Fake news detection models : Accuracy and generalization

Ahmed Khairaldin
ENSAE & Polytechnique
ahmed.khairaldin@ensae.fr

Imrane Zaakour
ENSAE & université Paris-Saclay
imrane.zaakour@ensae.fr

Abstract

The rapid spread of fake news poses a significant threat to public trust and informed decision-making in society. This project evaluates the performance and generalization ability of deep learning models-including RNN, LSTM, GRU, Transformer-based architectures, and zero-shot classification with pretrained language models-for detecting fake news in political articles. Using the ISOT and Kaggle datasets, we compare model accuracy and introduce a probabilistic framework to assess prediction confidence rather than relying solely on binary classification. Our results show that while models achieve high accuracy within individual datasets, their performance drops sharply-up to 50%-when tested across datasets, highlighting persistent challenges in generalizability. These findings emphasize the need for larger, more diverse datasets and the integration of linguistic and stylistic features to improve robustness. By focusing on model transparency and probabilistic outputs, this work contributes to the development of more reliable and ethically responsible fake news detection systems.

.

To understand this paper we put two important definitions given by [WD17] :

- *Information (or Influence) Operations* : Actions taken by governments or organized non-state actors to distort domestic or foreign political sentiment, most frequently to achieve a strategic and/or geopolitical outcome. These operations can use a combination of methods, such as false news, dis-information or networks of fake accounts aimed at manipulating public opinion (false amplifiers).
- *False News* : News articles that purport to be factual, but contain intentional misstatements of fact to arouse passions, attract viewership or deceive.

INTRODUCTION

The spread of fake news poses a major challenge for contemporary societies, impacting areas as diverse as public health, politics, and international security. False information can influence public opinion, disrupt elections, and even endanger lives, as seen with the widespread disinformation during the COVID-19 pandemic and recent conflicts [HK22]. In response to this threat, the automated detection of fake news using natural language processing (NLP) techniques has become an essential tool. However, while current models achieve promising performance in controlled settings, their ability to adapt to new contexts or different datasets—that is, their generalizability—remains a critical challenge for ensuring effectiveness in real-world applications. Understanding and improving this generalizability is therefore crucial for developing robust and reliable systems capable of combating large-scale disinformation effectively.

This project aims to implement the experiments outlined in the referenced paper while extending the research framework through several innovative dimensions. First, we plan to enhance the existing methodology by integrating advanced neural architectures like LSTM networks, Transformer-based models, and few-shot learning techniques with large language models (LLMs). These technical extensions seek to address the paper's identified limitation regarding model generalizability across datasets.

Moving beyond binary classification, our approach introduces a probabilistic evaluation framework that quantifies the likelihood of news articles being deceptive. This shift from discrete labels to continuous confidence scores provides critical insights into model certainty, particularly valuable for borderline cases where absolute classifications might be misleading. The probabilistic output also enables nuanced risk assessment – essential for real-world applications requiring graduated responses rather than binary decisions.

The ethical implications of automated deception detection drive our methodological refinements. By quantifying uncertainty through probability distributions rather than definitive judgments, we aim to create more transparent systems that acknowledge the complexity of truth verification. This aligns with the paper's emphasis on developing robust, real-world applicable solutions while addressing its observed performance drops (up to 50% accuracy reduction in cross-dataset testing). Our augmented methodology consequently operates at the intersection of technical innovation and responsible AI implementation.

RELATED WORK

Fake news detection represents a critical challenge in natural language processing (NLP), typically framed as a text classification problem. Current research employs diverse methodologies combining content analysis and contextual social signals[HK22]. Content-based approaches leverage textual features like N-grams, TF-IDF, word embeddings, linguistic patterns, stylometric markers, and psychological indicators [Hua23]. Contextual methods integrate social media interactions and user behavior data.

Technical implementations span classical machine learning algorithms to advanced deep learning architectures, including CNNs, RNNs, bidirectional LSTMs, and hybrid models. A significant trend involves Transformer-based pretrained models like BERT, GPT-2, and BART, which capture rich contextual representations through self-attention mechanisms. Reported performance varies widely across datasets and methods: bidirectional LSTM with GloVe embeddings achieved 99.95% accuracy on the ISOT dataset [DFHG19], while BERT-based models reached 91.96% accuracy in hybrid studies[SH25].

