# When words are not enough: Multi-modal context-based sarcasm detection using AI

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

*Abstract*— **The future of AI shows promising possibilities in field of communication, with a variety of applications in therapy, retail, customer service and more. Despite the significant progress in AI, there is a lack of research that explores the extent to which AI can be used to distinguish passive aggression or sarcasm using multimodal inputs. Thus, the aim of this study is to investigate the potential of AI in recognizing subtle forms of human expressions in conversations. This paper introduces an innovative framework that combines early fusion and late fusion with ensemble methods. The study demonstrates a 11.2% improvement in performance over existing research in this domain. Furthermore, two experimental approaches are setup to gain insights on the nature of sarcasm – whether sarcasm is speaker-dependent or not.**

***Keywords—Multimodal, Sarcasm, Passive-Aggressive, Context, AI***

## I. INTRODUCTION

Sarcasm is a complex form of verbal irony which is often a statement that conveys the opposite of its literal meaning, typically as humor, though at times it may also serve as a mode of aggression. It is powerful linguistic art, capable of adding depth and nuance to the conversation. A classic example could be someone commenting “Great weather we’re having” during a severe thunderstorm. Despite the positive phrasing, their tone, expression, and context contradict this statement, indicating the true sentiment is quite the contrary. Research indicates that sarcasm can also serve as a face-saving strategy, making the speaker appear less rude and unfair, particularly when expressing trivial criticism, often seen in a complaint or criticism (Jorgensen et al., 1996).

Given the inherently intricate nature of sarcasm, detecting it using natural language algorithms alone proves a challenge. In the previous example, the statement is only understood as sarcastic due to our ability to process the inconsistency between the speaker’s words and the prevailing conditions. To better recognize this human ability, it is evident that a multimodal approach is necessary for sarcasm to be detected consistently in artificial intelligence systems.

Given the complexity of detecting sarcasm, this research sets out to investigate a more comprehensive approach. We are focusing on the integration of multiple types of data: text, audio, and visual inputs. The idea is that using more than just words, by considering facial expressions and changes in tone along with intonations, we might lead to better understand and detect sarcasm. Furthermore, the model that doesn't just look at these multimodal features but also considers the overall context of the conversation. We have also proposed an architecture that leverages the power of ensemble techniques to generate more accurate predictions. By using this broader, more inclusive method, the goal of this paper is to make sarcasm detection in artificial intelligence systems more accurate and human-like, essentially bridging the gap between human comprehension and machine interpretation.

## II. LITERATURE REVIEW

Research in the field of sarcasm detection has mainly revolved around textual mode. Text based sarcasm detection have shown promise; however, they disregard the nuances offered by audio and visual data in conversations.

A milestone was set by Castro et al. (2019) where multi-modal sarcasm detection was introduced. The team created MUStARD dataset which is a subset of the EmotionLines (MELD) dataset (Poria et al., 2019). This dataset integrates textual, audio and visual data for detecting sarcasm. The results of their experiments supported the hypothesis that sarcasm is infact multimodal. They conducted multiple evaluations where multimodal models outperformed unimodal models. The main

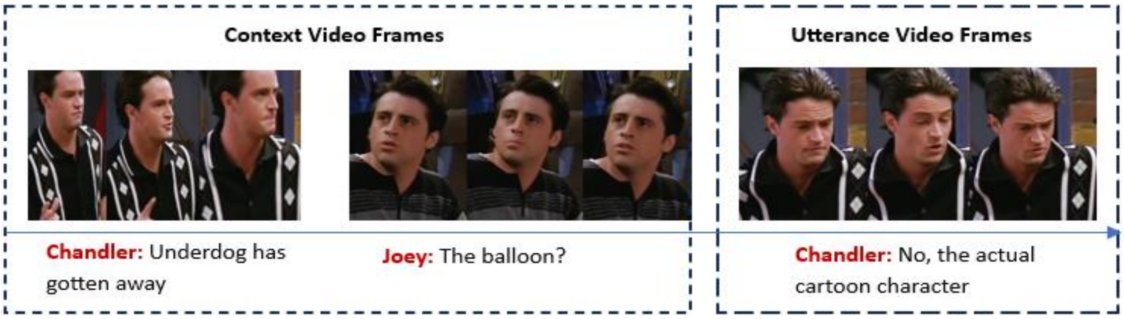

Context Video Frames
Utterance Video Frames
Chandler: Underdog has gotten away
Joey: The balloon?
Chandler: No, the actual cartoon character

Figure 1: An example from dataset (extracted from Castro et al. (2019))

concern in their work is the robustness of the results due to the small size of the MUStARD dataset.

