CHAT, IS THIS REAL? MM-DOUBLE CONFIRM! : A MULTIFOLD NETWORK FOR SINGAPORE-CONTEXT MISINFORMATION DETECTION

Felicia Tan Ee Shan[1], Goh Rou Hui Ashley[2], Adriel Kuek[3]
[1]Raffles Girls' School, 2 Braddell Rise, Singapore 318871
[2]River Valley High School, 6 Boon Lay Ave, Singapore 649961
[3]DSO National Laboratories, 12 Science Park Drive, Singapore 118225

BACKGROUND AND PURPOSE OF RESEARCH AREA

Introduction

Every day, a staggering amount of information floods the digital realm, with 500 million tweets, 294 billion emails, 4 million gigabytes of Facebook data, and 65 billion WhatsApp messages generated globally[1]. This vast volume of data in the digital age has facilitated the rapid spread of misinformation, holding the potential to significantly impact a large audience in a remarkably short time.

Amid this deluge of data, the escalating prevalence of misinformation stands as a pressing concern, having resulted in severe detrimental impacts on daily life, particularly in the challenging task of discerning trustworthy information on the internet. For instance, according to Islam et al. (2020), the dissemination of misinformation related to COVID-19 has been extensive, fostering widespread scepticism towards medical treatments and, in some cases, contributing to the public's hesitation towards vaccination.

The model framework that we have designed addresses this issue by facilitating an efficient and automated approach to fact-checking for images and textual content on social media, while promoting a deeper user understanding of the facts surrounding a given topic. In doing so, our project aligns with the broader objective of fostering an online environment space where individuals can confidently make informed decisions based on reliable and verified information, building up digital social literacy and resilience. Our model contributes to the well-being of the local population and can be adapted and applied to similar challenges globally.

[1] https://theconversation.com/the-worlds-data-explained-how-much-were-producing-and-where-its-all-stored-159964

The contributions of this paper are 4-folds:

(1) A large original multimodal dataset for Singapore-context misinformation content

(2) An automated framework for evidence retrieval and claim verification through external knowledge

(3) A custom satire text detector to evaluate exaggerated content based on language semantics

(4) Ensemble of multi-stage models, together with an out-of-context detection pretrained model to leverage on existing state-of-the-art Large Language Models (LLMs) for zero-shot misinformation prediction and explanation generation.

1. MULTIMODAL MISINFORMATION

In the context of automated fact-checking, the term "multimodal" refers to cases where "the claim and/or evidence are expressed through different or multiple modalities" (Hameleers et al., 2020; Alam et al., 2022; Biamby et al., 2022).

For the scope of our project, we focused on examples of Singapore-context misinformation containing text and image modalities. We did not consider audio and video modalities for our project scope.

Claim	The carpark in TTSH is converted to a hospital ward. The situation doesn't look good

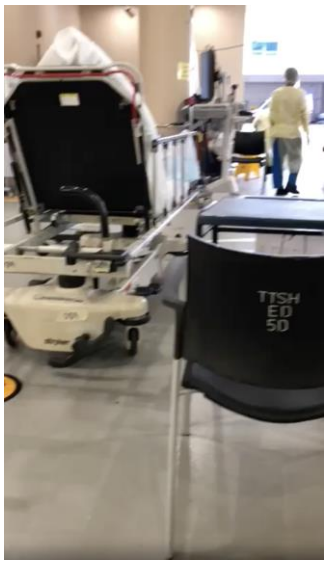
Image	

Table 1. An example of Singapore-context misinformation

2. MANUAL DETECTION OF MULTIMODAL MISINFORMATION

Human fact-checkers may be influenced by their own biases, such as confirmation bias, anchoring bias, or availability heuristic[2] (Park et. al, 2021). These biases can lead to a selective evaluation of evidence, known as "cherry-picking", undermining the accuracy of the fact-checking process.

Moreover, the overwhelming influx of information in the digital age can create an information overload[3] (Horrigan, 2016), making it challenging for users to thoroughly and efficiently verify every claim. This information overload, coupled with the time-consuming nature of manual fact-checking, often deters them to verify the information they come across on social media (Bergan et. al, 2022).[4]

[2] Park, S., Park, J. Y., Kang, J. H., & Cha, M. (2021). The presence of unexpected biases in online fact-checking. The Harvard Kennedy School Misinformation Review.

