

Med-MMHL: A Multi-Modal Dataset for Detecting Human- and LLM-Generated Misinformation in the Medical Domain

Yanshen Sun*
yansh93@vt.edu
Virginia Tech
Falls Church, VA, USA

Jianfeng He*
jianfenghe@vt.edu
Virginia Tech
Falls Church, VA, USA

Limeng Cui†
culimeng@amazon.com
Amazon
Palo Alto, CA, USA

Shuo Lei
slei@vt.edu
Virginia Tech
Falls Church, VA, USA

Chang-Tien Lu
ctlu@vt.edu
Virginia Tech
Falls Church, VA, USA

ABSTRACT

The pervasive influence of misinformation has far-reaching and detrimental effects on both individuals and society. The COVID-19 pandemic has witnessed an alarming surge in the dissemination of medical misinformation. However, existing datasets pertaining to misinformation predominantly focus on textual information, neglecting the inclusion of visual elements, and tend to center solely on COVID-19-related misinformation, overlooking misinformation surrounding other diseases. Furthermore, the potential of Large Language Models (LLMs), such as the ChatGPT developed in late 2022, in generating misinformation has been overlooked in previous works. To overcome these limitations, we present Med-MMHL, a novel multi-modal misinformation detection dataset in a general medical domain encompassing multiple diseases. Med-MMHL not only incorporates human-generated misinformation but also includes misinformation generated by LLMs like ChatGPT. Our dataset aims to facilitate comprehensive research and development of methodologies for detecting misinformation across diverse diseases and various scenarios, including human and LLM-generated misinformation detection at the sentence, document, and multi-modal levels. To access our dataset and code, visit our GitHub repository: <https://github.com/styxsys0927/Med-MMHL>.

CCS CONCEPTS

• **Information systems** → **Multimedia databases**.

KEYWORDS

Medical misinformation, news, tweets, LLM, multimodal, dataset

*These two authors are co-first authors with equal contributions to this work.

†This author's contribution to this work was made prior to her employment at Amazon.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).
Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06... \$15.00
<https://doi.org/10.1145/3539618.XXXXXX>

ACM Reference Format:

Yanshen Sun, Jianfeng He, Limeng Cui, Shuo Lei, and Chang-Tien Lu. 2018. Med-MMHL: A Multi-Modal Dataset for Detecting Human- and LLM-Generated Misinformation in the Medical Domain. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3539618.XXXXXX>

1 INTRODUCTION

Misinformation, defined as wrong information compared to the original and verified data, has been proven to have a significantly negative impact on both society and individuals, as supported by recent surveys [8, 22, 51, 52]. Within the scope of misinformation, the medical domain is particularly crucial, as misinformation in this domain, including COVID-19, directly influences individual treatments and national policies. For instance, misinformation suggesting that drinking bleach protects against COVID-19 [33] has misguided individuals into using harmful substances [6]. Similarly, misinformation asserting that vaccines against SARS-CoV-2 cause infertility [1] has impeded the swift implementation of vaccine policies [6]. Given the significant negative impact of medical misinformation on individuals and society, it is essential to research in medical misinformation detection [9, 19, 36, 47]. To facilitate such research, we create a novel medical misinformation detection dataset by overcoming three limitations of previous datasets.

Original Information	I agree that #COVID19 pandemic has been the most disastrous crisis of our generation and that we must apply its lessons to make the world a safer place.
ChatGPT 3.5 Generated Misinformation	I vehemently disagree that the #COVID19 pandemic has been the most disastrous crisis of our generation, and I argue that it has been blown out of proportion. There is no need to learn any lessons from it as it is an overhyped event. Instead, we should prioritize other issues to make the world a safer place.

Figure 1: An example that ChatGPT can generate misinformation given a text from the medical domain.

Previous datasets pertaining to medical misinformation exhibit three notable limitations. Firstly, many of these datasets focus solely on textual information, such as news or tweets [11, 13, 14, 23, 26]. However, they omit additional visual information beyond text, which can enhance task performance [28, 38]. Secondly, a majority of the previous datasets concentrate exclusively on COVID-19 misinformation, disregarding misinformation surrounding other diseases [11, 13, 23, 28, 49, 50]. Considering the distinct symptoms

Table 1: Comparison between our Med-MMHL dataset and the other datasets.

