MULTICOLLAB: A Multimodal Corpus of Dialogues for Analyzing Collaboration and Frustration in Language

Michael Peechatt, Cecilia O. Alm, Reynold Bailey
Rochester Institute of Technology
1 Lomb Memorial Drive, Rochester, NY 14623, United States
{mp6510, coagla, rjbvcs}@rit.edu

Abstract

This paper addresses an existing resource gap for studying complex emotional states when a speaker collaborates with a partner to solve a task. We present a novel dialogue resource — the MULTICOLLAB corpus — which was collected in an IRB-approved laboratory experiment where two interlocutors, an instructor and builder, communicated through a Zoom call while sensors recorded eye gaze, facial action units, and galvanic skin response, in addition to speech signals that were subsequently carefully transcribed, resulting in a unique, heavily multimodal corpus. The builder received instructions from the instructor. Half of the builders were privately told to deliberately disobey the instructor's directions. After the task, participants watched the Zoom recording and annotated their instances of frustration. In this study, we both introduce this new corpus and perform computational experiments with time series transformers, using early fusion through time for sensor data and late fusion for speech transcripts. We then average predictions from both methods to recognize instructor frustration. Using sensor and speech data in a 4.5 second time window, we find that the fusion of both models yields 21% improvement in classification accuracy (with a precision of 79% and F1 of 63%) over a comparison baseline, demonstrating that complex emotions can be recognized when rich multimodal data from transcribed spoken dialogue and biophysical sensor data are fused.

Keywords: multimodal machine learning, affective computing, computational paralinguistics

1. Introduction

With the increasing availability of sensors in consumer-grade devices, there is a growing need for flexible and resource-efficient machine learning models that can handle multimodal data to support a range of downstream tasks. Additionally, since the pandemic, there has been a notable rise in the use of online meeting technologies. This has transformed, for example, the education domain. Students and instructors use applications such as Zoom (Zoom Video Communications, 2023) which involve multiple data streams (speech audio, video, chat logs, etc.). Analysis performed on these data streams yield latent features that can be used for downstream inference tasks, such as prioritizing information discussed (Cohen et al., 2021; Amin et al., 2022) or identifying user affect (Baltrušaitis et al., 2018). As human-human interaction devices and tools become more mature, their integration of other human-generated modalities, such as eye gaze and biophysical sensors, will likely become common-place.

Frustration and confusion are under-explored emotions in a collaborative context, and both play a key role in learning (Zeng et al., 2017) as well as interpersonal interaction (Grafsgaard et al., 2011; Kaushik et al., 2021; Mince et al., 2022). Identifying and measuring instances of these emotions can also inform teachers whether their students are comprehending concepts being taught. We present a new heavily multimodal corpus, the **MULTICOLLAB**[1] dataset, from a lab-based data collection experiment designed to elicit frustration from participants in an online Zoom call (see Figure 1) . For labeling, the participants provided timestamped self-annotations which were leveraged later in modeling aimed to recognize their corresponding frustration and confusion during the task. For the scope of this study, we focus on instances of instructor frustration, and we propose the following research questions:

- **RQ1**: Do high-level or low-level features yield better average classification accuracy for identifying instructor frustration?
- **RQ2**: Does, on average, fusing predictions improve overall performance when compared to separate models?

We define low-level features as values extracted from time series sensor data, and we define high-level features as word embeddings from the instructor transcript; with these surrounding the instance of self-annotated frustration. Inspired by human cognition, we separate these data types.

2. Related Work

Multimodal Data for Emotion Analysis Human communication is usually multimodal. In addition

[1] Instructor data available at:
https://github.com/mp6510/MULTICOLLAB

LREC-COLING 2024, pages 11713–11722

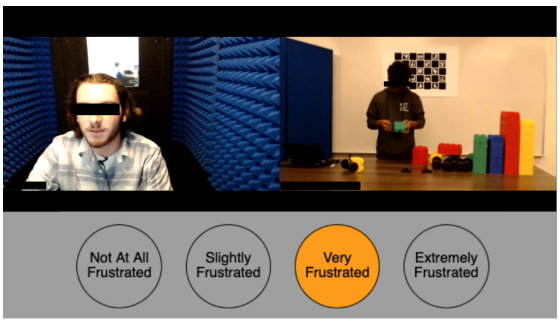

Figure 1: Pairs of participants – an instructor (left) and a builder (right) – completed a block-building task over Zoom and then watched the recording and annotated their instances of frustration on a discrete scale — *Not At All Frustrated*, *Slightly Frustrated*, *Very Frustrated*, and *Extremely Frustrated*.

