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

Tom ROSSA
ENSAE Paris
tom.rossa@ensae.fr

Abstract

Detecting fake news remains a major challenge in the field of Natural Language Processing (NLP). Beyond the development of high-performing models, one of the key issues highlighted in the literature is the need for high-quality, human-annotated datasets that allow models to capture the multiple dimensions—linguistic, syntactic, stylistic—that characterize misinformation. This project focuses precisely on that angle: understanding and modeling the linguistic cues that guide human annotators in identifying fake news. Our study is grounded in the *OBSINFOX* dataset, a richly annotated collection of French-language news articles containing a wide range of metadata and expert labels. We aim to investigate what kinds of linguistic features—both surface-level and high-level—contribute to the perception of an article as misleading or deceptive.

Contributions. We begin with a detailed exploration of the dataset's structure and labels to assess how informative each annotation is for the task of fake news detection. Next, we engineer a set of structural features directly from the raw textual content, in order to capture different stylistic and rhetorical dimensions of the articles. Finally, we integrate outputs from multiple pre-trained language models—fine-tuned for specific understanding tasks—to compare and contrast machine-based predictions with human annotations. Through a series of quantitative experiments, we evaluate whether the OBSINFOX dataset provides meaningful signals that can improve automated fake news detection and help bridge the gap between human judgment and computational analysis.

GitHub repository of the project.

1 Introduction

The automatic detection of fake news has emerged as a major challenge in natural language processing (NLP), particularly in the age of social media and the exponential growth of information sources. Both factual and misleading content now spread at unprecedented speed and scale. In response, a growing body of research has investigated the use of NLP techniques to detect fake news in textual data, which may appear in diverse forms such as online articles, social media posts, or political statements. This task is inherently difficult, as it often requires identifying false or misleading information solely based on the linguistic content of the text, without relying on external fact-checking sources. Indeed, building and maintaining a comprehensive, up-to-date knowledge base capable of verifying the factual accuracy of all online content would be extremely difficult in practice—especially for texts related to current events, where information evolves rapidly. Consequently, recent research efforts have focused on developing methods that rely exclusively on the internal features of the text—such as its vocabulary, phrasing, rhetorical structure, or the density of factual assertions—to estimate the likelihood that it conveys misinformation or bias. One of the key obstacles to such approaches lies in the availability of high-quality annotated datasets that allow researchers to better characterize fake

news and train reliable classifiers. This issue is precisely what the authors of [3] aim to address, by proposing a novel multilabel dataset specifically designed for the detection and categorization of fake news in French.

A Multidimensional Concept. One of the main challenges in detecting fake news using natural language processing (NLP) methods lies in the inherently multidimensional nature of fake news itself. Since the objective is to identify linguistic or semantic signals within a text that may suggest the presence of false or biased content, a wide array of factors must be considered. Two syntactically similar sentences in different news articles may be perceived by experts as either indicative of bias or entirely neutral, depending on subtle contextual elements. In this sense, fake news is not a monolithic phenomenon, but rather a multidimensional construct that may be shaped by factors such as historical context, the source of the article, the type of vocabulary used, rhetorical tone, or even the subject matter. Moreover, fake news can manifest in a variety of forms: it may involve complete fabrication of facts, deliberate exaggeration, factual errors, implicit framing to guide reader interpretation, or overtly biased opinions. Capturing this complexity is a significant challenge for NLP models—especially when the datasets on which these models are trained do not themselves reflect the multidimensionality of the phenomenon.

The Importance of High-Quality Datasets. In the literature, researchers have developed a range of datasets in multiple languages, comprising corpora of news articles and textual data annotated by humans to support research in automated fake news detection. The **ISOT** dataset [1] includes news articles labeled as either "real" or "fake", mainly sourced from American media outlets, and consists of well-structured and relatively long texts. Although ISOT covers a wide range of topics and article types, it provides only binary labels and thus fails to account for the various forms fake news can take. By contrast, the **LIAR** dataset [7] offers a more fine-grained labeling scheme, containing over 12,000 manually annotated political claims from the PolitiFact platform, categorized into six levels of truthfulness. This granularity makes it particularly well-suited for fine-grained classification tasks and for contextual approaches based on pre-trained language models such as BERT. In this context, the authors of [3] introduced the *OBSINFOX* dataset, which provides a rich, multilayered annotation of French news articles by multiple annotators. This project relies heavily on that dataset in an effort to identify which linguistic and structural cues annotators rely on when determining the veracity of a given article. The overarching objective is to establish a dialogue between human judgments and machine learning models, leveraging the dataset's granularity to identify reliable features for fake news detection in the press. Structural features will be engineered directly from the texts, while multiple instances of a fine-tuned CamemBERT model will be used to quantify various linguistic aspects of the articles and relate them to the human-provided labels.

2 Data Analysis: OBSFINFOX Dataset

The *OBSINFOX* dataset was introduced in [3] by a team of nine researchers. It is openly accessible via this GitHub repository.

2.1 Construction of the OBSINFOX Dataset

The *OBSINFOX* dataset constitutes a valuable resource for the study of misinformation in French-language news media. It comprises 100 articles sourced from 17 French media outlets identified as unreliable by watchdog organizations such as *NewsGuard* and *Conspiracy Watch*. The dataset was constructed through a multi-step selection process aimed at ensuring high data quality:

- From a large corpus of over 100,000 French news articles flagged as unreliable by these agencies, 17 sources were selected for further analysis.
- The authors applied the `TfidfVectorizer` method (see Appendix A), a text vectorization and transformation technique used to estimate the likelihood that an article contains fake news.
- Among an initial shortlist of 120 documents (ranked by predicted fake news probability and after eliminating uninterpretable texts), 100 final documents were selected.

Among the 100 retained documents, half were assessed by the `TfidfVectorizer` model as having a high probability of containing fake news (score > 0.8), while the other half were considered to have

a low probability (score < 0.2). Each document was manually annotated according to 11 distinct dimensions, capturing linguistic, stylistic, and factual characteristics that reflect various forms of misinformation, including factual inaccuracies, subjectivity, exaggeration, insinuation, and satirical tone.

Labels and Metadata in the OBSINFOX Dataset. The dataset comprises eleven carefully designed labels aimed at capturing a range of *factual*, *stylistic*, and *interpretative* properties of news content. Annotation was carried out by a panel of eight expert annotators, ensuring a high degree of quality and consistency. Metadata such as the article title, annotator ID, and article URL are also included. The label design reflects both the objective content of a text and the subjective framing often found in misleading or partisan articles. A detailed description of each label is provided in the following table.

