Integrative Sentiment Analysis: Leveraging Audio, Visual, and Textual Data

Jason S. Chu and Sindhu Ghanta

[1] Monta Vista High School, Cupertino, CA, USA
[2] AIClub, Mountain View, CA, USA

Abstract. Exploring the area of multimodal sentiment analysis, this paper addresses the growing significance of this field, driven by the exponential rise in multimodal data across platforms like YouTube. Traditional sentiment analysis, primarily focused on textual data, often overlooks the complexities and nuances of human emotions conveyed through audio and visual cues. Addressing this gap, our study explores a comprehensive approach that integrates data from text, audio, and images, applying state-of-the-art machine learning and deep learning techniques tailored to each modality. Our methodology is tested on the CMU-MOSEI dataset, a multimodal collection from YouTube, offering a diverse range of human sentiments. Our research highlights the limitations of conventional text-based sentiment analysis, especially in the context of the intricate expressions of sentiment that multimodal data encapsulates. By fusing audio and visual information with textual analysis, we aim to capture a more complete spectrum of human emotions. Our experimental results demonstrate notable improvements in precision, recall and accuracy for emotion prediction, validating the efficacy of our multimodal approach over single-modality methods. This study not only contributes to the ongoing advancements in sentiment analysis but also underscores the potential of multimodal approaches in providing more accurate and nuanced interpretations of human emotions.

Keywords: *Transformers, multi-modal, sentiment analysis*

1 Introduction

In the domain of human-computer interaction, sentiment analysis has traditionally emphasized textual data, often at the expense of the rich emotional nuances inherent in spoken language and visual cues. This focus has historically been driven by the ease of processing text in natural language processing (NLP) systems. However, the advent of advanced communication technologies and the increase in popularity of social media platforms, such as Facebook and YouTube, have catalyzed a significant shift. In the YouTube platform alone, around 3.7 million videos are uploaded daily. To put this in perspective, 500 hours of video content are uploaded to YouTube every minute [1][2]. The exponential increase in multimodal data – encompassing text, audio, and visual information – presents a compelling challenge and opportunity for sentiment analysis.

This data has intricate expressions of sentiment that surpass the predictions that can be made from text-only analysis. Beyond mere text, this new paradigm

David C. Wyld et al. (Eds): IOTBS, NLTM, AIMLA, DBDM -2024
pp. 155-169, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.1402011

aims to integrate audio and visual modalities, thereby capturing a more comprehensive spectrum of human sentiment. In applications ranging from human-machine conversation to autonomous driving, this multimodal approach to sentiment analysis is increasingly playing a pivotal role, moving the field towards a more inclusive and accurate interpretation of human emotions. This ability of computers and machines to understand emotions has always been crucial to assisting human requirements. As technology is increasingly used on various platforms, this understanding becomes even more essential. The ability to interpret subtle emotional cues can greatly enhance the interaction, making it more natural and effective.

Audio modality, characterized by variations in vocal attributes such as pitch and loudness, adds a layer of emotional characteristics that text alone may fail to capture. Similarly, visual cues from facial expressions and gestures provide vital sentiment information that complements textual analysis. The interaction between these modalities enriches the sentiment analysis process, enabling more accurate and nuanced emotion recognition. This multimodal approach is particularly beneficial in cases where text-based analysis faces ambiguity. For instance, the sentiment conveyed in spoken words can often be clarified through the speaker's intonations or expressions, which might be lost in textual transcription alone.

Research in the field of sentiment analysis and emotion recognition has been quite diverse and innovative. Studies have utilized a variety of text datasets, such as TED-LIUM release 2, Movie Review, and Twitter corpora, employing techniques like word2vec for converting text into numerical formats. These numerical representations are then analyzed using machine learning and deep learning models, including Support Vector Machines (SVMs) and convolutional neural networks (CNNs)[3].

Additionally, significant work has been conducted in recognizing emotions from audio data. One notable approach in this area has been the use of the robust wav2vec 2.0 system for dimensional emotion recognition, which has contributed significantly to the field [4].

In the area of emotion detection from visual data, researchers have made strides in classifying emotions using images of human faces sourced from platforms like Flickr or the face databases such as BU-3DFE [8][9]. These studies have experimented with a range of classification methods, from low level feature extraction to high-level features derived from architectures like VGG-ImageNet, ResNet, followed by classification using different algorithms such as SVM and multi-layer perceptron [6].

In contrast to the above work, there is a tremendous amount of interest in the field of multimodal sentiment analysis [7]. Researchers have been working on different combinations of text, audio, and image modalities that can enhance prediction accuracy [10]. This area of study focuses on various methods of integrating multimodal information, primarily through feature fusion and decision fusion, as

outlined in several studies [22][12][5]. The choice of method often depends on the specific application and the nature of the data being analyzed.

