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

Liam Hebert
liam.hebert@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

Gaurav Sahu
gaurav.sahu@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

Nanda Kishore Sreenivas
nksreenivas@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

Lukasz Golab
lgolab@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

Robin Cohen
rcohen@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada



## ABSTRACT

We present the Multi-Modal Discussion Transformer (mDT), a novel multi-modal graph-based transformer model for detecting hate speech in online social networks. In contrast to traditional text-only methods, our approach to labelling a comment as hate speech centers around the holistic analysis of text and images. This is done by leveraging graph transformers to capture the contextual relationships in the entire discussion that surrounds a comment, with interwoven fusion layers to combine text and image embeddings instead of processing different modalities separately. We compare the performance of our model to baselines that only process text; we also conduct extensive ablation studies. We conclude with future work for multimodal solutions to deliver social value in online contexts, arguing that capturing a holistic view of a conversation greatly advances the effort to detect anti-social behaviour.

**ACM Reference Format:**
Liam Hebert, Gaurav Sahu, Nanda Kishore Sreenivas, Lukasz Golab, and Robin Cohen. 2023. Multi-Modal Discussion Transformer: Integrating Text, Images and Graph Transformers to Detect Hate Speech on Social Media. In *Proceedings of Pre-Print*. ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

Social media has democratized public discourse, enabling billions of users worldwide to freely express their opinions and thoughts on a global scale. As of 2023, the social media giant Meta has reached 3 billion daily active users across its platforms [19]. While this level of connectivity and access to information is undeniably beneficial, it has also resulted in the alarming rise of hate speech, which refers to any form of communication that intends to belittle, intimidate, or discriminate against individuals or groups based on their race, ethnicity, religion, gender identity, sexual orientation, or other personal characteristics [3]. This pervasive spread of hateful rhetoric has caused significant mental and emotional harm to its targets [27] and has triggered social divisions and polarization [30]. As such, there is an urgent need for automated solutions that can effectively identify and combat hate speech in online communities.

Initially, automated hate speech detection models were limited to text-only approaches such as HateXplain [18], 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 text comment may be innocuous when taken alone, but the inclusion of 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 the context of a discussion about immigration or minority issues.

There is ongoing research to address these weaknesses. For example, multi-modal transformers such as VilT [13] can combine images and text for a richer representation of comments, but they do not account for the contextual nature of hate speech. On the other hand, Hebert et al. [8] do not discuss how to integrate the interpretation of images within hateful social media discussions, but they do address the concern of modeling context. This is done by adapting graph neural networks to model the relationships between comments, first creating text embeddings of comments and then aggregating those embeddings as nodes in a graph. However, the sequential nature of this architecture prevents text embeddings from being created in relation to other comments in a graph. That is, the initial semantic content encoded by a comment embedding may differ when considered together with different sets of comments versus in isolation.

To overcome the limitations of the existing graph and comment-only methods, we propose the Multi-Modal Discussion Transformer (mDT), a method to holistically encode comments in relation to the multi-model discussion context for hate speech detection. We make the following contributions.

(1) As the core of mDT, we propose a novel fusion mechanism that interweaves multi-modal fusion layers with graph transformer layers, allowing for multi-modal comment representations that are actively created in relation to the discussion context.


*Pre-Print, NA, NA*

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

(2) We propose a novel graph structure encoding specific to the conversational structure of social media discussions.
(3) We introduce a new dataset of over 8000 annotated discussions, totaling 18000 labeled comments, with complete discussion trees and images to evaluate the effectiveness of mDT. For this, we focus on the social platform Reddit, where discussions take place in branching tree structures where any user can reply to the comments of other users, forming separate sub-discussions.

We compare mDT against comment-only and graph methods [8] 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 behavior in online communities. We also propose future work towards more advanced multi-modal solutions that can better capture the nuanced nature of online communication and behavior, and that can adapt to the ever-changing landscape of social media. These efforts can be crucial in creating a safer, more inclusive, and positive online environment for all users. Our codebase, datasets, and pre-trained model weights will be found at https://github.com/liamhebert/MultiModalDiscussionTransformer.