Despite high intra-dataset performance (typically 86-94% accuracy on holdout sets), models face severe generalization challenges. Cross-dataset evaluations reveal accuracy drops up to 50% when testing on same-topic news from different sources[Hua23, HK22]. This limitation stems from dataset constraints - limited size, temporal bias, and insufficient diversity - coupled with feature robustness issues. Word-level representations prove particularly vulnerable compared to linguistic/stylistic features.

Current research directions address early detection strategies and data scarcity through weak supervision techniques [OQW20]. Transformer architectures optimized with text summarization demonstrate promise, with RoBERTa achieving 98.39% accuracy in recent experiments [SH25]. However, the field continues grappling with ethical implementation challenges, particularly regarding probabilistic uncertainty quantification versus binary classification.

In this project, we begin by presenting the training data, which plays a crucial role in our study. Building a high-quality dataset is a challenge in itself and remains a complex task. We then describe the models we explored and their specific features. In particular, we chose to focus on architectures that were not discussed in [HK22], with the goal of making our work a complementary contribution rather than a repetition of previous findings. We highlight the results before introducing a new approach that leverages transformers to estimate the probability that an article is fake news. This idea also raises ethical considerations, which we discuss later. Finally, we examine the ability of transformer-based models to generalize from one dataset to another.

DATASETS

To assess the generalizability of fake news detection models, it is essential to rely on well-documented and thematically consistent datasets. In this study, we use two prominent datasets focused specifically on intentionally deceptive political news, deliberately excluding clickbait and satirical content to maintain a clear scope.
The ISOT Fake News Dataset, compiled by the University of Victoria, comprises 44,898 political news articles, with a clear distinction between fake (23,481 articles) and real (21,417 articles) news. The fake news samples are sourced from unreliable websites identified by platforms like Politifact and Wikipedia, while the real news articles come from reputable sources such as Reuters. Each entry includes the article's title, full text, topic, and publication date, providing a rich resource for model training and evaluation.
The Kaggle Fake News Dataset contains 20,800 articles, evenly split between fake and real news. While thematically similar to ISOT, this dataset is more heterogeneous in its sources and is sometimes less thoroughly documented regarding data collection methods. It includes article titles, full text, and binary labels but may aggregate content from various origins, which can introduce additional variability.
Using both datasets allows for a robust evaluation of model performance across different but related sources, providing a meaningful test of how well detection algorithms generalize within the domain of political fake news.

We also present some information about the datasets. [HK22] highlights the importance of generalization. There are three types of generalization. A preliminary analysis of the data shows that it is difficult to study model generalization across different time periods (ISOT contains only articles from 2016–2017, while the Kaggle dataset lacks temporal information). Similarly, the topics covered are relatively close and are heavily focused on U.S. politics, as illustrated by the two word clouds, (fig6 and fig8). These articles generally contain enough information (in terms of text length) to assess the truthfulness of their content, based on the distribution of text lengths (fig5 and fig7). Ultimately, while both datasets are sufficient for building accurate models on their own (as shown in the final section), they lack the diversity needed to support a constructive study on the generalization capabilities of models across other datasets.

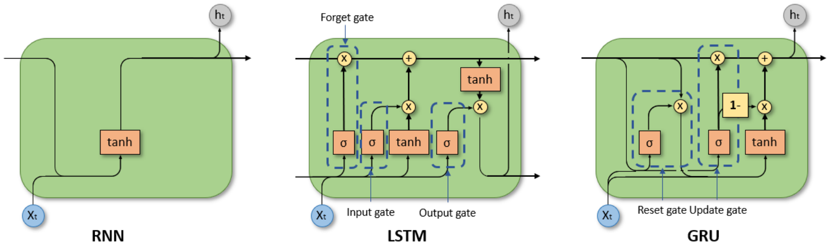

Figure 1: Comparison between RNN, LSTM, and GRU architectures, highlighting the differences in their memory cells and internal mechanisms. Source: adapted from a Medium article by Hassaan Idrees, *"RNN vs. LSTM vs. GRU: A Comprehensive Guide to Sequential Data Modeling"*, Medium, 2024.[2]

PROPOSED MODELS

In this section, we introduce several deep learning models to address the fake news detection task. Specifically, we implement five different approaches: a Recurrent Neural Network (RNN), two of its variants — Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), a Transformer-based model trained from scratch, and finally, a pretrained language model used for zero-shot binary classification.

