Detecting Harmful Memes and Their Targets

Shraman Pramanick[1] , Dimiter Dimitrov[2] , Rituparna Mukherjee[1] ,
Shivam Sharma[1,3], Md. Shad Akhtar[1], Preslav Nakov[4] , Tanmoy Chakraborty[1]
[1]Indraprastha Institute of Information Technology, Delhi, India
[2]Sofia University, Bulgaria
[3]Wipro AI Labs, India
[4]Qatar Computing Research Institute, HBKU, Doha, Qatar

Introduction:

- **Internet memes** are formed of an **image** and a **short piece of text** embedded on it.

Introduction:

- **Internet memes** are formed of an **image** and a **short piece of text** embedded on it.

- Memes are difficult to analyze: **multimodality**, **context**-dependency, **morphed** image, **noisy** text

Introduction:

- **Internet memes** are formed of an **image** and a **short piece of text** embedded on it.

- Memes are difficult to analyze: **multimodality**, **context**-dependency, **morphed** image, **noisy** text

- Increasing use of memes with bad intentions: hateful memes[1], offensive memes[2]

[1]The Hateful Memes Challenge, Kiela et al., NeurIPS’20
[2]Multimodal meme dataset for identifying offensive content, Suryawanshi et al., , LREC-TRAC ’20

Introduction:

- **Internet memes** are formed of an **image** and a **short piece of text** embedded on it.

- Memes are difficult to analyze: **multimodality**, **context**-dependency, **morphed** image, **noisy** text

- Increasing use of memes with bad intentions: hateful memes[1], offensive memes[2]

- **Harm** has a broader perspective compared to hate and offense

[1]The Hateful Memes Challenge, Kiela et al., NeurIPS'20
[2]Multimodal meme dataset for identifying offensive content, Suryawanshi et al., , LREC-TRAC '20

Introduction:

- **Internet memes** are formed of an **image** and a **short piece of text** embedded on it.
- Memes are difficult to analyze: **multimodality**, **context**-dependency, **morphed** image, **noisy** text
- Increasing use of memes with bad intentions: hateful memes[1], offensive memes[2]
- **Harm** has a broader perspective compared to hate and offense

NOT hateful, NOT offensive but HARMFUL to Donald Trump

[1]The Hateful Memes Challenge, Kiela et al., NeurIPS’20
[2]Multimodal meme dataset for identifying offensive content, Suryawanshi et al., , LREC-TRAC ’20

Our Contributions:

- We formally define the notion of harmful memes in contrast to hateful and offensive memes.

Our Contributions:

- We formally define the notion of harmful memes in contrast to hateful and offensive memes.
- We formulate two novel problems
  - ➔ Problem 1 **(Harmful meme detection)**: very harmful, partially harmful, or harmless
  - ➔ Problem 2 **(Target identification of harmful memes)**: individual, organization, community/country, or society/general public/others

Our Contributions:

- We formally define the notion of harmful memes in contrast to hateful and offensive memes.

- We formulate two novel problems

  ➔ Problem 1 **(Harmful meme detection)**: very harmful, partially harmful, or harmless
  ➔ Problem 2 **(Target identification of harmful memes)**: individual, organization, community/country, or society/general public/others

- We develop a large-scale dataset, **HarMeme**, containing 3,544 real memes related to COVID-19.

Our Contributions:

- We formally define the notion of harmful memes in contrast to hateful and offensive memes.
- We formulate two novel problems
  - ➔ Problem 1 **(Harmful meme detection)**: very harmful, partially harmful, or harmless
  - ➔ Problem 2 **(Target identification of harmful memes)**: individual, organization, community/country, or society/general public/others
- We develop a novel dataset, **HarMeme**, containing 3,544 real memes related to COVID-19.
- We experiment with ten state-of-the-art unimodal and multimodal models.

