Yale

Multi-Modal Content Moderation Systems with Dialogue Summarization and Argument Graphs

Faiaz S. Rahman ADVISED BY Dr. Dragomir Radev Department of Computer Science, Yale University

LILY Lab

Introduction

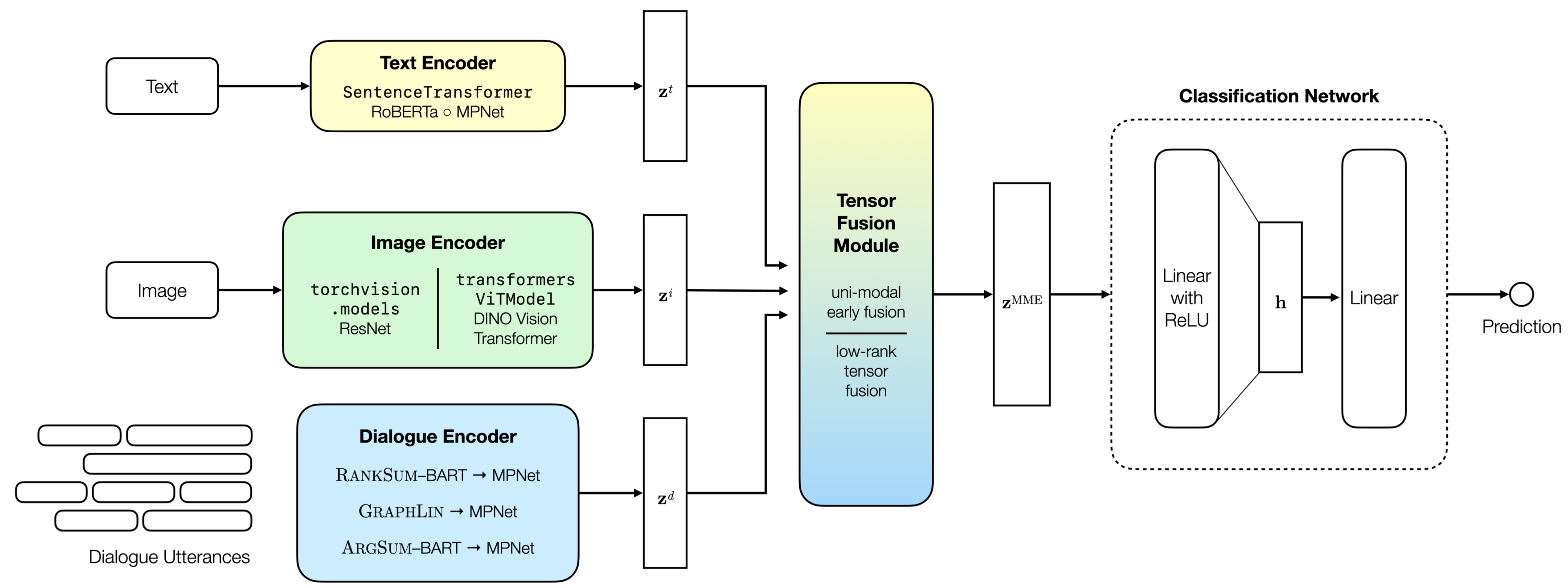
The proliferation of misinformation and hate speech online has created an era of digital disinformation, public mistrust, and even violence, particularly on social media platforms where users can engage in dialogue with such content. Fake news and hate speech exist not only in text form, but also include any accompanying images and video with the original post. Multi-modal models (i.e., those incorporating multiple modalities of data like text and images) offer a powerful approach in detecting such content. Prior work has both developed hate and misinformation datasets for experimentation, and examined different multi-modal representations in general, particularly for text–image data.

Given that user dialogue (e.g., comment threads, Tweet replies, etc.) can often give more insight into the integrity or hatefulness of a post (e.g., by indicating how extreme of a response was garnered, by introducing viewpoints beyond those of the original author, etc.), we investigate methods for modeling and incorporating the dialogue modality into multi-modal models. Specifically, we (1) develop **multi-modal models** for content moderation tasks (for the modalities of text, image, and dialogue), (2) improve dialogue modeling within those models by introducing **ArgSum**, an argument graph–based approach to dialogue summarization, and (3) improve the modeling of cross-modal interactions through the **multi-modal fusion** methods of uni-modal early fusion and low-rank tensor fusion.

Content Moderation Methods

Datasets For our content moderation tasks, we use Fakeddit for fake news detection and MMHS150K for hate speech detection.

Models Our multi-modal models consist of encoders for each data modality, a tensor fusion module to fuse the uni-modal embeddings, and a classification network for the content moderation task. For encoders, we experiment with RoBERTa and MPNet for text, and ResNet and DINO Vision Transformers for images; we initially use ranked dialogue summarization (RankSum) leveraging utterance metadata for dialogues.



The ArgSum Algorithm

Argument Graph Construction We segment each utterance its argumentative units, classify each unit as a claim, premise, or non-argumentative unit (using a RoBERTa-based classifier trained on AMPERSAND and Stab & Gurevych data), and create a node for each. We use a BERT-based entailment model trained on MNLI data for relationship type classification between nodes; we run premise-to-claim entailment to create subtrees of depth 1, then run claim-to-claim entailment to link related claims, greedily adding edges based on entailment scores without creating cycles. The major claims and zero-degree premises are linked to root.

Graph Linearization We use four heuristics, applied depth-first, to linearize the argument graph: greedy claim placement, semantic ordering, subtree size prioritization, and zero-degree premise tailing. The result is run through a BART summarization pipeline to produce the summary.