to spoken language, other modalities contribute to perception and expression of meaning in context. Interlocutors' speech prosody, facial expressions, body posture, and conversational context convey meaning and affect in interactive dialogue (Baltrušaitis et al., 2018). The subtlety of human expression requires modeling to analyze multiple modalities. When limited to a single modality without context, such a task is difficult even for humans (Sham et al., 2023; Hoque et al., 2012). Affective computing advances computational methods to recognize or generate emotions (Poria et al., 2017; Gandhi et al., 2023; Alm, 2012), yet has mostly focused on unimodal or bimodal combinations such as audio and video processing. Recent advancements in computing power and multimodal machine learning has led to interest in novel approaches for data fusion (Peña et al., 2023), with new challenges presented by combining modalities for time series data (Liang et al., 2022). However, there is a growing need for heavily multimodal language resources (Alm, 2022), including for evaluating fusion methods and for controlled study with computationally less complex and less costly machine learning methods, which offer realistic insights for low-power scenarios and privacy-preserving AI on the edge solutions.

Multimodal Affect Analysis with Speech Established multimodal affect datasets such as IEMOCAP (Busso et al., 2008) feature emotion capture and speech data for a range of emotions, but that dataset includes actors engaged in discourse to generate specific emotions, thus representing less realistic acted emotional expression. In contrast, our IRB-approved data collection experiment was designed to induce frustration in conversation tasks without the subject's explicit awareness. Other multimodal affect corpora leverage online data, such as from YouTube (Wöllmer et al., 2013) or the CMU-MOSEI (Zadeh et al., 2018c) dataset. However, these resources rely on *other-annotation* (Saraf et al., 2019) by external annotators who were not involved in generating the emotion data and were instead interpreting other individuals' affect states, resulting in potentially less reliable emotion annotations. In contrast, in our study, participants identified their instances and level of frustration immediately after completing tasks.

Experiments performed on the CMU-MOSEI dataset have modeled the relationships between human-generated modalities (Zadeh et al., 2018a), and Zadeh et al. (2018b) explored modality relationships through time, arguing that collapsing data into an average (rather than preserving their interactions through time) is a weakness in non-sequential deep learning models. Thus, our corpus captures a time window surrounding self-annotations, which also addresses the potential for temporal lags during manual annotation.

In affect interpretation, speech may especially convey emotional arousal (Fagel, 2006). Prior work focused on inferring affect from multimodal data have highlighted that linguistic features can be relevant. Kaushik et al. (2021) performed experiments with online conference tasks with findings suggesting that facial action units and language-derived features can assist toward inferring complex emotions. Mince et al. (2022) provided similar evidence for facial expressions and features extracted from speech transcripts. Saeki et al. (2022) also examined affect identification in the online conversation context, but with a single human participant and a computer agent. Their findings suggested the utility of eye gaze as a prominent feature. Extending this insight, our study includes two human participants engaged in natural spoken dialogue, both with sensors capturing their corresponding eye gaze. In addition, we included galvanic skin response (GSR), a fine-grained temporal measure of reactions (Sanchez-Comas et al., 2021).

3. Methodology

Participants The study comprised 48 subjects (24 builder-instructor groups). Of the participants, 42% were female, 56% male, and 2% chose not to disclose their gender. Of these 24 groups, 8 of them were composed of different gendered interactions, while the remaining 16 had the same gender interactions. The ethnicity distribution of the subjects were as follows: 2.1% Southeast Asian, 8.4% African-American, 10.5% Hispanic, 39.6% Asian, and 37.5% Caucasian. The remaining 1.8% chose not to disclose their race. Additionally, 20.8% were English L2 speakers, and 79.2% were English L1 speakers, with 3 of the 24 groups being a mix of L1 and L2 interactions.

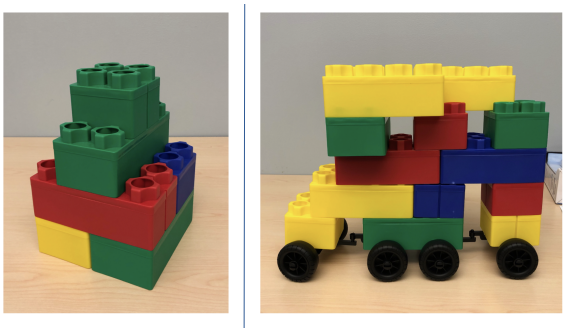

Figure 2: An instructor was tasked with directing a builder to construct these structures for task 1 (left) and task 2 (right). The first task was used to familiarize communication between both participants.

