Project Proposal: **Reaction Says it All**

Team: **Macrodata Refinement**

Garrard, Jeremy
jeremy.garrard@gatech.edu

ODonnell, Liam
lodonnell32@gatech.edu

Taskovski, Kelly
tlee459@gatech.edu

March 13, 2025

1 Project Summary

It has been said that a significant portion of human communication is nonverbal [4], and we aim to leverage this for more effective digital interactions. Our project seeks to analyze facial expressions to enhance empathetic communication.

We will pair detected emotions with a text prompt and a relationship label. By combining these three elements, we can infer how someone might feel when receiving a message based on their facial expression.

Using this information, our system will generate responses that encourage thoughtful, patient, and empathetic communication.

For example, imagine receiving a text from your mother asking if you will be careful on a trip for the hundredth time. You might feel slightly annoyed and either ignore the message or respond hastily. Our system will detect this frustration, interpret the context, and draft a more considerate response for user approval.

2 Approach

Our implementation consists of three key components:

2.1 Facial Emotion Recognition

We will train a convolutional neural network (CNN) on a labeled dataset of facial expressions to detect basic emotions from images. A key aspect of this model will be speed as eventually this solution could be implemented on mobile devices. The search for a state-of-the-art (SOTA) architecture for image classification, particularly for facial reactions, that has relatively fast time is a key part of the project. In *related work*, we discuss some SOTA approaches to speed up image classification. Given the potential highly diverse userbase, finding a training set with a wide range of faces, racial backgrounds, lighting, etc. is paramount. This will ensure that all users have the same experience.

2.2 Prompt Generation

A separate model will generate communication prompts based on three inputs: the user's **detected emotion**, the **received text**, and the **relationship label**. We may experiment with attention mechanisms to balance these inputs.

- Example input: `[frustrated, "I think you should get out more!", mother]`
- Expected prompt: *"Help me reassure my mother that I have a social life while appreciating her concern and maintaining a healthy boundary."*

If the individual's name is not indicative of the relationship with the user, we can further experiment with the context of previous messages to request a specific tone of the prompt `[friend, family, professional]`.

2.3 Response Generation

Using a transformer-based model (such as GPT-4), we will generate a suggested response based on the generated prompt. If implemented successfully, our system will intuit the emotions of the user and help facilitate more thoughtful digital communication.

3 Resources / Related Work

3.1 Facial Emotion Recognition

Classifying reactions have been an active area of research over the past decade. paperswithcode [6] provides **Figure 1** demonstrating that model accuracy improvements generally increase each year. Currently, the SOTA implementation is Norface [11] which is based on normalization of identity and expression. Additional implementations of facial expression classification are as follows:

- **RetinaFace** [1]: RetinaFace is a single-stage face detection model that uses a broader definition of face localization to provide accurate facial position information for all different scales.

- **Multi-task Cascaded Convolutional Networks** [9]: MTCNN adopts a cascaded structure with three stages of carefully designed deep convolutional networks that predict face and landmark location in a coarse-to-fine manner.

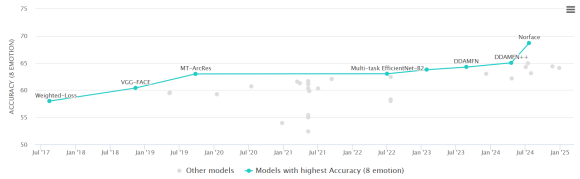

Figure 1: State-of-the-art reaction classification

3.2 Prompt Generation and Response Generation

- **InstructGPT** [7]: InstructGPT is a model that aligns language models with user intent on a wide range of tasks by fine-tuning with human feedback.

4 Datasets

A key point of discussion when deciding on a topic was the availability of datasets. The time spent labeling data sets would take away from model fine-tuning and creating a working solution. Many of the best performing models utilize labeled expression datasets such as BP4D+ [10], DISFA [3], AffectNet [5] , RAF-DB [2], and WIDERFACE [8] which will prove to be valuable data and we use a prebuild architecture and fine tune further. Additionally, we would like the generated text to sound natural and flow with the digital conversation. Using AI generated text datasets to ensure natural-sounding responses are generated will prove useful. The following Kaggle datasets may prove useful for this stage in the project:

- https://www.kaggle.com/datasets/thedrcat/daigt-v2-train-dataset
- https://www.kaggle.com/competitions/llm-detect-ai-generated-text/data

References

[1] Jiankang D. et al. “RetinaFace: Single-stage Dense Face Localisation in the Wild”. In: *ArXiv e-prints* (2020). arXiv: 1905.00641v2.

[2] S. Li, W. Deng, and J. Du. “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2852–2861.

[3] S.M. Mavadati et al. “Disfa: A spontaneous facial action intensity database”. In: *IEEE Transactions on Affective Computing* 4.2 (2013), pp. 151–160.

[4] Albert Mehrabian. *Nonverbal Communication*. Chicago, IL: Aldine-Atherton, 1972.

[5] A. Mollahosseini, B. Hasani, and M.H. Mahoor. “Affectnet: A database for facial expression, valence, and arousal computing in the wild”. In: *IEEE Transactions on Affective Computing* 10.1 (2017), pp. 18–31.

[6] Papers With Code. *Facial Expression Recognition on AffectNet*. `https://paperswithcode.com/sota/facial-expression-recognition-on-affectnet`. Retrieved [Insert Date Here]. n.d.

[7] Long Q. et al. “Training language models to follow instructions with human feedback”. In: *ArXiv e-prints* (2022). arXiv: 2203.02155.

[8] Shuo Yang et al. “WIDER FACE: A Face Detection Benchmark”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[9] Kaipeng Z. et al. “Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks”. In: *ArXiv e-prints* (2016). arXiv: 1604.02878.

[10] X. Zhang et al. “Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database”. In: *Image and Vision Computing* 32.10 (2014), pp. 692–706.

[11] Yu Zhang et al. “Learning Lightweight and Accurate Facial Expression Recognition with Multi-Granularity Knowledge Distillation”. Version v1. In: *arXiv preprint arXiv:2407.15617* (2024). arXiv: 2407.15617 [cs.CV].