Multi-Modal Discussion Transformer: Integrating Text, Images and Graph Transformers to Detect Hate Speech on Social Media

Liam Hebert, Gaurav Sahu, Yuxuan Guo, Nanda Kishore Sreenivas, Lukasz Golab, Robin Cohen

University of Waterloo
liam.hebert@uwaterloo.ca

Abstract

We present the Multi-Modal Discussion Transformer (mDT), a novel method for detecting hate speech on online social networks such as Reddit discussions. In contrast to traditional comment-only methods, our approach to labelling a comment as hate speech involves a holistic analysis of text and images grounded in the discussion context. This is done by leveraging graph transformers to capture the contextual relationships in the discussion surrounding a comment and grounding the interwoven fusion layers that combine text and image embeddings instead of processing modalities separately. To evaluate our work, we present a new dataset, HatefulDiscussions, comprising complete multi-modal discussions from multiple online communities on Reddit. We compare the performance of our model to baselines that only process individual comments and conduct extensive ablation studies.

Introduction

Social media have democratized public discourse, enabling users worldwide to freely express their opinions and thoughts. As of 2023, the social media giant Meta has reached 3 billion daily active users across its platforms (Meta 2023). While this level of connectivity and access to information is undeniably beneficial, it has also resulted in the alarming rise of hate speech (Das et al. 2020). This pervasive spread of hateful rhetoric has caused significant emotional harm to its targets (Vedeler, Olsen, and Eriksen 2019), triggered social divisions and polarization (Waller and Anderson 2021), and has caused substantial harm to the mental health of users (Wachs, Gámez-Guadix, and Wright 2022). There is an urgent need for a comprehensive solution to automate identifying hate speech as a critical first step toward combatting this alarming practice.

Initially, automated hate speech detection models were limited to text-only approaches such as HateXplain (Mathew et al. 2021), which classify the text of individual comments. Such methods have two significant weaknesses. First, social media comments have evolved to include images, which can influence the context of the accompanying text. For instance, a comment may be innocuous, but including an image may transform it into a hateful remark. Second, hate speech is contextual. Social media comments are often conversational and are influenced by other comments within the discussion thread. For example, a seemingly innocuous comment such as "That's gross!" can become hateful in a discussion about immigration or minority issues.

Ongoing research to address these weaknesses includes multi-modal transformers such as VilT (Kim, Son, and Kim 2021) that combine images and text for a richer representation of comments. Still, they do not account for the contextual nature of hate speech. Hebert, Golab, and Cohen (2022) model discussion context with graph neural networks, but they do not discuss how to integrate the interpretation of images within hateful social media discussions. Furthermore, the sequential nature of the proposed architecture prevents text embeddings from being grounded to other comments in a graph. The initial semantic content encoded by a comment embedding may differ when considered with different sets of comments versus in isolation.

To overcome the limitations of existing methods, we propose the Multi-Modal Discussion Transformer (mDT) to holistically encode comments with multi-model discussion context for hate speech detection. To evaluate our work, we also present a novel dataset, HatefulDiscussions, containing complete multi-modal discussion graphs from various Reddit communities and a diverse range of hateful behaviour.

We compare mDT against comment-only and graph methods and conduct an ablation study on the various components of our architecture. We then conclude by discussing the potential for our model to deliver social value in online contexts by effectively identifying and combating anti-social behaviour in online communities. We also propose future work towards more advanced multi-modal solutions that can better capture the nuanced nature of online behaviour.

To summarize our contributions: **1)** We propose a novel fusion mechanism as the core of mDT that interweaves multi-modal fusion layers with graph transformer layers, allowing for multi-modal comment representations actively grounded in the discussion context. **2)** We propose a novel graph structure encoding specific to the conversational structure of social media discussions. **3)** We introduce a dataset of 8266 annotated discussions, totalling 18359 labelled comments, with complete discussion trees and images to evaluate the effectiveness of mDT. Our work focuses on Reddit, which consists of branching tree discussions. Our codebase, datasets and further supplemental can be found at

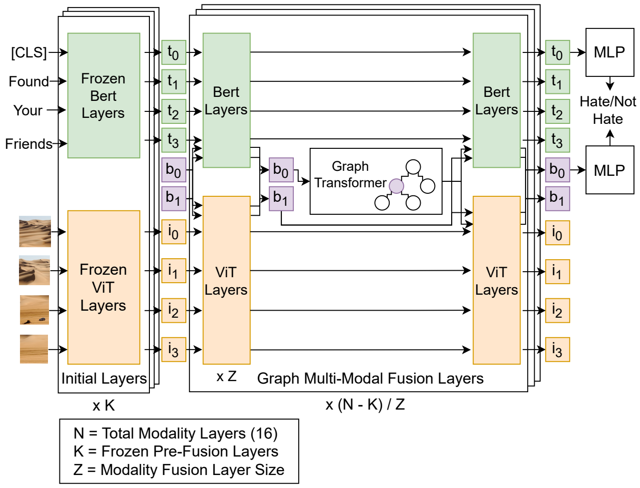

Figure 1: Multi-Modal Discussion Transformer

github.com/liamhebert/MultiModalDiscussionTransformer.

