Towards Automatic Annotation and Detection of Fake News

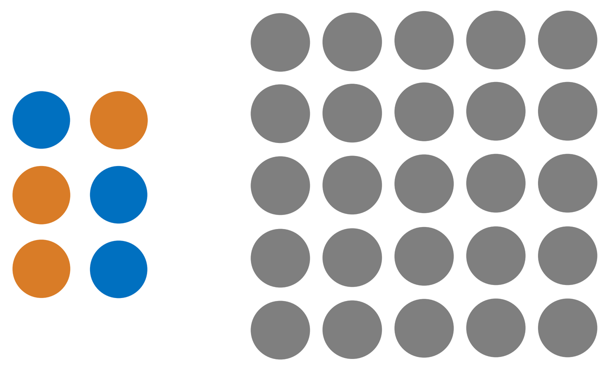
Mohammad Majid Akhtar†, Ishan Karunanayake†, Bibhas Sharma†, Rahat Masood†, Muhammad Ikram‡, Salil S. Kanhere† (†UNSW Sydney, ‡Macquarie University)

A Introduction and Problems

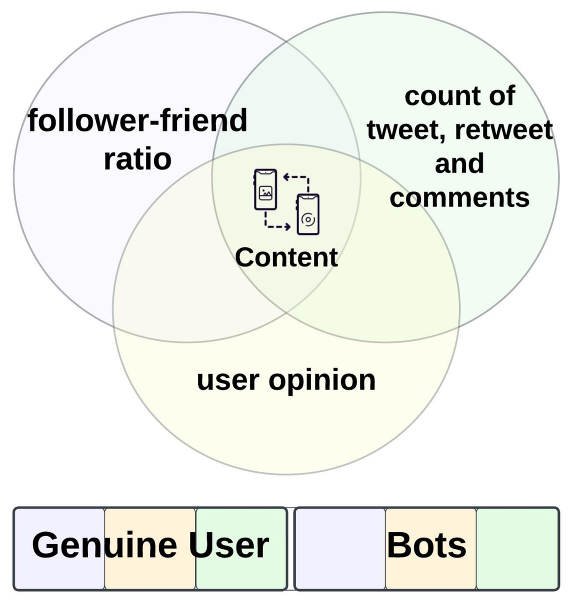
“ As misinformation spreads, it reinforces and amplifies our prejudices. It affects everyone and differently. ”

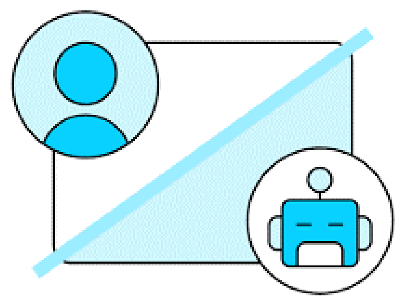
Problem 1: Relying on Manual Annotation. It is Hard to Scale and Incur Delays

Problem 2: Lack of Large Labeled Dataset to build a robust Fake news detection model



Problem 3: Existing work do not consider bots in the spread of false information





Bots can easily modify metadata around tweets such as by fake replies or fake likes.

