Emotion Recognition in Sarcastic Code-Mixed (Hinglish) Conversations

Ishita Rathore, Suhaani Sachdeva and Anjelica

Abstract

Emotion Recognition in Sarcastic, Code-Mixed (Hinglish) conversational data lies at the niche end of the field of Affective Computing. While the same problem in English (only) datasets has been explored, not much has been done in the code-mixed aspect. Through this project, we aim to address this missing avenue by building on top of an already existing Code-Mixed dataset, WITS (based on data collected from the Indian television program - Sarabhai vs Sarabhai). We labeled the WITS dataset manually, on the majority emotion leading to the use of sarcasm in a conversation. The emotions are of 4 types - Anger, Surprise, Ridicule and Sad. Following this, we used FastText text model, and RESNET-50 to give us text (100-D) and video (2048-D) embeddings respectively. Additionally, we also explored creating video embeddings using our custom CNN. These embeddings (FastText and ResNet-50) have been primarily passed through two models - Random Forest Classifier, which gave us an overall weighted average F1-score of 0.53 and SNN-AFM (Shallow Neural Network relying on Attention based Fusion Mechanism) following the Attention-based Fusion mechanism which combines text and video features by dynamically assigning attention weights to each of the modalities. Finally, a Classifier which is the final linear layer in the network, is used to output emotion class probabilities. This approach gave us a weighted F1-score of 0.51. Both these models outperform the nearest SOTA (Emotion Recognition for English language only) which is weighted F1-score of ~0.42. In summary, out of the two models used, Random Forest Classifier gives us the best performance, on recognizing emotions in code-mixed (Hinglish) data.

Introduction

Recognizing the underlying emotion in sarcastic, code-mixed (Hinglish) conversations is a nuanced task that is essential for building culturally aware AI systems. Sarcasm often makes it difficult for models to correctly decode the meaning of sentences, masking the speaker's true intent. When combined with code-mixing, the alternation between languages within a single conversation, the complexity deepens significantly.

Research in natural language processing has explored sarcasm and emotion detection extensively in monolingual English settings (Kumar et al., 2022). Sarcasm is a sophisticated linguistic phenomenon where surface meaning and underlying intent often diverge. Understanding sarcasm requires not just textual interpretation but also fine-grained analysis of audio-visual cues such as facial expressions, gestures, tone, and intonation. However, emotion recognition in sarcastic, code-mixed conversations remains largely unexplored, primarily due to the lack of specialized datasets and benchmark models.

To address this gap, we introduce WITS++, a manually annotated dataset designed for emotion recognition in sarcastic Hinglish dialogues. WITS++ is derived from the WITS corpus, which contains dialogues from the Indian sitcom Sarabhai vs Sarabhai. Using this dataset, we train and evaluate two baseline models for emotion recognition. Our contributions include: (1) WITS++, a manually annotated dataset for emotion recognition in sarcastic, code-mixed (Hinglish) conversations; (2) establishment of baseline performance on WITS++ across four emotion classes—Ridicule, Surprise, Anger, and Sad—using a Random Forest classifier and an attention-based neural network (SNN-AFM); and (3) an initial analysis highlighting the challenges of this novel task, particularly the significant class imbalance and the difficulty of recognizing underrepresented emotions in sarcastic contexts.

Background and Significance

The problem this paper addresses, pertains to Emotion Recognition in Sarcastic Code-Mixed (Hindi-English) Conversations by making use of multi-modal data i.e., Videos and Text transcripts. Emotion

Recognition in Conversations (ERC) is crucial in developing sympathetic human-machine interaction. In conversational videos, emotion can be present in multiple modalities, i.e., audio, video, and transcript. However, due to the inherent characteristics of these modalities, multi-modal ERC has always been considered a challenge to be tackled (Chudasama et al,. 2022). Moreover, an additional layer is recognizing emotions in conversations that are sarcastic and code-mixed. Indirect speech such as sarcasm achieves a constellation of discourse goals in human communication. While the indirectness of figurative language warrants speakers to achieve certain pragmatic goals, it is challenging for AI agents to comprehend such idiosyncrasies of human communication (Kumar et al,. 2022). Though sarcasm identification has been a well-explored topic in dialogue analysis now, for conversational systems to truly grasp a conversation's innate meaning and generate appropriate responses, simply detecting sarcasm is not enough; it is vital to recognize the emotion that majoritively occurs in a sarcastic conversation, most often leading to the use of sarcasm.

Therefore, this paper explores the aspect of recognizing the majority emotion in sarcastic code-mixed conversations through the video and text modalities. Emotion Recognition can lead to deeper insights for the conversational systems in understanding the intent of the speaker and since the current systems are only based on the text modality, adding the video modality will tune the systems more towards understanding humans better and giving correctly curated responses. This aspect has been completely unexplored in code-mixed (Hinglish) data, hence to the best of our knowledge, our model will be the first to step into this field for languages other than English, to provide better service to people belonging to different regions of the world.

Literature Review

The scope of this research resides at the intersection of several key disciplines - Natural Language Processing, Affective computing and Multimodal machine learning. We revolve around recognizing emotions in sarcastic code-mixed dialogues, therefore multilingual NLP is an emerging subarea of importance as it focuses on the computational and processing of human language. The affective computing aspect deals with recognizing, interpreting, processing and simulating human emotions, in our case, we focus on recognizing the majority underlying emotion leading to sarcasm. We also use the multimodal machine learning approach where more than one modality is leveraged to deal with such nuanced tasks, in order to capture the subtle differences between verbal and non-verbal cues.

The analysis of emotions and human speech has transformed over the years. During The 1990s, the earliest development of Natural Language Processing (NLP) was concerned with the most basic vocabulary and rule-based analysis of the text. These early approaches treated language meaning in a straightforward manner, without considering complex nuances like sarcasm or the wide spectrum of emotions (Shelke and Wagh, 2024). Consequently, early NLP models found it difficult to understand the wide spectrum of emotions in communication and the meanings derived from sarcastic remarks. By the 2000s, research within NLP shifted towards sentiment analysis, which aimed to classify speech and text into categories such as positive, negative, or neutral (Ahlgren et al,. 2016).