Label	Description
`Fake News`	The article contains at least one false or exaggerated claim.
`Places, Dates, People`	The article references at least one identifiable location, date, or person.
`Facts`	The article reports at least one factual element, whether true or false.
`Opinions`	The article expresses at least one opinion, judgment, or personal interpretation.
`Subjective`	Subjective content (opinions) outweighs verifiable facts in the article.
`Reported Information`	The article relays information attributed to an external source, without directly endorsing it.
`Sources Cited`	The article cites at least one source that supports or contextualizes a claim.
`False Information`	The article contains demonstrably false or factually incorrect content.
`Insinuation`	The article implies or suggests an interpretation without stating it explicitly.
`Exaggeration`	The article presents a real fact in a way that amplifies or distorts its significance.
`Offbeat Title`	The headline is sensational or misleading compared to the actual content of the article.

Table 1: Description of the labels used in the OBSINFOX dataset.

Some of these labels are relatively objective and can be directly identified (e.g., `Places, Dates, People`, `Facts`, `Sources Cited`), while others involve a higher degree of subjective interpretation by the annotators (e.g., `Insinuation`, `Subjective`, `Exaggeration`). This multi-label structure allows for a fine-grained, multidimensional characterization of news content.

Analyzing the co-occurrence patterns among labels provides valuable insights into how alternative or misleading narratives are constructed and perceived by human readers. It also offers a rich analytical framework for training and evaluating automated misinformation detection models. By relying exclusively on linguistic cues, the multi-label setup of OBSINFOX enables the exploration of weak signals of misinformation and editorial bias. As such, the dataset provides a robust foundation for developing supervised learning and linguistic analysis methods in the context of natural language processing applied to fake news detection.

2.2 Quality of Human Annotations

We now turn to evaluating the quality, structure, and relevance of the human annotations included in the *OBSINFOX* dataset. This analysis serves three main purposes:

- Assess the **consistency of the annotations** with respect to the dataset construction methodology,
- Examine the **class balance** across the different annotated labels,
- Measure the **level of agreement between annotators** (*inter-annotator agreement*).

This step is essential to our study, as it enables us to assess both the reliability and the informational value of the annotations provided by human experts. Understanding the linguistic and stylistic cues used by annotators to flag misinformation is a crucial prerequisite for comparing human and

model-based interpretations. In the later stages of the project, we aim to contrast these expert-driven annotation patterns with the predictions and internal logic of automated NLP models, with the ultimate goal of aligning algorithmic and human approaches to fake news detection.

As a starting point, we evaluate the extent to which annotators agree on the assignment of each label. The higher the degree of consensus among annotators, the more informative and trustworthy the dataset becomes. To do so, we begin by visualizing the frequency with which each label is assigned by individual annotators (see Appendix B for the detailed table).

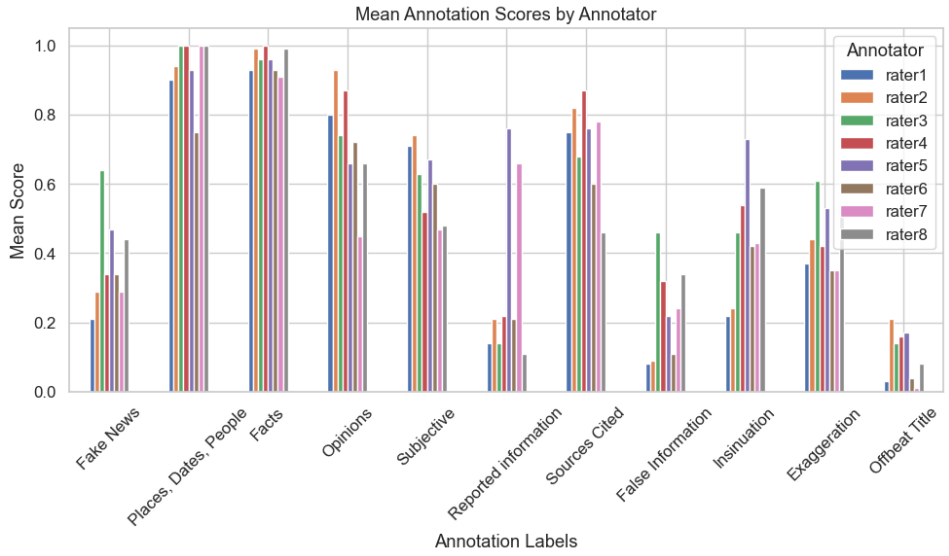

Figure 1: Average label attribution frequency by annotator.

We observe that for more *objective labels* such as `Places, Dates, People`, `Facts`, and `Sources Cited`, the distribution of label assignments is relatively consistent across annotators. However, for labels that rely more heavily on *subjective interpretation* and judgment, the variability is much greater—even though the label definitions were clearly communicated. For example, we see that Raters 5 and 7 are substantially more likely to assign the label `Reported Information`, while Raters 1 and 7 almost never assign the label `Offbeat Title`.

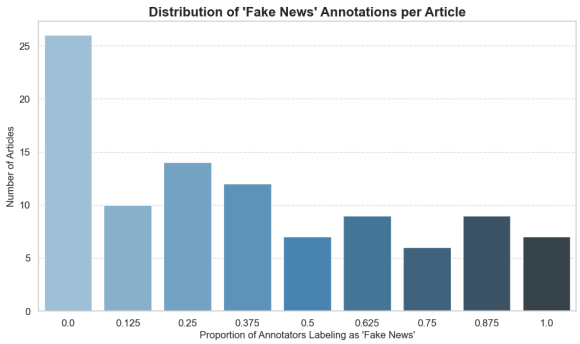

Figure 2: Barplot of voting proportions for the label `Fake News`.

Label	**Proportion $>$ 0.5**
Places, Dates, People	99%
Facts	98%
Sources Cited	73%
Opinions	69%
Subjective	60%
Insinuation	44%
Exaggeration	40%
Fake News	31%
Reported Information	14%
False Information	14%
Offbeat Title	2%

Table 2: Proportion of articles for which each label was assigned by a majority of annotators.

