A APPENDIX

A.1 Data Overview

A.1.1 Sample Data. We show the sample news with LLM-generated fake sentences in Figure 5, sample claims with their image in Figure 6, and sample tweets with their images in Figure 4.

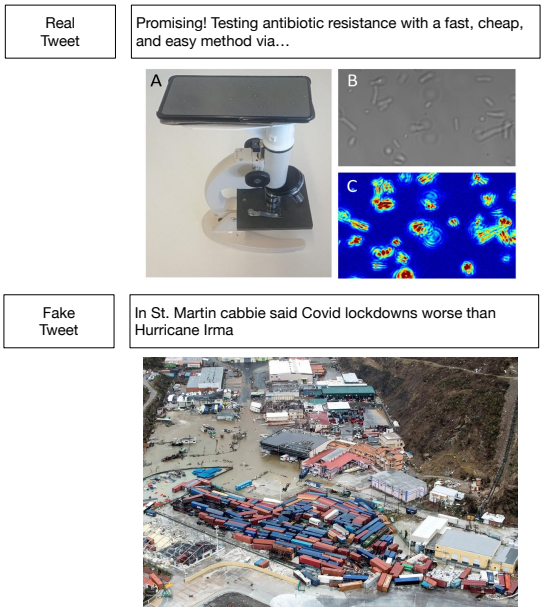


Figure 4: A diagram shows examples of a real tweet and a fake tweet as well as their images.

A.1.2 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 8a and Figure 8b, 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 7a, 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.

The observation in Figure 7a 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 7b. For analysis, we utilized TweetBERT embeddings [36]. 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 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. 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 [16]: A bi-directional transformer model pretrained on a large corpus of English data in a self-supervised fashion.
- BioBERT [15]: A sentence-transformers model built with medical dataset for fact-checking of online health information.
- Funnel Transformer [14]: 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 [45]: A BERT-based model recently finetuned on a Fake news classification dataset in 2023.
- sentenceBERT [40]: A sentence representation learning model pretrained using Siamese and triplet network structures.
- distilBERT [39]: A dual-encoder then dot-product scoring architecture BERT model. The version employed in this paper is

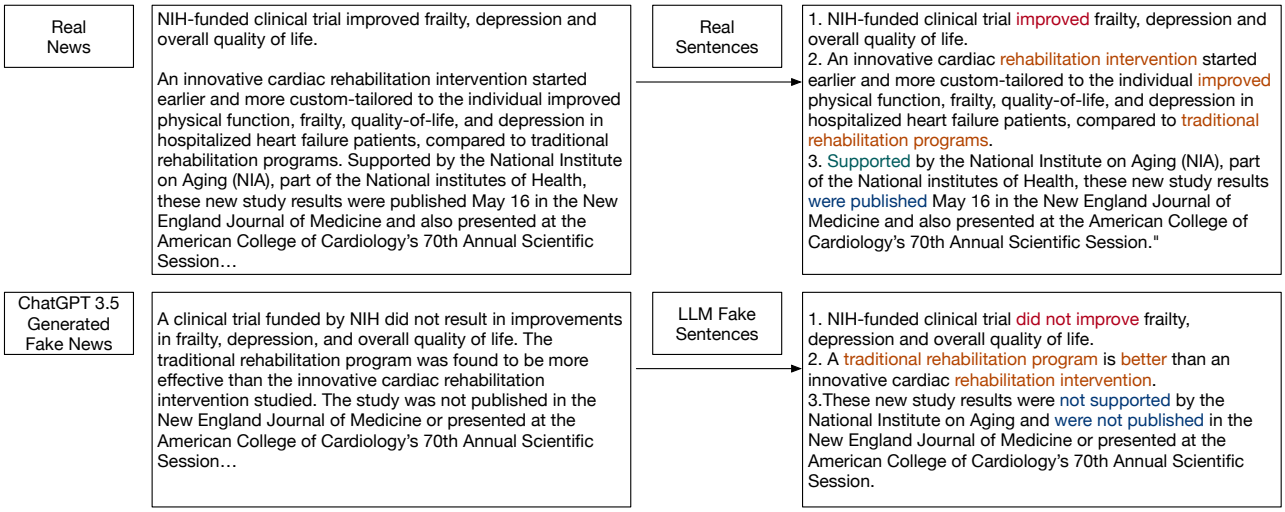


Figure 5: On the diagram, the left side represents real news and LLM-generated fake news. Its right side displays a collection of sampled sentences from the left side. Our dataset also collects human-generated news, which is not shown in the diagram. Plus, the news images are also not shown in this example.