The automatic detection of sarcasm has become a significant area of research in recent years, especially in social media platforms (Schifanella et al., 2016, p.1-2).Initial work on sarcasm detection, dating back to almost a decade, primarily focused on textual-data and utilized rule-based techniques (Veale and Hao, 2010), linguistic and lexical features (González-Ibáñez, 2011), stylistic features (Davidov et al., 2010), situational disparity (Riloff et al., 2013), or user-provided annotations such as hashtags (Liebrecht et al., 2013). The recent contributions in the field have moved towards multimodal methods to combine information from various sources (text, audio, video) to improve sarcasm detection (Castro et al., 2019).

Following the shift towards multimodal approaches which acknowledge that sarcasm often relies on cues beyond just text, we focus on understanding the nuanced nature of sarcasm and its interplay with emotion.

Sarcasm is defined as the use of words that mean the opposite of what you really want to say, especially in order to insult someone, to show irritation, or to be funny. Sarcasm being a sophisticated linguistic articulation, where the surface meaning differs from the underlying emotion, while this incongruity is the key element of sarcasm, the intent could be to appear humorous, ridicule someone, or to express contempt (Ray et al., 2022, p.1).

Emotion analysis aims to identify and interpret human emotional states. Detecting the specific emotion underlying a sarcastic expression is a non-trivial but important task. A single sarcastic expression can be associated with various underlying emotions, such as sadness ("I love being ignored") or frustration ("my mobile is fabulous with a battery backup of only 15 minutes!"). Sarcasm poses challenges to emotion recognition because the perceived emotion might be completely flipped due to the sarcasm. (Ray et al., 2022, p.1). Mixing languages, also known as code-mixing, is a norm in multilingual societies. Multilingual people, who are non-native English speakers, tend to code-mix using English-based phonetic typing and the insertion of anglicisms in their main language. In addition to mixing languages at the sentence level, it is fairly common to find the code-mixing behavior at the word level. This linguistic phenomenon poses a great challenge to conventional NLP systems, which currently rely on monolingual resources to handle the combination of multiple languages. So, Hinglish text is a mixture of English and Hindi language words written in English script i.e., words from Hindi vocabulary are written using English alphabets" (Singh, 2021, p.1). Visual sarcasm, involving facial expressions and body language is considered more universal than vocal cues as they vary significantly across languages. Therefore the concept of multimodality incorporates information from multiple sources, or modalities such as text, audio and video. Multimodal approaches towards sarcasm detection are considered relatively new compared to traditional text-based methods. Objectives of multi-modal analysis of sarcasm in a conversation are to leverage the context and extract the incongruity between the surface and expressed semantics (Bedi et al., 2021, p.1).

While the use and comprehension of sarcasm is a cognitively taxing process (Olkoniemi et al., 2016), psychological evidence advocate that it positively correlates with the receiver's theory of mind (Wellman, 2014), i.e., the capability to interpret and understand another person's state of mind. Early psycholinguistic research treated sarcasm as a critical form of verbal irony frequently found in written, computer-mediated communication, potentially more often than in face-to-face conversations (Olkoniemi et al., 2018). Sarcasm has been extensively researched by linguists and psychologists (Gibbs and Clark, 1992; Gibbs and Colston, 2007; Kreuz and Glucksberg, 1989), yet due to the limited availability of stimuli, sarcasm detection in text has relied chiefly on the recognition of stock patterns and lexical cues.

Earlier work on sarcasm detection was done by Kreuz and Caucci (2007) who studied the influence of adjectives, adverbs, interjections, and punctuation marks in sarcasm detection, and showed that their presence has positive correlation (though small) with the sarcastic text. Sarcasm often highlights failed expectations by engaging in a pragmatic pretense that is designed to be seen through (Campbell and Katz, 2012), so cues such as interjections, intensifiers, punctuation and markers of non-veridicality and hyperbole play a crucial role in recognizing sarcastic intent. Historically, automatic sarcasm detection systems were primarily built around English text, utilizing linguistic or lexical features to identify sarcasm markers like exclamation points, quotation marks (Joshi et al., 2017).

With the advancement of AI, particularly deep learning, researchers began applying neural networks to the task. Davidov et al. (2010) proposed a semi-supervised approach for sarcasm discovery in Amazon product reviews. The authors employed punctuation and pattern-based features to classify the unseen samples using a kNN classifier. A similar study on tweets was proposed by Tsur et al (2010). Other works claimed the presence of sentiment shift or the contextual incongruity to be an important factor in accurate sarcasm prediction (Joshi et al., 2015). Son et al. (2019) proposed a hybrid Bi-LSTM and CNN based neural architecture for the sarcasm detection. Most of the above studies involve sarcasm discovery in the standalone input - which are reasonably adequate for the sentence with explicit sarcastic clues. However, for the implicit case, more often than not, the context in which the sarcastic statement was uttered is of utmost importance. Khattri et al. (2015) exploited the historical tweets of a user to predict sarcasm in his/her tweet. They investigated the sentiment incongruity in the current and historical tweets and proposed it to be a strong clue in the sarcasm detection. In another work, Ghosh et al. (2017) employed an attention-based recurrent model to identify sarcasm in the presence of a context. The authors trained two separate LSTMs-with-attention for the two inputs (i.e., sentence and context), and subsequently, combined their hidden representations during the prediction. The availability of context was also leveraged by Ghosh and Veale (2017). The authors learned a CNN-BiLSTM based hybrid model to exploit the contextual clues for sarcasm detection. Additionally, they investigated the psychological dimensions of the user in sarcasm discovery using 11 emotional states (e.g., upbeat, worried, angry, depressed, etc.). Poria et al. (2016) have used sentiment and emotion features from pre-trained models on a text corpus to predict sarcasm via a Convolutional Neural Network. More recently, attention-based architectures have been proposed to harness the inter- and intra-sentence relationships in texts for efficient sarcasm identification (Tay et al., 2018; Xiong et al., 2019; Srivastava et al., 2020).