These statistics highlight the heterogeneity of perception among annotators, especially for subjectively defined categories. In particular, only 31% of articles received a majority vote for the `Fake News`

label, suggesting either genuine disagreement or the difficulty in reaching consensus on ambiguous cases. In contrast, nearly all annotators agree on the presence of objective features such as `Places, Dates, People` or `Facts`, reinforcing the idea that more subjective labels demand deeper contextual or rhetorical interpretation.

The distribution of voting proportions for the `Fake News` label reveals that a significant number of articles generate disagreement among annotators. For approximately one third of the dataset, between three and five annotators assigned the `Fake News` label—reflecting either a perfect split or only a weak majority. This highlights the subjectivity and ambiguity inherent in identifying misinformation. Interestingly, one might have expected a higher proportion of articles to be labeled as fake news, given that the dataset was initially constructed based on the output of a weakly supervised preselection model using a `TF-IDF Vectorizer`, which aimed for a 50% positive rate. In practice, however, only 31% of the articles received a majority vote for the `Fake News` label. This discrepancy suggests that human annotators are more cautious or conservative in applying the fake news label compared to the automated preselection process.

Agreement Scores. Several statistical metrics exist to evaluate the consistency of annotation decisions across multiple raters. One widely used measure is **Fleiss' Kappa**, which quantifies the degree of agreement among more than two annotators assigning categorical labels (e.g., binary labels 0/1) to a fixed number of items. Unlike Cohen's Kappa, which is limited to pairwise agreement, Fleiss' Kappa generalizes to any number of annotators under the following assumptions:

- Each item is rated by the same number of annotators,
- Labels are categorical (nominal),
- Agreement is computed based on how frequently annotators assign each category.

This metric allows us to estimate whether the observed agreement among raters is significantly better than what would be expected by chance, thereby offering a useful benchmark for the reliability of human-labeled data. The kappa statistic compares the *observed agreement* with the *agreement expected by chance*:

$$\kappa = \frac{\bar{P} - P_e}{1 - P_e}$$

where:

- $\bar{P}$ is the mean observed agreement across all items:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i \quad \text{with} \quad P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1)$$

- P_e is the expected agreement by chance:

$$P_e = \sum_{j=1}^{k} p_j^2 \quad \text{where} \quad p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}$$

Here, N is the number of items, n the number of annotators per item, k the number of categories, and n_{ij} the number of annotators who assigned category j to item i. Fleiss' Kappa ranges from -1 (complete disagreement) to 1 (perfect agreement), with 0 indicating agreement no better than random chance.

While Fleiss' Kappa is widely used to quantify inter-annotator agreement, it presents notable limitations—particularly its sensitivity to class imbalance. For example, the label *Places, Dates, People* is assigned by nearly all annotators to almost every article. This overrepresentation results in a high expected agreement by chance (P_e close to 1), thereby deflating the Kappa score despite consistent labeling. Consequently, Fleiss' Kappa may underestimate agreement for imbalanced labels, making interpretation less intuitive. In our case, the highest scores are observed for *Subjective*, *Exaggeration*, and *Fake News*, suggesting relatively high consensus among annotators on these dimensions. To address these limitations, we adopt a second, more interpretable metric that captures the degree of consensus using a rescaled proportion of label attribution. The method proceeds as follows:

- For each article and each label, compute the proportion x of annotators who assigned that label.
- Transform this value using the function $\alpha(x) = |2x - 1|$.

In this formulation, values of x close to 0 or 1 (i.e., strong consensus) yield an agreement score near 1, while $x = 0.5$ (complete disagreement) yields a score of 0. The overall agreement for a label is then computed as the average of $\alpha(x)$ across all articles. This metric provides a more intuitive reflection of annotator alignment, especially in the presence of class imbalance.

Label	**Fleiss' Kappa**
Subjective	0.54
Exaggeration	0.46
Fake News	0.39
Opinions	0.36
False Information	0.31
Insinuation	0.26
Sources Cited	0.25
Reported Information	0.17
Offbeat Title	0.12
Facts	0.11
Places, Dates, People	0.08

Figure 3: Inter-annotator Fleiss' Kappa Score of Agreement.

Label	**Mean Agreement**
Facts	0.92
Places, Dates, People	0.88
Offbeat Title	0.80
Subjective	0.71
False Information	0.68
Opinions	0.67
Exaggeration	0.63
Fake News	0.62
Sources Cited	0.59
Reported Information	0.55
Insinuation	0.51

Figure 4: Inter-annotator Rescaled Mean Agreement

The Fleiss' Kappa agreement scores indicate that, relative to chance-level expectations, the labels `Subjective`, `Exaggeration`, and `Fake News` exhibit the highest levels of inter-annotator agreement. This suggests that these labels contain meaningful signal rather than being purely noise, which is a positive sign regarding the overall quality of the dataset. Complementing this, the **Rescaled Mean Agreement** metric confirms, as expected, that annotators reach higher levels of consensus on more objective categories such as `Facts` and `Places, Dates, People`. These labels refer to concrete, observable elements of the text, making them easier to identify consistently across annotators. In contrast, the agreement score for `Fake News` reaches only 0.62, and drops further to 0.51 for `Insinuation`. While these values still indicate moderate agreement, they also reflect the inherent subjectivity of such labels. Even well-trained annotators may struggle to reliably detect insinuation or misinformation, as these judgments are highly sensitive to contextual subtleties, tone, and individual interpretation.

To further assess the nature of disagreement, we analyze the distribution of label proportions across all articles (see Appendix C). Figure 8 shows that objective labels such as `Facts` and `Places, Dates, People` display low variance in label attribution across annotators, suggesting consistent majority agreement. In contrast, labels such as `Fake News` and `Exaggeration` exhibit much higher variance. This indicates that some articles are labeled unanimously, while others provoke significant disagreement—highlighting the presence of ambiguous or polarizing articles within the dataset. For over one-third of the corpus, annotators appear highly divided, suggesting that even experts find it difficult to decisively determine whether these articles contain false or misleading information. Lastly, we explore the correlations between annotator decisions (see Appendix D). These inter-rater correlations help identify annotators whose labeling behavior deviates significantly from the others—either due to overly strict or overly lenient tendencies—which may introduce noise or bias into the dataset and influence downstream modeling.

3 State of the Art in Fake News Detection

Automatic detection of fake news has become an active area of research in the field of Natural Language Processing (NLP). Early approaches relied on traditional supervised learning methods applied to manually engineered features such as n-grams, word frequency, readability scores, or lexical indicators of subjectivity and polarity. These approaches enabled the construction of artificial representations of texts that helped models capture some dimensions of misinformation. However, despite being relatively easy to implement, such methods suffer from their inability to model contextual and semantic dependencies within the text.