While Castro et al. (2019) set the milestone, subsequent works like that of Chauhan, Ekbal and Bhattacharyya (2020) took a more sophisticated approach by employing multi-task deep learning frameworks. Recognizing the lack of comprehensive data, the authors went a step further by manually annotating the existing dataset with sentiment and emotion labels. The paper introduces two unique attention mechanisms to emphasize the role of context and speaker-specific information in detecting sarcasm. The results revealed that the model had an enhanced performance when data was complemented by emotion and sentiment data. This shows the interdependence of sarcasm, sentiment, and emotion. However, given the small size of the dataset, manually creating new features especially of emotion and sentiment would add bias to the model. Furthermore, the training and tuning of attention models require large datasets, the absence of which questions the reliability and robustness of the model. Their study also does not analyze the role of incongruity across modalities which was later research upon by Wu, Y., Zhao, Y., Lu, X., Qin, B., Wu, Y., Sheng, J. and Li, J. in 2021.

Building upon the MUStARD, a more extensive variant of this dataset, MUStARD++ was later developed. Ray, Mishra, Nunna, and Bhattacharyya (2022) constructed a comprehensive dataset tailored for emotion and sentiment analysis. They rectified several labels and introduced additional metadata including sarcasm types and arousal-valence labels. They significantly increased the size of dataset, affectively doubling it and explored the relationship between sarcasm and emotion, focusing on the type of emotion displayed in sarcasm. Their study also investigated how sarcasm labels could be useful in emotion recognition. The main drawback of this study is that their model depends on the emotion features to detect sarcasm, which is manually annotated. This is not readily available in the real-world scenario and as the previous study, could add bias to the model.

Following the above, Pramanick, Roy and Patel (2022), introduced a new approach named MuLOT. They paved the way for visual feature extraction in sarcasm dataset by using a pretrained action recognition model on the video data and ResNet-101 pretrained on ImageNet. They used self-attention and optimal transport to capture intra-modal and cross-modal dynamics. Their study takes an extra step by introducing multimodal attention fusion technique. This approach further improves the model performance. The results obtained provide strong evidence for the effectiveness of the proposed attention mechanisms and importance of transfer learning. While MuLOT model demonstrates impressive performance, it should be noted that the study primarily focuses on the algorithmic aspects and does not talk about the generalization or limitations of such a model. The feature extraction from video input using the proposed method can be computationally expensive and time consuming. This coupled with algorithm complexity of the attention model could present challenges as mentioned previously. Moreover, the performance improvement, although statistically significant, are relatively modest.

Solving the limitations of the above study, Sun et al. (2022) made notable advancements by addressing the problem of redundant information across video modalities. Their paper presented the EFAFN multimodal sarcasm detection where they used an Efficient Feature Adaptive Fusion Network. The study also reasoned that the background information in a conversation is irrelevant to the outcome, whether sarcasm or not. Hence, they used ResNet pretrained on ImageNet to generate embeddings on faces detected by the facial recognition algorithm provided by DLIB. The study also proposed data augmentation techniques to counteract overfitting during model training. Interestingly, the research showed that the models leveraging only the facial features in the visual data, outperformed those using entire image. This underlines the significance of facial information in sarcasm detection. Despite the positive outcomes, there are areas where the study could be improved or further explored. For instance, while the study does focus on removing irrelevant background information, it does not go into detail about how other modalities like speech and text might contribute to detecting sarcasm. They have over-relied on facial recognition algorithm and this might introduce biases as the audio and text data for the model is still lacking progress. Furthermore, data augmentation techniques, although effective in mitigating overfitting, do not provide solution to the issue of small sized dataset, possibly introducing bias. Finally, although the study indicates facial features are significant, it does fully explore this area because facial recognition algorithms do not generate embeddings that provide information on the feelings or the mood of a person, but rather the uniqueness in the structure of a person's face.