## 2 RELATED WORK

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based language representation model. It is pre-trained on large amounts of text and has been successful in a wide array of natural language processing tasks, including hate speech detection [4, 20]. Caselli et al. [1] introduced HateBERT, a BERT model re-trained on the Reddit Abusive Language dataset to detect hate speech. This dataset contains posts from communities that were banned for promoting hateful, abusive, and offensive content. A recent approach for text-only hate speech detection by Vidgen et al. utilized data augmentation to improve performance [29].

Hebert et al. [8] use contextual information (other comments in the discussion) to improve hate speech detection. They use BERT (fine-tuned on the HateXplain dataset [18]) to generate embeddings for each comment in a discussion. These are then aggregated and transformed by a modified Graphormer [31] architecture that predicts whether the conversation from that point onwards will lead to hate speech. The authors demonstrated noticeable improvement in predicting hate speech, compared to comment-only HateExplain [7]. As we mentioned earlier, this is, however, a text-only solution.

Given the increasing prevalence of images in online discussions, hate speech detection has become a multi-modal problem. Below we summarize some of these key models. mDT builds on this work, and additionally takes discussion context into account.

Kiela et al. [12] introduced the hateful memes challenge, where each sample contains an image/meme with a short text/caption, and the task was to predict if the image was hateful or not. VisualBERT [15] integrates pre-trained object proposals systems and BERT. The image features extracted using Faster-RCNN are passed as input tokens to the model along with the text. Thus, the image and text inputs are jointly processed by the transformer layers. ViLBERT [16] has separate transformers for image and text, but they interact through co-attentional transformer layers. Kiela et al. [12] benchmarked several methods on their dataset, and found that early fusion methods such as ViLBERT and VisualBERT significantly outperformed late fusion and other unimodal approaches.

Nagrani et al. [21] proposed the Multimodal Bottleneck Transformer (MBT), which uses fusion bottlenecks for multimodal fusion. Instead of pairwise self-attention at each layer, this model forces each modality to condense the information to only a few bottleneck tokens before sharing it with the other modality. This approach reduces computational costs while improving fusion performance.

Sahu et al. [24] used adaptive fusion techniques to combine visual and textual cues for multi-modal hate speech detection. Dosovitskiy et al. [5] proposed the Vision Transformer (ViT), which reshapes 2D images into a sequence of patches followed by simple linear projection before feeding them to the transformer. Kim et al. [13] proposed Vision-and-Language Transformer (ViLT). Unlike prior vision and language transformer-based models, ViLT is convolution-free and uses a similar approach to ViT (i.e., linear projection of flattened patches). ViLT was found to be significantly faster and performed better at several multi-modal tasks.

## 3 METHODS

### 3.1 Multi-Modal Discussion Transformer (mDT)

The mDT architecture consists of three components: Initial Pre-Fusion, Modality Fusion, and Graph Transformer (Figure 1). The description below expands upon the operations that assist with hate detection and outlines the inherently holistic nature of our solution.

*3.1.1 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 by leveraging pre-trained BERT and ViT models to encode text and images, respectively. Both models consist of $N$ layers with the same hidden dimension of $d$. In our experiments, we utilized BERT-base and ViT-base, which both 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, denoted as

$$t_c^k = Bert_{init}(t_c), i_c^k = ViT_{init}(i_c)$$

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

*3.1.2 Modality Fusion.* After creating initial 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 [21] and add $b$ shared modality bottleneck tokens $B \in R_{b \times d}$ to $t_c$ and $i_c$. The input sequence then becomes $[t_c^k \mid\mid B], [i_c^k \mid\mid B]$. We then define a modality fusion layer $l$ as

$$[t_c^{l+1}||B_{t,c}^{l+1}] = Bert_l([t_c^l||B_c^l])$$
$$[i_c^{l+1}||B_{i,c}^{l+1}] = ViT_l([i_c^l||B_c^l])$$
$$B_c^{l+1} = Avg(B_{t,c}^{l+1}, B_{i,c}^{l+1})$$

where both modalities can 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 there are no images attached to a comment then $B^{l+1} = B_t^{l+1}$.