In this project, our goal is to bridge the gap between automated language understanding models and human annotators by comparing model-derived linguistic indicators with expert labels. To do so, we make use of a set of state-of-the-art pretrained language models—each fine-tuned for specific downstream tasks—to explore whether they can reproduce, explain, or complement the human perception of misinformation.

BERT. With the emergence of pretrained language models based on Transformer architectures—most notably BERT [2]—performance on fake news detection tasks has seen significant improvement. BERT (*Bidirectional Encoder Representations from Transformers*) is a deep neural model that relies solely on the encoder stack of the Transformer architecture to produce rich contextualized representations of input text. Given a tokenized input sequence $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the model applies a stack of Transformer blocks to generate contextual embeddings:

$$\mathbf{z}^{(l)} = \text{TransformerBlock}(\mathbf{z}^{(l-1)}), \quad \text{with } \mathbf{z}^{(0)} = \text{Embeddings}(\mathbf{x})$$

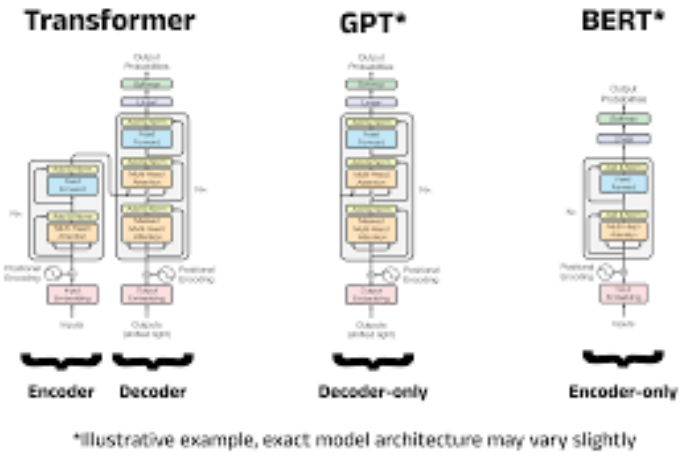

Figure 5: Comparison of Transformer architectures: Encoder (BERT), Decoder (GPT), and Full Transformer.

The original Transformer architecture, introduced by Vaswani et al. [6], consists of an encoder-decoder structure designed for sequence-to-sequence tasks such as machine translation. Each encoder block includes a multi-head self-attention mechanism followed by a position-wise feedforward network, with residual connections and layer normalization. The decoder additionally incorporates a masked (causal) attention layer to prevent information leakage from future tokens—making it suitable for conditional text generation.

BERT, however, only uses the encoder component of the Transformer and applies fully bidirectional self-attention, allowing each token to attend to all others in the input simultaneously. It is trained using two objectives: *Masked Language Modeling* (MLM), where random tokens are masked and predicted, and *Next Sentence Prediction* (NSP), which models sentence-level coherence. These objectives make BERT well-suited for text understanding tasks. In contrast, GPT (Generative Pretrained Transformer) leverages only the decoder part of the Transformer, using causal

(left-to-right) attention. Each token only attends to previous tokens, making GPT more suitable for generation tasks. It is trained autoregressively, i.e., to predict the next token at each step based on previous context. This architectural distinction has led to complementary applications: BERT excels in classification and comprehension tasks, while GPT is better suited for free-form text generation. Our study leverages both types of models—alongside their multilingual and task-specific variants—to explore their potential for detecting and characterizing fake news in the French-language *OBSINFOX* dataset.

mBERT. The multilingual BERT model (mBERT) [5], pretrained on the multilingual versions of Wikipedia, enables the application of language models across multiple languages, including French. It retains the exact architecture of BERT-Base (12 layers, 768 hidden dimensions, 12 attention heads) and does not rely on any explicit cross-lingual alignment. Despite the absence of supervised alignment between languages, mBERT has demonstrated strong zero-shot cross-lingual transfer capabilities, thanks to the shared multilingual WordPiece vocabulary and the emergent properties of joint multilingual training. While it offers robustness in multilingual settings, mBERT is not optimized for language-specific nuances, which may limit its effectiveness in capturing fine-grained linguistic phenomena in French.

CamemBERT. For French-language tasks, the CamemBERT model [4] has emerged as a high-performing alternative. Built upon the RoBERTa architecture, CamemBERT is pretrained on OSCAR—a large, high-quality French corpus—and has achieved state-of-the-art results on various downstream tasks, including misinformation detection. Its ability to capture linguistic subtleties specific to French—such as implicit meaning, subjectivity, or vague expressions—makes it a particularly suitable candidate for fake news classification. Recently, hybrid approaches have been proposed, combining deep learning models like CamemBERT with interpretable linguistic rules or heuristics (e.g., subjectivity markers, vagueness indicators, intensifiers) to improve both performance and explainability. These methods aim to support explainable AI by producing not only accurate predictions but also human-interpretable justifications.

SemEval. SemEval (Semantic Evaluation) is a prominent annual NLP competition designed to benchmark semantic understanding systems on complex tasks in multilingual and multimodal settings. In our study, we leveraged several models made available via the `GateCloud` API, which have been pretrained on large corpora for specific subtasks. In particular, Task 3 of SemEval-2023 [8] focused on detecting three dimensions in online news articles: journalistic genre, framing, and persuasion techniques. For Subtask 1 (News Genre), the top-performing system consisted of an ensemble of four mBERT-based models: three fully fine-tuned on multilingual training data, and one mBERT model with adapter modules. A majority voting strategy was used to aggregate the outputs. Due to significant class imbalance (notably between satire and reporting), the authors applied random oversampling without replacement to balance the training set. The proposed system achieved highly competitive results, including the best average rank among multilingual teams. In English, the ensemble outperformed standalone mBERT models by more than 25% in macro F1 score, demonstrating the effectiveness of the ensemble-based architecture. Multilingual models consistently outperformed their monolingual counterparts in Transformer-based configurations. Although English performance lagged slightly behind other languages, results indicate that the ensemble setup generalized robustly, even under zero-shot conditions.