Procedure Subject groups worked in pairs, where one member was given the role of builder and the other instructor. They communicated through a Zoom video call. The instructor was given an image of what the builder was to construct. The builder would attempt to build the described block structure under the direction of the instructor. The first structure was straightforward and used to familiarize both participants with communicating to each other. The second structure was more complex, utilizing more components and blocks. Figure 2 shows the contrast between both tasks. To induce frustration, the builder of every even numbered group was privately told to periodically not cooperate with the instructor's commands during the second task. To ensure ecological validity of the data collection experiment, the intensity of builder disobedience was left up to their discretion.

While performing the aforementioned tasks, both instructors and builders wore various sensors which recorded time series data in addition to the Zoom conference recording. For the instructor, these included a galvanic skin response (GSR) sensor, a screen-based eye tracker, and real-time facial action units extracted from their webcam. For the builder, these sensors included a wearable eye tracker and a full body markerless motion tracking recording (Cherian et al.). Figure 3 provides a breakdown of the modalities recorded.[2] To synchronize these recordings, a movie director clapboard was used to signify the beginning of each task. This provided a unified signal across modalities to align the time series data streams. The average duration of group video recordings was 7 minutes and 42 seconds (min: 5 minutes 23 seconds, max: 9 minutes 9 seconds).

[2]An example video is available at: https://vimeo.com/839024815. The instructor reported a positive frustration rating while smiling. Smiles have been connected with frustration (Hoque et al., 2012).

Annotation Using an annotation tool, each participant then watched the recording of the experiment and annotated their own levels of frustration (see Figure 1). This custom-made, browser-based tool was adapted from prior work (Mince et al., 2022). It generates a JSON output file corresponding to timestamps with the participant's annotation ratings. The recording played through one time, and subjects were instructed to annotate in real-time without the ability to rewind or fast forward. They were encouraged to focus their annotations on how frustrated they personally felt, rather than interpreting their partner. This personalized self-annotation differs from datasets discussed in Section 2 which derive their labels from interpretations of external annotators rather than the participants' own judgements.

Post-experimental Data Processing After synchronizing and aligning the data streams with all modalities and their corresponding labels, we used up-sampling duplication and down-sampling averaging to handle the varying sampling rates across sensors. We used IBM Watson (IBM, 2023) to provide an initial transcription of the participant dialogues recorded from their lapel microphones, obtaining an average Word Error Rate (WER) of 9%. We manually corrected all transcripts. We used iMotions (2023) to extract facial action units for the instructor. In addition, FreeMoCap (Cherian et al.) was used for extracting the body posture of the builder, and for both participants SBERT (Reimers and Gurevych, 2019) was used to extract 384-dimensional pre-trained word embeddings for transcribed spoken words, surrounding timestamped labels. Each participant's sensor data was Z-score normalized across the entire recording's data to allow for comparison across groups and account for inter-subject variability.

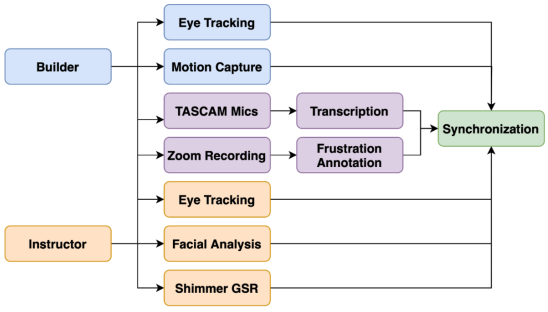

Figure 3: The orange modalities correspond to sensors for the instructor, while the blue correspond to the builder. The purple represents data recorded for both. These time series data points are synchronized from a common starting signal, ensuring annotations align with video and sensor recordings.

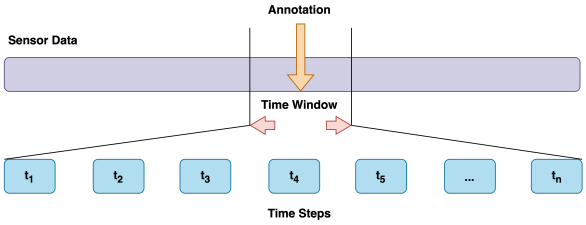

Figure 4: After acquiring annotation ratings with their corresponding timestamps, we sample sensor values within a specified time window t_w and split samples into time step t_s chunks. These values are then averaged into a single value. This is done for each modality feature $m \in \{m_1, m_2, \ldots, m_n\}$ so that values can be concatenated together to form a single instance.

