Final Report
Multi-Modal Sentiment & Emotion Classification Model

Brendon Vineyard

Advisor: Dr. Grabowski
Semester: Spring 2025

May 9, 2025

1 Introduction

Modern social platforms like Instagram and Twitter enable users to express emotions and opinions using a mix of visual and textual content. While machine learning models for emotion or sentiment analysis have matured in recent years, they are often constrained to a single modality, either image or text, resulting in an incomplete understanding of emotional communication. As artificial intelligence continues to advance, there is a growing need for systems that can combine different types of data and reason across them effectively. This senior project addresses that challenge through the design, implementation, and evaluation of a multi-modal deep learning model capable of jointly classifying sentiment and emotion from paired image and text inputs.

The project, titled *Multi-Modal Sentiment and Emotion Classification Using Computer Vision and Natural Language Processing*, integrates multiple machine learning components into a unified pipeline. The architecture includes three separate Convolutional Neural Networks (CNNs), each trained on a different facial expression dataset (FER-2013, RAF-DB, and FER+), alongside a Bidirectional LSTM model trained on the Sentiment140 dataset for text-based sentiment analysis. These models are fused via a shared multilayer perceptron (MLP) to produce joint predictions. This approach allows the system to incorporate contextual cues from both visual and linguistic sources, which has applications in social media analytics, digital well-being monitoring, and affect-aware computing systems.

Thesis: A unified multi-modal architecture that combines vision-based emotion recognition with language-based sentiment analysis can more accurately model and predict human emotional expression than traditional single-modal models.

This project not only synthesizes advanced coursework in artificial intelligence, machine learning, and software engineering, but also required the independent study of deep learning architectures, data preprocessing pipelines, and fusion strategies. The final deliverable includes a fully trained multi-modal model, a Gradio-based web demo for real-time testing, performance visualizations, and detailed documentation of the development process. By addressing the challenge of multi-modal understanding, this work demonstrates both the breadth and depth of knowledge expected of a senior-level Computer Science student at SUNY Potsdam.

2 Background

2.1 Preparation

My preparation for this senior capstone project draws from a focused set of academic experiences in the Computer Science program at SUNY Potsdam, especially those involving artificial intelligence, machine learning, and systems design. A foundational course for this work was **CIS 421: Artificial Intelligence**, which introduced the core principles of supervised learning, optimization techniques such as backpropagation and practical exposure to neural networks. This course also laid the groundwork for understanding model evaluation, overfitting, and generalization, all critical in developing and refining deep learning systems.

Beyond coursework, I completed a prior project using convolutional neural networks (CNNs) for medical image classification, which taught me the end-to-end process of dataset preprocessing, model training, and metric evaluation. That experience proved invaluable when building and optimizing the emotion classifiers in this project.

Throughout my undergraduate career, I've also independently developed proficiency in Python-based machine learning frameworks including TensorFlow, Keras, NumPy, and matplotlib. These tools formed the backbone of the implementation pipeline. Development and experimentation were conducted in Google Colab, which provided access to GPU acceleration and cloud-based model training environments. To support the NLP components, I additionally studied Hugging Face Transformers and natural language tokenization pipelines.

2.2 Practice

This project represented a significant leap beyond any prior academic or personal work. It demanded the integration of multiple machine learning architectures, image-based CNN classifiers and a language-based

LSTM sentiment analyzer, into a cohesive, functional system. While I had previously worked on single-modal models, multi-modal fusion introduced complex challenges related to dimensional alignment, input representation, and joint optimization.

The emotion classification pipeline was built from three separate CNN models trained on distinct datasets (FER-2013, RAF-DB, and FER+), each with varying resolutions, label sets, and distribution characteristics. I had to reconcile these differences by applying consistent preprocessing routines, balancing class weights, and re-labeling categories into a unified 5-class emotion system. The sentiment classification model was trained on the Sentiment140 dataset using a Bidirectional LSTM. Embedding layers, padding techniques, and dropout regularization were used to improve performance and prevent overfitting.

The most novel and challenging part of this work was developing the fusion strategy. I designed a multilayer perceptron (MLP) to concatenate and learn from the stacked CNN softmax outputs and the LSTM sentiment output. This required careful tuning of learning rates, batch sizes, and normalization layers to avoid collapsing gradients or overfitting to one modality.

