Figure 3: Examples of Multi-Modal Hate Speech

While there may be some validity to these predictions regarding their contentiousness, it does raise a concern about how to moderate these heated debates. However, in the case of the example in Table 4, the most inciteful comments (depth 2a, 2b, and 5) are appropriately labelled as such. In each of these cases, comment-only methods predicted each comment as neutral, even if the comments were hateful (depth 2b). We found that Graphormer was more sensitive to higher predictions than GAT when faced with these types of comments.

In each of the examples, we also evaluated the ability of graph networks to adapt to evolving social media conversations. Coping with this dynamic change is essential for the successful real-life implementation of AI for social impact. We evaluate this behaviour by iteratively predicting comments in the discussion graph in a depth-wise fashion, differing from **?**. As a result, we constrain graph models to predict labels from only the context provided by previous comments, mirroring how the system would be deployed in real situations. Despite this constraint, we still see that graph systems are able to make accurate predictions. This can best be seen in Table 1, where the graph models adapted their predictions to be less hateful once the conversation developed.

We also found that many examples we retrieved were entered around multi-modal posts. Such examples include the discussion in Table 1, involving an image of a tweet, and the discussion in Table 5, involving an article concerning Trump and the Ukraine war, among others. When investigating contextual hate speech, Table 3 presents an example where the image (Figure 2) provides important context to the comments that followed.

By analyzing this picture, it would be possible to understand that the discussion concerns an LGBTQ drag queen competing in an elaborate dress. However, without this context, we found that comment-only methods misclassified supportive speech using reclaimed LGBTQ vernacular as hateful. This is especially concerning given that these predictions could serve to suppress communities that are vitally important to the mental health of minority populations (**??**). Furthermore, memes sent on online platforms are often only hateful if one considers both the image and the text caption together, as seen in Figure 3 (**?**). By taking a holistic view of conversations by encoding images, text, and discussion structure together, we hypothesize those hate speech detection methods would be able to avoid many false predictions, such as the ones incurred in Table 3. Furthermore, following **?**, it would be possible to include user-level information into this graph representation.

Finally, it is also important to analyze the mental health impact given by a graph approach to hate speech. By reformulating hate speech as a graph prediction task, we are able to train systems that can leverage discussion context toward predicting the direction of conversations. This can allow moderators on social platforms to be alerted of potentially harmful comments and deploy mitigation strategies to shield users who are susceptible to mental health effects. We see an example of such a discussion in Table 4, where users that are susceptible to trauma from guns and race can be warned ahead of time by utilizing the proactive graph predictions. Furthermore, by utilizing an increasing ordinal scale (from zero to four) for predicting hate, users can select their level of comfort by choosing the intensity of contentious comments they are comfortable viewing. As the conversation develops, these predictions can then be updated with further context and revised accordingly. An example of where this would be useful is the discussion in Table 3, where further comments add credence to the innocence of previous comments. By providing these scores, platform owners can allow users to have control over the content they see through self-moderation. Another valued opportunity for deployment of our methods shown by qualitative analysis is in assiting platforms to curtail hate speech proliferation: greater prediction of impending escalating harm and caution in imposing penalites when discussion isn't hate can both be addressed.

Conclusion

In this work, we explored the impact of Graph Transformer Networks on hate speech detection (**?**). To do this, we performed an extensive qualitative analysis of graph and comment-only methods on conversations sampled from different communities on Reddit. When examining contextual hate speech, we found that Graph Transformer Networks can prevent both false positives and false negatives incurred by comment-only methods. In these cases, context played a key role in understanding the nature of analyzed comments. We also found similar gains in performance when analyzing discussions that concerned inciteful speech. However, we also found that debates were prone to high hate predictions despite being mostly civil.

Guided by this study, one promising direction for future work is to include more modalities to better contextualize comments. Among the examples we retrieved, many were centred around an image or article. We hypothesize that utilizing a holistic view of conversations by including all modalities can help prevent false positives. Most importantly, this approach could help catch the most pervasive hate speech of all - discourse.

Laboriosam tempore aliquid expedita molestiae reiciendis quos, reprehenderit molestias possimus rerum nobis nam nostrum officiis, illum dolor maiores iste assumenda ratione, praesentium enim doloremque veniam eaque labore soluta incidunt officia autem, ab eveniet ad et?Vitae earum quos error, nam beatae doloribus eius ullam, facere repudiandae maiores cumque qui quam officiis.Nesciunt dolorem rerum blanditiis incidunt fugiat, exercitationem molestiae recusandae eius, labore neque illo rem ducimus illum aperiam quis

similique numquam, iusto hic id aspernatur possimus reprehenderit nobis unde mollitia esse, maxime omnis nam sequi?Hic dolor nemo, minus inventore nam non illum