There is a diverse range of applications for multimodal sentiment analysis. For instance, spoken reviews and vlogs have been a key focus, utilizing the richness of multimodal data to gauge sentiment more effectively [13][15][16]. Another significant application area is in human-machine and human-human interactions, where understanding the sentiment is crucial for enhancing communication [18].

Additionally, visual sentiment analysis, which examines images and their associated tags on social media platforms, offers insightful perspectives on public sentiment and trends [14]. This approach is particularly relevant in the current digital age, where social media forms a substantial part of human interaction and expression. These studies illustrate the breadth and potential of multimodal sentiment analysis in understanding and predicting human emotions across various digital platforms and interaction scenarios.

Feature fusion is where the features from different modalities are fused to create a richer set of features based on which better predictions can be made. A method to combine text and audio features is proposed in [17] followed by a neural network to make predictions on the emotion category. Several other feature fusion techniques exist in literature to combine the features in interesting ways [20],[17].

Decision fusion on the other hand involves independently analyzing and classifying features from different modalities. The outcomes are then integrated into a unified decision vector for the final sentiment determination [24]. A notable application of this approach is demonstrated in the work of [19], where distinct models for each of the three modalities are combined to form an ensemble. This approach is developed in the context of visual data sourced from customer interactions on social media platforms like Instagram, Facebook, and Twitter, as well as from feedback forms and products, classifying sentiments as either positive or negative [19]. Another significant contribution in this field is from [23], where the researchers conducted a comprehensive analysis using a smaller dataset of 47 videos. In this study, both feature-level and decision-level fusion techniques were employed, showcasing the versatility and effectiveness of multimodal sentiment analysis in processing complex data sources. These studies highlight the growing sophistication in sentiment analysis methodologies, where integrating multiple modalities offers a more nuanced understanding of user sentiments, especially in the context of social media and customer feedback.

Building on the significant progress made in sentiment analysis and emotion recognition, our study seeks to enhance these efforts by focusing on decision fusion across three modalities: text, audio, and visual. We utilize the expansive MOSEI dataset, which contains multimodal content from YouTube, comprising over 23,500 spoken sentence videos and totaling more than 65 hours. This dataset provides a diverse and comprehensive platform for our analysis where each sample is classified

into six emotions. Our approach parallels the evolving trends in multimodal data analysis and aims to develop more sophisticated models.

2 Materials and Methodology

2.1 Dataset

In this study, we utilized the MOSEI dataset, as developed by [27] in 2018, for training, testing, and validating our model. This extensive dataset includes more than 23,500 videos of spoken sentences, amounting to over 65 hours of content. It was meticulously organized at the sentence level, with each sentence transcribed and supplemented with audio, visual, and textual features. These features, along with the raw video footage, were made accessible through a software development kit released by Zadeh et al. in 2018. The dataset uniquely categorizes each video for sentiment analysis on a scale from -3 to 3 and identifies six distinct emotions for emotion analysis, all scored by human evaluators. The distribution of the dataset is shown (Figure 1). In this study, we focus on the prediction of the six distinct emotions happiness, sadness, anger, disgust, surprise and fear. Note that this is a multi-label classification problem, which means that each video clip could be positive for several of the emotions.

2.2 Methods

In this section, we present the methods used for training the machine learning and deep learning models for text and audio features. We also describe the model used for visual emotion recognition.

For text data, experiments were conducted using the BERT (Bidirectional Encoder Representations from Transformers) model. As a preliminary step, textual data must be transformed into numerical representations to be processed by deep learning models. To accomplish this, we utilized AutoTokenizer from the Transformers library, which encodes the text with techniques such as padding and truncation, adhering to a specified maximum length to standardize input sizes. Additionally, label matrices were generated and appended to these encoded representations.

Our model underwent hyperparameter tuning, where we experimented with various values for the number of epochs (ranging from 30 to 100) and the learning rate (ranging from 0.0001 to 0.01). The model with the highest validation accuracy was selected and saved for subsequent predictions on the test subset of the data.

For audio analysis, we used the librosa library to extract relevant features. Specifically, we focused on Mel-frequency cepstral coefficients (MFCCs) derived from the audio signal. These MFCCs were processed to compute their mean and standard deviation along the time axis and were used as features for each audio track.