Recurrent Neural Networks[1]

The first deep learning model we propose is a Recurrent Neural Network (RNN) [RHW+85]. RNNs are known as the counterparts of feedforward neural networks for sequential data. Since neural networks are universal approximators, RNNs are expected to perform well on sequence-based tasks such as ours. However, as we will observe in the next section, RNNs suffer from the vanishing gradient problem. During backpropagation through time, the gradient is propagated across many time steps and can become very small, especially for earlier tokens in the sequence making it hard to learn long-term dependencies.

To address this issue, we propose using two variants of RNNs specifically designed to mitigate this problem.

Long Short-Term Memory Model[1]

The Long Short-Term Memory model (LSTM) [HS97] is motivated by the vanishing gradient issue encountered in usual RNNs. Thus, it leverages memory cells that learns when to remember and when to forget information preserving gradient descent without its complete attenuation.

Figure 1 shows the LSTM memory cell, where the hidden states are passed between two tokens. The LSTM model uses a forget gate, an input gate, and an output gate, each of which is a simple neural network with its own set of weights. This introduces more parameters compared to traditional RNNs, allowing the model to learn long-term dependencies by selectively forgetting or remembering relevant information using the appropriate weights. Moreover, these gates perform linear operations (point-wise multiplication and addition) that allow for better gradient flow, preventing the gradients from vanishing.

Gated Recurrent Units[1]

Like LSTM, GRU [CGCB14] aims to learn long-term dependencies using a memory mechanism, but with a different internal structure. It uses only two gates — the reset gate and the update gate — unlike LSTM, which employs three. Moreover, GRU does not use a separate memory cell (as LSTM does with its cell state); instead, all information is stored and processed directly within a single hidden state, as illustrated in Figure 1. Each of the two gates has its own trainable weights.

[2] `https://medium.com/@hassaanidrees7/rnn-vs-lstm-vs-gru-a-comprehensive-guide-to-sequential-data-modeling-03aab16647bb`

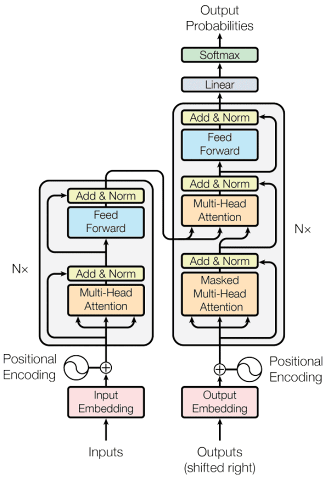

Figure 1: The Transformer - model architecture.

Figure 2: Architecture of the Transformer with several encoder and decoder blocks.

Moreover, while LSTMs tend to perform better on long sequences due to their ability to capture very long-term dependencies, GRUs are generally more efficient on shorter sequences. They require fewer parameters and consume less memory, making them faster to train.

Transformers[2]

We also propose to use a Transformer-based model, as this architecture has revolutionized Natural Language Processing (and even Computer Vision) with the rise of large language models (LLMs), which rely heavily on the attention mechanism first introduced by [VSP+17].

Our model uses multi-head self-attention and multiple encoder layers. In this setup, the model applies self-attention several times in parallel through different attention heads, allowing it to focus on different parts of the sequence simultaneously. The outputs of these heads are then concatenated and passed through a linear transformation. Since the Transformer has no recurrence, positional encoding is added to the input embeddings to inject information about the token order into the model.