While substantial progress has been made in this field, certain gaps still remain. To address these shortcomings, this study will leverage state-of-the-art transfer learning techniques to generate high-quality embeddings, an approach that has shown considerable promise in the above studies. While transfer learning has been applied in textual and visual inputs, its potential for audio feature extraction remains a mystery. Accordingly, we will compare the performance of the models trained on the traditional acoustic features with that of transfer learning embeddings for audio inputs. Furthermore, for visual data, we will employ a model specifically trained for facial expression recognition. This is expected to yield more insightful embeddings compared to those generated by the facial recognition models used in the previous study. Additionally, given that Late Fusion has shown lower performance compared to Early Fusion (Ding et al., 2022), this study will employ ensemble techniques on Late Fusion predictions and then ensemble these with the Early Fusion predictions to further enhance the performance of the model. This approach not only employs the power of ensembles, but also leverages the latest advancements in individual modal models through state-of-the-art transfer learning techniques.

## III. CONTRIBUTION

**Employment of VGGish Transfer Learning for Audio Data:** A significant gap in existing literature involves the limited exploration of audio features for sarcasm detection. This research bridges this gap by employing the VGGish transfer learning model to generate high-quality audio embeddings. The results indicate that the features derived from the VGGish model outperform the manually extracted audio features, highlighting the efficacy of transfer learning in this context.

**Adoption of a State-of-the-Art Facial Expression Recognition Algorithm:** This study utilizes one of the top-performing models trained on the Facial Expression Recognition (FER) dataset to generate facial embeddings (WuJie et al., 2020). This approach enables the extraction of subtle emotional cues from facial images present in the video data, further enriching the multimodal context for sarcasm detection.

**New Architecture called MFEF: Multimodal Fusion Ensemble Framework:** This is the proposed architecture which leverages the power of ensemble models on early fusion and late fusion.

## IV. FEATURE EXTRACTION

### *A. Features in Text data*

The proposed architecture is designed to process conversational data, distinguishing between the contextual setup leading to a sarcastic comment ('Sentence A') and the concluding utterance, which could be sarcastic or non-sarcastic ('Sentence B'). Inorder to preserve speaker-specific nuances, each sentence is prefixed with speaker name for the speaker-dependent setup.

To extract features from text data, transfer learning methodology is employed, specifically utilizing a case-sensitive BERT (Bidirectional Encoder Representations from Transformers) model. BERT has shown high potential in detection of sarcasm in recent studies (Savini et al., 2020). BERT was also used for sarcasm detection in Arabic text in recent research (Abuteir et al., 2021). The rationale behind choosing case sensitivity lies in the observation that capitalized words in subtitles often serve to emphasize tonal depth or intensity of audio cues, which are commonly used in sarcastic speech. BERT's bidirectional nature also allows it to understand the contextual significance of each word in relation to its surrounding words, thus enhancing semantic understanding.

The architecture generates 768-dimensional embeddings, which are once again concatenated with one-hot encoded labels representing the speaker. This enriches the feature set by incorporating speaker-specific information. The resultant feature set encompasses conversational context, potential sarcastic utterances, and speaker information which would aid in detecting sarcasm.

### *B. Features in Audio data*

The audio data must first be extracted from the associated video files, as it is not directly accessible in the dataset. Given that the sarcasm in speech generally depends on the final utterance sentences in a conversation rather than the contextual

