End-semester Review

Team Valkyrion

Shubham Goel, Yatharth Nehra, Manan Gadesha

Problem Statement

"Building an ML model to caption videos by decoding emotions from faces, voices, and transcripts using Emotional Sentiment Analysis."

Potential Applications

1. Accessibility and Inclusion

- Emotionally Intelligent Captions: Traditional captions are limited to the type of content provided. Emotion-aware captions can communicate emotional tone, facial expressions, and vocal nuances. This benefits deaf and neurodivergent users by making video content more engaging.

2. Content Recommendation and Personalization

- Emotion-based Recommendations: By understanding the emotional context of videos, platforms can recommend content that understands a user's current mood, leading to more personalized and satisfying user experiences.

3. Human-Computer Interaction

- Empathetic AI and Virtual Agents: Emotionally aware captions and responses can make AI-driven assistants, chatbots, and virtual avatars more empathetic, natural, and trustworthy in their interactions with users.

Potential Impact of the Solution

A. Social and Cultural Impact

- *Greater Inclusion: Emotionally nuanced captions bridge communication gaps, particularly for marginalized or differently-abled populations, fostering a more inclusive digital environment15.*
- *Cultural Sensitivity: Models trained on diverse datasets can better interpret and represent emotions across cultures, reducing bias and miscommunication1.*

B. Technological Advancement

- *State-of-the-Art Multimodal AI: Integrating facial, vocal, textual, and (eventually) body language cues pushes the boundaries of multimodal machine learning, setting new benchmarks for accuracy and expressiveness in video understanding341.*
- *Improved Model Robustness: By moving beyond unimodal systems, the model is less likely to misinterpret or miss subtle emotional cues, offering richer, context-aware outputs134.*

C. Economic and Market Impact

- *New Business Opportunities: Emotionally intelligent video captioning opens new markets in accessibility tech, personalized media, education, and healthcare.*
- *Enhanced User Engagement: Platforms with richer, more relatable content can see increased user retention and satisfaction.*

Literature Review

GAP 1 - Lack of multimodal and diverse datasets

The Problem -
Many models, regardless of their performance have very limited modalities included due to the lack of use of multimodal datasets.
Case -
SentiCap - Although it has achieved a very respectable accuracy of 88% in emotion classification tasks, its captions remain emotionally neutral as it cannot infer minute emotional changes from pure text and punctuation

Das, R., & Singh, T. D. (2023). Multimodal sentiment analysis: A survey of methods, trends, and challenges. ACM Computing Surveys, 55(13s), 1–38. https://doi.org/10.1145/3586075

Mathews, A. P., Xie, L., & He, X. (2016). SentiCap: Generating image descriptions with sentiments. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30, No. 1).

Literature Review

GAP 2 - Emotion Recognition beyond facial expressions

The Problem -
Most existing systems focus narrowly on facial expressions and speech, neglecting the broader nonverbal cues conveyed by body posture and gestures
Case -
Liu, H. (2015). Emotion detection through body gesture and face. Frontiers in Psychology **-** By integrating pose features through the usage of OpenPose, he was able to prove that emotions are more than just the facial features, especially in the context of video data for ESA.

Liu, H. (2024). Emotion detection through body gesture and face. arXiv. https://arxiv.org/abs/2407.09913

Literature Review

GAP 3 - Emotion-Specific Captioning

The Problem -

Existing models may be very astute when it comes to classifying emotions with a high accuracy, but even they come up short when it comes to generating authentic, emotionally driven captioning.

Case -

SentiCap **-** it employs a switching Recurrent Neural Network (RNN) architecture to generate image captions with positive or negative sentiments. It uses a sentiment classifier to guide the caption generation process. However, the model is limited to binary sentiment classification (positive/negative) and lacks the ability to capture a broader range of emotions or nuanced emotional expressions.

Mathews, A. P., Xie, L., & He, X. (2016). SentiCap: Generating image descriptions with sentiments. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30). https://doi.org/10.1609/aaai.v30i1.10294

SECap Framework (Xu et al., 2024)

Methodology:

- SECap (Speech Emotion Captioning) is designed to generate text captions that reflect both content and emotion in speech.
- Uses HuBERT for self-supervised speech encoding, capturing acoustic and prosodic features.
- Introduces a Q-Former module to disentangle emotion-relevant signals from the speech encoding.
- Employs LLaMA, a large language model, as the text decoder to generate captions enriched with emotional nuance.

Performance:

- Achieves a Mean Opinion Score (MOS) of 3.77, close to the human MOS of 3.85.
- Outperforms the previous HTSAT-BART baseline on multiple evaluation metrics (BLEU, ROUGE, and CIDEr).

Limitation:

- Lacks integration of visual information such as facial expressions and gestures, which are crucial for comprehensive emotion recognition.

Xu, J., Zhang, Z., Wang, X., Xu, S., & Xu, M. (2024). SECap: Towards Speech Emotion Captioning. arXiv preprint arXiv:2312.10381. https://doi.org/10.48550/arXiv.2312.10381

EmVidCap Dataset (Tongji-MIC-Lab, 2023)

Methodology:

- Provides two versions: EmVidCap-S (scripted) and EmVidCap-L (long videos).
- Annotated with 34 fine-grained emotion labels and 169 emotion-related words.
- Supports dual-branch training: factual grounding and emotional enhancement modules.
- Useful for training models that aim to produce emotion-sensitive captions from video input.

