VerbaNexAI Lab at SemEval-2024 Task 10: Emotion recognition and reasoning in mixed-coded conversations based on an NRC VAD approach

Santiago Garcia and **Elizabeth Martinez** and **Juan Cuadrado**
and **Juan Carlos Martinez-Santos** and **Edwin Puertas**
Universidad Tecnologica de Bolivar, Cartagena Colombia
epuerta@utb.edu.co

Abstract

This study introduces an innovative approach to emotion recognition and reasoning about emotional shifts in code-mixed conversations, leveraging the NRC VAD Lexicon and computational models such as Transformer and GRU. Our methodology systematically identifies and categorizes emotional triggers, employing Emotion Flip Reasoning (EFR) and Emotion Recognition in Conversation (ERC). Through experiments with the MELD and MaSaC datasets, we demonstrate the model's precision in accurately identifying emotional shift triggers and classifying emotions, evidenced by a significant improvement in accuracy as shown by an increase in the F1 score when including VAD analysis. These results underscore the importance of incorporating complex emotional dimensions into conversation analysis, paving new pathways for understanding emotional dynamics in code-mixed texts.

1 Introduction

Exploring emotion recognition in textual and multimodal conversations is crucial within Natural Language Processing (NLP) and Artificial Intelligence (AI). This domain addresses the complexity of human emotional expression, particularly challenged by the interlacing of multiple languages in code-mixed texts. Such code-switching, prevalent in digital communication, necessitates innovative computational strategies to decipher the embedded emotional substrates (Wang et al., 2022), presenting unique challenges for emotion recognition and understanding.

Recent advancements have highlighted the potential of complex neural architectures, like hierarchical transformers, to dissect the nuanced interplay between linguistic codes. This approach indicates a broader NLP trend that prioritizes models capable of parsing linguistic structures and decoding emotional cues within them (Cuadrado et al., 2023a). Significantly, the dynamic nature of conversational emotion and the phenomenon of emotion flips in multi-party interactions call for adaptive models that can trace these shifts accurately (Puertas et al., 2022).

Moreover, multimodal approaches that integrate visual, textual, and auditory cues are pivotal in capturing the essence of code-mixed interactions. These strategies convey sentiment and intention, underscoring the significance of non-verbal cues (Martinez et al., 2023). Additionally, the exploration of large language models for understanding complex conversational patterns has led to an evolving AI research landscape, where the efficacy of models like GPT in nuanced tasks such as sarcasm explanation and affect understanding in dialogues is rigorously evaluated (Cuadrado et al., 2023b).

Analyzing sociolinguistic features in digital social networks further enriches the discourse on digital communication's implications for emotion recognition and conversational AI. It includes bot detection, gender profiling, and community detection through sociolinguistic cues analysis (Moreno-Sandoval et al., 2019; Puertas et al., 2021, 2019). Moreover, the precision application of NLP methodologies, such as phonetic detection techniques for identifying hate speech spreaders on Twitter, showcases the necessity for targeted approaches to specific social media phenomena (Puertas and Martinez-Santos, 2021).

Our research aims to advance the understanding of NLP's multifaceted applications in digital interactions' integrity and authenticity. Through the analysis of polarity, emotion, and user statistics for fake profile detection, alongside multimodal emotion-cause pair extraction in conversations, we seek to significantly improve the comprehension of the complex interrelations between emotional expressions and their triggers (Moreno-Sandoval and Alvarado-Valencia, 2020; Wang et al., 2022).

In evaluating the incorporation of Valence,

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1332–1338

Arousal, and Dominance (VAD) scores from the NRC VAD Lexicon into our computational models, we observed a marginal performance improvement. Specifically, the inclusion of VAD scores resulted in F1 scores of 0.34 for Emotion Flip Reasoning (EFR) and 0.23 for Emotion Recognition in Conversation (ERC), compared to models without VAD scores, which achieved F1 scores of 0.32 and 0.20 for EFR and ERC respectively. These results underscore the nuanced challenges of accurately capturing emotional shifts in code-mixed conversations, paving the way for future research to refine and enhance emotion recognition systems in complex conversational contexts. Find here the GitHub repository[1]

2 Related Work

The recognition of emotions in code-mixed text and multimodal conversations has garnered increasing attention within the natural language processing (NLP) and artificial intelligence (AI) communities. The growing prevalence of code-switching in digital communication fuels this surge in interest and the multifaceted nature of human emotional expression.