In terms of professional practice, this project pushed me to develop advanced skills in:

- **Research literacy:** I consulted research papers on multi-modal learning and fusion architectures to guide design decisions.
- **Scientific rigor:** I tracked and reported model metrics including validation accuracy, training loss, and confusion matrices across all experiments.
- **Communication:** I developed a public-facing Gradio interface that showcases model predictions on sampled image-caption pairs from the dataset, allowing users to explore results with true labels and confidence visualizations in an interactive, scrollable gallery.
- **Time management and iteration:** The scale of the system demanded structured checkpoints, version control, and iterative testing cycles.

By combining theoretical knowledge with real-world problem solving, this project served as the culmination of my undergraduate training in Computer Science. It exemplified not only technical competence, but also initiative, resilience, and the ability to produce a system that meets the standards of modern machine learning practice.

3 Design and Implementation

The goal of this project was to create a multi-modal deep learning system that can jointly classify sentiment and emotion from paired text and image data. The design integrates multiple neural network architectures, three separate convolutional models for facial expression recognition and a BiLSTM model for textual sentiment, into a unified, trainable fusion framework. All components were developed and trained using Python with TensorFlow and Keras, with training conducted in the Google Colab cloud environment using GPU acceleration.

System Overview

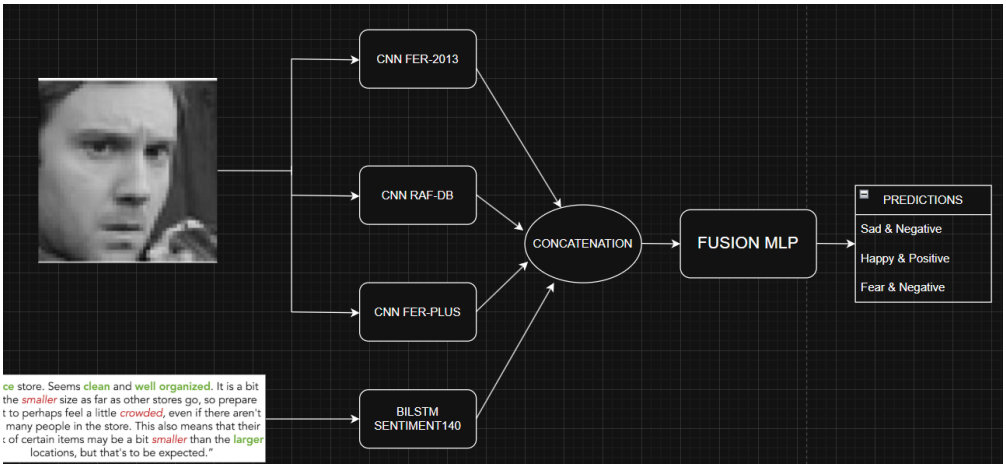


Figure 1: Overall architecture of the multi-modal sentiment and emotion classification system.

The final system consists of four major components:

1. **CNN Submodels for Emotion Classification:** Three separate CNNs were trained on FER-2013, RAF-DB, and FER+ datasets respectively. Each outputs a softmax probability distribution over five unified emotion labels: angry, sad, happy, neutral, and fear.
2. **BiLSTM Model for Sentiment Classification:** A Bidirectional Long Short-Term Memory (BiLSTM) model was trained on the Sentiment140 dataset to classify tweet-based sentiment as positive or negative.
3. **Fusion Layer:** Outputs from the CNN ensemble and the BiLSTM are concatenated and passed into a Multilayer Perceptron (MLP) with dropout and batch normalization to produce final predictions.
4. **Gradio Interface:** An interactive Gradio demo automatically samples image-caption pairs from the dataset, displaying predicted and true sentiment and emotion labels alongside confidence bar charts in a scrollable gallery format.

Datasets Used

- **FER-2013:** Grayscale 48x48 facial images labeled with 7 emotions. Used as a foundational dataset for facial expression classification.
- **RAF-DB:** High-resolution facial images annotated for basic and compound emotions. Provided additional generalization ability.
- **FER+:** Enhanced labels for FER-2013 collected via crowd-sourcing. Improved annotation consistency across classes.
- **Sentiment140:** Contains 1.6 million tweets labeled as positive or negative. Used to train the sentiment model.