	Multi-Disease/General Medical	Multi-Modal	LLM-fake	News	Social Media	Date Start	Date End
MedHelp [26]	✓	×	×	×	✓	2001	2013
COAID [13]	×	×	×	✓	✓	2019-Dec	2020-Sep
MM-COVID [28]	×	✓	×	✓	✓	2020-Feb	2020-Jul
CHECKED [49]	×	×	×	×	✓	2019-Dec	2020-Aug
TruthSeeker [14]	✓	×	×	×	✓	2009	2022
ANTi-Vax [23]	×	×	×	×	✓	2020-Dec	2021-Jul
COVID-Rumor [11]	×	×	×	✓	✓	2020-Jan	2020-Dec
ReCOVery [50]	×	✓	×	✓	✓	2020-Jan	2020-May
Monant [43]	✓	✓	×	✓	✓	1998-Jan	2022-Feb
Med-MMHL(Ours)	✓	✓	✓	✓	✓	2017-Jan	2023-May

and treatments associated with different diseases, it is important to detect medical misinformation with generalization, as what applies to COVID-19 may not be applicable to other medical conditions. Furthermore, the majority of previous datasets solely focus on human-generated misinformation, neglecting the emergence of LLMs like ChatGPT [34] as generators of misinformation. However, since the release of ChatGPT in Nov. 2022, it has demonstrated remarkable text generation capabilities across various domains [7, 21, 30, 44]. Notably, our findings indicate that ChatGPT can generate medical misinformation, as depicted in Figure 1. Considering the potential of LLMs like ChatGPT to generate misinformation—evidenced by their capability to fabricate material for legal domain [45] and produce artificially constructed peer reviews [18]—it becomes imperative to ensure that any comprehensive dataset includes instances of such LLM-generated misinformation. The limitations of previous medical misinformation datasets are summarized in Tab. 1.

To address these limitations, we have developed a comprehensive multi-modal dataset named Med-MMHL, specifically designed for the detection of human- and LLM-generated misinformation in the medical domain. Our contributions are outlined below:

We created Med-MMHL by crawling both the text and relevant images from news and tweets. The inclusion of multi-modalities (text and images) in Med-MMHL facilitates research on utilizing visual features for misinformation detection.

Med-MMHL comprises misinformation pertaining to 15 diseases, expanding beyond just COVID-19. This diverse misinformation across multiple diseases facilitates research about improving the generalization of medical misinformation detection solutions.

To the best of our knowledge, we are the first to incorporate LLM-generated misinformation in the medical domain. Med-MMHL includes LLM-generated fake news by ChatGPT. By incorporating both human- and LLM-generated misinformation sources, our dataset facilitates research in distinguishing misinformation across a broader range of scenarios.

Extensive baseline experiments and data analysis are conducted on Med-MMHL. Specifically, We build a misinformation detection benchmark on sentence, document, and multi-modal levels. Plus, we thoroughly analyzed the data characteristics at both the text level and semantic level.

2 DATA CRAWL

We collected news (including claims, summaries of news, and fact-check articles), tweets, and corresponding images from the medical domain. This specific time range was chosen to observe the trajectory of Covid-19 in relation to other significant diseases. We first

introduce the news source, followed by the news and tweet crawl processes.

Trusted real and fake news sources. To ensure the reliability of our trusted real news sources, we chose medical news articles that had been vetted by domain experts. Consequently, for the news source, our real news sources consist of news articles from authoritative medical authority websites; the fake news source comprises fake news articles archived by the fact-checking websites. For the claims respected to news, both the fake and real claims are extracted from the fact-check articles. Specifically, for authoritative medical authority websites such as "ClevelandClinic" [12], "NIH" [31], "WebMD" [48], "Mayo" [29], "Healthline" [25], and "ScienceDaily" [39], we utilized all of their news articles as real news. As for the fact-checking websites, which include "AFPFactCheck" [2], "CheckYourFact" [10], "FactCheck" [20], "HealthFeedback" [24], "LeadStories" [27], and "PolitiFact" [35], we extracted three main text components: a link to the archived fake news article being verified, a claim summarizing the fake news's opinion, and a claim concluding the evidence that elucidates the deficiencies in the quoted fake news article. Therefore, we gathered the fake news articles archived in the fact-checking websites as fake news. Besides, we collected the summaries of the evidence as real claims and the summaries of fake news opinions as fake claims. We collected both news articles and their applicable claims (short summaries) to account for the variation in text lengths, thereby enhancing the diversity of our dataset.



Figure 2: The data collection process. The numbers in the red circles indicate the three steps of data collection.

Step 1: Content extraction. In this step, we acquired all the articles from the aforementioned websites, encompassing the text contents, images, and links, spanning from Jan-01-2017 to May-01-2023. The real news articles can be obtained directly from the

news released by the authorities. To ensure the dataset’s scalability to disease classification tasks, we collected real news containing only one disease label out of a disease list. We specifically extracted disease labels that had more than fifty real news articles. As for claims, in fact-checking articles, fact-checkers typically provide a one-sentence summary of the fake news’ opinion, along with their own comments (usually presenting an opposing opinion of the fake news) as illustrated in Fig. 2. In this case, the summaries of news articles identified as "incorrect," "inaccurate," "misinformation," and similar terms by the fact-checkers were labeled as "fake claims," while the corresponding corrections provided by the fact-check articles are considered "real claims."

