.95** |
| Hybrid Lite | **79.45** | **88.54** | 34.24 | 31.42 | 67.12 | 79.66 | 56.16 | 69.81 |
| Text Only | **79.45** | **88.54** | 20.54 | 0.0 | **79.45** | **88.54** | 20.54 | 0.0 |

Table 2: Results for propaganda detection with different data splits and different ablations.

proach performs local explanations on each sample by determining which features are the most important for the final prediction. However, the produced explanations are local and dataset-dependent, they do not explain the model's behavior more generally.

To get a global representation of what a model has learned, we can average absolute SHAP values over the different splits. We grouped SHAP values by categories for better readability. For each sample of each split, we calculate the absolute value of the sum of all text-encoding SHAP values, and similarly for the embeddings of genre, topic, and persuasion techniques. Mean SHAP values by split group and category are shown in Figure 5.

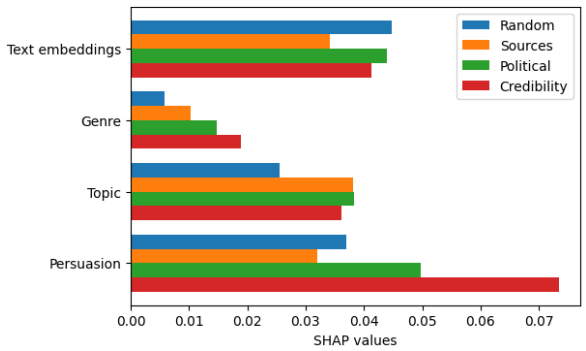


Figure 5: Mean SHAP values for the Hybrid method depending on split.

The figure shows that for **Random**, the text embeddings contribute the most to the decision classification. For **Credibility**, the persuasion embeddings contribute comparatively the most. For **Sources** and **Political**, the results are more mixed across groups of features.

In the **Random** case, this is likely due to the fact that by choosing articles randomly, the training set is better aligned with validation and test, so they are informative enough for classification. This is confirmed by the results, which show that they are better with textual features only.

However, when the test distribution is different from the training distribution, the method tends to use at least as many conceptual embeddings when compared with text embeddings. In particular, for the **Credibility** and **Sources** splits, relying on conceptual features is what makes the model generalize better to new sources and to papers whose reliability is questionable. This situation corresponds to the case in which an article is suspicious and comes from unknown sources, making this approach suitable for propaganda detection.

## 7. Conclusion

This paper introduced a propaganda detection method that integrates textual embeddings with conceptual features extracted through cross-comparison of PPN and MAINSTREAM. The corpora differ significantly in vocabulary and persuasion strategies, suggesting that models trained on a single source may miss corpus-specific signals. These observations motivated the inclusion of conceptual features to better capture propaganda patterns.

By designing biased splits in our datasets, which correspond to the exposure of the model to new types of articles, we have shown that adding conceptual information extracted from the texts improves detection performance, especially in cases where there are new sources of variable credibility ratings. Experiments also suggest that political diversity in the training set is essential for propaganda detection, as the addition of conceptual features significantly degrades performance in this case.

Further explainability analyses show that the added features were indeed used by the model when the splits were biased, allowing the model to correctly detect propaganda when simple textual embeddings are not informative enough.

However, the experiments were run on a corpus centered on one main theme: the Russia-Ukraine conflict. Additional experiments could be conducted on other recent themes, such as recent elections, or other conflicts. The PPN and MAINSTREAM corpora were also only processed in English, and similar experiments should be conducted in other languages to identify potential language-specific differences.

Finally, other types of conceptual features could be used based on other expert knowledge systems or even human operators. In other experiments, an expert vagueness estimation system was successfully combined with a language model for the task of subjectivity detection (Casanova et al., 2024). It may be possible to add an estimate of the document's source reliability to a classification model, to enhance the classification performance of a text-only classifier.

## Acknowledgments

We thank two anonymous reviewers for helpful comments and feedback. This work was supported by the programs HYBRINFOX (ANR-21-ASIA-0003) and TRUSTEDNEWS (ANR-25-ASM2-0003). PE thanks the Department EEE of the University of Melbourne, and the Department of Philosophy of Monash University, for additional support.

## 8. References

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.

João Pedro Baptista and Anabela Gradim. 2022. A working definition of fake news. *Encyclopedia*, 2(1).

