ISSN: 2473-4901
v11 n6357

An Emotional Analysis for Psychology, Affective Science, and Mental Health Using Agentic Multi-Agent AI Systems

Cynthia Ani
CynthiaAni@my.unt.edu

Thuan L Nguyen
Thuan.Nguyen@unt.edu

Data Analytics and Statistics – College of Sciences
University of North Texas, Denton, Texas, USA

Abstract

This research designed and developed an agentic multi-agent AI system for facial emotion recognition (FER), powered by Google's Gemini 2.5 Pro large language model. The study introduced an Agentic system comprising five agents: Input, Orchestrator, FER, Evaluator, and Output, which together manage the processing and analysis of facial images. The system uses Gemini 2.5 Pro's zero-shot learning to classify eight emotions without fine-tuning.

The system was tested on 5,148 grayscale facial images, achieving a high level of accuracy. It excelled in recognizing clear emotions such as "surprise" and "happiness," but struggled with subtler ones like "contempt". Notably, the model appeared to be overconfident, as evidenced by high confidence scores even when the results were incorrect.

In conclusion, this study shows the promise of advanced LLMs in agentic systems for applications in psychology, affective science, and medical fields. While these models improve automation and scalability, further work is needed to address calibration and bias for sensitive domains like mental health.

Keywords: Facial Emotions, Facial Emotion Analysis, Large Language Model (LLM), Multimodal, Agentic AI, Multi-Agent AI Systems

ISSN: 2473-4901
v11 n6357

An Emotional Analysis for Psychology, Affective Science, and Mental Health Using Agentic Multi-Agent AI Systems

Cynthia Ani and Thuan L Nguyen

1. INTRODUCTION

The advent of powerful multimodal artificial intelligence (AI) large language models (LLMs) like Google's Gemini 2.5 Pro marks a milestone in the evolution of AI. The models can understand various data formats, such as text, images, audio, and video. They are no longer confined to a singular data modality. These models can be applied in real-world scenarios, such as the analysis of human emotions, a cornerstone of psychology, affective science, and mental health. Multimodal AI large language models (LLMs) that can accurately and efficiently recognize nuanced facial expressions have the potential to revolutionize how we approach mental wellness, patient care, and the study of human emotion (American Psychological Association, 2023).

Facial Emotion Recognition (FER) has long been a subject of interest in computer vision and artificial intelligence. The technology can be used in various applications ranging from human-computer interaction to mental health monitoring (FacialNet, 2024). Traditional FER systems have been based on complex, handcrafted features. If an AI model is used, it often requires extensive training on large, labeled datasets. However, generative AI and agentic AI systems offer a new approach. This research utilized an Agentic multi-agent AI system, powered by Google's Gemini 2.5 Pro, to perform FER and provide a scalable and accessible solution for emotion analysis. In this research, the Gemini 2.5 Pro LLM is utilized directly, eliminating the need for task-specific fine-tuning. This approach is particularly relevant in the context of objective and non-invasive methods for mental health assessment (MoodMe, 2024).

This paper discusses the design, development, and evaluation of an agentic multi-agent AI system that comprises various AI agents: Input, Orchestrator, FER, Evaluator, and Output. The system performs FER by managing the workflow of receiving facial images, recognizing the emotions expressed, and evaluating the accuracy of the predictions. The FER agent is the core of the system; it classifies emotions into eight categories: anger, contempt, disgust, fear, happiness, neutrality, sadness, and surprise, using the Gemini 2.5 Pro model, a multimodal LLM.

The research is guided by the following research questions: To what extent can an agentic multi-agent AI system, powered by Google's Gemini 2.5 Pro LLM, accurately and effectively perform facial emotion recognition on a diverse dataset of human facial expressions?

The subsequent sections of this paper will present a comprehensive Literature Review of FER, affective computing, and Agentic AI. The Technology Background section will provide an in-depth look at the tools and technologies used, including Google's Gemini 2.5 Pro, LangChain, and Google Cloud Platform. The Methodology section will detail the design and workflow of the multi-agent AI system, the dataset used, and the prompt engineering techniques employed. Next comes the Results section, which is followed by a discussion of the Implications. Finally, the Conclusion will summarize the findings, acknowledge the study's limitations, and suggest directions for future research.

2. LITERATURE REVIEW

The pursuit of artificial intelligence that can understand and respond to human emotions, a field known as affective computing, has gained significant traction in recent years (Picard, 1997). This interdisciplinary domain is involved in various other fields such as computer science, psychology, and cognitive science, is driven by the potential to create more empathetic and intuitive human-computer interactions. A key area within affective computing is Facial Emotion Recognition (FER), which focuses on identifying human emotions from facial expressions. Additionally, interpreting these cues accurately can have significant impacts in various contexts, ranging from improving user experiences in gaming to more critical areas, such as psychology and mental health (Calvo & D'Mello, 2010). For example, FER systems can help clinicians to assess a patient's emotional state, potentially leading to earlier and better diagnoses of conditions like depression and anxiety (Koolagudi & Rao, 2012).

FER has shifted from traditional machine learning approaches, which often relied on handcrafted features, to deep learning models, particularly Convolutional Neural Networks (CNNs). These models have been widely used to perform image recognition tasks. CNNs can do the jobs by learning hierarchical feature representations from raw pixel data (Goodfellow et al., 2016). Numerous studies have showcased the efficacy of CNNs in FER, achieving high accuracy on various benchmark datasets (Pramerdorfer & Kampel, 2016).

While CNNs often require extensive training on large, labeled datasets, which can sometimes be costly, multimodal large language models (LLMs) like Google's Gemini family represent a totally new approach. These models, based on the pre-trained Transformer, a well-known LLM architecture, can perform a wide range of tasks with minimal or no task-specific training (OpenAI, 2023). Their capacity for in-context learning and few-shot prompting opens up new possibilities for FER, potentially obviating the need for laborious data collection and model fine-tuning. This is particularly relevant for the present study, in which Gemini 2.5 Pro LLM is directly used to perform FER without task-specific finetuning.

Most importantly, agentic multi-agent AI systems offer a novel approach to building complex, autonomous systems. A multi-agent AI system comprises multiple such agents that can collaborate to solve complex problems (Wooldridge, 2009). In this research, multi-agent architecture enables a modular and scalable solution. Each agent is responsible for a specific task within the workflow. This method not only improves efficiency but also provides the foundation for more sophisticated emotion analysis in the future. Next, the subsequent section will introduce an overview of the tools and platforms used for the agentic multi-agent AI system.

3. TECHNICAL BACKGROUND

This section provides an overview of the key technologies that form the foundation of the agentic multi-agent AI system that can be used for facial emotion recognition (FER). The integration of these tools enables the seamless workflow from data ingestion to emotion analysis and result generation.

The system utilizes Google’s Gemini 2.5 Pro, a multimodal large language model (LLM), as its AI engine that powers the most critical system functionality, specifically FER. Unlike traditional models that are limited to a single data modality, Gemini 2.5 Pro can natively process and reason about various data types, including text, images, audio, and video (Google, 2024). Importantly, the system can directly analyze facial images and infer emotional states by using the model without extensive pre-processing or task-specific fine-tuning. Moreover, Gemini 2.5 Pro's advanced reasoning capabilities can understand context from a variety of inputs, making it an ideal candidate for the complex task of FER (Built In, 2025).