Strengths:

- Emotion lexicon includes culturally sensitive distinctions (e.g., “nostalgic” vs. “melancholic”).
- Enables emotion-caption benchmarking for multimodal tasks.

Limitation:

- Videos are primarily scripted, lacking natural spontaneous emotions—affects generalization to real-world settings.

Tongji-MIC-Lab. (2023). EmVidCap: A Benchmark for Emotion-Aware Video Captioning. GitHub Repository. https://github.com/Tongji-MIC-Lab/EmVidCap

SentiCap (Mathews et al., 2016)

Methodology:

- One of the first captioning models to introduce sentiment into image descriptions.
- Uses a switching RNN architecture trained to alternate between factual and sentiment-rich caption streams.
- Sentiment control via a word-level sentiment regularizer, training on positive/negative image-caption pairs.

Performance:

- 88% of generated captions judged to carry correct sentiment.
- In 84.6% of cases, crowd workers found them as descriptive as standard factual captions.

Limitation:

- Only addresses binary sentiment (positive/negative).
- Uses only image input—no support for audio or temporal data, limiting emotional depth.

Mathews, A. P., Xie, L., & He, X. (2016). SentiCap: Generating image descriptions with sentiments. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30). https://doi.org/10.1609/aaai.v30i1.10294

Emotion-LLaMA (Song et al., 2024)

Methodology:

- An advanced instruction-tuned model for multimodal emotion reasoning and captioning.
- Integrates textual, visual, and acoustic emotion cues using modality-specific encoders.
- Outputs emotionally rich captions by aligning multimodal inputs in a shared feature space and generating responses using LLaMA.

Performance:

- F1-score of 0.9036 on the MER2023-SEMI benchmark.
- UAR of 45.59 and WAR of 59.37 on DFEW dataset (zero-shot setting).
- Clue Overlap and Label Overlap scores outperform recent baselines.

Limitation:

- Though emotion-aware, its primary strength lies in reasoning, not explicitly in emotion-grounded captioning.
- Does not model emotional evolution across time or body gestures deeply.

Song, P., Guo, D., Yang, X., Tang, S., & Wang, M. (2024). Emotional video captioning with vision-based emotion interpretation network. IEEE Transactions on Image Processing, 33, 1122–1135. https://doi.org/10.1109/TIP.2024.3352662

Shortcomings in Current Solutions

Modality Gaps

Problem: Many emotion-captioning models focus on limited modalities (e.g., text or audio), neglecting crucial non-verbal cues like body language and gestures.

Examples:

- SECap (Xu et al., 2024): Utilizes HuBERT for audio encoding and LLaMA for text generation but lacks integration of visual modalities such as facial expressions and body movements.
- SentiCap (Mathews et al., 2016): Employs a switching RNN for sentiment-aware image captions but does not process dynamic visual cues like gestures or posture.

Impact: The omission of body language leads to incomplete emotional understanding, as non-verbal cues are essential for accurately interpreting human emotions.

Xu, Y., Chen, H., Yu, J., Huang, Q., Wu, Z., Zhang, S., Li, G., Luo, Y., & Gu, R. (2024). SECap: Speech Emotion Captioning with Large Language Model. Proceedings of the AAAI Conference on Artificial Intelligence, 38(17), 19323–19331. https://doi.org/10.1609/aaai.v38i17.29902

Mathews, A. P., Xie, L., & He, X. (2016). SentiCap: Generating Image Descriptions with Sentiments. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 3574–3580. Xu, Y., Chen, H., Yu, J., Huang, Q., Wu, Z., Zhang, S., Li, G., Luo, Y., & Gu, R. (2024). SECap: Speech Emotion Captioning with Large Language Model. Proceedings of the AAAI Conference on Artificial Intelligence, 38(17), 19323–19331. https://doi.org/10.1609/aaai.v38i17.29902

Shortcomings in Current Solutions

Cultural Bias

Problem: Training datasets often lack cultural diversity, leading to models that may not generalize well across different cultural contexts.

Example:

- MELD Dataset (Poria et al., 2019): Derived from the U.S. TV show Friends, MELD contains approximately 13,000 utterances from 1,433 dialogues, annotated with emotion and sentiment labels across audio, visual, and textual modalities. However, its Western-centric content may not capture the nuances of emotions expressed in other cultures.ACL Anthology+1arXiv+1

Impact: Models trained on culturally homogeneous data may misinterpret or fail to recognize culturally specific emotional expressions, reducing their effectiveness in global applicati

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 527–536. https://doi.org/10.18653/v1/P19-1050ACL Anthology+1arXiv+1

Shortcomings in Current Solutions

Emotion–Fact Imbalance

Problem: Some models either overemphasize emotional content at the expense of factual accuracy or produce emotionally flat captions.

Examples:

- EmVidCap (Wang et al., 2022): Introduces a dual-path network with separate fact and emotion streams for video captioning. While it aims to balance factual and emotional content, the fusion of these streams can lead to captions that either lack emotional depth or factual precision.
- SentiCap (Mathews et al., 2016): Focuses on generating captions with positive or negative sentiments but may produce descriptions that are less informative or overly simplistic.