We follow this with a decoder architecture that includes both multi-head and masked multi-head self attention as shown in Figure 1 . The decoder uses masked self-attention to prevent each position from attending to future tokens, ensuring that predictions are made autoregressively and only based on previously generated outputs. Finally, a linear layer followed by a sigmoid activation performs the binary classification task.

Zero-shot classification[2]

In addition to all these models, we propose leveraging pretrained language models for our fake news detection task. For this, we use zero-shot classification, which refers to models predicting labels they haven't been explicitly trained on (in our case, "Real" or "Fake"). Since these models are pretrained on large textual corpora, they can reason and understand both the texts and the labels, allowing them to infer the correct label for each text.

RESULTS

We evaluate the performance of the proposed models on the same test set for both datasets: the ISOT and the Kaggle "Real or Fake" datasets. During preprocessing, we removed special characters

[1] This part was implemented and tested by Ahmed.

[2] This part was implemented and tested by Imrane.

and converted all text to lowercase. Additionally, we limited the vocabulary to the 30,000 most frequent tokens and truncated each article to a maximum length of 512 tokens, due to computational constraints.

We implement RNN, LSTM, and GRU models, each using an embedding layer followed by a single hidden layer, with the hyperbolic tangent (tanh) as the activation function. For our Transformer-based model, we use two layers of encoder-decoder architecture with 4 heads in the multi-head self-attention mechanism for both modules. Finally, we leverage pretrained language models from Hugging Face, specifically the RoBERTa model, for zero-shot classification. This model is open-source and well-suited for our task.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
RNN	0.49	0.50	0.50	0.49
LSTM	0.99	0.99	0.99	0.99
GRU	0.99	1.00	0.99	0.99
Transformer	1.0	1.0	1.0	1.0
Zero-shot (RoBERTa)	0.56	0.67	0.58	0.50

Table 1: Performance comparison of the proposed models on the ISOT test set.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
RNN	0.58	0.60	0.58	0.56
LSTM	0.80	0.81	0.80	0.80
GRU	0.84	0.84	0.84	0.84
Transformer	0.93	0.93	0.92	0.92
Zero-shot (RoBERTa)	0.53	0.56	0.52	0.44

Table 2: Performance comparison of the proposed models on the Kaggle test set.

Tables 1 and 3 present the results of the proposed deep learning models on the ISOT and Kaggle datasets, respectively. Interestingly, zero-shot classification performs the worst among all models for the fake news binary classification task. Our experiments show that, while it tends to predict real articles accurately, it often fails to correctly identify fake ones. This suggests that even a powerful pretrained language model does not fully capture the subtle nuances of fake news. A fine-tuned version of such a model on our specific training data would likely yield significantly better results.

On the other hand, the Transformer-based model achieves the best performance on both datasets. In this case, the attention mechanism outperforms the memory cells used in LSTM and GRU models, as it captures contextual information more effectively.

GENERALIZABILITY

The ability to generalize across datasets is critical for fake news detection models to maintain real-world relevance, as overfitting to dataset-specific patterns risks missing evolving disinformation tactics. [HK22] reveals 50% accuracy drops when models trained on political news datasets like ISOT are tested on different same-topic datasets, despite achieving 80-99% holdout accuracy. Key factors driving this poor generalizability include:

- Dataset limitations: Small sizes (<50k articles) and source biases (e.g., ISOT's strict publisher-based labeling) create skewed feature distributions
- Feature sensitivity: Word-level representations (TF-IDF, BERT) capture topical biases rather than deception patterns, while linguistic features (readability scores, punctuation cues) show 15% better cross-dataset stability

Now, we propose the same study for transformers (which yielded the best results before). We observe similar effects when training on the Kaggle dataset and evaluating on ISOT. However, the effects are sometimes less pronounced than 50%. It is also worth noting that the results can vary from one iteration of the code to another. The training process is quite lengthy (approximately 20 minutes), which prevented us from repeating the experiment multiple times to compute an average. The best result yielded a generalization with an F1 score of 0.69 (a 23% decrease), but overall, we obtained the following scores:

Train	Test	Accuracy	Precision	Recall	F1 Score
Kaggle	ISOT	0.52	0.52	0.52	0.52
ISOT	Kaggle	0.52	0.52	0.52	0.52

Table 3: Performance of the Transformer for generalization question.

