Improving Generalization for Multimodal Fake News Detection

ESR 18: Sahar Tahmasebi
TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
sahar.tahmasebi@tib.eu

Motivation

- Increase of multimodal misinformation and its alarming impact on society

Existing datasets for multimodal fake news detection:

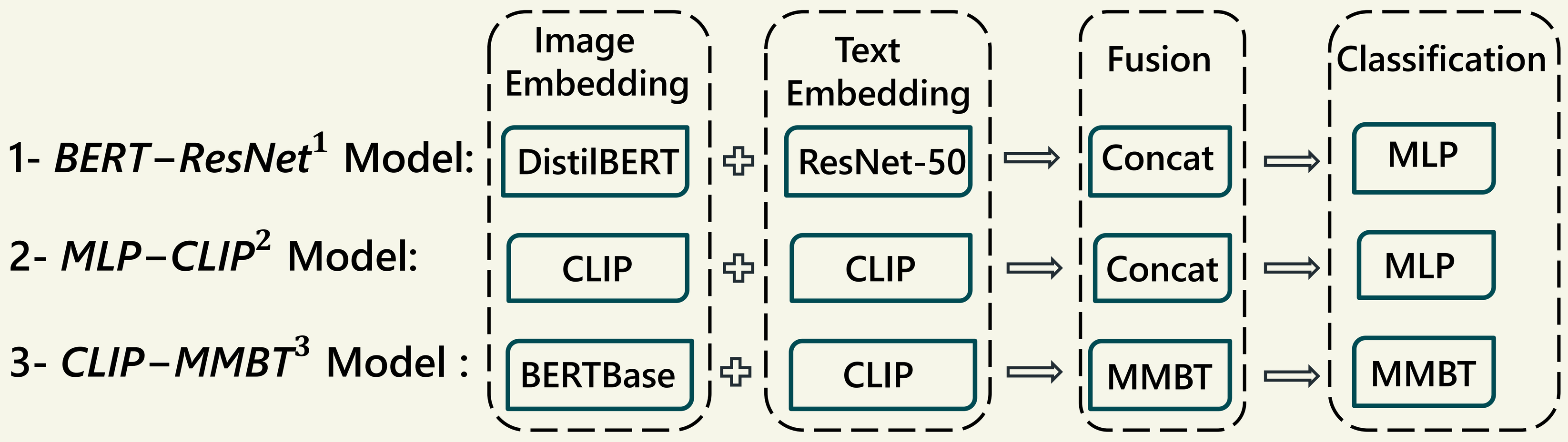
- Rather small size
- Limited set of specific topics

As a consequence:

- Poor generalization capabilities of models
- Not applicable to real-world data

Proposed Models

- Three multimodal approaches for effective fake news detection
- Based on state-of-the-art multimodal transformers
- Get a text-image pair as input and predict *fake* or *real*.



[1]Bidirectional Encoder Representations from Transformers – Residual Network (BERT–ResNet)
[2]Multi Layer Perceptron – Contrastive Language–Image Pre–training (MLP–CLIP)
[3]Contrastive Language–Image Pre–training – MultiModal BiTransformers(CLIP–MMBT)

Experimental Setup and Results

- Compared our models with reproduced BDANN [3] and Spotfake [4]
- Used MediaEval 2015 and MediaEval 2016 for the comparison
- MLP-CLIP outperformed our models and reproduced models in both datasets

New Test Scenarios

Motivation: Test model generalization in realistic use cases

Idea: Manipulate the content and evaluate model performance on new test set

How? Manipulation of *real* posts from MediaEval (ME) 2015 [1] dataset (Figure 1) :

- *Event Replacement* : Events have been randomly replaced wih other events in the dataset. This changes the ground truth of all samples from *real* to *fake*.
- *Event Removal:* All events have been removed from text. As the ground truth can be both real or fake, one expert manually annotated the samples.
- *Replacement with Fake Image:* Images have been replaced with other images depicting a different event in the test set. The new ground truth is fake.
- *Replacement with Real Image:* image have been replaced with similar image depicting same event. All samples remain real after the manipulation.