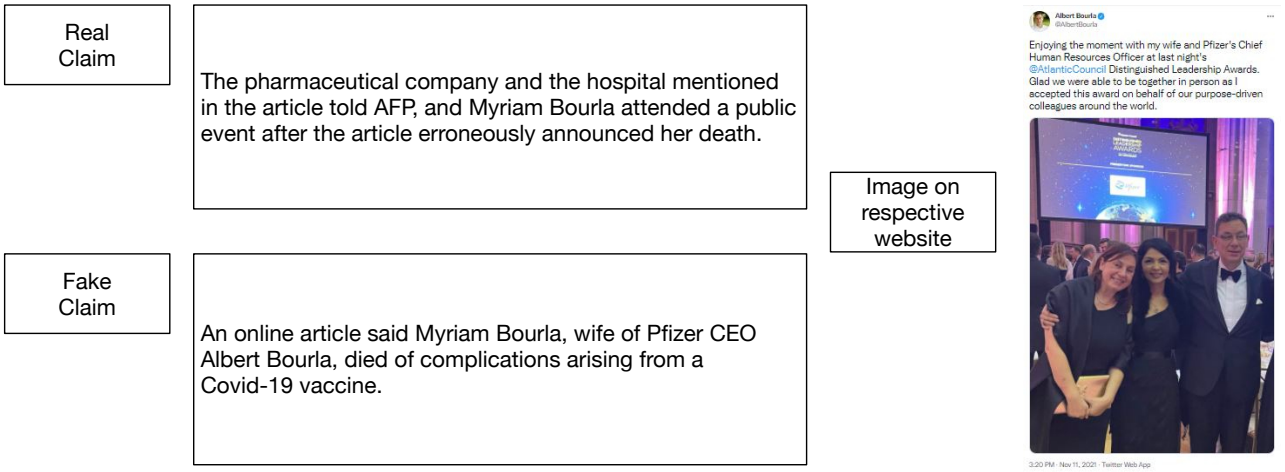


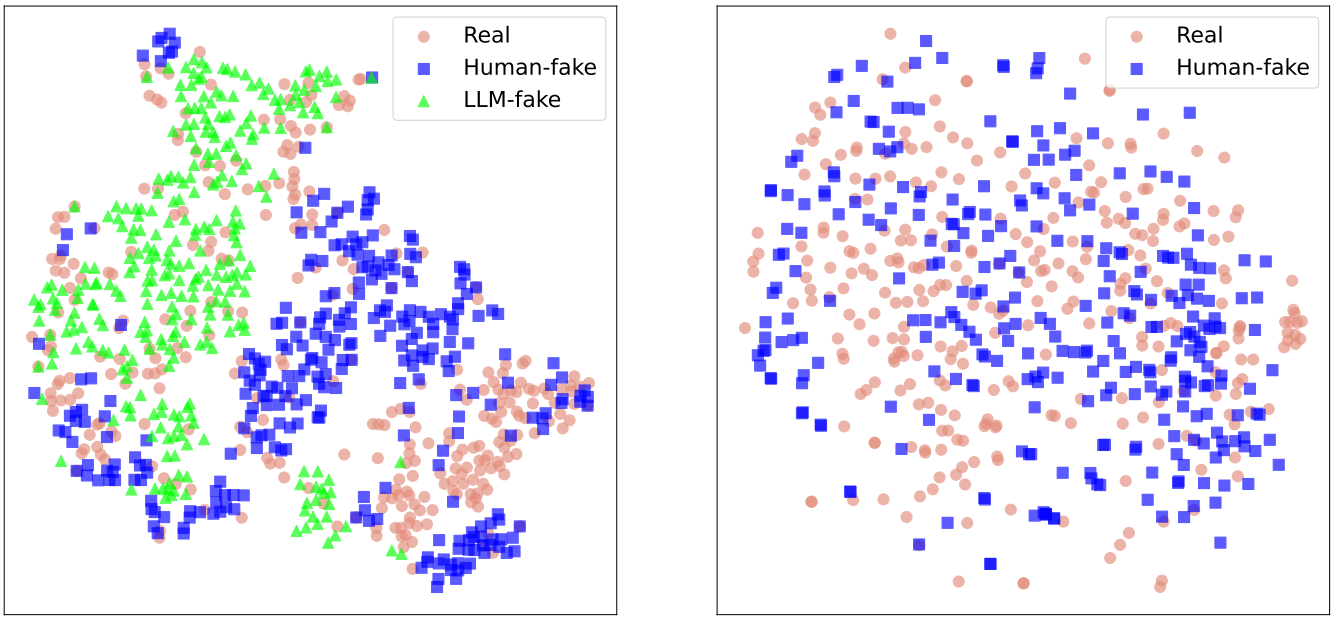
Figure 6: A diagram shows a pair of a real claim and a fake claim from a website. The right image is also from the website.