Alberto Barrón-Cedeño, Ismaeel Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Morgane Casanova, Julien Chanson, Benjamin Icard, Géraud Faye, Guillaume Gadek, Guillaume Gravier, and Égré Paul. 2024. HYBRINFOX at CheckThat! 2024 - task 2: Enriching bert models with the expert system VAGO for subjectivity detection. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, CLEF '2024, Grenoble, France.

Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. 2019. A topic-agnostic approach for identifying fake news pages. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 975–980, New York, NY, USA. Association for Computing Machinery.

Limeng Cui and Dongwon Lee. 2020. Coaid: COVID-19 healthcare misinformation dataset. *CoRR*, abs/2006.00885.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019a. Findings of the nlp4if-2019 shared task on fine-grained propaganda detection. In *Proceedings of the NLP4IF Workshop*. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of SemEval-2020*. Association for Computational Linguistics.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, pages 4826–4832. Survey track.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019b. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Mansour Davoudi, Mohammad Moosavi, and M. Sadreddini. 2022. DSS: A hybrid deep model for fake news detection using propagation tree and stance network. *Expert Systems with Applications*, 198:116635.

Géraud Faye, Benjamin Icard, Morgane Casanova, Julien Chanson, François Maine, François Bancilhon, Guillaume Gadek, Guillaume Gravier, and Paul Égré. 2024. Exposing propaganda: an analysis of stylistic cues comparing human annotations and machine classification. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 62–72, Malta. Association for Computational Linguistics.

Paul Guelorget, Benjamin Icard, Guillaume Gadek, Souhir Gahbiche, Sylvain Gatepaille, Ghislain

Atemezing, and Paul Égré. 2021. Combining vagueness detection with deep learning to identify fake news. In *2021 IEEE 24th International Conference on Information Fusion (FUSION)*.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: An interdisciplinary literature review. *International Journal on Digital Libraries*, 20(2):391–415.

Hans WA Hanley. 2025. Tracking and identifying international propaganda and influence networks online. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29263–29264.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):759–766.

Bo Hu, Zhendong Mao, and Yongdong Zhang. 2025. An overview of fake news detection: From a new perspective. *Fundamental Research*, 5(1):332–346.

Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. 2022. Deep learning for fake news detection: A comprehensive survey. *AI Open*, 3:133–155.

Benjamin Icard, François Maine, Morgane Casanova, Géraud Faye, Julien Chanson, Guillaume Gadek, Ghislain Atemezing, François Bancilhon, and Paul Égré. 2024. A multi-label dataset of french fake news: Human and machine insights. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 812–818.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jan-David Krieger, Timo Spinde, Terry Ruas, Juhi Kulshrestha, and Bela Gipp. 2022. A domain-adaptive pre-training approach for language bias detection in news. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL '22)*. ACM.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada. Association for Computational Linguistics.

Dan S. Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3141–3153. Association for Computing Machinery.

Liangming Pan, Yunxiang Zhang, and Min-Yen Kan. 2023. Investigating zero- and few-shot generalization in fact verification. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–524, Nusa Dua, Bali. Association for Computational Linguistics.

Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabbany. 2021. The surprising performance of simple baselines for misinformation detection. In *Proceedings of the web conference 2021*, pages 3432–3441.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE:

Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *ArXiv*, abs/1707.03264.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 797–806, New York, NY, USA. ACM.

Noureddine Seddari, Abdelouahid Derhab, Mohamed Belaoued, Waleed Halboob, Jalal Al-Muhtadi, and Abdelghani Bouras. 2022. A hybrid linguistic and knowledge-based analysis approach for fake news detection on social media. *IEEE Access*, 10:62097–62109.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Amila Silva, Yi Han, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021a. Propagation2vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management*, 58:102618.

Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021b. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 557–565.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural media bias detection using distant supervision with BABE - bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023. Fake news detectors are biased against texts generated by large language models.

Abhijit Suprem, Sanjyot Vaidya, and Calton Pu. 2022. Exploring generalizability of fine-tuned models for fake news detection. In *2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC)*, pages 82–88. IEEE.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake news detection in social networks via crowd signals. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 517–524, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

William Yang Wang. 2017. "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 (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 849–857. ACM.

Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policymaking.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5).