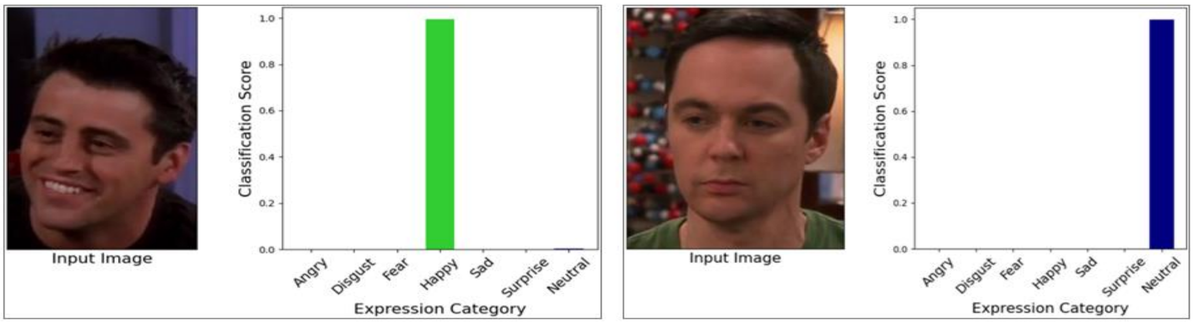


Figure 2: VGG-FER visualization on dataset

speech,only the audios corresponding to these concluding utterances are extracted.

Information about the ending timestamp for each speaker's audio utterance is provided in the dataset. Using this, we create an additional feature called 'START_TIME', representing the timestamp for the beginning of each utterance. This feature combined with the ending timestamp and transcript, enables the calculation of 'Words per Minute' as one of the features.

In terms of acoustic properties, traditional features such as Mel Frequency Cepstral Coefficients (MFCCs), Mel-spectrograms, spectral centroids, and their delta values, along with intensity and zero-crossing rate, are extracted which have been seen relevant in sarcasm detection (Cheang et al., 2008). Since these features are frame-specific, their average values are calculated to serve as representative metrics. To capture more nuanced information, standard deviations and medians of these features are also included. These features were also used by Zheng Lin Chia, Ptaszynski, M., Masui, F., Gniewosz Leliwa and M Wroczynski in 2021, to detect and classify cyberbullying in online game voice chats.

Furthermore, the architecture also employs a transfer learning approach, utilizing the VGGish model to obtain rich embeddings using the audio data. The decision to employ VGGish over other audio feature extraction models was multi-faceted. VGGish is trained on a larger dataset compared to other audio transfer learning models. This allows for it to capture subtle cues in audio that are not captured by traditional audio feature extraction methods. The robustness of the VGGish model, as evidenced by its successful application in numerous audio classification tasks, also promises a high level of reliability (plakal, 2023).

### *C. Features in Video data*

The dataset includes two types of video data: contextual videos and utterance videos. Contextual videos capture the lead-up to the sarcastic utterance, whereas the utterance videos focus solely on the moment sarcasm is delivered. For the purpose of this study, we concentrate on the latter, since the facial expressions that accompany the utterance are generally more indicative of sarcasm.

To extract facial features, faces in the utterance videos are identified using the Multi-Task Cascaded Convolutional Networks (MTCNN) provided by DLIB library. A transfer learning model based on VGG-Face for Facial Expression Recognition (VGG-FER) is then employed to generate embeddings for these faces (WuJie, 2020). The visualization of this model can be seen in figure 2. Given that facial expressions across a given room are likely to be similar, these embeddings are averaged over time.

In addition to these general embeddings, we generate a separate set of images that are speaker-dependent. Here, only the face of the speaker delivering the sarcastic or non-sarcastic utterance is considered. Faces are manually cropped from the videos using labelImg software (Tzutalin et al., 2022), and embeddings are generated using a pretrained VGG-FER model.

## V. EXPERIMENTAL SETUPS

To investigate the role of the speaker in sarcasm versus its inherent nature, two experimental setups are employed. One focuses on Speaker-Dependent sarcasm data and the other on Speaker-Independent sarcasm data. As mentioned earlier, the latter will not contain any information regarding the information of the speaker. The results of these experiments will provide insight into the nature of sarcasm.