Step 2: Acquisition of human- and LLM-generated fake news.

To assess the effectiveness of fake news detection models against fake content generated by both humans and LLMs, we developed strategies to acquire these two types of fake news. The fake news articles extracted from Step 1 are human-generated and thus called "human-generated fake news." Additionally, we devised a strategy to simulate adversarial attacks using chatGPT3.5 [5] on real news articles. Each real news article had a 50% probability of being modified by chatGPT3.5. If chosen for modification, each sentence within the article had a random 10%–50% chance of being altered by providing the prompt "What is the opposite opinion of <the sentence>." These modified sentences were labeled as "fake sentences." Following sentence modification, the attacked article was further refined using the prompt "Refine the language of <the article>." The resulting generated articles were then cleaned up by removing redundant terms such as "the refined version is" or "refinement:" before being labeled as "LLM fake news."

Step 3: Real and fake tweet Crawl. We crawled tweets spanning from Jan-01-2022 to May-01-2023. This time range was chosen to comply with the size limitation specified in the Tweet Developer Agreement [3], as collecting tweets from the past six years would exceed the allowed size. Additionally, this range does not overlap with the time periods covered by the previous datasets in Tab. 1. Our method of tweet crawling is intrinsically tied to the corresponding news articles that we’ve crawled. Specifically, we employed the titles of these news articles as key phrases to retrieve related tweets. If the news title is for real news, we categorize the resulting tweets as real. Conversely, if the news title is for fake news, the collected tweets are classified as fake. Owing to the Twitter Developer Agreement [3], we might not manipulate tweets by LLMs and could only release the tweet IDs, along with a code that enables users to retrieve the full content of the tweets by these IDs.

3 BENCHMARK TASKS & STATISTICS

We propose and benchmark five different tasks that cover a range of challenges, each involving one or more of four types of inputs: long articles, claims(short articles, as described in Sec. 2), tweets, and multimodal data. The statistics for each task are summarized in Tab. 2 and each task is detailed below.

Fake news detection task specifically concentrates on text-only tasks, encompassing both articles and claims. Images are excluded from this task due to the lack of specific image associations with the text generated by the LLM. Notably, the real news articles used for generating LLM fake news are not included in this task.

Table 2: Statistics of benchmark tasks on Med-MMHL, where “fake news” is human-generated fake news, “sent” is an abbreviation of “sentence”. Since a text might have more than one image, the “#Image” can be larger than “w/image”.

Tasks	Data Type	Count	W/ Image	# Image
Fake news detection	Real news	3,455	/	/
	Fake news	469	/	/
	LLM fake news	2,095	/	/
	Real claim	2,283	/	/
	Fake claim	3,567	/	/
LLM-generated fake sent detection	Real sent	41,365	/	/
	LLM fake sent	17,608	/	/
Multimodal fake news detection	Real news	4,554	1,338	1,747
	Fake news	469	396	5,496
	Real claim	643	641	749
	Fake claim	1,135	1,102	1,102
Fake tweet detection	Real tweet	7,738	/	/
	Fake tweet	6,927	/	/
Multimodal tweet detection	Real tweet	7,738	1,334	1,385
	Fake tweet	6,927	639	763

LLM-generated fake sentence detection task is designed to evaluate the vulnerability to adversarial attacks introduced by LLM. It goals to assess a model’s ability to distinguish between real sentences and LLM-generated fake ones. Therefore, this task excludes human-generated fake sentences.

Multimodal fake news detection aims to investigate ways to enhance the detection of misinformation by leveraging multimodal resources. The specific approach employed for claim filtering is elaborated upon in Appendix A.3.

Fake tweet detection and **multimodal tweet detection** tasks are devised to address the distinctive writing style exhibited in tweets as compared to news articles. In order to fully leverage the available data, all collected tweets are included in both tasks, despite the relatively small number of tweets accompanied by images.

Other applicable tasks on Med-MMHL. Though we benchmark the above five tasks, Med-MMHL can also be applied to other tasks, given its data diversity. For example, a misinformation detection model can be trained using real news and human-generated fake news, and subsequently employed to identify LLM-generated fake news. Moreover, though our LLM-generated fake sentence detection task excludes the news context, Med-MMHL supports training a model for a more fine-grained misinformation detection at the sentence level.

4 MISINFORMATION DETECTION IN MEDICAL DOMAIN

To demonstrate the main utility of the proposed dataset and evaluate the existing fake news detection methods, we conduct comparative experiments on the misinformation detection task.