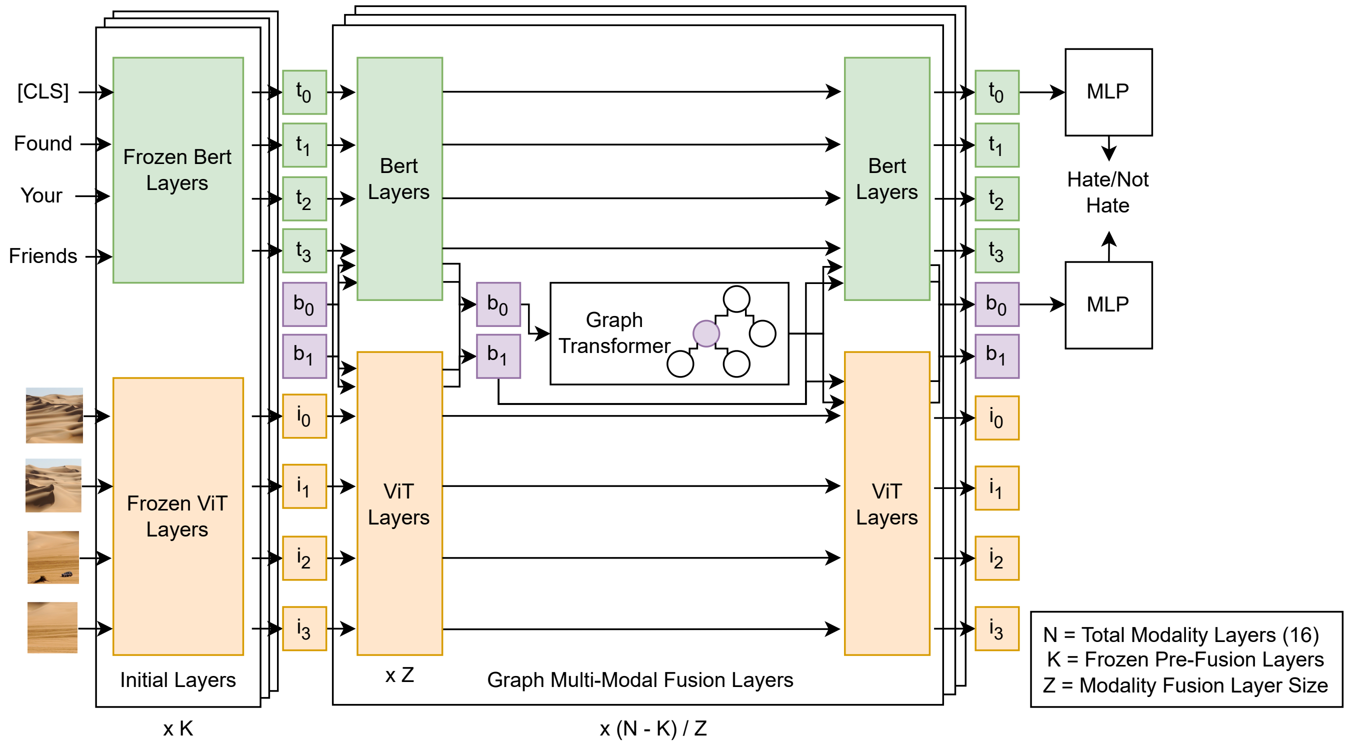


**Figure 1: Multi-Modal Discussion Transformer**

*3.1.3 Graph Transformer.* Then, after $Z$, where $Z < (N - K)$, modality fusion layers, we deploy Graph Transformer layers to aggregate contextual information from the other comments in the graph. For this, we modify the Graphormer Graph Transformer mechanism proposed by [31]. 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 [26], we include two learned structure encodings. The first is Centrality Encoding, denoted $z$, which encodes the degree of nodes in the graph [31]. Since social media discussion graphs are bidirectional, 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