Recent advancements in understanding code-mixed language semantics have underscored the potential of hierarchical transformer models to grasp the nuanced interplay between different linguistic codes (Sengupta et al., 2022). Through their ability to capture deep semantic relationships, these models offer a promising avenue for more accurate emotion recognition in code-mixed conversations. Such approaches align with the broader trend of employing sophisticated neural architectures to tackle the complexities of multilingual text processing.

Studies focusing on multiparty interactions have specifically addressed emotion flip in conversations, where the emotional trajectory can shift dramatically due to a single utterance or interaction (Kumar et al., 2023a, 2024a,b, 2022b). These studies highlight the dynamic nature of conversational emotion and the need for models that can adaptively reason about these shifts to maintain coherence and accuracy in emotion recognition tasks.

Multimodal approaches to sarcasm detection and humor classification in code-mixed conversations further illustrate the rich potential of integrating visual, textual, and auditory cues to enhance the understanding of conversational context and emotional undertones (Bedi et al., 2021). This multimodal perspective is critical in fully capturing the essence of code-mixed interactions, where non-verbal cues significantly convey sentiment and intention.

Exploring large language models' capability in logical reasoning and understanding complex conversational patterns points towards an evolving landscape in AI. Researchers have tested models like GPT (Generative Pre-trained Transformer) for their efficacy in nuanced tasks such as sarcasm explanation and effect understanding in dialogues (Xu et al., 2023). These inquiries into the logical capabilities of large models contribute to a deeper understanding of their potential applications in conversational AI and emotion analysis.

Moreover, research on explaining sarcastic utterances to enhance affected understanding in multimodal dialogues sheds light on the importance of context and the subtleties of human communication. Such work suggests that describing a particular emotional expression's underlying intent or cause beyond detecting sarcasm or emotion is crucial for advanced AI systems for naturalistic human-computer interaction (Kumar et al., 2023b).

The development of comprehensive datasets like MELD, which provides a multimodal multiparty dataset for emotion recognition in conversations, has been instrumental in advancing research in this area (Poria et al., 2018). These datasets not only facilitate the training and testing of sophisticated models but also enable the exploration of new methodologies for emotion recognition across diverse conversational settings.

As we move forward, the integration of insights from masked memory networks, transformer models, and intent-conditioned counter speech generation into the realm of emotion recognition and conversational AI promises to open new avenues for research and application (Poria et al., 2018; Kumar et al., 2022c; Christ et al., 2023). The collective efforts in these areas underscore the ongoing pursuit of more empathetic, contextually aware, and linguistically versatile AI systems capable of navigating the complexities of human emotion and communication.

3 Methodology

This section details the methodology adopted for analyzing emotion causes in multimodal conversations. Our approach, grounded in integrating