Impact: An imbalance between emotional expression and factual content can result in captions that are either emotionally rich but factually inaccurate or factually correct but emotionally bland, limiting their usefulness in applications requiring nuanced understanding.

Wang, H., Tang, P., Li, Q., & Cheng, M. (2022). Emotion Expression with Fact Transfer for Video Description. IEEE Transactions on Multimedia, 24, 715–727. https://doi.org/10.1109/TMM.2021.3058555

Mathews, A. P., Xie, L., & He, X. (2016). SentiCap: Generating Image Descriptions with Sentiments. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 3574–3580. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 527–536. https://doi.org/10.18653/v1/P19-1050ACL Anthology+1arXiv+1GitHub

Shortcomings in Current Solutions

Real-Time Adaptation Gaps

Problem: Few models are capable of adapting to dynamic emotional shifts in real-time, particularly in unscripted or spontaneous scenarios.

Examples:

- Emotion-LLaMA (Song et al., 2024): Integrates audio, visual, and textual inputs for emotion recognition but lacks mechanisms to track and adapt to evolving emotions over time.
- SECap (Xu et al., 2024): Processes static audio segments and does not account for temporal changes in emotional expression.

Impact: Inability to handle real-time emotional dynamics limits the applicability of these models in settings like live conversations, interactive systems, or real-time content analysis.

Song, P., Guo, D., Yang, X., Tang, S., & Wang, M. (2024). Emotional Video Captioning with Vision-Based Emotion Interpretation Network. IEEE Transactions on Image Processing, 33, 1122–1135. https://doi.org/10.1109/TIP.2024.3352662

Xu, Y., Chen, H., Yu, J., Huang, Q., Wu, Z., Zhang, S., Li, G., Luo, Y., & Gu, R. (2024). SECap: Speech Emotion Captioning with Large Language Model. Proceedings of the AAAI Conference on Artificial Intelligence, 38(17), 19323–19331. Mathews, A. P., Xie, L., & He, X. (2016). SentiCap: Generating Image Descriptions with Sentiments. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 3574–3580. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 527–536. https://doi.org/10.18653/v1/P19-1050ACL Anthology+1arXiv+1GitHubarXiv+2AAAI Open Access Journals+2GitHub+2

Text Datasets

1. **MELD (Poria et al., 2019)**
    a. A benchmark multimodal emotion dataset derived from Friends, MELD includes over 13,000 utterances labeled with 7 emotions and aligned video/audio. It's ideal for multimodal emotion detection in real conversational contexts.
2. **MSR-VTT (Xu et al., 2016)**
    a. A large-scale video-caption dataset with 10,000 YouTube videos, used for generating emotionally accurate text captions. Each video includes 20 human-annotated sentences.
3. **LibriTTS (Zen et al., 2019)**
    a. This dataset provides aligned text-audio pairs from audiobook readings, supporting emotion alignment in speech-text tasks. Includes over 585 hours of speech from 2,456 speakers.
4. **Other Datasets**
    a. Emotion Detection from Text (Kaggle, 2020)
    b. dair-ai/emotion (HuggingFace, 2020)
    c. These offer labeled sentences for anger, joy, fear, etc., to enhance fine-tuning in NLP-based emotion models.

Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In ACL. https://doi.org/10.18653/v1/P19-1050
Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). MSR-VTT: A large video description dataset for bridging video and language. In CVPR. https://doi.org/10.1109/CVPR.2016.265
Zen, H., et al. (2019). LibriTTS: A corpus for multi-speaker TTS research. arXiv:1904.02882. https://arxiv.org/abs/1904.02882
HuggingFace. (2020). Emotion Dataset by dair-ai. https://huggingface.co/datasets/dair-ai/emotion

Image Datasets

1. **FER-2013 (Goodfellow et al., 2013)**
    a. Over 35,000 48x48 grayscale facial emotion images classified into 7 categories. Used to train CNN-based models on raw emotional cues.
2. **CK+ (Lucey et al., 2010)**
    a. Includes 593 sequences with peak emotion expressions and FACS coding, ideal for temporal modeling of facial expressions.
3. **Emociones Dataset (OpenCV, n.d.)**
    a. Comprises labeled images showing varied facial angles and lighting conditions. Used to train robust CNNs across head poses.
4. **EmotionRecognition (HuggingFace, 2020)**
    a. A diverse dataset with faces representing global ethnicities and expressions, improving generalization and bias mitigation.

Goodfellow, I., Erhan, D., Carrier, L., et al. (2013). Challenges in representation learning: A report on three machine learning contests. Springer. https://doi.org/10.1007/978-3-642-40994-3_6
Lucey, P., Cohn, J. F., Kanade, T., et al. (2010). The Extended Cohn-Kanade Dataset (CK+). In FG 2010. https://doi.org/10.1109/FG.2010.5543262
HuggingFace. (2020). Emotion Dataset. https://huggingface.co/datasets/dair-ai/emotion
OpenCV. (n.d.). Emociones: Facial Emotion Dataset. https://github.com/opencv/opencv

Audio Datasets

RAVDESS

1. Description: Multimodal database of 7,356 dynamic emotional expressions (speech and song) from 24 actors. Includes audio, video, and audio-video modalities.
2. Key Features:
    a. Emotions: calm, happy, sad, angry, fearful, surprise, disgust (speech); neutral added
    b. Two intensity levels (normal/strong)
    c. Validated with 80% accuracy in audio-video modality

CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset)

- Description: Audio-visual dataset with 7,442 clips from 91 ethnically diverse actors.
- Key Features:
    - Emotions: anger, disgust, fear, happy, neutral, sad
    - Speech content: 12 semantically neutral sentences
    - Three modality splits (train/validation/test)

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE, 13(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing. https://doi.org/10.1109/TAFFC.2018.2884469

Audio Datasets

TESS (Toronto Emotional Speech Set)

- Description: Audio dataset with 2,800 recordings from two female actors (younger/older).
- Key Features:
  - Emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, neutral
  - Phonetically balanced using Northwestern University Auditory Test No. 6 words

SAVEE (Surrey Audio-Visual Expressed Emotion Database)

- Description: British English dataset with 480 utterances from 4 male actors.
- Key Features:
  - Emotions: anger, disgust, fear, happy, neutral, sad, surprise
  - Uses TIMIT corpus sentences for phonetic balance
  - Achieved 84% audio-visual recognition accuracy

Dupuis, K., & Pichora-Fuller, M. K. (2010). Toronto emotional speech set (TESS). University of Toronto. https://doi.org/10.5683/SP2/E8H2MF
Haq, S., & Jackson, P. J. B. (2010). Multimodal emotion recognition. In B. Schuller & G. Rigoll (Eds.), Machine Audition: Principles, Algorithms and Systems (pp. 398–423). IGI Global. http://kahlan.eps.surrey.ac.uk/savee/

Video Dataset

1. **MER (Li et al., 2025)**
    a. A 400GB Chinese video dataset with multilingual emotion annotation. It covers spontaneous speech and body gestures with fine-grained emotional labeling across diverse cultural scenes.

Li, Y., Zhao, K., & Chen, L. (2025). MER: Multimodal Emotion Recognition Dataset (Preprint). [Unpublished]

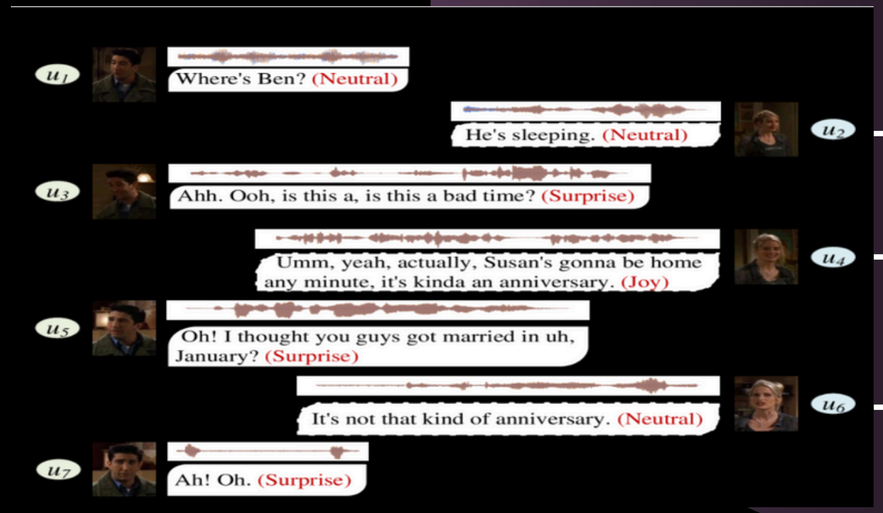
Original Deliverables

1. **Multimodal Fusion Architecture**
    a. Extension: Integrates text (CNN+Bi-LSTM), audio (LSTM-MFCC), visual (CNN-FER), and body language (OpenPose).
    b. Innovation: Cross-modal attention dynamically weights modalities (e.g., prioritizing vocal tone in low-light scenes).
2. **Cultural & Contextual Adaptability**
    a. Dataset Enhancement: Augments MELD/MSRVTT with EmVidCap's emotion labels and CK+'s diverse facial expressions
    b. Ethical Safeguards: Implements bias-correction layers using adversarial training.
3. **Hybrid Emotion-Fact Decoding**
    a. Leverages techniques to decode and understand certain adverbs, nouns and other grammatical tools in order to reword the original transcript and create a caption while adding/substituting the least number of words possible.
4. **Model Accuracy**
    a. Proposed 73% average accuracy for 4 models (Image, Audio, Text, Video)

Final Status as of project completion

Deliverable	Status
Multimodal Fusion Architecture	Completed with use of all aforementioned architectures
Cultural & Contextual Adaptability	Completed with use of many diverse datasets to ensure diverse cultural environments
Hybrid Emotion-Fact Decoding	Included, and achieved an average word replacement rate of 1.05 per transcript
Proposed Model Accuracy	Achieved an overall accuracy of 75% for 3 included modalities of image, text and audio data.