[3] Horrigan, J. B. (2016). Information overload | Pew Research Center. Pew Research Center: Internet, Science & Tech. https://www.pewresearch.org/internet/2016/12/07/information-overload/

[4] Ma, S., Bergan, D. E., Ahn, S., Carnahan, D., Gimby, N., McGraw, J., & Virtue, I. (2022). Fact-checking as a deterrent? A conceptual replication of the influence of fact-checking on the sharing of misinformation by political elites. Human Communication Research, 49(3), 321–338. https://doi.org/10.1093/hcr/hqac031

Furthermore, contemporary misinformation is increasingly challenging to detect due to its multimodal nature. For instance, earlier studies indicate that visual content demonstrates a "photo truthiness" phenomenon (Newman and Zhang, 2020), influencing readers to perceive a claim as more truthful when presented alongside visuals compared to a similar claim conveyed in text alone. Moreover, the increasingly common use of out-of-context images, altered or employed misleadingly, contributes to the complexity of misinformation. In the absence of original context or specialised image analysis tools, these images can be highly persuasive (Morstatter et. al, 2019).[5]

Misconstruing satire as factual information further introduces complexity, as the humorous, ironic, or exaggerated elements in satirical content make detection challenging, given the varied interpretations among individuals.

The difficulty of the task of human fact-checking is evident in a recent online survey by Ipsos – while four in five Singaporeans express confidence in identifying fake news, approximately 90% of participants incorrectly identified at least one out of five false headlines as genuine.[6]

This highlights the need for advanced detection methods capable of addressing the diverse challenges posed by misinformation.

3. AUTOMATED DETECTION OF MULTIMODAL MISINFORMATION

There has been a growing interest in employing artificial intelligence (AI) to detect and combat misinformation. A wide variety of approaches have been adopted, ranging from utilising BERT-based models to decision tree classifiers.

[5] Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media: definition, manipulation, and detection. ACM SIGKDD explorations newsletter, 21(2), 80-90

[6] Huiwen, N. (2018, September 27). 4 in 5 Singaporeans confident in spotting fake news but 90 per cent wrong when put to the test: Survey. The Straits Times. https://www.straitstimes.com/singapore/4-in-5-singaporeans-confident-in-spotting-fake-news-but-90-per-cent-wrong-when-put-to-the

Literature Review

1. Text-only Models

Conventional text-based detection models, such as GEAR (Zhou et. al, 2019) are challenged by the multimodal nature of the types of misinformation in modern society as they fail to capture the complex interplay of information across multiple modalities.

2. Multimodal Models

MOCHEG (Yao et. al, 2022) is a model addressing evidence retrieval, claim verification, and explanation generation. It employs Sentence-BERT (SBERT) for evidence retrieval, utilising cosine similarity to rank the top-1000 candidate evidence sentences and BERT-based re-ranking for refinement. Claim verification utilises pre-trained Contrastive Language-Image Pretraining (CLIP) (Radford et. al, 2021) to encode text and image evidence, employs stance detection through cross-attention and fusion operations, and predicts truthfulness by optimising concatenated representations with a cross-entropy objective. Lastly, explanation generation is achieved by concatenating the input claim, predicted truthfulness label, and text evidence, and optimising the BART-based generation model through reinforcement learning with a truthfulness reward from a pre-trained BERT-based classification model. However, performance is poor, with low precision scores of 4.71% for image retrieval and 14.92% for text retrieval, and a claim verification F-score of 44.06%.

The Truthformer model (Chaitanya et. al, 2022) only addresses claim verification and integrates text and image data. It employs separate embedding blocks fine-tuned on FACTIFY[7](Mishra et al., 2021) data and fuses representations using Conv1D layers (FusConv1D) or self-attention (FusAttn). Textual embeddings use mBERT, and image embeddings use Vision Transformer (ViT). The model is trained simultaneously, optimising all blocks and enhancing performance through pseudo-labelling. Performance is strong, with a final F1-score of 76.819% but is untrained on Singapore-context misinformation.