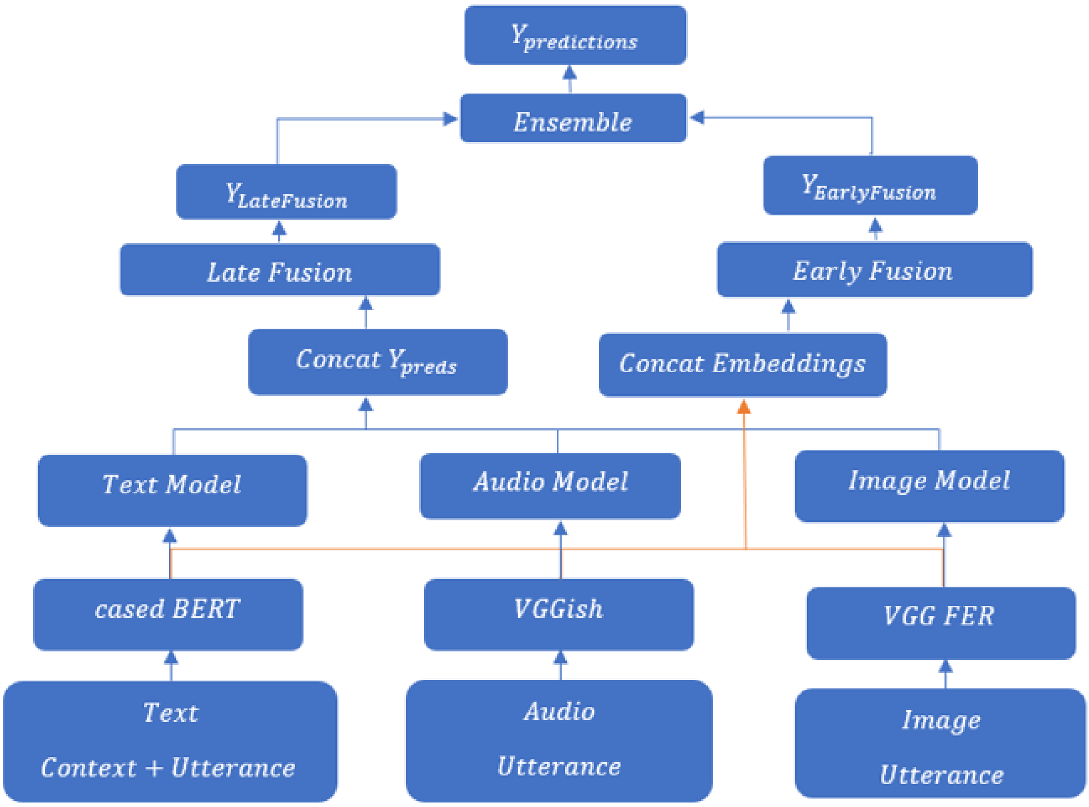


Figure 3. Architecture of MFENet

## VI. NOTATIONS

The notations in section 6 are described in Table 1 for ease of use for the reader.

TABLE I: NOTATIONS

| | Notation | Description |
|---|---|---|
| Embeddings | w | Word |
| | a | Audio |
| | i | Image |
| Inputs | $T_u$ | Utterance Text |
| | $T_c$ | Context Text |
| | $A_u$ | Utterenace Audio |
| | $I_u$ | Utterance Image |
| | S | Speaker information |
| Models | $T_m$ | Text |
| | $A_u$ | Audio |
| | $I_u$ | Image |
| | $F_{early}$ | Early Fusion |
| | $F_{late}$ | Late Fusion |
| | E | Ensemble |

## VII. FRAMEWORK

This section describes the proposed architecture of the framework, which incorporates multiple layers and models explained in the subsections below.

### *A. Textual Model*

For speaker-dependent textual model, embeddings are generated using a case-sensitive BERT model. These embeddings represent Speaker information (S), Contextual Text ($T_c$) and Utterance Text ($T_u$). These BERT embeddings are then concatenated with speaker names and to serve as input for textual model.

$$\{w_1, w_2, w_3 \dots w_{768}\} = \mathrm{BERT}(S + T_c + T_u) \quad (1)$$

$$T_m = \mathrm{CAT}(\{w_1, w_2 \dots w_{768}\}, \{S_1, S_2 \dots S_{27}\}) \quad (2)$$

For speaker-independent setup, the speaker information is omitted from the inputs. The speaker-independent model for textual data also does not contain speaker information.