[1]https://github.com/VerbaNexAI/EmoVAD.git

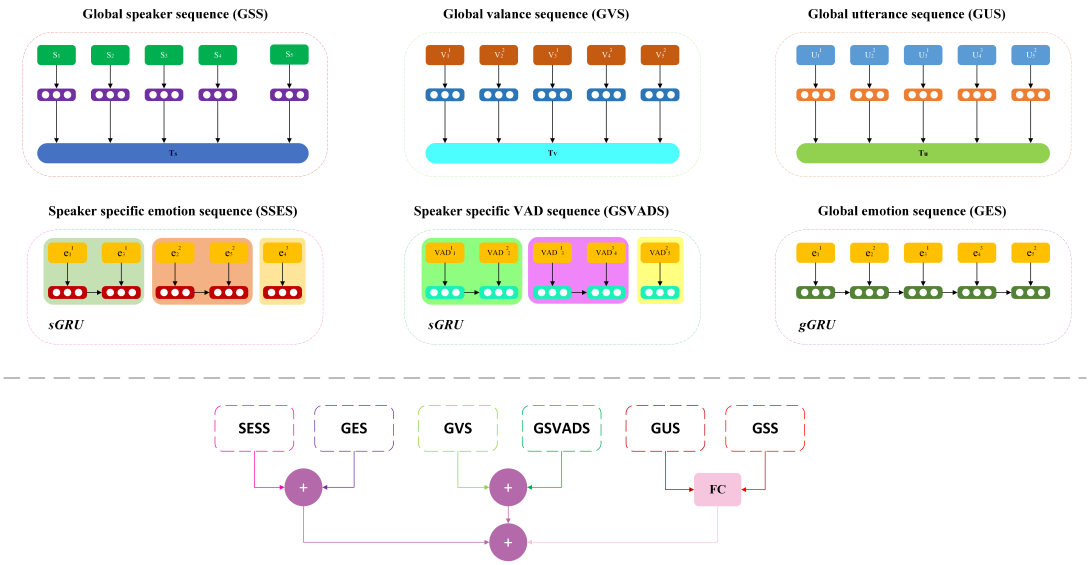

Figure 1: System General Pipeline

the NRC VAD Lexicon and computational models, aims to systematically identify and categorize emotion triggers. The methodology encompasses Emotion Flip Reasoning (EFR) for detecting shifts in conversation emotions and Emotion Recognition in Conversation (ERC), focusing on classifying these emotions accurately. By employing a combination of Transformer and GRU models, we enhance the analysis of valence, arousal, and dominance (VAD) scores, contributing to a nuanced understanding of emotional dynamics in conversations. All scripts and data related to this study are available at SemEval 2024 VerbaNex AI Repository.

3.1 Emotion Flip Reasoning

This section shows the proposed model and how this can identify the trigger for the corresponding emotional flip in the conversation; the way to identify it is by analyzing each utterance in the sequence; that is, the task is essential for a binary classification because we're trying to categorize each utterance responsible or not for the emotion flip. (Kumar et al., 2023a) propose the TGIF model, which contains the context of utterances, speakers, and emotions. This model explains how to process these three inputs through the pipeline. They propose four modules:

- **Global Utterance Sequence:** They use a Transformer (Vaswani et al., 2017) encoder architecture to push the $U = u_2, u_1, ..., u_i$ utterance distribution into a latent space, capture the global context of the dialogue
- **Global Emotion Sequence:** In this approach, we use GRU for the emotions processing, due to there are just a few (Ekman, 1992) emotions {*disgust, joy, surprise, anger, fear, sadness*} encoded in one hot.
- **Speaker-Specific Emotion Sequence:** This time, they also process the emotions, but concerning the speakers, each speaker has their own GRU for the speaker's emotions
- **Global Speaker Sequence:** For the speakers processing, they also use a Transformer approach encoded in one hot.

The original task in (Kumar et al., 2023a) was to predict the instigator(s) label(s) for each emotion flip; for example, they assign 'nervousness' and 'adoration' instigators to the trigger utterances u_2 and u_3, the instigator is the reason why the emotion flip occurs. We worked on a simple task: identify the trigger in the conversation and which utterance was the cause of the emotion flip.

We propose a new serial of data input to contribute to the model performance; this data is (Mohammad, 2018) NRC VAD Lexicon. This Lexicon contains the {*valence, arousal, dominance*}, with the valence the positive/negative or pleasure/displeasure dimension, arousal is the excited/calm or active/passive dimension. Dominance is the powerful/weak or 'have full control'/' have no control' dimension. We compare the words between the NRC VAD Lexicon dictionary and the words in each utterance and build a personal-

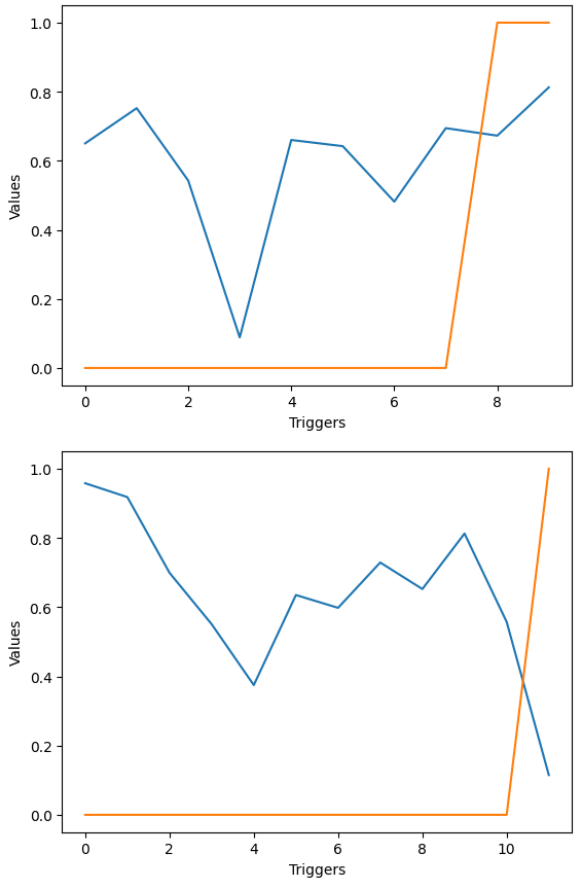

Figure 2: These two graphics describe the relationship between the valence behavior and triggers