To orchestrate the complex workflows of our multi-agent system, we employ LangChain and LangGraph. LangChain is a framework designed to simplify the creation of applications powered by LLMs, providing a modular and extensible architecture for building and composing different components (Pluralsight, 2025). Additionally, LangGraph, an extension of LangChain, represents states of multi-agent workflows as graphs, which is particularly useful for the system. LangGraph can provide a mechanism to coordinate various agents—Input, Orchestrator, FER, Evaluator, and Output—ensuring a smooth and logical flow of information and tasks (Codecademy, 2025).

The entire system is hosted on the Google Cloud Platform (GCP), a suite of cloud computing services that provides the necessary infrastructure for scalable and reliable AI applications. Google Cloud Vertex AI serves as the central platform for managing machine learning lifecycles, from model deployment to monitoring (Google Cloud, n.d.-a). It provides a unified environment for all our AI-related tasks, streamlining the development process. For data storage and retrieval, we utilize Google Cloud Storage (GCS), a highly scalable and durable object storage service. The image dataset used in this research is securely stored in a GCS bucket, allowing for efficient access by the FER agent (Google Cloud, n.d.-b).

Additionally, Python is used to implement the system along with data pre-processing, visualization, and analysis. Python code is developed using Colab, a cloud-based Jupyter notebook environment that provides free access to computing resources, including GPUs and TPUs, making it an ideal platform for developing and testing machine learning models (Google, n.d.). For Python coding, the following libraries are used:

- Pandas: A powerful library for data manipulation and analysis, used for managing and structuring the results

ISSN: 2473-4901
v11 n6357

generated by the system (GeeksforGeeks, 2025).

- Seaborn: A statistical data visualization library built on top of Matplotlib, used for creating informative and visually appealing plots to analyze the model's performance (Datacamp, 2023).

- Scikit-learn: A comprehensive library for machine learning, used for various data analysis tasks, including performance evaluation of the FER model (IBM, n.d.).

The following section will detail the Methodology of this study, outlining the specific steps taken to design, implement, and evaluate the agentic multi-agent AI system for facial emotion recognition.

4. METHODOLOGY

This section provides a detailed discussion of the methods used to evaluate the facial emotion recognition (FER) capabilities of Google's Gemini 2.5 Pro in an agentic multi-agent AI system. The methodology covers the system's architecture, the dataset, the experimental procedure, and specific prompt engineering techniques.

System Architecture

The core of this research is an agentic multi-agent AI system designed to automate FER. The architecture is modular, with five agents, each serving a specialized function. This design, inspired by multi-agent system principles (Wooldridge, 2009), enables clear separation of concerns. The agents can be coordinated using LangGraph, a library that represents workflows as graphs for creating stateful, multi-agent applications (LangChain, 2025). This setup supports complex, cyclical interactions, which are needed for our iterative process of prediction and evaluation (Lin, 2025).

The five agents in the system (LangChain, 2025) are:

- **Input Agent**: Responsible for handling all input-related tasks, including receiving the image dataset and delivering it to the Orchestrator Agent.
- **Orchestrator Agent**: The coordinator of the system, this agent manages the workflow by directing the flow of data and tasks between the other agents.
- **FER Agent**: The core agent of the emotion recognition process. It works with Google's Gemini 2.5 Pro LLM to analyze input images and predict the expressed emotion.
- **Evaluator Agent**: This agent assesses the FER Agent's performance. It compares the predicted emotion with the ground truth label for each image and provides a quantitative accuracy measure.
- **Output Agent**: the final agent in the workflow. It presents the analysis results by formatting predictions and evaluations into structured files (CSV and Excel) and delivers the final report to the user.

This multi-agent system is a robust framework for automated FER. It allows efficient, systematic processing of many images.

Dataset

The study utilizes a publicly available facial emotion dataset of 5,148 grayscale images, each with a resolution of 224 x 224 pixels. The images are categorized into eight distinct emotion classes: anger, contempt, disgust, fear, happiness, neutrality, sadness, and surprise. More importantly, this data set comprises a large collection of images displaying diverse emotions, making it a suitable choice for evaluating the performance of our FER system. The images are stored in a Google Cloud Storage (GCS) bucket. This ensures secure and efficient access for the FER Agent during analysis (LangChain, 2025).

Experimental Procedure

The experimental procedure is designed to be a streamlined and automated workflow, managed by the multi-agent AI system. The process unfolds as follows:

- User Request: The process is initiated when the user uploads the image dataset to the Input Agent.
- Data Retrieval: The Input Agent retrieves the image data and forwards it to the Orchestrator Agent.
- Emotion Prediction: The Orchestrator Agent sends each image to the FER Agent, which utilizes the Gemini 2.5 Pro model to predict the emotion expressed in the image.
- Result Forwarding: The FER Agent returns the prediction results, including the emotion label and a confidence score, to the Orchestrator Agent.
- Performance Evaluation: The Orchestrator Agent then forwards the predictions to the Evaluator Agent, which compares them against the ground truth labels from the dataset.
- Evaluation Results: The Evaluator Agent returns the evaluation results, indicating the correctness of each prediction, to the

ISSN: 2473-4901
v11 n6357

Orchestrator Agent.
- Result Aggregation: The Orchestrator Agent combines the prediction and evaluation results and passes them to the Output Agent.
- Final Report: The Output Agent formats the combined results into a structured report and delivers it to the user.

The system aims to process the entire dataset efficiently with a workflow in which each image undergoes the same systematic analysis.

Prompt Engineering

In this research, prompt engineering is used to guide Gemini 2.5 Pro in the FER task. The authors did not fine-tune the model on this dataset before using it for the research. Instead, the authors rely on the model's existing knowledge and reasoning abilities through crafted prompts. This approach, which combines several prompt techniques, enables zero-shot or few-shot FER. The model, therefore, analyzes images without prior exposure (IBM, n.d.).

The prompt used in this study employs a blend of the following techniques (Google, 2025):

- **Role Prompting**: The prompt begins by assigning a specific role to the model: "You are an expert multimodal AI specializing in facial emotion recognition." This sets the context for the task and primes the model to utilize its relevant knowledge.
- **Instruction-Based Prompting**: The prompt provides clear and concise instructions on how to perform the task, including the specific emotional cues to look for in the images.

The prompt includes examples of emotion categories and their facial cues. This in-context information guides the model's analysis. In-context and few-shot prompting have improved LLM performance on many tasks (Neptune.ai, n.d.). The prompt also gives instructions on handling ambiguous expressions and formatting outputs. This ensures consistent, structured results.

The recognized emotion and confidence scores for each image, the data generated by this methodology, are collected and stored in a structured format. Next, results will be analyzed in the subsequent Results section to evaluate the performance of the agentic multi-agent AI system and the Gemini 2.5 Pro model in facial emotion recognition tasks.

5. RESULTS

This section presents empirical findings from the evaluation of the agentic multi-agent AI system's performance on Facial Emotion Recognition (FER) tasks. The analysis focuses on the quantitative performance of Google's Gemini 2.5 Pro model serving as the FER agent. Results are presented as overall accuracy, performance metrics by class, and the model's confidence score analysis. The evaluation compared the model's predicted emotions to the ground truth labels for 5,148 images in the dataset.