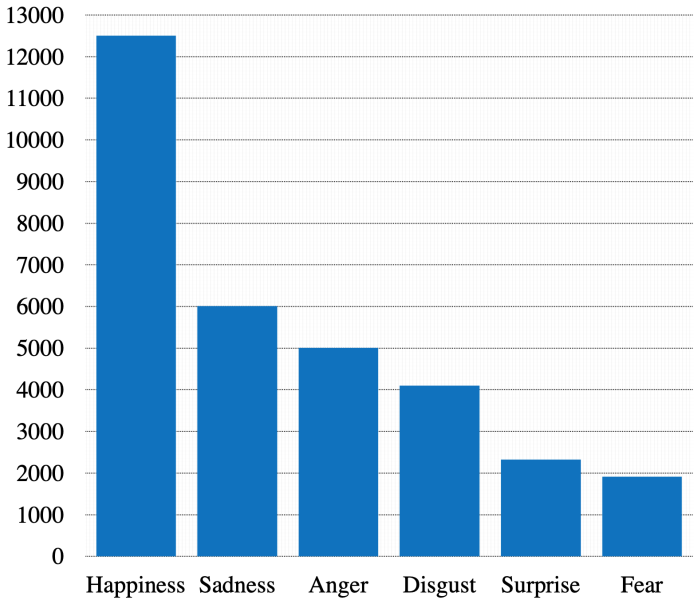


Fig. 1: The distribution of the dataset is shown. A skewing towards the more common emotions is present. This histogram was taken from the CMU MOSEI dataset paper [28].

Our approach to predicting emotions from audio features involved two distinct techniques. In the first technique, we employed two algorithms: Support Vector Classifier (SVC) and Random Forest Classifier (RFC), within a multi-output classifier framework to enable the prediction of all six emotion labels using a single model. We fine-tuned hyperparameters, such as max depth and max width (ranging from 1 to 9) for RFC, as well as kernel type and degree for polynomial kernels in SVC. The performance of these models was evaluated on the validation dataset.

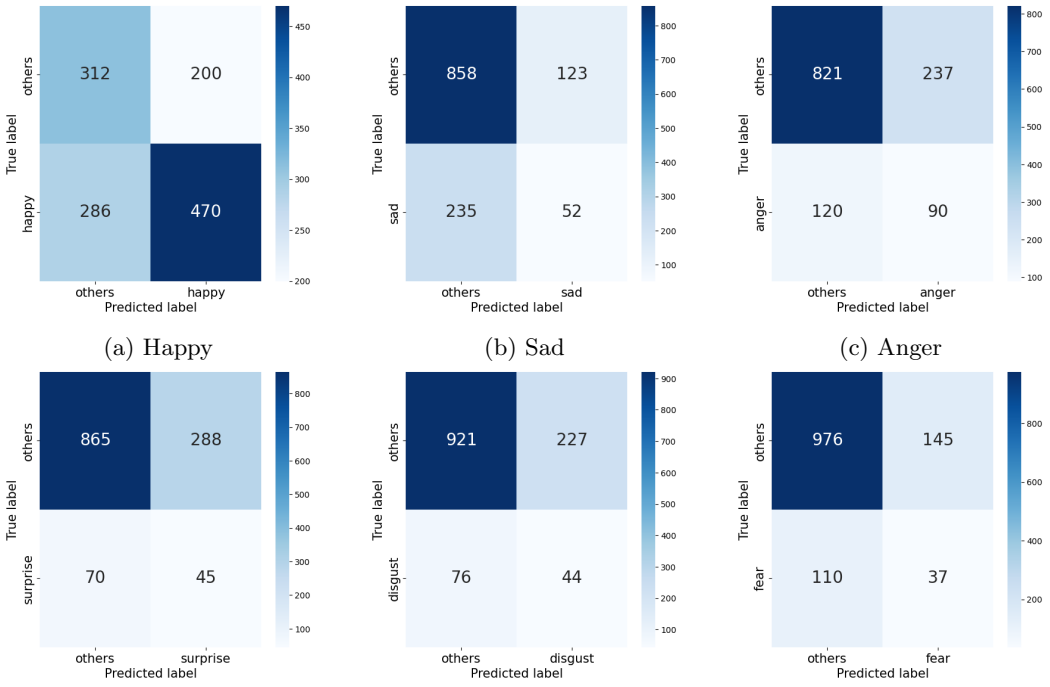
In the second technique, we pursued a different strategy by building a separate model for each emotion category. This approach was motivated by the dataset's inherent skew, which proved challenging to rectify when using a single model in the first approach. However, by creating individual binary classifiers for each emotion, we balanced the training set before model training. For this purpose, we explored three algorithms: K-Nearest Neighbors (KNN), RFC, and SVC. In the case of KNN, we varied the hyperparameter representing the number of neighbors between 2 and 14. Meanwhile, for RFC, we adjusted max depth and the number of trees, exploring values ranging from 10 to 100 and 1 to 7, respectively. When using SVC, we experimented with two different kernel types: Radial Basis Function (RBF) and polynomial. For polynomial kernels, we tested a range of degrees from 1 to 13.