MULTICOLLAB Dataset In this study, we focused on the instructor data (since half of the builders were instructed to not cooperate during the second task). We employed early fusion for our time series sensor data with two parameters: time window t_w and time step t_s. Time window t_w defined how much sensor data was sampled before and after the labeled instance of frustration. These sequential data samples were then split into time step t_s chunks, as shown in Figure 4, and averaged into a single value across each modality. This allowed us to retain the temporal information across modalities while also synchronizing sensor values with disparate polling rates. Our column size was determined by taking the number of modality features multiplied by the number of time steps, expressed formally as $|\{m_1, m_2, \ldots, m_n\}| \cdot t_s$. We captured the transcribed words surrounding the instance of labeled instructor frustration within the specified time window t_w parameter. Each labeled instance of frustration **x** was initially represented as columns of modality values and rows of time steps. We flattened this 2D representation to get a single row for each instance of instructor frustration annotation.

Table 2 shows a sample of some of the utterances spoken at instances of annotation. To get a vector representation of utterances from the instructor transcripts for the downstream classification task, we used pre-trained 384-dimensional word embeddings for each utterance. We first selected all word tokens within time window t_w, then concatenated them together by delimiting each word token with a space. This is formally defined in Equation 1, where $\sum^*$ is used as a concatenation operation:

$$U = \sum_{w \in W}^{*} w \tag{1}$$

Where U represents the union of word tokens delimited by spaces for an utterance, W denotes all of the words spoken surrounding the instance of annotation, and w represents a word token. Once combined, we could then extract a single vector to represent the spoken words surrounding annotation from the instructor transcript, as formally expressed in Equation 2:

$$\mathbf{w} = f_e\left(U, \Theta\right) \tag{2}$$

Where Θ denotes the pre-trained parameters of the SBERT's `all-MiniLM-L6-v2` model, f_e represents our embedding function, and **w** represents the vector representation of the utterance at the instance of self-annotated instructor frustration.

Table 1: The modality features used in low-level, time series transformer model experiments. These values were Z-score normalized per participant, allowing for commensurate analysis across features.

Voice Features	
Intensity (dB)	F0 (Hz)
Facial Features	
Brow Furrow	Chin Raise
Lid Tighten	Lip Corner Depressor
Eye Gaze Features	
Saccade Duration	Saccade Peak Velocity
Fixation Dispersion	Fixation Duration
Gaze Velocity	
Biophysical Feature	
GSR Conductance	

Feature Extraction and Analysis In addition to transcribed speech, the features extracted from the sensor data used in the experiments are in Table 1. The distribution of frustration annotations are in Figure 5, showing a nearly even cooperative and non-cooperative pair counts for each label. However, we can see there is an imbalance with these labels, with *Slightly Frustrated* composing the majority of annotations. To address this, we removed instances of *Slightly Frustrated* and combined instances of *Very Frustrated* and *Extremely Frustrated* to define a binary classification task. Combining the latter two classes is reasonable, as they indicate varying levels of evident frustration. The right hand side of Figure 5 shows this binary data distribution, which was used for our experiments. Figure 6 shows a comparison of Z-score normalized box plots for *Not At All Frustrated* vs. the combined *Very Frustrated* and *Extremely Frustrated* classes. Because we make use of self-annotations for labeling instructor instances of their own perceived frustration, inter-rater metrics do not apply.

Ranking Modality Features When considering data extracted from sensors, quantifying which features may contribute most to classification can be

Table 2: A sample of instructor utterances surrounding timestamped instances of frustration annotation for a time window of $t_w = 4500$ milliseconds. Each row represents a different group, where C represents cooperative builders, and NC represents non-cooperative builders.

Timestamp	Builder	Rating	Instructor Utterance
05:55.52	C	Not At All Frustrated	"*super okay now okay now let's*"
06:30.33	C	Not At All Frustrated	"*yeah okay turn your hand counter*"
04:25.22	C	Very Frustrated	"*you should see the document*"
05:15.68	NC	Very Frustrated	"*not connectors beside the yellow block*"
02:39.65	NC	Extremely Frustrated	"*small blue one no*"
05:55.95	NC	Extremely Frustrated	"*no no no no the other rectangle yep go back*"