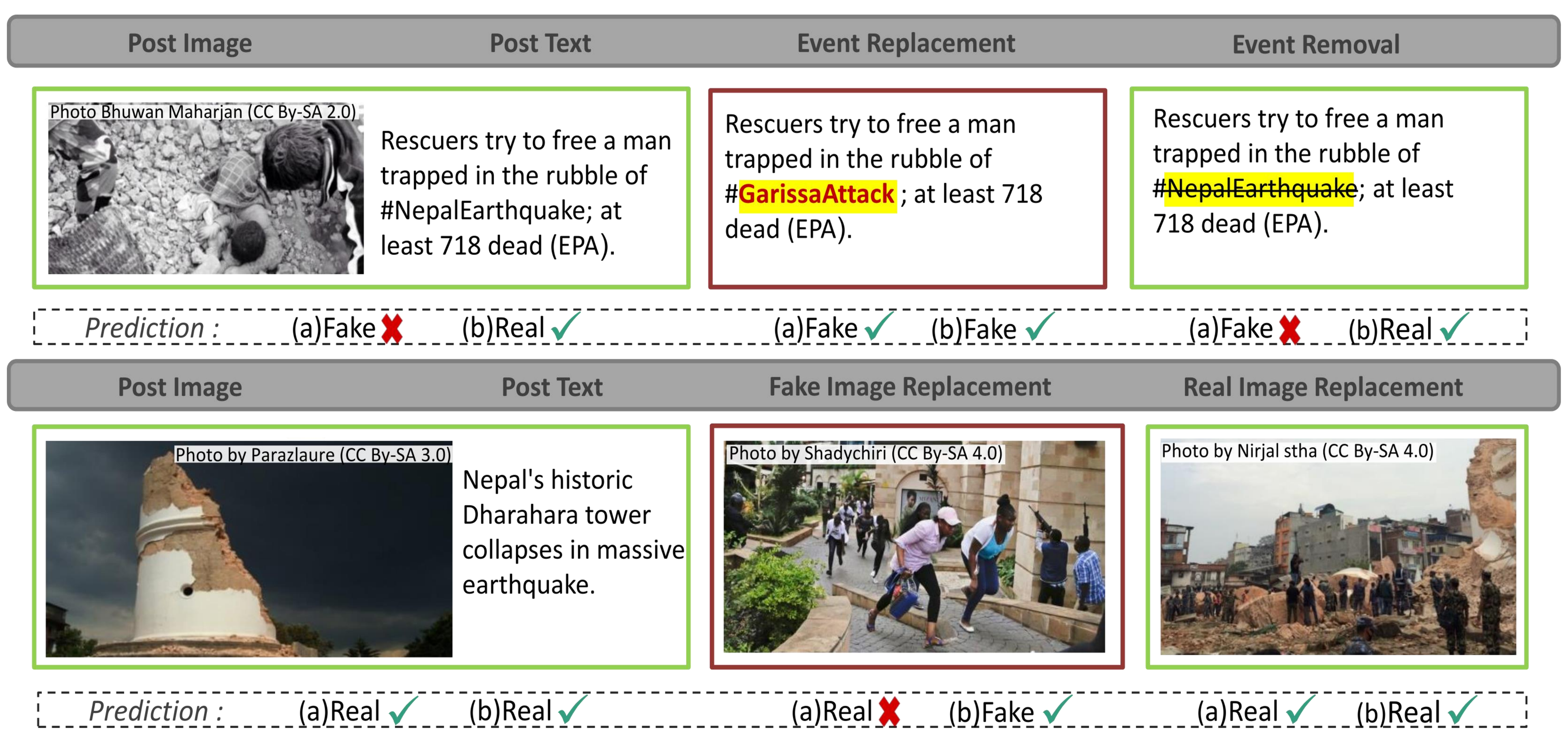


Figure 1. Manipulation techniques and results of (a) Spotfake (b) MLP-CLIP(Ens). The border color denotes the ground truth (green: real, red: fake). Images are replaced with similarones due to licensing issues.

Training Strategy to Improve Generalization

To reduce the bias and improve model generalization:

- VNME dataset: an extension of the ME dataset with Visual News (VN) dataset [2]
- More samples from many domains, topics, and events
- Using real images and their associated captions of VN as *real* samples
- Creating fake samples using the aforementioned manipulation techniques as follow:

Dataset	Manipulation Strategy	VNME (Img)	VNME (Evt)	VNME (All)
Visual News	Original	✓	✓	✓
	EvRep	×	✓	✓
	FakeIm	✓	×	✓
MediaEval	Original	✓	✓	✓
	EvRep	×	✓	✓
	FakeIm	✓	×	✓

- ➢ We train our best model *(MLP-CLIP)* based on three above training data variants to evaluate their impact.
- ➢ We evaluate the *MLP-CLIP (VNME-Ens)* that combines the outputs of the previous models by majority voting.