Multi-Modal Fusion Methods

Uni-Modal Early Fusion We apply tensor concatenation to the uni-modal tensors and embed the result.

▲ $\mathbf{z}^{\text{MME}} = \text{ReLU}\left(W\left([\mathbf{z}^a, \mathbf{z}^b, \mathbf{z}^c]\right) + b\right)$ ▼ $\mathbf{z}^{\text{MME}} = \begin{bmatrix} \mathbf{z}^t \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{z}^i \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{z}^d \\ 1 \end{bmatrix}$

Low-Rank Tensor Fusion We use a differentiable outer product to model cross-model interactions, first reducing the dimension for tri-modal settings for compute feasibility.

Results

Modality	Models	Results		
		2-way	**3-way**	**6-way**
Text	RoBERTa	0.7309	0.7406	0.6127
	MPNet	0.7853	0.7434	0.6189
Image	DINO ViT	0.7043	0.6863	0.6004
	ResNet	0.7260	0.7074	0.6086
Text + Image	RoBERTa + DINO ViT	0.7213	0.6972	0.6871
	MPNet + DINO ViT	0.7971	0.7559	0.6992
	RoBERTa + ResNet	0.8087	0.7816	0.7071
	MPNet + ResNet	0.8232	0.7844	0.7288
Text + Image + Dialogue	RoBERTa + DINO ViT + RankSum-BART	0.7902	0.7994	0.7422
	MPNet + DINO ViT + RankSum-BART	0.8475	0.8568	0.7550
	RoBERTa + ResNet + RankSum-BART	0.8837	0.8921	0.8259
	MPNet + ResNet + RankSum-BART	**0.9104**	**0.9036**	**0.8665**

Modality	Models	Results		
		2-way	**3-way**	**6-way**
Text + Image + Dialogue	MPNet + ResNet + RankSum-BART	0.9104	0.9036	0.8665
	MPNet + ResNet + GraphLin-MPNet*	0.9125	**0.9326**	0.9116
	MPNet + ResNet + ArgSum-BART	**0.9208**	0.9216	**0.9172**

Modality	Models	Uni-Modal Early Fusion			Low-Rank Tensor Fusion		
		2-way	**3-way**	**6-way**	**2-way**	**3-way**	**6-way**
Text + Image	MPNet + ResNet	0.8232	0.7844	0.7288	0.8325	0.8328	0.7473
Text + Image + Dialogue	MPNet + ResNet + RankSum-BART	0.9104	0.9036	0.8665	0.9248	0.9238	0.8988
	MPNet + ResNet + GraphLin-MPNet	0.9208	0.9216	0.9172	0.9301	0.9475	0.9227
	MPNet + ResNet + ArgSum-BART	0.9125	0.9326	0.9116	0.9312	0.9409	0.9298

Algorithm 1 Argument graph construction for ArgSum

Input: D, the set of dialogue utterances
Output: $G = (V, E)$, an argument graph

let $U \leftarrow$ [] ▷ Utterance segmentation
for $d \in D$ **do**
  $u \leftarrow$ `UtteranceToArgumentativeUnitSegmenter`(d)
  $U \leftarrow [...U, u]$
end for

Load trained AUC model and its tokenizer ▷ Argumentative unit classification
for batch of $u \in U$ **do**
  Pass batch of argumentative units through AUC model to get classification predictions
  Create an `ArgumentativeUnitNode` *for each argumentative unit*
end for

let $G \leftarrow$ *Instantiate an* `ArgumentGraph` *object* ▷ Relationship type classification

Load trained RTC model and its tokenizer ▷ (a) premise-to-claim entailment
for $p \in$ all premise nodes $P \subseteq U$ **do**
  Compute the entailment score for this premise against all claims (passing in batches to the RTC model to get the relationship type predictions) and get the claim c_{max} with the maximum score
  if the maximum score is above the minimum entailment score threshold **then**
    Create a support `RelationshipTypeEdge` *from the premise p to the claim c_{max}*
  end if
end for

First, store all potential claim-to-claim edges ▷ (b) claim-to-claim entailment
for $c \in$ all claims $C \subseteq U$ **do**
  let $C' \leftarrow C \setminus c$
  Compute the entailment score for this claim c against all other claims $c' \in C'$ (again passing in batches to the RTC model to get the relationship type predictions) and get the claim c'_{max} with the maximum score
  if the maximum score is above the minimum entailment score threshold **then**
    Store a potential support edge from c to c'_{max} (but do not add it to G yet)
  end if
end for

Next, greedily add edges from the stored potential edges in order of decreasing entailment score, only if it does not create a cycle in the graph

let root node $r \leftarrow$ `ArgumentativeUnitNode` with classification `ROOT` ▷ Root node linking
for $u \in U$ **do**
  if the node u does not entail any other nodes **then**
    Create an edge from the node to the root, with relationship type `TO_ROOT`
  end if
end for

Method	Results		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**
Baseline BART	31.65	11.93	28.32
ArgSum-BART	32.27	13.12	28.46

Conclusion

Our experiments find that (1) the incorporation of the dialogue modality in multi-modal models improves performance on fake news detection, (2) modeling argumentative structures in dialogues via ArgSum improves both summarization quality and multi-modal model performance, and (3) low-rank tensor fusion is able to better model cross-modal interactions than early fusion. Additionally, we release a public codebase including all of our PyTorch models, our ArgSum software package, and our experiment configuration files, built with an extensible design for future work on hate and misinformation detection.