For images, we imported the pre-trained FER model and did not conduct any training using our dataset [26].

To arrive at the final emotion prediction for each sample in the test set, we combined predictions from all three modalities. This involved averaging the predictions from each modality to calculate the ultimate prediction for each emotion label.

The evaluation of these models involved the use of several metrics, including overall accuracy values and the generation of confusion matrices for individual emotion categories.

3 Results

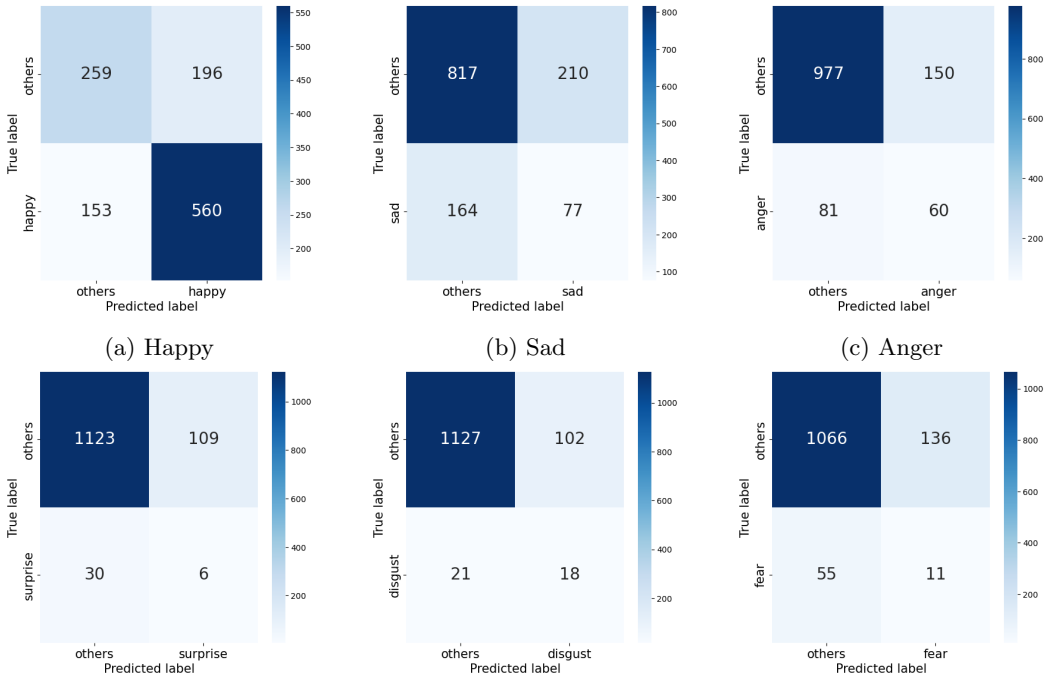


(a) Happy (b) Sad (c) Anger

(d) Surprise (e) Disgust (f) Fear

Fig. 2: Confusion matrix based on predictions of the combinations of the optimal audio models for each emotion.

In this section, we present a comprehensive analysis of our experiments across different modalities: audio, text, and images.



(a) Happy (b) Sad (c) Anger

(d) Surprise (e) Disgust (f) Fear

Fig. 3: Confusion matrix based on predictions of the optimal BERT model on the test subset of the data.

3.1 Audio Results

In the first set of experiments with audio, we utilized Support Vector Classifier (SVC) and Random Forest algorithms. For SVC, we utilized three different kernel types including, linear, polynomial, and Radial Basis Function (RBF). For the polynomial type, we experimented with varying degrees. Ultimately, the highest validation accuracy reached with SVC was 40.85%, with a polynomial of 3rd degree and the RBF kernel. In addition to SVC, we used Random Forest Classifier (RFC), with a range of 5 to 35 for max depth, and 1 to 9 for max width. Varying results were achieved, with the highest validation accuracy reaching 41.88% with a max depth of 7 and a max width of 20. However, the results were similar to the accuracies of the SVM model, although it did outperform slightly.

Next, we employed three algorithms as binary classifiers for each emotion to enhance the accuracy of the audio modality. We tested K-Nearest Neighbors (KNN), RFC, and SVC on all six emotions. The best-performing model for each emotion and its corresponding hyperparameters are summarized below:

For the “happy” emotion, the KNN model achieved the highest validation accuracy of 63.69% with 2 neighbors as reported in Table 1. For the “sad” emotion,

# Neighbors	2	3	4	5	6	7	8	9	10	11	12	13	14
Accuracy	0.58	0.64	0.61	0.64	0.62	0.63	0.61	0.63	0.63	0.63	0.62	0.63	0.62

Table 1: Accuracy values for different numbers of neighbors in KNN classification for the emotion happy

the SVC model performed the best, achieving a validation accuracy of 74.13% with a polynomial kernel of degree 10 as reported in Table 2. The KNN model outperformed other models for the "anger" emotion, reaching a validation accuracy of 73.74% with 2 neighbors as reported in Table 3.