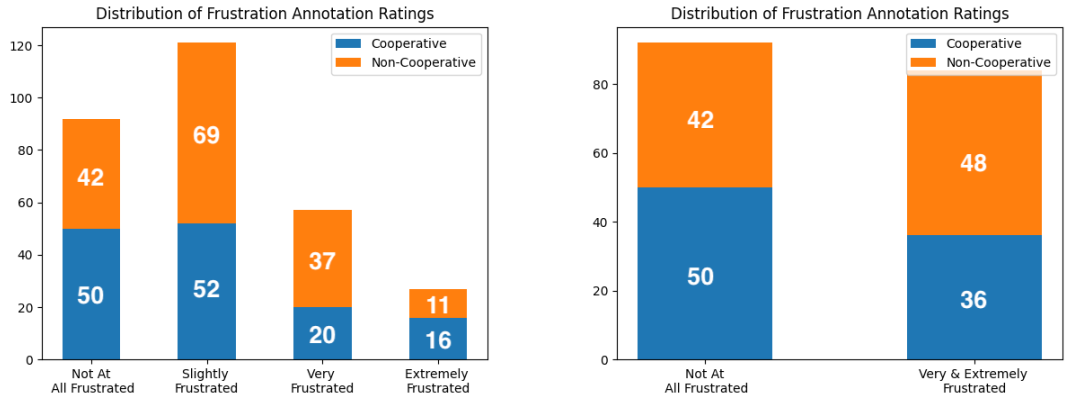

Figure 5: The label distribution of instructor frustration ratings. There is an imbalance of labels, with *Slightly Frustrated* composing the majority of labels (left). When we remove *Slightly Frustrated* instances and combine *Very and Extremely Frustrated* instances into a single class, we get a more balanced label set for binary classification (right). We use the latter distribution for our computational experiments.

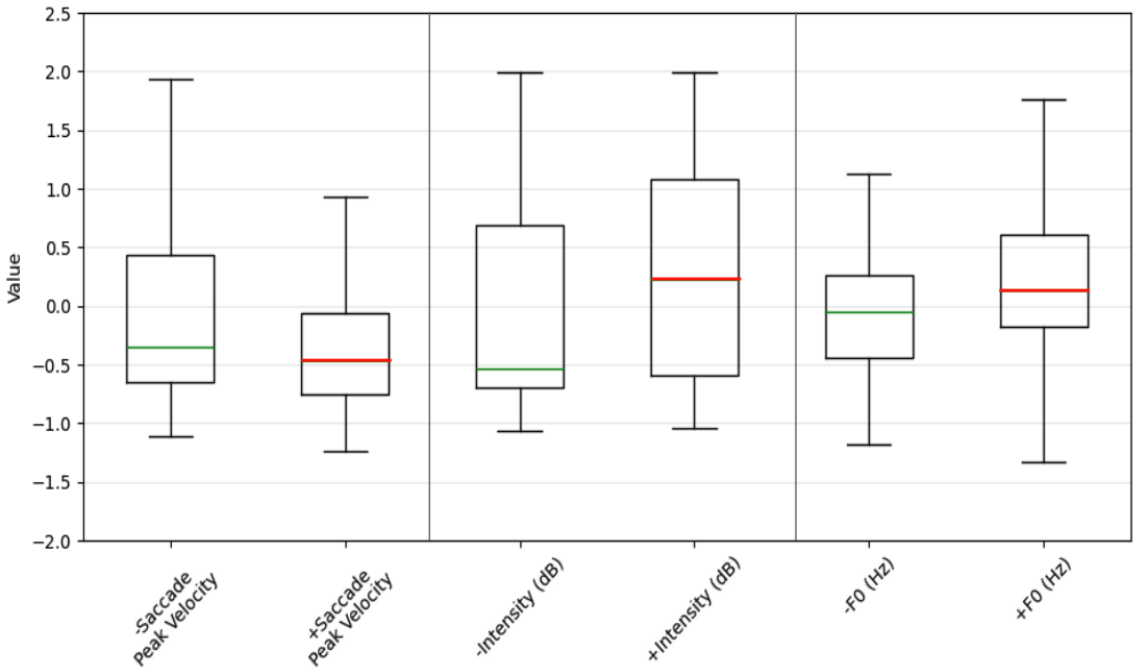

Figure 6: Box plot distribution of Z-score normalized modality values for instructors in a 400 ms window from annotation. The left box plots show *Not At All Frustrated* annotations, with a - sign next to the label. The right shows show *Very Frustrated and Extremely Frustrated* annotations combined, with a + sign next to the label. Certain modality medians stray from the *Not At All Frustrated* median. Notably, voice intensity (dB) and F0 (Hz) are higher when the instructors are frustrated.