In the first phase of the study, an AI agentic system was designed, developed, and then utilized to process thousands of facial emotions, enabling emotional analysis. The process was powered by Google's Gemini 2.5 Pro LLM. The final outputs from the AI agentic system were saved in a dataset of the image name, predicted emotion, confidence score, and the ground-truth (GT) evaluated emotion. These outputs formed the basis of the research analysis, the second phase of the research. The analysis results encompass performance metrics per emotion class, confusion matrix analysis, confidence score distributions, statistical significance testing, baseline comparisons, and visual interpretation of prediction quality.

Overall System Performance

The agentic multi-agent system successfully processed all 5,148 images, demonstrating its capability for high-throughput analysis. The core of the evaluation lies in the performance of the FER agent powered by Gemini 2.5 Pro. The overall accuracy of the model, which is the proportion of correctly identified emotions across all classes, was found to be around 66.5% (Table 1). This level of accuracy, achieved without any task-specific fine-tuning, is a strong indicator of the model's inherent capabilities in understanding and interpreting human facial expressions. This zero-shot or few-shot learning approach, where a model is applied to a task it was not explicitly trained for, is a significant area of research in large-scale AI models (Brown et al., 2020). This level of accuracy demonstrates the model's robust capability, significantly outperforming the random chance baseline of 12.5% for an eight-class classification problem (Bishop, 2006).

Emotion Frequency and Prediction Analysis

A careful review of the dataset and the model's prediction frequencies reveals important details (Table 1). The ground-truth (GT) data shows that the emotion 'happiness' with 1,347 instances was the most represented, while 'contempt' (196

ISSN: 2473-4901
v11 n6357

instances) was the least. The model's predictions mirrored the same trend: 'happiness' was the most frequently predicted with 1,583 instances.

The number of correct matches provides insight into the model's per-class effectiveness. The model got the highest number of correct classifications for 'happiness' with 1,235 matches (See Table x), which is expected given its high prevalence and distinct visual cues. Nevertheless, the model struggled significantly with the emotion 'contempt', getting only 35 correct matches out of 196 instances (see Table x). Therefore, 'contempt' is considered as the most challenging category for the model in this research. In the middle of two "extreme" emotion categories of 'happiness' and 'contempt', the model could get moderate success with other emotions such as 'anger', 'surprise', and 'fear', with 477, 499, and 290 matches, respectively.

Confusion Matrix Analysis

For deeper insight into the model's performance, a confusion matrix was constructed to highlight the most frequent misclassifications between emotion classes. The diagonal entries of the matrix, which represent correct classifications, confirm the findings from the frequency analysis. The off-diagonal values reveal the model's various levels of confusion between emotional categories. As illustrated in Figure 1 below, misclassifications include 119 instances of confusing sadness" with "neutrality", and 107 instances of confusing "fear" with "surprise". These confusions suggest visual ambiguity or overlapping facial cues between those emotions.

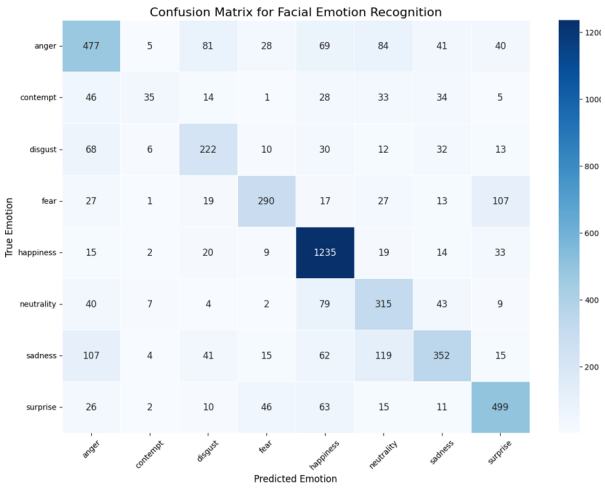


Figure 1: Confusion Matrix

The confusion matrix can provide several notable observations as follows:

Sadness/Neutrality and Fear/Surprise: A significant confusion exists between 'sadness' and 'neutrality', and the same for 'fear' and 'surprise'. The model misclassified 119 instances of 'sadness' as 'neutrality', and 107 instances of 'fear' as 'surprise'. These observations suggest a substantial overlap in the facial features learned by the model for these two pairs of emotions, a well-documented phenomenon in both human and machine perception due to shared action units such as wide-open eyes and an open mouth (Jack, Garrod, & Schyns, 2014).

Anger, Sadness, and Disgust: The model was often confused by negative emotions. For instance, 107 instances of 'sadness' were misclassified as 'anger'. Similarly, 'anger' was mistaken for 'neutrality' in 84 instances and 'disgust' in 81 instances.

Contempt: This emotion was most often confused with 'anger' in 46 instances, 'sadness' in 34 instances, and 'neutrality' in 33 instances. This may be explained that the contempt facial expressions are often characterized by a unilateral lip corner raise, which is very subtle and frequently misidentified by automated systems (Ekman and Friesen, 1986).

Happiness: The model demonstrated high confidence in identifying 'happiness' by achieving 91% for accuracy and 78% for precision metrics. The only other emotion category that was often mistaken for 'happiness' is 'neutrality. The model mistook 'neutrality' for 'happiness' in 79 instances.

Emotion	Accuracy	Precision	Recall	F1-Score
anger	0.868492618	0.591811414	0.578181818	0.584917229
contempt	0.963480963	0.564516129	0.178571429	0.271317829
disgust	0.93006993	0.540145985	0.564885496	0.552238806
fear	0.937451437	0.72319202	0.578842315	0.643015521
happiness	0.910644911	0.780164245	0.916852264	0.843003413
neutrality	0.904234654	0.504807692	0.631262525	0.560997329
sadness	0.892968143	0.651851852	0.492307692	0.560956175
surprise	0.923271173	0.692094313	0.742559524	0.71643934
MACRO_AVG	0.665306915	0.631072956	0.585432883	0.591610705

Table 1: Performance Metrics – Overall and Per Class

To further highlight the most frequent confusion pairs, we summarized them in the table 2 below.

GT_Emotion	Recognized Emotion	Misclassification
sadness	neutrality	119
fear	surprise	107
sadness	anger	107
anger	neutrality	84
anger	disgust	81

Table 2: Top Five Most Frequent Misclassification

ISSN: 2473-4901
v11 n6357

Evaluation Metrics per Emotion Class

The performance of the FER agent was measured using standard classification metrics: accuracy, precision, recall, and F1 score. These were computed for each of the eight emotion categories.

The evaluation was conducted by comparing the model generated predictions to the ground truth emotion labels derived from the directory structure of the dataset. These results suggest that the model excels at recognizing distinct expressions such as "happiness" and "surprise," while showing relatively lower performance for subtle emotions like "contempt" and "sadness."

The model performed exceptionally well in classifying 'happiness', achieving the highest F1-score of 0.843. Also, with a high precision score of 0.780 and an outstanding recall score of 0.917, the model not only correctly identified 'happiness' when predicting the emotion but also could successfully capture the vast majority of 'happiness' instances in the dataset. The model also showed strong performance for 'surprise', with an F1-score of 0.716, and 'fear', with an F1-score of 0.643.