Recently it has been observed that sarcasm requires some shared knowledge between the speaker and the audience; it is a profoundly contextual phenomenon [2]. Bamman et al. (2015) use information about the authors, their relationship to the audience and the immediate communicative context to improve prediction accuracy. Rajadesingan et al. (2015) adopt psychological and behavioral studies on when, why, and how sarcasm is expressed in communicative acts to develop a behavioral model and a set of computational features that merge user's current and past tweets as historical context. Joshi et al. (2015) propose a framework based on the linguistic theory of context incongruity and introduce inter-sentential incongruity for sarcasm detection by considering the previous post in the discussion thread. Khattri et al.(2015) present a quantitative evidence that historical tweets by an author can provide additional context for sarcasm detection. They exploit the author's past sentiment on the entities in a tweet to detect the sarcastic intent. Wang et al. (2015) focus on message-level sarcasm detection on Twitter using a context-based model that leverages conversations, such as chains of tweets. They introduce a complex classification model that works over an entire tweet sequence and not on one tweet at a time. In the same direction, our work is based on the integration between linguistic and contextual features extracted from the analysis of visuals embedded in multimodal posts.

In 1974, Ekman conducted extensive studies on emotion recognition research over 6 basic emotions: anger, disgust, fear, happiness, sadness, and surprise. His study shows it is possible to detect emotions given enough features (Ekman et al., 1987). Later studies on text-based emotion recognition are mainly divided into three categories: keyword-based, learning-based, and hybrid recommendation approaches (Kao et al., 2009). Recently, emotion recognition research on text focuses on learning-based methods. Kim proposed CNN (Convolutional neural network) text classification, which is widely used for extracting sentence information (Kim, 2014). However, single sentence emotion recognition is a lack of contextual emotion flow within a dialogue. Therefore, contextual LSTM (Long short-term memory) architecture is proposed to measure the interdependence of utterances in the dialogue (Poria et al., 2017).

In recent times, the use of multi-modal sources of information has gained significant attention to the researchers for affective computing. Mai et al. (2019) proposed a new two-level strategy (Divide, Conquer, and Combine) for feature fusion through a Hierarchical Feature Fusion Network for multi modal affective computing. Chauhan et al. (2019) exploits the interaction between a pair of modalities through an application of Inter-modal Interaction Module (IIM) that closely follows the concepts of an auto-encoder for the multi-modal sentiment and emotion analysis. Ghosal et al. (2018) proposed a contextual inter-modal attention-based framework for multi-modal sentiment classification. In other work (Akhtar et al., 2019), an attention-based multi-task learning framework has been introduced for sentiment and emotion recognition.

The focus of sarcasm detection has shifted from text-based uni-modal analysis to multi-modal analysis as using multi-modal inputs help the model to understand the intent and the sentiment of the speaker with more certainty. Thus, in the context of a dialogue, multi-modal data such as video (visual frames) along with text helps to understand the underlying emotion of the speaker, in a sarcastic conversation (Chauhan et al., 2020). Early multimodal explorations included coupling of textual features with cognitive features such as the gaze-behavior of readers (Mishra et al., 2016) or electro/magneto-encephalographic (EEG/MEG) signals (Filik et al., 2014; Thompson et al., 2016). In the bimodal setting, sarcasm identification with tweets containing images has also been well explored (Cai et al.,2019; Xu et al., 2020; Pan et al., 2020).

Although a significant number of studies on sarcasm detection have been conducted in English, only a handful attempts have been made in Hindi or other Indian languages. One of the prime reasons for limited works is the absence of sufficient dataset on these languages. Bharti et al. (2017) developed a sarcasm dataset of 2,000 Hindi tweets. For the baseline evaluation, they employed a rule-based approach that classifies a tweet as sarcastic if it contains more positive words than the negative words, and vice-versa. In another work, Swami et al. (2018) collected and annotated more than 5,000 Hindi-English code-mixed tweets. They extracted n-gram and various Twitter-specific features to learn SVM and Random Forest classifiers.

In the conversational setting, MUStARD, a multimodal, multi speaker dataset compiled by Castro et al. (2019) is considered the benchmark for multimodal sarcasm identification. Chauhan et al. (2020) leveraged the intrinsic interdependency between emotions and sarcasm and devised a multi-task framework for multimodal sarcasm detection. Currently, Hasan et al. (2021) performed the best on this dataset with their humor knowledge enriched transformer model. MUStARD is a subset of Multi-modal Emotion Lines Dataset (MELD) (Poria et al., 2018) and MELD is a multimodal extension of textual dataset EmotionLines (Chen et al., 2018). MELD contains about 13,000 utterances from the TV series Friends, labeled with one of the seven emotions (anger, disgust, sadness, joy, neutral, surprise, and fear) and sentiment. EmotionLines (Chen et al., 2018) and EmoryNLP (Zahiri and Choi, 2017) are textual datasets with conversational data, the former containing data from the TV show Friends and private Facebook messenger dialogues, while the latter was also curated from the series Friends. Iemocap (Busso et al., 2008) is a well-known multimodal, dyadic dataset with 151 recorded videos annotated with categorical emotion labels, as well as dimensional labels such as valence, activation, and dominance. However, none of them have sarcastic utterances. Along with categorical classification of basic emotions, seminal works (Russell, 1980; Plutchick, 1980) also propose dimensional models of emotion (Ex: Valence, Arousal, Dominance), which could help in capturing the complicated nature of human emotions better. (Zafeiriou et al., 2017) created a database of 298 videos (non-enacted, in-the-wild) and captured facial affect in their subjects in terms of valence arousal annotations ranging between -1 to +1. Similar work was undertaken in (Preotiuc-Pietro et al., 2016) where valence and arousal were annotated on a nine-point scale on Facebook data. They also release bag-of-words regression models trained on their data which outperform popular sentiment analysis lexicons in valence-arousal prediction tasks.

