MoodMosaic: Mapping Emotions and Personality Through Language

Hanshitha Mahankali
George Mason University
hmahanka@gmu.edu

1 Introduction

MoodMosaic was a small research prototype that looked into the various ways that emotions, tones and personality could be inferred from everyday use of language. The goal of MoodMosaic is to take free-form written text and condense it down into a smaller, more digestible form for the user. This includes creating a visual map of the users' emotional richness, seeing if the user exhibits any tone of politeness and building a 5-dimensional personality model based on the Big Five. While we do not intend to claim any psychological validity, we see MoodMosaic as a helpful analytic tool that can be used for self-reflection, communication skills training and small exploratory studies.

Our initial plan consisted of developing a complete pipeline using large publicly available datasets, performing numerous ablation studies that we planned to run over the course of the research project. This report outlines what we were able to produce by the original deadline - three neural-networks: one for detecting emotion, one for detecting tone and one for detecting personality; an aggregation layer to consolidate/report out these neural-networks; a personality-recognition model that is trained on a public dataset using the Big Five inventory and a user-friendly front-end display, built using Streamlit. We will be transparent regarding the number of simplifications we made due to computational resources and timing constraints; therefore, the original eventful pipeline that we laid out in our project plan will serve as a roadmap for any further development or future work.

2 Background and Motivation

The classical approach to sentiment analysis relies on the calculation of positive or negative polarity with an intermediate neutral point. Recent studies have demonstrated that deep neural approaches, such as BERT and RoBERTa, improve the prediction of sentiment and emotion detection when large amounts of annotated data are available ((Hu and Liu, 2018; Poria et al., 2019). Additionally, the area of computational pragmatics investigates the relationship between politeness marking, hedging behaviour, and socialisation and power dynamics ((Danescu-Niculescu-Mizil et al., 2013). A third area of research focuses on how language reflects human personality, with one notable example being the Essays corpus analysed using the Big Five personality model (Mairesse et al., 2007).

MoodMosaic combines all three of these research avenues. Rather than producing a single state-of-the- art sentiment prediction tool, we utilise a number of lightweight models to create one unified representation of a user and/or conversation. The end goal of MoodMosaic is to be a tool that acts as a reflective mirror for the user; in that as the user inputs text, they are able to view a visual representation of their mood without being forced to contend with the uncertainty of a black box.

3 Data and Tasks

3.1 Emotion: GoEmotions

The GoEmotions dataset provides 34 different emotions, in addition to 'neutral', which we will be using for the emotion classification task. For our training, we are using Hugging Face's Simplified Split of the data. The Simplified Split provides a Reddit comment as an example, along with labels designated by an index. Once we remove all but the 27 most common emotions used in our research (along with 'neutral'), we have a total of 43,410 training samples; 5,426 validation samples; and 5,427 test samples.

We will treat all samples as a multi-label item, with each sample's emotion being assigned a probability by our model, and we will apply a threshold for the final label assignment. For the dashboard component, we are displaying the most probable emotion(s) given by the model for each sample and also providing an overall summary of the distribution of the emotions across the entire test dataset.

3.2 Tone: Stanford Politeness

Using the Stanford Politeness Corpus, which has been cleaned and made available for download via the Cleanlab repository, we selected thank you cards based on their labels within an impolite/polite classification scheme. The split as indicated by Hugging Face shows that each unique request/response pair has been assigned a label of either or where the label of "impolite" corresponds to "0" and the label of "polite" corresponds to "1".

Once we had established the split file, we created both the training and validation portions of our dataset by selecting examples at random from within their respective strata (i.e. stratified sampling); thus the training set consists of 542 items and the validation set consists of 136 items. While representing only a small subset of the total available data, the items chosen from within the training and validation datasets include naturally occurring requests and responses that demonstrate the multiple ways in which people communicate requests and responses, thus providing a reasonable basis for demonstrating a simple tone classifier on the dashboard application.

3.3 Personality: Essays Big Five

For the personality scores, we utilized the Essays Big Five corpus from Hugging Face `jingjietan/essays-big5`. The dataset contains essays written by volunteers, along with scores (from 0 - 100) for each of the Big Five traits—openness, conscientiousness, extraversion, agreeableness, neuroticism—on a continuous scale. The authors of the essays partitioned the dataset into training, validation, and test according to the distribution specified in the original dataset. To create our training loop, we used the training and validation sets, which consisted of 1,578 essays for training and 395 essays for validation. We used these datasets to build both offline experiments and to develop and evaluate the personality assessment models used in the Streamlit dashboard. In the user interface, the models take all messages in the session, concatenate them together, and return a five-dimensional vector representing the user's personality profile, which is then normalized and visualized.

