RELATED WORK

Sarcasm can be tricky to detect as it often depends on how something is said, not just the words themselves which were delivered. For improving sarcasm detection, researchers have found that combining text, voice, and facial expressions—i.e., a multimodal approach—can help computers understand it better.

Wu et al. worked on detecting sarcasm by focusing on the differences between what people say and how they express it. Their paper, *Modeling Incongruity between Modalities for Multimodal Sarcasm Detection,* shows how sarcasm is often a mismatch between positive words and negative expressions or tone. They developed a model that looks at text, voice, and facial expressions together to find these mismatches and identify sarcasm more accurately.

Sangwan et al. also explored this idea in their work, *I Didn't Mean What I Wrote! Exploring Multimodality for Sarcasm Detection.* They used separate models for text, voice, and facial expressions, combining them to get better results. Their study showed that using multiple types of data provides better sarcasm detection than relying on just one, like text alone.

Castro et al., in *Towards Multimodal Sarcasm Detection (An Obviously Perfect Paper),* used a more advanced method. They combined text, voice, and facial expression data with an attention mechanism, which helped their model focus on the most important parts of each. This allowed their system to catch sarcasm more effectively by weighing the information from each source differently.

Pramanick et al., in their paper *Multimodal Learning using Optimal Transport for Sarcasm and Humor Detection,* took a different approach. They used a method called optimal transport to ensure that features from text, voice, and facial expressions worked together smoothly. This helped their model detect both sarcasm and humor more effectively.

Other researchers, like Cai et al., also examined combining multiple types of data. Their work used hierarchical fusion to bring together text, images, and videos for sarcasm detection in social media. This method worked better because it took into account both the specific features of each type of data and the way they relate to each other.

Mittal et al. showed the importance of context in sarcasm detection in their paper, *Multimodal Sentiment Analysis using Hierarchical Fusion with Context Modeling.* Their model combined text, voice, and facial expressions while also considering the situation where these are used. This made their model more accurate in detecting sarcasm.

In summary, researchers agree that using multiple types of data—text, voice, and facial expressions—greatly improves sarcasm detection. Combining these types of data with techniques like attention mechanisms and optimal transport improves the accuracy of sarcasm detection models.