ized dictionary for our dataset. We get the following distributions, $u_i = w_0, w_1, ..., w_l$, l begin the number of words in a given utterance, and each word has this distribution $w_l = v_l, a_l, d_l$ (or in the case of the articles, pronouns, etc., the value will be 0), the aim here is to calculate the values of valence, arousal, and dominance that represent the whole utterance, then we apply $1/l \sum_{i=1}^{l} v_i$, $1/l \sum_{i=1}^{l} a_i$, $1/l \sum_{i=1}^{l} d_i$, so the final shape is $u_i = v_{avg}, a_{avg}, d_{avg}$, the VAD values will be lower so we scale it into [0,1]

We took all these features to help the model in the classification task; we found a relationship between the VAD values peak and the trigger sequence location in several samples. see Figure 2

To contribute to the model, we added a new Transformers Encoder for the valence, arousal, and dominance values and a GRU for VAD speaker-specific values. Like the Speaker-Specific Emotion Sequence, we compute a particular speaker's VAD sequence and go through lineal classification layers, see Figure 1.

3.2 Emotion Recognition in Conversation

The architecture for emotion recognition parallels EFR's, with modifications to accommodate emotion as the primary label. This adjustment allows for a direct correlation between VAD scores and emotion classification, eliminating the need for separate emotion modules. The methodological choice to employ GRU models for processing speaker-specific emotional sequences facilitates a refined analysis, enabling the identification of diverse emotional expressions within the conversational context.

4 Experiments

This section presents the experimental setup, including dataset description, data preprocessing techniques, and model evaluation. Utilizing the MELD dataset, we describe our approach to representing conversational utterances through advanced embedding methods. The experiments aim to validate the effectiveness of our methodology in identifying and classifying emotional causes within conversations. Our findings, evaluated against established metrics, indicate a promising direction for future research in multimodal emotion analysis.

4.1 Dataset

MELD. (Poria et al., 2019) is an extension and enhancement of (Chen et al., 2018) EmotionLines. MELD contains dialogues from the TV series Friends. Each utterance is annotated with emotion and sentiment labels and encompasses audio, visual, and textual modalities. The SemEval 2024 Task 10 Subtask 1 presents a variation of MELD, providing *speakers, utterance, emotion* as only textual features and *triggers* as labels.

(Kumar et al., 2023a) identify a set of trigger utterances that cause the emotion to flip at the target. They mark each utterance that acts as a trigger as 'Yes' and the ones not contributing as 'No'.

MaSaC. (Bedi et al., 2023) develop a Hindi-English code-mixed dataset for the multi-modal sarcasm detection and humor classification in conversational dialog. Like MELD, SemEval modifies the dataset for two tasks (ERC and EFR) and only textual data(Kumar et al., 2023c). ERC uses *emotions* as labels, and EFR uses *triggers*

NRC-VAD Lexicon. The National Research Council Canada Valence, Arousal, and Dominance (NRC-VAD) Lexicon (Mohammad, 2018) includes a list of more than 20,000 English words and their valence, arousal, and dominance scores. For a given word and a dimension (V/A/D), the scores range from 0 (lowest V/A/D) to 1 (highest V/A/D). The lexicon with its fine-grained real-valued scores was created by manual annotation using Best-Worst Scaling. The lexicon is markedly more significant than any of the existing VAD lexicons.

4.2 Data Preprocessing

Utterances. To represent the sentences in a dense numerical vector, we use Sentence Embedding pre-trained models. Specifically (Song et al., 2020) the model MPNet, due we handle a sequence of sentences and not a sequence of words is necessary to put all the utterance meaning in just one vector that the Transformer Encoder could process. In the case of Hindi-English code-mixed, we use paraphrase-multilingual-mpnet-base-v2 - Multilingual version of MPNet, trained on parallel data for 50+ languages.

Emotions and Speakers. We use One-Hot Encoder in both cases; the speaker's sequence has a tensor shape of max sequence length and max number of unique speakers in a dialogue in the whole dataset. We assign a one-hot vector for each emotion.

One of the other steps is padding the dataset for every sequence by the maximum sequence length.

