A Multi-Label Dataset of French Fake News: Human and Machine Insights

Benjamin Icard[1,2], **François Maine**[1,3], **Morgane Casanova**[4], **Géraud Faye**[5,6]
Julien Chanson[7], **Guillaume Gadek**[5], **Ghislain Atemezing**[8]
François Bancilhon[9], **and Paul Égré**[2]

[1]LIP6, Sorbonne Université, CNRS, France - [2]Institut Jean-Nicod, CNRS, ENS-PSL, EHESS, France
[3]Freedom Partners, France - [4]Université de Rennes, CNRS, Inria, IRISA, France
[5]Airbus Defence and Space, France - [6]Université Paris-Saclay, CentraleSupélec, MICS, France
[7]Mondeca, France - [8]European Union Agency for Railways, France
[9]Observatoire des Médias, France

Abstract

We present a corpus of 100 documents, named OBSINFOX, selected from 17 sources of French press considered unreliable by expert agencies, annotated using 11 labels by 8 annotators. By collecting more labels than usual, by more annotators than is typically done, we can identify features that humans consider as characteristic of fake news, and compare them to the predictions of automated classifiers. We present a topic and genre analysis using GATE Cloud, indicative of the prevalence of satire-like text in the corpus. We then use the subjectivity analyzer VAGO, and a neural version of it, to clarify the link between ascriptions of the label Subjective and ascriptions of the label Fake News. The annotated dataset is available online at the following url: https://github.com/obs-info/obsinfox

Keywords: Fake News, Multi-Labels, Subjectivity, Vagueness, Detail, Opinion, Exaggeration, French Press

1. Introduction

One of the challenges raised by fake news is that the very notion of fake news is multidimensional. It includes fabrication, satire, but also mistaken reports, and often just biased or partisan information (Tandoc et al., 2018; Gelfert, 2018; Zhou and Zafarani, 2018).

Notwithstanding that complexity, algorithms trained to detect fake news typically rely on datasets involving just two labels, such as "biased" vs "legitimate" (viz. ISOT[1]), with no indication of the type of fake news in question, let alone the cues used to explain the labels. In order to get reliable detectors of fake news, however, it matters to use datasets and labels that are sufficiently precise in order to inform classifiers along several dimensions. Some multi-label fake news datasets are available, such as LIAR (Wang, 2017) (6 labels), or the Brazilian dataset of (de Morais et al., 2019) (4 labels). In the former, labels qualify levels of truth (following the politifact.com guidelines) and in the in the latter, the *legitimate-biased* distinction is crossed with the presence or absence of satire. However, except for that feature, these labels do not pertain to stylistic information.

In this paper, we report on the constitution and annotation of a corpus of French press named OBSINFOX, selected from websites categorized by expert organizations as unreliable, and so as good candidates to include biased, exaggerated, or even factually false statements. While OBSINFOX is limited in size (100 documents), our goal was to obtain a rich dataset, by considering 11 labels for annotation, and then by asking 8 annotators to annotate it.

The aim was twofold: on the one hand, we intend to identify which labels are most informative of the status of a text. On the other, we are interested in finding the cues in those texts that best explain their classification by humans and then by machines as containing fake news or not.

Section 2 explains the selection of the corpus OBSINFOX, the choice of the labels, and the guidelines and method for collection of the annotations. Section 3 gives an analysis of the topics and genres of the corpus using GATE Cloud, and Section 4 presents an analysis of the human annotations and their relations. In Section 5, finally, we examine the way in which the label "Fake News" is ascribed in relation to other labels, in particular to "Subjective", "False" and "Exaggerated". Toward that goal, we use the text analyzer VAGO to relate scores of linguistic subjectivity with human scores on the label "subjective". We validate this approach by using a neural version of VAGO, trained on a distinct corpus "FreSaDa" (Ionescu and Chifu, 2021), of satirical news.

[1]https://onlineacademiccommunity.uvic.ca/isot/#datasets

LREC-COLING 2024, pages 812–818
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

2. Corpus and Labels

- **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 of the article 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 of the article.

Figure 1: Description of the 11 labels selected for the annotation task.