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

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 structural relationship between two nodes $c, v$ in the graph. 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

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

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 $c, v$ in the graph [31]. 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

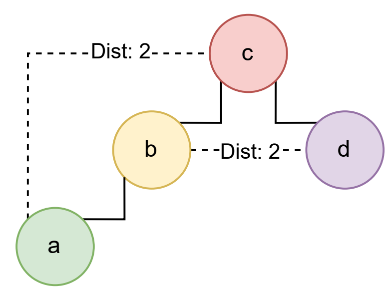


**Figure 2: Example Discussion Structure**

graph. However, $b$ and $d$ represent direct replies to the same parent post whereas $a$ is two comments underneath $c$.

To account for this, we propose a novel hierarchical spatial encoding based on Cantor's pairing function [10]. Cantor's pairing function uniquely maps sets of two numbers into a single number $\mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$. We utilize this function to encode structure as follows: 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 between $a$ and $d$ would be $u_{(a,b)} = 2, d_{(a,b)} = 1$. We then compress both numbers into a single index using the proposed position-independent variant of Cantor's pairing:

$$
\begin{aligned}
s_{(c,v)} &= s_{(v,c)} \\
&= Cantors(u_{(c,v)}, d_{(c,v)}) \\
&= \frac{(u_{(c,v)} + d_{(c,v)})(u_{(c,v)} + d_{(c,v)} + 1)}{2} + min(u_{(c,v)}, d_{(c,v)})
\end{aligned}
$$

which 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 layers and fusion layers through modality bottleneck tokens ensures that fusion models create representations that are grounded in the discussion context.

### 3.2 HatefulDiscussions Dataset

To train our model, we require a diverse dataset of complete discussion graphs with multi-modal comments and a wide range of hateful content. To ensure that our dataset met these requirements, we merged several existing datasets that featured labeled hateful comments (described below). For each labeled comment, we retrieved the corresponding complete discussion tree using the Pushshift Reddit API and downloaded all associated images. To refine our dataset, we filtered out conversations without any 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 are able to reduce computational complexity and focus the discussion on the most relevant parts of the conversation [22].

The first dataset we utilized was the Slurs corpus [14], which contained annotated comments with both derogatory and non-derogatory slurs. We retrieved comments from the non-derogatory slur (NDG), derogatory slur (DEG), and homonym (HOM) categories. We chose this dataset because we believed that understanding the meaning of slurs would be enhanced by considering their discussion context. The second dataset we employed was the Contextual Abuse Dataset [28], which included comments with fine-grained hate speech labels that were annotated with respect to prior comments in the discussion. We retrieve comments from the

**Table 1: Label Distribution of the Hateful Discussions Dataset**

| Label | Count |
|---|---|
| Derogatory Slur (DEG) | 4258 |
| Not Derogatory Slur (NDG) | 2385 |
| Homonym (HOM) | 361 |
| LTI Normal | 4083 |
| LTI Hate | 1295 |
| Neutral | 4876 |
| Identity Directed Abuse | 700 |
| Affiliation Directed Abuse | 273 |
| Normal | 11705 |
| Hateful | 6526 |

**Table 2: mDT Model Hyperparameters**

| Hyper Parameter | Value |
|---|---|
| Pre-Fusion Layers (K) | 4 |
| Modality Fusion Layers (Z) | 4 (12 total) |
| Graph Transformer Layers (G) | 2 (6 total) |
| Bottleneck Size (B) | 4 |
| Max Spatial Attention | 5 |
| Learning Rate | $3e^{-5} \rightarrow 3e^{-7}$ |
| Learning Rate Scheduler | Linear |
| Hidden Dimension (d) | 768 |
| Graph Attention Heads | 12 |
| Modality Attention Heads | 12 |
| Batch Size | 48 |

Neutral, AffiliationDirectedAbuse, and IdentityDirectedAbuse categories. Finally, we also used the Learning to Intervene (LTI) Dataset [23], which was created by labelling multiple comments from the same conversation as either hateful or not. By incorporating expanded data from many datasets, we are able to train our system on a much wider breadth of hateful discussions that contain multi-modal elements. We believe that providing this dataset publicly can enable future research into robust graph-based methods for hate speech detection.