Related Work

Transformer-based encoding models such as BERT have significantly improved natural language processing due to their ability to capture textual semantics (Devlin et al. 2019). Inspired by these developments, methods such as HateXplain (Mathew et al. 2021) and HateBERT (Caselli et al. 2021) have been introduced to discern hateful comments on social platforms, focusing on text alone. The effectiveness of these efforts is intrinsically tied to the diversity of datasets they are trained on. For instance, HateXplain utilized a specialized dataset from diverse social platforms like Twitter and Gab, emphasizing interpretable hate speech detection. Other noteworthy datasets include Gong et al. (2021), studying heterogeneous hate speech (comments containing mixed abusive and non-abusive language), and Founta et al. (2018), which crowdsourced annotation of Twitter abusive content. Finally, Zampieri et al. (2019) collected hateful Twitter posts collated through a collaborative semi-supervised approach.

While text is essential, images also contribute to the semantic context. CLIP introduced an approach to align text and image representations via contrastive pre-training (Radford et al. 2021). ViLBERT (Lu et al. 2019) conceptualized distinct transformers for each modality—images and text, which are then amalgamated through co-attentional transformer layers. Subsequent works such as ViLT (Kim, Son, and Kim 2021) and Nagrani et al. (2021) have devised novel inter-modality fusion mechanisms, unifying both modality transformers into one. This integration of multi-modal language grounding has also enriched hate speech detection, as evidenced by the HatefulMemes challenge (Kiela et al. 2020). Additional works such as Liang et al. (2022) employ graph convolutional networks to merge text and images, primarily for sarcasm detection. Meanwhile, Sahu, Cohen, and Vechtomova (2021) leverage generative adversarial networks to encode these modalities, facilitating congruent representations of comments. Cao et al. (2022) pursue a unique strategy by mapping the paired image to text descriptors, appending the comment text, and then predicting with a generative language model. Finally, (Singh et al. 2022) incorporate image and text representations of product reviews to accurately disambiguate complaints.

Despite the progress, many of these techniques overlook a vital modality: the context of discussions. The prevailing emphasis remains on datasets and techniques that analyze singular comments, bypassing the contextual significance of the prior discussion. By extending Graphormer (Ying et al. 2022)—a graph transformer network tailored for molecular modelling—Hebert, Golab, and Cohen (2022) consolidate learned comment representations to predict the trajectory of hateful discussions. However, this work has limitations; it neglects the influence of images and, owing to the absence of complete discussion-focused hate speech datasets, resorts to approximating ground truth labels using a ready-made external classifier. Our work addresses both limitations, including interleaving comment and discussion layers and human ground truth data.

Methodology

Multi-Modal Discussion Transformer (mDT)

The mDT architecture consists of Initial Pre-Fusion, Modality Fusion, and Graph Transformer (Figure 1). The description below outlines the holistic nature of our solution.

Initial Pre-Fusion Given a discussion D with comments $c \in D$, each represented with text t_c and optional image i_c, we start with pre-trained BERT and ViT models to encode text and images, respectively. Both models contain N layers with the same hidden dimension of d. In our experiments, we utilized BERT-base and ViT-base, which have $N = 16$ layers and $d = 768$ hidden dimensions. Given these models, the Initial Pre-Fusion step consists of the first K layers of both models with gradients disabled (frozen), denoted as

$$t_c^k = BERT_{init}(t_c), i_c^k = ViT_{init}(i_c) \tag{1}$$

where $K < N$. This step encodes a foundational understanding of the images and text that make up each comment.

Modality Fusion After creating initial text and image embeddings t_c, i_c for all comments $c \in D$ in the discussion, we move to the Modality Fusion step. We adopt the bottleneck mechanism proposed by Nagrani et al. (2021) to encode inter-modality information. We concatenate b shared modality bottleneck tokens $B \in R_{b \times d}$ to t_c and i_c, transforming the input sequence to $[t_c^k \mid\mid B], [i_c^k \mid\mid B]$. We then define a modality fusion layer l as

$$[t_c^{l+1}||B_{t,c}^{l+1}] = BERT_l([t_c^l||B_c^l]) \tag{2}$$

$$[i_c^{l+1}||B_{i,c}^{l+1}] = ViT_l([i_c^l||B_c^l]) \tag{3}$$

$$B_c^{l+1} = Avg(B_{t,c}^{l+1}, B_{i,c}^{l+1}) \tag{4}$$

where both modalities only share information through the B bottleneck tokens. This design forces both modalities to compress information to a limited set of tokens, improving performance and efficiency. If no images are attached to a comment, then $B_c^{l+1} = B_t^{l+1}$.