The dataset OBSINFOX was compiled from online sources of French press presented as unreliable and prone to propagating fake news by NewsGuard[2] and Conspiracy Watch[3] in particular. The time period covered goes from 2010 to 2023, but is mostly focused on the 3 last years. Exactly 100 articles were selected for the study. That sample originated in a larger corpus of 54,845 online articles, itself the result of keeping only the 17 most popular French sources among the 40 involved in an original corpus of 101,200 articles. A pilot study involved the selection of 906 articles within the 54,845 articles in order to conduct a first human annotation task with 4 annotators and 26 labels. We used the TfidfVectorizer transformer to pre-select 120 articles among the 54,845 articles, half of which with a probability of reporting fake news above .8 according to the predictor, half with a probability below .2. Among those 120, 100 were retained after elimination of 20 articles too short or uninterpretable, with 49 predicted to be fake news, and 51 not. A detailed list of the press sources included can be found in the README file available on the OBSINFOX repository.

For the labels, Figure 1 presents them in the order in which annotators had to mark them, with a summary of their definition. The selection of the labels was based on prior meetings, during which the annotators iteratively discussed the procedure and the annotation manual including the definitions provided in Figure 1. The annotators eventually agreed on 11 labels after discussing a broader set of 26 labels coming from the pilot study on 906 articles mentioned in section 2. Another decision was to allow only binary responses instead of more answer types (such as “I don’t know”), but participants were authorized to leave personal comments (eventually removed from the dataset).

The labels distinguish “Fake News” from “False Information", and define the former more widely (as involving falsity or exaggeration), to distinguish plain falsities (“Obama was not born in the USA”) from cases of exaggeration involving partially true facts (“inflation skyrockets everywhere"). The label “Offbeat Title" is linked to clickbait detection, usually a marker of exaggeration or distortion. The label “Opinion” looks for the occurrence of at least one opinion sentence, “Subjective” concerns whether opinions are prevalent over objective reports. The labels “Places, Dates, People" and “Facts" are used to assess whether an article reports at least a factual piece of information or only the author’s opinion. The presence of location, temporal, or nominal information allows fact-checking systems to process the article, in contrast to articles containing only opinions (Guo et al., 2022). “Reported Information” and “Sources Cited” are related and can help identify if the information is directly endorsed by the writer and if it comes from secondary sources. Beside “Exaggeration”, the label “Insinuation” was also included to detect indirect derogatory techniques (such as dog-whistle).

The 8 annotators included the designers of the experiment (7 male, age range from 29 to 76). Only one of them had seen the texts prior to annotating, in order to upload them on the platform. Annotators didn’t have access to the url to avoid bias by source, unlike in other datasets (ISOT or Horne and Adali 2017). The resulting dataset does not present aggregate data (as do PolitiFact and GossipCop, Shu et al. 2018) but includes individual annotations grouped by annotator, to give access to individual variability and allow for more refined analyses.

3. Topic and Genre Analysis

Analyses of the corpus were conducted after selection, using pretrained tools made available by

[2] https://www.newsguardtech.com/
[3] https://www.conspiracywatch.info/

GATE Cloud.[4] More precisely, the topics and genres of articles were detected using an ensemble of mBERT models (Wu et al., 2023). These tools have been chosen because of their accessibility and the good performance they achieved during the SemEval 2023 Task 3. The distributions of topics is shown in Figure 2, left. Topics are diverse, but nearly half of the articles deal with Politics and with Health and Safety, followed by Security Defense and Well-being, and Religious, Ethical and Cultural topics.

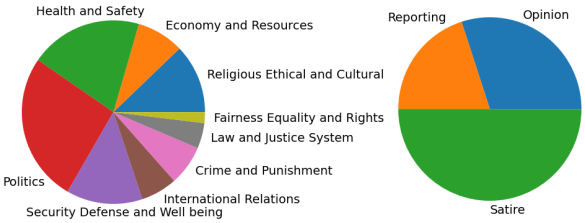


Figure 2: Topic and genre distribution in the corpus.