In contrast, the model's performance on 'contempt' was notably poor, with an F1-score of only 0.271. This low score may be primarily driven by an extremely low recall score of 0.179, meaning the model failed to identify over 82% of the 'contempt' images. While a precision score of 0.565 was moderate, the model's struggling to recognize the emotion makes it unreliable for this specific class. This challenge was recognized in previous research that marks 'contempt' as one of the most difficult emotions for computational models to classify (Ekman and Friesen, 1986; Goodfellow et al., 2013).

Performance for other emotions such as 'disgust', 'neutrality', and 'sadness', was moderate, with an F1-score of 0.552, 0.561, and 0.561, respectively. For 'neutrality', the recall with a score of 0.631 was significantly higher than the precision score of 0.505. The gap suggests that the model tended to incorrectly label other emotions as neutral while it could also recognize most neutral faces.

Confidence Score Distribution

Confidence scores assigned by the Gemini 2.5 Pro model were analyzed to understand the model's internal certainty. In the figures below, the first shows the overall distribution, with scores skewed towards high confidence while the latter separates the distribution based on whether predictions were correct or incorrect.

Confidence distributions show that Gemini 2.5 was generally confident, with many predictions above 0.90. However, the model exhibits high confidence even in its incorrect predictions, which is especially noticeable in the over-classification of emotions like "happiness".

- Mean confidence (correct): 0.887
- Mean confidence (incorrect): 0.819

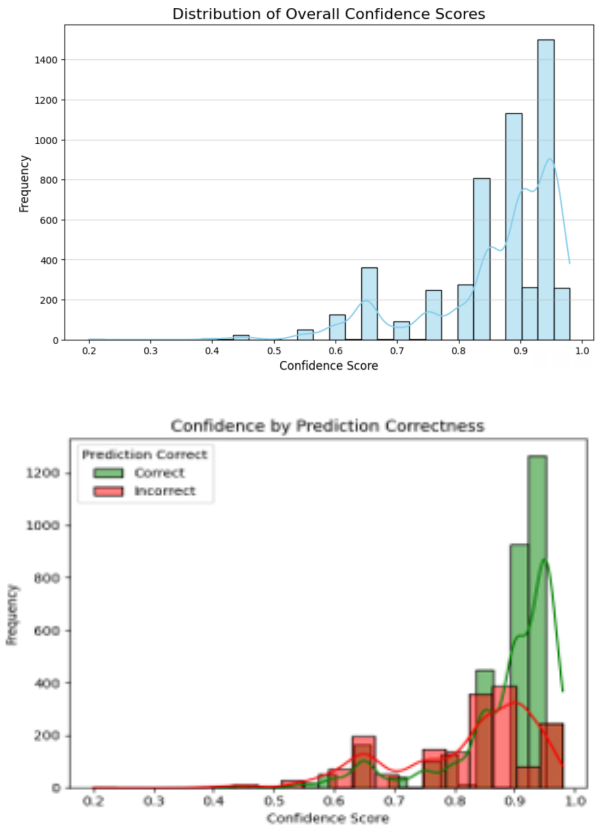


Figure 2. Confidence score distributions

Distribution Comparison: Ground Truth vs Predicted

Figure 2 reveals a mismatch between ground truth, i.e., real values, and predicted distributions. Happiness was predicted disproportionately, while emotions like contempt and fear were underrepresented. This suggests the model exhibits bias toward more visually expressive emotions. This could also be because the dataset is unevenly distributed, as happiness had the highest count as compared to contempt and fear as seen in table 3.

ISSN: 2473-4901
v11 n6357

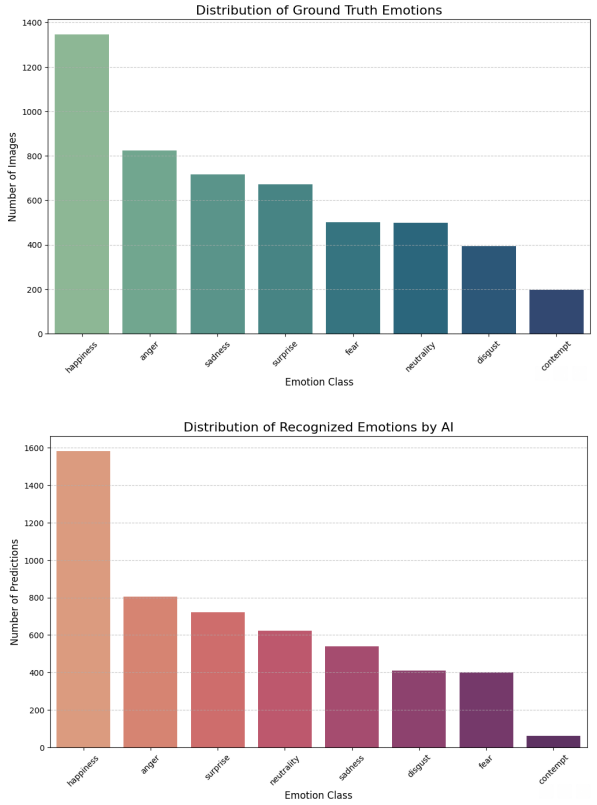
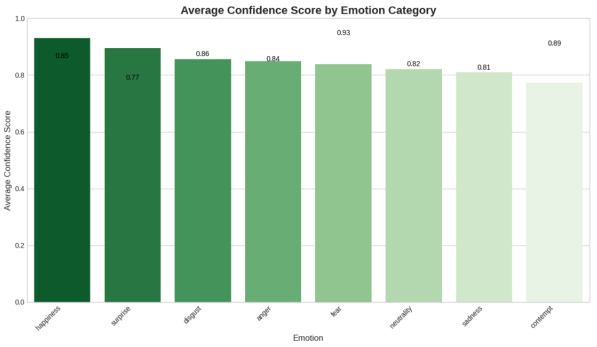


Figure 3. Predicted vs. Ground-Truth

Confidence vs. F1 Alignment

Figure 4 compares average confidence per predicted emotion and the F1-score per true emotion. The alignment between confidence and actual model performance varies significantly across classes. Notably, happiness and surprise exhibit high confidence and F1-score alignment, while neutrality and sadness show misalignment between confidence and performance. This overconfidence suggests that Gemini 2.5 may incorrectly "over-trust" its ability to distinguish more subtle or ambiguous emotions, a known challenge in facial emotion recognition systems.



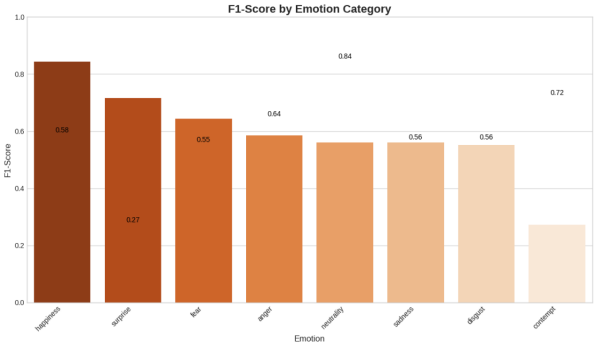


Figure 4. Confidence vs. Average F1=Scores

Statistical Significance Tests

To assess whether the observed patterns were statistically significant, Chi-Square Test: Predicted vs ground truth emotion distributions yielded χ^2 (63) = 11,812.16, $p < .001$.