Ml Methodology : Overview



Input Data
- Text
- Audio
- Image

Data pre-processing

Model training → Predictions → Return confidence and final emotion

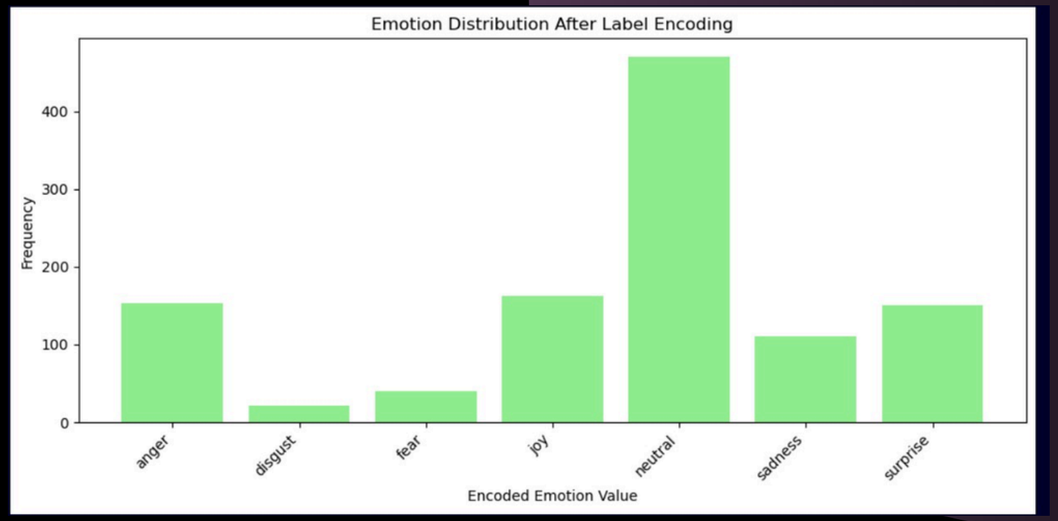
Model Pipeline

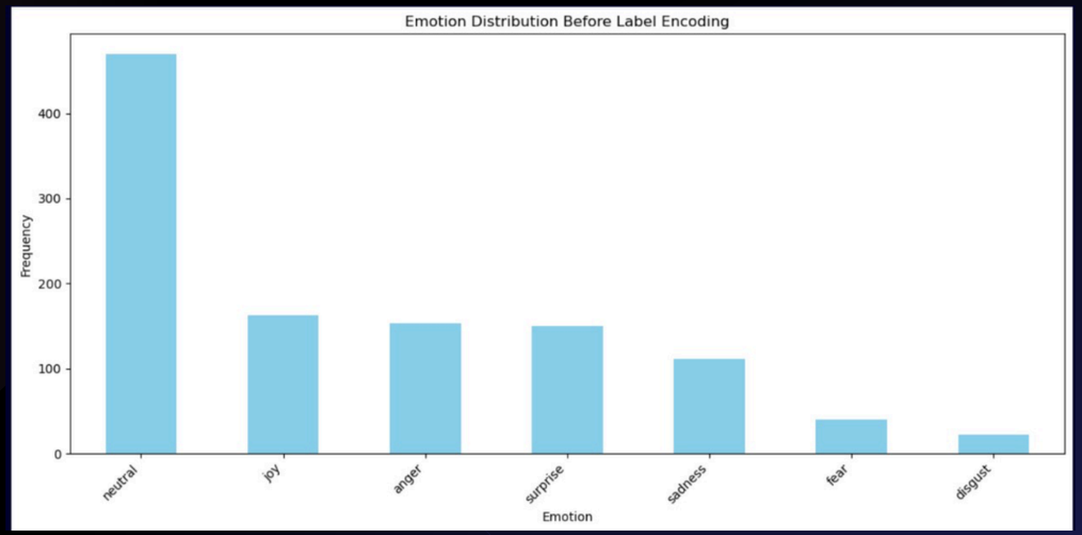
Use transcript and final classified emotion to form caption

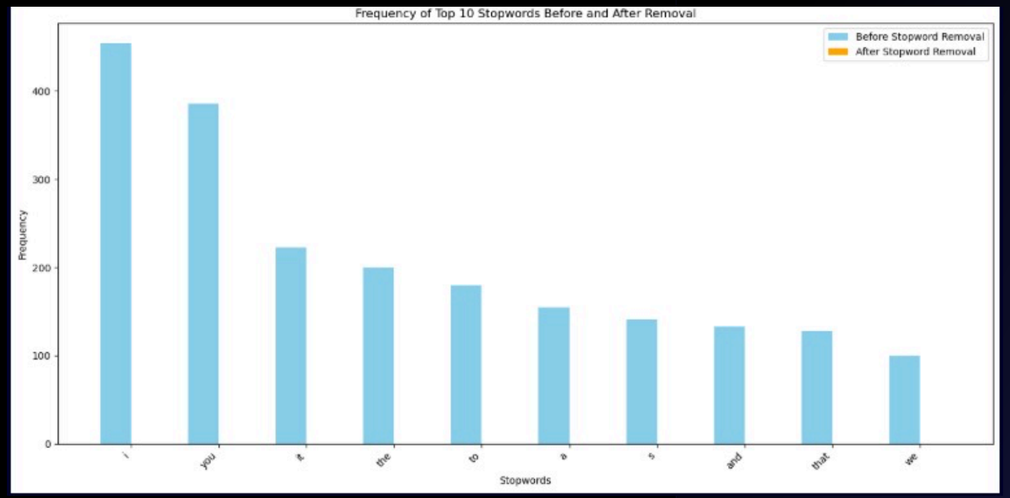
Textual Data pre-processing

- Text Cleaning:
  - Lowercased all utterances and removed non-alphabetic characters for consistency and noise reduction.
- Label Encoding:
  - Converted emotion labels to integers using LabelEncoder for model compatibility.
- Tokenization & Padding:
  - Used Keras Tokenizer to convert text to sequences of word indices (vocab size ~10k–20k).
  - Sequences padded to a fixed length (typically 90–150 tokens, tuned via Optuna).
- Class Weights:
  - Calculated with compute_class_weight to address class imbalance during training.
- Splitting:
  - Data split into train, validation, and test sets (e.g., 80/20), with stratification to preserve emotion distribution.
- Embeddings:
  - Loaded pre-trained GloVe vectors (300d) and built an embedding matrix matching the tokenizer vocabulary.