NEW APPROACH : PROBABILITIES

What does it really mean for an article to be labeled as fake news? In practice, models output a probability and apply a threshold to make a final decision. Yet, articles with probabilities of 0.49 and 0.51 are nearly identical in uncertainty—only one ends up being censored. Probabilistic models represent a critical advancement in fake news detection by expressing uncertainty rather than issuing binary labels. This allows for the identification of borderline cases (with probabilities close to 0.5) that warrant human review, helping to prevent both unjustified censorship and the spread of misinformation. Ethically, this approach avoids overly definitive judgments that might infringe on free speech, while also improving transparency in algorithmic decision-making. This also provides technical insight into the models, allowing us to distinguish them based on their confidence in predictions. Two different models may both achieve perfect classification on the same dataset, yet one may be significantly more certain in its outputs. Ultimately, the choice of model depends on the stakeholder and their specific priorities—whether it's caution, performance, minimizing false positives, or other concerns. We explore this concern with the example of transformers fig3 and fig4.

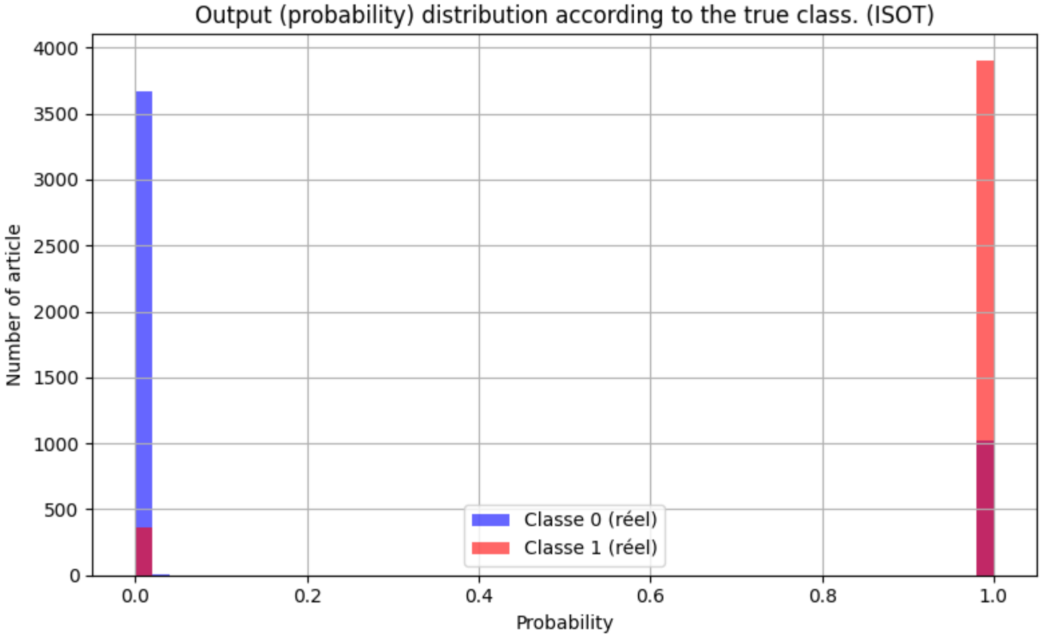

Figure 3: Distribution of probability according to the true class on ISOT dataset

