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 the field of communication such as therapy, retail, customer service, reception desks etc. However, despite the significant progress, there is a lack of research that explores the extent to which AI can be used to distinguish passive aggression or sarcasm. Thus, the aim of this project is to investigate the potential of AI in detecting this subtle human expressiveness. This paper introduces an innovative framework that combines ensemble methods with late fusion and early fusion techniques. Using two distinct experimental approaches and an optimized data-splitting strategy, the study demonstrates a 10% performance improvement over existing methods at the ensemble stage. Additionally, the paper investigates the nature of sarcasm, and the findings unveil an intriguing truth – the sarcasm is, in fact, speaker-independent.**

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

1. 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, 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.

2. Related Work:

2.1 Multi-modal sarcasm dataset

While much of the research on sarcasm detection centers around text-based datasets, the concept of a multi-modal approach in this domain is relatively new. However, developing multi-modal sarcasm datasets presents significant challenges, followed by concerns surrounding piracy and revenue generation, resulting in a scarcity of research in this specific area. In 2019, a notable step was taken with the creation of Multimodal Sarcasm Detection Dataset (MUStARD) (Castro et al., 2019) which is a subset of Multimodal Extension of textual dataset

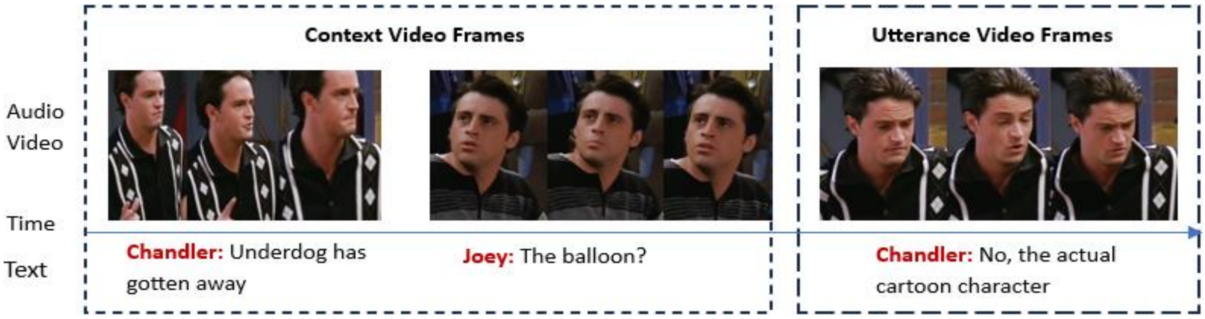

Figure 1: Sample of dataset

EmotionLines (MELD) dataset (Poria et al., n.d.). The size of the MUStARD dataset is very small given the difficulty to create such data, however a variant known as MUStARD++ was developed and offers double the amount of data and additional features (Ray et al., n.d.). This study will leverage the MUStARD++ dataset to further explore and advance the field of multi-modal sarcasm detection.

2.2 Text-focused detection

Twitter and reddit often serve as the primary source of data for sarcasm detection and the datasets are typically created either through manual annotation or through the utilization of hashtags (Bamman & A. Smith. 2015; Amir et al., 2016). The approaches for sarcasm detection in text typically fall into three categories: Rule-based methods, Statistical methods, and Deep learning methods (Zheng Lin Chia et al., 2021). Among these, deep learning methods have emerged consistently, yielding significantly improved results (Zhou, Zhang and Wu, 2020; Savini and Caragea, 2022).

2.3 Sarcasm detection in speech

Sarcasm detection in speech is a complex task due to the intricacies and nuances of human communication, necessitating the interpretation of both verbal and non-verbal cues. These cues include factors such as mean fundamental frequency, standard deviation, amplitude characteristics, speech rate, harmonics-to-noise ratio, and one-third octave spectral values (Cheang and Pell, 2008). Building on advancements in deep learning and multi-modal processing, recent research has explored the combination of acoustic and linguistic information for sarcasm detection (Liang et al., 2018). Looking at the research that's been done, there exists a noticeable gap on detecting sarcasm through speech alone. This underexplored area in presents a compelling opportunity with the potential to significantly enrich our understanding of sarcasm in speech.