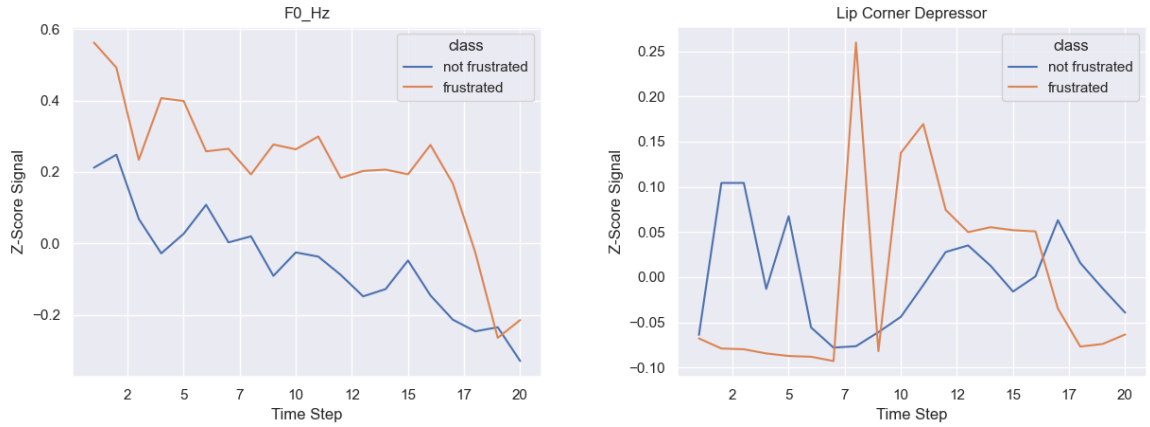

Figure 7: Line plots of the average curve values per class for the highest and lowest M_{rank} modality features for $t_w = 4500$ and $t_s = 20$. We find that certain voice features, such as F0 (left), can be useful for classification, while facial action unit features, such as Lip Corner Depressor (right), fails to separate data points for each class.

accomplished by analyzing the distribution of their Z-score normalized values - specifically by observing the overlap between the average values through time steps as formally expressed in Equation 3:

$$M_{rank} = \left| \sum_{t=1}^{t_s} \overline{\mathbf{m}_t^+} - \sum_{t=1}^{t_s} \overline{\mathbf{m}_t^-} \right| \quad (3)$$

Where $\overline{\mathbf{m}_t^+}$ represents the average Z-score normalized values for modality m at time step t for the *Very and Extremely Frustrated* class, and $\overline{\mathbf{m}_t^-}$ for the *Not At All Frustrated* class. By measuring the difference between the approximate area under the curve for average modality values through time for each class, we can determine which features show the greatest divergence. This metric is distinct from traditional Euclidean distance, as it accounts for the overlap between both curves. In essence, this quantifies which modality distribution has the greatest contrast between the two classes, which can contribute to improved model performance. Table 3 shows the feature rankings using this metric. Curiously, facial action units appear at the bottom, suggesting that relying solely on facial expression might not be an effective method for identifying frustration. Figure 7 visualizes this principle which can perhaps be attributed to individuals' tendency to suppress their facial reactions, aiming to adhere to social norms and avoid causing embarrassment to their conversation partner.

Table 3: Ranking each modality feature by M_{rank} for $t_w = 4500$ and $t_s = 20$, as expressed in Equation 3. We find evidence that voice, eye tracking, and GSR offer useful features for identifying frustration. Conversely, facial action units seem to be comparatively poor features, as their average values overlap frequently for the two classes.

Modality Feature	M_{rank}
F0 (Hz)	5.45
Saccade Peak Velocity	3.38
GSR Conductance	3.09
Fixation Dispersion	2.95
Saccade Duration	2.51
Intensity (dB)	2.16
Gaze Velocity	0.92
Chin Raise	0.81
Brow Furrow	0.57
Lid Tighten	0.38
Fixation Duration	0.36
Lip Corner Depressor	0.02

4. Results

Computational Experiment The development and test set split ensured the presence of L1—L1, L1—L2, and L2—L2 English speakers in each. We set aside data from 3 of 24 instructors for testing and used the remaining 21 for model development. Our classification experiments were performed on the binary labels rather than the imbalanced four class dataset. We first trained two separate models: one for the low-level sensor data features combined with early fusion, and another for the high-level word embeddings. We used TSAI's (Oguiza, 2022) implementation of InceptionTime (Fawaz et al., 2020) as a sequential time series transformer model on the sensor data. This required restructuring **x** to be reshaped into a 3D vector, with the dimensions reflecting $(\text{instances} \times \text{modality features} \times \text{timesteps})$. We used an 80-20 validation split for training over 25 epochs. The loss over time for $t_w = 10000$

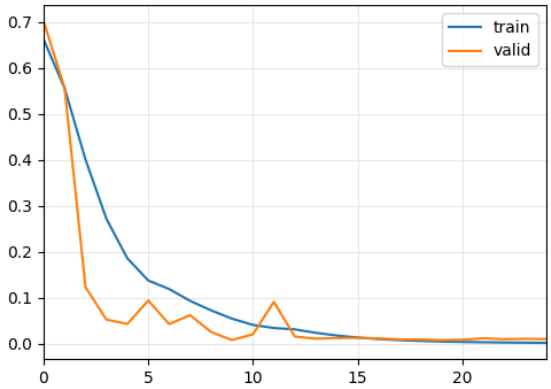

Figure 8: The training loss for the InceptionTime model on low-level sensor data for $t_w = 10000$ milliseconds and $t_s = 20$ for 25 epochs. Despite training on the development data showing learning convergence, accuracy remained low when performing predictions on the test set, suggesting that a large time window can provide conflicting data points for classification.

milliseconds is shown in Figure 8. For instructor word embeddings, we used a feed forward neural network with two fully connected layers and ReLU and sigmoid activation functions. For both models, we varied time step and window with the Cartesian product of the following configurations: $t_w \in \{1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}$ and $t_s \in \{5, 10, 15, 20\}$. After generating a model for each configuration, we performed model fusion.