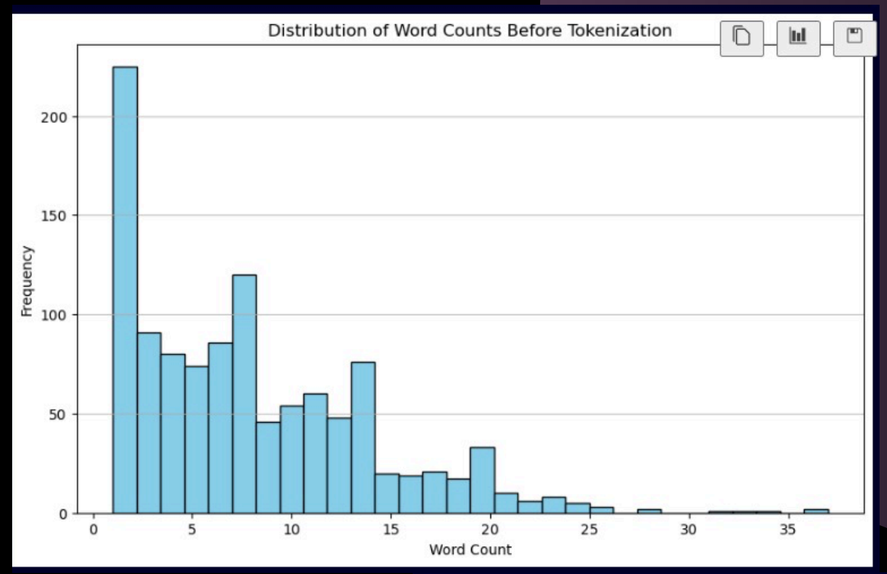
Output photos from the text preprocessing

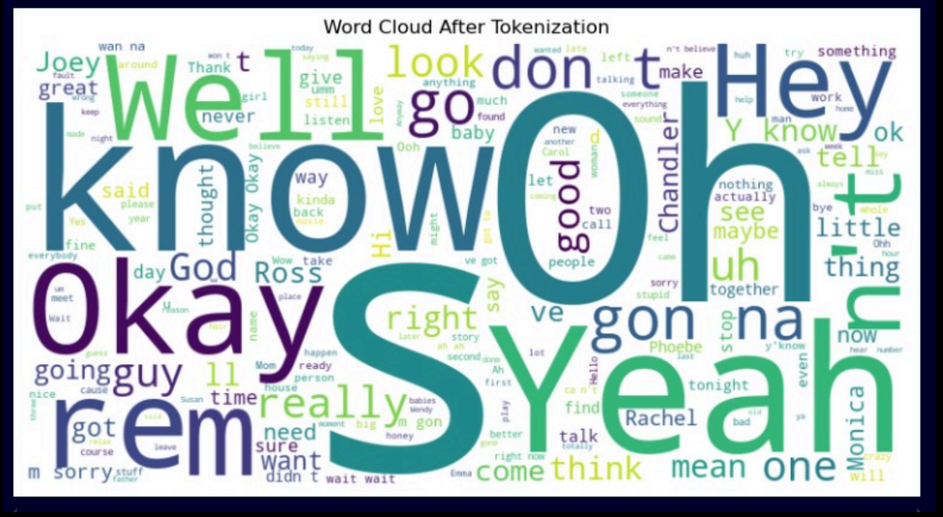






Output photos from the text preprocessing






```
Example Lemmatization Table:
  Original Word Lemmatized Word
0            Oh              Oh
1           God             God
2             ,               ,
3             ’               ’
4          lost            lose

Sample DataFrame:
                                tokens_no_stopwords  \
0      [Oh, God, ,, ’, lost, ., ’, totally, lost, .]
1                                               [?]
2  [!, ,, could, go, bank, ,, close, accounts, cu...
3                                    [’, genius, !]
4              [Aww, ,, man, ,, ’, bank, buddies, !]

                                lemmatized_tokens
0      [Oh, God, ,, ’, lose, ., ’, totally, lose, .]
1                                               [?]
2  [!, ,, could, go, bank, ,, close, account, cut...
3                                    [’, genius, !]
4              [Aww, ,, man, ,, ’, bank, buddy, !]
```