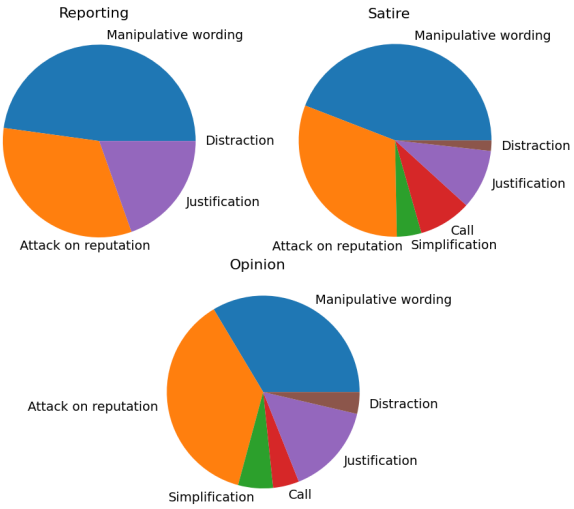


Figure 3: Persuasion techniques by genre.

To further our analysis, we categorized the different articles into genres, following the three-way news categorization proposed by (Piskorski et al., 2023) into *opinion* pieces, pieces aiming at objective news *reporting*, and *satire* pieces, also using the tools provided by GATE (Figure 2, right). Half of the chosen articles are written in a satire-like style (only stylistically, as no real satire involving humor is present in the corpus). This confirms the observations made in (Horne and Adali, 2017) about the prevalence of caricature and exaggeration in fake news. Within each type, we looked at the manipulative persuasion techniques inventoried in (Piskorski et al., 2023), based on the taxonomy proposed by (Da San Martino et al., 2020) for propaganda. They include 23 techniques in total, falling into 6 main groups, including so-called manipulative wording, distraction, attack on reputation, call (to act or think), simplification, and [partisan or biased] justification (see Figure 3).

The distributions of persuasion techniques are approximately the same for opinion articles and for satire-like articles. They differ by the number of persuasion techniques used by articles, with opinion and satire-like containing respectively a mean of 3.4 and 4.6 persuasions techniques by article. In the reporting articles, less diverse persuasion techniques are found, which is to be expected as they are more factual. However, the number of persuasion techniques found is relatively high (2.3 per article) for a factual content, even if it is lower than for opinion pieces.

4. Human Annotations

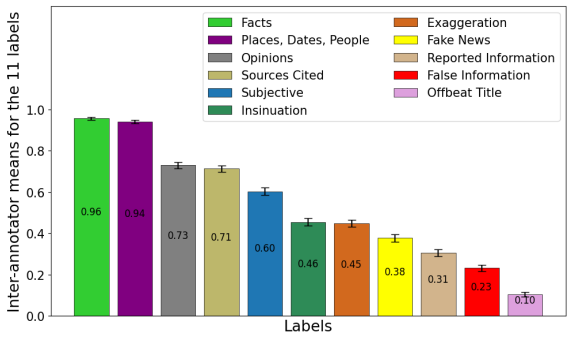
In order to assess the quality of annotations, we measured the inter-annotator agreement among the 8 human annotators, using two distinct measures. First we calculated Fleiss's kappa for each document, in order to shed light on the overall reliability of their collective judgments. For the 11 annotation labels given in Figure 1, we obtained a mean value per document of $\kappa = 0.4659074$, showing moderate agreement between annotators of the panel.

The second method we used, displayed in Figure 4b, consisted of rescaling the percentage of agreement between annotators: for each document, we computed the proportion x of answers equal to 1, rescaled by the function returning the value $\alpha = |2x - 1|$. The rescaling implies that when only half of the annotators agree on a label, the level of agreement is 0. When 75% go in the same direction, agreement is 0.5, and a value of .7 or above indicates 85% of agreement or more. Compared to Fleiss's kappa, the rescaling method is easier to interpret and sheds light on the inter-annotator agreement per article for each of the 11 labels. In addition, Pearson's calculation shows that both metrics are very well correlated ($r = 0.94$, $p = 3.06e - 46$).