[7] Mishra, S., S, S., Bhaskar, A., & Ahuja, C. (2021, November 18). *FACTIFY: A Multi-Modal Fact Verification Dataset*. ResearchGate. https://www.researchgate.net/publication/356342935_FACTIFY_A_Multi-Modal_Fact_Verification_Dataset

Existing automated fact-checking models, while valuable, exhibit notable limitations in their design:

1) No existing model that addresses Singapore-context misinformation
2) Scarcity of end-to-end fact-checking models
3) The prevalent binary classification system, offering only "True" or "False" labels, poses a challenge for users seeking a nuanced understanding of the veracity of claims, such as discerning whether the content is satire.
4) Limited knowledge and thus poor performance for novel topics due to being trained on a fixed dataset with limited topics
5) Impractical in real world setting as evidence retrieval is not automated

Aims and Objectives

Considering the limitations of current automated fact-checking models, our project sets out to develop an *end-to-end* deep learning model that (1) detects *Singapore-context* multimodal misinformation (2) detect satire and out-of-context image usage and provide explanations so as to promote user understanding (3) automates evidence retrieval and retrieves evidence updated in real time so as to be practical and functional in a real-world context.

Key criteria include accuracy of the evidence retrieval module, claim verification module, the satire detector, the out-of-context image usage detector.

A constraint is the lack of a Singapore-context image-text misinformation dataset.

METHODS

Pipeline

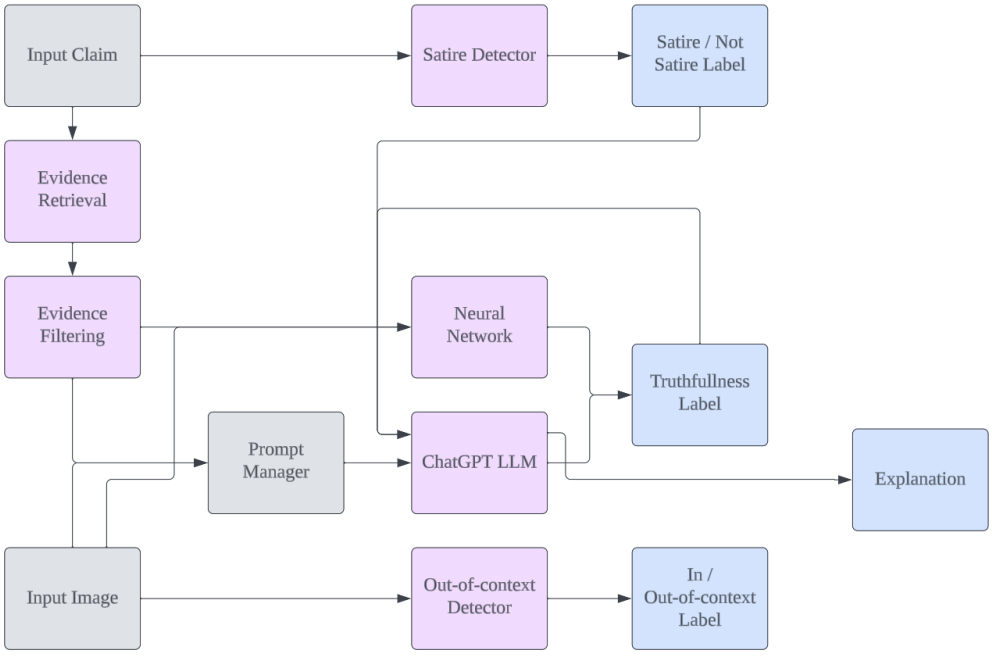


Fig.1. A diagram of our pipeline

Mass Dataset Collection

1. Singapore-context multimodal misinformation dataset

Data was collected from a variety of sources to ensure diversity and thus mirror the complexity and diversity of misinformation found in the real-world context. A mass original dataset for Singapore-context misinformation was created by obtaining examples from fact-checking websites Black Dot Research [8]and Factually[9]. Additionally, we manually collected and labelled examples that were circulated on social media platforms Facebook and Whatsapp. “True” examples were also compiled by taking the headlines and photos of articles by credible news

[8] https://blackdotresearch.sg/
[9] https://www.gov.sg/factually

outlets like Straits Times[10] and Channel News Asia[11]. A total of 295 sets of claims and their corresponding image. The dataset set was split with a 80/20 ratio of train data and test data. We spent a total of 100 hours manually searching, cleaning and annotating the dataset. We had to search for Singapore context specifically through the many articles found online. The textual evidence was manually done up by searching for news articles relating to the claims and retrieving the most relevant sentences for each of 3 articles. Images were downloaded and renamed while file paths were manually added. Labels were also added manually. After data collection, we had to clean the dataset to double check and ensure that there was no missing data or irrelevant and unnecessary data such as those not in Singapore context.