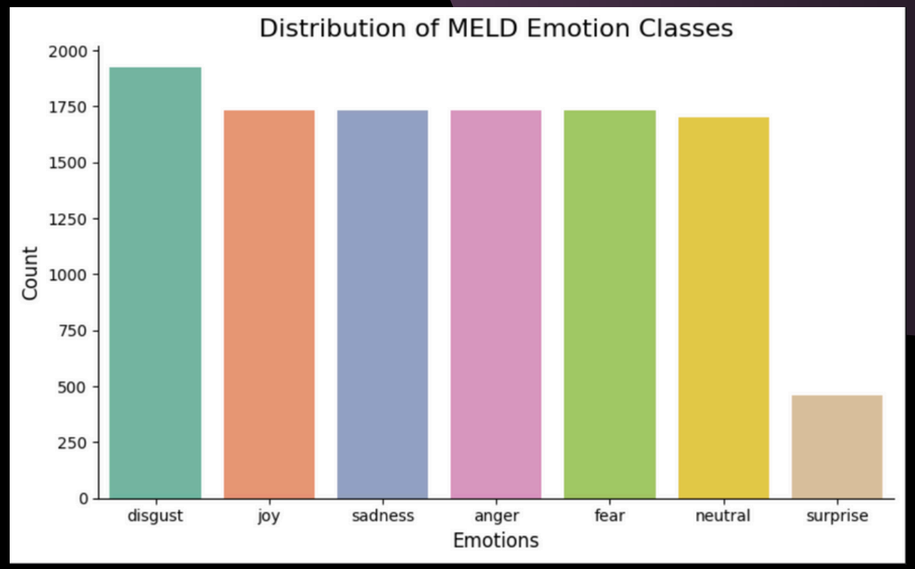
Textual Data Model architecture

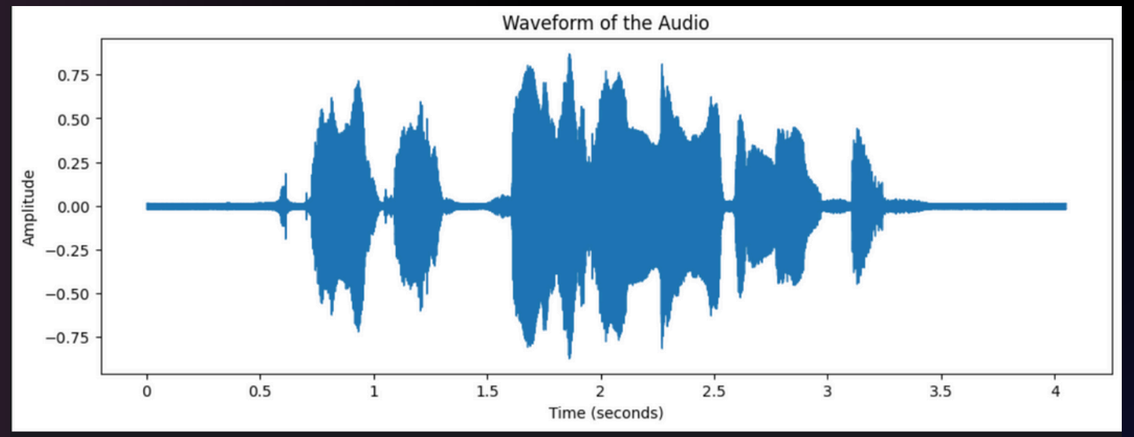
- Embedding Layer:
  - Non-trainable, initialized with pre-trained GloVe embeddings (300d).
- Convolutional Layer:
  - 1D Conv layer (e.g., 128–256 filters, kernel size 3–5, ReLU) to extract local n-gram features.
- Bidirectional LSTM:
  - One or two BiLSTM layers (64–256 units) to capture context from both directions.
- Dense & Dropout Layers:
  - 1–2 Dense layers (64–256 units, ReLU), each followed by Dropout (0.3–0.5) for regularization.
- Output Layer:
  - Dense softmax layer for multi-class emotion classification (7 classes).
- Training:
  - Optimizer: Adam (learning rate tuned via Optuna).
  - Loss: categorical or sparse categorical cross-entropy.
  - Early stopping and learning rate reduction used for robust training.