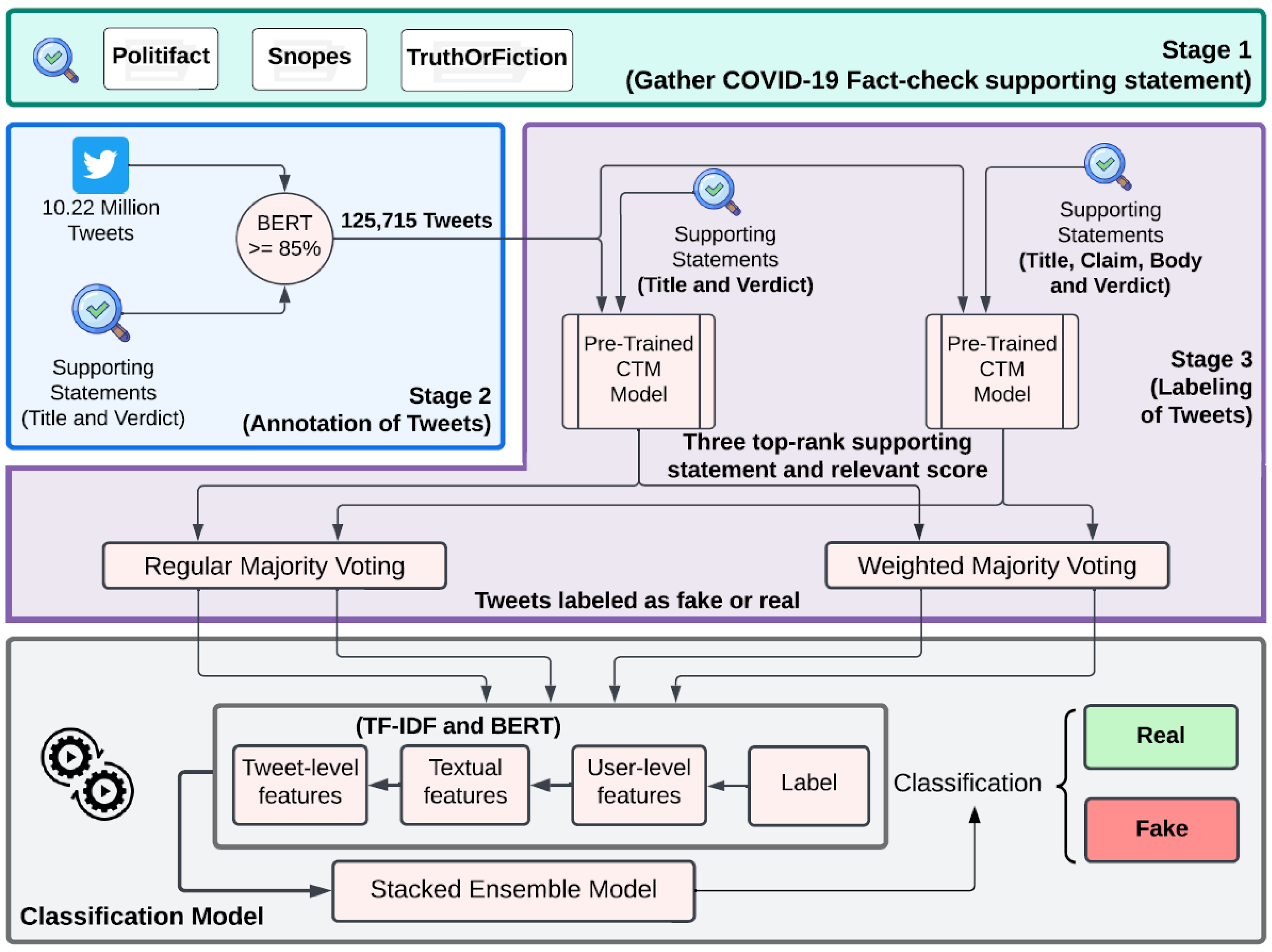
B Contributions

C1. We propose an **automatic annotation model** to annotate the textual content of posts (tweets) using verified fact-checked statements (supporting statements).

C2. We design and develop an **ensemble stack model** to detect fake information.

C3. We investigate the **impact of bots** in our dataset with regards to misinformation.

C Methodology



From fact-checking to classification: A multi-stage journey in debunking false claims. In *Stage 1*, we gather fact-checked statements related to COVID-19 (taken as a case study). These verified statements are claims that have been debunked by organizations like PolitiFact. In *Stage 2*, we leverage BERT-based cosine similarity to identify tweets debunked by fact-checked statements, using a comparison with statements and tweets. The filtered tweets are then passed to our labeling algorithms in *Stage 3*. Finally, we use the newly annotated tweets to train and test our classification model to detect fake news.

D Results

Result C1 **Table 1: Annotation result using our model. Table also shows the split 80:20 used for training of the model.**

Labeled Data Description		Labeled *	Total Tweets	Train (80%)		Test (20%)	
				Fake	Real	Fake	Real
cosine distance >= 0.85		Data 0	125,715	86,738	13,834	21,688	3455
Without body text	Regular Majority	Data 1	125,709	74,638	25,929	18,710	6,432
	Weighted Majority	Data 2	125,709	74,008	26,559	18,611	6,531
With body text	Regular Majority	Data 3	125,709	77,200	23,367	19,261	5,881
	Weighted Majority	Data 4	125,709	75,697	24,870	18,903	6,239

Result C2 **Table 2: Performance of our ensemble stack model.**

Labeled Data Description		Without body text						With body text					
		Regular Majority (Labeled Data 1)			Weighted Majority (Labeled Data 2)			Regular Majority (Labeled Data 3)			Weighted Majority (Labeled Data 4)		
		TF-IDF	BERT	C-BERT	TF-IDF	BERT	C-BERT	TF-IDF	BERT	C-BERT	TF-IDF	BERT	C-BERT
Stacked_{DT}	Precision	77%	77%	76%	77%	77%	76%	82%	81%	80%	81%	80%	79%
	Recall	98%	98%	99%	97%	98%	97%	96%	97%	98%	96%	97%	97%
	F1-score	86%	86%	86%	86%	86%	86%	89%	89%	88%	88%	88%	87%
	FPR	~2%	~2%	<1%	~3%	~2%	~1%	~4%	~3%	~2%	~4%	~3%	~3%
Stacked_{SVM}	Precision	78%	77%	76%	78%	76%	75%	83%	81%	80%	82%	80%	79%
	Recall	96%	99%	99%	96%	98%	100%	96%	96%	97%	96%	95%	96%
	F1-score	86%	86%	86%	86%	86%	86%	89%	88%	88%	88%	87%	87%
	FPR	~4%	~1%	~1%	~4%	~2%	<1%	~4%	~3%	~3%	~4%	~5%	~4%

Result C3 **Table 3: Number of bots generated in four labeled dataset**

Data	# Fake Tweets	# Bot-generated Tweets	%
Labeled Data 1	93,348	9,885	~11%
Labeled Data 2	92,619	9,894	~11%
Labeled Data 3	96,461	9,574	~10%
Labeled Data 4	94,600	9,417	~10%

E Analysis

- Our model achieved a **precision score of 83%, recall score of 96%** and a false positive rate of 4% when utilizing TF-IDF for extracting the tweet's textual features as shown in Table 2.
- Additionally, we provided evidence that bots play an active role in disseminating misinformation i.e., **bots generate approximately 10% misinformation tweets** as shown in Table 3.
- We also showed that bots behavior changes over time, depicting that **bots are more active during misinformation campaigns.**

F Conclusion

We proposed annotation model for creating large datasets using COVID-19 as a case study and a machine learning classifier. We also show bots play an active role in disseminating misinformation and changes behavior.

1. Akhtar, M. M., Karunanayake, I., Sharma, B., Masood, R., Ikram, M., & Kanhere, S. S. (2023, October). Towards Automatic Annotation and Detection of Fake News. In 2023 IEEE 48th Conference on Local Computer Networks (LCN) (pp. 1-9). IEEE.
2. Akhtar, M. M., Masood, R., Ikram, M., & Kanhere, S. S. (2023). False Information, Bots and Malicious Campaigns: Demystifying Elements of Social Media Manipulations. arXiv preprint arXiv:2308.12497.

UNSW Institute for Cyber Security

UNSW SYDNEY