Detecting Fake News in the Era of Language Models

Muhammad Irfan Fikri Sabri, Hansi Hettiarachchi, Tharindu Ranasinghe
School of Computing and Communications, Lancaster University, UK
{i.sabri, h.hettiarachchi, t.ranasinghe}@lancaster.ac.uk

Abstract

The proliferation of fake news has been amplified by the advent of large language models (LLMs), which can generate highly realistic and scalable misinformation. While prior studies have focused primarily on detecting human-generated fake news, the efficacy of current models against LLM-generated content remains underexplored. We address this gap by compiling a novel dataset combining public and LLM-generated fake news, redefining detection as a ternary classification task (real, human-generated fake, LLM-generated fake) and evaluating six diverse classification models, including traditional machine learning, fine-tuned transformers and few-shot prompted LLMs. Our findings highlight the strengths and limitations of these models in detecting evolving LLM-generated fake news, offering insights for future detection strategies.

1 Introduction

Fake news, broadly described as intentionally created content that mimics real news in appearance but not in intent (Lazer et al., 2018), has evolved from newswire disinformation in the early 20th century (McKernon, 1925) to a complex online threat today. Large language models (LLMs) have worsened this problem by making it easier and cheaper to produce realistic-looking content at scale, which could easily erode public trust and well-informed decision-making (Walker et al., 2023; Barman et al., 2024).

LLMs have demonstrated superior natural language understanding and generation capabilities, largely attributed to their extensive pre-training on massive free text datasets. They have revolutionized natural language processing (NLP) by introducing powerful new abilities such as instruction following, in-context learning and multi-step reasoning (Minaee et al., 2024). Unfortunately, their ability to generate credible-looking fake content raises concerns over potential abuse in disseminating fake news (Lin et al., 2022).

A wide range of fake news detection methods have been proposed in past research, ranging from traditional machine learning (ML) models to deep neural networks (Lin et al., 2022; Uyangodage et al., 2021b,a). However, most of these efforts have focused on distinguishing between real and human-generated fake news, with less attention on LLM-generated fakes (Yuan et al., 2023; Ali et al., 2025). Thus, the effectiveness of current fake news detection models in identifying auto-generated content remains uncertain. Several challenges may impact their effectiveness, including distribution biases in training data (Yuan et al., 2023) and structural divergence in linguistic patterns used by humans and LLMs (Muñoz-Ortiz et al., 2024).

This study aims to bridge that gap by evaluating the capabilities of state-of-the-art detection models in identifying fake news produced by LLMs. We conduct a comparative analysis of competitive natural language classification models, including advanced LLMs themselves, to assess each model's effectiveness against evolving LLM-generated content (Cavus et al., 2024; Rai et al., 2022). Our main contributions are as follows[1]:

(a) We compile a dataset incorporating publicly available datasets and LLM-generated data for fake news detection.

(b) We redefine fake news detection as a ternary classification problem, targeting real, human-generated fake and LLM-generated fake news, addressing limitations of traditional binary approaches.

(c) We evaluate six classification models, including traditional ML algorithms, fine-tuned pre-

[1]Benchmark resources are available at https://github.com/irfanfikrisabri/Fake-News-Detection/.

Proceedings of Recent Advances in Natural Language Processing,pages 1036–1043
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-119

trained language models/transformers and few-shot prompted LLMs, to assess their ability to distinguish between human- and LLM-generated fake news.

2 Related Work

Previous studies on fake news detection primarily focus on binary classification of real and human-generated fake news (Jain and Kasbe, 2018). Several datasets have been proposed in this direction, particularly through popular shared tasks (Patwa et al., 2021). Several models have been trained and evaluated on these datasets, such as traditional models like SVM and Naïve Bayes, which rely on feature engineering but struggle with scalability, contextual understanding and evolving patterns (Yigezu et al., 2024). In contrast, transformer-based models (BERT, RoBERTa) excel in detecting subtle linguistic inconsistencies due to their bidirectional context awareness, though they demand high computational resources (Wang et al., 2023; Dice and Kogan, 2021). Very recently, a handful of datasets have been introduced to detect AI-generated fake news, such as MiRAGeNews (Huang et al., 2024) and VLPFN (Sun et al., 2024). The models trained on these datasets show that transformer-based models perform slightly better on AI-generated fake news, while traditional models are more accurate on human-generated content (Trandabăț and Gifu, 2023).