SemEval 2019. SemEval 2019 introduced the *Hyperpartisan News Detection* task, which aimed to assess whether systems could detect hyperpartisan argumentation in news articles—that is, content exhibiting blind, biased, or irrational allegiance to a political party, individual, or ideology. The subjectivity of this task made it particularly challenging, even for human annotators. The GATE team developed a system based on sentence-level representations derived from ELMo (*Embeddings from Language Models*) embeddings, which were then fed into a Convolutional Neural Network (CNN) with Batch Normalization. Sentence embeddings were computed as the average of their constituent word embeddings, and each document was modeled as a sequence of such sentence-level vectors. The final system relied on ensemble averaging to generate predictions. It achieved an accuracy of 0.822 on the manually annotated evaluation set, ranking first on the final leaderboard.

4 Numerical Experiments

In this section, we present the various computational experiments conducted as part of this project using the *OBSINFOX* dataset. As previously mentioned, the primary objective is to understand the linguistic and stylistic cues that human annotators rely on to identify fake news, and to explore the extent to which NLP models and selected metrics can also capture the multidimensional nature of misinformation in order to detect it effectively.

4.1 Informative Relationships Between Annotated Labels

To gain deeper insights into how annotators perceive and judge the presence of misinformation, we now turn to an analysis of the relationships between the different labels in the dataset. In particular, we aim to identify and disentangle potential *correlations* and *co-occurrences* among labels. The goal is to understand which combinations of stylistic, linguistic, or factual cues are most indicative of an article being perceived as *Fake News*. By examining the correlation structure between label attributions across articles, we seek to uncover the implicit reasoning patterns that annotators may rely on. This multidimensional perspective is crucial for understanding the diverse criteria humans use when evaluating the credibility of information. Ultimately, it can inform models that aspire to replicate or complement human judgment in detecting fake news. To anticipate the next steps of the project, we begin by analyzing the correlation matrix between label attributions. This matrix reveals how often labels co-occur and whether they tend to reinforce or exclude one another.

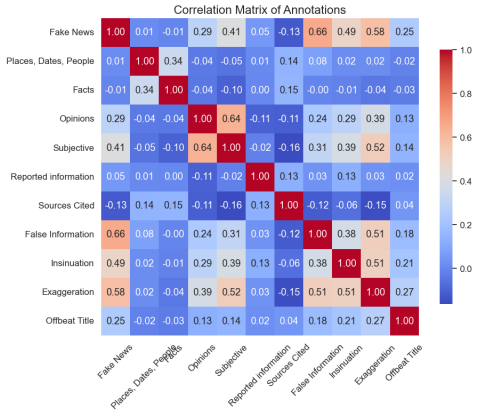

Figure 6: Correlation Matrix Between Labels

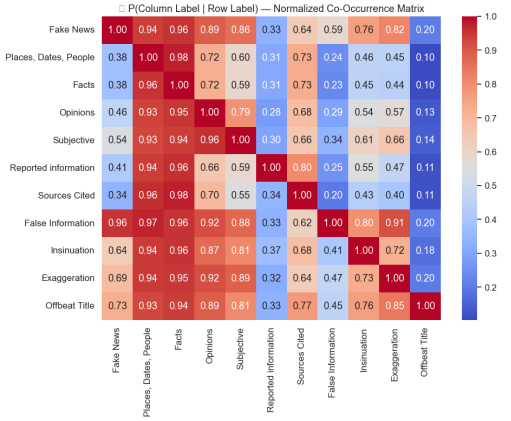

Figure 7: Co-Occurence Matrix of Labels

We observe that the *Fake News* label is positively correlated with labels such as *False Information*, *Insinuation*, and *Exaggeration*. This suggests that annotators tend to associate fake news with articles that are factually incorrect, implicitly misleading, or emotionally charged. These stylistic cues may serve as red flags that prompt suspicion. Conversely, *Fake News* is negatively correlated with the *Sources Cited* label. This indicates that articles which reference sources explicitly are less likely to be flagged as fake news—consistent with the idea that transparency and verifiability enhance perceived trustworthiness. These findings lay the groundwork for further analysis into the multi-faceted nature of fake news detection, emphasizing that it is not a single-label decision but rather a complex synthesis of linguistic and contextual clues.

The matrix reveals several notable insights regarding the informativeness of different labels in the dataset. Each row of the matrix represents the conditional proportion of other labels given the presence of the label corresponding to that row. For instance, 59% of the articles labeled as *Fake News* are also annotated as explicitly containing *False Information*. This suggests that, in the eyes of annotators, a fake news article does not always have to rely on direct falsehoods or lies—it may also involve more subtle forms of manipulation or distortion of the truth. Conversely, 96% of the articles labeled as *False Information* are also labeled as *Fake News*, indicating that when a statement is perceived as objectively false, it is almost systematically associated with misinformation. We also observe strong co-occurrence patterns between the *Fake News*, *Exaggeration*, *Insinuation*, and

Offbeat Title labels. These associations suggest that fake news tends to adopt particular rhetorical or stylistic devices—such as emotional exaggeration or suggestive headlines—that go beyond the mere presence or absence of factual content. Interestingly, the presence of factual elements—such as named entities (places, dates, people) or verifiable facts—is not significantly lower in fake news articles compared to others. Therefore, one cannot conclude that the abundance of factual or quantifiable content alone is sufficient to rule out the presence of misinformation. This underscores the importance of considering not only what is said, but how it is framed and contextualized.

4.2 Fake News Detection with Structural Features

In this section, we shift our focus from annotation-level metadata to the full textual content of the news articles themselves. Since the original dataset only included the URLs of the articles, we automatically retrieved their content using the `newspaper3k` Python library. This tool enables rapid and reliable extraction of the main body text from a given news webpage by combining HTML parsing heuristics with lightweight natural language processing. For each article, we followed the standard pipeline: downloading the page, parsing its structure, and extracting the cleaned main text via the `.text` attribute. Out of the 100 original articles, 79 were successfully processed in this way. The remaining 21 could not be extracted—most likely due to irregular HTML formatting or unsupported web structures.

Once the text was collected, we applied a standard preprocessing pipeline to clean and normalize it. From the cleaned texts, we manually constructed a set of interpretable, structural features aimed at characterizing each article's form. These include the total number of words, total number of characters, average word length, number of digits (figures), number of sentences, number of unique words, number of stopwords, and the ratio of stopwords to total words. We then trained a logistic regression model to predict the `Fake News` label based solely on these features. The objective was to assess whether simple, surface-level structural cues could help discriminate between real and fake news as perceived by human annotators. Despite the simplicity of the model and the hand-crafted nature of the features, the results offer initial insights into which textual patterns may correlate with perceived misinformation.