Claim	Wolbachia mosquitoes to be released at five more sites in Singapore to combat dengue
Image	

Table 2. A "True" example

2. Satire dataset

We collected 184 examples of Singapore-context satirical claims and 104 examples of Singapore-context non-satirical fake claims.

[10] https://www.straitstimes.com/
[11] https://www.channelnewsasia.com/

To circumvent the lack of Singapore-context satirical claims available online, we leveraged ChatGPT to automate the process of generating Singapore-context satirical claims.

We included a reference example and prompted it to generate satirical claims that were specifically in Singapore's context. Refer to Table 3 for the reference example provided and examples of ChatGPT-generated Singapore-context satirical claims.

Singapore-context satirical claims	
Reference Example	HDBs set to receive 'genius windows' that can predict rain before meteorologists even get the memo.
ChatGPT-generated	Government launches 'Makan Masterclass' to train tourists in the ancient art of hawker mind-reading. Changi Airport introduces 'self-flying suitcases' that navigate travelers to their destinations like magical luggage fairies.

Table 3. Examples of reference and ChatGPT-generated claims

The dataset was further supplemented with 203 examples of satirical news articles and 283 examples of non-satirical news articles from the Fake News vs Satire: A Dataset and Analysis [12](Golbeck et al., 2018) dataset. Our final dataset contained a 1:1 ratio of satirical claims and non-satirical fake claims with 387 examples for each class.

[12] Golbeck, J., Mauriello, M. L., Auxier, B. E., Bhanushali, K. H., Bonk, C., Bouzaghrane, M. A., Buntain, C., Chanduka, R., Cheakalos, P., Everett, J. B., Falak, W., Gieringer, C., Graney, J., Hoffman, K. M., Huth, L., Ma, Z., Jha, M., Khan, M. A., Kori, V., . . . Visnansky, G. (2018, May 15). Fake News vs Satire. https://doi.org/10.1145/3201064.3201100

1. Automated Evidence Retrieval

The H2-keywordextractor transformer model was used for keyword extraction from the user-input claim. The extracted keywords were further refined through a cleaning process. Firstly, Named Entity Recognition (NER) was applied to extract names, locations and time phrases from the keywords. These entities were deemed as important and preserved. Subsequently, common stop words such as 'the,' 'and,' and 'is' were excluded from the keywords to focus on the more meaningful and distinctive terms. This process aimed to refine the keyword set and improve the accuracy of subsequent searches.

The cleaned keywords were then used to query NewsAPI[13] for relevant articles. NewsAPI searches through articles published by over 80,000 news sources and blogs in the last 5 years and is updated live which provides our model with a comprehensive and dynamically updated evidence bank. Textual content was scrapped and extracted from each article's URL using Beautifulsoup[14] and a pre-trained SBERT model from the *paraphrase-MiniLM-L6-v2* weights was utilised to encode both the extracted keywords and article content into vector embeddings. Using the same model, the similarity score to the keywords for each article was calculated as the mean cosine similarity score between the article content and each extracted keyword. For the top 25 articles with the highest similarity scores to the keywords, the relevance score for each article was computed as the cosine similarity score between the article content and the claim as a whole. From each of the three articles with the highest relevance score, two sentences with the highest cosine similarity score to the claim were selected to form the textual evidence for that particular claim.

2. Claim Verification

2.1. CLIP multimodal Custom Truthfulness Classifier

Contextual representations for the claim, image and text evidence are obtained by tokenising them using the CLIP model. Stance representations between the claim and text evidence and between the image and text evidence are computed by applying attention mechanisms between the text evidence representations and both the claim representation and the image representation. Attention weights are calculated using the softmax function on the dot product of claim representations with

[13] https://newsapi.org/

[14] https://www.crummy.com/software/BeautifulSoup/

text evidence and image representations. The claim representation is updated based on these attention weights for both text evidence and images. A CLIP multimodal Custom Truthfulness Classifier consisting of two fully connected layers with a ReLU activation in between is developed for the task of claim verification. Sigmoid activation is applied to the logits for binary classification, and the function returns the loss, logits, and true labels. Binary cross-entropy loss was chosen as the loss function.