Kernel	**rbf**	**poly**											
Degree	**N/A**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Accuracy	0.57	0.57	0.60	0.66	0.69	0.71	0.72	0.74	0.74	0.74	0.74	0.74	0.74	0.74

Table 2: Accuracy values for the emotion sad when support classification model hyper-parameters are tuned for the emotion sad

# Neighbors	2	3	4	5	6	7	8	9	10	11	12	13	14
Accuracy	0.74	0.60	0.7	0.60	0.69	0.61	0.68	0.60	0.66	0.59	0.65	0.58	0.64

Table 3: Accuracy values for different numbers of neighbors in KNN classification for the emotion anger

# Neighbors	2	3	4	5	6	7	8	9	10	11	12	13	14
Accuracy	0.72	0.53	0.65	0.50	0.62	0.51	0.6	0.5	0.6	0.51	0.58	0.50	0.58

Table 4: Accuracy values for different numbers of neighbors in KNN classification for the emotion surprise

Similarly, for the "surprise" emotion, the KNN model attained the highest validation accuracy of 72.08% with 2 neighbors as reported in Table 4. For the "disgust" emotion, the KNN model once again outperformed other models with a validation accuracy of 76.74% using 2 neighbors as reported in Table 5. Finally, for the "fear" emotion, the SVC model demonstrated the best performance, reaching a validation accuracy of 80.44% with the RBF kernel as reported in Table 6.

# Neighbors	2	3	4	5	6	7	8	9	10	11	12	13	14
Accuracy	0.77	0.6	0.71	0.59	0.69	0.59	0.68	0.61	0.68	0.61	0.68	0.6	0.67

Table 5: Accuracy values for different numbers of neighbors in KNN classification for the emotion disgust

Kernel	**rbf**	**poly**											
Degree	**N/A**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Accuracy	0.49	0.47	0.47	0.54	0.43	0.47	0.59	0.61	0.68	0.77	0.80	0.80	0.79	0.77

Table 6: Accuracy values for the emotion sad when support classification model hyper-parameters are tuned for the emotion fear

We selected the best models for each emotion based on their respective validation accuracies and employed them to make predictions on the test data. The resulting confusion matrix is illustrated in Figure 2.

3.2 Text Results

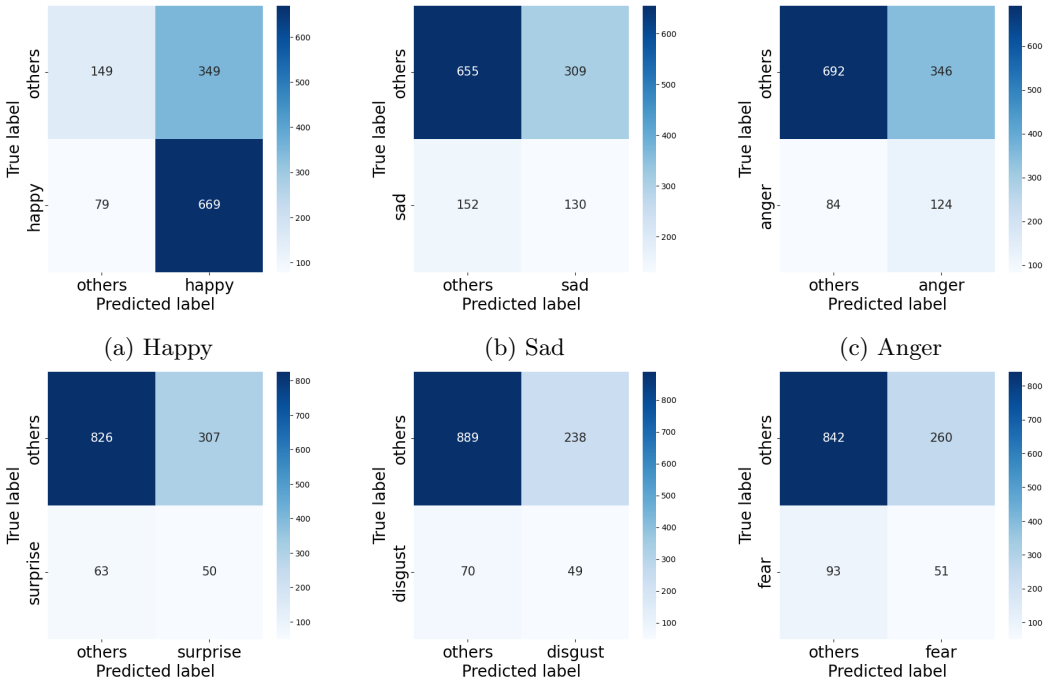
Table 7: F1 Scores based on Different Learning Rates and Weight Decay Values, when run for 30 Epochs.