The logistic regression results suggest that some of the manually engineered structural features significantly contribute to explaining the likelihood that a given text was labeled as "Fake News" by human annotators. For example, a high ratio of *stopwords* — commonly used function words that carry little semantic meaning on their own — appears to be positively associated with the Fake News label. A high proportion of these words within a text may indicate a writing style that is more verbose, emotionally charged, or rhetorically inflated, rather than factually dense or information-rich. This stylistic pattern could reflect strategies used in misleading or manipulative content, where surface-level fluency is prioritized over informational depth. As such, the stopword ratio may serve as a useful proxy for detecting texts with lower informational density — a potential marker of fake news.

Variable	**Coef.**	**Std. Err.**	**z**	**P>|z|**	**[0.025, 0.975]**
Intercept (const)	-8.9289	2.562	-3.485	0.000	[-13.951, -3.907]
Number of Words	0.0002	0.004	0.041	0.968	[-0.008, 0.008]
Number of Characters	-0.0004	0.000	-1.405	0.160	[-0.001, 0.000]
Average Word Size	0.6013	0.193	3.112	0.002	[0.223, 0.980]
Number of Figures	-0.0012	0.004	-0.349	0.727	[-0.008, 0.006]
Number of Sentences	0.0172	0.008	2.159	0.031	[0.002, 0.033]
Number of Unique Words	0.0075	0.002	3.071	0.002	[0.003, 0.012]
Number of Stopwords	-0.0017	0.006	-0.266	0.790	[-0.014, 0.011]
Ratio of Stopwords	10.8908	5.576	1.953	0.051	[-0.038, 21.819]

Table 3: Logistic regression results for predicting the "Fake News" label based on textual features.

4.3 Link between Human and Machine Insights for Fake News Detection

To further enrich our dataset with high-level linguistic and stylistic annotations, we leverage the **GATE Cloud API**, a suite of web-based natural language processing services developed for large-scale, multilingual text analysis. GATE Cloud provides access to a variety of pre-trained and fine-tuned models via RESTful APIs, enabling efficient and language-aware processing of raw text in real-world scenarios. These services are especially useful for augmenting textual datasets with semantic metadata without requiring local model deployment.

In this study, we utilize two specific GATE Cloud models:

- **Genre Classification**: Automatically infers the genre of a news article, distinguishing between types such as opinion pieces, reports, or editorials.
- **Hyperpartisan Detection**: Identifies whether a text exhibits politically biased or hyperpartisan language, indicative of extreme or one-sided rhetoric.

Our objective is twofold. First, we aim to explore how these model-generated labels correlate with the human annotations in our dataset—particularly the `Fake News` label. This may help reveal whether misinformation tends to be associated with specific genres or rhetorical tones. Second, we investigate the distribution of topics and styles across the dataset, assessing whether certain article types or persuasion techniques are disproportionately represented in fake news instances. All models were accessed programmatically via the GATE Cloud REST API using authenticated requests. The API processes unstructured raw text, taking into account full sentence structure and punctuation. The outputs were integrated into our data pipeline and analyzed alongside the manually constructed features.

In the final part of our analysis, we leverage a pre-trained **CamemBERT** model — a French-language adaptation of RoBERTa — originally trained on a large general-purpose corpus for tasks involving textual understanding and the detection of linguistic and stylistic features. This model was then *fine-tuned* on a corpus of French movie reviews for the task of sentiment analysis, specifically to classify whether the tone of a text is predominantly *positive* or *negative*. The model outputs a score between 0 and 1 representing the estimated degree of negativity in the text. We apply this model to our collection of news articles in order to explore whether sentiment polarity can help reveal patterns related to misinformation. More precisely, we aim to investigate whether articles that exhibit highly negative or pejorative language are more likely to be labeled as `Fake News`, or whether they co-occur with other subjective labels such as `Exaggeration`, `Insinuation`, or `Opinion`.

This exploration is based on the hypothesis that fake news articles often rely on emotional intensity, shock value, or caricatured language to attract and hold the reader's attention. Detecting an association between sentiment polarity and fake news classification could therefore offer an additional signal to enrich misinformation detection pipelines. The sentiment model we use is openly available and documented in the following repository: here

Results. We now turn to interpreting the outputs of the pre-trained models applied earlier in the analysis. One of the most striking observations concerns the relationship between the genre predicted by the GATE Cloud classifier and the likelihood of a text being considered hyperpartisan. Specifically, the *Satire* genre is consistently and overwhelmingly identified as hyperpartisan by the model, followed by a relatively high proportion for *Opinion* articles. In contrast, articles classified as *Reporting* exhibit a much lower rate of hyperpartisanship. This consistency across genres suggests a degree of internal coherence in the model's interpretation of textual style and tone, reinforcing the idea that such external predictions can be meaningfully integrated into downstream analyses. Another noteworthy finding is that articles labeled as *Satire* also display a substantially higher prevalence of the human-annotated labels `Fake News`, `Exaggeration`, and `False Information`. This observation aligns with previous research showing that genre—and particularly the use of rhetorical or exaggerated devices—is a strong proxy for misinformation cues. In this context, the detection of genre may act as a valuable high-level semantic signal that complements structural and lexical features in identifying potentially misleading content.

Group	Category	Sentiment	Hyperpartisan	Fake News	Exaggeration	Insinuation
Gate Genre	Opinion	0.737	0.241	0.370	0.442	0.463
	Satire	0.779	0.000	0.446	0.589	0.446
	Reporting	0.655	0.667	0.313	0.340	0.403
Hyperpartisan	Non-hyperpartisan	–	–	0.322	0.410	0.405
	Hyperpartisan	–	–	0.455	0.480	0.540

Table 4: Descriptive statistics of New Features

In addition, we incorporated the outputs of the pre-trained models—namely, the predicted *Genre*, *Hyperpartisan Label*, and *Negativity Score*—as new features in our dataset. These were used as input variables in a series of logistic regression models targeting key annotator-provided labels: `False Information`, `Exaggeration`, and `Fake News`. The aim of this analysis is to assess whether features derived from external language models can serve as significant explanatory variables for the presence of fake news, thus bridging the gap between automated textual interpretation and human judgment.