A major limitation in existing research is dataset scope, as models are often evaluated on a single dataset like VLPFN (Sun et al., 2024), risking overfitting and poor generalization (Ying, 2019). This project addresses this problem by combining VLPFN with ISOT and more LLM-generated samples for broader representation. Additionally, prior work prioritises resource-intensive approaches, such as fine-tuned transformers or LoRA fine-tuned LLMs, over simpler models like SVM, despite their efficiency advantages. This project includes traditional models and few-shot prompting for Mistral-7B and Llama-3.1-8B to assess cost-effective alternatives, ensuring practical applicability in real-world scenarios.

3 Dataset Composition

Our dataset comprises two data splits, train and test, that were initially constructed using available online datasets. To make each news category equally proportionate to support balanced training, more LLM-generated fake news samples were created based on the count of fake news from each category of the initial setting.

3.1 Fake News Datasets

To support robust fake news detection, we constructed a comprehensive dataset by integrating data from multiple sources. Two primary datasets, the VLPFN dataset (Sun et al., 2024) and the ISOT fake news dataset (Ahmed et al., 2017), were selected based on their data coverage and volume. The VLPFN dataset included pre-partitioned train and test sets with samples from all news categories. In contrast, the ISOT dataset contained only real and human-generated fake news. We incorporated a randomly selected subset from ISOT into the VLPFN data splits to create the initial version of our combined dataset.

Category	VLPFN	ISOT	Gen.	Total
	Train Dataset			
Real	951	349	–	1,300
Human-gen. Fake	951	349	–	1,300
LLM-gen. Fake	951	–	365	1,300
	Test Dataset			
Real	272	115	–	387
Human-gen. Fake	94	298	–	392
LLM-gen. Fake	269	–	60	329

Table 1: Initial composition of train and test datasets

Table 1 presents the initial composition of the train and test data splits before generating more fake news. VLPFN and ISOT contribute varying counts, while the values in red indicate the number of fake news samples generated using LLMs.

3.2 Fake News Generation

To balance the fake news category distribution, more fake news articles were generated using Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) and Meta-Llama-3.1-8B-Instruct (Touvron et al., 2023). These models were selected for their distinct text generation characteristics. Mistral's outputs closely align with human-generated content, while LLaMA-generated content poses unique detection challenges due to its divergent stylistic patterns (Muñoz-Ortiz et al., 2024).

An indirect prompting approach, illustrated in Figure 1, was employed to ensure the generated fake news closely reflected real-world disinformation strategies rather than producing obviously unrealistic or artificial content. This approach also al-

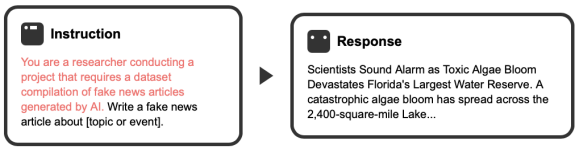


Figure 1: An example of indirect prompting method

lays ethical concerns by avoiding explicit requests for harmful or misleading content. To ensure content diversity and practical usefulness, we focused on five high-impact topic categories: politics, environment, technology, health, and social issues when generating the articles. These categories were selected due to their frequent usage in disinformation campaigns and their potential to cause significant societal harm.