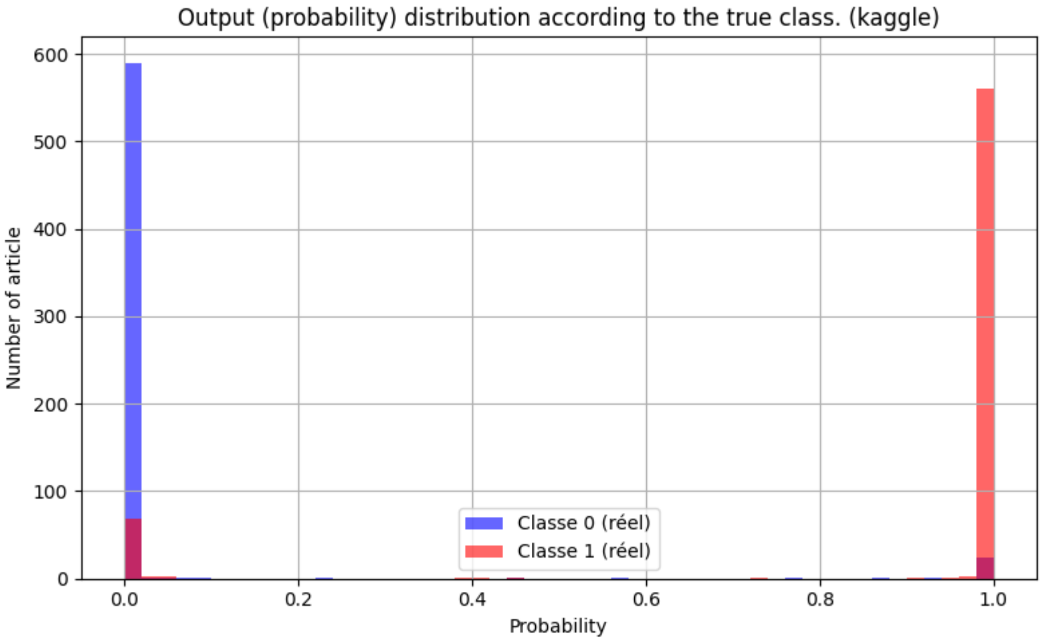

Figure 4: Distribution of probability according to the true class on kaggle dataset

Finally, with transformers, we observe that the probability values are mostly extremely close to 1 or 0. A key hypothesis for this phenomenon is the binary nature of the target, coded as 0 or 1. It would be interesting to revisit this protocol with data evaluated by a panel of experts who assign a probability to an article, rather than categorizing it directly. However, the results are still promising, as the model remains a good classifier. There is, nonetheless, a small number of articles that have ambiguous probabilities around 0.5.

CONCLUSION

Our results highlight that, while deep learning models such as Transformers can achieve near-perfect accuracy when evaluated on a single dataset, their performance drops sharply-sometimes by more than 30%-when tested on a different but thematically similar dataset. This underscores a central challenge: current models often overfit to dataset-specific patterns rather than learning generalizable features of deception. Improving generalizability will require larger, more diverse datasets and new approaches that go beyond word-level representations, such as incorporating linguistic or stylistic cues. Looking ahead, a key question remains: can we design fake news detection systems that remain robust and reliable when faced with new sources and evolving tactics of misinformation? A good approach could be to use articles published before a certain date for training, and those published after that date for testing, all from the same dataset.

References

[CGCB14] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[DFHG19] Ahlem Drif, Zineb Ferhat Hamida, and Silvia Giordano. Fake news detection method based on text-features. 08 2019.

[HK22] Nathaniel Hoy and Theodora Koulouri. Exploring the generalisability of fake news detection models. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5731–5740, 2022.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[Hua23] Lu Huang. Deep learning for fake news detection: Theories and models. In *Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering*, EITCE '22, page 1322–1326, New York, NY, USA, 2023. Association for Computing Machinery.

[OQW20] Ray Oshikawa, Jing Qian, and William Yang Wang. A survey on natural language processing for fake news detection. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France, May 2020. European Language Resources Association.

[RHW$^+$85] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning internal representations by error propagation, 1985.

[SH25] Belhadef H. Guessas A. Saadi, A. and O Hafirassou. Enhancing fake news detection with transformer models and summarization. In *Engineering, Technology Applied Science Research*, pages 23253–23259, 2025.

[VSP$^+$17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[WD17] Claire Wardle and Hossein Derakhshan. *INFORMATION DISORDER : Toward an interdisciplinary framework for research and policy making Information Disorder Toward an interdisciplinary framework for research and policymaking*. 09 2017.