Task	Train	Val	Test
Emotion (GoEmotions)	43410	5426	5427
Tone (Stanford Politeness)	542	136	0
Personality (Essays Big Five)	1578	395	494

Table 1: Summary of the labeled datasets used in MoodMosaic. Counts show the number of instances in each split. For tone we use a train and validation split derived from one partition and do not reserve a separate test set.

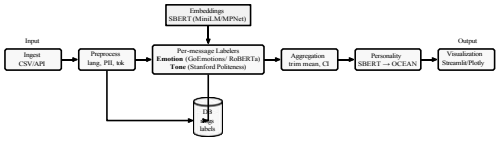

Figure 1: Pipeline: ingest, preprocess, per message labelers for emotion and tone, aggregation, personality regression, visualization.

4 Methods and Implementation

4.1 Overall Pipeline

Figure 1 gives an overview of the full system. A batch of messages enters the pipeline through the dashboard input. Each message is preprocessed using standard tokenization from Hugging Face and spaCy. The emotion and tone models run at the message level and produce probabilities or labels. An aggregation module then pools the emotion scores across messages, and a personality head processes the concatenated text to produce a five dimensional OCEAN profile. The final view presents three elements: an emotion radar chart, a personality radar chart, and a table with per message predictions.

4.2 Emotion Model

To classify emotions, we fine-tuned a RoBERTa base encoder. The classifier head consisted of a linear layer that received the pooled hidden state as input. We used binary cross entropy loss with logits for training. This is a standard approach

when doing multilabel predictions. Instead of using the high-level Trainer class, we opted to write our training loop directly in PyTorch. For the optimizer, we selected AdamW, with an initial learning rate of $2 \cdot 10^{-5}$, and a batch size of 16 and trained for three epochs. All of our training was performed on a CPU in our local environment. While this process is not as fast as training on a GPU, it was adequate for our small experimental needs.

Throughout the process of training, we monitored validation loss, macro F1 score, and accuracy. The best checkpoint based on the macro F1 score was saved so that we could later include it in the MoodMosaic pipeline. We used a sigmoid to get predicted probabilities and then applied a binary threshold to determine which emotions were present during inference.

4.3 Tone Model

The training data consists of 678 labeled examples from the test split of the Stanford Politeness Release; we performed a stratified split into 542 training examples and 136 validation examples. We trained for three epochs using AdamW Optimizer, with a learning rate of $2 \cdot 10^{-5}$.

As we trained, we logged both training loss and validation loss, as well as accuracy. We stored the checkpoint with the highest validation accuracy as our tone model, which can be accessed through the Streamlit Dashboard and used to classify all lines of text as either polite or impolite.

4.4 Personality Model

For personality we adopt a two stage approach. We use the Sentence Transformers model `all-mpnet-base-v2` to obtain a fixed length embedding for each essay or block of text. This embedding is then passed into a small multilayer perceptron regressor with one hidden layer of size one hundred twenty eight and an output layer of size five. The model predicts continuous values for the OCEAN traits.

We load the Essays Big Five dataset splits from Hugging Face, which provide one thousand five hundred seventy eight training essays and three hundred ninety five validation essays. The sentence encoder is kept frozen and only the multilayer perceptron head is trained. We use mean squared error as the loss function and AdamW with a learning rate of $2 \cdot 10^{-5}$ and batch size eight, training for five epochs. The checkpoint with the lowest validation loss is stored as the personality model.

Component	Encoder	Objective
Emotion	RoBERTa base	Multi-label BCE
Tone	DistilBERT base	Cross entropy
Personality	all-mpnet-base-v2 + MLP	Mean squared error

Table 2: Backbones and losses for the three MoodMosaic components. BCE denotes binary cross entropy.

Our tone classifier is based on DistilBERT's base uncased model and has a very small classification head. The label space consists of 2 very distinct and different classes - impoliteness and politeness. We tokenized all requests/messages with the standard uncased tokenizer for this task. Our model was trained to optimize the cross-entropy loss.