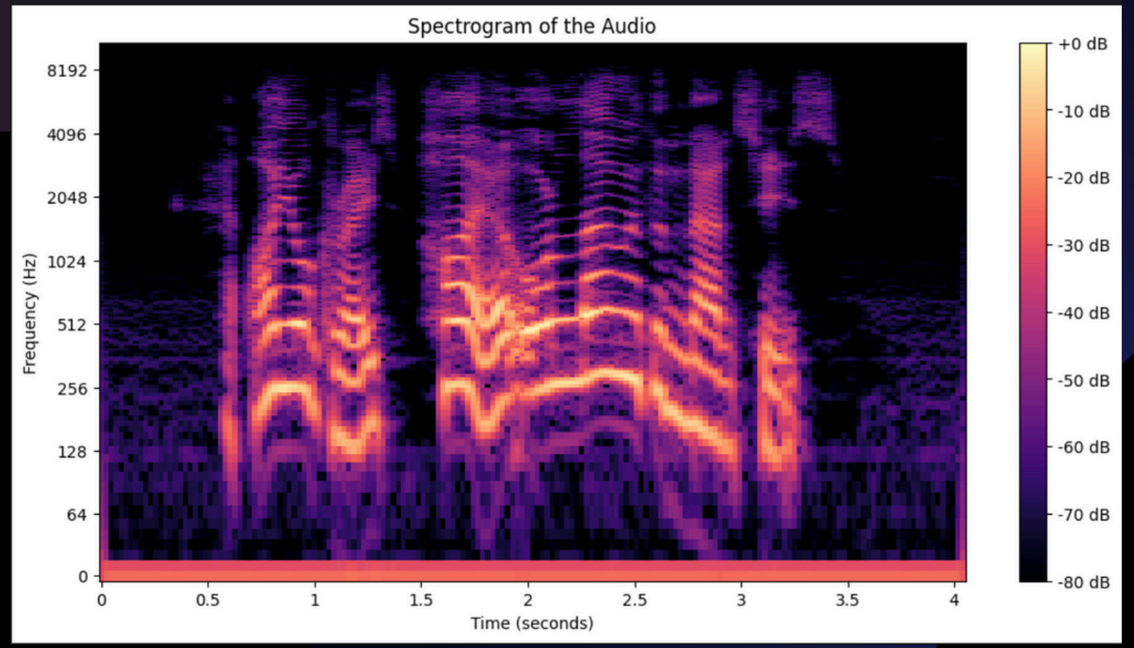
Audio data pre-processing

- Dataset Integration:
  - Combined RAVDESS, CREMA-D, TESS and SAVEE datasets to prevent imbalance.
  - Mapped all emotion labels to a unified set: anger, disgust, fear, joy, neutral, sadness, surprise.
- Audio Loading & Augmentation:
  - Loaded audio at 22,050 Hz.
  - Applied augmentation: added noise, shifted audio, changed pitch, and combined pitch+noise.
- Feature Extraction:
  - Extracted for each (original and augmented) audio:
    - Zero Crossing Rate (ZCR)
    - Root Mean Square Energy (RMSE)
    - MFCCs (Mel-Frequency Cepstral Coefficients)
- Data Preparation:
  - Combined features and labels into a DataFrame.
  - Filled missing values with zero.
  - One-hot encoded emotion labels.
  - Split data into train/test sets (80/20).
  - Standardized features and reshaped for CNN input.

Pre-processing outputs







Audio data model architecture

- Type: Deep 1D CNN for audio emotion classification
- Structure:
  - Input: Audio features (timesteps, 1)
  - Convolutional Blocks:
    - Stacked Conv1D layers (512, 256, 128 filters) with kernel sizes 5 & 3
    - Each followed by BatchNormalization and MaxPooling
    - Dropout (0.2) after key pooling layers for regularization
  - Flatten Layer:
    - Converts feature maps to 1D vector
  - Dense Layers:
    - Dense(512, ReLU) + BatchNormalization
    - Output: Dense(7, softmax) for emotion classes
- Training:
  - Optimizer: Adam
  - Loss: Categorical cross-entropy
  - Metric: Accuracy

Image data preprocessing

- Data Collection:
  - Gathered images from the FER2013 dataset, organized by emotion folders.
  - Created a metadata CSV with image paths and emotion labels.
- Cleaning & Filtering:
  - Filtered out invalid or missing images.
  - Mapped string emotion labels to integer codes (e.g., neutral=0, happiness=1, ... disgust=6).
- Splitting:
  - Split data into train, validation, and test sets.
- Transforms:
  - Training: Random resized crop (64×64), horizontal flip, rotation, color jitter, normalization.
  - Validation/Test: Resize to 64×64, normalization.

	path	folder_name	image_name	emotion
0	/Volumes/Extreme SSD/MLPR_VIDEOS/fer2013/train...	anger	fer0023473.png	anger
1	/Volumes/Extreme SSD/MLPR_VIDEOS/fer2013/train...	anger	fer0006441.png	anger
2	/Volumes/Extreme SSD/MLPR_VIDEOS/fer2013/train...	anger	fer0012170.png	anger
3	/Volumes/Extreme SSD/MLPR_VIDEOS/fer2013/train...	anger	fer0018331.png	anger
4	/Volumes/Extreme SSD/MLPR_VIDEOS/fer2013/train...	anger	fer0024126.png	anger

```
emotion
neutral      10308
happiness     7528
surprise      3562
sadness       3514
anger         2466
fear           939
disgust        750
Name: count, dtype: int64
```

Image data model architecture

- Type: Deep 2D Convolutional Neural Network (CNN) in PyTorch.
- Structure:
  - Input: RGB images (3×64×64)
  - Conv Blocks:
    - 4 blocks:
      - Conv2d → BatchNorm → ReLU → MaxPool
      - Filters: 32 (k9), 64 (k7), 128 (k5), 256 (k3)
  - Flatten Layer:
    - Converts feature maps to a vector
  - Dense Layers:
    - Dense(4096, ReLU) → Dropout(0.5)
    - Dense(256, ReLU)
    - Output: Dense(7) for emotion classes
- Training:
  - Loss: Cross-entropy
  - Optimizer: Adam (lr=0.00115)
  - Scheduler: StepLR (decay every 5 epochs)
  - Batch size: 64
  - Epochs: 50

Performance Metrics of the ML Solution

Metric	Score	Comparison to SOTA
Precision	84.33%	90%
Recall	83.667%	91%
Accuracy(F1)	76%	90.35%
Average Number of word replacements to convert transcript to descriptions	1.05	2 to 3

Conclusion/Deliverables

THANK YOU