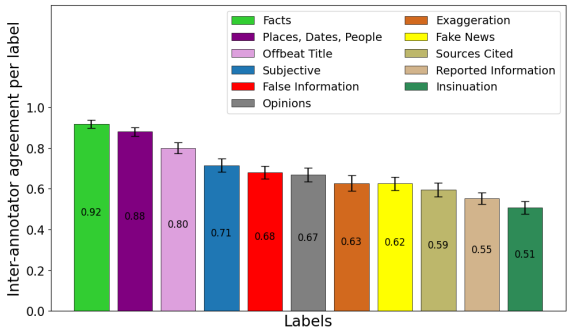
As Figure 4b shows, all labels reached a mean value above .5. The label "Facts" shows the highest agreement, and the label "Insinuation" the lowest. Other labels of particular interest for us, concretely "Fake news", "False Information", "Opinions", and "Subjective", all reach a mean score above .6, with the highest level for "Subjective" (0.715).

Figure 4a shows how much on average a label was used across the 100 articles by each annotator (mean of means). The label "Fake News" reaches a mean of .38, hence below the proportion predicted at selection, and below the percentage of Satire

[4] https://cloud.gate.ac.uk/



(a) Mean inter-annotator scores per label.



(b) Mean inter-annotator agreement per label.

Figure 4: Mean scores and agreement by label (error bars=standard error of the mean).

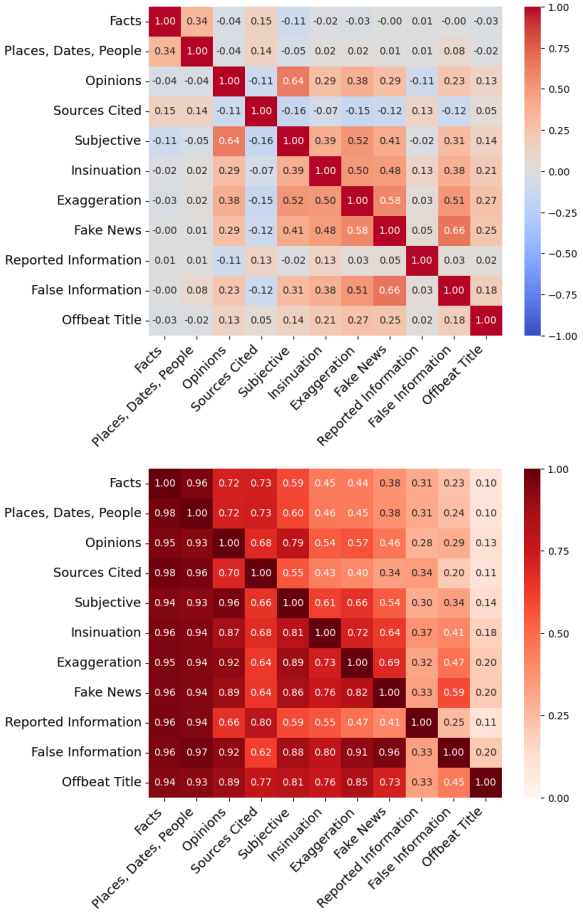


Figure 5: Correlation matrix between the 11 labels (top), and percentage of a row label satisfying a column label (bottom).

found by GATE Cloud.[5] "False Information", with a mean of .23, is ascribed less than "Fake News", consistently with their definition.

Figure 5 (top) displays the correlation between the 11 labels and shows that the labels that correlate the most are "Subjective", "Opinion", "Insinuation", "Exaggeration", "Fake News", and "False Information". Figure 5 (bottom) also reports, from the 800 judgment profiles, the proportion of a row label A that is (asymmetrically) associated with a column label B. 59% of items tagged as "Fake news" are tagged as "False Information", versus 96% of "False Information" tagged as "Fake News". The proportions of "Exaggeration"/"Fake News"/"False Information" tagged as "Subjective" are 89%, 86%, 88%. Conversely, the proportion of "Subjective" documents tagged as "Exaggeration"/"Fake News"/"False Information" is 66%, 54%, 34%. This indicates that while the inference from "Fake" or even "False" to "Subjective" is strong, the converse inference from "Subjective" to "Fake" and "False" is weaker.

[5]However, a comparison between the labels of the predictor used for selection and majorities on the label "Fake News" shows low accuracy of 0.40, 0.40 and 0.36 respectively for levels of agreement $\alpha = .5, = 0.75$ and $= 1$.