Graph Transformer Then, after $Z\ (< (N-K))$ modality fusion layers, we deploy Graph Transformer layers to aggregate contextual information from the other comments in the discussion[1]. Given that the tokens in B_c encode rich intermodality information, we innovate by leveraging these representations to represent the nodes in our discussion graph. Using $b_c^0 \in B_c$ to represent each comment $c \in D$, we aggregate each embedding using a transformer model to incorporate discussion context from other comments. Our novel utilization of bottleneck tokens to represent graph nodes allows modality models to maintain a modality-specific pooler token ([CLS]) as well as a graph context representation (b_0).

Since transformer layers are position-independent, we include two learned structure encodings. The first is Centrality Encoding, denoted z, which encodes the degree of nodes in the graph (Ying et al. 2022). Since social media discussion graphs are directed, the degree of comments is equivalent to the number of replies a comment receives plus one for the parent node. We implement this mechanism as

$$h_c^{(0)} = b_c^0 + z_{deg(c)} \tag{5}$$

where $h_c^{(0)}$ is the initial embedding of b_c^0 in the graph and $z_{deg(c)}$ is a learned embedding corresponding to the degree $deg(c)$ of the comment.

The second structure encoding is Spatial Encoding, denoted $s_{(c,v)}$, which encodes the graph's structural relationship between two nodes, c and v. This encoding is added as an attention bias term during the self-attention mechanism. That is, we compute the self attention $A_{(c,v)}$ between nodes c, v as

$$A_{(c,v)} = \frac{(h_c \times W_Q)(h_v \times W_K)}{\sqrt{d}} + s_{(c,v)} \tag{6}$$

where W_Q and W_K are learned weight matrices and d is the hidden dimension of h.

In previous graph transformer networks, $s_{(c,v)}$ is encoded as a learned embedding representing the shortest distance between nodes c and v in the graph (Ying et al. 2022; Hebert, Golab, and Cohen 2022). However, this metric does not lend itself well to the hierarchical structure of discussions, where equivalent distances can represent different interactions. This is best seen in the example discussion illustrated in Figure 2. When utilizing the shortest distance to encode structure, the distance between nodes a and c is the same as the distance between nodes b and d in this graph. However, b and d represent direct replies to the same parent post, whereas a is two comments underneath c.

We propose a novel hierarchical spatial encoding based on Cantor's pairing function to account for this. Cantor's pairing function uniquely maps sets of two numbers into a single number $\mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$. Given comments a and b, we first calculate the number of hops upward $u_{(a,b)}$ and hops downward $d_{(a,b)}$ to reach b from a. In the example above, the distance

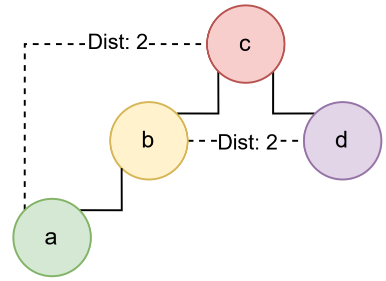

Figure 2: Example Discussion Structure. Each node in the discussion tree represents a comment. The shortest distance between (a, c) and (b, d) is equivalent, demonstrating a lack of expressiveness towards hierarchy.

between a and d is $u_{(a,b)} = 2, d_{(a,b)} = 1$. We then compress both numbers into a single index using the proposed position-independent variant of Cantor's pairing:

$$\begin{aligned} s_{(c,v)} &= s_{(v,c)} && (7)\\ &= Cantors(u, d) && (8)\\ &= \frac{(u+d)(u+d+1)}{2} + min(u, d) && (9) \end{aligned}$$

This uniquely maps $\mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$ such that $s_{c,v} = s_{v,c}$. We utilize this function to index learned spatial embeddings in the self-attention mechanism.

After G graph transformer layers, the final representation of h_c^G replaces b_c^0 for the next set of Z modality fusion layers. We denote the combination of Z Modality Fusion and G Graph Transformer layers as a Graph Multi-Modal Fusion module. Finally, after $(N-K)/Z$ Graph Multi-Modal Fusion modules, we predict logits using the final embedding of b_c^0 and the [CLS] embedding of t_c. This novel interweaving of graph transformer and fusion layers through modality bottleneck tokens ensures that fusion models create representations grounded in the discussion context. Notably, this differs from previous approaches that utilize graph neural networks, which sequentially process individual comments before applying a set of graph layers.