Detecting Harmful Memes and Their Targets

Data Collection & Annotation:

- Collection: Google Image, Instagram, Facebook
- Deduplication
- Annotation Guidelines
- Annotation Process
    - ➔ Dry run
    - ➔ Final annotation
    - ➔ Consolidation

Baselines:

- **Unimodal Models**
  - Text Only
    - TextBERT
  - Image Only
    - VGG19
    - DenseNet
    - ResNet
    - ResNeXt
- **Multimodal Models (Image + Text)**
  - Unimodal Pre-training
    - Late Fusion
    - Concat BERT
    - MMBT
  - Multimodal Pre-training
    - ViLBERT CC
    - VisualBERT COCO

Access our dataset and implementation using this QR

The full dataset and the source code of the baseline models are available at http://github.com/di-dimitrov/harmeme

Evaluation:

Modality	Model	Harmful Meme Detection: 2-Class Classification Acc ↑	P ↑	R ↑	F1 ↑	MAE ↓	MMAE ↓	3-Class Classification Acc ↑	P ↑	R ↑	F1 ↑	MAE ↓	MMAE ↓
	Human†	90.68	84.35	84.19	83.55	0.1760	0.1723	86.10	67.35	65.84	65.10	0.2484	0.4857
	Majority	64.76	32.38	50.00	39.30	0.3524	0.5000	64.76	21.58	33.33	26.20	0.4125	1.0
Text Only	TextBERT	70.17	65.96	66.38	66.25	0.3173	0.2911	68.93	48.49	49.15	48.72	0.3250	0.5591
Image Only	VGG19	68.12	60.25	61.23	61.86	0.3204	0.3190	66.24	40.95	44.02	41.76	0.3198	0.6487
	DenseNet-161	68.42	61.08	62.10	62.54	0.3202	0.3125	65.21	41.88	44.25	42.15	0.3102	0.6326
	ResNet-152	68.74	61.86	62.89	62.97	0.3188	0.3114	65.29	41.95	44.32	43.02	0.3047	0.6264
	ResNeXt-101	69.79	62.32	63.26	63.68	0.3175	0.3029	66.55	42.62	44.87	43.68	0.3036	0.6499
Image + Text (Unimodal Pre-training)	Late Fusion	73.24	70.28	70.36	70.25	0.3167	0.2927	66.67	44.96	50.02	45.06	0.3850	0.6077
	Concat BERT	71.82	71.58	72.23	71.82	0.3033	0.3156	65.54	42.29	45.42	43.37	0.3881	0.5976
	MMBT	73.48	68.89	68.95	67.12	0.3101	0.3258	68.08	51.72	51.94	50.88	0.3403	0.6474
Image + Text (Multimodal Pre-training)	ViLBERT CC	78.53	78.62	**81.41**	78.06	0.2279	0.1881	**75.71**	48.89	49.21	48.82	**0.2763**	0.5329
	V-BERT COCO	**81.36**	**79.55**	81.19	**80.13**	**0.1972**	**0.1857**	74.01	**56.35**	**54.79**	**53.85**	0.3063	**0.5303**

Performance for harmful meme detection. For two-class classification, we merge very harmful and partially harmful into a single class. † This row reports the human accuracy on the test set.

- Multimodal systems consistently outperform unimodal ones.
- Sophisticated fusion techniques yield better results than simple concatenation.
- The best baseline is still far from human accuracy, indicating the potential for enriched multimodal models for meme analysis.

Evaluation:

Modality	Model	Target Identification of Harmful Memes					
		Acc ↑	**P ↑**	**R ↑**	**F1 ↑**	**MAE ↓**	**MMAE ↓**
	Human†	87.55	82.28	84.15	82.01	0.7866	0.3647
	Majority	46.60	11.65	25.00	15.89	1.2201	1.5000
Text (T) only	TextBERT	69.35	55.60	54.37	55.60	1.1612	0.8988
Image (I) only	VGG19	63.48	53.85	54.02	53.60	1.1687	1.0549
	DenseNet-161	64.52	53.96	53.95	53.51	1.1655	1.0065
	ResNet-152	65.75	54.25	54.13	53.78	1.1628	1.0459
	ResNeXt-101	65.82	54.47	54.20	53.95	1.1616	0.9277
I + T (Unimmodal Pre-training)	Late Fusion	72.58	58.43	58.83	58.43	1.1476	0.6318
	Concat BERT	67.74	54.79	49.65	49.77	1.1377	0.8879
	MMBT	72.58	58.43	58.83	58.35	1.1476	0.6318
I + T (Multimodal Pre-training)	ViLBERT CC	72.58	59.92	55.78	57.17	1.1671	0.8035
	V-BERT COCO	**75.81**	**66.29**	**69.09**	**65.77**	**1.1078**	**0.5036**