$$\{w_1, w_2, w_3 \dots w_{768}\} = \mathrm{BERT}(T_u + T_c) \quad (3)$$

$$T_m = \{w_1, w_2, w_3 \dots w_{768}\} \quad (4)$$

### *B. Audio Model*

VGGish, developed by google is employed for generating audio embeddings resulting in a 128-dimensional audio vector for each audio file. These embeddings are then concatenated with speaker names to form the input for the audio model. This concatenation allows for the capture of speaker-specific patterns.

$$\{a_1, a_2, a_3 \dots a_{128}\} = \mathrm{VGGish}(A_u) \qquad (5)$$

$$A_m = \mathrm{CONCAT}(\{a_1, a_2 \dots a_{128}\}, \{S_1, S_2 \dots S_{27}\}) \quad (6)$$

For speaker-independent setup, the speaker information is not included.

$$A_m = \{a_1, a_2 \dots a_{128}\} \qquad (7)$$

We also extracted the traditional acoustic features mentioned in previous sections, resulting in a 253-dimensional vector for each audio sample. Upon evaluating models using these features, it became clear that they did not contribute meaningfully to the model's performance. In contrast, embeddings generated through pre-trained models like VGGish consistently demonstrated superior performance.

### *C. Visual Model*

The image consists of the face of the speaker. VGG model trained on Facial Expression Recognition (FER) dataset is used to generate embeddings (WuJie, 2020). These embeddings are then concatenated with speaker names. This allows the model to understand speaker specific facial expressions.

$$\{i_1, i_2, i_3 \dots i_{256}\} = \mathrm{VGG\ FER}(I_u) \qquad (8)$$

$$I_m = \mathrm{CONCAT}(\{i_1, i_2 \dots i_{256}\}, \{S_1, S_2 \dots S_{27}\}) \quad (9)$$

For the speaker-independent setup, embeddings are generated for all faces in the video and averaged since the facial expressions are generally similar across participants.

$$\{i_1, i_2, i_3 \dots i_{256}\} = \mathrm{VGG\ FER}(I_u + I_c) \qquad (10)$$

$$I_m = \{i_1, i_2, i_3 \dots i_{256}\} \qquad (11)$$

### *D. Early Fusion Model*

The embeddings generated from the Textual, Audio, Visual models are concatenated and used as the input for the Fusion model.

$$F_{early} = \mathrm{CONCAT}(\{w_{1:768}\}, \{a_{1:128}\}, \{i_{1:256}\}) \quad (12)$$

The Early Fusion model structure remains the same for both experimental setups.

### *E. Late Fusion Model*

The predictions generated by the Textual, Audio and Visual models are concatenated and used as the input for a logistic regression model. The Late Fusion model structure also remains the same for both experimental setups.

$$F_{late} = \mathrm{CONCAT}(T_m, A_m, I_m) \qquad (13)$$

### *F. Ensemble Model*

The early fusion and late fusion techniques are similar to ensemble techniques, the only difference being that the inputs to these models arise from different modality of data. The ensemble model averages the probability outputs from the Early and Late Fusion models.

$$E = \frac{F_{early} + F_{late}}{2} \qquad (14)$$

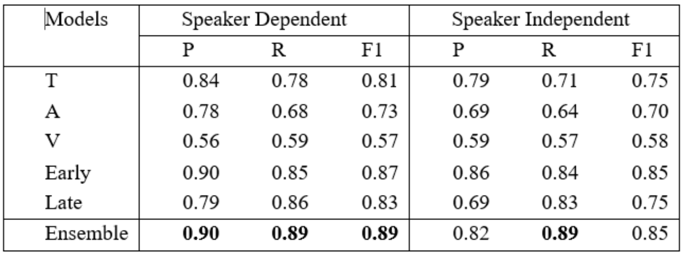
The threshold for classification is tuned for optimal performance. The structure of the ensemble model remains same across both experimental setups.

## VIII. EXPERIMENTAL FINDINGS

### *A. Analysis of Results*

Our experimental evaluations, presented in Table 2, provide valuable insights into the performance of the proposed models.