HatefulDiscussions Dataset

To train our model, we require a complete multi-modal discussion graph dataset. However, datasets used by other works (Mathew et al. 2021; Zampieri et al. 2019; Kiela et al. 2020) consist of individual labelled comments and are predominately text-only. To address this issue, we curated a novel benchmark comprising multiple datasets that used human annotators, which we augmented to include complete multi-modal discussion graphs. Our final dataset comprises 8266 Reddit discussions with 18359 labelled comments from 850 communities. Note that our architecture can extend to other platforms, such as Facebook, Twitter and TikTok, as they also discuss in tree structures. Since discussions on these platforms are typically smaller with less complexity, using Reddit allows us to best stress-test our model.

The first type of hate speech included in our benchmark is Identity-Directed and Affiliation-Directed Abuse. We re-

[1]Our implementation can handle discussions up to 516 comments, as we only require a single graph transformer pass to evaluate all comments. The above limit can be exceeded via efficient attention mechanisms such as sparse or flash attention.

Label	Count
Derogatory Slur	4297
Not Derogatory Slur (NDG)	2401
Homonym (HOM)	364
LTI Person Directed Neutral	4116
LTI Person Directed Hate	1313
CAD Affiliation Neutral	4892
CAD Identity Directed Hate	701
CAD Affiliation Directed Hate	275
Neutral	11773
Hateful	6586

Table 1: Label Distribution of Hateful Discussions

trieved labelled examples of this from the Contextual Abuse Dataset (CAD) developed by Vidgen et al. (2021a). According to the authors, Identity-Directed abuse refers to content containing negative statements against a social category, encompassing fundamental aspects of individuals' community and socio-demographics, such as religion, race, ethnicity, and sexuality, among others. On the other hand, Affiliation-Directed abuse is defined as content expressing negativity toward an affiliation, which is described as a voluntary association with a collective, such as political affiliation and occupations (Vidgen et al. 2021a). We selected both of these forms of abuse from CAD due to the similarity in their definitions—abuse that is directed at aspects of a person's identity rather than a specific individual directly.

Next, slurs form the second type of hateful content within our dataset, sampled from the Slurs corpus (Kurrek, Saleem, and Ruths 2020). Notably, historically derogatory slurs can undergo re-appropriation by specific communities, such as the n-slur in African American Vernacular, transforming them into non-derogatory terms. Therefore, we hypothesize that understanding the contextual nuances surrounding the use of slurs becomes essential in distinguishing between non-derogatory and derogatory instances.

The last type of hateful content we include is person-directed abuse, hate speech or offensive content that specifically targets and attacks an individual or a group of individuals. We source labelled examples from the Learning to Intervene (LTI) dataset by Qian et al. (2019) to include examples of this abuse requiring context.

For each labelled comment, we retrieved the corresponding complete discussion tree using the Pushshift Reddit API and downloaded all associated images[2]. To refine our dataset, we filtered out conversations without images and constrained comments to have a maximum degree of three and conversations to have a maximum depth of five. By trimming the size of the discussion tree, we reduce computational complexity and focus on the most relevant parts of the conversation (Parmentier et al. 2021). We map each retrieved label to Hateful or Normal and treat the problem as a binary classification. The distribution of each label can be seen in Table 1.

[2]At the time of writing, Reddit has suspended access to the Pushshift API; however, our dataset remains complete.

Hyperparameter	Value
Pre-Fusion Layers (K)	4 - [0, 2, 4, 6]
Modality Fusion Layer Stack (Z)	2 (8 total) - [1, 2, 4]
Graph Layer Stack (G)	2 (8 total) - [1, 2, 4]
Bottleneck Size (B)	4 - [1, 4, 8, 16, 32]
Max Spatial Attention	5 - [2, 5, 4096 (inf)]
Attention Dropout	0.3 - [0, 0.1, 0.3, 0.5]
Activation Dropout	0.3 - [0, 0.1, 0.3, 0.5]
Graph Dropout	0.4 - [0, 0.2, 0.4, 0.6]
Hidden Dimension (d)	768
Graph Attention Heads	12 - [4, 6, 12, 24]
Modality Attention Heads	12
Optimizer	Adam
Batch Size	48
Epochs	10 w/ Early Stopping
Learning Rate	$3e^{-5} \rightarrow 3e^{-7}$
Learning Rate Scheduler	Polynomial Decay
Warm up Updates	500
Total Updates	3350
Positive Class Weighting	1.5 - [1, 1.5]
Negative Class Weight	1
Freeze Initial Encoders	Yes - [Yes, No]

Table 2: mDT Model Hyperparameters. The search space for each parameter is denoted by [...]

Most images in HatefulDiscussions, such as the root post, appear in the discussion context rather than directly attached to labelled comments. In our case, only 424 labelled instances have an image attached. Still, all 8000 discussions have an image in the prior context. Therefore, the challenge becomes how to interpret and incorporate multi-modal discussion context to disambiguate the meaning of comments that may not contain those modalities.