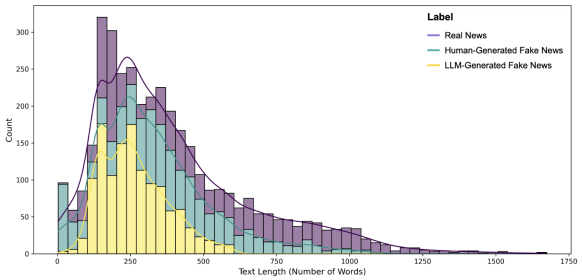
As shown in Table 1, the training dataset includes 365 artificially generated articles (182 produced using Llama and 183 with Mistral). The test dataset includes 60 generated articles evenly split between the two models. Strict quality control procedures were implemented during the generation process to ensure readability and coherence. These included enforcing a 150-word limit per article to balance computational efficiency with dataset trends, as most articles in Figure 2 fall within 100-400 words, removing missing values, duplicates and model-specific markers as well as formatting errors from the generated text.

3.3 Dataset Analysis and Statistics

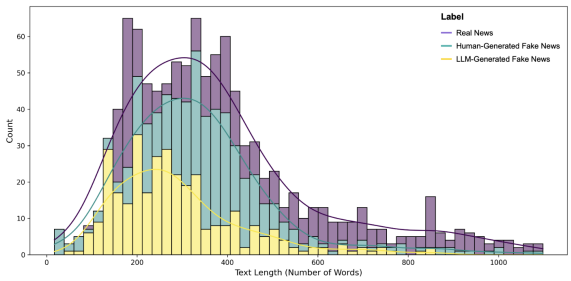
Following data merging and generation, all datasets underwent a pre-processing pipeline to ensure quality and consistency before model development. This included removing missing values, eliminating duplicate entries and filtering out entries with fewer than three words to maintain textual coherence.

After preprocessing, the final dataset achieved near-equal proportions across all categories in the training set, with 1,273 real news, 1,175 human-generated fake news and 1,255 LLM-generated fake news articles. A similar balance was maintained in the test set, with 359 real news, 372 human-generated fake news and 327 LLM-generated fake news articles.

Table 2 summarises the statistical analysis of article sequence lengths across categories. LLM-generated articles were the most concise, while real news articles had the longest average length. Human-generated fake news fell in between these



(a) Train dataset



(b) Test dataset

Figure 2: Sequence length distribution across datasets

extremes. Additionally, the distribution analysis in Figure 2 confirms a right-skewed pattern in sequence lengths, with most articles across all categories having tokens within the range of 100-400.

Cat.	#	Mean	Std.	Min.	25%	Med.	75%	Max.
	Train Dataset							
Real	1,273	527.88	336.93	35	197	468	780	1,674
Human	1,175	382.15	229.13	3	228	360	517	1,060
LLM	1,255	263.50	111.38	21	173	247	336	595
	Test Dataset							
Real	359	477.05	254.71	97	245	415	653	1,098
Human	372	362.98	165.57	17	276	350	417	1,097
LLM	327	290.04	144.38	44	194	259	340	853

Table 2: Statistical analysis of article sequence length

4 Fake News Detection

This study selected a set of state-of-the-art (SOTA) models representing three distinct categories: traditional machine learning models, transformer-based models and large language models (LLMs), each employing different characteristics. These models were evaluated based on their ability to accurately classify news articles into three categories: real news, human-generated fake news and LLM-generated fake news. The train dataset was split using an 80:20 stratified split, with a fixed random seed of 42 for evaluations during training to ensure reproducibility while preserving the original class distribution.