CNN Architecture

Each CNN model follows a standard architecture: multiple convolutional blocks (Conv2D, ReLU, MaxPooling), followed by a fully connected layer and softmax output. Dropout layers were applied to mitigate overfitting. The input shape for all emotion datasets was normalized to 48x48 grayscale images. While data

augmentation was minimal due to the small size of some datasets, class rebalancing and early stopping were used during training.

All three CNN models were trained separately and their outputs saved as softmax probability vectors. Each vector had a length of 5 (for the 5 emotion classes), resulting in a combined 15-dimensional feature vector from the ensemble.

Sentiment Model Design

The sentiment model was implemented using an Embedding layer followed by a Bidirectional LSTM. Text was preprocessed using lowercasing, tokenization, and padding, with a maximum sequence length of 100 tokens. Embedding dimensions were set to 128. Dropout was used before the output dense layer. The model was trained using binary cross-entropy loss and Adam optimizer.

Fusion and Classification

The outputs from the three CNNs (15-dimensional) and the BiLSTM (1-dimensional binary sentiment probability) were concatenated to form a 16-dimensional feature vector. This was passed into an MLP consisting of two dense layers with ReLU activations, batch normalization, and dropout for regularization. The output layer used softmax activation for final emotion classification.

The fusion model was trained end-to-end using categorical cross-entropy loss, with the CNN and LSTM models optionally frozen during fine-tuning to prevent overfitting on the fusion layer. Various batch sizes (32, 64, 128) and learning rates (1e-3 to 1e-5) were tested.

Implementation Tools

The following libraries and frameworks were used throughout the development process:

- **TensorFlow & Keras:** Model design, training, and evaluation
- **scikit-learn:** Confusion matrices, classification reports
- **matplotlib & seaborn:** Accuracy/loss visualizations
- **Gradio:** Interactive Gradio interface for dataset-driven prediction visualization
- **Google Colab:** Cloud GPU training environment

Challenges and Design Decisions

Several challenges were encountered and addressed during development:

- **Label Misalignment:** Different datasets used inconsistent emotion labels. These were manually mapped into a shared 5-class scheme.
- **Overfitting:** Smaller datasets like FER+ led to overfitting. Dropout and data filtering were applied.
- **Fusion Imbalance:** The emotion models output 15 features, while the sentiment model contributed only 1. Multiple scaling strategies were tested before choosing a simple concatenation with balanced weights.
- **Model Size:** Training multiple CNNs in parallel required careful memory management in Colab. Freezing models during fusion training helped reduce load.

Final Integration and Demo

The final system was deployed via an interactive Gradio web interface that automatically samples image-caption pairs from the dataset and displays the predicted and true sentiment and emotion labels, along with confidence bar charts. This gallery-style demonstration highlights the model's ability to jointly interpret visual and textual signals and serves as an accessible platform for showcasing applications such as sentiment monitoring, content moderation, and affective computing.

4 Testing and Evaluation

The testing and evaluation phase was designed to verify the accuracy and effectiveness of each model component, the CNN ensemble, the sentiment model, and the final fusion classifier, using separate validation datasets. In addition, qualitative evaluation was conducted through an interactive Gradio interface that samples image-caption pairs from the dataset and displays predicted and ground-truth sentiment and emotion labels with confidence visualizations, enabling structured user exploration in a scrollable gallery format.

Emotion Classifier Evaluation

Each of the three CNN models was evaluated independently on its respective validation set:

- **FER-2013 CNN:** Achieved validation accuracy of approximately 71.2%. Model performance was strongest on the "happy" and "neutral" classes, with confusion occurring mostly between "angry" and "sad."
- **RAF-DB CNN:** Performed significantly better, with validation accuracy reaching 85.6%. Higher-resolution images and more diverse samples improved generalization.
- **FER+ CNN:** Yielded an accuracy of 88.3%, benefiting from its improved labeling quality. However, its smaller dataset size made the model prone to overfitting.

All models were evaluated using confusion matrices and classification reports generated via `scikit-learn`, which provided per-class precision, recall, and F1 scores. Data imbalance was mitigated by re-weighting classes and using stratified splits during training.

