EXMULF
An Explainable Multimodal Content-based Fake News Detection System

Sabrine Amri, **Dorsaf Sallami**, and Esma Aımeur

FPS 2021: The 14th International Symposium on Foundations & Practice of Security

Plan

- Introduction
- State of the art
- Proposed method
- Experiments and results
- Conclusion and future works

1

Introduction

- ❑ Digital era.
- ❑ World Wide Web.
- ❑ Share data across the globe.

- ➢ What is “fake news”?
- ➢ Fake news rapid propagation.
- ➢ Fake news impact on OSN users.

2

State of the art

Multimodal Content-based Fake News Detection

Multimodal approaches: textual data and visual data extracted from the news content

Techniques:

- Correlation between the attached images and the credibility of the news text
- Various techniques ranging from neural networks
- Semantic analysis
- Sentiment analysis
- Web scraping

Table 1: A comparison between the multimodal fake news detection approaches.

Reference	Techniques used	Datasets used
Xue et al.	BERT, ResNet50, cosine similarity.	MCG-FNeWS, PolitiFact, Twitter.
Zeng et al.	VGG model, multimodal variational autoencoder.	Twitter, Weibo.
Zhang et al.	BERT, VGG19.	Twitter, Weibo.
Kumari et al.	ABS-BiLSTM, ABM-CNN–RNN, MFB.	Twitter, Weibo.
Mangal et al.	VGG, Word2Vec, LSTM, cosine similarity.	Collected 1000 images from Google, Kaggle and onion for fake or real images with text.
Meel et al.	Hierarchical Attention Network (HAN), Caption and Headline matching (CHM), Noise Variance Inconsistency (NVI), Error Level Analysis (ELA).	Fake News Detection by Jruvika, All Data, Fake News Sample by Guilherme Pontes.
Giachanou et al.	BERT, VGG-16, cosine similarity.	FakeNewsNet.
Giachanou et al.	Word2Vec, VGG19, LBP.	MediaEval, PolitiFact, GossipCop.
Singhal et al.	BERT,VGG19.	Twitter MediaEval, Weibo.
Zhou et al.	Text-CNN, Text-CNN, image2sentence, cosine similarity.	PolitiFact, GossipCop.
Qian et al.	BERT, ResNet, attention mechanism.	Twitter, Weibo.
Yuan et al.	BERT, VGG19, Bi-LSTM, Graph-attention layer.	Twitter, Weibo.
Vishwakarma et al.	Optical Character Recognition (OCR), Web scraping.	A dataset of thousands of images collected from Google Images, the Onion, and Kaggle.
Shah et al.	Sentiment Analysis, Cultural Algorithms (CA).	Twitter, Weibo.

Explainable Fake News Detection

To achieve transparency in many applications such as fake news detection in online social networks.

Techniques:

- Attention neural network.
- SHAP.
- Tsetlin Machine (TM).
- MIMIC, ATTN, PERT. ..