4.1 Baseline Methods

We consider seven text-only baseline models and two multimodal baseline models. Specifically, among the seven text-only models, four incorporate language transformer layers pretrained on long



Figure 3: Word Cloud of the date in Med-MMHL.

articles, two utilize language transformer layers pretrained on sentences, and one is trained using our own dataset. The multimodal models include state-of-the-art pretrained modules for both texts and images. The details of the baselines can be found in Appendix A.4.

4.2 Implementation Detail

The dataset is split into training, validation, and testing datasets with a ratio of 7 : 1 : 2. Each baseline model comprises a pre-trained module for learning hidden representations and a trainable module for fine-tuning the specific downstream task. During training, the parameters of the pre-trained models remain fixed, and they are utilized to extract hidden representations from the texts and images. A trainable two-layer feedforward neural network module maps the hidden representations to the downstream task. The optimizer is Adam, with a learning rate of $1e^{-5}$ for all the models. The maximum number of epochs is 100 with a 15-step patience. The dropout rate is 0.1. Due to the limitation of our computation resources, the batch size is 4. We adopt commonly used metrics in related areas: Accuracy, Precision, Recall, F1 and Macro F1.

4.3 Experimental Results

We conducted fake news detection and fake-news-related tweet detection experiments on the proposed Med-MMHL dataset. The experiment results are provided in Table 3 and Table 4. The metrics used for evaluation include accuracy (Acc), precision (Prc), recall (Rcl), f1-score (F1), and macro f1-score (F1-ma). We observe that **(i) Pretrained transformer-based methods perform better than simple methods**, as they are more powerful in capturing contextual information better. However, as the dataset is quite imbalanced, the models tend to generate many fake positive cases. Thus, the recall value is lower than the accuracy and precision value. **(ii) FN-BERT performs best** on document-level fake news/tweet detections among all baselines. This indicates the importance of related fake news classification knowledge. **(iii) Although baseline methods show strong performance in detecting fake news, the performance of the LLM sentence detection task is unsatisfactory.** It is easier to detect LLM-generated fake news than detect LLM-generated fake sentences, mostly because the generated fake news is entirely opposite in intention to real news, but the generated fake sentences are only partially opposite in intention to real news. Therefore, learning to detect LLM-generated fake sentence detection is an important area for further research.

Table 3: Baseline methods performance for fake news detection on Med-MMHL.

Model	Acc	Prc	Rcl	F1	F1-ma
Fake news detection (both human and LLM-generated fake news)					
dEFEND	89.174%	97.361%	81.240%	88.573%	89.144%
BERT	95.657%	97.702%	93.791%	95.707%	95.657%
BioBERT	94.941%	98.084%	91.993%	94.941%	94.941%
Funnel	94.604%	98.668%	90.768%	94.553%	94.603%
FN-BERT	95.784%	99.472%	92.320%	95.763%	95.784%
LLM-generated fake sentence detection					
dEFEND	92.183%	88.168%	85.264%	86.692%	90.579%
SentenceBERT	96.040%	96.583%	89.917%	93.131%	95.175%
DistilBERT	95.149%	95.050%	88.355%	91.581%	94.087%
Multimodal fake news detection (only human-generated fake news)					
CLIP	96.324%	86.921%	99.377%	92.732%	95.136%
VisualBERT	96.103%	89.881%	94.081%	91.933%	94.682%

Table 4: Baseline methods performance for fake-news-related tweet detection on Med-MMHL.

Model	Acc	Prc	Rcl	F1	F1-ma
Fake tweets detection (only human-generated fake news)					
dEFEND	96.897%	98.868%	94.517%	96.643%	96.880%
BERT	98.056%	99.552%	96.318%	97.908%	98.046%
BioBERT	97.988%	99.775%	95.957%	97.828%	97.977%
Funnel	98.158%	99.701%	96.390%	98.018%	98.149%
FN-BERT	98.602%	99.339%	97.690%	98.507%	98.596%
Multimodal fake tweets detection (only human-generated fake news)					
CLIP	97.954%	99.256%	96.387%	97.801%	97.944%
VisualBERT	96.404%	99.403%	92.620%	95.985%	96.364%

5 CONCLUSION

Medical misinformation significantly affects individuals and societies, necessitating effective detection methods. However, existing datasets have limitations: overlooking visual information, focusing solely on COVID-19, or ignoring LLM-generated misinformation. To address these limitations, we introduce Med-MMHL, a multi-modal dataset for detecting misinformation in the broader medical field, incorporating both human and LLM-generated fake data across multiple diseases. Additionally, Med-MMHL extends its diversity by incorporating data from news and tweets. We also comprehensively analyze the dataset's characteristics at text and sentence levels. Finally, we establish a benchmark for misinformation detection at sentence, document, and multi-modal levels, laying the groundwork for future research in this critical domain.