5. Ascriptions of "Fake News"

In light of the associations between the labels "Fake News", "Exaggeration", and "Subjective", and to understand the linguistic cues picked by annotators, we used an automated detector of subjectivity in texts, the VAGO tool, applied on larger corpora to relate the occurrence of subjective lexicon in text with the detection of fake news (Guélorget et al., 2021; Icard et al., 2023b). For a given text, VAGO computes three scores, a score of vagueness, a score of opinion, and a score of relative detail compared to vagueness. VAGO does not incorporate any world-knowledge, but checks for the occurrence of markers of subjectivity (including first-person pronouns, exclamation marks, and terms of exaggeration or slurs among evaluative adjectives), as well as markers of objectivity (named entities in particular).

VAGO produces a score of linguistic subjectivity that previous studies have found positively correlated with the label "biased" in news articles (Guélorget et al., 2021; Icard et al., 2023a), and a score of detail-vs-vagueness that previous studies have

found negatively correlated with the label "satirical" (Icard et al., 2023a). Hence, we hypothesized that larger VAGO scores of opinion should predict higher use of the labels "Subjective", and "Opinions". For "Fake News", however, we expect a weaker association, since falsity is a separate component of that label as defined in the annotation guide.

To test those hypotheses, we calculated the correlation between the VAGO scores for each document in the corpus and the mean inter-annotator scores for the labels "Subjective", "Opinions", "Exaggeration", "Fake News", and finally, "False Information". We used two sets of VAGO scores: those produced by the expert system VAGO, and the scores produced by a neural clone VAGO-N. This neural version VAGO-N combines the "CamemBERT-base" French version (Martin et al., 2019) of the BERT model (Devlin et al., 2018) (*Batch Size*=5, *Learning Rate*=1e-05, *Epochs*=5) with 3 regression layers and 3 MSE loss functions to predict the scores of vagueness, opinion and detail of sentences. Building the model consisted of using the VAGO scores on the 141,137 sentences of the French corpus "FreSaDa"[6] (Ionescu and Chifu, 2021), making the following random selection: 99,022 sentences for training, 21,219 sentences for validation and 21,219 sentences for test. We obtained high performance for the three scores as indicated by the Root Mean Square Error (RMSE) measures: 0.026 for vagueness, 0.028 for opinion, and 0.083 for detail. We obtained similar performances by comparing VAGO with VAGO-N on the 100 articles (see Table 1), and generally, found close correlation scores between the two versions of VAGO across the labels tested (Figure 6).

RMSE	sentences level	articles level
vagueness	0.048	0.012
opinion	0.036	0.010
detail	0.118	0.046

Table 1: Root Mean Square Error between VAGO scores and VAGO-N scores, at sentence level (N=2,445) and at article level (N=100).

Regarding our hypotheses, we computed correlations between the three VAGO scores and the mean scores for the labels "Subjective", "Opinions", "Exaggeration", "Fake News" and "False Information", using both VAGO and VAGO-N. Here we report the VAGO-N case only, as both versions give very similar results. As shown in Table 2, we found positive correlations between scores of vagueness and opinion and the labels "Subjective", "Opinions", and "Exaggeration", but not for "Fake News" and "False Information". For all cases, however, we found a

[6] https://github.com/adrianchifu/FreSaDa

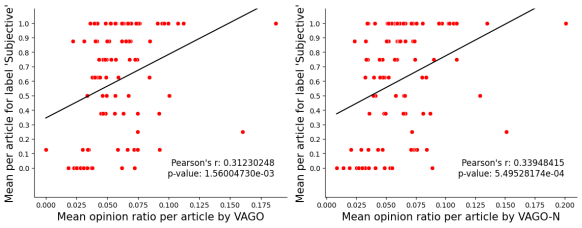


Figure 6: Pearson correlations between the mean opinion score per article provided by VAGO and VAGO-N and the mean inter-annotator score for the label "Subjective".

negative correlation between the score of detail to vagueness and the labels. The correlations are weak to moderate, but in the order of magnitude found in previous studies, and even higher in the labels "Opinions" and "Subjective" directly connected to VAGO's opinion score. These results confirm that there is a stronger association between VAGO markers of opinion and assessments of texts as "Subjective" than between those scores and assessments as "Fake News", a distinction not visible in (Guélorget et al., 2021)'s analysis of the ISOT-False corpus, in which "Biased" was an annotation used indistinctly to refer to both "Fake News" and "Opinions" pieces that may not be fake.