Table 2. Accuracy (*Acc*) and number of samples predicted as fake (N_F) and real (N_R) for different models and test data manipulations (number of fake / real ground-truth samples). Models denoted with ‡ are solely trained on ME 2015. Note that models with * are reproduced and that VNME-Ens is an ensemble of MLP-CLIP models trained on VNME.

Method	ME 2015 717 / 1,215	Original 0 / 100			FakeIm 100 / 0			RealIm 0 / 100			EvtRep 100 / 0			EvtRem 6 / 94			Total 206 / 294
	Acc	N_F	N_R	*Acc*	N_F	N_R	*Acc*	N_F	N_R	*Acc*	N_F	N_R	*Acc*	N_F	N_R	*Acc*	*Acc*
BDANN*, ‡	0.76	16	84	0.84	12	88	0.12	15	85	0.85	19	81	0.19	17	83	0.77	0.55
Spotfake*, ‡	0.84	37	63	0.63	30	70	0.30	18	82	0.82	37	63	0.37	37	63	0.61	0.54
BERT-ResNet, ‡	0.87	28	72	0.72	25	75	0.25	21	79	0.79	28	72	0.28	28	72	0.68	0.54
CLIP-MMBT, ‡	0.75	3	97	**0.97**	10	90	0.10	2	98	0.98	4	96	0.04	4	96	**0.90**	0.59
MLP-CLIP, ‡	**0.93**	27	73	0.73	40	60	0.40	31	69	0.69	51	49	0.51	39	61	0.41	0.54
• VNME-Img	0.69	3	97	**0.97**	90	10	0.90	5	95	0.95	24	76	0.24	16	84	0.80	0.77
• VNME-Evt	0.70	6	94	0.94	20	80	0.20	19	81	0.81	75	25	**0.75**	47	53	0.51	0.64
• VNME-All	0.68	21	79	0.79	100	0	**1.00**	0	100	**1.00**	61	39	0.61	38	62	0.60	0.80
• VNME-Ens	0.70	6	94	0.94	100	0	**1.00**	3	97	0.97	62	38	0.62	35	65	0.63	**0.83**

Experimental Findings :

Performance drop on manipulated test variants for models trained only on ME. → Poor Generalization

Improved Generalization ← Performance robustness for models trained with a modality-specific data manipulation on manipulated sets specific to that modality.

Best overall performance averaged over all test sets for MLP-CLIP (VNME-Ens) which is an ensemble of all models trained with all modifications. → Most Reliable in Applications

Summary

1. Proposed three multimodal fake news detection models
2. Our MLP-CLIP outperformed baselines on the MediaEval 2015 dataset
3. Create more diverse test scenario by content manipulation
4. Provide a solution to improve model generalization

Future work

1. Explore different kinds of manipulation techniques
2. Different fusion strategies for the ensemble mode

References

[1] Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, and Yiannis Kompatsiaris. Verifying Multimedia Use at MediaEval 2015.

[2] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual News: Benchmark and Challenges in News Image Captioning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.

[3] Tong Zhang, Di Wang, Huanhuan Chen, Zhiwei Zeng, Wei Guo, Chunyan Miao, and Lizhen Cui. 2020. BDANN: BERT-Based Domain Adaptation Neural Network for Multi-Modal Fake News Detection. In International Joint Conference on Neural Networks, IJCNN 2020

[4] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. SpotFake: A Multi-modal Framework for Fake News Detection. In IEEE International Conference on Multimedia Big Data, BigMM 2019

[5] Sahar Tahmasebi, Sherzod Hakimov, Ralph Ewerth, and Eric Müller-Budack. Improving Generalization for Multimodal Fake News Detection. In International Conference on Multimedia Retrieval (ICMR '23), 2023.

TIB LEIBNIZ INFORMATION CENTRE FOR SCIENCE AND TECHNOLOGY UNIVERSITY LIBRARY

CLEOPATRA

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project "CLEOPATRA - Cross-lingual Event-centric Open Analytics Research Academy ".