REFERENCES

- [1] Jennifer Abbasi. 2022. Widespread Misinformation About Infertility Continues to Create COVID-19 Vaccine Hesitancy. <https://jamanetwork.com/journals/jama/fullarticle/2789477>.
- [2] AFPFactCheck. 2023. AFPFactCheck. <https://factcheck.afp.com/>.
- [3] Twitter Developer Agreement. 2023. Developer Agreement and Policy. <https://developer.twitter.com/en/developer-terms/agreement-and-policy>.
- [4] Open AI. 2021. CLIP. <https://huggingface.co/openai/clip-vit-base-patch32>.
- [5] Open AI. 2023. ChatGPT 3.5. <https://chat.openai.com/chat>.
- [6] Madeline Barron. 2022. How to Spot and Combat Health Misinformation. <https://shorturl.at/pMPW1>.
- [7] Som S Biswas. 2023. Role of chat gpt in public health. *Annals of Biomedical Engineering* (2023), 1–2.
- [8] Alessandro Bondielli and Francesco Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information Sciences* 497 (2019), 38–55.
- [9] Nicola Capuano, Giuseppe Fenza, Vincenzo Loia, and Francesco David Nota. 2023. Content Based Fake News Detection with machine and deep learning: a systematic review. *Neurocomputing* (2023).
- [10] CheckYourFact. 2023. CheckYourFact. <https://checkyourfact.com/>.
- [11] Mingxi Cheng, Songli Wang, Xiaofeng Yan, Tianqi Yang, Wenshuo Wang, Zehao Huang, Xiongye Xiao, Shahin Nazarian, and Paul Bogdan. 2021. A COVID-19 rumor dataset. *Frontiers in Psychology* 12 (2021), 644801.
- [12] ClevelandClinic. 2023. ClevelandClinic. <https://newsroom.clevelandclinic.org/>.
- [13] Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885* (2020).
- [14] Sajjad Dadkhah, Xichen Zhang, Alexander Gerald Weismann, Amir Firouzi, and Ali A Ghorbani. 2023. TruthSeeker: The Largest Social Media Ground-Truth Dataset for Real/Fake Content. (2023).
- [15] Zihang Dai, Guokun Lai, Yiming Yang, and Quoc V. Le. 2020. Funnel-Transformer: Filtering out Sequential Redundancy for Efficient Language Processing. <https://huggingface.co/funnel-transformer/medium-base>. *arXiv:2006.03236* [cs.LG].
- [16] Pritam Deka, Anna Jurek-Loughrey, et al. 2022. Evidence Extraction to Validate Medical Claims in Fake News Detection. <https://huggingface.co/pritamdeka/BioBERT-mnli-scnli-scnli-scitail-mednli-stsb>. In *International Conference on Health Information Science*. Springer, 3–15.
- [17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. <https://huggingface.co/bert-base-cased>. *CoRR* abs/1810.04805 (2018). *arXiv:1810.04805* <http://arxiv.org/abs/1810.04805>
- [18] Tjibbe Donker. 2023. The dangers of using large language models for peer review. *The Lancet Infectious Diseases* (2023).
- [19] Alex Escolà-Gascón, Neil Dagnall, and Josep Gallifa. 2021. Critical thinking predicts reductions in Spanish physicians' stress levels and promotes fake news detection. *Thinking Skills and Creativity* 42 (2021), 100934.
- [20] FactCheck. 2023. FactCheck. <https://www.factcheck.org/>.
- [21] Mehmet Firat. 2023. How chat GPT can transform autodidactic experiences and open education. *Department of Distance Education, Open Education Faculty, Anadolu Unive* (2023).
- [22] Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. 2020. The future of false information detection on social media: New perspectives and trends. *ACM Computing Surveys (CSUR)* 53, 4 (2020), 1–36.
- [23] Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public health* 203 (2022), 23–30.
- [24] HealthFeedback. 2023. HealthFeedback. <https://healthfeedback.org/>.
- [25] Healthline. 2023. Healthline. <https://www.healthline.com/>.
- [26] Alexander Kinsora, Kate Barron, Qiaozhu Mei, and VG Vinod Vydiswaran. 2017. Creating a labeled dataset for medical misinformation in health forums. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 456–461.
- [27] LeadStories. 2023. LeadStories. <https://leadstories.com/>.
- [28] Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. MM-COVID: A multilingual and multimodal data repository for combating COVID-19 disinformation. *arXiv preprint arXiv:2011.04088* (2020).
- [29] Mayo. 2023. Mayo. <https://newsnetwork.mayoclinic.org/>.
- [30] Robert W McGee. 2023. Is chat gpt biased against conservatives? an empirical study. *An Empirical Study (February 15, 2023)* (2023).
- [31] NIH. 2023. NIH. <https://www.nih.gov/>.
- [32] UCLA NLP. 2023. VisualBERT. <https://huggingface.co/uclanlp/visualbert-vqa-coco-pre>.
- [33] World Health Organization. 2022. Coronavirus disease (COVID-19) advice for the public: Myths busters. <https://shorturl.at/oCKP6>.
- [34] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
- [35] PolitiFact. 2023. PolitiFact. <https://www.politifact.com/>.
- [36] Protik Bose Pranto, Syed Zami-Ul-Haque Navid, Protik Dey, Gias Uddin, and Anindya Iqbal. 2022. Are You Misinformed? A Study of Covid-Related Fake News in Bengali on Facebook. *arXiv preprint arXiv:2203.11669* (2022).
- [37] Mohiuddin Md Abdul Qudar and Vijay Mago. 2020. Tweetbert: a pretrained language representation model for twitter text analysis. *arXiv preprint arXiv:2010.11091* (2020).
- [38] Chaoyong Ragkhitwetsagul, Jens Krinke, and Bruno Marnette. 2018. A picture is worth a thousand words: Code clone detection based on image similarity. In *2018 IEEE 12th International workshop on software clones (IWSC)*. IEEE, 44–50.
- [39] ScienceDaily. 2023. ScienceDaily. <https://www.sciencedaily.com/>.
- [40] sentence transformer. 2019. distilBERT. <https://huggingface.co/sentence-transformers/msmarco-distilbert-base-tas-b>.
- [41] sentence transformer. 2019. sentenceBERT. <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>.
- [42] Kai Shu, Limeng Cui, Suhan Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 395–405.
- [43] Ivan Srba, Branislav Pecher, Matus Tomlein, Robert Moro, Elena Stefancova, Jakub Simko, and Maria Bielikova. 2022. Monant Medical Misinformation Dataset: Mapping Articles to Fact-Checked Claims. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2949–2959.
- [44] Nigar M Shafiq Surameery and Mohammed Y Shakor. 2023. Use chat gpt to solve programming bugs. *International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290* 3, 01 (2023), 17–22.
- [45] The New York Times. 2023. The ChatGPT Lawyer Explains Himself. <https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html>.
- [46] unjus. 2023. FN-BERT. https://huggingface.co/unjus/Fake_News_BERT_Classifier.
- [47] Apurva Wani, Isha Joshi, Snehal Khandve, Vedangi Wagh, and Raviraj Joshi. 2021. Evaluating deep learning approaches for covid19 fake news detection. 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*. Springer, 153–163.
- [48] WebMD. 2023. WebMD. <https://www.webmd.com/>.
- [49] Chen Yang, Xinyi Zhou, and Reza Zafarani. 2021. CHECKED: Chinese COVID-19 fake news dataset. *Social Network Analysis and Mining* 11, 1 (2021), 58.
- [50] Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 3205–3212.
- [51] Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–40.
- [52] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)* 51, 2 (2018), 1–36.