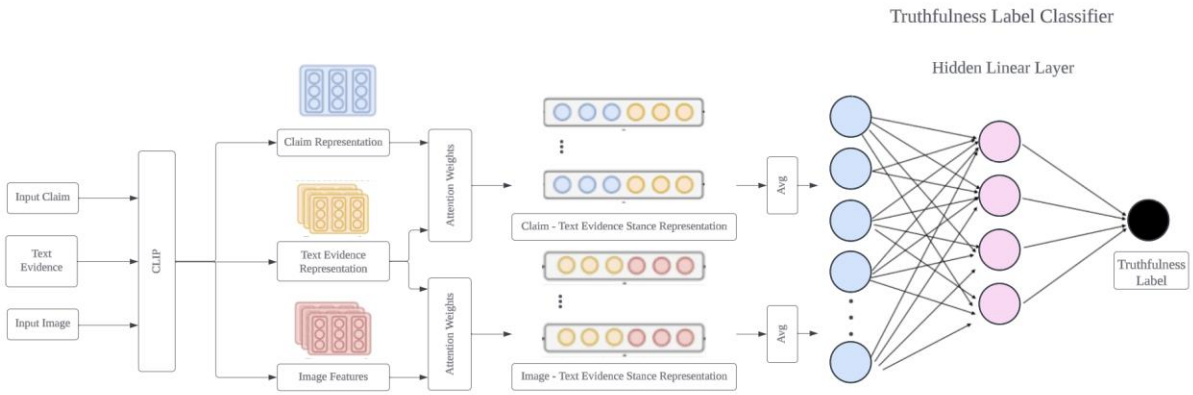


Fig.2. A diagram of our CLIP multimodal Custom Truthfulness Classifier

2.2. ChatGPT LLM

Current state-of-the-art Large Language Models (LLMs) which are trained on large-scale corpus have shown promising emergent abilities on various tasks (Wei et al., 2022a), including those requiring sophisticated reasoning and evaluation. Particularly, the effectiveness of LLMs has been enhanced through innovative prompt-engineering techniques, enabling impressive performance in zero-shot or few-shot tasks.

Thus, we aimed to leverage the strength of these LLMs by implementing a parallel LLM-involved misinformation detection system that leverages OpenAI's ChatGPT-4-32k. The system utilises a structured query format template where context placeholders are provided to insert modelling outputs from previous stages to formalise a holistic description to the language model:

Query Input

> “Given a claim {{{test_claim}}}, with the accompanying context {{{', '.join([article['title'] for article in wikipedia_articles])}}}, with the following accompanying news article evidences {{{evidence_part}}}, predict a truthfulness label given {{true, false}}and provide a reasoning explanation for the prediction."

Contextual information is obtained from the Wikipedia articles relating to each keyword in the claim using WikipediaAPI. The same evidence from the automated evidence retrieval module is used. The resulting misinformation label prediction is presented with a reasoning explanation for the prediction.

3. Satire Detector

Employing 'not satire’ (0) and 'satire' (1) labels, we developed a binary classifier for satire detection by fine-tuning the pre-trained bert-base-uncased model. To minimise class imbalance, we ensured approximately equal distribution between satire examples and non-satirical fake news examples.The dataset was split into training and testing sets with a 80/20 train-test split. Our training dataset contains 310 examples of satire and 310 examples of non-satirical fake news while our testing dataset contains 77 examples of satire and 77 examples of non-satirical fake news. BERT's tokenizer was applied for data tokenization. The tokenized sequences were converted to token IDs and padded to a maximum length of 512. The model contains a dropout layer and a linear layer with sigmoid activation for binary classification. The training loop uses 100 epochs, and the model was optimised using the Adam optimizer to adjust parameters. BERT model layers 10 and 11 were unfrozen to allow gradient computation and updates during training. Binary Cross-Entropy Loss was chosen as the loss function.