	vague	opinion	detail
Subjective	0.294^{**}	0.339^{***}	-0.380^{***}
Opinions	0.266^{**}	0.342^{***}	-0.358^{***}
Exaggeration	0.217^{*}	0.300^{**}	-0.232^{*}
Fake News	0.134	0.201	-0.261^{**}
False Information	0.080	0.139	-0.303^{**}

Table 2: Pearson correlations between labels' mean scores and VAGO-N scores (*, **, and *** indicate p-value $< .05, < .01, < 0.001$).

6. Conclusion

With only 100 documents, the corpus presented here is limited to train a classifier, but it is valuable in virtue of its rich set of annotations, and it can be used for further regression analyses concerning the ascription of the label "Fake News" relative to other labels. The analyses confirm that linguistic markers of subjectivity explain part of the variance in the ascription of labels such as "Subjective", "Opinion", "Exaggeration", but also "Fake News". Some labels in our study turns out to be uninformative ("Facts", "Places"), while others could be included ("Satirical", to check for presence of humor). We refer the readers to the follow-up study (Faye et al., 2024), in which an adjusted set of 11 labels is used to analyze propaganda press.

Limitations

All annotators have a higher-education degree (5 years or more after graduation), not necessarily representative of a larger and more diverse population.

Acknowledgements

We thank three anonymous reviewers for helpful comments and feedback, and Guillaume Gravier for his support. This work was supported by the programs HYBRINFOX (ANR-21-ASIA-0003), FRONTCOG (ANR-17-EURE-0017), THEMIS (n°DOS0222794/00 and n°DOS0222795/00) and PLEXUS (Marie Skłodowska-Curie Action, Horizon Europe Research and Innovation Programme, grant n°101086295). PE thanks Monash University for hosting him during the writing of this paper.

Declaration of contribution

All the authors contributed to the design, analysis, and discussion of the results. BI, GF, MC, and PE wrote the paper, which all authors read and revised together. Correspondence: benjamin.icard@lip6.fr, paul.egre@ens.psl.eu.

References

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Janaína Ignácio de Morais, Hugo Queiroz Abonizio, Gabriel Marques Tavares, André Azevedo da Fonseca, and Sylvio Barbon Jr. 2019. Deciding among fake, satirical, objective and legitimate news: A multi-label classification system. In *Proceedings of the XV Brazilian Symposium on Information Systems*, pages 1–8.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*.

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 EACL Third Workshop on Understanding Implicit and Underspecified Language (UnImplicit 2024)*.

Axel Gelfert. 2018. Fake news: A definition. *Informal logic*, 38(1):84–117.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Paul Guélorget, 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 *IEEE 24th International Conference on Information Fusion (FUSION)*, pages 1–8.

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. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.

Benjamin Icard, Vincent Claveau, Ghislain Atemezing, and Paul Égré. 2023a. Measuring vagueness and subjectivity in texts: from symbolic to neural VAGO. In *IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2023)*.

Benjamin Icard, Vincent Claveau, Ghislain Atemezing, and Paul Égré. 2023b. Un traitement hybride du vague textuel: du système expert VAGO à son clone neuronal. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2023)*.

Radu-Tudor Ionescu and Adrian-Gabriel Chifu. 2021. FreSaDa: A French satire data set for cross-domain satire detection. In *The International Joint Conference on Neural Network, IJCNN 2021*, IJCNN2021.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. CamemBERT: a tasty French language model. *arXiv preprint arXiv:1911.03894*.

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.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Edson Tandoc, Zheng Wei Lim, and Richard Ling. 2018. Defining “fake news” a typology of scholarly definitions. *Digital journalism*, 6(2):137–153.

William Yang Wang. 2017. Liar, Pants on Fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Ben Wu, Olesya Razuvayevskaya, Freddy Heppell, João A. Leite, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. SheffieldVeraAI at SemEval-2023 task 3: Mono and multilingual approaches for news genre, topic and persuasion technique classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1995–2008, Toronto, Canada. Association for Computational Linguistics.

Xinyi Zhou and Reza Zafarani. 2018. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*, 2.