2.3 Multi-modal detection

Multimodal sarcasm detection is a relatively novel research area. Castro et al. (2019) pioneered this field, emphasizing the integration of audiovisual and textual cues to boost sarcasm detection efficacy. Their work underscored the potential of multimodal data in revealing the complex linguistic phenomenon of sarcasm. Chauhan, Ekbal, and Bhattacharyya (2020) further enriched the field by developing a multi-task deep learning framework. By simultaneously interpreting sarcasm, sentiment, and emotion, they emphasized the interrelation of these elements, broadening the scope of sarcasm detection methodologies.

More recent advancements in the field have focused on examining the incongruity between modalities. Wu et al. (2021) developed the Incongruity-aware Attention Network (IWAN), which scrutinizes word-level incongruity for sarcasm detection. Pramanick, Roy, and Patel (2021) proposed the Multimodal Learning using Optimal Transport (MuLOT), a system utilizing self-attention and optimal transport mechanisms to identify cross-modal correspondence.

Following this, Sun et al. (2022) advocated for the close examination of facial expressions in images, arguing that they carry pivotal emotional cues. This approach highlighted the essentiality of non-verbal cues in sarcasm detection, though the reliance on clear facial imagery may limit its application. Chauhan et al. (2022) proposed incorporating emojis into textual data. While their emoji-aware multimodal approach displayed effectiveness, its real-world applicability may be restricted considering emojis are seldom used in verbal communication.

Lastly, Ding, Tian, and Yu (2022) presented a multi-level late-fusion learning framework with residual connections for sarcasm detection, offering an innovative fusion model that has demonstrated superior performance. This approach offers a glimpse into the promising potential of advanced fusion techniques in this domain.

2.4 Contribution:

The main contributions of this paper are:

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 not only employing advanced audio features but also leveraging the VGGish transfer learning model to generate 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, 2020). This approach enables the extraction of sublte emotional cues from visual 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 with fusion techniques.

3. Feature Extraction:

3.1 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'). To preserve speaker-specific nuances, each sentence is prefixed with speaker name.

To extract features from text data, transfer learning methodology is employed, specifically utilizing a case-sensitive BERT (Bidirectional Encoder Representations from Transformers) model. The rationale behind choosing case sensitivity lies in the observation that capitalized words in subtitles often serve to emphasize tonal depth or intensity. BERT's bidirectional nature 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 then concatenated with one-hot encoded labels representing the speaker. This further enriches the feature set by incorporating speaker-specific information. The resultant comprehensive feature set encompasses conversational context, potential sarcastic utterances, and speaker information which facilitates more nuanced binary classification of sarcasm.

3.2 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 sentence in a conversation rather than the contextual setup, 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 introduce an additional feature called 'START_TIME', representing the timestamp for the beginning of each utterance. This feature combined with the ending timestamp and transcript data, enables the calculation of 'Words per Minute' as one of the features.

In terms of acoustic properties, 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. 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.

Furthermore, the architecture employs a transfer learning approach, utilizing the VGGish model to obtain rich embeddings for the audio features. The decision to employ VGGish over other audio feature extraction models was multi-faceted. VGGish's architecture is optimized for capturing a wide array of audio features, which is essential for understanding the tonal complexities associated with sarcasm. The robustness of the VGGish model, as evidenced by its successful application in numerous audio classification tasks, promises a high level of reliability (plakal, 2023).

3.3 Features in Video data:

The dataset includes two types of videos: context videos and utterance videos. Context videos capture the lead-up to the sarcastic utterance, whereas utterance videos focus solely on the moment the 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). A transfer learning model based on VGG-Face for Facial Expression Recognition (VGG-FER) is then employed to generate embeddings for these faces. 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 that is 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, 2022), and embeddings are generated using a pretrained VGG-FER model. This set of features is specifically used for experiments that focus on speaker-dependent sarcasm detection.