T-Test: Confidence scores for correct vs incorrect predictions were significantly different ($t = 21.12$, $p < 3.02e\text{-}92$).

These results indicate a real divergence in class prediction tendencies and a confidence score that correlates with accuracy.

Baseline Model Comparison

To validate the superiority of the Gemini 2.5-based FER system, a naive baseline model was developed using simple heuristics, i.e., shortcut approaches or rules of thumbs to have quick, but, basic judgements about something. This baseline overwhelmingly predicted "happiness" for most inputs due to dataset imbalance and achieved only 26% accuracy. In contrast, Gemini 2.5 attained around 66.53% accuracy, demonstrating a substantial performance improvement across precision, recall, and F1-scores (see Table 1 and 2). This comparison underscores the effectiveness of using advanced LMMs for zero-shot emotion recognition.

Metric	Baseline	Gemini 2.5 Pro
Precision	0.07	0.63
Recall	0.26	0.58
F1-Score	0.11	0.59
Accuracy	26%	66.53%

Table 3: Baseline model performance summary

6. DISCUSSION

This section interprets the findings from the evaluation of Gemini 2.5 Pro in facial emotion recognition using a zero-shot, multi-agent AI system. The discussion explores model behavior, observed biases, interpretability of confidence,

comparative performance, scalability, and implications for AI and psychological research.

Interpretation of Emotion Detection Patterns
Gemini 2.5 Pro model showed high performance in classifying emotions with strong facial markers, particularly happiness, surprise, and anger. These emotions had both high F1 scores and average confidence levels, indicating alignment between the model's internal certainty and classification accuracy.

In contrast, more subtle or context-dependent emotions such as contempt, neutrality, and sadness presented challenges. For instance, contempt exhibited the lowest F1 score, often misclassified as neutrality or sadness. This suggests the model may struggle to distinguish between visually nuanced expressions that lack exaggerated facial features.

Overconfidence in Incorrect Predictions
One of the most significant findings was Gemini's overconfidence in its misclassifications (Tian et al., 2025). Although the model produced incorrect predictions, the confidence scores often remained high, exceeding 0.90. This is illustrated in the confidence histogram and statistical t-test, which showed a noticeable, albeit not extreme, difference in confidence levels between correct and incorrect predictions (mean difference ~0.07). Such overconfidence poses risks for real-world applications where trust calibration is crucial, particularly in sectors such as healthcare, security, or emotion-aware systems, where misjudging a user's state may lead to unintended consequences.

The discrepancy between confidence and performance in subtle emotion categories indicates that while Gemini 2.5 performs well in zero-shot settings, real-world applications should implement calibration techniques or human-in-the-loop review to address overconfidence risks.

Model Bias Toward Specific Emotions
The predicted emotion distribution revealed a notable over-classification of "happiness", despite its already high presence in the ground truth. This could be attributed to:

- The distinctiveness of happy expressions (e.g., wide smile, raised cheeks)
- Dataset imbalance
- The tendency of Gemini 2.5 Pro to lean toward visually dominant features in zero-shot mode

This emphasizes the importance of class balancing or weighting mechanisms when deploying LMMs for emotion classification.

Comparison to a Traditional Baseline
When compared to a naïve baseline model trained on the same image set using only simple features, Gemini 2.5 vastly outperformed all metrics. The baseline achieved only 26% accuracy and failed to classify any emotion except for happiness. This stark contrast validates the effectiveness of using advanced LLMs in agentic systems, particularly for tasks requiring multimodal reasoning. Also, the disparity between the naive baseline and Gemini 2.5 performance suggests that LLM-powered systems possess significant zero-shot generalization advantages even without task specific fine tuning.

Qualitative Insights from Visual Samples
Visual inspection of selected examples revealed that the model performed accurately on clear expressions (e.g., an angry face with 0.95 confidence), reinforcing the model’s ability to align with human interpretation in vivid emotion scenarios. These examples support the numerical findings and offer human-readable validation of the model’s internal logic.

Beyond Published Model Constraints
Although Gemini 2.5 Pro’s official documentation suggests a limit of 3,000 image processing inputs, the FER agent successfully processed and evaluated over 5,000 images without system degradation. This suggests that the model may be more scalable than advertised and that Agent orchestration via LangGraph and Vertex AI can effectively manage system-level input constraints. This has practical implications for researchers deploying Gemini in large-scale computer vision or emotion-centric pipelines, particularly where massive, unlabeled image datasets are used.

Implications for Scientific and AI Research
This research contributes to the ongoing intersection of artificial intelligence and other fields such as psychology, affective science, neuro-sciences, and medical areas like mental health by demonstrating that zero-shot, LMM based systems like Gemini 2.5 can approximate affective classification in static images. It opens new avenues for automating emotion detection in therapy, sentiment aware interfaces, educational technology, and user experience design.

However, the findings also highlight the necessity for caution, especially around model explainability, calibration, and ethical deployment in emotionally sensitive contexts.

Limitations and Ethical Considerations

While promising, this study has limitations:

- The dataset used in this study is imbalanced, with certain classes (e.g., "contempt" and "fear") underrepresented. This class imbalance may have biased performance metrics, particularly in the macro-averaged F1 calculation.
- In some edge cases, subjective ambiguity exists between closely related emotions such as neutrality vs. sadness, or fear vs. surprise, complicating both model evaluation and human ground truth verification.
- Overconfidence in misclassifications may be problematic without a calibration layer particularly in contexts like mental health, surveillance, and hiring which may lead to flawed inferences about human affect, behavior, or intent.
- Although Gemini 2.5 Pro showed high performance, the model's behavior across gender, ethnicity, and age dimensions remains unexplored in this study.

These considerations call for careful application and transparency in deploying FER technologies powered by multimodal LLMs like Google's Gemini 2.5 Pro.

7. CONCLUSION

This research demonstrates the viability and effectiveness of using Google's Gemini 2.5 Pro within a multi-agent AI framework for facial emotion recognition (FER). By employing Gemini as the core inference engine (FER Agent) and embedding it within a coordinated system of Input, Evaluation, and Output Agents, we showed that Gemini 2.5 could autonomously classify emotions in static facial images with high accuracy, confidence alignment, and statistical robustness; all without prior training or image preprocessing. The agentic architecture enabled a seamless workflow from raw image ingestion to prediction evaluation, highlighting the power of large multimodal models (LMMs) in applied psychological and affective science tasks.

A particularly striking finding was Gemini's ability to scale beyond its published input limits. Although the model documentation caps image input at 3,000 per prompt, the system processed over 5,000 images. Despite this high throughput, the system maintained reliable prediction accuracy and confidence levels across all eight emotion categories. Visual and statistical analyses confirmed that the model not only performed well on easily distinguishable emotions like happiness and surprise but also handled subtler categories like sadness and neutrality with meaningful granularity. This outcome strengthens the case for deploying LLMs like Gemini in emotion-centered psychological research, user behavior modeling, and adaptive human-computer interaction (HCI) systems.