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	62	84	144	1,305	793	320	42	724	69	49	276	44	72	49	254	4,287
Fake news	0	1	0	27	304	1	2	114	1	0	4	3	0	2	10	469
LLM fake news	18	35	60	568	428	158	20	330	30	24	132	19	35	10	104	1,97
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	88	130	227	2,359	5,701	548	69	2,105	125	76	438	92	127	85	423	12,593
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

pre-trained with the TAS-Balanced method on the MSMARCO standard.

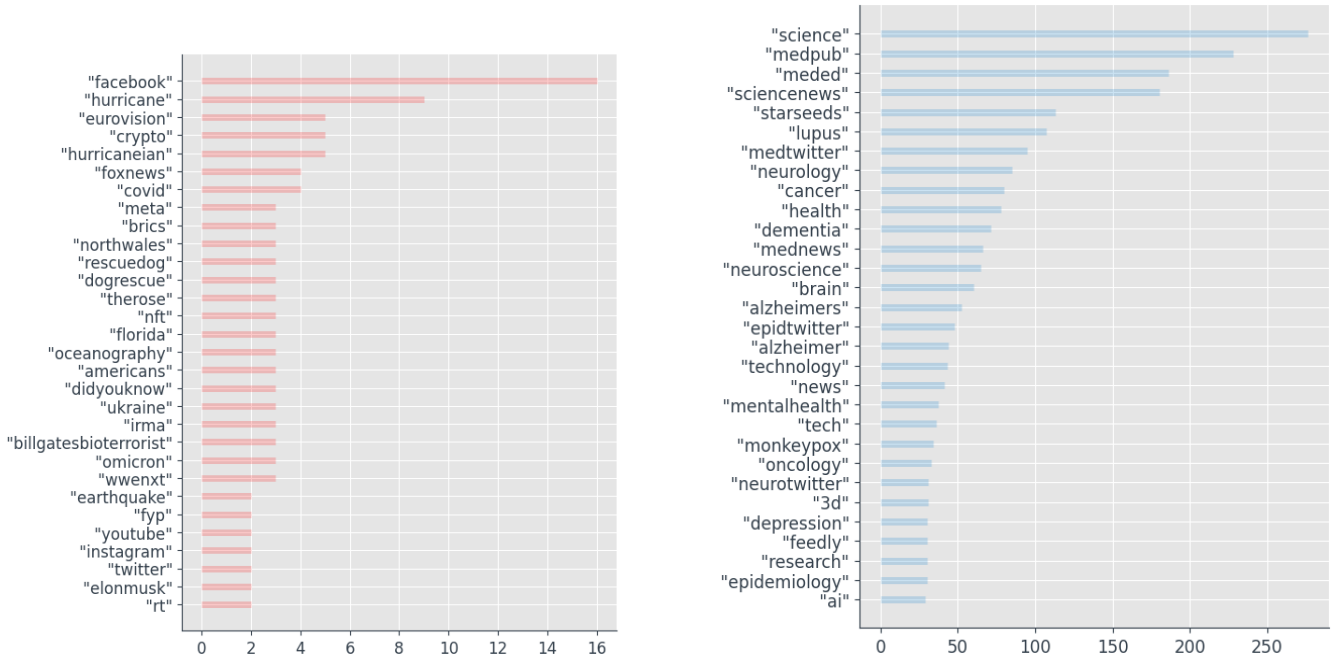
- dEFEND [41] utilizes the hierarchical attention network to model article content for misinformation detection.



(a) News embedding level distribution.

(b) Tweet embedding level distribution.

Figure 7: (a). A t-SNE figure of randomly sampled real news, human-fake news, and LLM-fake news. Each category consists of 300 samples, and BERT embeddings are utilized. (b). A t-SNE figure of randomly sampled 300 real tweets, and 300 human-fake tweets, where TweetBERT [36] embeddings are used. Due to the Tweet Develop Policy [3], we cannot use ChatGPT to generate LLM-fake tweets.



(a) Fake News.

(b) True News.

Figure 8: Frequency of hashtags in tweets about fake and true news articles.

- CLIP [4]: A multi-modal vision and language model pretrained on 400 million image-text pairs.
- VisualBERT [31]: A multi-modal vision and language model. It uses a BERT-like transformer to prepare embeddings for image-text pairs.