4. Out-of-Context Image Detector

Our framework employs the trained Catching Out-of-Context Misinformation using Self-Supervised Learning (COSMOS) model as our out-of-context image detector. The COSMOS model lacks specific codes for dataset collection and preparation, but does state the usage of Google’s Cloud Vision API, Spacy NER, SBERT and detectron2. Google’s Cloud Vision API helps us to reverse search the input image to get a second caption. Spacy NER replaces proper nouns in sentences with general items instead. Detectron2 provides us with detection and

segmentation algorithms that form bounding boxes for the images. Our implementation involved the execution of custom code for Spacy NER, the all-MiniLM-L6-v2 model for SBERT, and detectron2. The functions of Spacy NER, SBERT, and detectron2 encompass the augmentation of modified captions and entities list, computation of the BERT base score, and generation of bounding boxes, respectively, within the JSON files housing comprehensive data. The json file also includes the original two captions, the image path, the context label, and the article link.

5. Explanation Generator

For explanation generation, the ChatGPT model is prompted with a structured query format:

> Input Query (Explanations)
> "Given that this claim {{{test_claim}}}has been classified as {{{truthfulness_label}}}, {{{satire_label}}} , based on the evidence {{{evidence_list}}}, provide a reasoning explanation for the truthfulness label for the claim."

Input Parameters for ChatGPT-4-32k	
Temperature	0.4
Max_tokens	200
Top_p	0.8
Frequency_penalty	0.0
Presence_penalty	0.0

Table 4. Optimised input parameters for ChatGPT

RESULTS AND DISCUSSION

Evidence Retrieval

As a benchmark, we manually collected evidence for 10 claims based on our human perception of relevance. The six manually-retrieved sentences and the six sentences retrieved by the module were then compared.

When only the similarity score was used for filtering, the model showed poor performance. Out of the 60 evidence sentences selected by the model, 39 were irrelevant.

When both similarity score and relevance score were used for filtering, the performance of the model improved significantly. The model was able to identify 58 relevant sentences that were similar in ideas to those that were manually selected for all, with only two irrelevant sentences retrieved. Variations in sentences were mainly due to differences in phrasing and not ideas.

Discussion

From our results, the second filter using the relevance score between the article content and the claim in its entirety is crucial. We hypothesise that this is because when only the similarity score was used, the model likely focused more on surface-level lexical and syntactic similarities, leading to the inclusion of sentences that contained the same keywords albeit with different meaning and context.

This method of automating evidence retrieval by using real-time APIs can provide the following improvements to automated fact-checking models:

1) Improved performance due to access of a large evidence bank
2) Utility in larger range of topics
3) Ability to stay up-to-date

1.1. ChatGPT LLM

The ChatGPT-enabled misinformation detection system was tested with the same test set as the CLIP multimodal Custom Truthfulness Classifier. In the absence of supporting evidence, the ChatGPT-enabled misinformation detection system shows a poor F1-score of 62.9%. Without evidence, the system mainly predicts a truthfulness label based on the credibility of the alleged source in the claim and the level of detail provided. For instance, when evaluated on the false claim "Singapore Police Force about 2 years ago: The Police would like to clarify that the video circulating online that showed a mass brawl involving workers did indeed happen in Singapore.", the system inaccurately predicted the claim to be true, citing the reason that "The claim is made by the Singapore Police Force, a reliable and official source of information. There is no reason to doubt the authenticity of this claim." However, this approach of relying solely on the perceived credibility of sources is impractical as individuals with malicious intent can easily exploit the reputation of credible sources, employing their names to propagate misinformation. This highlights the need for a more robust and evidence-driven approach to effectively combat misinformation in the real-world context.

However, with evidence, the ChatGPT-enabled misinformation detection system achieved an F1-score of 95.4% on the same test set. The improvement in accuracy from 62.9% to 95.4% in F1-score highlights our proposed framework for automated evidence retrieval as effective in significantly improving performance. Notably, when provided with evidence, the system made only two errors, wherein the explanations provided were accurate, but the assigned labels were erroneously assigned. For example, for the claim "Mechanic cheated friend into thinking he could buy Mercedes-Benz car for S$140,000 including COE", ChatGPT provided an accurate explanation "The claim states that the mechanic cheated his friend into thinking he could buy a Mercedes-Benz car for S$140,000 including COE, which is consistent with the information provided in the news article evidence. The mechanic lied to his friend about being able to procure a heavily discounted Mercedes-Benz car, and the friend did not receive the car. Therefore, the claim is not misinformation." but incorrectly predicted the label as "False". This highlights the system's proficiency in comprehending and identifying Singapore-context misinformation, with the rare misclassification arising from labelling discrepancies rather than substantive misunderstandings.

