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

INTRODUCTION

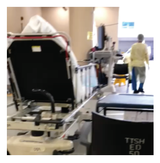
500 MILLION tweets
65 BILLION WhatsApp messages

→ **MISINFORMATION** spreads **faster** and more **extensively**

→ Public **skepticism**

→ Social **polarisation**

Example of Singapore-context misinformation



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

RESEARCH OBJECTIVE

To develop a **multi-stage** model with **model chaining** to **aggregate output** that 1) detects **Singapore-context multimodal misinformation** (**text & image**) 2) detect **satire** and **out-of-context image usage** and provides **explanations** to promote user understanding 3) **automates evidence retrieval** updated in real-time to be practical and functional in real-world context

LIMITATIONS OF CURRENT MODELS

1) No existing model for **Singapore-specific** misinformation
2) **Scarcity** of **end-to-end** fact-checking models
3) Prevalent **binary classification** system does not promote **understanding** of nuanced veracity of claims
4) **Fixed** training dataset with **limited** topics, causing **poor** performance for **novel topics**
5) **Impractical** as **evidence retrieval** is **not automated**

METHODOLOGY

Mass Dataset Collection

Diverse sources to mirror real-world complexity

295 claim-image pairs

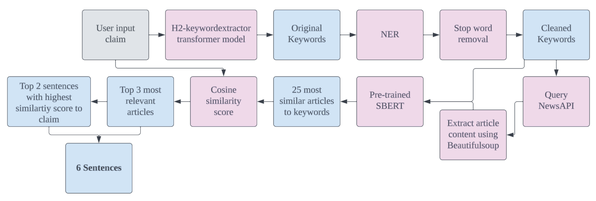
True	False
Straits Times headlines	Facebook
Channel News Asia headlines	WhatsApp
	Black Dot Research
	Factually

Satire Dataset Collection

	Satirical	Non-satirical
Self-collected	184	104
Golbeck et. al dataset	203	283

To circumvent the **lack** of Singapore-context satirical claims online, **ChatGPT** was leveraged on to **automate** the claim-generation process

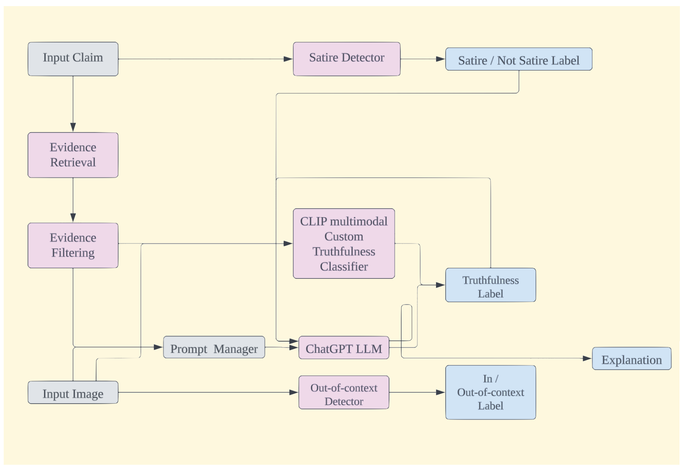
Automated Evidence Retrieval



Satire Detector

100 Epochs **ADAM** Optimiser
BCE Loss **1 E-3** Learning Rate

Pre-trained bert-base-uncased model (layers 10 and 11 unfrozen) → Binary classifier for satire detection → **'satire' (1) or 'not satire' (0) label**



ChatGPT LLM

"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."

CLIP CTC

2 fully connected layers **ReLU** activation
500 Epochs **1 E-3** Learning Rate

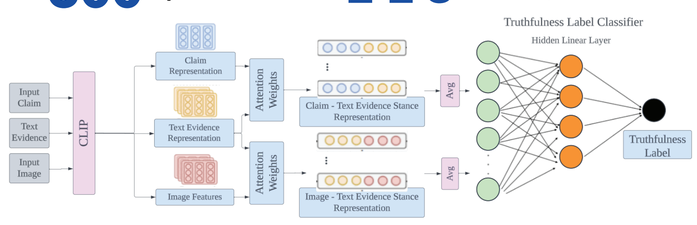
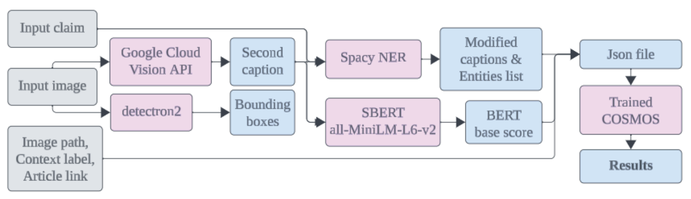


Image adapted from MOCHEG (Yao et. al)

OoC Detector



RESULTS & DISCUSSION

Automated Evidence Retrieval

10 claims, 6 manually-retrieved VS 6 module-retrieved sentences

	Accuracy
Similarity Filter	**65%**
Similarity + Relevance Filters	**96.7%**

Model focused on surface-level **lexical** and **syntactic** similarities

Relevance filter is **important**

✓ Improved performance due to access to **large evidence bank**
✓ Utility in a **larger range** of **novel** topics due to **dynamic updates**
✓ Ability to stay **up-to-date**

Satire Classifier

77 test claims

"Satire" prediction	"Not Satire" prediction
HDBs set to receive 'genius windows' that can predict rain before meteorologists even get the memo.	The new oximeter can be used to measure blood oxygen in your soft toys

Strong performance in identifying **nuances** between both classes

F1-score **93.2%**

Explanation Generation

Gold explanations VS **ChatGPT-generated**

	Max	Min	Avg
BLEU	0.841	1.80 E-231	**0.361**
METEOR	0.945	0.156	**0.450**
CIDEr	0.004	0.002	**0.001**
ROUGE	0.913	0.264	**0.609**

Strong performance in producing **high-quality** explanations → **Enhances user understanding**

CLIP CTC

ChatGPT LLM

Same test set with 59 claim-image pairs

An example of a **correctly-predicted** example

Image

Claim: "Staff lost control of migrant worker in dorm!!"

F1-score **95.2%**

→ **"False"**

	Without Evidence	With Evidence
Precision	0.59	**0.97**
Recall	0.67	**0.94**
Accuracy	0.60	**0.95**
F1-Score	0.63	**0.95**

Predict label based on the **credibility** of alleged source and **level of detail** → **Impractical**

2 errors, **accurate explanations** but erroneous labels → Proficient, no substantive misunderstanding

ChatGPT-generated explanation	Gold explanation
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.

With Evidence	Without Evidence
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 truthfulness label for the claim is "false". The evidence directly contradicts the claim that no part of Fullerton Road will be closed as a security measure for the event.

Evidence is important in reasoning process

FUTURE WORK

NewsAPI Free Plan provides only articles up to **1 month old**, with a **24h latency**	Premium paid access for updated content
Absence of Singapore-specific dataset	Expand our dataset
Two-layer CLIP CTC structure may not provide the depth required to learn intricate representations	1. Explore complex model architectures 2. Increase input dimension

SOCIAL IMPACT

DIGITAL SOCIAL WELLBEING **DECISION CONFIDENCE** **ONLINE SAFETY**

REFERENCES

1. All images were self-created unless stated otherwise
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. 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.

Members:
Goh Rou Hui Ashley, River Valley High School
Felicia Tan Ee Shan, Raffles Girls' School
Mentor:
Adriel Kuek, DSO National Laboratories

YDSP Young Defence Scientists Programme
DSTA Defence Science & Technology Agency
DSO NATIONAL LABORATORIES