Feelings Behind Words

Eeshan Kandikattu Caleb Kim Adhitya Vootukuru Kshitij Kulshrestha Vaishnavi Pasumarthi
Thejaswin Kumaran Sahil Sidheekh Saurabh Sanjay Mathur Dr. Sriraam Natarajan

Introduction

Political discourse is essential for informed democratic decision-making but can be susceptible to manipulation through persuasive techniques and misinformation. The rise of social media (5.04 billion users in 2024) exacerbates the spread of misinformation within online communities, making moderation challenging. In this project, we consider a sub-problem within the larger problem of large-scale moderation, namely, that of automatic detection of persuasion techniques employed in social media content. We present a multi-modal Artificial Neural Network (ANN) based model for detecting persuasion techniques in social media posts. Our model consists of two sub-models that analyze the text and image content of a social media post. We combine the predicted probabilities from each model by averaging them. This work demonstrates the effectiveness of a multimodal approach (text and image analysis) in analyzing social media content.

Model

Model 1 (Sentiment Analysis): The process of analyzing text to determine the emotional tone of the message. Initially, textual data is tokenized and adapted to facilitate analysis, enabling feature extraction to represent the data in the form of numeric vectors through the use of BERT. These vectors are subsequently employed to classify the sentiment via TensorFlow.

Model 2 (Image Classification): Distinguish/detect objects in an image by transforming an image into a digital form and analyzing the pixels to perform pattern recognition. Images are often pre-processed and undergo the process of feature extraction through the use of CNN's.

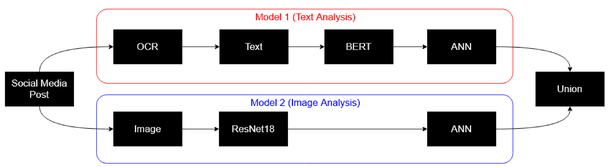

Figure 1. Multimodel Architecture

Combined Model: Through the use of late fusion, a technique that combines the outputs of multiple models, our approach integrates sentiment analysis and image classification. This comprehensive methodology enables us to effectively infer persuasive techniques from both texts and images. Furthermore, by combining the outputs of Model 1 and Model 2, our integrated approach enhances precision in discerning persuasive techniques from both textual and visual content.

Results

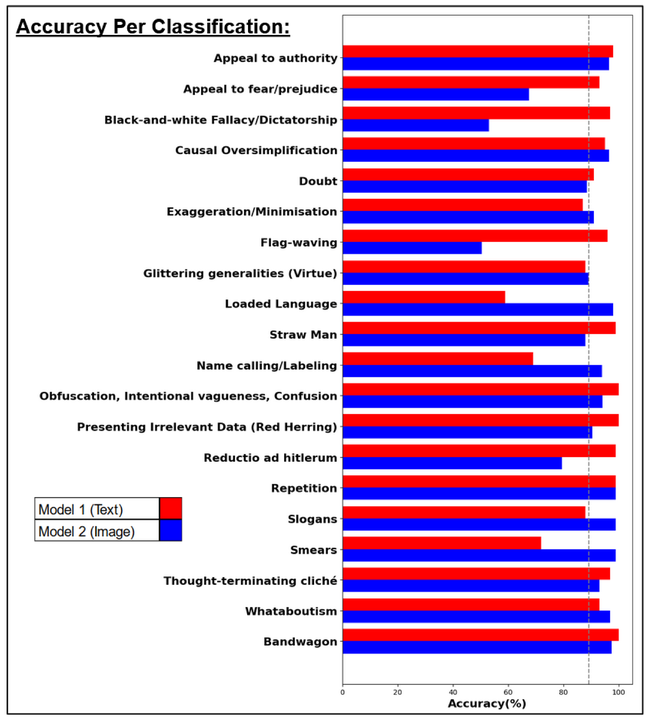

Figure 2. Accuracy Per Classification

Analysis

The overall accuracy of the combined model is 89.2% and the bar plot shows the label-wise accuracy's (which is represented by the dotted vertical line). Of the 20 persuasion techniques, our model could predict appeal to authority, repetition, and bandwagon with high accuracy (+90%) while it struggles to predict flag-waving, loaded language, and name calling (<60% accuracy). Large disparities in accuracy scores between the models indicate that both perspectives inherently give better insight of identifying rhetoric for each classification.

Data Collection

Our dataset comprises 888 data points obtained from the 15th International Workshop on Semantic Evaluation (SemEval 2021). Each data point corresponds to a social media post containing an image and the text content of the image. Each data point is labeled with the persuasion technique (out of 22 possible) corresponding to various rhetorical devices (e.g., black-and-white fallacy, whataboutism, appeal to authority). Data points permit multi-labeling, reflecting the presence of multiple persuasive techniques within a single post. We used the Pandas & Pillow libraries to pre-process the labels and the image data. The data set was split into 688 data points for the training set & 200 for the testing set.

Conclusion

By analyzing online media, our model provides valuable insights into the presence of manipulative content, along with minimizing the spread of propaganda. It identifies harmful rhetoric and could allow people to think twice before making a decision. For example, users may be able to utilize our detection services to better understand advertising during election periods in order to further their awareness of the candidates. With further development time, a more advanced union algorithm could be used such as early fusion or sketch. Contextual analysis could also be improved with the inclusion of CLIP, which could be used to annotate information in images. Our work serves as a foundation for additional research, and with further development, this technology has the potential to allow people to be more aware of how the media they consume is being used against them.

References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT. https://arxiv.org/pdf/1810.04805.pdf

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. CVPR. arXiv preprint arXiv:2105.09284 (2021)

He, K., Zhang, X., Ren, S., & Sun, J. (2015, December 10). Deep residual learning for image recognition. arXiv.org. https://arxiv.org/abs/1512.03385v1

Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal machine learning: A survey and taxonomy." IEEE transactions on pattern analysis and machine intelligence 41.2 (2018)

GoDataDriven. (2019, January 30). Keras: Multi-label classification with image-datagenerator. Xebia. https://xebia.com/blog/keras-multi-label-classification-