Results

Experimental Setup

We conduct a 7-fold stratified cross-validation with a fixed seed (1) and report the average performance for each model. We report overall accuracy (Acc.) and class-weighted Precision (Pre.), Recall (Rec.) and F1 to account for label imbalance. Underlined values denote statistical significance using Student's t-test with p-value < 0.05 in all results. Our hyperparameter search space can be seen in Table 2. All experiments were run using 2xA100-40GB GPUs, a 12-core Intel CPU, 80GB of RAM, and Linux.

Text-only Methods vs. Discussion Transformers

To assess the performance of mDT, we compared it against several state-of-the-art hate speech detection methods. For comment-only approaches, we evaluated BERT-HateXplain

Method	Acc.	Pre.	Rec.	F1
BERT-HateXplain	0.742	0.763	0.742	0.747
Detoxify	0.687	0.679	0.696	0.677
RoBertA Dynabench	0.811	0.822	0.811	0.814
BERT-HatefulDiscuss	0.858	0.858	0.858	0.858
Graphormer	0.735	0.594	0.759	0.667
mDT (ours)	**0.880**	**0.880**	**0.880**	**0.877**

Table 3: Performance of mDT against Text-Only Methods

Bottleneck Size	Acc.	Pre.	Rec.	F1
4	**0.880**	**0.880**	**0.880**	**0.877**
8	0.863	0.864	0.863	0.863
16	0.864	0.850	0.853	0.852
32	0.874	0.872	0.874	0.872

Table 4: Effect of Bottleneck Size on mDT Performance

(Mathew et al. 2021), Detoxify (Hanu and Unitary team 2020), and RoBertA Dynabench (Vidgen et al. 2021b). We also compared mDT against a BERT model trained on the training set of HatefulDiscussions, referred to as BERT-HatefulDiscuss. To compare against previous graph-based approaches, we evaluated the text-only Graphormer model proposed by (Hebert, Golab, and Cohen 2022).

Our results (Table 3) show that mDT outperforms all evaluated methods across all metrics. Specifically, mDT achieves 14.5% higher accuracy and 21% higher F1 score than Graphormer. This indicates that our approach to including graph context significantly improved over the previous approach incorporating this modality. Although the performance gap between BERT-HatefulDiscussions and mDT is narrower, we still perform better against all text-only methods. We observed F1 score improvements of 20%, 13%, and 6.3% over Detoxify, BERT-HateXplain, and RoBertA Dynabench, respectively.

Effect of Bottleneck Size

Next, we investigated the impact of increasing the number of bottleneck interaction tokens (B) in mDT, added during the modality fusion step. Adding more bottleneck tokens reduces the amount of compression required by the BERT and ViT models to exchange information. Table 4 presents the results, where we find that using four bottleneck tokens leads to the best performance. We also observe a slight drop in performance when we increase the number of bottleneck tokens beyond four, indicating the importance of compression when exchanging modality encodings between models.

Effect of Constrained Graph Attention

A recent study by Hebert et al. explored the limitations of graph transformers for hate speech prediction, finding that discussion context can sometimes mislead graph models into making incorrect predictions (Hebert et al. 2023). In light of this, we explore the impact of constraining the attention mechanism of our graph transformer network to only attend to nodes within a maximum number of hops away from a source node. We report the results in Table 5 and find that constraining the attention window to 5 hops achieves better performance. However, we also observed that performance gains from the 5-hop constraint were lost when we further constrained the attention to only two hops. Our findings suggest a balance is required when constraining graph attention for optimal performance.

Attention Window	Acc.	Pre.	Rec.	F1
2	0.866	0.866	0.866	0.866
5	**0.880**	**0.880**	**0.880**	**0.877**
∞	0.870	0.861	0.850	0.855

Table 5: Effect of Constraining Graph Attention

Fusion Layers	Acc.	Pre.	Rec.	F1
6	0.868	0.856	0.854	0.855
8	0.872	0.871	0.844	0.855
10	0.866	0.867	0.866	0.862
12	**0.880**	**0.880**	**0.880**	**0.877**

Table 6: Effect of Fusion Layers

Effect of Fusion Layers

Next, we investigate the effect of increasing the number of Multi-Modal Fusion Layers (Z) in our mDT model. To ensure full utilization of the 16 available layers, any unused layers were allocated to the Initial Pre-Fusion step (K). Our results in Table 6 indicate that utilizing 12 fusion layers leads to the best performance. Interestingly, the performance gains did not follow a linear trend with the number of fusion layers. Specifically, we observed that eight fusion layers outperformed ten but were still inferior to 12. Further research should explore the potential benefits of scaling beyond 12 fusion layers using larger modality models.