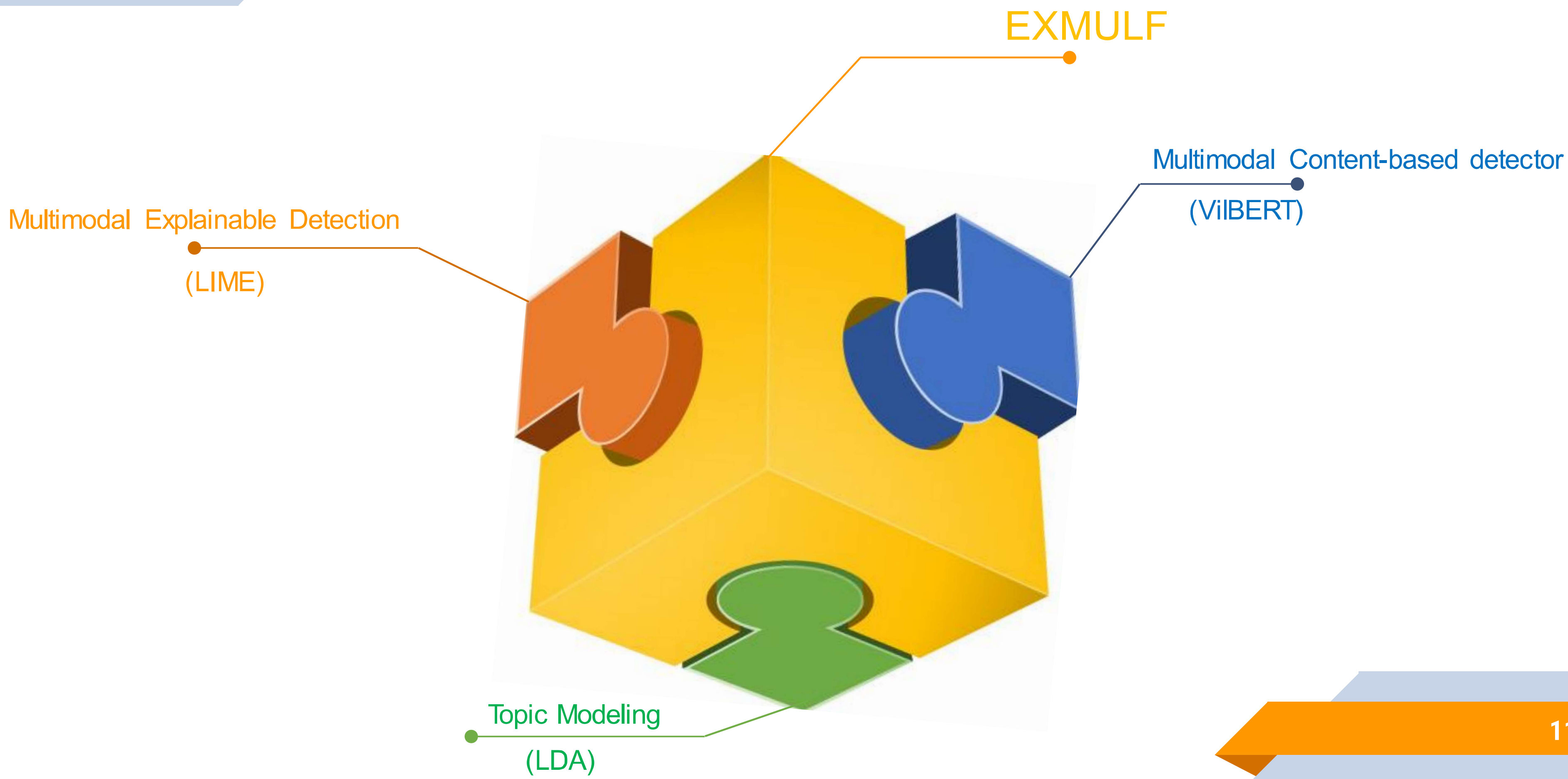
Table 2: A comparison between the explainable fake news detection approaches.

Reference	Approach	Techniques used	Datasets used
Shu et al.	DEFEND.	Attention neural network.	PolitiFact, GossipCop.
Reis et al.	–	SHAP.	BuzzFace.
Yang et al.	XFake.	MIMIC, ATTN, PERT.	An annotated benchmark dataset in the German language.
Lu et al.	GCAN.	Co-Attention Network.	Twitter datasets: Twitter15, Twitter16.
Przybyła et al.	–	Machine learning: linear method trained on stylometric features, a recurrent neural network method.	Fake News Corpus dataset.
Bhattarai et al.	TM framework.	Tsetlin Machine (TM).	PolitiFact, GossipCop.
Denaux et al.	–	NLP: semantic similarity and stance detection.	Clef18, FakeNewsNet, coinform250.
Silva et al.	Propagation2Vec.	Network embedding learning.	PolitiFact, GossipCop.