4.5 Model Summary

To make the architecture easier to compare across tasks, Table 2 lists the main encoder and training objective used by each component.

4.6 Dashboard and Deployment

The interactive front end is built in Streamlit. The main file `src/dashboard/app.py` loads the emotion model through a shared `MoodMosaicPipeline` class and then loads the tone and personality models directly from their checkpoints.

Two input methods are available for users to use input data into their Dashboard. If users choose to enter data using text entry, they can either write or copy/paste a comment to the Text Entry area, where only one comment may be entered per row. Users can also view a limited number of default comments in the sidebar (i.e., "I had a good day," "I had a bad day, I was angry and frustrated," "I was polite, but impolite," or a random combination of short characteristics based on the Big Five traits). Once selected, these comments will automatically populate the comment field in the Text Entry area. When uploading via CSV upload (by clicking on the CSV button), users may upload a comma-separated values file, indicate which column contains their comment(s), and run analyses on each row(s) within that column. Users can analyze a small amount of publicly available datasets (e.g., surveys, chat logs) without requiring developers to write new code.

The Dashboard has two sections (columns).

Element	Description
Input mode	Free text box or CSV upload with column selector
Presets	Predefined positive, frustrated, mixed tone and synthetic personality examples
Overview tab	Emotion radar and Big Five radar over all messages
Details tab	Per message table with dominant emotion, tone, filters, and CSV export

Table 3: Main user facing features in the MoodMosaic dashboard.

The left column has the controls for entering text or uploading a CSV file and an Analyze Now button. In the right column, you'll see session details including the type of device providing access and information about how many emotional labels were generated when analyzing the data using the dashboard and what tone classifications and personality trait classifications were generated based on that data.

The Outcomes section includes a total of 2 Tabs. The first Tab (Overview Tab) includes a Plotly-generated Emotion Mosaic Radar Chart, which represents the total distribution of emotional labels generated from sorting through GoEmotions Labels and displays the Five Big Five Personality Traits as a Personality Profile Radar Chart. The second Tab (Per Message Details Tab) includes a grid/table listing each message sent/received in the order that it was sent/received, including the dominant emotion and tone label associated with each of those messages. This Tab will allow users to filter for specific tones/emotions and view all the associated labels together in one table. In addition, there is a separate button on the Overview Tab to download the full list of emotion and tone label predictions as a CSV file for further analysis, when needed, using either a spreadsheet or notebook.

Table 3 summarizes the main user facing features of the dashboard.

We run this app locally during development and also host the code on GitHub. For classroom sharing we run the interface on a personal laptop so that the results match the locally fine tuned models.

4.7 Implementation Details and Reproducibility

To keep the prototype reproducible, we record a few practical details. All training and inference for this project are performed on a single laptop without a dedicated GPU. The training scripts live under `src/training` and are invoked as Python modules, for example:

```
python-msrc.training.
train_emotion_goemo
python-msrc.training.
train_tone_stanford
python-msrc.training.
train_personality_big5
```

Each script writes the best checkpoint under `experiments/checkpoints` together with a short log file.

We fix a single random seed for NumPy and PyTorch inside each script so that runs are locally stable. The environment is managed through a virtual environment with pinned versions of `transformers`, `sentence-transformers`, `pandas`, and `streamlit`. The repository includes a requirements file so that a new user can recreate the environment and retrain or reuse the models with a few commands.

5 Experiments and Results

5.1 Evaluation and Metrics

Each component of the model was evaluated with a held out validation dataset taken from the same dataset. For both classification tasks, we report macro F1 score and accuracy; for the personality regressor, we report MSE and RMSE. The key findings of this study are summarised in Table 4.

The emotion classifier represents a fine-tuned RoBERTa model trained on the GoEmotions dataset, which was then evaluated on the official GoEmotions dataset validation split. The resulting model obtained a Macro F1 score of 0.488, with an accuracy of 0.433. Considering the nature of the label space (very fine-grained emotional categories), as well as the training epoch limit (three) and hardware constraints (CPU), this score appears reasonable. A review of the output illustrated that the model captured a strong overall trend of positive versus negative, but often associated the wrong emotional category due to confu-

sion between the two most related categories e.g., admiration versus gratitude.