In order to train our models, we map each of the retrieved labels to either Hate or Normal and treat the problem as a binary classification. The final distribution of each label can be seen in Table 1.

## 4 RESULTS

### 4.1 Experimental Setup

In our experiments, we conduct a 7-fold stratified cross-validation (equivalent to a 14% test split) and report the average performance for each model. By utilizing 7-fold, we allow for a larger diversity of labels across each fold, as opposed to 10-fold validation. We report overall accuracy and class-weighted Precision, Recall and F1 to account for label imbalance. Unless otherwise noted, the model hyperparameter configuration we use for mDT can be seen in Table 2.

**Table 3: Performance of mDT against Text-Only Methods**

| Method | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| Bert-HateXplain [18] | 0.742 | 0.763 | 0.742 | 0.747 |
| Detoxify [6] | 0.687 | 0.679 | 0.696 | 0.677 |
| RoBertA Dynabench [29] | 0.811 | 0.822 | 0.811 | 0.814 |
| Bert-HatefulDiscussions | 0.858 | 0.858 | 0.858 | 0.858 |
| Graphormer [8] | 0.735 | 0.594 | 0.759 | 0.667 |
| mDT | **0.880** | **0.880** | **0.880** | **0.877** |

**Table 4: Effect of Bottleneck Size on mDT Performance**

| Bottleneck Size | Accuracy | Precision | Recall | 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 |

## 4.2 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 [18], Detoxify [6], and RoBertA Dynabench [29]. We also compared mDT against a Bert model trained on the training set of HatefulDiscussions (Section 3.2), which we refer to as Bert-HatefulDiscussions. To compare against previous graph-based approaches, we evaluated text-only Graphormer [8].

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

## 4.3 Effect of Bottleneck Size

Next, we investigated the impact of increasing the number of bottleneck interaction tokens ($B$) in mDT, which are added during the modality fusion step. By adding more bottleneck tokens, we reduce 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 tokens, indicating the importance of compression when exchanging modality encodings between models. We assume that this reduction is due to the importance of compressed information to represent comments in the graph transformer network.

**Table 5: Effect of Constraining Graph Attention**

| Attention Window | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| 2 | 0.866 | 0.866 | 0.866 | 0.866 |
| 5 | **0.880** | **0.880** | **0.880** | **0.877** |
| $\infty$ | 0.870 | 0.861 | 0.850 | 0.855 |

**Table 6: Effect of Fusion Layers**

| Total Fusion Layers | Accuracy | Precision | Recall | 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 7: Effect of Incorporating Images**

| Usage of Images | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| With Images | **0.880** | **0.880** | **0.880** | **0.877** |
| Without Images | 0.832 | 0.835 | 0.822 | 0.828 |

## 4.4 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 [7]. 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 2 hops. Our findings suggest that a balance is required when constraining graph attention for optimal performance.

## 4.5 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, presented in Table 6, indicate that utilizing 12 fusion layers leads to the best performance. Interestingly, we found that the performance gains did not follow a linear trend with the number of fusion layers. Specifically, we observed that 8 fusion layers outperformed 10 layers, but were still inferior to 12 layers. We believe that further research in this area should explore the potential benefits of scaling beyond 12 fusion layers using larger modality models.

## 4.6 Effect of Images

We also investigated the impact of including images in mDT. Our findings (Table 7) support the hypothesis that images provide crucial contextual information for detecting hateful content. Specifically, we observed that incorporating images into mDT led to a 4.8% improvement in accuracy and a 4.9% improvement in the F1 score.