Recently, Bedi et al. (2021) proposed a code-mixed multi-party dialogue dataset, MASAC, for sarcasm and humor detection in a conversational dialog. They also propose an attention-based architecture named MSH-COMICS for enhanced utterance classification. Kumar et al. (2022) proposed the novel task of Sarcasm Explanation in Dialogue (SED), to generate a natural language explanation for sarcastic conversations by curating WITS (an extension of MASAC) - a novel multimodal, multiparty, code-mixed, dialogue dataset. Ray et al. (2022) introduced the MUStARD++ dataset, an extended version of MUStARD with added labels for emotion, valence, arousal, and sarcasm-type, and benchmarked multimodal fusion models for emotion detection in sarcasm. These two dominant models form the basis for our contemporary debate. Ray et al. (2022) specifically address the task of emotion recognition in sarcasm. They point out that while sentiment and emotion analysis have been extensively studied, the relationship between sarcasm and emotion has largely remained unexplored. They note that affective dimensions like valence and arousal are important indicators of emotional intensity. Kumar et al. (2022) argue that understanding sarcasm, particularly in conversational settings, involves the acute comprehension of audio-visual cues, as well as contextual and speaker-dependent information. To support this task, they curated the WITS dataset, containing sarcastic dialogues from a popular Indian TV show, code-mixed (Hindi-English) dialogue dataset augmented with natural language explanations for sarcastic conversations. Both papers deal with conversational contexts, which are considered essential for detecting sarcasm.

Ray et al.'s (2020) work involves multimodal emotion recognition in sarcasm. The limitation however is that the MUStARD++ dataset comprises monolingual conversations from famous sit-coms: Friends, Big Bang Theory, The Golden Girls, Burnistoun, The Silicon Valley. Hence it would not be able to adapt to the linguistic subtleties and cultural hints in code-mixed speech as processing code-mixed content directly using monolingual models can lead to a loss of context, as sarcasm often depends on language-specific forms as pointed out by Bedi et al. (2021). Kumar et al.'s (2022) work comprises code-mixed conversations from the Indian show 'Sarabhai v/s Sarabhai', however it focuses on explaining sarcasm and does not deal with emotions in any regard directly.

Based on the literature, a fundamental gap exists at the intersection of sarcasm detection, emotion recognition in code-mixed languages. While sarcasm detection has been explored in both monolingual (primarily English) and code-mixed environments, and emotion recognition has been analyzed separately, specific work on classifying the emotional content within sarcastic utterances in code-mixed contexts has been notably neglected. Previous research has not adequately explored how emotions are realized in sarcastic speech when speakers code-switch within the same conversation. Critically, there is a lack of datasets that consistently integrate emotion labels, sarcastic intent, and code-mixed language use. This absence makes it difficult to determine if current models can effectively learn and represent the subtleties of multilingual sarcastic communication. Although some datasets exist for multi-modal sarcasm detection, their primary objective has not been the explicit recognition of emotions contained within that data. This deficiency highlights the need for systems specifically tailored to understand emotions in sarcastic code-mixed interactions using multimodal information.

Therefore, after having found the research gap, our research aims to solve the problem - "How can we recognize the underlying majority emotion in sarcastic, code-mixed (Hindi-English) conversations using multimodal data?"

Methodology

Current models primarily focus on conversations held solely in the English language. We aim to extend the idea to code-mixed (hinglish) conversations. This will diversify the usage and service to other languages as well. To the best of our knowledge, no work has been done on Emotion Recognition in sarcastic code-mixed conversations, however ideas related to sarcasm have been extensively explored.

The proposed solution consists of a multimodal architecture comprising of PCA-based dimensionality reduction, shallow encoders for each modality (text & videos), and an attention-based fusion mechanism to combine the embeddings of the above-mentioned modalities and giving weights (attention) dynamically to each of the input samples based on the combined target-context relationship.

Existing emotion-recognition systems do not address sarcastic code-mixed conversational videos where textual content often contradicts the visual cues, in addition to being in a combination of more than one language.

The uniqueness of our project lies in the absence of work done in this niche arena. While a lot of researches have involved emotion recognition, sarcasm detection and code-mixed conversational data, none have so far combined these three dimensionalities to create what our model aims to do. We aim to address this missing element from the discipline, so as to diversify the usage of conversational systems by including multiple languages, where our starting point lies at Hinglish (Hindi+English).

To begin with we created two models- a simple random Forest Classifier and a custom model-SNN-AFM (Shallow Neural Network relying on Attention based Fusion Mechanism).

Both our models begin with creating the text and video embeddings after the preprocessing and cleaning of the dataset. For the text embeddings, we used FastText Supervised Model which is adapted to the use of code-mixed data (Bedi et al., 2021). This model generated 100-dimensional vector embeddings for each instance of our text (target and context utterances combined). For our video dataset, we generated key frames for each instance, using OpenCV library, and then converted each key frame into vector embeddings using ResNET50. These embeddings were further averaged, in order to get one embedding of 2048 dimensions per instance.

For our custom model we then perform dimensionality reduction using PCA to condense text embeddings from 100 to 64 dimensions and video embeddings from 2048 to 256 dimensions. The next steps include using a Text Encoder, which is a simplified neural network that transforms reduced text features (from PCA) into a 32-dimensional hidden representation, and a Video Encoder which has a similar structure that transforms reduced video features (from PCA) into a 64-dimensional representation. This follows the Attention-based Fusion mechanism which combines text and video features by dynamically assigning attention weights to each of the modalities. Finally, a Classifier which is the final linear layer in the network, is used to output emotion class probabilities. For the simple Random Forest approach we applied the simple concatenation method to the video and text embeddings hence creating a 2148-dimensional vector. This vector is then sent to the classifier to predict the emotions.