TABLE II: EXPERIMENTAL RESULTS



| Models | Speaker Dependent | | | Speaker Independent | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| T | 0.84 | 0.78 | 0.81 | 0.79 | 0.71 | 0.75 |
| A | 0.78 | 0.68 | 0.73 | 0.69 | 0.64 | 0.70 |
| V | 0.56 | 0.59 | 0.57 | 0.59 | 0.57 | 0.58 |
| Early | 0.90 | 0.85 | 0.87 | 0.86 | 0.84 | 0.85 |
| Late | 0.79 | 0.86 | 0.83 | 0.69 | 0.83 | 0.75 |
| Ensemble | **0.90** | **0.89** | **0.89** | 0.82 | **0.89** | 0.85 |

The experimental results suggests that final ensemble model produces the highest performance compared to the unimodal and fusion models, showing a 2% performance increase relative to the second-best performing model (Early Fusion).

Comparing the results of the two experimental setups, it is evident that even though speaker-dependent setup has an overall increase of 3% in performance. While this might initially seem to suggest that speaker information is important, the small magnitude of the difference argues against this interpretation. Rather, the data implies that sarcasm is largely a speaker-independent phenomenon.

### *B. Comparison of Results*

Table 3 offers a comparative analysis of our work against existing studies. Given that some previous research does not account for speaker-independence, we focus solely on speaker-dependent results for this comparison.

TABLE III: COMPARISON OF RESULTS

| **Methods** | **P** | **R** | **F1** |
|---|---|---|---|
| (Chauhan et al., 2020) | 73.40 | 72.75 | 72.5 |
| (Ray et al., 2022) | 74.2 | 74.2 | 74.2 |
| (Chauhan et al., 2022) | 77.9 | 76.9 | 76.7 |
| (Sun et al., 2022) | 79.0 | 79.0 | 79.0 |
| (Sun et al., 2022) – ViViT | 69.3 | 88.4 | 77.7 |
| MFEF - Ours | **90.2** | **89.2** | **89.7** |
| **Performance Increase** | **11.2%** | **0.8%** | **10.7%** |

Our MFEF model excels with a noticeable performance increase of up to 11.2%. It should be highlighted that our model benefits from the larger, more robust MUStARD++ dataset, which is a superset of MUStARD dataset, and also from the state-of-the-art feature extraction techniques across all modalities of input, which collectively contribute to its standout performance.

## IX. CONCLUSION AND FUTURE WORK

In this research paper, we introduced MFEF, a multimodal architecture designed to identify sarcasm in conversations. We used a bigger dataset (MUStARD++) than most previous studies and by took advantage of transfer learning techniques to improve the model performance. Our Ensemble model emerged as the most effective, outperforming unimodal and fusion models. Most notably, the model achieved superior results in comparison with existing research in this domain, setting a new standard for detecting sarcasm using multimodal data.

The experimental results also seemed to support the possibility that sarcasm is mostly a speaker-independent phenomenon. However, this cannot be claimed as of now because the robustness of the model also remains questionable due to the limited size of the dataset.

Despite the promising results, there are several areas for improvement and further exploration. One key limitation is the modest dataset size. Even though we use larger dataset than most previous studies, it is still not nearly enough to be used for training the neural networks. During the experimental phase, we found that neural networks excelled in individual modalities, sometimes even outperforming the final ensemble model. However, this was only seen on the training data. The model suffered from overfitting. Enlarging the dataset could alleviate this issue, leading to a more reliable and precise model. Future research should concentrate on expanding the dataset to enhance the model's reliability and robustness, as well as investigating the viability of real-time sarcasm detection. Additionally, if new multi-modal transfer learning models come out, especially ones trained on large conversational datasets, it could mark a new milestone in this domain.

## X. References:

Abuteir, M.M. and Eltyeb S. A. Elsamani (2021). *Automatic Sarcasm Detection in Arabic Text: A Supervised Classification Approach*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/354054553_Automatic_Sarcasm_Detection_in_Arabic_Text_A_Supervised_Classification_Approach [Accessed 25 Jun. 2023].

Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R. and Poria, S. (2019). *Towards Multimodal Sarcasm Detection (An Obviously Perfect Paper)*. [online] pp.4619–4629. Available at: https://aclanthology.org/P19-1455.pdf.

