Paper Title:
Emotion Based Hate Speech Detection using Multimodal Learning

Paper Link:
https://arxiv.org/pdf/2202.06218.pdf

1. Summary

1.1 Motivation
The paper addresses the pressing need for a more comprehensive approach to hate speech detection, considering the limitations of text-only analysis. It explores the integration of both text-based classification and speech-based emotion prediction to enhance hate speech detection in multimedia content.

1.2 Contribution
The study contributes a multimodal learning framework that combines linguistic content analysis with vocal signals to detect hate speech more effectively. It presents three models: text-based hate speech classification, speech-based emotion attribute prediction, and a multimodal model combining text and audio embeddings.

1.3 Methodology
The approach involves utilizing transformer networks pre-trained on offensive tweets for text-based hate speech classification. Speech-based emotion attribute prediction is performed through a multi-task deep learning model, predicting emotions as continuous values. The multimodal model combines text and audio embeddings, leveraging joint representations through neural networks for hate speech classification.

1.4 Conclusion
The experiments demonstrate that incorporating speech features significantly enhances hate speech detection compared to text-only approaches. The study underscores the potential benefits of multimodal approaches in understanding hate speech within multimedia content.

2. Limitations

2.1 First Limitation
The paper acknowledges the limitations in the size and diversity of the dataset used, impacting the generalizability of the models to a broader range of hate speech instances.

2.2 Second Limitation
Another limitation highlighted is the absence of visual modalities in the multimodal model, which could further improve hate speech detection, such as detecting facial expressions or identifying specific objects related to hate speech.

3. Synthesis
The paper presents a significant advancement in hate speech detection by integrating text and speech features. However, it acknowledges the need for larger and more diverse datasets and emphasizes the potential inclusion of visual modalities for a more comprehensive hate speech detection system in multimedia content.