5 Results

In our experimental investigations, we meticulously evaluated various configurations of our model and adjusted hyperparameters, alternating between Transformer and GRU modules. For the Emotion Flip Reasoning (EFR) task, we utilized sigmoid neurons with binary cross-entropy loss for binary classification. For Emotion Recognition in Conversation (ERC), we employed softmax neurons with cross-entropy loss for multi-class classification. The F1 score was chosen as the primary metric for evaluation, reflecting the balanced consideration of precision and recall in our assessments.

The integration of Valence, Arousal, and Dominance (VAD) values, crucial emotional dimensions discussed in Section 3, was meticulously analyzed to optimize the model configuration. Drawing on the methodology proposed by (Kumar et al., 2022a), we processed the VAD values with a Transformer Encoder and amalgamated them with other Transformer modules via a linear layer. Similar to processing emotion-specific data, we treated VAD values in a speaker-specific manner using several GRUs, subsequently integrating them with emotion modules through straightforward concatenation.

As depicted in Table 1, the results underscore the significance of including VAD in the model. With VAD integration, the model achieved F1 scores of 0.34 for EFR and 0.23 for ERC, demonstrating improvements of approximately 13% and 5%, respectively, over the configurations without VAD. These findings were consistent across both MELD and MaSaC datasets, highlighting VAD's contribution to enhancing model performance in identifying emotion flips and recognizing emotions in code-mixed conversations.

Our analysis revealed a notable observation regarding the model's distribution loss function. The model's propensity to predict triggers at the beginning of sequences was identified, with ROC and AUC analyses suggesting an optimal threshold of 0.3. For the ERC task, a tendency to predict the 'neutral' category was observed, possibly due to the low deviation of most VAD values from the mean. However, considering the broadly spaced combinations of valence, arousal, and dominance led to slight but discernible improvements in model performance.

Comparatively, while showing promise, our model's performance indicates room for further refinement, especially when juxtaposed with other participants in the shared task. It underlines the necessity for ongoing enhancements and reevaluating the methodological approach to emotion recognition in complex, code-mixed conversational contexts.

Limitations of our current work, including potential dataset biases and the model's generalizability across varied types of code-mixed text, warrant further investigation. Future research directions could encompass exploring additional linguistic features and incorporating other dimensions of emotional reasoning, aiming to build on the foundational insights gained from this study.

6 Conclusion

This study ventured into the complex domain of emotion recognition and reasoning in code-mixed

Model	F1 Score	
	MELD	MaSaC
With VAD	0.34	0.23
Without VAD	0.32	0.20

Table 1: Model Results

conversations, with a particular focus on the tasks of Emotion Flip Reasoning (EFR) and Emotion Recognition in Conversation (ERC). Our primary contribution has been the integration of Valence, Arousal, and Dominance (VAD) values into computational models, aiming to enrich the models' understanding of emotional shifts within dialogues. Despite the modest improvements observed in our experimental results, our work underscores the nuanced challenge of effectively identifying emotion flips and recognizing emotions in code-mixed texts. The incremental advancements achieved, particularly the slight enhancements in F1 scores with VAD values, highlight the potential of incorporating emotional dimensions into NLP models for a deeper understanding of conversational dynamics.

7 Future Work

The pathway forward from this investigation is twofold. Firstly, there is a pressing need to explore additional linguistic and emotional features that could enhance the accuracy and robustness of emotion recognition models. It includes delving deeper into the complexities of code-mixing phenomena and how they influence emotional expression and perception. Secondly, our findings advocate for developing more sophisticated model architectures capable of handling the intricacies of multimodal data and the multifaceted nature of human emotions. Future research should also consider the implications of dataset biases and the challenge of generalizing models across diverse code-mixed contexts. By addressing these areas, subsequent work can build upon the foundational insights provided by this study, contributing to the advancement of NLP and AI's capability to navigate the rich tapestry of human emotions in digital communications.

Acknowledgments

To the SemEval contest, sponsored by the SIGLEX Special Interest Group on the Lexicon of the Association for Computational Linguistics. To the master's degree scholarship program in engineering at the Universidad Tecnológica de Bolivar (UTB) in Cartagena, Colombia.

We would like to express our gratitude to the team at the VerbaNex AI Lab [2] for their dedication, collaboration, and ongoing support of our research endeavors.

References

Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*.

Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*, 14(2):1363–1375.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Ting-Hao, Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations.

Lukas Christ, Shahin Amiriparian, Alice Baird, Alexander Kathan, Niklas Müller, Steffen Klug, Chris Gagne, Panagiotis Tzirakis, Lukas Stappen, Eva-Maria Meßner, et al. 2023. The muse 2023 multimodal sentiment analysis challenge: Mimicked emotions, cross-cultural humour, and personalisation. In *Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation*, pages 1–10.