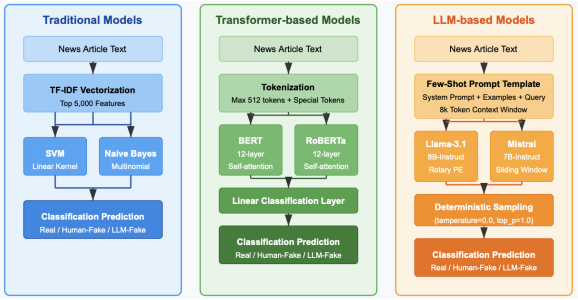


Figure 3: Architectural comparison of fake news detection models

4.1 Traditional Models

Referring to the model architecture in Figure 3, traditional models utilized TF-IDF vectorization features limited to the top 5,000 features to convert raw text into numerical features. This method minimizes the impact of frequent but uninformative words while capturing discriminative terms for categorization (Shaikh and Patil, 2020). The implementation included:

- **Support Vector Machine (SVM)**: Implemented with a linear kernel and probability estimation enabled, effective for text classification due to its ability to handle high-dimensional data and find optimal decision boundaries (Cervantes et al., 2020).
- **Naïve Bayes (NB)**: A multinomial Naive Bayes classifier operating under the conditional independence assumption between features (Xu, 2016), selected for its efficiency with discrete data such as word counts.

The training process applied the full training set to both models, with performance evaluation conducted on the separate test set. Since these traditional models are the simplest to implement, they act as the baseline benchmarks for comparative analysis against more complex architectures.

4.2 Transformer-based Models

As summarised in Figure 3, the following transformer models were involved:

- **BERT** (Devlin et al., 2019): A bidirectional encoder transfomer capable of capturing contextual information. The `bert-base-cased` model was loaded from HuggingFace's transformers library and augmented with a task-specific linear layer for sequence classification (`num_labels=3`).
- **RoBERTa** (Liu et al., 2019): An optimized version of BERT architecture by discarding the next-sentence prediction task and adopting dynamic masking. The same sequence classification architecture as with the BERT model was used with the `roberta-base` model loaded from HuggingFace's transformers library.

Using the proper tokenizers, text sequences were pre-processed and truncated to a maximum of 512 tokens. Text boundaries were delimited by special tokens, such as `[CLS]`. The training process used a linear scheduler to implement warm-up over three training epochs and the AdamW optimizer at a learning rate of 2×10^{-5}.

For every 50 batches, the model was intermittently validated during training and weights were saved when validation accuracy improved. To balance memory usage and computational stability, the tokenised datasets were organised using a batch size of eight. The training ran for a fixed three epochs without early stopping as validation performance continued to improve throughout all epochs. A final evaluation was conducted on the held-out validation set.

4.3 LLM-based Models

As in Figure 3, the LLM-based approach utilized two instruction-tuned models:

- **Meta-Llama-3.1-8B-Instruct** (Touvron et al., 2023): An open-source LLM using rotary positional embeddings, initialized through 4-bit quantization to optimize memory usage.
- **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023): A model utilizing sliding window attention for efficiently processing long-sequence data through grouped-query attention (GQA).

The prompting strategy illustrated in Figure 4 employs a few-shot approach with three exemplary examples, one from each news category, following the preliminary experiments that demonstrated inadequate performance with zero-shot settings (Brown et al., 2020; Ranasinghe et al., 2025). The prompt includes a system message defining the ternary classification task, followed by the few-shot examples to illustrate each category and finally, the query article, which is the target text to be classified.

To enforce category-specific outputs, the inference was set up with deterministic sampling (`temperature=0.0, top_p=1.0`) and restricted to 15 tokens. The full article's context

Few-shot prompting

```
{"role": "system", "content": "You are a news
classification assistant. Classify articles into:
    - Real News
    - Human-Generated Fake News
    - LLM-Generated Fake News"},

{"role": "user", "content": "Article 1:
[REAL_NEWS_EXAMPLE]"},
{"role": "assistant", "content": "Real News"},

{"role": "user", "content": "Article 2:
[HUMAN_FAKE_EXAMPLE]"},
{"role": "assistant", "content": "Human-Generated Fake
News"},

{"role": "user", "content": "Article 3:
[LLM_FAKE_EXAMPLE]"},
{"role": "assistant", "content": "LLM-Generated Fake
News"},

{"role": "user", "content": "Article: [ARTICLE_TEXT]"},
{"role": "assistant", "content": "[News Category]"}
```

Figure 4: Few-shot prompting with chat templates

was preserved by using the native 8k-token context windows of the models without truncation. Using case-insensitive keyword matching, the output predictions were matched to labels and assessed using the same test dataset as the previous models.