Weight Decay	**Learning Rate**				
	0.001	**0.0005**	**0.0001**	**0.00005**	**0.00001**
0.05	0.4204	0.5462	0.5462	0.4994	0.4957
0.01	0.5462	0.5462	0.5462	0.5005	0.499
0.001	0.5462	0.5462	0.5462	0.5097	0.5462
0.0001	0.5462	0.5462	0.5462	0.5497	0.4947

For the text modality, we tested the BERT model at varying learning rates and weight decay values, reaching accuracies around 50% (Table 2). Results were run with 30 epochs for each experiment. The confusion matrices for each class of the test dataset are also shown (Figure 2).

3.3 Image Results

The last model we used was Facial Expression Recognition (FER), a pre-trained model for emotion detection from images. After running our test dataset through FER, we reached an overall accuracy of 64.74%. The accuracy reached per class is as follows: 55.63% for happy, 52.42% for sad, 88.86% for disgust, 53.40% for anger, 79.00% for surprise, and 59.14% for fear.



(a) Happy (b) Sad (c) Anger

(d) Surprise (e) Disgust (f) Fear

Fig. 4: Confusion matrix based on predictions of the ensemble model on the test subset of the data.

3.4 Emsemble Method

Since the individual modalities did not show the most promising accuracies, the last step was to combine the modalities and form an ensemble model. The final results were obtained by adding the results from the individual modality's predictions on the test dataset and comparing the final value with a threshold. The best threshold was 0.11, which resulted in final accuracies of 65.65% for happy, 63.30% for sad, 65.49% for anger, 69.19% for surprise, and 71.67% for fear, and 75.28% for disgust.

4 Discussion and Conclusion

Initially, we were able to successfully utilize machine learning models for each modality, although the accuracies and results reached were not the most desirable.

For audio, we used two approaches, one where a single model was used to predict all 6 emotion and the other where six different binary classifiers were used, one for each emotion. The first approach of using a single model did not yield good results. We attribute this to the skew in data which is very hard to correct in a multi-label classifier.

To overcome this, a second strategy where a single model per category of the emotion was utilized. For each emotion, a separate model was trained and tuned. Table 1 illustrates the relationship between the number of neighbors used in a K-Nearest Neighbors (KNN) classification algorithm and the corresponding accuracy in identifying the emotion "happy". A noteworthy observation is that the accuracy does not follow a linear improvement with the increase in the number of neighbors. The highest accuracy observed is 0.64, achieved with 3 and 5 neighbors. The table indicates that increasing neighbors beyond a certain point does not necessarily enhance the accuracy, as seen with the fluctuating accuracy values (ranging from 0.61 to 0.64) across different neighbor counts.

For the emotion sad, Table 2 compares the accuracy of two kernel types: radial basis function (rbf) and polynomial (poly), across various polynomial degrees. Here, the radial basis function kernel and the polynomial kernel with a degree of 1 yield the same accuracy (0.57). However, as the degree of the polynomial kernel increases, there is a consistent improvement in accuracy, up to a degree of 6. Beyond degree 6, the accuracy plateaus at 0.74, indicating that further increasing the polynomial degree does not contribute to additional gains in accuracy. This pattern suggests that for this specific task of classifying the emotion "sad", a polynomial kernel with a degree between 6 and 13 is optimal. The plateauing of accuracy beyond a certain degree highlights the phenomenon of diminishing returns, where further complexity in the model does not yield proportional improvements.

Table 3 shows an interesting pattern of accuracy in the KNN algorithm as the value of K is varied. The highest accuracy is observed with 2 neighbors (0.74), which is notably higher than the accuracies for other neighbor counts. This could suggest that for this specific classification task, a smaller number of neighbors is more effective. However, the accuracy decreases to 0.60 with 3 neighbors and continues to oscillate between 0.59 and 0.70 as the number of neighbors increases. This inconsistency might be indicative of the sensitivity of the KNN algorithm to the number of neighbors in this particular context.

Similarly, for the emotion surprise, Table 4 shows a noticeable fluctuation in accuracy as the number of neighbors changes, indicating a complex relationship between the number of neighbors and the model's performance. In the case of the disgust emotion, Table 5 shows the highest accuracy is achieved with 2 neighbors (0.77), indicating a strong performance with a very localized neighborhood. However, as the number of neighbors increases to 3, there is a noticeable drop in accuracy to 0.60. This pattern of fluctuation continues as the number of neighbors increases, with accuracy values oscillating between 0.59 and 0.71.

