# Reflective Essay on Multi-modal Sarcasm Detection Research

Student Name: Vijeth Rai
Student Number: 220228042
Supervisor: Steve Uhlig
MSc Data Science and Artificial Intelligence

## Overview:

In this reflective essay, I offer an in-depth look into my journey through my dissertation project, which deals with the challenging area of multi-modal sarcasm detection. Bridging the gaps in existing literature, the project used ensemble models on fusion techniques to outperform previous models. This essay is about my journey in this domain capturing the full scope of the project, from conceptualization, challenges and lessons learned, highlighting both its academic and practical implications. I have also discussed the future research scope, serving as a guide for those venturing into this domain.

## The Idea

The initial hurdle was identifying gaps in the existing research. Originally, I had considered working on a project requiring either biological expertise or intensive coding skills. Luckily, my supervisor guided me towards a more suitable research focus. I spent many days trying to find unexplored or under explored areas of research in my domain of expertise, Mechanical Engineering. However, I kept finding either too many existing papers on recurring topics or a lack of useful data. It wasn't until a sarcastic comment on social media momentarily stumped me that I wondered about the use of AI in sarcasm detection. While researching existing papers in this topic, I noticed the recurring and lacking trend again. After further research, I discovered an underexplored area in sarcasm detection, which had an intriguing paper that happened to be one of the founding studies in multimodal sarcasm detection. The idea of capturing important features across different modalities of input data captivated both my me and my supervisor, leading me to choosing this domain for dissertation.

After surveying all the existing papers in this domain, which were only a handful, I promptly observed that none had employed transfer learning techniques in audio feature extraction and this mode of input in general, and there was an observable trend of sticking to basic traditional audio features like averaged MFCCs and mel-spectrograms. Moreover, I noticed the deficit in leveraging task-specific transfer learning models to generate visual embeddings. Not only did I find an under-explored area of research, I also found the gaps in existing studies. I discovered my niche quite quickly, downloaded the relevant datasets and was excited to start working on the project.

## Challenges

The dataset itself presented its own set of challenges, as it comprised solely of videos and dataframe of subtitles and ground truths. I realized that extracting all the required audio and visual features from videos will be a formidable task. The videos were short clips, mostly from well-known TV shows like F.R.I.E.N.D.S. and Big Bang Theory. Gathering the images was straightforward as I utilized the available face detection models to capture screenshots when a human face was detected. I then realized the immense value of gathering localized facial images of the final speaker, as this person is the one delivery the last sentence in the video, which would either be sarcastic or non-sarcastic. I dedicated weeks to isolating images of individual speakers by manually going through each video frame. Gathering the text embeddings was the easiest followed by audio feature extractions.

I had already found areas of improvement in existing works. One of them was to use a task-specific transfer learning model. In this case, I chose a model specialized in capturing facial expressions and emotions to generate embeddings for visual data. The most efficient way of performing this operation was to go through frames of the video at a certain skip interval and use facial recognition model to recognize whether there are any face present. Incase there were, localize each face and pass it to the facial expression recognition model. While the logic seemed to be flawless, my computer's RAM limitations proved otherwise. I again had to wait for days breaking down the video data into multiple smaller parts to generate these embeddings. After the embeddings were obtained, I was enlightened that facial detection models are computationally expensive because of their inherent size. This taught me a valuable lesson; never directly assume the code runs slow because of my setup.

Similarly, I used BERT and VGGish as transfer learning models to generate embeddings for text and audio data respectively. After cleaning and preprocessing all the features sets, I was faced with a new challenge, the task of testing performance of multiple models and hyper-tuning. This was also a very lengthy process. After identifying the best performing model, I faced the issue of overfitting. Manual adjustments to the parameters were continuously made until. It felt like every obstacle had lined up against me. The dissertation seemed relentless in its challenges.

One hurdle after another, I persevered. I had finally achieved results that would set the standards for multimodal sarcasm detection. This was the most rewarding part was the dissertation. The ensemble model, using both early and late fusion techniques, outperformed not just unimodal models but also other leading methods. My models had 11.2% increase in precision compared to the top performing model in previous studies. Accuracy, recall and f1-score also had similar improvements. This success validated my approach of utilizing state-of-the-art task-specific transfer learning models which generate high-quality embeddings.

# Areas for Future Work and Reflections

Data scarcity was a prevalent issue, causing the models to overfit in neural networks. The neural networks showed great promise when it comes to performance in train set however, the lack of data makes them unusable due to overfit. Data augmentation could be a potential solution, however excessive manipulation could undermine the model's ability to generalize to new data. Initially, I had intended to gather more data for the project to overcome this problem. However, there is no availability of episodes or shows where sarcasm is frequently used. Moreover, those that did exist were not easily accessible for download and editing, raising ethical concerns while venturing into data creation. Even though short clips can be used for research and educational purposes in many jurisdictions, the inability to download these episodes legally posed a significant constraint. Additionally, the time and resources needed to find, download, and preprocess suitable clips would have been immense. Therefore, I was restricted not only by resource availability but also by the timeframe of the project.

A significant contribution of this project, which will most likely be over-looked is the evaluation of acoustic features which are created without the use of transfer learning. The experimental results were poor, the test results resembling a random model. This sheds light on why previous works neglected audio feature generation. On the bright side, this result can serve as a valuable guidance for future research, which can now safely bypass traditional methods of audio feature extraction in favour of transfer learning techniques that yield richer acoustic features. After observing the traditional acoustic feature performance on various models, my expectations for the embedded audio features were quite low. The results were a surprise for sure, but a welcome one. The same models that struggled with traditional features excelled when utilizing transfer-learned embeddings, confirming the efficacy of transfer learning for audio feature generation.

An unknown lie in the ability of the model to carry on the same results into the real-world scenarios. Although my study provides significantly high results, it is not claiming that sarcasm will remain speaker-independent in real world in a more varied contexts and conversations. This can only be analysed with a greater size of dataset.

One promising avenue for future research is the integration of Generative Pre-trained Transformers (GPT) for generating embeddings of multi-modal input data. Recent advancements in GPT models have extended their capabilities to process visual inputs. Should a future GPT model be developed that is extensively trained on multi-modal data, leveraging such a model through transfer learning could yield even more robust and accurate results in sarcasm detection.

# Summary

In summary, this dissertation has been an intensely educational experience. From identifying gaps in existing research to troubleshooting model overfits, each phase had its challenges and rewards. It has left me with a deep sense of accomplishment and a clear roadmap for future research endeavours. Most importantly, the problem-solving skills, time management skills and project scoping gained are not merely academic; they are transferable life skills that will aid me in future scholarly and professional pursuits.

Thus, while my dissertation may have reached its conclusion, it marks the beginning of further exploration and potential breakthroughs in the fascinating arena of multi-modal sarcasm detection for future researchers.