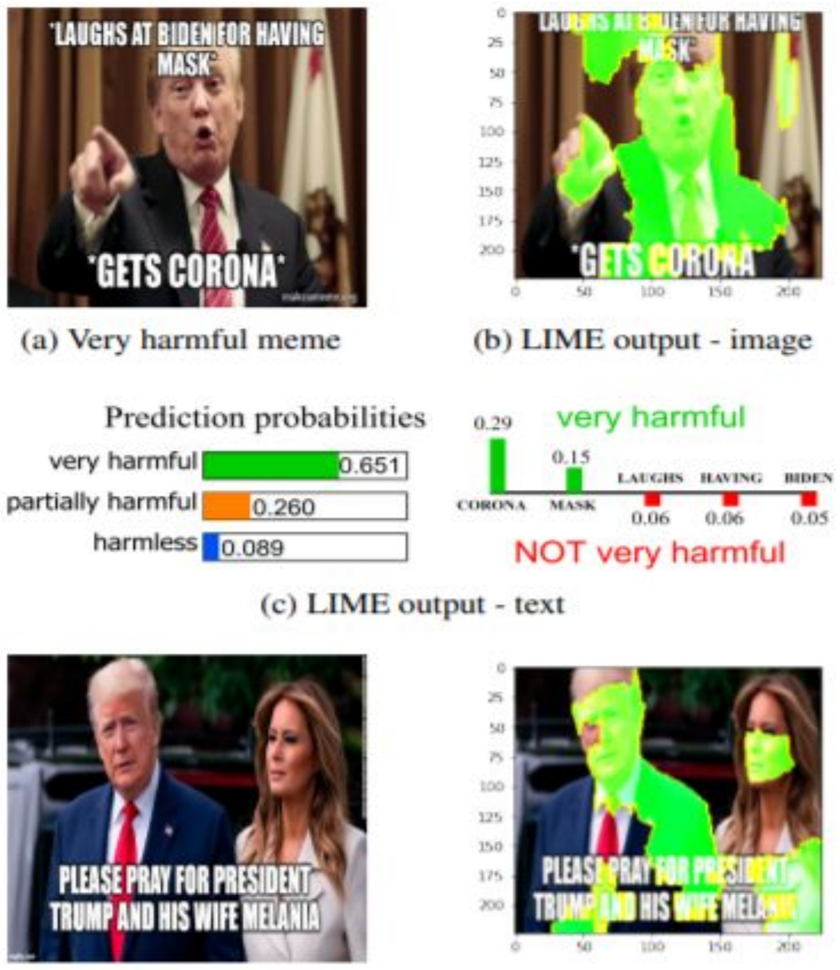
- Similarly for target identification, multimodal systems perform well.

Detecting Harmful Memes and Their Targets

Evaluation:

Modality	Model	Target Identification of Harmful Memes					
		Acc ↑	P ↑	R ↑	F1 ↑	MAE ↓	MMAE ↓
	Human†	87.55	82.28	84.15	82.01	0.7866	0.3647
	Majority	46.60	11.65	25.00	15.89	1.2201	1.5000
Text (T) only	TextBERT	69.35	55.60	54.37	55.60	1.1612	0.8988
Image (I) only	VGG19	63.48	53.85	54.02	53.60	1.1687	1.0549
	DenseNet-161	64.52	53.96	53.95	53.51	1.1655	1.0065
	ResNet-152	65.75	54.25	54.13	53.78	1.1628	1.0459
	ResNeXt-101	65.82	54.47	54.20	53.95	1.1616	0.9277
I + T (Unimmodal Pre-training)	Late Fusion	72.58	58.43	58.83	58.43	1.1476	0.6318
	Concat BERT	67.74	54.79	49.65	49.77	1.1377	0.8879
	MMBT	72.58	58.43	58.83	58.35	1.1476	0.6318
I + T (Multimodal Pre-training)	ViLBERT CC	72.58	59.92	55.78	57.17	1.1671	0.8035
	V-BERT COCO	**75.81**	**66.29**	**69.09**	**65.77**	**1.1078**	**0.5036**

- Similarly for target identification, multimodal systems perform well.
- Interpretability analysis shows the presence of bias even in the best baseline system.



(a) Very harmful meme (b) LIME output - image
(c) LIME output - text
(d) Harmless meme (e) LIME output - image

Conclusion:

- In this work, we formally define the notion of **harmful mems** which is much broader than hate and offense.

- We present **HarMeme**, the first large-scale benchmark dataset for the **detection of harmful memes** and **identification of their targets**.

- Our analysis shows that **off-the-shelf multimodal systems** are **not adequate** to understand the underlying semantics of harmful memes.

- In future work, we plan to design new multimodal models for meme analysis and extend HarMeme with more examples.

Thank You!