Finally, Table 6 shows that as the degree of the polynomial kernel increases, there is a noticeable improvement in accuracy. This trend is particularly significant from degree 8 onwards, where the accuracy jumps to 0.68 and continues to increase, peaking at 0.80 for degrees 10 and 11. The data indicates that higher-degree poly-

nomial kernels are more effective for classifying the emotion "fear" in this context. The increase in accuracy with higher degrees may be attributed to the model's enhanced ability to capture more complex patterns in the data, which lower-degree polynomials or the rbf kernel might not effectively model.

For text, the BERT model utilized the different learning rates and epochs resulting in a wide range of validation F1 values. It is evident from the data that the F1 scores are relatively consistent across a wide range of learning rates, primarily hovering around the 0.54 mark for most configurations. This stability in performance suggests that the model is robust to changes in learning rate within the tested range. Notably, the highest learning rate tested (0.05) resulted in a noticeably lower F1 score of 0.42. This decrease might indicate that at higher learning rates, the model's optimization process overshoots optimal solutions, leading to less effective learning. The variation in weight decay values appears to have a negligible impact on the model's F1 scores. This could imply that the regularization effect, which weight decay is intended to provide, may not be a crucial factor for the model's learning process in this particular task. It suggests that the model's performance is more significantly influenced by other factors, such as the nature of the dataset or the inherent architecture of the BERT model. The most optimal settings, in terms of balancing the learning rate and weight decay, seem to converge around a learning rate of 0.00005 with a weight decay of 0.0001, yielding an F1 score of 0.55. This specific combination offers a slight improvement over other configurations, indicating its potential as the best setting for our model's training on this dataset. The uniformity in F1 scores across various learning rates and weight decays suggests that the model is effectively capturing the complexity of the task at hand. For images, we used a pre-trained FER model and hence did not conduct any experiments.

To address the overall low accuracies achieved throughout the three single modalities, we completed the final step of combining the models to result in the final predictions, where we were able to increase the individual accuracy, precision and recall values across all classes. For the emotions, happy and sad, the ensemble model confusion matrix shows a reduction in false positives and false negatives compared to the single models. This improvement indicates that the ensemble model better captures the nuances of "happy" and "sad" emotions, possibly by effectively combining audio, visual and textual cues. The confusion matrix for "anger" in the ensemble model shows a marked improvement in distinguishing "anger" from similar emotions like "disgust" or "fear". This improvement could be due to the ensemble's ability to balance the intensity captured in audio with specific linguistic cues in the text. Similarly, for surprise, disgust and fear, a much better balance is observed in the confusion matrix between false positives and true positives. This highlights the usefulness of such a multimodal methodology, as the significant improvement reflects the combined usage of each of the trends of classification that

each model was able to extract. This integrated approach is particularly beneficial in distinguishing between emotions that are closely related or often exhibit overlapping characteristics in single-mode analyses. The ensemble model's effectiveness highlights the importance of multimodal emotion recognition systems in complex real-world applications.

However, many different approaches could be utilized in future work to further improve the accuracy beyond our multimodal approach. Firstly, for audio, increasing the number of features used in featurization could better depict the patterns of the words and emotions expressed for the audio model. Testing different methodologies for featurization other than the python librosa package and MFCCS used would also be a viable option to increase the accuracy by attempting to capture richer information from the audio data. In addition, using different techniques such as neural networks or deep learning techniques instead of the KNN, RFC and SVC models used could also benefit. Testing different models for text, or exploring ensemble techniques specific to text classification such as combining diverse text models, including transformer-based models like BERT with traditional machine learning models, might also improve the overall predictive power of the text analysis component. Various models and architectures, including different CNNs, could be explored for the image modality. Finally, different fusion techniques could be tested other than averaging the results across modalities.

References

1. 100+ YouTube Statistics in 2024: Users, Revenue More. (n.d.). Notta. Retrieved January 13, 2024, from https://www.notta.ai/en/blog/youtube-statistics
2. Adavelli, M. (2023, April 27). Frequently Asked Questions. Techjury. https://techjury.net/blog/how-many-videos-are-uploaded-to-youtube-a-day/
3. Bertero, D.; Siddique, F.B.; Wu, C.S.; Wan, Y.; Ho, R.; Chan, Y.; Fung, P. Real-Time Speech Emotion and Sentiment Recognition for Interactive Dialogue Systems. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 1042–1047.
4. Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., Schuller, B.W. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. arXiv 2022, arXiv:2203.07378.
5. Poria, S., Chaturvedi, I., Cambria, E., Hussain, A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 12–15 December 2016; pp. 439–448.
6. Gajarla, V., Gupta, A. (2015). Emotion detection and sentiment analysis of images. Georgia Institute of Technology, 1, 1-4.
7. Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. F., Pantic, M. (2017). A survey of multimodal sentiment analysis. Image and Vision Computing, 65, 3-14.
8. Padgett, C., Cottrell, G. (1996). Representing face images for emotion classification. Advances in neural information processing systems, 9.
9. Zheng, W., Tang, H., Lin, Z., Huang, T. S. (2010). Emotion recognition from arbitrary view facial images. In Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI 11 (pp. 490-503). Springer Berlin Heidelberg.

