Towards a Multi-agent Sytem for Online Hate Speech Detection

Gaurav Sahu, Robin Cohen, Olga Vechtomova
{**gaurav.sahu**, rcohen, ovechtomova}@uwaterloo.ca
David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada

Objectives

- Detect hate speech in online social platforms from multimodal posts
- Propose a multi-agent system to moderate hateful content online

Introduction

To tackle hate speech online, a system should:

- have intelligent agents for each modality working together *and*
- ultimately recommend a human user what actions to take in the real world

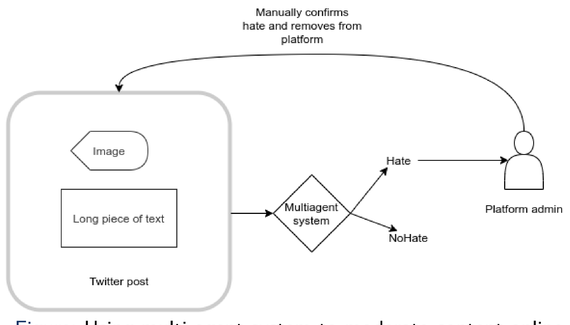

Figure: Using multi-agent system to moderate content online

Approach

Proposed an end-to-end pipeline for hate speech detection using adaptive fusion methods:

- **Auto-Fusion:** Train an autoencoder model to capture intermodal dynamics by maximizing correlation between multimodal inputs.
- **GAN-Fusion:** Train adversarial networks to align cross-modal embeddings.

Given a social media post with text, and visual input (p_t, p_v), we first obtain their respective latent feature vectors z_t, z_v. Next, we obtain z_{fuse} using a fusion mechanism. Finally, we use z_{fuse} to **classify** the post as hateful or not.

Proposed System

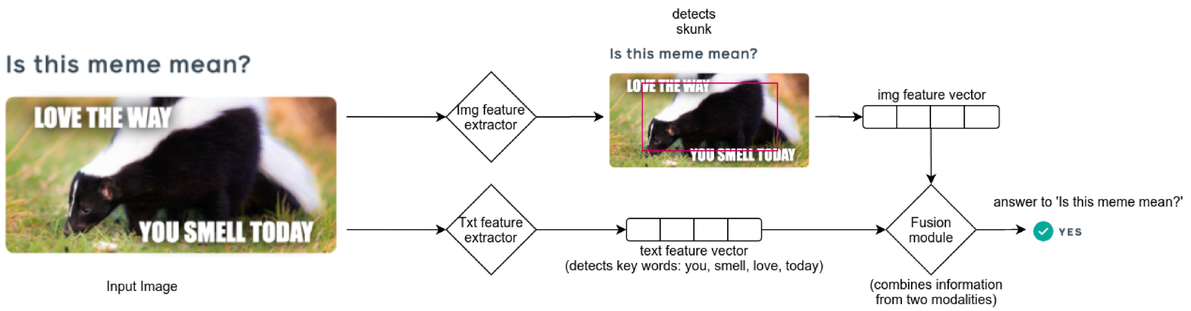

Figure: End-to-End pipeline of the proposed multi-agent system for hate speech classification.

Important Result

- Introducing input from multiple modalities enhances the efficacy of hate speech detection.
- Cross-modal adversarial alignment (GAN-fusion) best models context in a multi-modal sample.

Results

Model	Classes	Input modes	Fusion type	P	R	F
BiL	binary	text	none	70.08	63.31	66.52
TKM	binary	image+text+caption	Concat	-	-	70.1
SCM	binary	image+text+caption	Concat	-	-	70.2
FCM	binary	image+text+caption	Concat	-	-	70.4
BiL	multi	text	none	45.18	33.4	38.41
BiL	multi	text+caption	none	45.38	33.67	38.67
VBiL	multi	image+text+caption	Concat	55.27	35.54	43.04
VBiL	multi	image+text+caption	Auto-Fusion	59.65	43.87	50.56
VBiL	multi	image+text+caption	GAN-Fusion	**61.33**	**51.34**	**55.89**

Table: Precision (P), Recall (R), F1-score (F) for multimodal hate speech detection on MMHS150K [1].

Model	Input modes	Fusion type	P	R	F	A
E1	audio	none	57.3	57.3	57.3	56.6
E2	text	none	71.4	63.2	67.1	64.9
BiL1	text	none	53.2	40.6	43.4	43.6
BiL2	audio+text	Concat	66.1	65.0	65.5	64.2
E3	audio+text	Concat	72.9	71.5	72.2	70.1
MDRE	audio+text	Concat	-	-	-	71.8
MHA-2	audio+text	Concat	-	-	-	76.5
M1	audio+text	Auto-Fusion	75.3	77.4	76.3	77.8
M2	audio+text	GAN-Fusion	**77.3**	**79.1**	**78.2**	**79.2**

Table: Precision (P), Recall (R), F1-score (F), and Accuracy (A) for emotion recognition on the IEMOCAP dataset. The results confirm general value of the multimodal architecture.

Loss Functions

- **Auto-Fusion:** The MSE loss is given by:

$$J_{tr} = || \hat{\boldsymbol{z}}_{\boldsymbol{m}}^{\boldsymbol{k}} - \boldsymbol{z}_{\boldsymbol{m}}^{\boldsymbol{k}} ||^2 \quad (1)$$

- **GAN-Fusion:** Overall adversarial loss: $J_{adv} = J_{adv}^{t} + J_{adv}^{v}$ where,

$$\min_G \max_D J_{adv}^{i}(D, G) = \mathbb{E}_{x \sim p_{z_j}(x)}[\log D(x)] + \mathbb{E}_{\boldsymbol{z} \sim p_{z_i}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \quad (2)$$

where $i \neq j \ \forall \ i, j \in \{t, v\}$.

Conclusion & Future Work

- We sketch a solution to the problem of hate speech detection through a multi-agent system.
- We demonstrate the effectiveness of adaptive fusion techniques for coordinating text processing and image processing.
- The proposed system can be used by a human user to moderate content online.
- It aligns with previous work [2] which proposes intelligent recommendation systems as social good for the management of social networks.
- We plan to embed external knowledge base to enhance context capturing ability of the system.

References

[1] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas.
Exploring hate speech detection in multimodal publications.
In *IEEE/CVF WACV*, pages 1470–1478, 2020.

[2] Amulya Yadav, Hau Chan, Albert Xin Jiang, Haifeng Xu, Eric Rice, and Milind Tambe.
Using social networks to aid homeless shelters: Dynamic influence maximization under uncertainty.
In *AAMAS*, volume 16, pages 740–748, 2016.

UNIVERSITY OF WATERLOO