Juan Cuadrado, Elizabeth Martinez, Juan Carlos Martinez-Santos, and Edwin Puertas. 2023a. team utb-nlp at finances 2023: financial targeted sentiment analysis using a phonestheme semantic approach. -.

Juan Cuadrado, Elizabeth Martinez, Anderson Morillo, Daniel Peña, Kevin Sossa, Juan Martinez-Santos, and Edwin Puertas. 2023b. Utb-nlp at semeval-2023 task 3: Weirdness, lexical features for detecting categorical framings, and persuasion in online news. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1551–1557.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

Shivani Kumar, Md Shad Akhtar, Erik Cambria, and Tanmoy Chakraborty. 2024a. Semeval 2024 – task 10: Emotion discovery and reasoning its flip in conversation (ediref). In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

[2] https://github.com/VerbaNexAI

Shivani Kumar, Shubham Dudeja, Md Shad Akhtar, and Tanmoy Chakraborty. 2023a. Emotion flip reasoning in multiparty conversations. *arXiv preprint arXiv:2306.13959*.

Shivani Kumar, Shubham Dudeja, Md Shad Akhtar, and Tanmoy Chakraborty. 2024b. Emotion flip reasoning in multiparty conversations. *IEEE Transactions on Artificial Intelligence*, 5(3):1339–1348.

Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. 2022a. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. *arXiv preprint arXiv:2203.06419*.

Shivani Kumar, Ishani Mondal, Md Shad Akhtar, and Tanmoy Chakraborty. 2023b. Explaining (sarcastic) utterances to enhance affect understanding in multimodal dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12986–12994.

Shivani Kumar, Ramaneswaran S, Md Akhtar, and Tanmoy Chakraborty. 2023c. From multilingual complexity to emotional clarity: Leveraging commonsense to unveil emotions in code-mixed dialogues. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9638–9652, Singapore. Association for Computational Linguistics.

Shivani Kumar, Anubhav Shrimal, Md Shad Akhtar, and Tanmoy Chakraborty. 2022b. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowledge-Based Systems*, 240:108112.

Shivani Kumar, Anubhav Shrimal, Md Shad Akhtar, and Tanmoy Chakraborty. 2022c. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowledge-Based Systems*, 240:108112.

Elizabeth Martinez, Juan Cuadrado, Juan Carlos Martinez-Santos, Daniel Peña, and Edwin Puertas. 2023. Automated depression detection in text data: leveraging lexical features, phonesthemes embedding, and roberta transformer model. -.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Luis Gabriel Moreno-Sandoval and Jorge Andres Alvarado-Valencia. 2020. Assembly of polarity, emotion and user statistics for detection of fake profiles. In -.

Luis Gabriel Moreno-Sandoval, Edwin Puertas, Flor Miriam Plaza-del Arco, Alexandra Pomares-Quimbaya, Jorge Andres Alvarado-Valencia, and L Alfonso. 2019. Celebrity profiling on twitter using sociolinguistic. *CLEF (Working Notes)*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations.

Edwin Puertas and Juan Carlos Martinez-Santos. 2021. Phonetic detection for hate speech spreaders on twitter. -.

Edwin Puertas, Juan Carlos Martinez-Santos, and Pablo Andrés Pertuz-Duran. 2022. Presidential preferences in colombia through sentiment analysis. In *2022 IEEE ANDESCON*, pages 1–5.

Edwin Puertas, Luis Gabriel Moreno-Sandoval, Flor Miriam Plaza-del Arco, Jorge Andres Alvarado-Valencia, Alexandra Pomares-Quimbaya, and L Alfonso. 2019. Bots and gender profiling on twitter using sociolinguistic features. *CLEF (Working Notes)*, pages 1–8.

Edwin Puertas, Luis Gabriel Moreno-Sandoval, Javier Redondo, Jorge Andres Alvarado-Valencia, and Alexandra Pomares-Quimbaya. 2021. Detection of sociolinguistic features in digital social networks for the detection of communities. *Cognitive Computation*, 13:518–537.

Ayan Sengupta, Tharun Suresh, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. A comprehensive understanding of code-mixed language semantics using hierarchical transformer. *arXiv preprint arXiv:2204.12753*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2022. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, pages 1–12.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023. Are large language models really good logical reasoners? a comprehensive evaluation from deductive, inductive and abductive views. *arXiv preprint arXiv:2306.09841*.