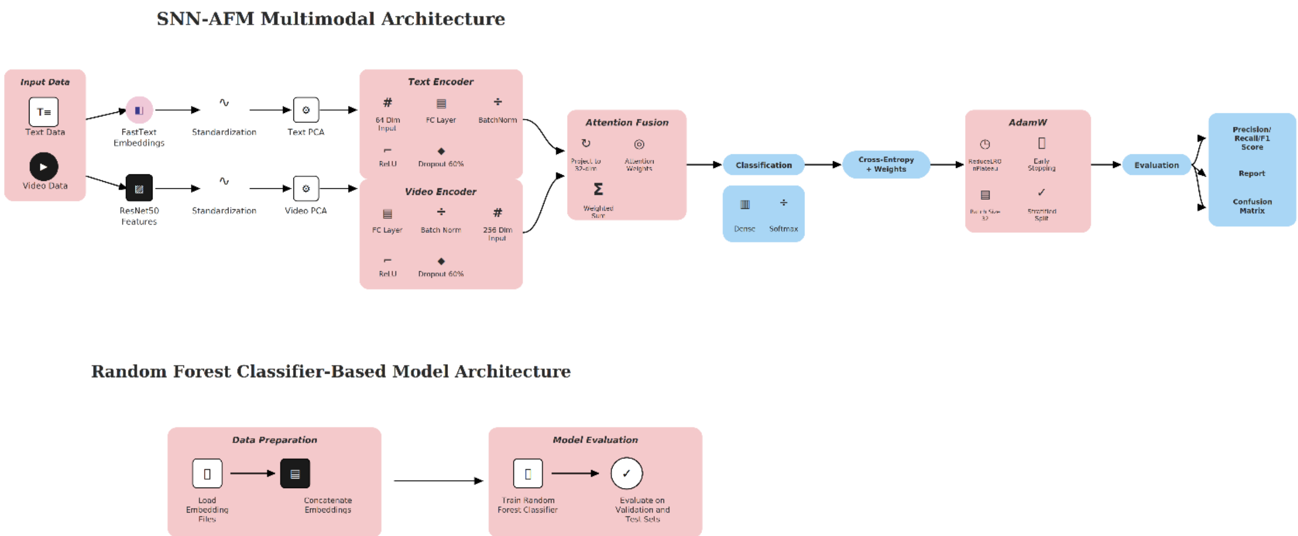

Figure 1: Complete model architecture used in the project.

For Evaluation, we are calculating F1-score for each emotion class. We are using the weighted-average F1-score metric to find an equal balance between precision and recall. This is very important since our dataset has highly imbalanced emotion distributions.

Compared to the contemporary models, we have used a few different and efficient approaches. Instead of using large transformer-based models like BERT, our model employs a lightweight supervised FastText model offering faster embedding generation and significantly reducing computational overhead. Fusion has been handled through an attention-based mechanism that dynamically assigns weights to the text and video features, unlike simple techniques such as Concatenation and early or late fusion techniques (Castro et al., 2019, p.7). Additionally, our training process has been made more robust incorporating AdamW optimization, weighted cross-entropy loss, learning rate scheduling, and early stopping, techniques that are a unique combination, catered to our project. These also tend to be our unique features, used specifically for our set of constraints and requirements.

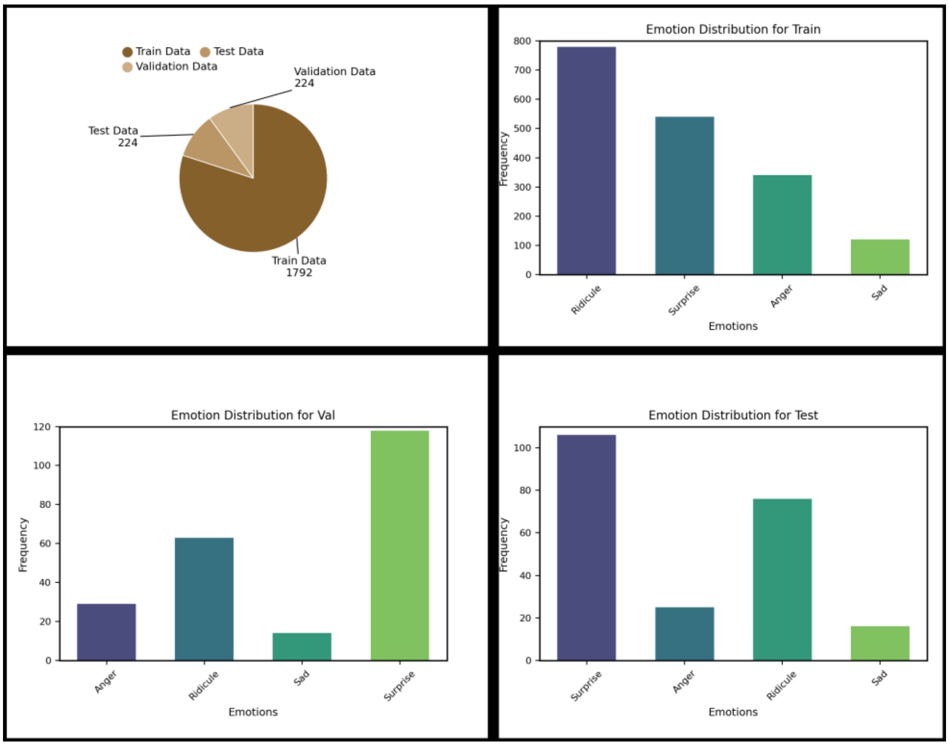

Figure 2: Distribution of emotions in our hand-annotated dataset (WITS++)

Our train dataset comprises 1792 instances of English–Hindi code-mixed utterances as texts in Json files, and sequence of videos for each instance. We also have validation and test datasets, each having 224 instances each (texts & videos). All the above-mentioned datasets have been manually assigned emotion labels. The emotions used are Ridicule, Surprise, Sad and Anger. The emotion labeling followed a procedure wherein each member of the team annotated a set of instances and the emotion chosen by the majority of the team was assigned to that instance.

This methodology, hence delivers a computationally efficient pipeline that targets the challenge of emotion recognition in sarcastic, code-mixed conversations.

Results and Findings

Data Analysis and Main Findings

On our test dataset of 224 sarcastic, English–Hindi code mixed utterances (text & video), the proposed SNN-AFM achieved a weighted-average F1-score of 0.51 (51 %). Per class F1 scores were 0.40 for Anger, 0.52 for Surprise, 0.59 for Ridicule, and 0.30 for Sad. Our Random Forest Classifier achieved a weighted-average F1-score of 0.53 (53 %). Per class F1 scores were 0.36 for Anger, 0.54 for Surprise, 0.58 for Ridicule, and 0.48 for Sad. These results are summarized in Table 1.