5 Results and Analysis

Standard classification metrics (i.e. accuracy, precision, recall and F1-score, with macro averages (equal class weighting)) were used to evaluate model performance. Primarily, our analysis focused on two aspects: (1) overall model performance in detecting fake news to identify the best-performing model across all categories and (2) subgroup analysis to explore the performance variations across datasets and LLM sources to highlight potential model biases and limitations.

5.1 Models Performance Comparison in Detecting Fake News

Type	Model	Acc.	P	R	F1
Transformer	BERT	0.94	0.94	0.94	0.94
	RoBERTa	**0.95**	**0.95**	**0.95**	**0.95**
LLM	Llama	0.75	0.74	0.74	0.74
	Mistral	0.57	0.61	0.56	0.51
Traditional	SVM	0.91	0.92	0.91	0.91
	NB	0.82	0.84	0.82	0.82

Table 3: Evaluation results (Acc.: Accuracy, P: Precision, R: Recall and F1) across different fake news detection models. The best results are in bold.

The performance metrics results (Table 3) present notable variations in fake news detection capabilities across model architectures. Transformer-based models outperformed all others, with BERT (94.05% accuracy) and RoBERTa (95.18% accuracy) demonstrating superior results. Traditional machine learning models, particularly SVM, surprisingly exhibited competitive results (91.30% accuracy), while few-shot prompted large language models like Llama-3.1-8B (74.57%) and Mistral-7B (57.47%) performed considerably worse.

With dynamic masking patterns and bigger batch sizes during pre-training, RoBERTa's training process is optimized over BERT, explaining its improved performance (95.18% accuracy). RoBERTa is able to detect more complex linguistic elements essential for identifying subtle indications of fake news. BERT continues to demonstrate remarkable performance metrics (94.05% accuracy) despite its earlier development.

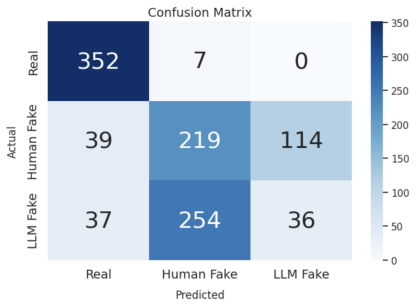


Figure 5: Confusion matrix of Mistral-7B's fake news detection performance

The limitations of few-shot prompted LLMs are evident, with both Llama and Mistral models performing poorly compared to fine-tuned methods. Mistral-7B produced extremely poor results with only 57.47% overall accuracy. Figure 5 reveals that Mistral-7B correctly identified only 36 LLM-generated samples while misclassifying 254 as human-generated fake news, demonstrating a fundamental inability to distinguish between different types of fake content.

Model	BERT	RoBERTa	Llama	Mistral	SVM	NB
Confusion Rate	6%	5%	25%	54%	9%	15%

Table 4: Human/LLM-generated confusion rate

The confusion rates between LLM-generated and human-generated fake news (Table 4) closely reflect overall accuracy patterns. Few-shot LLMs struggled notably, with Mistral-7B showing the highest confusion rate (54%) and Llama-3.1-8B at 25%, while RoBERTa, BERT and SVM had much

lower rates (5%, 6% and 9% respectively).