4 Experimental Setups:

To investigate the role of the speaker in sarcasm versus its inherent nature, two experimental setups are employed: one focusing on Speaker-Dependent sarcasm and the other on Speaker-Independent sarcasm. The results of these experiments will provide insight into the nature of sarcasm.

5 Notations:

The notations in section 6 are described in Table 1 for reader convenience.

	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

Table 1: Notations

6. Framework:

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

6.1 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).

$$\{\mathrm{w}_1, \mathrm{w}_2, \mathrm{w}_3 \ldots \mathrm{w}_{768}\} = \mathrm{BERT}(\mathrm{S} + \mathrm{T_c} + \mathrm{T_u})$$

These BERT embeddings are then concatenated with speaker names and to serve as input for textual model.

$$\mathrm{T_m} = \mathrm{CONCAT}(\{\mathrm{w}_1, \mathrm{w}_2 \ldots \mathrm{w}_{768}\}, \{\mathrm{S}_1, \mathrm{S}_2 \ldots \mathrm{S}_{27}\})$$

For speaker-independent setup, the speaker information is omitted from the inputs.

$$\{\mathrm{w}_1, \mathrm{w}_2, \mathrm{w}_3 \ldots \mathrm{w}_{768}\} = \mathrm{BERT}(\mathrm{T_u} + \mathrm{T_c})$$

$$\mathrm{T_m} = \{\mathrm{w}_1, \mathrm{w}_2, \mathrm{w}_3 \ldots \mathrm{w}_{768}\}$$

6.2 Audio Model:

VGGish, developed by google is employed for generating audio embeddings resulting in a 128 dimensional audio vector for each audio file.

$$\{\mathrm{a}_1, \mathrm{a}_2, \mathrm{a}_3 \ldots \mathrm{a}_{128}\} = \mathrm{VGGish}(\mathrm{A_u})$$

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.

$$\mathrm{A_m} = \mathrm{CONCAT}(\{\mathrm{a}_1, \mathrm{a}_2 \ldots \mathrm{a}_{128}\}, \{\mathrm{S}_1, \mathrm{S}_2 \ldots \mathrm{S}_{27}\})$$

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

$$\mathrm{A_m} = \{\mathrm{a}_1, \mathrm{a}_2 \ldots \mathrm{a}_{128}\}$$

We also manually extracted acoustic features, resulting in a 253-dimensional vector for each audio sample. Upon evaluating various 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 efficacy.

6.3 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).

$$\{\mathrm{i}_1, \mathrm{i}_2, \mathrm{i}_3 \ldots \mathrm{i}_{256}\} = \mathrm{VGG\ FER}(\mathrm{I_u})$$

These embeddings are then concatenated with speaker names. This allows the model to understand speaker specific facial expressions.

$$\mathrm{I_m} = \mathrm{CONCAT}(\{\mathrm{i}_1, \mathrm{i}_2 \ldots \mathrm{i}_{256}\}, \{\mathrm{S}_1, \mathrm{S}_2 \ldots \mathrm{S}_{27}\})$$