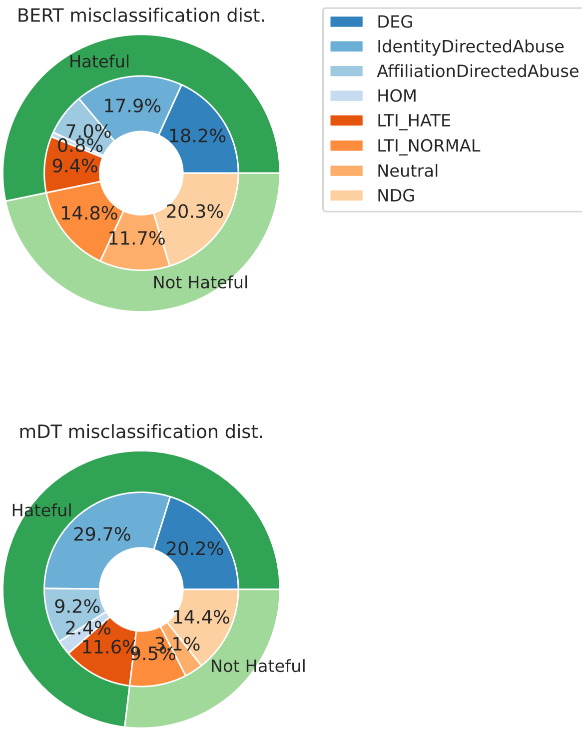


**Figure 3: Fine-grained distribution of BERT and mDT misclassifications.**

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

### 4.7 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 instances. We further note that BERT and mDT predictions disagree on 264 test instances, out of 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 major reduction in false positives (fewer misclassifications for the 'Not Hateful' class). However, BERT and mDT struggle to detect the presence of hate speech in derogatory (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 first example in Table 8. The word "tranny" is a common acronym for "transmission" on social media, but considering the context, it is clearly an abusive discussion directed toward the transgender community. 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 two examples in Table 8, both primary text and comments in the context are non-abusive. The only clear indicator of hate speech is an abusive image attached to the discussions. This suggests that while adding context results in a net decrease in misclassifications, majorly neutral context might also fool the model, given that we average the text embedding logit and the node embedding ($b_c^0$) to obtain the final classification.

## 5 FUTURE WORK

While we find mDT to be an effective method for analyzing discussions on social media, we have identified some limitations and areas for improvement. First, mDT is prone to be misled when the discussion context contains comments that are predominately neutral, as discussed in Section 4.7. To address this, future work could explore filtering and weighing some comments in the discussion to reduce noise. For example, a first-stage text ranker could be deployed to compute semantic relevance between comments and filter unrelated messages accordingly.

Secondly, there are still many contextual signals in social media discussions beyond text, images, and discussion structure that remain untapped. User modeling techniques could be employed to create a richer understanding of a user's background, especially in regards to understanding the usage of re-appropriated slurs and homonyms. Additionally, incorporating named entity recognition techniques to identify and expand named entities mentioned in the discussion could enable the model to leverage real-world knowledge and provide a richer context for hate speech detection [9].

Finally, the versatility of mDT's core mechanisms makes it a promising tool for a wide range of applications beyond hate speech detection. The rich and contextual multi-modal representations of discussions it generates can provide valuable contextual information for tasks such as information retrieval and recommendation systems. For example, mDT could be used to surface relevant discussions or related content to users based on their interests, preferences, or search queries. Furthermore, the approach could be extended to other domains such as online product reviews [11], political discourse analysis [17], and popularity analysis [2, 25], where understanding the discussion context is critical for accurate interpretation.