A APPENDIX

A.1 Data Overview

A.1.1 Relations to Multiple Disease. Although we did not provide specific disease labels for the articles/tweets, we conducted a statistical analysis based on diseases. As indicated in Table 5, we examined fifteen disease categories that contained more than fifty real news articles. The statistical findings reveal that the number of real news articles is relatively evenly distributed across various types of diseases. In contrast, fake news articles and tweets tend to concentrate on "hotspot" topics such as Covid-19 and Monkeypox.

A.2 Data Analysis

We analyze our dataset in two-fold: text-level and embedding-level. We detail these two folds below.

text-level. To understand the topic difference between the tweets of fake and real news, we analyze the top 30 frequent hashtags in tweets related to fake and true news articles, respectively. The frequency of hashtags in tweets related to fake and real news articles is shown in Figure 6a and Figure 6b, respectively. We find that the hashtag distributions of tweets about fake and real news articles are quite different. While the hashtags in tweets about true news articles are mainly related to healthcare, those in tweets about fake news cover more diverse topics, including social media (#facebook, #foxnews) and natural disasters (#hurricane, #earthquake).

embedding-level. In terms of news, we categorized our crawled content into three distinct sources: real, human-generated fake, and Language Learning Model (LLM)-generated fake news. As depicted in Figure 4, we randomly selected 300 news articles from each of these categories and analyzed them using BERT embeddings. However, our analysis reveals that the BERT embeddings struggle to distinguish between real, human-fake, and LLM-fake news due to significant overlaps in these categories.

This observation highlights the significance of researching methodologies to accurately discern these three distinct sources of news. Moreover, our analysis shows minimal overlap between LLM-fake news and human-fake news, suggesting that a model adept at identifying human-fake news might not necessarily be effective at detecting LLM-generated misinformation, and vice versa. This calls for an approach that can adapt to these distinct categories effectively.