In order to classify politeness, we trained the DistilBERT model on the cleaned Stanford Politeness dataset. We created a stratified random split from the provided test partition; 542 samples were used for training, and 136 samples were used for validation. The best model checkpoint from this validation dataset had an accuracy of 1.0 with extremely low validation loss. However, this unusually high accuracy should be taken into account because it results from the small sample size and only using one random split, so it represents a maximum potential value rather than a reliable measure of the model's performance. Looking at the qualitative data from the dashboard, we can see that the model behaves as we would expect it to — that is, for simple polite requests, it includes phrases such as hesitating and thanking and that for impolite requests, it makes very short and direct statements.

To derive the personality profile, we used the SBERT algorithm and a multilayer perceptron (MLP) regression model on the Essays (Big Five) dataset. The best model found on the validation split achieved an average mean squared error of 0.259 and an average root mean squared error of 0.509 over the five traits. We consider this to represent the de facto moderate level of error for this analysis; therefore, users should interpret any results reported using this system as general indicators of comparisons of the strengths of certain traits to others rather than precise psychological quantifications of each trait.

In our informal testing of the model (using very short phrases or sentences) we found that the resulting personality profiles assigned a greater degree of Openness when presented with creative types of writing (e.g. texts on creativity and new discoveries) while they added the highest degree of Neuroticism when engaging with written material that signified mental stress (e.g. writing about worry or anxiety). This is consistent with a common understanding of the associations between traits.

In conjunction with the intrinsic measures mentioned in the previous paragraph, we also conducted an exploratory study using an interactive dashboard of our model. In this test we used handwritten templates which represented positive wording. Our templates represented all four combinations of positive/negative wording and sarcastic/honest. Using the templates, we generated "dummy" data and created mockups of personal profiles that would be indicative of personality types typically found within the Big Five model. We utilized the dashboard's functionality to cross-verify the representative profile with the dominant emotions, tone indicators, and trait rankings. This comparison informed our evaluation of the dashboard. While this evaluation will provide insight into whether the three components of the interface work together coherently, we wish to point out that this evaluation is not a formal user study.

5.2 Training Outcomes

Table 4 summarizes the main quantitative results we obtained on the validation splits for each task. The emotion and personality rows are based on real public corpora. The tone evaluation uses a held out slice from the Stanford Politeness test partition and should be read with some caution because the total data size remains modest.

The emotion model has produced a macro F1-statistic of approximately 0.48 along with an accuracy of 0.43 on the validation actions. This makes sense, considering that we have a very fine-grain label space, and we have used only three epochs on a CPU basis for our study. The classifier seems to be good at identifying broad positive vs. negative feelings, but it is not as good at distinguishing closely related emotions (e.g., admiration vs. gratitude).

The politeness classifier has a validation accuracy of one along with a very low validation loss. Although this sounds excellent, it must be interpreted with caution because the training model has been trained with a small slice of the total training corpus. On qualitative evaluation, we notice that it recognises phrases with explicit gratitude or hedged requests as polite and short direct commands as impolite and meets our needs for demonstration dashboards.

The frozen SBERT plus multilayer perceptron model on the Essays Big Five dataset achieves an approximate mean squared error of 0.26 and a root mean squared error of 0.51 on the validation set over all of the traits. As a way of investigating how the model behaves, we supplied it with a small number of sample sentences that contained descriptions of common behaviours associated with each of the five personality traits. As

Task	Model	Train size	Val size	Main metric	Score
Emotion (multi label)	RoBERTa base classifier	43 410	5 426	Macro F1 / Accuracy	0.488 / 0.433
Tone (binary politeness)	DistilBERT base classifier	542	136	Val loss / Accuracy	0.008 / 1.000
Personality (OCEAN, Essays)	SBERT + MLP regressor	1 578	395	Val MSE / RMSE	0.259 / 0.509

Table 4: Validation results for the MoodMosaic components. The emotion model uses the GoEmotions corpus and gives reasonable scores for a simple baseline. The tone model is trained on a split of the Stanford Politeness data and reaches very high accuracy on the small validation set. The personality regressor is trained on the Essays Big Five corpus and achieves a moderate mean squared error across the five traits.