3

Proposed method

The proposed approach



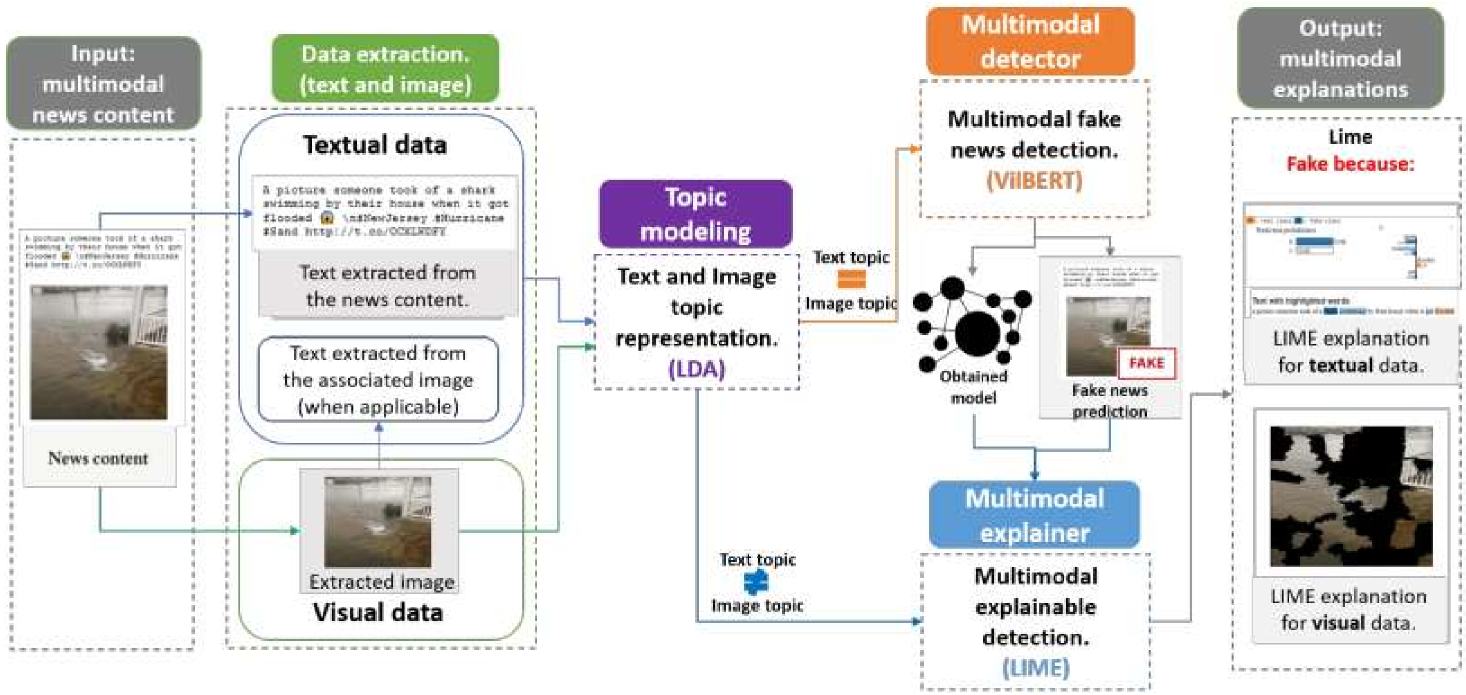
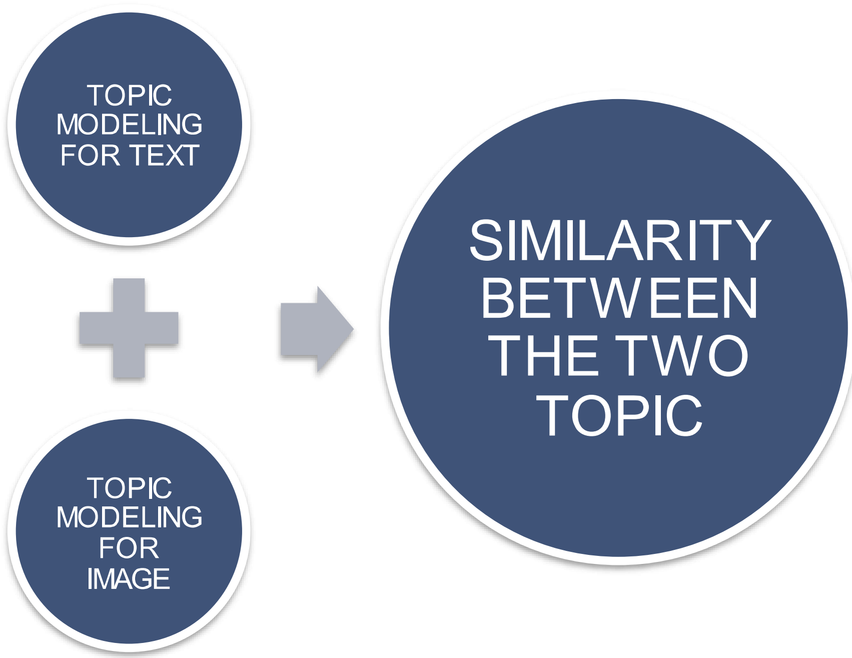


Fig. 1:EXMULF methodology overview

Topic Modeling



Why Vision-and-Language BERT (VilBERT)?