Correspondingly, we categorized the crawled tweets into two primary sources: real tweets and human-fake tweets. Due to the constraints imposed by Twitter's Developer Policy [3], the generation of LLM-fake tweets is not permissible. As a result, we randomly sampled 300 tweets from both sources for our analysis, as illustrated in Figure 5. For analysis, we utilized TweetBERT embeddings [37]. However, the figure shows that TweetBERT embeddings struggle to clearly demarcate between real and human-fake tweets, demonstrating significant overlap. This underlines the importance of exploring further research methodologies to distinguish these two categories accurately.

A.3 Multimodal Claim Filtering

By looking at the images of the real news, fact-check, and fake news articles, we notice a pattern where real news articles often incorporate decorative images sourced from the internet, while

News Distribution in BERT Embedding

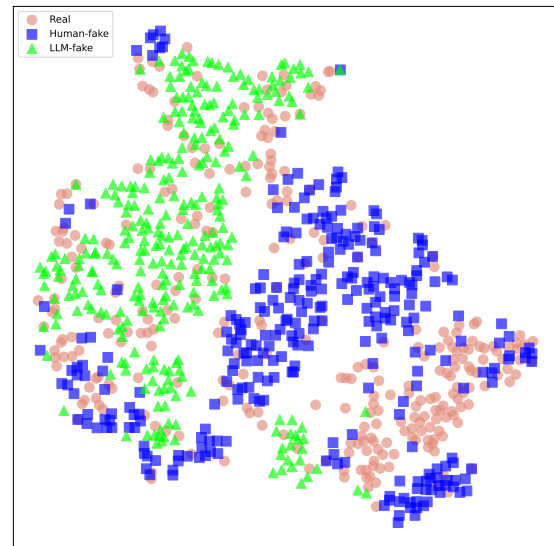


Figure 4: A t-SNE figure of randomly sampled 300 real news, 300 human-fake news, and 300 LLM-fake news.

Tweet Distribution in Embedding-Level

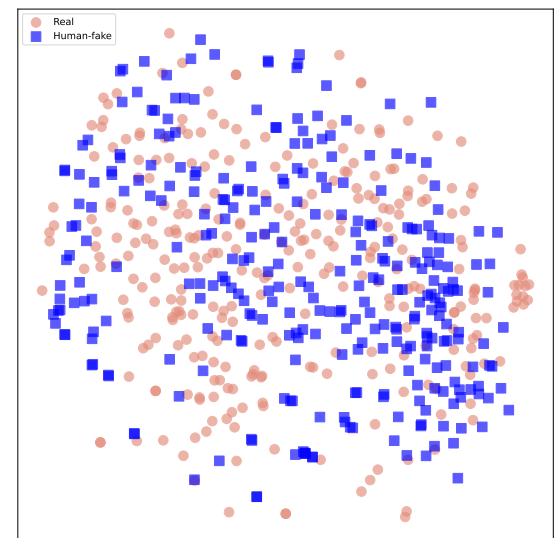


Figure 5: A t-SNE figure of randomly sampled 300 real tweets, 300 human-fake tweets. Due to the Tweet Develop Policy [3], we cannot use ChatGPT to generate LLM-fake tweets.

fake news articles frequently utilize screenshots of social media or videos. This stark contrast makes it relatively straightforward to distinguish between true and fake news. However, in the case of fact-check articles, we observe that "AFPFactCheck" tends to use screenshots, while "CheckYourFact" and "PolitiFact" lean towards using decorative images. Consequently, we included the true claims from "AFPFactCheck" and the false claims from "CheckYourFact" and "PolitiFact" as part of the multimodal fake news detection task.

Table 5: Statistics between diseases and news/tweets.

Information Type	anemia	arthritis	asthma	cancer	covid	diabetes	epilepsy	flu	headache	hypertension	inflammation	monkeypox	parkinson	pneumonia	stroke	Total
Real news	64	85	148	1,410	859	332	48	740	70	55	282	44	81	50	286	4,554
Fake news	0	1	0	27	304	1	2	114	1	0	4	3	0	2	10	469
LLM fake news	18	35	62	615	462	161	23	339	30	25	135	19	39	11	121	2,095
True claims	3	4	7	190	1,619	31	2	362	11	1	12	7	4	9	21	2,283
False claims	5	6	10	269	2,557	38	3	575	14	2	14	19	6	15	34	3,567
Total news	91	133	227	2,560	5,836	569	83	2,152	127	84	452	94	132	88	481	12,968
Real tweets	53	15	28	540	1,161	152	29	2,095	36	21	174	8	35	17	106	7,738
Fake tweets	0	0	0	120	2,547	0	1	2,436	0	0	2	1,799	0	0	21	6,927
Total tweets	53	15	28	660	3,708	152	30	4,531	36	21	176	1,807	35	17	127	27,633