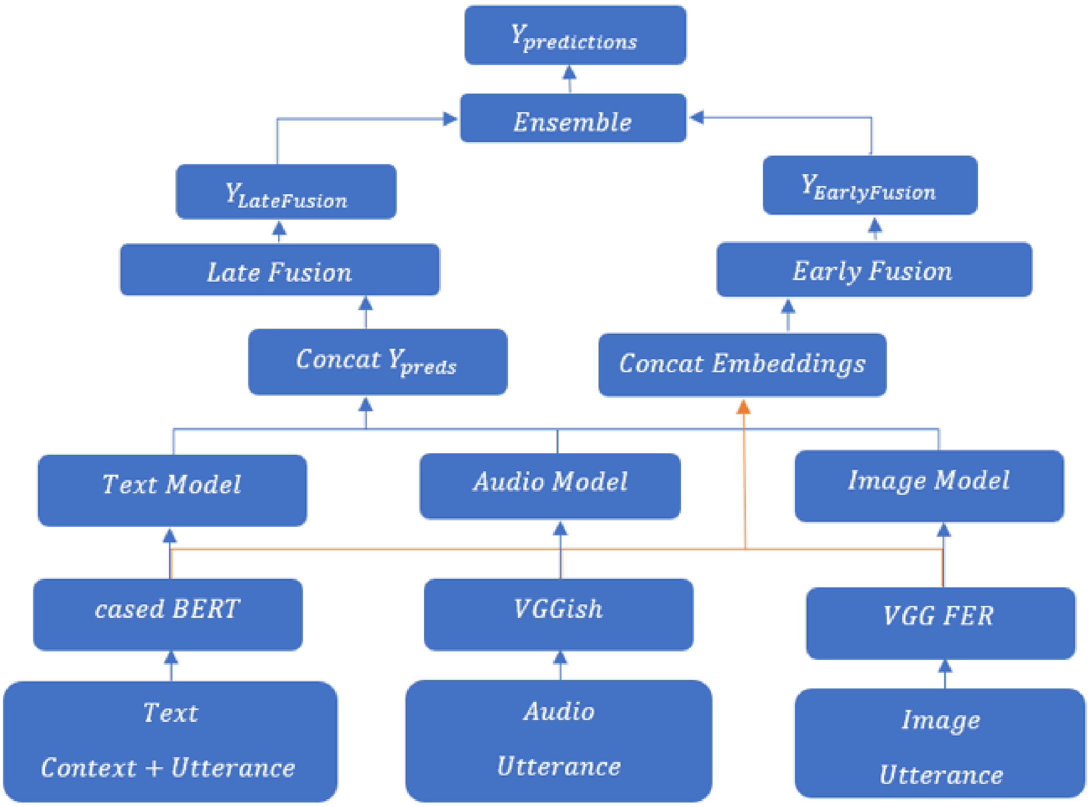

Figure 1. Architecture of MFENet

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.

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

$$I_m = \{i_1, i_2, i_3 \dots i_{256}\}$$

6.4 Early Fusion Model:

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

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

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

6.5 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.

$$F_{late} = \text{CONCAT}(T_m, A_m, I_m)$$

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

6.6: Ensemble Model:

The ensemble model averages the probability outputs from the Early and Late Fusion models.

$$E = \frac{F_{early} + F_{late}}{2}$$

The threshold for classification is tuned for optimal performance. The Ensemble model structure remains consistent across both experimental setups.

7. Experimental Findings:

7.1 Analysis of Results:

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

Models	Speaker-Dependent				Speaker-Independent			
	A	P	R	F1	A	P	R	F1
T	0.79	0.84	0.78	0.81	0.71	0.79	0.71	0.75
A	0.69	0.78	0.68	0.73	0.63	0.69	0.64	0.70
V	0.56	0.56	0.59	0.57	0.55	0.59	0.57	0.58
Early	0.86	0.90	0.85	0.87	0.84	0.86	0.84	0.85
Late	0.80	0.79	0.86	0.83	0.71	0.69	0.83	0.75
Ensemble	**0.88**	**0.90**	**0.89**	**0.89**	0.85	0.82	**0.89**	0.85

Table 2: Experimental Results

The experimental results suggests that final ensemble model produces the highest performance compared to other 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.

It's crucial to acknowledge that the constrained dataset size introduces an approximate 5% margin of error in our results. To make more definitive claims about sarcasm's characteristics, a more extensive dataset is necessary.

7.2 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.

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%**

Table 3: Comparison of Results

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 and state-of-the-art feature extraction techniques, which collectively contribute to its standout performance.