The Random Forest classifier represents a simple early-fusion baseline based on concatenated text and video features, while SNN-AFM integrates the same modality encodings through an attention-based fusion mechanism. As shown in Table 1, both models demonstrate comparable overall performance, with the Random Forest slightly outperforming SNN-AFM in weighted F1-score (0.53 vs. 0.51). However, SNN-AFM achieves notably better performance on the Ridicule class (F1 = 0.59), suggesting that attention-based fusion may be particularly effective at capturing this emotion. Both models struggle with the Sad class, with SNN-AFM performing particularly poorly (F1 = 0.30), indicating the inherent difficulty in recognizing this emotion from textual and visual cues alone in code-mixed sarcastic contexts.

Model	Weighted F1 Score	Anger	Surprise	Ridicule	Sad
Random Forest Classifier	0.53	0.36	0.54	0.58	0.48
SNN-AFM	0.51	0.40	0.52	0.59	0.30

Table 1: Weighted and per-class F1-scores of the proposed models. How main findings answer the research question

Our research question—"How can we recognize the underlying majority emotion in sarcastic, code-mixed (Hindi-English) conversations using multimodal data?"—is addressed through the performance benchmarks established by our two models. Our analysis shows that combining textual and visual cues through multimodal fusion enables emotion recognition in this challenging task, with weighted F1-scores reaching 0.51–0.53. The relatively strong performance on Ridicule (F1 = 0.59 for SNN-AFM, 0.58 for Random Forest) suggests that visual and textual features can effectively capture this emotion even in code-mixed sarcastic utterances. This indicates that in code-mixed sarcasm, visual cues may carry significant emotional signals that complement linguistic information.

However, the results also reveal substantial room for improvement, particularly for underrepresented emotions like Sad (F1 = 0.30–0.48) and Anger (F1 = 0.36–0.40). The performance gap across emotion classes suggests that certain emotions may require additional modalities or more sophisticated feature representations to be accurately recognized in code-mixed sarcastic contexts.

Limitations and Future Work

We acknowledge that several limitations exist in our current approach. First, we did not evaluate unimodal baselines (text-only and video-only models), which would have provided clearer insights into the individual contribution of each modality to the overall performance. Such ablation studies would help determine whether both modalities are truly necessary or if one modality dominates the predictive signal.

Second, the omission of the audio modality represents a significant gap, as prosodic features such as pitch, and tone are known to carry important emotional information, particularly in sarcastic speech. We plan to incorporate audio features and conduct comprehensive unimodal and multimodal comparisons to establish a more complete understanding of emotion recognition in code-mixed sarcastic dialogues.

Third, our dataset exhibits significant class imbalance, with a disproportionate number of instances labelled as Ridicule and Surprise compared to Sad and Anger. This imbalance is reflected in the performance metrics shown in Table 1, where the underrepresented classes (Sad: F1 = 0.30–0.48, Anger: F1 = 0.36–0.40) are poorly recognized compared to the majority classes (Ridicule: F1 = 0.58–0.59, Surprise: F1 = 0.52–0.54). The class imbalance likely contributes to the models' difficulty in learning robust representations for minority emotion classes.

Finally, the dataset size is relatively small (224 test instances) and domain-specific, which may limit the generalizability of our findings. The subjective nature of sarcasm also introduces ambiguity in emotion labelling, as sarcasm's inherent nature often overlaps with ridicule, making clear emotional distinctions challenging.

Our future work would be focussed on addressing these limitations by: (1) incorporating audio features and conducting comprehensive unimodal and multimodal ablation studies; (2) collecting additional data to balance class distributions and increase dataset size; (3) exploring techniques such as oversampling, data augmentation, or class-weighted loss functions to mitigate the impact of class imbalance; and (4) extending the approach to other code-mixed language pairs and conversational contexts to establish broader applicability.

Overall, our results establish initial performance benchmarks on the novel task of emotion recognition in code-mixed sarcastic conversations using the WITS++ dataset. While the achieved F1-scores of 0.51–0.53 demonstrate the feasibility of this approach, they also highlight the complexity of the task and the need for continued research incorporating additional modalities, addressing data imbalance, and employing more sophisticated fusion mechanisms.

Conclusion

This paper addresses an unexplored gap in the literature by investigating emotion recognition in sarcastic, code-mixed (Hinglish) dialogues through a multimodal approach. The core research question—"How can we recognize the underlying majority emotion in sarcastic, code-mixed (Hindi-English) conversations using multimodal data?"—was addressed using text and video modalities from the WITS++ dataset.

Our proposed architectures combine PCA-based dimensionality reduction, shallow modality-specific encoders, and two fusion strategies: an attention-based mechanism (SNN-AFM) and a simpler concatenation-based Random Forest classifier. Results demonstrate that multimodal fusion enables emotion recognition in this challenging context, with weighted F1-scores of 0.51–0.53 establishing initial performance benchmarks on this novel task. The Random Forest classifier, despite its simplicity, slightly outperformed the attention-based approach (0.53 vs. 0.51), while SNN-AFM showed particular strength in recognizing Ridicule (F1 = 0.59).

This work establishes a foundation for emotion recognition in multilingually rich and contextually complex conversational settings. By demonstrating that multimodal approaches can capture emotional signals in code-mixed sarcastic dialogue—a previously unexplored intersection of linguistic code-switching, sarcasm, and emotion—we open pathways for future research in multilingual affective computing and cross-cultural communication analysis.