Effect of Images

We also investigated the impact of removing images in mDT. Our findings (Table 7) support the hypothesis that images provide crucial contextual information for detecting hateful content: excluding images led to a 4.8% decrease in accuracy and a 4.9% decrease in the F1 score. It is worth noting that even without images, mDT outperformed Graphormer (Table 3), indicating that our approach provides substantial gains over previous graph-based methods for hate speech detection beyond just including images. The results of this experiment underscore the importance of considering multiple

Usage of Images	Acc.	Pre.	Rec.	F1
With Images	**0.880**	**0.880**	**0.880**	**0.877**
Without Images	0.832	0.835	0.822	0.828

Table 7: Effect of Excluding Images

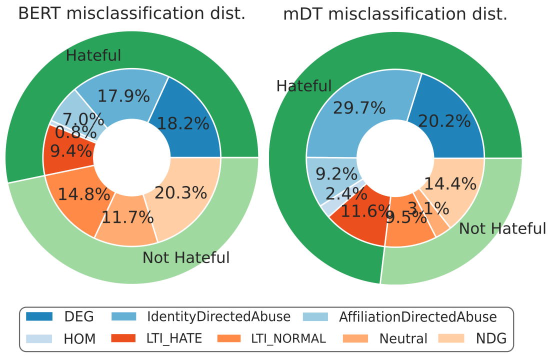

Figure 3: Fine-grained distribution of BERT and mDT misclassification. (Acronyms above as in Table 1)

Figure 4: An image present in the discussion context of example 3 (Table 8), seen only by mDT, contextualizing comments as potentially hateful

modalities for hate speech detection and suggest that future research should explore further improvements by leveraging additional types of contextual information.

Qualitative Analysis: BERT vs. mDT

We next perform a qualitative comparison of the text-only BERT model and the proposed mDT architecture. We find that the text-only BERT model misclassifies 385/2717 test instances. Upon passing those test instances through mDT, we found that it corrected BERT's labels in 161/385 cases. We further note that BERT and mDT predictions disagree on 264 test instances, which mDT is correct on 161 (61%). Figure 3 shows a fine-grained distribution of misclassified test examples by class. Using mDT results in an overall decrease in misclassifications (385 $\rightarrow$ 327), with a significant reduction in false positives (fewer misclassifications for the 'Not Hateful' class). However, we notice that BERT and mDT both struggle to detect the presence of hate speech in derogatory slur (DEG) and identity-directed (IdentityDirectedAbuse) comments.

Table 8 shows some hateful test instances misclassified by the two models. We note that the main text under consideration (an individual comment) may not exhibit hate speech on its own; however, considering it with the context (rest of the discussion thread+image) helps mDT correctly classify the test instances as hate speech. Consider the second example in Table 8. The word "tranny" is a common acronym for "transmission" on social media, but considering the context, it is an abusive discussion directed toward the transgender community. This is further contextualized by the accompanying image in the discussion, leading to evidence of hateful interpretations (Figure 4). Images within the discussions for each example provide similar contextual evidence, such as the third example and Figure 5.

We also found some intriguing test examples where adding context proved misleading for the model, while BERT confidently classified the main text as hateful. For instance, in the last example in Table 8, comments in the context are largely non-abusive, misinterpreting the primary text as non-abusive. This suggests that while adding context results in a net decrease in misclassifications, majorly neutral context might also fool the model. This is likely since we emphasize the discussion context when we obtain the final classification by averaging the text embedding logit with the discussion node embedding (b_c^0).

Future Work

While we find mDT to be an effective method for analyzing discussions on social media, we have pointed out how it is challenged when the discussion context contains predominately neutral comments. To address this, we propose using a first-stage text ranker to compute semantic relevance between comments to filter unrelated messages.

We also note that many contextual signals in social media discussions remain untapped beyond text, images, and discussion structure. Incorporating named entity recognition techniques to integrate deeper analysis of real-world knowledge would be well-supported by the contextual nature of mDT(Mishra et al. 2022).

Perhaps the most exciting step forward would be to expand our analysis of individual communities toward learning indicators of their propensity for hateful conduct. We are interested in capturing the culture of specific platforms containing diverse communities, including marginalized communities, which exchange unique reclaimed vernacular that should not be misinterpreted as hate. In addition to the example given earlier of the African American community, there are particular usages as well that arise among platforms supporting LGBTQ users. The contextual nature of mDT captured by graph transformers provides much promise for advancing these extensions.

Finally, the versatility of mDT's core mechanisms makes it a promising tool for many applications beyond hate speech detection. This approach could be applied to other domains such as online product reviews (Jagdale, Shirsat, and Deshmukh 2019), political discourse analysis (Lyu and Luo 2022), and popularity analysis (Chen et al. 2022; Tan et al. 2022), where understanding the discussion context is critical for accurate interpretation.