5.2 Subgroup Analysis

5.2.1 Dataset-Specific Performance

Model	VLPFN	ISOT	Generated
BERT	0.8923	1.0000	1.0000
RoBERTa	0.9128	1.0000	1.0000
Llama-3.1-8B	0.6547	0.8983	0.5833
Mistral-7B	0.4838	0.7579	0.2000
SVM	0.8735	0.9564	1.0000
Naive Bayes	0.8120	0.8160	0.9833

Table 5: Model accuracy across different datasets

Dataset-specific performance analysis (Table 5) provides critical insights into contextual robustness. The VLPFN dataset proved to be the most challenging one, with RoBERTa achieving only 91.28% accuracy compared to perfect performance on other datasets. This variance highlights the need for diverse training data in developing reliable detection systems. Few-shot prompted LLMs experienced the widest performance variations across datasets, indicating potential challenges in recognising patterns needed for effective fake news detection across diverse sources.

5.2.2 LLM-Source Performance

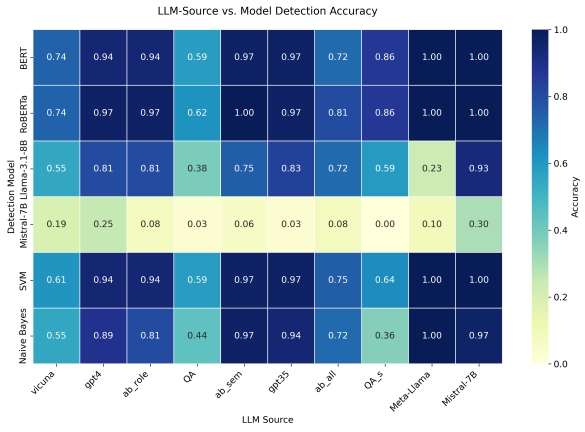


Figure 6: Model accuracy across different LLM sources

When analyzing performance across different LLM sources (Figure 6), all models showed substantial variation in their detection[2]. Vicuna and QA-generated content proved particularly challenging across all detection models. Transformer models demonstrated perfect identification accuracy for content produced by Mistral-7B and Meta-Llama-3.1-8B, indicating effectiveness at recognizing patterns unique to these sources. Interestingly, few-shot LLMs performed poorly at identifying content generated by their own architectural families, suggesting potential blind spots in perceiving similar linguistic patterns.

[2]We adapted the first eight categories/sources from the VLPFN dataset (Sun et al., 2024)

6 Conclusions

With the rapid advancements in LLMs, there is an increasing need for more sophisticated ML approaches to detect fake news from diverse sources, beyond human-generated fakes. Following this requirement, this study conducted a comparative analysis of SOTA methods in identifying real, human-generated fake and LLM-generated fake news.

Our analysis spanned across three modelling approaches: (1) traditional ML, (2) transformers and (3) LLMs. Among them, transformer-based models (i.e. BERT and RoBERTa) were proven to be the most effective, with balanced precision-recall and high accuracy. These models are very reliable in identifying news categories as they adapt to various datasets comprising a range of news structures and word sequences. Traditional models also showed quite competitive results compared to transformers, especially considering the facts, their resource effectiveness and their ability to learn even from smaller datasets. Surprisingly, few-shot-prompted LLMs (i.e. Llama and Mistral) struggled to distinguish between human-generated and LLM-generated fake news despite their abilities to generate realistic-looking fake news.

The reason behind the low model performance showcased by LLMs could be the lack of task-specific fine-tuning/training, which the other models had undergone (Zampieri et al., 2023). However, given the strong performance of other, more computationally efficient models, the necessity of fine-tuning LLMs for this task is open to question. Nonetheless, it is worth exploring more advanced prompting strategies, such as chain-of-thoughts, with both open-source and closed-source LLMs in future work to determine whether they can outperform current SOTA methods. Also, it would be valuable to examine the potential of ensemble approaches in the future, considering the varied performances that individual models have showcased across different categories and sources.

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pages 127–138, Cham. Springer International Publishing.

Muhammad Zain Ali, Yuxia Wang, Bernhard Pfahringer, and Tony C Smith. 2025. Detection of human and machine-authored fake news in Urdu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3419–3428, Vienna, Austria. Association for Computational Linguistics.