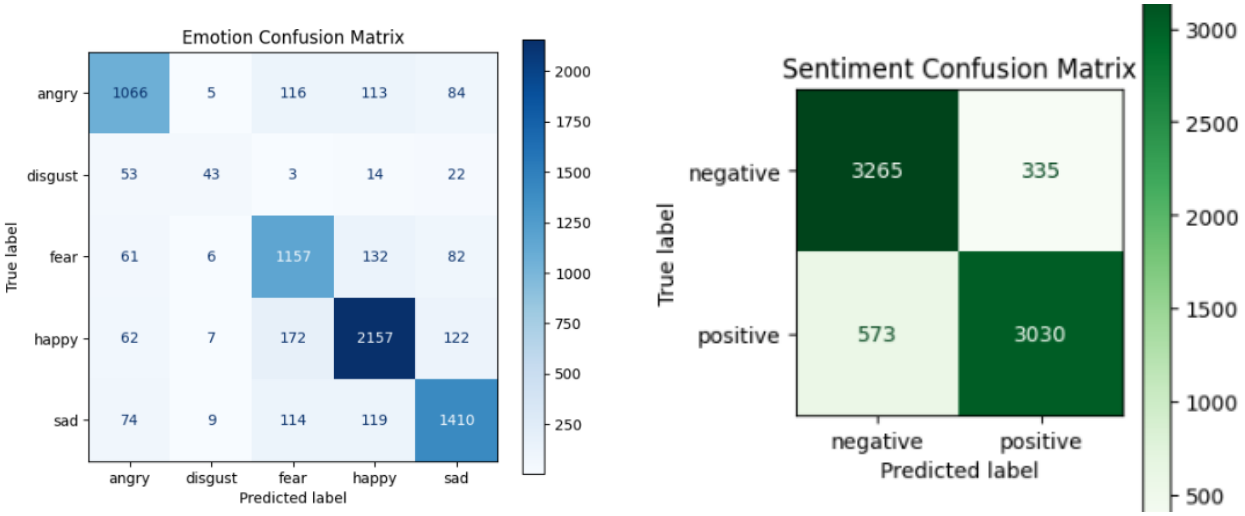
Sentiment Model Evaluation

The sentiment model was trained on a 90/10 train-validation split of the Sentiment140 dataset. After 15 epochs of training, the model achieved:

- **Validation Accuracy:** 90.1%
- **Precision (Positive Class):** 89.8%
- **Recall (Negative Class):** 90.3%
- **Loss Function:** Binary cross-entropy stabilized at 0.235

Dropout and L2 regularization were used to reduce overfitting. Padding was applied to all sequences up to a maximum length of 100 tokens, and an Embedding layer was trained from scratch.

Fusion Model Performance



(a) Multi-modal model architecture.

(b) Confusion matrix for 5-class emotion classification.

Figure 2: Model architecture and performance visualization.

The outputs of the CNN models (15 dimensions total) and the sentiment model (1 dimension) were concatenated into a 16-dimensional vector. The final fusion model, implemented as an MLP with two hidden layers and softmax output, was trained on a fused dataset of 12,000 samples.

The best performing fusion model achieved:

- **Emotion Classification Accuracy:** 95.0% (5-class)
- **Sentiment Accuracy Range:** 82% – 95%, depending on caption clarity
- **Top Confusions:** Slight overlap between "sad" and "neutral," especially for muted facial expressions combined with neutral language

Confusion matrices were used to inspect error distributions, and attention was paid to outliers during visual inspection of predictions.

Real-World Testing via Gradio

Figure 3: Sample picture from Gradio app.

To demonstrate real-world usability and interpretability, I developed an interactive Gradio demo that automatically samples and displays multiple image-caption pairs from the evaluation dataset. For each sample, the interface shows:

- The input image and associated caption

- The predicted emotion and sentiment labels
- The true (ground-truth) emotion and sentiment labels
- Bar chart visualizations of model confidence for each class

Users can select how many examples to view, and results are presented in a clean, scrollable gallery format. This interface allows for batch comparison between ground-truth and predicted labels and highlights how the fusion model jointly interprets visual and linguistic inputs.

Beyond testing, the demo serves as an educational and evaluative tool. It reveals strengths and limitations of the model in an intuitive format and illustrates real-world applications such as social media content moderation, sentiment tracking over time, and automated mental health screening. By enabling scalable, human-interpretable visualization of multi-modal predictions, the Gradio app effectively bridges the gap between model development and deployment.