Chauhan, D., Ekbal, A. and Bhattacharyya, P. (2020). *Sentiment and Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis*. pp.4351–4360.

Chauhan, D.S., Singh, G.V., Arora, A., Ekbal, A. and Bhattacharyya, P. (2022). An emoji-aware multitask framework for multimodal sarcasm detection. *Knowledge-Based Systems*, 257, p.109924. doi:https://doi.org/10.1016/j.knosys.2022.109924.

Cheang, H.S. and Pell, M.D. (2008). The sound of sarcasm. [online] 50(5), pp.366–381. doi:https://doi.org/10.1016/j.specom.2007.11.003.

Ding, N., Tian, S. and Yu, L. (2022). A multimodal fusion method for sarcasm detection based on late fusion. *Multimedia Tools and Applications*, 81(6), pp.8597–8616. doi:https://doi.org/10.1007/s11042-022-12122-9.

Jorgensen, J.C. (1996). The functions of sarcastic irony in speech. [online] 26(5), pp.613–634. doi:https://doi.org/10.1016/0378-2166(95)00067-4.

plakal (2023). VGGish [online] GitHub. Available at: https://github.com/tensorflow/models/blob/master/research/audioset/vggish/README.md [Accessed 10 Aug. 2023].

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E. and Mihalcea, R. (2019). *MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations*. [online] Available at: https://arxiv.org/pdf/1810.02508.pdf.

Pramanick, S., Roy, A. and Patel, V.M. (2021). *Multimodal Learning using Optimal Transport for Sarcasm and Humor Detection*. [online] arXiv.org. Available at: https://arxiv.org/abs/2110.10949 [Accessed 30 Jun. 2023].

Ray, A., Mishra, S., Nunna, A. and Bhattacharyya, P. (2022). *A Multimodal Corpus for Emotion Recognition in Sarcasm*. [online] Available at: https://arxiv.org/pdf/2206.02119v1.pdf [Accessed 25 Jun. 2023].

Santosh Kumar Bharti, Gupta, R.K., Prashant Kumar Shukla, Wesam Atef Hatamleh, Tarazi, H. and Stephen Jeswinde Nuagah (2022). Multimodal Sarcasm Detection: A Deep Learning Approach. *Wireless Communications and Mobile Computing*, [online] 2022, pp.1–10. doi:https://doi.org/10.1155/2022/1653696.

Savini, E. and Caragea, C. (2022). Intermediate-Task Transfer Learning with BERT for Sarcasm Detection. [online] 10(5), pp.844–844. doi:https://doi.org/10.3390/math10050844.

Sun, Y., Zhang, H., Yang, S. and Wang, J. (2022). EFAFN: An Efficient Feature Adaptive Fusion Network with Facial Feature for Multimodal Sarcasm Detection. *Applied Sciences*, 12(21), p.11235. doi:https://doi.org/10.3390/app122111235.

Tzutalin (2022). *labelImg*. [online] GitHub. Available at: https://github.com/HumanSignal/labelImg/tree/master [Accessed 22 Jun. 2023].

Wu, Y., Zhao, Y., Lu, X., Qin, B., Wu, Y., Sheng, J. and Li, J. (2021). Modeling Incongruity between Modalities for Multimodal Sarcasm Detection. *IEEE MultiMedia*, 28(2), pp.86–95. doi:https://doi.org/10.1109/mmul.2021.3069097.

WuJie (2020). *WuJie1010/Facial-Expression-Recognition.Pytorch: A CNN based pytorch implementation on facial expression recognition (FER2013 and CK+), achieving 73.112% (state-of-the-art) in FER2013 and 94.64% in CK+ dataset*. [online] GitHub. Available at: https://github.com/WuJie1010/Facial-Expression-Recognition.Pytorch [Accessed 12 Aug. 2023].

Zheng Lin Chia, Ptaszynski, M., Masui, F., Gniewosz Leliwa and M Wroczynski (2021). Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection. [online] 58(4), pp.102600–102600. doi:https://doi.org/10.1016/j.ipm.2021.102600.