- ✓ Model for learning task-agnostic joint representations of image content and natural language.
- ✓ Two training objectives, masked multimodal learning and image text alignment prediction.
- ✓ High performance on a variety of visiolinguistic tasks.
- ✓ Learn semantic alignment/association between visual and language features through pretraining.

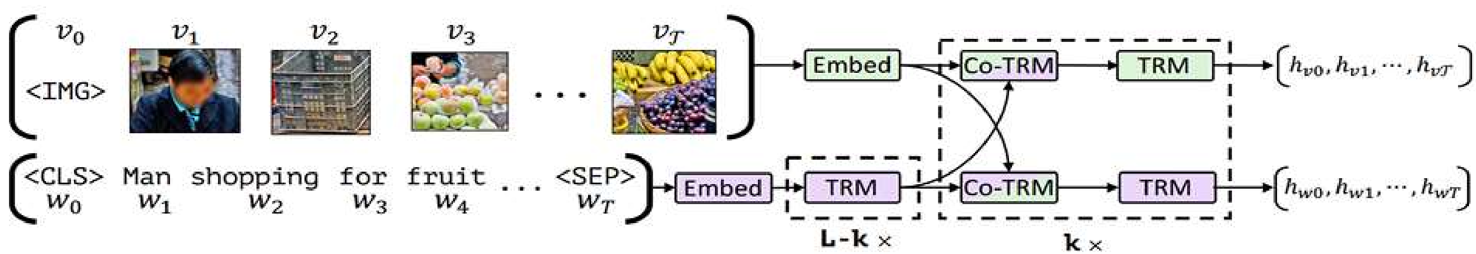


Fig. 2: ViLBERT Architecture

Why Local Interpretable Model-Agnostic Explanations (LIME)?

- Accessibility and simplicity.
- Model agnosticism: it can be used with any machine learning model.
- Gives local explanations: explanations for each observation instead of just the model itself.
- Interpretable: explanations based on the input features instead of abstract features

4

Experiments and results

Datasets used

Table 3:Statistics of the datasets used.

Dataset	Train		Test	
	Fake	Real	Fake	Real
Twitter	6841	5009	2564	1217
Weibo	3748	3783	1000	996

Data preprocessing

- **Removal of single modality instances**
- **Preprocessing of textual data:**
  - Removal of punctuation, symbols and emoji
  - Translating non-English text into English (just for Twitter dataset)
- **Preprocessing of images:**
  - Resizing all images to the same equal size
  - Extracting the text within the image (when applicable)

How have we used Vision-and-Language BERT (VilBERT)?

ViLBERT is applicable in the multimodal fake news detection task through fine-tuning on the datasets used

Learn visually grounded language understanding in the fake news context to help classify the news content.

Fine-tuning:
passing the element-wise product of the final image and text representations into a learned classification layer

Table 4: Results.