## 6 CONCLUSION

In this paper, we presented a holistic approach to detecting hate speech in social media using our mDT model. Our model leverages graph transformers together with text and image transformers to reason about entire threads of discussion. Core to our approach

**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 [. . . ]. Finally, we have not included the images from the discussion due to their profane nature.**

| **Primary Text** | **Context (only seen by mDT)** | BERT pred. | mDT pred. |
|---|---|---|---|
| I'm a tranny chaser (throwaway account) and I'm quite familiar with "the tuck," but my mind didn't even go there. I thought it was an oddly dislodged tampon, or something. | **[. . . ]** *f-slur* that guy **[. . . ]** I'm not a bro, and I *c-slur* take all the puns here!. . . *c-slur*, hes had a hard life, give him a break **[. . . ]** | Not Hateful | Hateful |
| 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 |
| "That *n-slur* was on PCP Johnson" Lmao | **[. . . ]** 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 |
| whoa brah.. leave my tranny out of this | **[. . . ]** that's *f-slur* retarded **[. . . ]** Just spit my drink **[. . . ]** | Not Hateful | Hateful |
| Is like the lovechild of the kkk and a vietnong that got possessed by a ghost. | **[. . . ]** Rule 34? **[. . . ]** `anonymized_username` werry like the fashurn, do you know it? **[. . . ]** looks like an assassin's creed character | Hateful | Not 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 |

is the introduction of hierarchical spatial encodings and coupling of text, image and graph transformers through a novel bottleneck mechanism to produce an integrated solution specific to social discussions. We also present 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.

One significant contribution is demonstrating how multi-modal analysis can improve the detection of anti-social behavior online. Experimental results, compared with several key competitors, provide important quantitative metrics; an initial effort to present examples to show how the lack of holistic multi-modal analysis will compromise success introduces a valued qualitative perspective as well. These steps with analysis, measured against our proposed dataset of multi-modal discussions, provide practitioners with additional insights into where the challenges lie in order to deliver social good in our current online environment, by embracing a multi-modal viewpoint. Our results overall demonstrate a significant advancement in the application of graph networks for hate speech detection.

Another important outcome of our work is highlighting the value of graph transformers when dealing with online content that has a notable emotional nature, and where it is clearly insufficient to simply examine comments in isolation. While graph transformers are gaining important momentum within the artificial intelligence community, revealing their power by efficiently incorporating a multi-modal context may offer new inspirations for both applied and theoretical investigations. This in turn may help to provide much-valued attention to our community of researchers, who are devoted to research on multi-modal approaches to AI.

In addition to the theme of engaging users of multimedia with social signals within emotional contexts, and benefiting society through the experience of multi-modal solutions, a third topic of interest is of examining new insights into how to achieve multi-modal fusion and embedding to better understand multimedia content. The unique architecture sketched in this paper for our particular application may be of use to researchers who are examining companion issues, such as information retrieval and recommender systems on social platforms. Overall, we believe that our approach presents a promising 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.

## REFERENCES

[1] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for Abusive Language Detection in English. In

*Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. Association for Computational Linguistics, Online, 17–25. https://doi.org/10.18653/v1/2021.woah-1.3
[2] Weilong Chen, Chenghao Huang, Weimin Yuan, Xiaolu Chen, Wenhao Hu, Xinran Zhang, and Yanru Zhang. 2022. Title-and-Tag Contrastive Vision-and-Language Transformer for Social Media Popularity Prediction. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 7008–7012. https://doi.org/10.1145/3503161.3551568
[3] Mithun Das, Binny Mathew, Punyajoy Saha, Pawan Goyal, and Animesh Mukherjee. 2020. Hate Speech in Online Social Media. *SIGWEB Newsl.* Autumn, Article 4 (nov 2020), 8 pages.
[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]
[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2023. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://openreview.net/forum?id=YicbFdNTTy&utm_campaign=f86497ed3a-EMAIL_CAMPAIGN_2019_04_24_03_18_COPY_01&utm_medium=email&utm_source=Deep%20Learning%20Weekly&utm_term=0_384567b42d-f86497ed3a-72965345
[6] Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.
[7] Liam Hebert, Hong Yi Chen, Robin Cohen, and Lukasz Golab. 2023. Qualitative Analysis of a Graph Transformer Approach to Addressing Hate Speech: Adapting to Dynamically Changing Content. *arXiv preprint arXiv:2301.10871* (2023).
[8] Liam Hebert, Lukasz Golab, and Robin Cohen. 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. https://doi.org/10.1109/WI-IAT55865.2022.00012
[9] Liam Hebert, Raheleh Makki, Shubhanshu Mishra, Hamidreza Saghir, Anusha Kamath, and Yuval Merhav. 2022. Robust Candidate Generation for Entity Linking on Short Social Media Texts. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*. Association for Computational Linguistics, Gyeongju, Republic of Korea, 83–89. https://aclanthology.org/2022.wnut-1.8
[10] John E Hopcroft, Rajeev Motwani, and Jeffrey D Ullman. 2001. Introduction to automata theory, languages, and computation. *Acm Sigact News* 32, 1 (2001), 60–65.
[11] Rajkumar S Jagdale, Vishal S Shirsat, and Sachin N Deshmukh. 2019. Sentiment analysis on product reviews using machine learning techniques. In *Cognitive Informatics and Soft Computing: Proceeding of CISC 2017*. Springer, 639–647.
[12] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. https://doi.org/10.48550/arXiv.2005.04790 arXiv:2005.04790 [cs].
[13] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 5583–5594. https://proceedings.mlr.press/v139/kim21k.html ISSN: 2640-3498.
[14] Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 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*. Association for Computational Linguistics, Online, 138–149. https://doi.org/10.18653/v1/2020.alw-1.17
[15] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. https://doi.org/10.48550/arXiv.1908.03557 arXiv:1908.03557 [cs].
[16] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Number 2. Curran Associates Inc., Red Hook, NY, USA, 13–23.
[17] Hanjia Lyu and Jiebo Luo. 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* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 4072–4082. https://doi.org/10.1145/3503161.3547898
[18] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14867–14875.
[19] 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.
[20] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. In *Complex Networks and Their Applications VIII (Studies in Computational Intelligence)*, Hocine Cherifi, Sabrina Gaito, José Fernendo Mendes, Esteban Moro, and Luis Mateus Rocha (Eds.). Springer International Publishing, Cham, 928–940. https://doi.org/10.1007/978-3-030-36687-2_77
[21] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention Bottlenecks for Multimodal Fusion. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 14200–14213. https://proceedings.neurips.cc/paper_files/paper/2021/file/76ba9f564ebbc35b1014ac498fafadd0-Paper.pdf
[22] Alexandre Parmentier, Jason P'ng, Xiang Tan, and Robin Cohen. 2021. Learning Reddit User Reputation Using Graphical Attention Networks. In *Future Technologies Conference (FTC) 2020, Volume 1*. 777–789.
[23] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 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)*. Association for Computational Linguistics, Hong Kong, China, 4755–4764. https://doi.org/10.18653/v1/D19-1482
[24] Gaurav Sahu, Robin Cohen, and Olga Vechtomova. 2021. Towards a multi-agent system for online hate speech detection. *Second Workshop on Autonomous Agents for Social Good (AASG), AAMAS, 2021* (2021).
[25] YunPeng Tan, Fangyu Liu, BoWei Li, Zheng Zhang, and Bo Zhang. 2022. An Efficient Multi-View Multimodal Data Processing Framework for Social Media Popularity Prediction. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 7200–7204. https://doi.org/10.1145/3503161.3551607
[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[27] Janikke Solstad Vedeler, Terje Olsen, and John Eriksen. 2019. Hate speech harms: a social justice discussion of disabled Norwegians' experiences. *Disability & Society* 34, 3 (2019), 368–383. https://doi.org/10.1080/09687599.2018.1515723
[28] Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. 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*. Association for Computational Linguistics, Online, 2289–2303. https://doi.org/10.18653/v1/2021.naacl-main.182
[29] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. In *ACL*.
[30] Isaac Waller and Ashton Anderson. 2021. Quantifying social organization and political polarization in online platforms. *Nature* 600, 7888 (01 Dec 2021), 264–268. https://doi.org/10.1038/s41586-021-04167-x
[31] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2022. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems* 34 (2022), 28877–28888.