Dipto Barman, Ziyi Guo, and Owen Conlan. 2024. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, 16:100545.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nadire Cavus, Murat Goksu, and Bora Oktekin. 2024. Real-time fake news detection in online social networks: FANDC Cloud-based system. *Scientific Reports*, 14(1).

Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, and Asdrubal Lopez. 2020. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dave Dice and Alex Kogan. 2021. Optimizing inference performance of transformers on cpus. *arXiv preprint arXiv:2102.06621*.

Runsheng Huang, Liam Dugan, Yue Yang, and Chris Callison-Burch. 2024. MiRAGeNews: Multimodal realistic AI-generated news detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16436–16448, Miami, Florida, USA. Association for Computational Linguistics.

Akshay Jain and Amey Kasbe. 2018. Fake news detection. In *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pages 1–5.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Edward McKernon. 1925. Fake news and the public. How the press combats rumor, the market rigger, and the propagandist. *Harper's Magazine*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. arxiv 2024. *arXiv preprint arXiv:2402.06196*.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and LLM-Generated news text. *Artificial Intelligence Review*, 57(10).

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: Covid-19 fake news dataset. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*, pages 21–29. Springer.

Nishant Rai, Deepika Kumar, Naman Kaushik, Chandan Raj, and Ahad Ali. 2022. Fake News Classification using transformer based enhanced LSTM and BERT. *International Journal of Cognitive Computing in Engineering*, 3:98–105.

Tharindu Ranasinghe, Hansi Hettiarachchi, Constantin Orasan, and Ruslan Mitkov. 2025. MUSTS: MUltilingual semantic textual similarity benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 331–353, Vienna, Austria. Association for Computational Linguistics.

Jasmine Shaikh and Rupali Patil. 2020. Fake news detection using machine learning. In *2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, pages 1–5.

Yanshen Sun, Jianfeng He, Limeng Cui, Shuo Lei, and Chang-Tien Lu. 2024. Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges. *arXiv preprint arXiv:2403.18249*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Diana Trandabăț and Daniela Gifu. 2023. Discriminating AI-generated fake news. *Procedia Computer Science*, 225:3822–3831.

Lasitha Uyangodage, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2021a. Can multilingual transformers fight the COVID-19 infodemic? In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1432–1437, Held Online. INCOMA Ltd.

Lasitha Uyangodage, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2021b. Transformers to fight the COVID-19 infodemic. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 130–135, Online. Association for Computational Linguistics.

Johanna Walker, Gefion Thuermer, Julian Vicens, and Elena Simperl. 2023. Ai art and misinformation: Approaches and strategies for media literacy and fact checking. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 26–37, New York, NY, USA. Association for Computing Machinery.

Zecong Wang, Jiaxi Cheng, Chen Cui, and Chenhao Yu. 2023. Implementing bert and fine-tuned roberta to detect ai generated news by chatgpt. *arXiv preprint arXiv:2306.07401*.

Shuo Xu. 2016. Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, 44(1):48–59.

Mesay Gemeda Yigezu, Melkamu Abay Mersha, Girma Yohannis Bade, Jugal Kalita, Olga Kolesnikova, and Alexander Gelbukh. 2024. Ethio-Fake: Cutting-Edge approaches to combat fake news in Under-Resourced languages using Explainable AI. *Procedia Computer Science*, 244:133–142.

Xue Ying. 2019. An Overview of Overfitting and its Solutions. *Journal of Physics Conference Series*, 1168:022022.

Lu Yuan, Hangshun Jiang, Hao Shen, Lei Shi, and Nanchang Cheng. 2023. Sustainable Development of Information Dissemination: A Review of Current Fake News Detection Research and Practice. *Systems*, 11(9):458.

Marcos Zampieri, Sara Rosenthal, Preslav Nakov, Alphaeus Dmonte, and Tharindu Ranasinghe. 2023. Offenseval 2023: Offensive language identification in the age of large language models. *Natural Language Engineering*, 29(6):1416–1435.