Conclusion

This paper presented mDT: a holistic approach to detecting hate speech in social media. Our model leverages graphs, text and images to reason about entire discussion threads.

Primary Text	**Context (only seen by mDT)**	BERT pred.	mDT pred.
Now imagine if virtuous keyboard sjws had their way? Their mascot should be Ralph Wiggum.	**[. . .]** Preferred pronouns: go/*f-slur*/yourself **[. . .]** If the Chinese in my corner of NZ only sold to Chinese they'd starve by Thursday. **[. . .]** They just wanna *b-slur* about something because their own life sucks.	Not Hateful	Hateful
whoa brah.. leave my tranny out of this	[image] (Figure 4) **[. . .]** that's *f-slur* retarded **[. . .]** Just spit my drink **[. . .]**	Not Hateful	Hateful
"That *n-slur* was on PCP Johnson" Lmao	[image] (Figure 5) **[. . .]** Its' a common pattern when dealing with these shootings. * Kill black dude **[. . .]** * Wingnut welfare kicks in as racist *f-slur* create gofundme of over half a million *f-slur* dollars for cops family **[. . .]**	Not Hateful	Hateful
uwu owo uwu	**[. . .]** That is not even close to what feminism is. What you are talking about is radical Feminism **[. . .]** Got banned from my sexual minority subreddit (r/bisexual) for not believing that all bisexuals should actually be pansexuals **[. . .]**	Hateful	Not Hateful

Table 8: Text instances misclassified by BERT and mDT. Note: The ground truth for all the examples shown here is "Hateful". We have also redacted chunks of text from the context in the interest of space. The redacted content is shown by [. . .].

Core to our approach is the introduction of hierarchical spatial encodings and coupling of text, image, and graph transformers through a novel bottleneck mechanism. We also presented a new dataset of complete multi-modal discussions containing a wide spectrum of hateful content, enabling future work into robust graph-based solutions for hate speech detection[3].

One significant contribution is demonstrating how discussion-oriented multi-modal analysis can improve the detection of anti-social behaviour online. Compared with several key competitors, our experimental results demonstrate the quantitative improvements stemming from our method. Notably, we see a 21% improvement in F1 over previous methods to include discussion context, such as (Hebert, Golab, and Cohen 2022). Furthermore, our initial qualitative analysis demonstrates the valuable impact of our holistic multi-modal approach.

Beyond enhanced holistic discussion analysis, our work enables a rich understanding of conversational dynamics, enabling community-centric prediction. This is primarily powered by our novel improvements to graph transformers, a method gaining momentum in AI molecular modelling, revealing their potential to capture the relationships in complex multi-modal discussions. We hypothesize that this expressiveness in capturing context can aid in disambiguating false positives, preventing further marginalization of communities, and proactively mitigating hateful behaviours.

Overall, our approach presents a path forward for addressing the issue of hate speech on social media and encourages the exploration of holistic graph-based multi-modal models to interpret online discussions. We believe our research can help foster healthier and more inclusive environments, improving mental health for individuals online.

[3]We leave additional comparisons with image-text multi-modal methods to future work to resolve fairness considerations; these competitors require comments to be multi-modal, whereas we can notice relevant discussion responses with images.

Figure 5: An image present in discussion context of example 2 (Table 8), seen only by mDT, contextualizing comments as potentially hateful

Ethics Statement

Our data collection efforts to create HatefulDiscussions are consistent with Reddit's Terms of Service and with approval. We removed all user-identifying features and metadata from all posts and images to ensure privacy and refrained from contacting users. To best comply with their terms of service, we will be open-sourcing our dataset and pre-trained models under the CC-BY-NC 4.0 license, with the model code under the permissive MIT license.

Acknowledgements

The authors thank the Natural Sciences and Engineering Research Council of Canada, the Canada Research Chairs Program and the University of Waterloo Cheriton Scholarship for financial support. We are also grateful to the reviewers for their valued feedback on the paper.

References

Cao, R.; Lee, R. K.-W.; Chong, W.-H.; and Jiang, J. 2022. Prompting for Multimodal Hateful Meme Classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 321–332. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Caselli, T.; Basile, V.; Mitrović, J.; and Granitzer, M. 2021. HateBERT: Retraining BERT for Abusive Language Detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, 17–25. Online: Association for Computational Linguistics.

Chen, W.; Huang, C.; Yuan, W.; Chen, X.; Hu, W.; Zhang, X.; and Zhang, Y. 2022. Title-and-Tag Contrastive Vision-and-Language Transformer for Social Media Popularity Prediction. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 7008–7012. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Das, M.; Mathew, B.; Saha, P.; Goyal, P.; and Mukherjee, A. 2020. Hate Speech in Online Social Media. *SIGWEB Newsl.*, (Autumn).