Dataset	Model		Accuracy	Fake News			Real News		
				Precision	Recall	F1	Precision	Recall	F1
Twitter	Text only	$BERT_T$	0.572	0.602	0.586	0.597	0.543	0.553	0.544
		$BERT_{T+IT}$	0.577	0.612	0.574	0.598	0.551	0.564	0.556
	Image only	ResNet-34	0.624	0.712	0.567	0.6	0.558	0.72	0.62
		VGG-19	0.596	0.698	0.522	0.593	0.531	0.698	0.597
	Multi-modal	Fusion	0.7695	0.820	0.726	0.779	0.719	0.798	0.748
		SpotFake [22]	0.7777	0.751	0.900	0.82	0.832	0.606	0.701
		AMFB [8]	0.883	0.89	**0.95**	0.92	**0.87**	0.76	0.741
		HMCAN [15]	0.897	**0.971**	0.801	0.878	0.853	**0.979**	**0.912**
		BDANN [30]	0.830	0.810	0.630	0.710	0.830	0.930	0.880
		VilBERT	**0.898**	0.934	0.92	**0.926**	0.859	0.88	0.869
Weibo	Text only	$BERT_T$	0.680	0.731	0.715	0.709	0.667	0.676	0.669
		$BERT_{T+IT}$	0.682	0.739	0.72	0.71	0.672	0.684	0.673
	Image only	ResNet-34	0.694	0.701	0.634	0.698	0.698	0.711	0.699
		VGG-19	0.633	0.640	0.635	0.637	0.637	0.641	0.639
	Multi-modal	Fusion	0.8152	0.865	0.734	0.88	0.764	0.889	0.74
		SpotFake [22]	0.8923	0.902	**0.964**	0.932	0.847	0.656	0.739
		AMFB [8]	0.832	0.82	0.86	0.84	0.85	0.81	0.83
		FND-SCTI [29]	0.834	0.863	0.780	0.824	0.815	0.892	0.835
		HMCAN [15]	0.885	0.920	0.845	0.881	0.856	0.926	**0.890**
		BDANN [30]	0.842	0.830	0.870	0.850	0.850	0.820	0.830
		VilBERT	**0.9204**	**0.946**	0.948	**0.946**	**0.879**	**0.893**	0.885

A picture someone took of a shark swimming by their house when it got flooded 😱 \n#NewJersey #Hurricane #Sand http://t.co/OCXLWDFY

Fig. 3:Input tweet example.

Fig. 4: LIME explanations for image data. (a) presents the original fake tweet (b) shows the superpixels that are generated using the quickshift segmentation algorithm (c) shows the area of the image that produced the prediction of the class (fake, in our case)