10. Cai, L., Hu, Y., Dong, J., Zhou, S. (2019). Audio-textual emotion recognition based on improved neural networks. Mathematical Problems in Engineering, 2019, 1-9.
11. Gao, S., Chen, X., Liu, C., Liu, L., Zhao, D., Yan, R. (2020, April). Learning to respond with stickers: A framework of unifying multi-modality in multi-turn dialog. In Proceedings of the Web Conference 2020 (pp. 1138-1148).
12. Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion 37 (2017), 98–125.
13. Morency, L. P., Mihalcea, R., Doshi, P. (2011, November). Towards multimodal sentiment analysis: Harvesting opinions from the web. In Proceedings of the 13th international conference on multimodal interfaces (pp. 169-176).
14. Borth, D., Ji, R., Chen, T., Breuel, T., Chang, S. F. (2013, October). Large-scale visual sentiment ontology and detectors using adjective noun pairs. In Proceedings of the 21st ACM international conference on Multimedia (pp. 223-232).
15. Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., Morency, L. P. (2013). Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intelligent Systems, 28(3), 46-53.
16. V. Pérez Rosas, R. Mihalcea and L. -P. Morency, "Multimodal Sentiment Analysis of Spanish Online Videos," in IEEE Intelligent Systems, vol. 28, no. 3, pp. 38-45, May-June 2013, doi: 10.1109/MIS.2013.9.
17. Zhou, S., Jia, J., Wang, Q., Dong, Y., Yin, Y., Lei, K. (2018, April). Inferring emotion from conversational voice data: A semi-supervised multi-path generative neural network approach. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).
18. C. Langlet and C. Clavel, "Adapting sentiment analysis to face-to-face human-agent interactions: From the detection to the evaluation issues," 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 2015, pp. 14-20, doi: 10.1109/ACII.2015.7344545.
19. Vasanth, K., Srideviponmalar, P., Shete, V., Ravi, C. N. (2022). Dynamic Fusion of Text, Video and Audio models for Sentiment Analysis. Procedia Computer Science, 215, 211-219.
20. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250.
21. Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., Morency, L. P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. arXiv preprint arXiv:1806.00064.
22. Yang, K., Xu, H., Gao, K. (2020, October). Cm-bert: Cross-modal bert for text-audio sentiment analysis. In Proceedings of the 28th ACM international conference on multimedia (pp. 521-528).
23. Poria, S., Cambria, E., Howard, N., Huang, G. B., Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing, 174, 50-59.
24. Dobrišek, Simon, et al. "Towards efficient multi-modal emotion recognition." International Journal of Advanced Robotic Systems 10.1 (2013): 53.
25. Liang, P. P., Salakhutdinov, R., Morency, L. P. (2018, July). Computational modeling of human multimodal language: The mosei dataset and interpretable dynamic fusion. In First Workshop and Grand Challenge on Computational Modeling of Human Multimodal Language (Vol. 113, pp. 116-125).
26. Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., Bengio, Y. (2013). Challenges in Representation Learning: A report on three machine learning contests. arXiv:1307.0414.
27. Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., & Morency, L.-P. (2018). Multi-attention recurrent network for human communication comprehension. In Thirty-Second AAAI Conference on Artificial Intelligence.

28. Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., Morency, L. P. (2018, July). 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) (pp. 2236-2246).

Authors

Jason Chu is a junior at Monta Vista High School, with extensive interest and passion for machine learning algorithms and their applications. He has conducted previous research on the usage of deep learning models for osteosarcoma (cancer) detection in images using a variety of algorithms and methods. He also has research experience with computer vision, natural language processing and audio based machine learning.

Sindhu Ghanta received the M.S. degree from Texas Tech University in 2010 and the Ph.D. degree in electrical and computer engineering from Northeastern University, Boston, USA, in 2014. She was a Post-Doctoral Fellow with BIDMC and the Department of Pathology, Harvard Medical School, where she was involved in detection and classification of features from histopathological (breast cancer) images. She worked as a research scientist with Parallel Machines on monitoring the health of machine learning algorithms in production and has many publications on ML innovations. She currently works as the Head of Machine Learning in AIClub.