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Founta, A.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Gong, H.; Valido, A.; Ingram, K. M.; Fanti, G.; Bhat, S.; and Espelage, D. L. 2021. Abusive language detection in heterogeneous contexts: Dataset collection and the role of supervised attention. In *Proceedings of the AAAI Conference on Artificial Intelligence (AISI Track)*, volume 35, 14804–14812.

Hanu, L.; and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Hebert, L.; Chen, H. Y.; Cohen, R.; and Golab, L. 2023. Qualitative Analysis of a Graph Transformer Approach to Addressing Hate Speech: Adapting to Dynamically Changing Content. *Workshop in Artificial Intelligence for Social Good (AISG) at AAAI.*

Hebert, L.; Golab, L.; and Cohen, R. 2022. Predicting Hateful Discussions on Reddit using Graph Transformer Networks and Communal Context. In *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 9–17.

Jagdale, R. S.; Shirsat, V. S.; and Deshmukh, S. N. 2019. Sentiment analysis on product reviews using machine learning techniques. In *Cognitive Informatics and Soft Computing: Proceeding of CISC 2017*, 639–647. Springer.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 5583–5594. PMLR. ISSN: 2640-3498.

Kurrek, J.; Saleem, H. M.; and Ruths, D. 2020. Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for Online Slur Usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 138–149. Online: Association for Computational Linguistics.

Liang, B.; Lou, C.; Li, X.; Yang, M.; Gui, L.; He, Y.; Pei, W.; and Xu, R. 2022. Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1767–1777. Dublin, Ireland: Association for Computational Linguistics.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2, 13–23. Red Hook, NY, USA: Curran Associates Inc.

Lyu, H.; and Luo, J. 2022. Understanding Political Polarization via Jointly Modeling Users, Connections and Multimodal Contents on Heterogeneous Graphs. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 4072–4082. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AISI Track)*, volume 35, 14867–14875.

Meta. 2023. Meta Reports First Quarter 2023 Results. Meta Investor Relations. https://investor.fb.com/investor-news/press-release-details/2023/Meta-Reports-First-Quarter-2023-Results/default.aspx.

Mishra, S.; Saini, A.; Makki, R.; Mehta, S.; Haghighi, A.; and Mollahosseini, A. 2022. Tweetnerd-end to end entity linking benchmark for tweets. *Advances in Neural Information Processing Systems*, 35: 1419–1433.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention Bottlenecks for Multimodal Fusion. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 14200–14213. Curran Associates, Inc.

Parmentier, A.; P'ng, J.; Tan, X.; and Cohen, R. 2021. Learning Reddit User Reputation Using Graphical Attention

Networks. In *Future Technologies Conference (FTC) 2020, Volume 1*, 777–789. ISBN 978-3-030-63128-4.

Qian, J.; Bethke, A.; Liu, Y.; Belding, E.; and Wang, W. Y. 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4755–4764. Hong Kong, China: Association for Computational Linguistics.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Sahu, G.; Cohen, R.; and Vechtomova, O. 2021. Towards a multi-agent system for online hate speech detection. *Second Workshop on Autonomous Agents for Social Good (AASG), AAMAS, 2021*.

Singh, A.; Dey, S.; Singha, A.; and Saha, S. 2022. Sentiment and Emotion-Aware Multi-Modal Complaint Identification. *Proceedings of the AAAI Conference on Artificial Intelligence (AISI Track)*, 36(11): 12163–12171.

Tan, Y.; Liu, F.; Li, B.; Zhang, Z.; and Zhang, B. 2022. An Efficient Multi-View Multimodal Data Processing Framework for Social Media Popularity Prediction. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 7200–7204. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Vedeler, J. S.; Olsen, T.; and Eriksen, J. 2019. Hate speech harms: a social justice discussion of disabled Norwegians' experiences. *Disability & Society*, 34(3): 368–383.

Vidgen, B.; Nguyen, D.; Margetts, H.; Rossini, P.; and Tromble, R. 2021a. Introducing CAD: the Contextual Abuse Dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2289–2303. Online: Association for Computational Linguistics.

Vidgen, B.; Thrush, T.; Waseem, Z.; and Kiela, D. 2021b. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1667–1682. Online: Association for Computational Linguistics.

Wachs, S.; Gámez-Guadix, M.; and Wright, M. F. 2022. Online hate speech victimization and depressive symptoms among adolescents: The protective role of resilience. *Cyberpsychology, Behavior, and Social Networking*.

Waller, I.; and Anderson, A. 2021. Quantifying social organization and political polarization in online platforms. *Nature*, 600(7888): 264–268.

Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; and Liu, T.-Y. 2022. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems*, 34: 28877–28888.

Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; and Kumar, R. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1415–1420. Minneapolis, Minnesota: Association for Computational Linguistics.