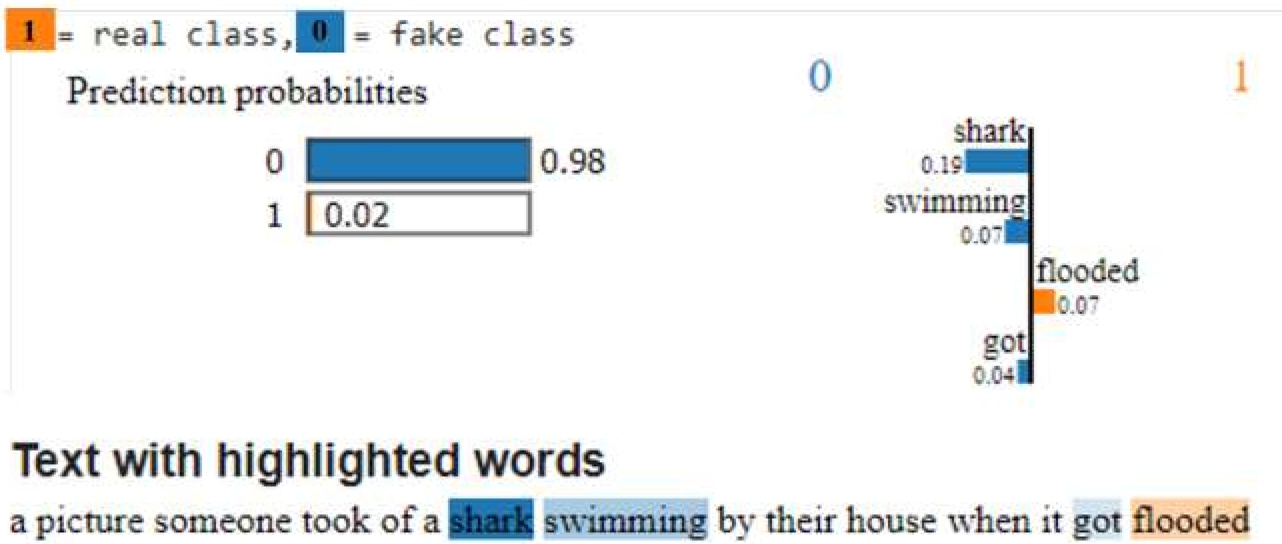


Fig. 5: LIME explanations for textual data

5

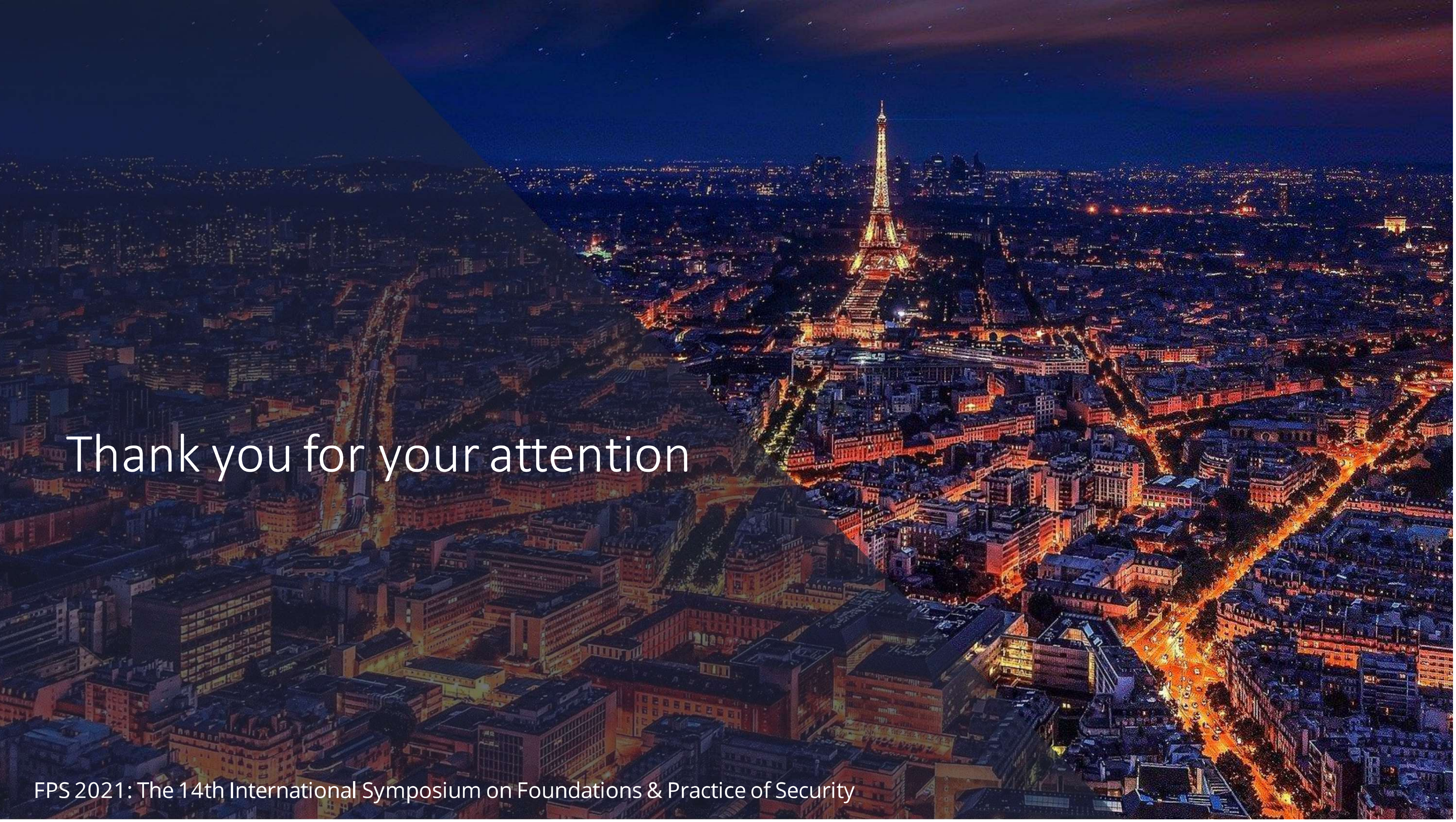
Conclusion and future works

EXMULF:

- ✓ takes as input the textual and the visual information within the content of the online news post
- ✓ detects whether this post is fake or real
- ✓ and explains the reasoning behind system decisions to OSN users

Future work:

- include audio and video as multimodal input data
- expand the visual representations (the effectiveness of explainability provided to OSN users)


Thank you for your attention
FPS 2021: The 14th International Symposium on Foundations & Practice of Security