APPENDICES

ISOT

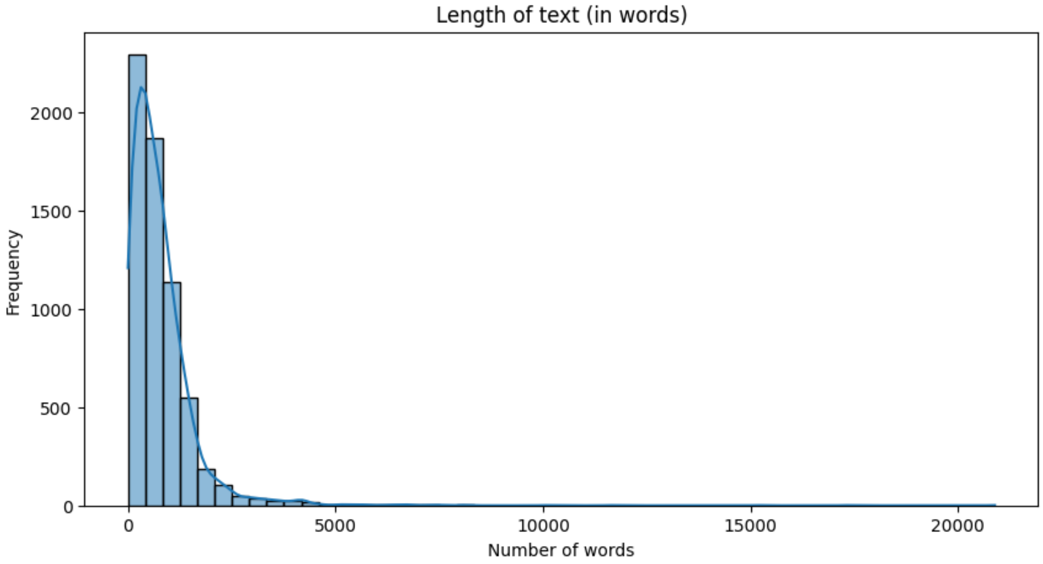

Figure 5: Length frequency for articles in ISOT

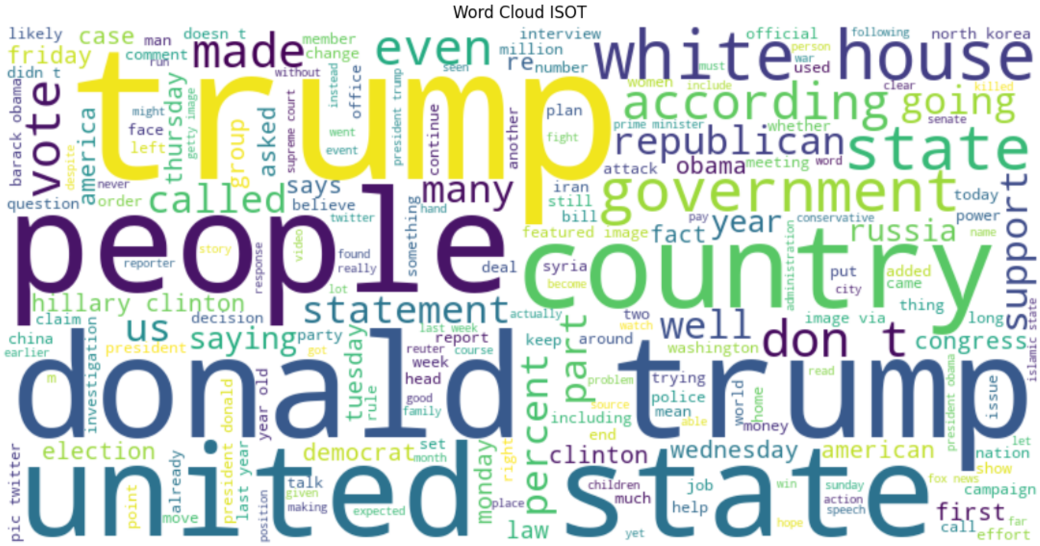

Figure 6: Word cloud for ISOT