For the same example about the Singapore Police Force, when provided with evidence, the system was able to predict a truthfulness label based on the correct reasoning "The claim states that the Singapore Police Force clarified that a video of a mass brawl involving workers did happen in Singapore. This is refuted by the accompanying news article evidence, which confirms that the police stated the video did not occur in Singapore and was actually taken at a Shanghai shopping centre. Therefore, the claim is false according to the provided evidence."

	Without Evidence	**With Evidence**
Precision	0.59459	0.96875
Recall	0.66667	0.93939
Accuracy	0.60000	0.95385
F1-Score	0.62857	0.95385
Confusion Matrix	[[17 15] [11 22]]	[[31 1] [2 31]]

Table 5. Results for ChatGPT-enabled misinformation detection system, with and without evidence

2.Satire Classifier

The satire classifier model had a final F1 score of 93.18%, indicating strong performance.

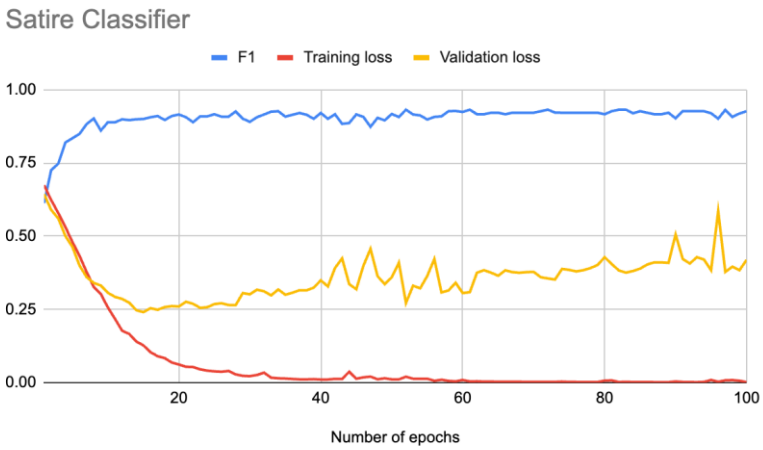


Fig 3. Results of our Satire Classifier

3.Out-of-Context Detector

We queried 10 Singapore context claims to test the accuracy of the trained model in-domain and verified that it was able to attain satisfactory performance for the domain transfer task.

4.Explanation Generation

Refer to Table 6. for examples of ChatGPT-generated explanations and our gold explanations. We employed BLEU, METEOR, CIDEr and ROUGE as the evaluation metrics for assessing the performance of our explanation generation system leveraging ChatGPT. When evaluated against the gold explanations, our system demonstrated strong performance in producing high-quality explanations. This is seen from the high average score for each metric, as shown in Table 3.

ChatGPT-generated explanation	Gold explanation
The claim is supported by the evidence provided in the news article. The article states that the number of Chinese travellers entering Singapore is expected to increase due to the mutual 30-day visa-free travel agreement between both countries. This is expected to push inbound travel volume closer to pre-pandemic levels. Furthermore, the article mentions that during the peak season between July and August, travel bookings returned to 85 percent of pre-pandemic levels. The Singapore Tourism Board	The claim is labelled as true because the evidence supports it. The evidence states that the number of Chinese travellers entering Singapore is expected to increase due to the mutual 30-day visa-free travel agreement between both countries. This is expected to push inbound travel volume closer to pre-pandemic levels. Additionally, the evidence also mentions that during the peak season, travel bookings returned to 85% of pre-pandemic levels, and tourist arrivals from China are climbing back up to pre-pandemic levels. Therefore,

also confirms that tourist arrivals from China are climbing back up to pre-pandemic levels. Therefore, the claim is true.	the claim that Singapore's visa-free agreement with China may push inbound travel closer to pre-pandemic levels is true.
The truthfulness label for the claim is "false". The news article evidence clearly states that part of Fullerton Road will be closed to traffic from 4pm on New Year's Eve until 5am the next day due to the Marina Bay countdown activities. This directly contradicts the claim that no part of Fullerton Road will be closed as a security measure for the event.	The claim is labelled as false because the evidence states that part of Fullerton Road will be closed to traffic from 4pm on New Year's Eve until 5am the following day due to the countdown activities at Marina Bay, refuting the claim that no part of Fullerton Road will be closed for the event.