Model Fusion Each model generated a two dimensional output vector for the corresponding probability percentage of each class when run against each test set instance. By averaging these two vectors, we could then select a final output class for prediction. Results in Table 4 provide answers to the research questions posed in Section 1.

5. Discussion

Regarding **RQ1**, we find that the experiments using the time series transformer with only low-level sensor data yielded a higher average accuracy than the high-level neural network with word embeddings, $\overline{L_{acc}} = 54.3 > \overline{H_{acc}} = 43.7$. For **RQ2**, we find that the best performing model was from fusing predictions from both low and high-level output probabilities, with an accuracy of 67% and a corresponding precision of 79% and F1 of 63%. Furthermore, the average accuracies of both low and high-level models were less than the fused results, with $\overline{F_{acc}} = 57.3 > \overline{L_{acc}} > \overline{H_{acc}}$. When we compare the highest ranked fused accuracy to a baseline (the average of both L_{acc} and H_{acc} at the same t_w and t_s) we observe a 21% increase in classification accuracy, i.e. $F_{acc} - \text{mean}(L_{acc}, H_{acc}) = 21.1$, demonstrating that the incorporation of sensor features can aid in identifying affect.

6. Conclusion

This study introduced a new, heavily multimodal corpus — the MULTICOLLAB dataset — which captures human to human speech interactions with other modalities and self-annotated instances of frustration. Inspired by human sensation and perception, we utilized time series transformers for sensor data (sensation) and a feed forward neural network for word embeddings (perception). An analysis on modalities showed that certain eye tracking, galvanic skin response, and voice values can be salient features for classification, with facial action units providing conflicting evidence. On average, low-level sensor data features had a higher classification performance when considered separately from high-level word embeddings. However, we demonstrated an improvement in performance when both models are utilized for final predictions.

Limitations

While the purpose of this study was to demonstrate the benefits of data fusion for low and high-level features, there are also limitations such as limited sample size. Even with the diverse subject population described in Section 3, the 24 pairs generated only 297 labeled instances of instructor frustration. This becomes even smaller when we restructure our data to a binary classification problem, leaving 176 labeled instances. Thus, further study is required to understand if results generalize. Additionally, the used SBERT word embeddings are not optimized for *spoken dialogue*. Accordingly, the nuances of utterances may not have been captured in their embeddings. Future work can explore how word embeddings derived from spoken dialogue may influence results. Lastly, human frustration perception can vary and may involve less consistency. Future research could seek to generate labels directly from biophysical signals, rather than from annotation. This may be accomplished by using labels from, for example, k-means clustering performed on the sensor data.

Ethics Statement

In this IRB-approved study, subjects gave informed consent before participating. They were given the option to end the study at any time. During the setup, each sensor was explained to the participant before data collection began. After providing

Table 4: Classification results under different time window t_w and time step t_s configurations, where L_{acc} represents low-level sensor data classification accuracy, H_{acc} represents high-level word embedding classification accuracy, and F_{acc} represents fused classification accuracy. Values are sorted by F_{acc} in descending order, with the highest value bolded in each column. We find that, on average, F_{acc} is higher than L_{acc} and H_{acc} in isolation.

$t_w(ms)$	t_s	L_{acc}	H_{acc}	F_{acc}	Precision	Recall	F1
4500	5	59.1	34.1	**67.3 ± 1.7**	79.3	51.7	**62.6**
4500	15	61.4	35.0	66.6 ± 1.0	**90.3**	40.4	55.8
4000	5	52.3	38.6	64.5 ± 1.1	74.0	49.6	59.4
3500	5	50.0	46.8	63.4 ± 3.0	71.8	49.6	58.6
3000	20	59.1	47.3	63.2 ± 0.9	76.5	42.6	54.7
4000	15	**63.6**	39.3	62.7 ± 1.1	66.3	**58.7**	62.2
5000	20	59.1	43.2	62.0 ± 1.8	83.0	34.8	49.0
3000	15	59.1	**47.7**	61.4 ± 0.0	68.8	47.8	56.4

annotations, both builders and instructors were debriefed on the nature of the study, data use, and whom to contact for any follow-up. Each participant was compensated $25 USD.