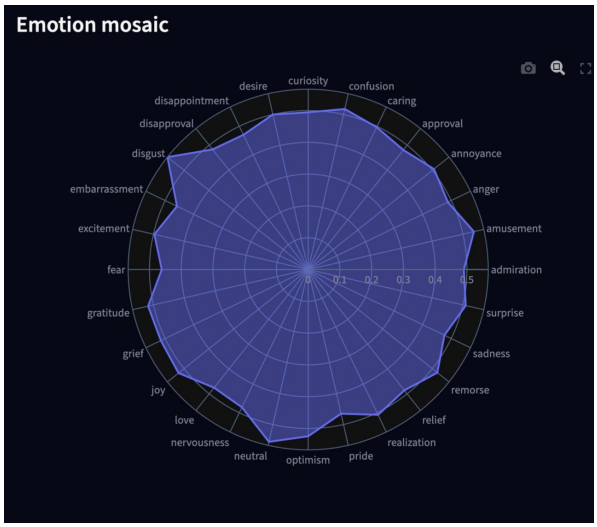

Figure 2: Example emotion radar visualization from the Streamlit dashboard for four short messages. The user can inspect which emotions are most active overall.

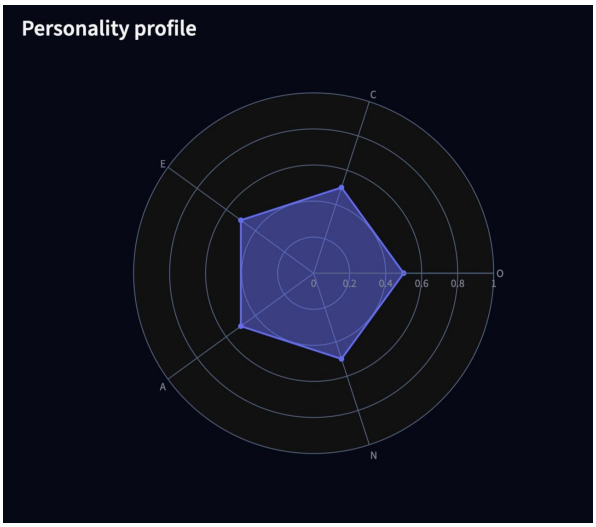

Figure 3: Personality radar for the joined text of the same example. This view aggregates across all messages and displays the five Big Five traits.

an example, one of the sample sentences describes someone's enjoyment of new ideas and creative projects and scores highest on the Openness dimension when compared to the remaining personality traits. Whereas a sample sentence that mentions being concerned about outcomes and having a high level of anxiety when faced with uncertain situations generates the highest Neuroticism score. The scores generated by this model fall in a moderate range and should be viewed as low-confidence indicators rather than definitive identifiers of a person's true personality.

5.3 Dashboard Behaviour

Figure 2 shows a screenshot of the emotion mosaic for a simple four line example that includes a joyful sentence, a frustrated complaint, a grateful thank you, and a critical request. The radar chart places most mass on joy, gratitude, annoyance, and disapproval, which matches human intuition.

Figure 3 shows the personality profile for the same text block under the Essays Big Five model. The scores suggest moderate openness and extraversion, with somewhat lower neuroticism. We stress that these values are not clinically meaningful; they simply demonstrate how such a profile can be visualized.

In the lower section of the dashboard, there is a small table showing a summary of all your messages. In my example, it shows that the positive message used a friendly and pleasant way to express joy, the upset message used an angry and aggressive expression of annoyance, the response to "Thank You" used a warm and kind tone, and the request for feedback used an offensive expression of disapproval. This allows users to identify exactly how they might be speaking to someone and thus allow for appropriate responses. The presets and filters available in the sidebar allow users to quickly check the various tones and emotions of their messages while they are giving live demonstrations. Users will also be able to use the upload CSV capability and the download of predictions button for more realistic workflow options; for ex-

ample, uploading a small public dataset and then viewing and exporting the predictions so they can further analyze them statistically in the future.

6 Discussion and Limitations

The MoodMosaic prototype currently integrates three forms of analysis in a single integrated interface. The prototype's emotion model has been developed using a large amount of real data and can provide useful granular insights into emotions. The politeness classifier and the personality regressor based on essays have been developed using public datasets and have been incorporated within the dashboard. The prototype's architecture is general-purpose, making it simple for classroom use.

However, several limitations exist. The most important limitation is that the politeness dataset used for calibration is comparatively small and therefore the high degree of validation accuracy will almost certainly reflect inflated results on new domains. The second limitation is that the calibration and uncertainty assessment functions contained in the original project plans were not delivered as part of the final product. Thirdly, there are no rigorous evaluation methods provided within the current results, which limits our ability to accurately assess reliability and uncertainty.