Acknowledgements

Shivani Kumar (IIITD) generously provided us with access to the required WITS dataset, which contained raw Text and Video data, which has been used (with some changes) to train our models. GitHub: https://github.com/LCS2-IIITD/MAF?tab=readme-ov-file

Data Availability Statement

Data sets generated during the current study are available on reasonable request. A mail to any one of the co-authors, along with the reason for data requirement will lead to data availability. The complete codebase for this project is openly available at: https://github.com/IshitaR0/EmotionsInOurSarcasm

References

1. Veale, T., & Hao, Y. (2010). Detecting ironic intent in creative comparisons. In Frontiers in artificial intelligence and applications. https://doi.org/10.3233/978-1-60750-606-5-765
2. Shelke, N. P. P. (2024). Enhanced sarcasm and emotion detection through unified model of transformer and FCNEts. Journal of Electrical Systems, 20(3), 551–565. https://doi.org/10.52783/jes.2982
3. Poria, S., Cambria, E., Hazarika, D., & Vij, P. (2016). A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. International Conference on Computational Linguistics, 1601–1612. https://www.aclweb.org/anthology/C16-1151.pdf
4. Tay, Y., Tuan, L. A., Hui, S. C., & Su, J. (2018). Reasoning with Sarcasm by Reading In-

between. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia. Association for Computational Linguistics, pp 1010-1020. https://doi.org/10.48550/arxiv.1805.02856

5. Xiong, T., Zhang, P., Zhu, H., & Yang, Y. (2019). Sarcasm Detection with Self-matching Networks and Low-rank Bilinear Pooling. In In The World Wide Web Conference, WWW '19, page 2115–2124, New York, NY, USA. Association for Computing Machinery. (pp. 2115–2124). https://doi.org/10.1145/3308558.3313735
6. Srivastava, H., Varshney, V., Kumari, S., & Srivastava, S. (2020). A Novel Hierarchical BERT Architecture for Sarcasm Detection. In Proceedings of the Second Workshop on Figurative Language Processing, pages 93–97, Online. Association for Computational Linguistics. (pp. 93–97). https://doi.org/10.18653/v1/2020.figlang-1.14
7. Hasan, M. K., Lee, S., Rahman, W., Zadeh, A., Mihalcea, R., Morency, L., & Hoque, E. (2021). Humor Knowledge Enriched Transformer for understanding Multimodal humor. Proceedings of the AAAI Conference on Artificial Intelligence, 35(14), 12972–12980. https://doi.org/10.1609/aaai.v35i14.17534
8. Bedi, M., Kumar, S., Akhtar, M. S., & Chakraborty, T. (2021). Multi-Modal sarcasm detection and humor classification in Code-Mixed conversations. IEEE Transactions on Affective Computing, 14(2), 1363–1375. https://doi.org/10.1109/taffc.2021.3083522
9. Mishra, A., Kanojia, D., & Bhattacharyya, P. (2016). Predicting readers' sarcasm understandability by modeling gaze behavior. Proceedings of the AAAI Conference on Artificial Intelligence, 30(1). https://doi.org/10.1609/aaai.v30i1.9884
10. Filik, R., Leuthold, H., Wallington, K., & Page, J. (2014). Testing theories of irony processing using eye-tracking and ERPs. Journal of Experimental Psychology Learning Memory and Cognition, 40(3), 811–828. https://doi.org/10.1037/a0035658
11. Thompson, D., Mackenzie, I. G., Leuthold, H., & Filik, R. (2016). Emotional responses to irony and emoticons in written language: Evidence from EDA and facial EMG. Psychophysiology, 53(7), 1054–1062. https://doi.org/10.1111/psyp.12642
12. Cai, Y., Cai, H., & Wan, X. (2019). Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2506–2515, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1239
13. Xu, N., Zeng, Z., & Mao, W. (2020). Reasoning with Multimodal Sarcastic Tweets via Modeling Cross-Modality Contrast and Semantic Association. In In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3777–3786, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.349
14. Pan, H., Lin, Z., Fu, P., Qi, Y., & Wang, W. (2020). Modeling Intra and Inter-modality Incongruity for Multi-Modal Sarcasm Detection. In In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1383–1392, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.124
15. Kumar, S., Kulkarni, A., Akhtar, M. S., & Chakraborty, T. (2022). When did you become so smart, oh wise one?! Sarcasm Explanation in Multi-modal Multi-party Dialogues. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). https://doi.org/10.18653/v1/2022.acl-long.411
16. Ray, A., Mishra, S., Nunna, A., & Bhattacharyya, P. (2022). A multimodal corpus for emotion recognition in sarcasm. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2206.02119

17. Swami, S., Khandelwal, A., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). A corpus of English-Hindi Code-Mixed tweets for sarcasm detection. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1805.11869

18. Bharti, S. K., Babu, K. S., & Jena, S. K. (2017). Harnessing online news for sarcasm detection in Hindi tweets. In Lecture notes in computer science (pp. 679–686). https://doi.org/10.1007/978-3-319-69900-4_86

19. Tsur, O., Davidov, D., & Rappoport, A. (2010). ICWSM — a great catchy name: Semi-Supervised Recognition of Sarcastic Sentences in online product reviews. Proceedings of the International AAAI Conference on Web and Social Media, 4(1), 162–169. https://doi.org/10.1609/icwsm.v4i1.14018

20. Joshi, A., Sharma, V., & Bhattacharyya, P. (2015). Harnessing Context Incongruity for Sarcasm Detection. ACL , Pp. 757–762. https://doi.org/10.3115/v1/p15-2124

21. Son, L. H., Kumar, A., Sangwan, S. R., Arora, A., Nayyar, A., & Abdel-Basset, M. (2019b). Sarcasm detection using soft Attention-Based bidirectional Long Short-Term Memory model with Convolution Network. IEEE Access, 7, 23319–23328. https://doi.org/10.1109/access.2019.2899260

22. Khattri, A., Joshi, A., Bhattacharyya, P., & Carman, M. (2015). Your Sentiment Precedes You: Using an author's historical tweets to predict sarcasm. In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2015, pp. 25–30. (pp. 25–30). https://doi.org/10.18653/v1/w15-2905

23. Ghosh, A., & Veale, T. (2017). Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/d17-1050

24. Ghosh, D., Fabbri, A. R., & Muresan, S. (2017). The Role of Conversation Context for Sarcasm Detection in Online Interactions. In Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, 2017, pp. 186–196. https://doi.org/10.18653/v1/w17-5523

25. Chauhan, D. S., Akhtar, M. S., Ekbal, A., & Bhattacharyya, P. (2019). Context-aware Interactive Attention for Multi-modal Sentiment and Emotion Analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5651–5661, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/d19-1566

26. Ghosal, D., Akhtar, M. S., Chauhan, D., Poria, S., Ekbal, A., & Bhattacharyya, P. (2018). Contextual inter-modal attention for multi-modal sentiment analysis. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/d18-1382

27. Akhtar, M. S., Chauhan, D., Ghosal, D., Poria, S., Ekbal, A., & Bhattacharyya, P. (2019). Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 370–379, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1034

28. Mai, S., Hu, H., & Xing, S. (2019). Divide, Conquer and Combine: Hierarchical Feature Fusion Network with Local and Global Perspectives for Multimodal Affective Computing. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp.