Looking forward, future work will focus on expanding the emotional and demographic diversity of the dataset, incorporating real-time video emotion analysis, and exploring the ethical dimensions of FER technologies in higher-risk domains such as healthcare and education. With continued improvements in prompt engineering and agent orchestration, the role of multimodal LLMs like Gemini 2.5 is poised to grow in both technical capacity and social relevance. This study contributes to that trajectory by showing that an experimental, zero-shot model when properly embedded in a multi-agentic design can deliver actionable, scalable insights into human affect.

8. REFERENCES

American Psychological Association. (2023). APA Dictionary of Psychology. Retrieved from https://dictionary.apa.org/emotion

Built In. (2025). What Is Gemini 2.5 Pro?. Retrieved from https://builtin.com/artificial-intelligence/google-gemini-2-5-pro

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Codecademy. (2025). How to Build Agentic AI with LangChain and LangGraph. Retrieved from https://www.codecademy.com/article/agentic-ai-with-langchain-langgraph

Datacamp. (2023). Python Seaborn Tutorial For Beginners: Start Visualizing Data. Retrieved from https://www.datacamp.com/tutorial/seaborn-python-tutorial

FacialNet. (2024): Facial emotion recognition for

ISSN: 2473-4901
v11 n6357

mental health analysis using UNet segmentation with transfer learning model. (2024). Frontiers in Computational Neuroscience. Retrieved from https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2024.1485121/full

GeeksforGeeks. (2025). Pandas Tutorial. Retrieved from https://www.geeksforgeeks.org/pandas/pandas-tutorial/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Google. (n.d.). Google Colab. Retrieved from https://colab.research.google.com/

Google. (2024). Gemini 2.5 Pro. Retrieved from https://deepmind.google/models/gemini/pro/

Google. (2025). Google for Developers. Retrieved from https://developers.google.com/machine-learning/resources/prompt-eng

Google Cloud. (n.d.-a). Vertex AI. Retrieved from https://cloud.google.com/vertex-ai

Google Cloud. (n.d.-b). Cloud Storage. Retrieved from https://cloud.google.com/storage

IBM. (n.d.). Prompt Engineering Techniques. Retrieved from https://www.ibm.com/think/topics/prompt-engineering-techniques

IBM. (n.d.). What is Scikit-Learn (Sklearn)? Retrieved from https://www.ibm.com/think/topics/scikit-learn

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. International journal of speech technology, 15(2), 99-117.

LangChain. (2025). How and when to build multi-agent systems. Retrieved from https://blog.langchain.com/how-and-when-to-build-multi-agent-systems/

Lin, K. (2025). LangGraph: A Framework for Building Stateful Multi-Agent LLM Applications. Medium. Retrieved from https://medium.com/@ken_lin/langgraph-a-framework-for-building-stateful-multi-agent-llm-applications-a51d5eb68d03

MoodMe. (2024). How Emotion Detection AI is Revolutionizing Mental Healthcare. Retrieved from https://www.mood-me.com/how-emotion-detection-ai-is-revolutionizing-mental-healthcare/

Neptune.ai. (n.d.). Zero-Shot and Few-Shot Learning with LLMs. Retrieved from https://neptune.ai/blog/zero-shot-and-few-shot-learning-with-llms

OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

Pramerdorfer, C., & Kampel, M. (2016). Facial expression recognition using convolutional neural networks: a survey. arXiv preprint arXiv:1612.02903.

Pluralsight. (2025). How to use LangChain and LangGraph for Agentic AI. Retrieved from https://www.pluralsight.com/resources/blog/ai-and-data/langchain-langgraph-agentic-ai-guide

Wooldridge, M. (2009). An introduction to multiagent systems. John Wiley & Sons.

Tian, Z., et al. (2025). Overconfidence in LLM-as-a-Judge: Diagnosis and Confidence-Driven Solution. Retrieved from https://arxiv.org/abs/2508.06225

ISSN: 2473-4901
v11 n6357

Appendices and Annexures

APPENDIX A

Emotion	Accuracy	Precision	Recall	F1-Score
anger	0.868492618	0.591811414	0.578181818	0.584917229
contempt	0.963480963	0.564516129	0.178571429	0.271317829
disgust	0.93006993	0.540145985	0.564885496	0.552238806
fear	0.937451437	0.72319202	0.578842315	0.643015521
happiness	0.910644911	0.780164245	0.916852264	0.843003413
neutrality	0.904234654	0.504807692	0.631262525	0.560997329
sadness	0.892968143	0.651851852	0.492307692	0.560956175
surprise	0.923271173	0.692094313	0.742559524	0.71643934
MACRO_AVG	0.665306915	0.631072956	0.585432883	0.591610705

Table 1: Performance Metrics – Overall and Per Class

GT_Emotion	Recognized Emotion	Misclassification
sadness	neutrality	119
fear	surprise	107
sadness	anger	107
anger	neutrality	84
anger	disgust	81