Another key ethical limitation of this project is that utilizing written text to speculate on someone's personality is subject to unethical use. We restrict our focus of this research to the educational context, provide clear information stating results cannot be used as psychologic analysis and will avoid the misuse of such results in sensitive situations. Any future work based on this project will have to go through a formal ethics review process.

7 Future Work

There are several natural next steps that build on this prototype.

- Train the politeness classifier on the full Stanford corpus with more careful cross validation and regularization, and evaluate on held out domains.
- Explore light calibration for the emotion and tone models so that aggregate plots better reflect relative uncertainty and avoid overconfident labels.
- Add richer visualizations in the dashboard such as time series plots for long running chat logs or heat maps across conversation turns.
- Extend the aggregation logic to handle entire conversations, with options for per speaker and per time window summaries and comparison across different threads.
- Conduct a small user study to measure how understandable and helpful the visualizations are for students or research collaborators, and use this feedback to refine the interface.

8 Conclusion

By using a single transparent pipeline and user interface as a way of combining emotion, tone and personality analysis together into one system, MoodMosaic has demonstrated the feasibility of such an approach. While the models currently available are relatively simple compared to more advanced models, they provide an easily accessible means of observing the relation between language and affect/style. With way more robust training data along with additional careful evaluation, the MoodMosaic system offers significant opportunities for supporting teaching and research related to language, emotion and personality.

9 Team Work Clarification

The MoodMosaic system was developed collaboratively by all three members through the co-development of the model's pipeline, the selection/non-theoretical basis of its datasets, the selection of its architectures, and the finalization of the components by which emotional, tonal, and personality predictive data can be fed into the system. In order to ensure that the project would progress collaboratively and consistently through the feasibility analysis of each development phase, all three members of the MoodMosaic team participated in discussions and project planning.

Abhimanyu Singh was part of the MoodMosaic team in creating a foundational model structure based on a core set of transformers by reviewing and analysing previous models (e.g., RoBERTa, DistilBERT, Sentence-BERT). Under his direction, the team outlined how to use existing models such as RoBERTa, DistilBERT, and Sentence-BERT to perform predictions of emotions, tones and personalities of users. He provided assistance with decisions related to the overall design and

development of each model's prediction workflow within the MoodMosaic system.

Hanshitha Mahankali is focused on understanding and preparing the dataset used throughout the project. She is analyzing the structure and appropriateness of GoEmotions, the Stanford Politeness Corpus, and the Essays Dataset for Big Five traits. It is also her responsibility to define the preprocessing required for each of these and assure they are in line with their respective model components.

Sravani Kadavakollu contributed to the overall architecture design of the system. She is also involved in planning for the interactive dashboard that visualizes the output, such as radar charts and emotional heat maps. Her contributions include integrating the output of the models into the User Interface design and helping to define the overall evaluation plan and metrics for testing the system.

Source Code

https://github.com/abhimanyusingh00/moodmosaic

References

Veronique Hoste Carlo Strapparava Orphee De Clercq Alexandra Balahur, Roman Klinger. 2020. Exploiting bert for empathy and emotion classification. In *Proceedings of the 10th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Minqing Hu and Bing Liu. 2018. Text sentiment analysis: A review. *International Journal of Data Mining and Knowledge Management Process*.

Vinod Kumar, Akshit Bansal, and Umang Gupta. 2023. Comparative analysis of text based emotion detection on goemotions dataset. In *2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, Greater Noida, India. IEEE.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for personality recognition. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. 2020. Recent trends in deep learning based personality detection. In *Proceedings of the Workshop on Computational Personality Recognition at EMNLP*.

James W Pennebaker, Martha E Francis, and Roger J Booth. 1999. Linguistic inquiry and word count (liwc): A computerized text analysis program. *Erlbaum Publishers*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*.

Fan Xing, Humaira Ashraf, Uswa Ihsan, Navid Ali Khan, and Rajesh Bahuguna. 2024. Social media text sentiment analysis: Exploration of machine learning methods. In *Proceedings of the 2024 IEEE 1st Karachi Section Humanitarian Technology Conference (KHI-HTC)*, Tandojam, Pakistan. IEEE.