481–492. https://doi.org/10.18653/v1/p19-1046

29. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 527–536, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1050
30. Hsu, C., Chen, S., Kuo, C., Huang, T. K., & Ku, L. (2018). EmotionLines: An emotion corpus of Multi-Party Conversations. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1802.08379
31. Zahiri, S. M., & Choi, J. D. (2017). Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks. National Conference on Artificial Intelligence, 44–52. https://aaai.org/ocs/index.php/WS/AAAIW18/paper/view/16434
32. Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335–359. https://doi.org/10.1007/s10579-008-9076-6
33. Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161–1178. https://doi.org/10.1037/h0077714
34. Plutchik, R. (1982). A psychoevolutionary theory of emotions. Social Science Information, 21(4–5), 529–553. https://doi.org/10.1177/053901882021004003
35. Zafeiriou, S., Kollias, D., Nicolaou, M. A., Papaioannou, A., Zhao, G., & Kotsia, I. (2017). Aff-Wild: Valence and Arousal 'In-the-Wild' Challenge. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 1980–1987). https://doi.org/10.1109/cvprw.2017.248
36. Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., & Shulman, E. (2016). Modelling Valence and Arousal in Facebook posts. In In Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis, pages 9–15 (pp. 9–15). https://doi.org/10.18653/v1/w16-0404
37. Bamman, D., & Smith, N. (2021). Contextualized sarcasm detection on Twitter. Proceedings of the International AAAI Conference on Web and Social Media, 9(1), 574–577. https://doi.org/10.1609/icwsm.v9i1.14655
38. Rajadesingan, A., Zafarani, R., & Liu, H. (2015). Sarcasm Detection on Twitter. In In WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining (pp. 97-106). (WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining). Association for Computing Machinery. https://doi.org/10.1145/2684822.2685316
39. Wang, Z., Wu, Z., Wang, R., & Ren, Y. (2015). Twitter Sarcasm detection exploiting a Context-Based model. In Lecture notes in computer science (pp. 77–91). https://doi.org/10.1007/978-3-319-26190-4_6
40. Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K., Tomita, M., & Tzavaras, A. (1987b). Universals and cultural differences in the judgments of facial expressions of emotion. Journal of Personality and Social Psychology, 53(4), 712–717. https://doi.org/10.1037/0022-3514.53.4.712

41. Kao, E. C., Liu, C., Yang, T., Hsieh, C., & Soo, V. (2009). Towards Text-based Emotion Detection A Survey and Possible Improvements.In In 2009 International Conference on Information Management and Engineering, pages 70–74, April. (pp. 70–74). https://doi.org/10.1109/icime.2009.113

42. Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1408.5882

43. Olkoniemi, H., Johander, E., & Kaakinen, J. K. (2018b). The role of look-backs in the processing of written sarcasm. Memory & Cognition, 47(1), 87–105. https://doi.org/10.3758/s13421-018-0852-2

44. Chudasama, V., Kar, P., Gudmalwar, A., Shah, N., Wasnik, P., & Onoe, N. (2022). M2FNET: Multi-modal fusion Network for emotion recognition in conversation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 4651–4660. https://doi.org/10.1109/cvprw56347.2022.00511

45. Schifanella, R., De Juan, P., Tetreault, J., & Cao, L. (2016). Detecting sarcasm in multimodal social platforms. Proceedings of the 30th ACM International Conference on Multimedia, 1136–1145. https://doi.org/10.1145/2964284.2964321

46. Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. Conference on Computational Natural Language Learning, 107–116. https://www.aclweb.org/anthology/W10-2914.pdf

47. Liebrecht, C., Kunneman, F., & Van Den Bosch, A. (2013). The perfect solution for detecting sarcasm in tweets #not. North American Chapter of the Association for Computational Linguistics, 29–37. https://repository.ubn.ru.nl/bitstream/handle/2066/112949/112949.pdf

48. Olkoniemi, H., Ranta, H., & Kaakinen, J. K. (2015). Individual differences in the processing of written sarcasm and metaphor: Evidence from eye movements. Journal of Experimental Psychology Learning Memory and Cognition, 42(3), 433–450. https://doi.org/10.1037/xlm0000176

49. Ahlgren, O. (2016). Research on Sentiment Analysis: The first decade. IEEE 16th International Conference on Data Mining Workshops. https://doi.org/10.1109/icdmw.2016.0131

50. Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S. (2019). Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper). In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1455

51. Singh, G. (2021). Sentiment Analysis of Code-Mixed Social Media Text (Hinglish). arXiv (Cornell University). https://doi.org/10.48550/arxiv.2102.12149

52. Wellman, H. M. (2014). Making Minds: How Theory of Mind Develops. http://ci.nii.ac.jp/ncid/BB18602092

53. Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. Journal of Memory and Language, 31(2), 183–194. https://doi.org/10.1016/0749-596x(92)90010-u

54. Gibbs, R. W., & Colston, H. L. (2007). Irony in language and thought: A Cognitive Science Reader. Psychology Press. https://books.google.ie/books?isbn=0805860622

55. Kreuz, R. J., & Glucksberg, S. (1989). How to be sarcastic: The echoic reminder theory of verbal irony. Journal of Experimental Psychology General, 118(4), 374–386. https://doi.org/10.1037/0096-3445.118.4.374

56. Kreuz, R. J., & Caucci, G. M. (2007). Lexical influences on the perception of sarcasm. In

Proceedings of the Workshop on Computational Approaches to Figurative Language (pp. 1–4). https://doi.org/10.3115/1611528.1611529

57. Campbell, J. D., & Katz, A. N. (2012). Are there necessary conditions for inducing a sense of sarcastic irony? Discourse Processes, 49(6), 459–480. https://doi.org/10.1080/0163853x.2012.687863