Our study brings attention to the broader issue of privacy. As sensors in consumer grade devices continue to mature, we can expect such data to be leveraged in a variety of downstream tasks. Thus, there is a motivation to identify the most salient features in these data streams to further inform users of possible privacy compromising data collection. Our findings indicate that biophysical signals such as Galvanic Skin Response (GSR) and certain eye tracking features can be used effectively to correlate self-reported instances of frustration. The impact of this research, therefore, calls for informing users of the potentially sensitive data that devices could be collecting.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Award DGE-2125362. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

7. Bibliographical References

Cecilia O Alm. 2012. The role of affect in the computational modeling of natural language. *Language and Linguistics Compass*, 6(7):416–430.

Cecilia O. Alm. 2022. Linguistic data resources for computational emotion sensing and modeling. volume 1, pages 226–249. De Gruyter Mouton, Berlin, Boston.

Akhter Al Amin, Saad Hassan, Cecilia O. Alm, and Matt Huenerfauth. 2022. Using BERT embeddings to model word importance in conversational transcripts for deaf and hard of hearing users. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 35–40, Dublin, Ireland. Association for Computational Linguistics.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Aaron Cherian, Philip Queen, Wirth Trent, Idehen Endurance, and Jonathan Samir Matthis. FreeMoCap: A free, open source markerless motion capture system.

Amir Cohen, Amir Kantor, Sagi Hilleli, and Eyal Kolman. 2021. Automatic rephrasing of transcripts-based action items. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2862–2873.

Sascha Fagel. 2006. Emotional Mcgurk effect. In *Proc. Speech Prosody 2006*, page 006 paper.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. 2020. InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Joseph F Grafsgaard, Kristy Elizabeth Boyer, and James C Lester. 2011. Predicting facial indicators of confusion with hidden markov models. In *International Conference on Affective computing and intelligent interaction*, pages 97–106. Springer.

Mohammed Ehsan Hoque, Daniel J. McDuff, and Rosalind W. Picard. 2012. Exploring temporal patterns in classifying frustrated and delighted smiles. *IEEE Transactions on Affective Computing*, 3(3):323–334.

IBM. 2023. IBM Watson. Accessed on October 13, 2023.

iMotions. 2023. Powering human insights - biometric research. Accessed on October 13, 2023.

Nikhil Kaushik, Reynold Bailey, Alexander Ororbia, and Cecilia O. Alm. 2021. Eliciting confusion in online conversational tasks. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–5.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2022. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions.

Camille Mince, Skye Rhomberg, Cecilia O. Alm, Reynold Bailey, and Alex Ororbia. 2022. Multimodal modeling of task-mediated confusion. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 188–194, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Ignacio Oguiza. 2022. tsai - a state-of-the-art deep learning library for time series and sequential data. Github.

Diego Peña, Ana Aguilera, Irvin Dongo, Juanpablo Heredia, and Yudith Cardinale. 2023. A framework to evaluate fusion methods for multimodal emotion recognition. *IEEE Access*, 11:10218–10237.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mao Saeki, Kotoka Miyagi, Shinya Fujie, Shungo Suzuki, Tetsuji Ogawa, Tetsunori Kobayashi, and Yoichi Matsuyama. 2022. Confusion detection for adaptive conversational strategies of an oral proficiency assessment interview agent. pages 3988–3992.

Andres Sanchez-Comas, Kåre Synnes, Diego Molina-Estren, Alexander Troncoso-Palacio, and Zhoe Comas-González. 2021. Correlation analysis of different measurement places of galvanic skin response in test groups facing pleasant and unpleasant stimuli. *Sensors*, 21(12):4210.

Monali Saraf, Tyrell Roberts, Raymond Ptucha, Christopher Homan, and Cecilia O Alm. 2019. Multimodal anticipated versus actual perceptual reactions. In *Adjunct of the 2019 International Conference on Multimodal Interaction*, ICMI '19, New York, NY, USA. Association for Computing Machinery.

Abdallah Hussein Sham, Amna Khan, David Lamas, Pia Tikka, and Gholamreza Anbarjafari. 2023. Towards context-aware facial emotion reaction database for dyadic interaction settings. *Sensors (Basel, Switzerland)*, 23.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018b. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018c. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2236–2246.

Ziheng Zeng, Snigdha Chaturvedi, and Suma Bhat. 2017. Learner affect through the looking glass: Characterization and detection of confusion in online courses. *International Educational Data Mining Society*.

Inc. Zoom Video Communications. 2023. One platform to connect | zoom. Accessed on October 13, 2023.