Table 2: Top Five Most Frequent Misclassification

ISSN: 2473-4901
v11 n6357

APPENDIX B

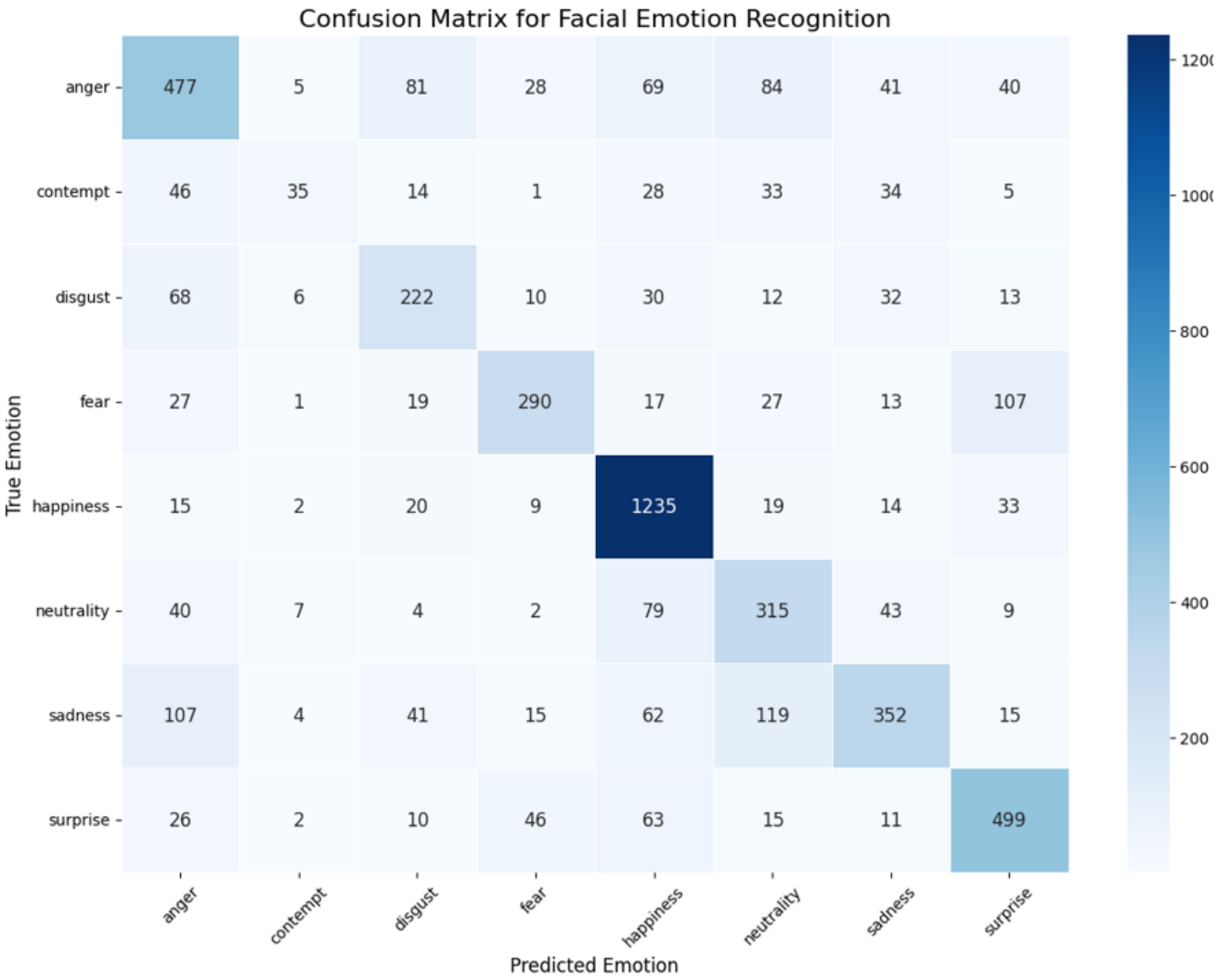
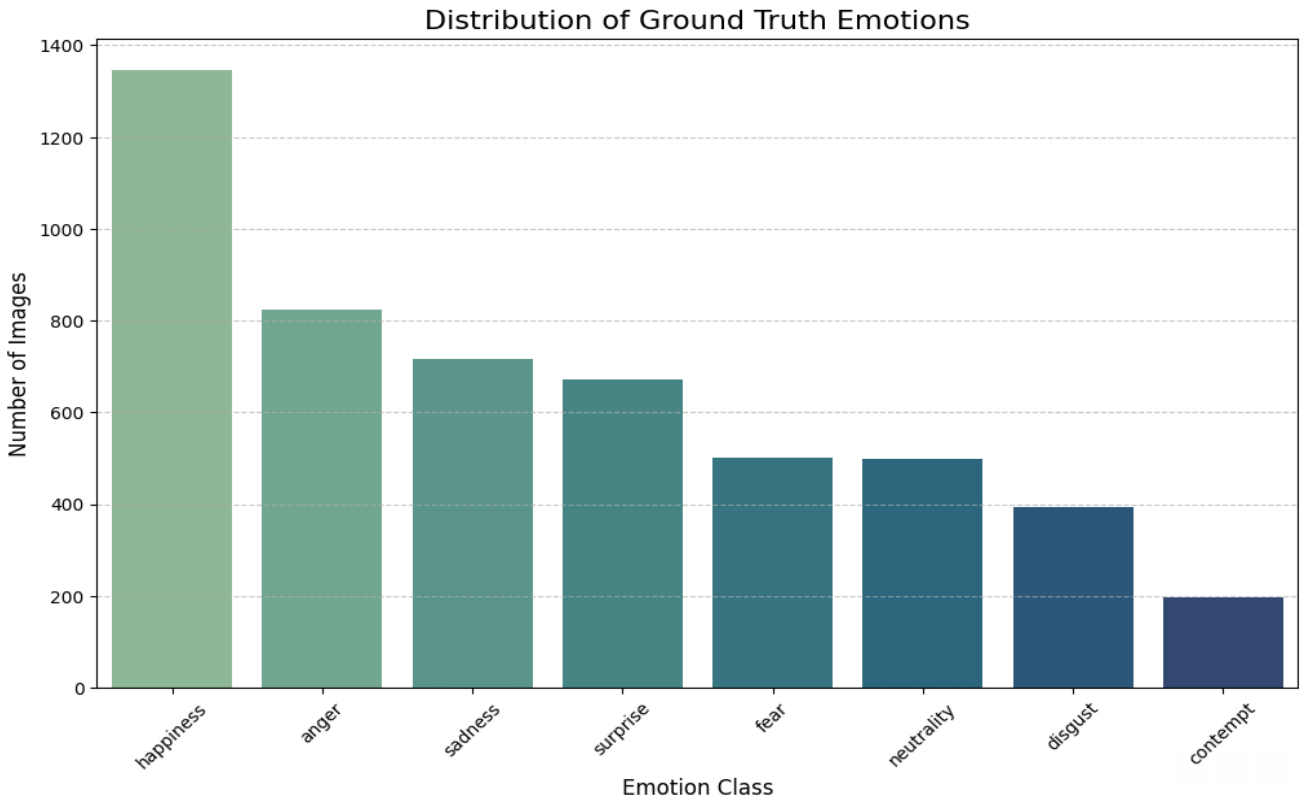
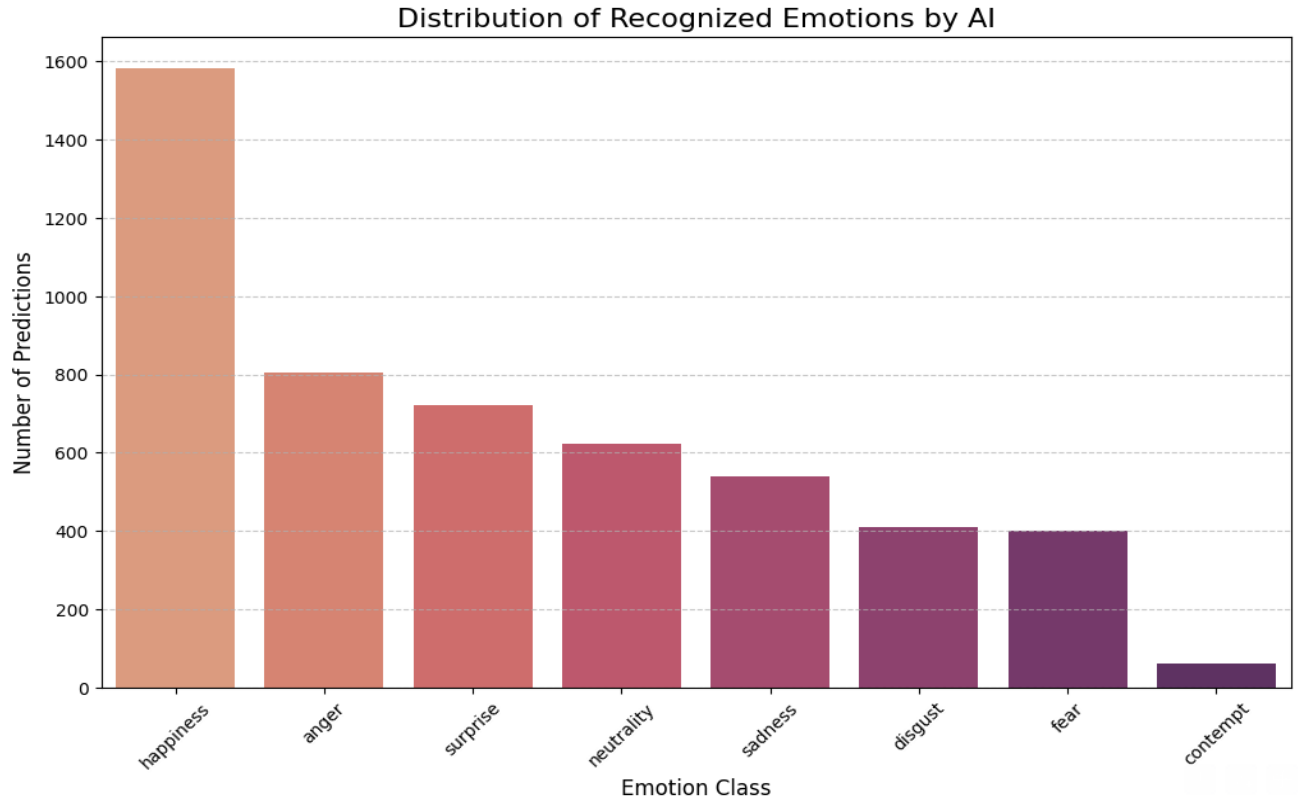