Table 6. Examples of ChatGPT-generated explanations and our gold explanations

	Maximum	**Minimum**	**Average**
BLEU	0.84050	1.80E-231	0.36110
METEOR	0.94511	0.15620	0.44984
CIDEr	0.00419	0.00182	0.00140
ROUGE	0.91304	0.26374	0.60931

Table 7. Results for Explanation Generation

FUTURE RECOMMENDATIONS

In the current limitations, the model is constrained by the availability of news articles from the NewsAPI free plan to articles up to only 1 month old and subjected to a 24-hour latency. We note that this would be easily resolved with a premium paid access to news aggregating sources for up-to-date content. Secondly, the absence of an existing dataset specific to the Singapore context for multimodal misinformation poses a challenge to model training and performance evaluation as most of the misinformation content presents itself as unimodal (text) focus. Looking ahead, we hope to expand our training dataset for Singapore-context misinformation and explore the integration of deep learning models capable of comprehending graphical relationships for misinformation detection.

CONCLUSION

In conclusion, we have collected a mass original dataset for Singapore-context misinformation, developed and implemented an end-to-end model for detection of Singapore-context misinformation. Our model accomplishes (1) the identification of multimodal misinformation specific to the Singapore context, (2) the inclusion of a custom satire recognition detector that is able to differentiate exaggerated content commonly present in social media, (3) the recognition of instances of out-of-context image usage, (4) the generation of reasoning explanations that enhance user comprehension, and (5) the automation of evidence retrieval with real-time updates, ensuring practicality and functionality in real-world scenarios. Our model has achieved strong performance across all functions and is a novel model that specifically addresses Singapore-context misinformation.

REFERENCES

1. The Presence of Unexpected Biases in Online Fact-checking. (n.d.). https://doi.org/10.37016/mr-2020-53

2. Park, S., Park, J. Y., Kang, J. H., & Cha, M. (2021). The presence of unexpected biases in online fact-checking. The Harvard Kennedy School Misinformation Review.

3. Horrigan, J. B. (2016). Information overload | Pew Research Center. Pew Research Center: Internet, Science & Tech. https://www.pewresearch.org/internet/2016/12/07/information-overload/

4. Ma, S., Bergan, D. E., Ahn, S., Carnahan, D., Gimby, N., McGraw, J., & Virtue, I. (2022). Fact-checking as a deterrent? A conceptual replication of the influence of fact-checking on the sharing of misinformation by political elites. Human Communication Research, 49(3), 321–338. https://doi.org/10.1093/hcr/hqac031

5. Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media: definition, manipulation, and detection. ACM SIGKDD explorations newsletter, 21(2), 80-90

6. Aggarwal, S., Sahu, P., Gupta, T., & Das, G. (2022). GPTs at Factify 2022: Prompt aided fact-verification. In Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR

7. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. International Conference on Machine Learning.

8. Gao, J., Hoffmann, H. F., Oikonomou, S., Kiskovski, D., & Bandhakavi, A. (2021). Logically at Factify 2022: Multimodal Fact Verification. arXiv preprint arXiv:2112.09253.

9. Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media: definition, manipulation, and detection. ACM SIGKDD explorations newsletter, 21(2), 80-90.

10. Zhou, J., Han, X., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2019). GEAR: Graph-based evidence aggregating and reasoning for fact verification. arXiv preprint arXiv:1908.01843.

11. Menglong Yao, B., Shah, A., Sun, L., Cho, J. H., & Huang, L. (2022). End-to-End Multimodal Fact-Checking and Explanation Generation: A Challenging Dataset and Models. arXiv e-prints, arXiv-2205.

12. Chaitanya, B. S. N. V., Prathyush, P., & Rutvik, V. (2021). Truthformers at Factify 2022: Evidence aware Transformer based Model for Multimodal Fact Checking.