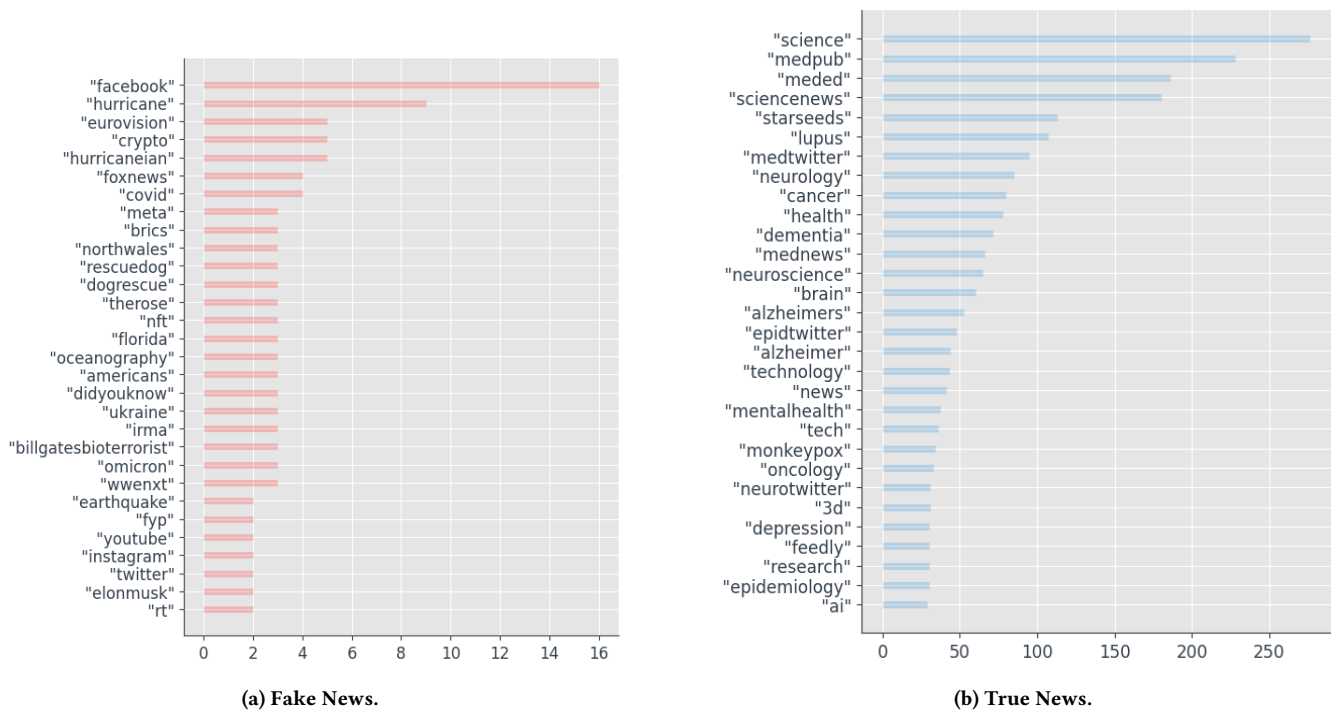


Figure 6: Frequency of hashtags in tweets about fake and true news articles.

This ensures that the models trained on our dataset do not get misled by features that are irrelevant to the content of the articles.

A.4 Baseline Models

The following baseline fake news detection methods are considered for medical misinformation detection:

- BERT [17]: A bi-directional transformer model pretrained on a large corpus of English data in a self-supervised fashion.
- BioBERT [16]: A sentence-transformers model built with medical dataset for fact-checking of online health information.
- Funnel Transformer [15]: An efficient bidirectional transformer model by applying a pooling operation after each layer, akin to convolutional neural networks, to reduce the length of the input.
- FN-BERT [46]: A BERT-based model recently finetuned on a Fake news classification dataset in 2023.

- sentenceBERT [41]: A sentence representation learning model pretrained using Siamese and triplet network structures.
- distilBERT [40]: A dual-encoder then dot-product scoring architecture BERT model. The version employed in this paper is pre-trained with the TAS-Balanced method on the MSMARCO standard.
- dEFEND [42] utilizes the hierarchical attention network to model article content for misinformation detection.
- CLIP [4]: A multi-modal vision and language model pretrained on 400 million image-text pairs.
- VisualBERT [32]: A multi-modal vision and language model. It uses a BERT-like transformer to prepare embeddings for image-text pairs.