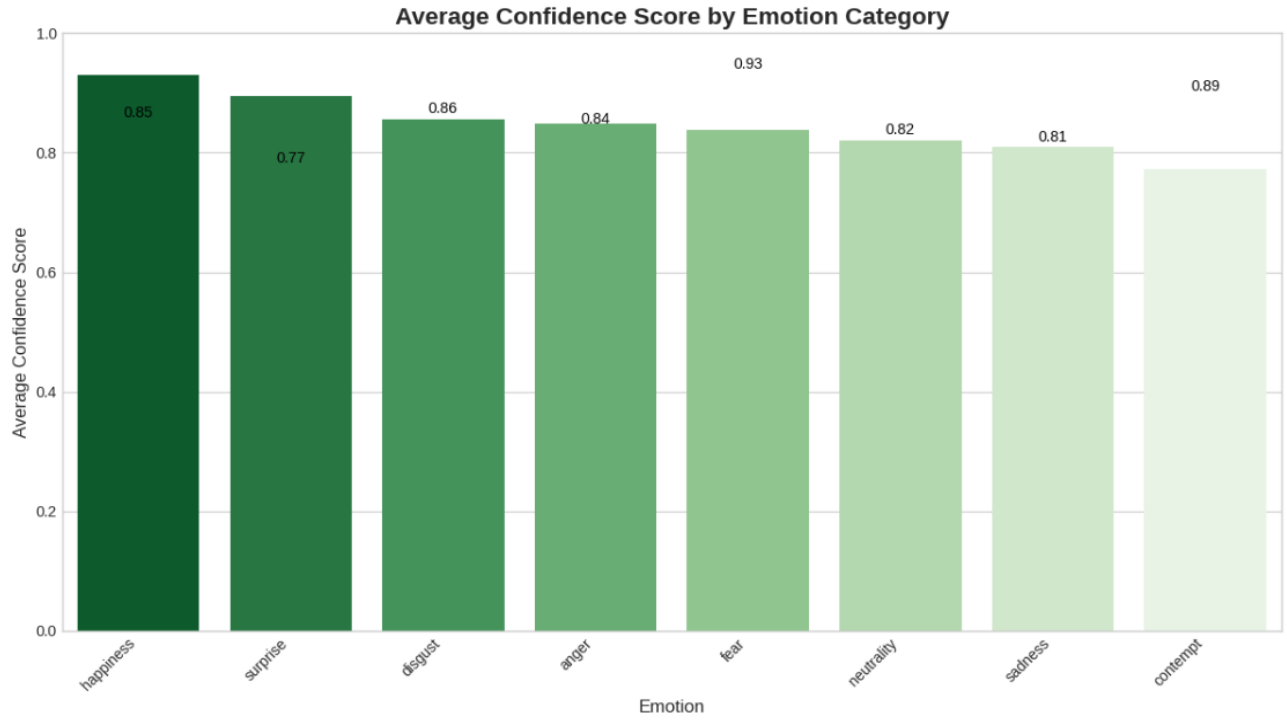
Figure 1: Confusion Matrix

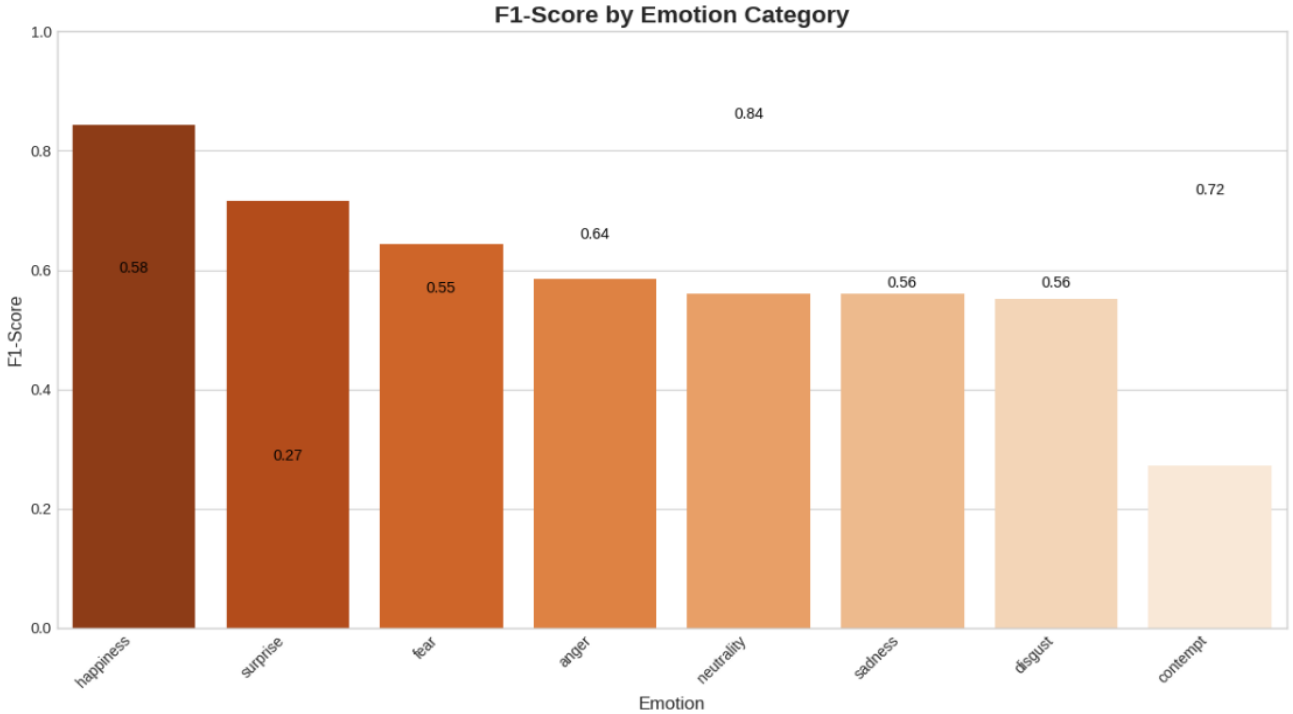
ISSN: 2473-4901
v11 n6357





ISSN: 2473-4901
v11 n6357





ISSN: 2473-4901
v11 n6357