Kaggle

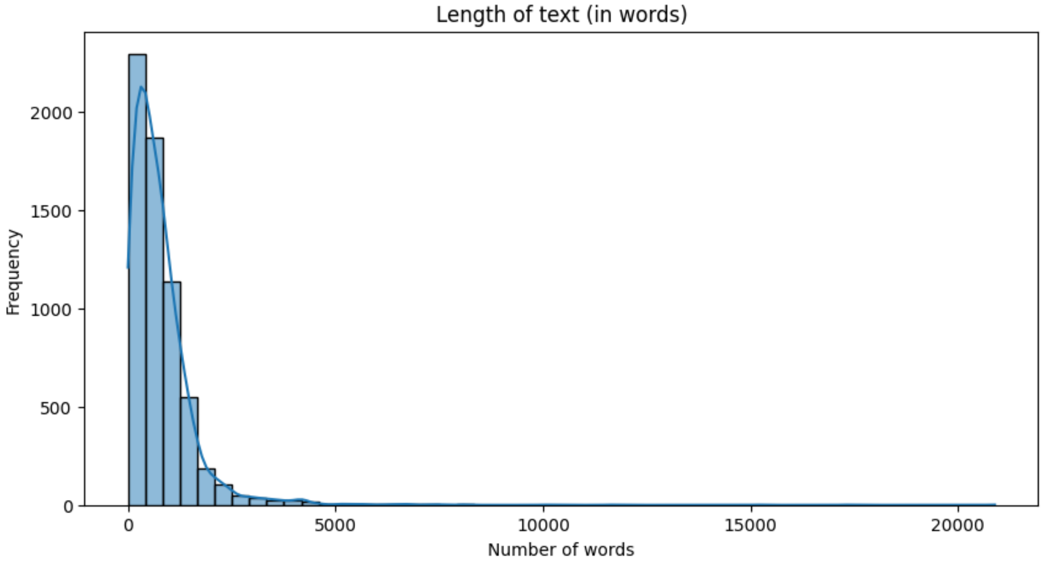

Figure 7: Length frequency for articles in kaggle

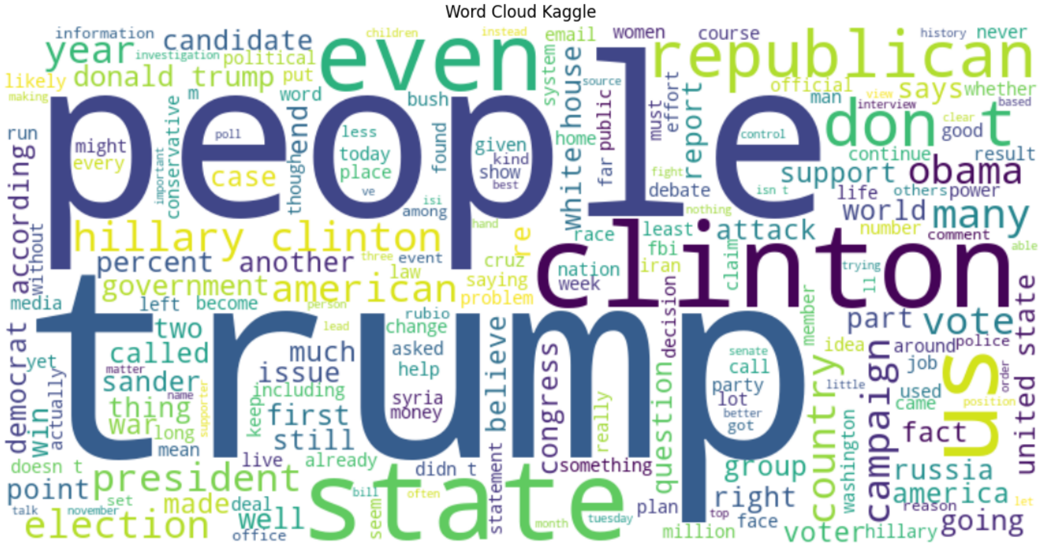

Figure 8: Word cloud for kaggle