Variable	Coef.	Std. Err.	z	P>|z|	[0.025, 0.975]
Intercept (const)	-8.933	2.839	-3.146	0.002	[-14.497, -3.368]
Number of Words	0.0024	0.005	0.519	0.604	[-0.007, 0.011]
Number of Characters	-0.0005	0.000	-1.341	0.180	[-0.001, 0.000]
Average Word Size	0.3437	0.238	1.441	0.150	[-0.124, 0.811]
Number of Figures	0.00006	0.003	0.016	0.987	[-0.007, 0.007]
Number of Sentences	0.0221	0.008	2.610	0.009	[0.006, 0.039]
Number of Unique Words	0.0052	0.003	1.974	0.048	[0.00004, 0.010]
Number of Stopwords	-0.0049	0.007	-0.719	0.472	[-0.018, 0.009]
Ratio of Stopwords	12.3570	5.868	2.106	0.035	[0.857, 23.857]
Sentiment Analysis	1.2679	0.846	1.498	0.134	[-0.391, 2.927]
Gate Hyperpartisan Label	0.8467	0.253	3.347	0.001	[0.351, 1.343]
Gate Genre	-0.2824	0.149	-1.900	0.057	[-0.574, 0.009]

Table 5: Logistic regression on `Fake News` label from structural and model-based features

The logistic regression models allow us to investigate which features significantly contribute to predicting the human-annotated labels `Fake News`, `Exaggeration`, and `False Information`. While each model captures slightly different dynamics, several consistent patterns emerge. First, the average word size appears to be a good predictor across all three models, with a positive and statistically significant coefficient. This suggests that texts labeled as misleading or exaggerated tend to use longer words on average—possibly reflecting a more rhetorical or elaborate writing style. Similarly, the number of unique words is positively associated with all three labels, indicating that lexical diversity may correlate with more stylistically charged or less fact-based content.

In contrast, the number of characters shows a negative association with both `Exaggeration` and `False Information`, hinting that more concise wording may be a marker of deceptive or manipulative language. Importantly, the ratio of stopwords—previously hypothesized to reflect verbosity or rhetorical excess—is positively and significantly associated with both `Fake News` and `Exaggeration`, lending further support to the idea that fake news may favor stylistic over informative density. The outputs from the pre-trained models are also informative. The hyperpartisan label predicted by GATE Cloud significantly predicts both `Fake News` and `False Information`, suggesting that texts perceived as politically biased are more likely to be flagged as misleading. Overall, these results confirm that both manually constructed textual features and high-level outputs from pre-trained NLP models can help explain the presence of misinformation-related labels, and highlight the multifaceted nature of fake news—at the intersection of style, structure, and rhetorical intent.

5 Conclusion

In summary, our approach combined standard text preprocessing with the extraction of structural features to quantitatively characterize each article. Beyond these foundational metrics, we leveraged the richness and diversity of fake news annotations present in the dataset to explore how human intuition and model-driven insights can complement each other. By integrating interpretable features derived from human-labelled content with signals extracted from pre-trained models—originally trained on diverse corpora and tasks—we aimed to capture the multifaceted nature of textual manipulation. This dialogue between human judgment and automated perception allowed us to assess the potential of hybrid feature construction for detecting factual distortion, ultimately offering a scalable and generalizable framework for misinformation analysis.

References

[1] H. Ahmed, I. Traore, and S. Saad. "Detecting opinion spams and fake news using text classification". In: *Proceedings of the International Conference on Security and Privacy in Communication Systems (SecureComm)*. 2017.

[2] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

[3] Benjamin Icard et al. *A Multi-Label Dataset of French Fake News: Human and Machine Insights*. 2024. arXiv: 2403.16099 [cs.CL]. URL: https://arxiv.org/abs/2403.16099.

[4] Louis Martin et al. "CamemBERT: a Tasty French Language Model". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. DOI: 10.18653/v1/2020.acl-main.645. URL: http://dx.doi.org/10.18653/v1/2020.acl-main.645.

[5] Telmo Pires, Eva Schlinger, and Dan Garrette. *How multilingual is Multilingual BERT?* 2019. arXiv: 1906.01502 [cs.CL]. URL: https://arxiv.org/abs/1906.01502.

[6] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[7] William Y. 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, pp. 422–426.

[8] Ben Wu et al. "SheffieldVeraAI at SemEval-2023 Task 3: Mono and Multilingual Approaches for News Genre, Topic and Persuasion Technique Classification". In: *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics, 2023, pp. 1995–2008. DOI: 10.18653/v1/2023.semeval-1.275. URL: http://dx.doi.org/10.18653/v1/2023.semeval-1.275.

6 Appendix A: TF-IDF Vectorizer for Fake News Qualification

The `TfidfVectorizer` is a classic yet powerful method for text vectorization, widely used in supervised classification tasks, particularly for fake news detection. It is based on the principle of *Term Frequency-Inverse Document Frequency* (TF-IDF), which aims to quantify the relative importance of a word in a document, adjusted by its rarity across the entire corpus.

TF-IDF Principle

TF-IDF weighting combines two components:

- **TF (Term Frequency)**: the frequency of a term t in a document d, calculated as:

$$tf(t,d) = \frac{\text{Number of occurrences of } t}{\text{Total number of words in } d}$$

- **IDF (Inverse Document Frequency)**: a measure of how rare a term is across the document set, defined as:

$$idf(t) = \log\left(\frac{N}{1 + df(t)}\right)$$

  where N is the total number of documents and $df(t)$ is the number of documents containing the term t.

The product of these two quantities yields the final score:

$$\text{TF-IDF}(t,d) = tf(t,d) \times idf(t)$$

This representation reduces the weight of frequent but uninformative words (e.g., "the", "is"), while emphasizing discriminative terms that are more specific to individual documents.

Application to Fake News Detection

In the context of fake news detection, the `TfidfVectorizer` is used to transform news articles or online content into numerical vectors that can be fed into machine learning models (e.g., logistic regression, support vector machines). Typical terms associated with misinformation (e.g., "hidden truth", "they don't want you to know") are automatically assigned higher importance when they appear frequently in fake news documents but remain rare in legitimate ones.

While TF-IDF does not capture semantic context like Transformer-based models, it remains highly competitive for classification tasks on well-annotated and relatively homogeneous corpora. It also offers key advantages in terms of training speed, interpretability, and resistance to overfitting.

The `TfidfVectorizer` thus provides a robust and efficient method for text-to-vector transformation in fake news detection pipelines. It serves as a valuable initial step before applying more complex architectures, and can even produce reliable results on its own when combined with well-tuned linear classifiers.