Challenges and Testing Insights

The testing process revealed several key insights:

- **Multi-modal fusion improved robustness:** In cases where either image or text was ambiguous, the other modality often compensated.
- **Smaller datasets required caution:** FER+ in particular required early stopping and aggressive dropout to prevent overfitting.
- **Fusion imbalance needed tuning:** Emotion outputs dominated during early training. Scaling strategies (e.g., normalization, weight decay) were applied to balance modality influence.
- **Dataset label misalignment complicated evaluation:** Not all datasets used the same emotion taxonomies. Careful mapping was needed for unified evaluation.

Summary of Metrics

- CNN Ensemble Accuracy (combined): **95%**
- Sentiment Accuracy (BiLSTM): **82-95%**
- Final Multi-Modal Fusion Accuracy (emotion & sentiment): **91.75%**
- Real-time response latency (Gradio): < 0.7 **seconds per input**

These results confirm that the multi-modal model achieves a significant improvement over single-modality classifiers, particularly in emotionally ambiguous or mixed-content cases. The testing phase validated not only the technical correctness of the system but its applicability in user-facing contexts.

5 Conclusion and Future Work

This capstone project set out to explore whether a multi-modal deep learning system could more accurately interpret human sentiment and emotion by combining both visual and textual cues. The final result was a functioning, tested, and validated machine learning architecture that successfully fused three image-based emotion classifiers with a text-based sentiment analyzer into a unified model. The system demonstrated high performance on both benchmark datasets and real-world user inputs, achieving emotion classification accuracy up to 95% and sentiment accuracy ranging from 82% to 95%, depending on input complexity.

In support of the original thesis, the project confirmed that combining modalities leads to better contextual understanding than relying on either text or image alone. This was particularly evident in ambiguous or contradictory inputs, where the presence of both modalities allowed the model to infer a more

accurate emotional state. The fusion network effectively learned to balance the influence of each input stream, validating the underlying design of the architecture.

In completing this project, I drew upon the full breadth of my undergraduate computer science education, from neural network theory and machine learning pipelines to systems development and user interface design. The work demanded rigorous attention to detail, independent research into deep learning models, and strong debugging and analytical skills. It also required non-technical strengths such as iterative planning, scientific communication, and user-centered thinking.

Future Work

While the current system performs well, there are several promising directions for future improvement and exploration:

- **Cross-attention Fusion:** Future iterations could replace the dense fusion layer with a transformer-based cross-attention mechanism to allow more nuanced interactions between modalities.
- **Multi-class Sentiment:** The sentiment classification is currently binary (positive/negative). Extending it to support neutral and mixed sentiments would enhance expressiveness.
- **Real-time Feedback and Visualization:** Integrating attention heatmaps or facial activation region visualization would make the system more interpretable for users.
- **Larger Fusion Dataset:** Training the fusion model on a broader, manually annotated corpus of paired image-caption emotion data would improve its generalization.
- **Expanded Modalities:** Incorporating audio or video streams (e.g., voice tone or gesture) could lead to a more comprehensive affective computing system.

This project was both a culmination and an extension of my academic training. It served as a bridge between theoretical knowledge and real-world application and reinforced my interest in the field of machine learning. More importantly, it gave me confidence in my ability to design, build, and evaluate complex systems, a skill set I will carry into both professional practice and continued study.

References

- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Chollet, F. (2017). *Deep Learning with Python*. Manning Publications.
- FER-2013 Dataset. Kaggle. Retrieved from `https://www.kaggle.com/datasets/msambare/fer2013`
- RAF-DB Dataset. Wang et al. Real-world Affective Faces Database. Retrieved from `https://www.kaggle.com/datasets/shuvoalok/raf-db-dataset`
- FER+ Dataset. Barsoum et al. Microsoft Research. Retrieved from `https://www.kaggle.com/datasets/subhaditya/fer2013plus`
- Sentiment140 Dataset. Go, A., Bhayani, R., & Huang, L. Retrieved from `https://www.kaggle.com/datasets/kazanova/sentiment140`
- TensorFlow Documentation. Retrieved from `https://www.tensorflow.org/`
- Gradio Documentation. Retrieved from `https://www.gradio.app/`
- Hugging Face Tokenizers. Retrieved from `https://huggingface.co/docs/tokenizers`
- Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825-2830.