8. Conclusion and Future Work:

In this research paper, we introduced MFEF, a multi-modal architecture designed to identify sarcasm in conversational data. By leveraging advanced feature extraction techniques and transfer learning, our model integrates textual, auditory, and visual inputs to make nuanced determinations about the presence of sarcasm. Our Ensemble model emerged as the most effective, outperforming all other models and showing resilience to speaker variations. The model achieved superior results in comparison with existing research, particularly in a speaker-dependent context

Despite the promising results, there are several areas for improvement and further exploration. One key limitation is the modest dataset size, which restricts the model's overall performance. During the experimental phase, we found that neural networks excelled on individual input types and even outperformed the final ensemble model, but suffered from significant 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 classification datasets, it could mark a new milestone in this domain.

References:

1. 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].

2. Amir, S., Wallace, B., Lyu, H., Carvalho, P. and Silva, M. (2016). *Modelling Context with User Embeddings for Sarcasm Detection in Social Media*. [online] Available at: https://arxiv.org/pdf/1607.00976.pdf.

3. Bamman, D. and A. Smith, N. (2015). *Contextualized Sarcasm Detection on Twitter*. [online] Aaai.org. Available at: https://ojs.aaai.org/index.php/ICWSM/article/view/14655/14504 [Accessed 25 Jun. 2023].

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

9. Ghosh, A. and Veale, T. (2016). *Fracking Sarcasm using Neural Network*. [online] Association for Computational Linguistics, pp.161–169. Available at: https://aclanthology.org/W16-0425.pdf.

10. 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.

11. Joshi, A., Sharma, V. and Bhattacharyya, P. (2015). *Harnessing Context Incongruity for Sarcasm Detection*. [online] Association for Computational Linguistics, pp.757–762. Available at: https://aclanthology.org/P15-2124.pdf.

12. Joshi, A., Tripathi, V., Bhattacharyya, P. and Carman, M. (2016). *Harnessing Sequence Labeling for Sarcasm Detection in Dialogue from TV Series 'Friends'*. [online] Association for Computational Linguistics, pp.146–155. Available at: https://aclanthology.org/K16-1015.pdf [Accessed 25 Jun. 2023].

13. Jyoti Godara, Batra, I., Aron, R. and Shabaz, M. (2021). Ensemble Classification Approach for Sarcasm Detection. [online] 2021, pp.1–13. doi:https://doi.org/10.1155/2021/9731519.

14. Kumar, A., Shubham Dikshit and Gupta, D. (2021). Explainable Artificial Intelligence for Sarcasm Detection in Dialogues. [online] 2021, pp.1–13. doi:https://doi.org/10.1155/2021/2939334.

15. Liang, P.P., Liu, Z., Zadeh, A. and Morency, L.-P. (2018). *Multimodal Language Analysis with Recurrent Multistage Fusion*. [online] arXiv.org. Available at: https://arxiv.org/abs/1808.03920 [Accessed 25 Jun. 2023].

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

17. 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].

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

19. Rolandos Alexandros Potamias, Georgios Siolas and Stafylopatis, A. (2020). A transformer-based approach to irony and sarcasm detection. [online] 32(23), pp.17309–17320. doi:https://doi.org/10.1007/s00521-020-05102-3.

20. 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.

21. 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.

22. 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.

23. Zhang, Y., Liu, Y., Li, Q., Joel, Wang, B., Li, Y., Hari Mohan Pandey, Zhang, P. and Song, D. (2021). CFN: A Complex-Valued Fuzzy Network for Sarcasm Detection in Conversations. [online] 29(12), pp.3696–3710. doi:https://doi.org/10.1109/tfuzz.2021.3072492.

24. 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.

25. Zhou, Q., Zhang, Z. and Wu, H. (n.d.). NLP at IEST 2018: BiLSTM-Attention and LSTM-Attention via Soft Voting in Emotion Classification. [online] doi:https://doi.org/10.18653/v1/P17.