7 Appendix B: Label Attribution Frequency Across Annotators

Annotator	Fake News	Places Dates, People	Facts	Opinions	Subjective	Reported Information	Sources Cited	False Information	Insinuation	Exaggeration	Offbeat Title
rater1	0.21	0.90	0.93	0.80	0.71	0.14	0.75	0.08	0.22	0.37	0.03
rater2	0.29	0.94	0.99	0.93	0.74	0.21	0.82	0.09	0.24	0.44	0.21
rater3	0.64	1.00	0.96	0.74	0.63	0.14	0.68	0.46	0.46	0.61	0.14
rater4	0.34	1.00	1.00	0.87	0.52	0.22	0.87	0.32	0.54	0.42	0.16
rater5	0.47	0.93	0.96	0.66	0.67	0.76	0.76	0.22	0.73	0.53	0.17
rater6	0.34	0.75	0.93	0.72	0.60	0.21	0.60	0.11	0.42	0.35	0.04
rater7	0.29	1.00	0.91	0.45	0.47	0.66	0.78	0.24	0.43	0.35	0.01
rater8	0.44	1.00	0.99	0.66	0.48	0.11	0.46	0.34	0.59	0.52	0.08

Table 6: Proportion of articles assigned each label by annotator

The figure below presents a typical example of how different annotators labeled the same article. We observe that for factual labels such as `Facts`, `Opinions`, or `Places, Dates, People`, annotators tend to reach unanimous agreement. In contrast, labels that require greater subjective interpretation by the annotator exhibit more variation and disagreement.

ID	Fake News	Places, Dates, People	Facts	Opinions	Subjective	Reported Information	Sources Cited	False Information	Insinuation	Exaggeration	Offbeat Title
0	0	1	1	1	1	1	0	1	0	0	0
100	0	1	1	1	1	1	0	1	0	0	0
200	0	1	1	1	1	1	0	1	0	0	0
300	0	1	1	1	1	1	0	1	1	0	0
400	0	1	0	1	1	1	1	1	0	0	0
500	0	1	1	1	1	1	0	1	0	0	0
600	0	1	1	1	1	1	1	1	1	0	0
700	1	1	1	1	1	1	0	0	1	1	0

Table 7: Example of binary label annotation per article in the dataset.

8 Appendix C: Distribution of Labels Scores Across 100 Articles

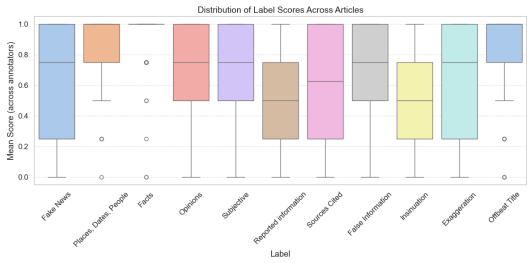

Figure 8: Distribution of Labels Across articles

9 Appendix D: Correlation Matrix Between Annotators' Decisions

	0	1	2	3	4	5	6	7
0	1.000000	0.373875	0.284387	0.407369	0.350735	0.407369	0.373875	0.235431
1	0.373875	1.000000	0.387501	0.285646	0.502048	0.332169	0.514327	0.499020
2	0.284387	0.387501	1.000000	0.450347	0.539305	0.450347	0.479326	0.371014
3	0.407369	0.285646	0.450347	1.000000	0.592993	0.509804	0.378691	0.341920
4	0.350735	0.502048	0.539305	0.592993	1.000000	0.592993	0.502048	0.497283
5	0.407369	0.332169	0.450347	0.509804	0.592993	1.000000	0.518257	0.384447
6	0.373875	0.514327	0.479326	0.378691	0.502048	0.518257	1.000000	0.454623
7	0.235431	0.499020	0.371014	0.341920	0.497283	0.384447	0.454623	1.000000

Figure 9: Inter-annotators Correlations

10 Appendix E: Logistic Regressions on `Exaggeration` and `False Information` Labels

Variable	Coef.	Std. Err.	z	P>|z|	[0.025, 0.975]
Intercept (const)	-12.273	2.925	-4.196	0.000	[-18.005, -6.540]
Number of Words	0.0017	0.005	0.380	0.704	[-0.007, 0.011]
Number of Characters	-0.0013	0.000	-3.181	0.001	[-0.002, -0.000]
Average Word Size	0.8368	0.243	3.445	0.001	[0.361, 1.313]
Number of Figures	0.0011	0.003	0.320	0.749	[-0.006, 0.008]
Number of Sentences	0.0390	0.009	4.322	0.000	[0.021, 0.057]
Number of Unique Words	0.0111	0.003	4.006	0.000	[0.006, 0.016]
Number of Stopwords	0.0017	0.007	0.248	0.804	[-0.012, 0.015]
Ratio of Stopwords	14.4577	5.922	2.441	0.015	[2.851, 26.065]
Sentiment Analysis	1.1925	0.840	1.419	0.156	[-0.454, 2.839]
Gate Hyperpartisan Label	0.2921	0.253	1.154	0.248	[-0.204, 0.788]
Gate Genre	-0.3568	0.152	-2.352	0.019	[-0.654, -0.060]

Table 8: Logistic regression on `Exaggeration` label from structural and model-based features.

Variable	Coef.	Std. Err.	z	P>|z|	[0.025, 0.975]
Intercept (const)	-11.9318	3.664	-3.257	0.001	[-19.113, -4.751]
Number of Words	0.0013	0.005	0.286	0.775	[-0.008, 0.011]
Number of Characters	-0.0008	0.000	-2.126	0.034	[-0.002, -0.00006]
Average Word Size	0.8441	0.262	3.228	0.001	[0.332, 1.357]
Number of Figures	-0.0014	0.005	-0.287	0.774	[-0.011, 0.008]
Number of Sentences	0.0183	0.010	1.874	0.061	[-0.001, 0.037]
Number of Unique Words	0.0090	0.003	2.939	0.003	[0.003, 0.015]
Number of Stopwords	0.0007	0.007	0.096	0.923	[-0.013, 0.015]
Ratio of Stopwords	12.4006	7.726	1.605	0.108	[-2.742, 27.544]
Sentiment Analysis	-0.0889	1.028	-0.087	0.931	[-2.104, 1.926]
Gate Hyperpartisan Label	0.7262	0.289	2.509	0.012	[0.159, 1.293]
Gate Genre	-0.3733	0.184	-2.025	0.043	[-0.735, -0.012]

Table 9: Logistic regression on `False Information` label from structural and model-based features.