UR-112

Monarch: A Privacy-focused NLP Model for Emotional Pattern Detection

Abstract

Introducing: Monarch — a privacy-focused deep learning model that interprets emotional patterns in text. Monarch is trained on large, lexicon-based datasets and uses fine-tuned NLP models (BERT) to identify patterns associated with sadness, worry, anger, and distress. It runs entirely offline with no data collection, making it ideal for private use. Monarch evaluates text and returns clear, readable probability scores across emotional categories, giving users insight into emotional trends. Monarch is interpretive, not diagnostic, displaying results based on scientifically backed linguistic patterns. Its potential use in schools could help flag early signs of distress, giving educators a chance to support those in need. Monarch is also suitable for research in linguistics, mental health, and ethical AI implementations.

Introduction

Text-based emotional analysis in modern day plays a crucial role in understanding signals within individuals, especially when they express distress, anger, or sadness through online messages, or forum posts.

Existing models rely on cloud-based infrastructure that invade your data privacy. This can be problematic and may raise ethical concerns in private or school settings. On the other hand, Monarch runs completely offline and local. Focusing only on interpretive emotional feedback.

Monarch uses the BERT NLP model that is tuned on emotion-labeled datasets to appropriate a categorical response within 4 categories; sadness, worry, anger, and distress. Monarch explores how patterns within the way we talk can reveal subtle emotional signals, allowing early, private, and interpretable insights into our mental and emotional state.

Research Question(s)

- What words most frequently correlate with emotional distress?
- How accurate is emotion classification when models are trained on lexicon-tagged emotional data like GoEmotions or Reddit?
- Can a fine-tuned deep learning model identify emotional cues (e.g., sadness, worry, anger, distress) in text-based language?

Materials and Methods

- We began by sourcing three major datasets: GoEmotions, a labeled dataset from Google containing over 58k Reddit comments tied to 27 emotion categories, and a custom Reddit-scraped dataset scored using anxiety-related lexicons and keyword heuristics.
- All text was cleaned using a custom pipeline that removed noise (e.g., punctuation, links, emojis) and normalized content for model usage.
- Texts were then labeled into five categories: sadness, worry, anger, distress, and joy.
- We used **BERT** (Bidirectional Encoder Representations from Transformers) as our main deep learning model for lexicon-based sentiment comparisons. BERT was fine-tuned on our cleaned emotional dataset using HuggingFace's Trainer pipeline.
- The model was trained and validated entirely offline using PyTorch, HuggingFace Transformers, and with local GPU/CPU.
- Visualizations were generated using Seaborn/Matplotlib, and predictions were displayed as clear emotion probability scores across categories.
- All emotional processing is handled directly on the user's device to maintain complete privacy. A lightweight, Raspberry Pi-compatible version will showcase the portability, and locality of this model.

Results

Results:
- The Monarch model was trained on a balanced dataset of 3,860 text entries distributed equally across five emotional categories: **sadness, anger, distress, joy, and worry**.
- Each label was represented with **772** samples.
- The model used a fine-tuned BERT-base architecture, trained for **5** epochs using mixed-precision (fp16) on an NVIDIA 2070 Super GPU.

Training and Validation:
- Final training loss: **0.07** (See Fig. 2)
- Validation loss plateaued at approximately **0.55** by epoch 3
- Total training duration: approximately **12** minutes (GPU-accelerated)
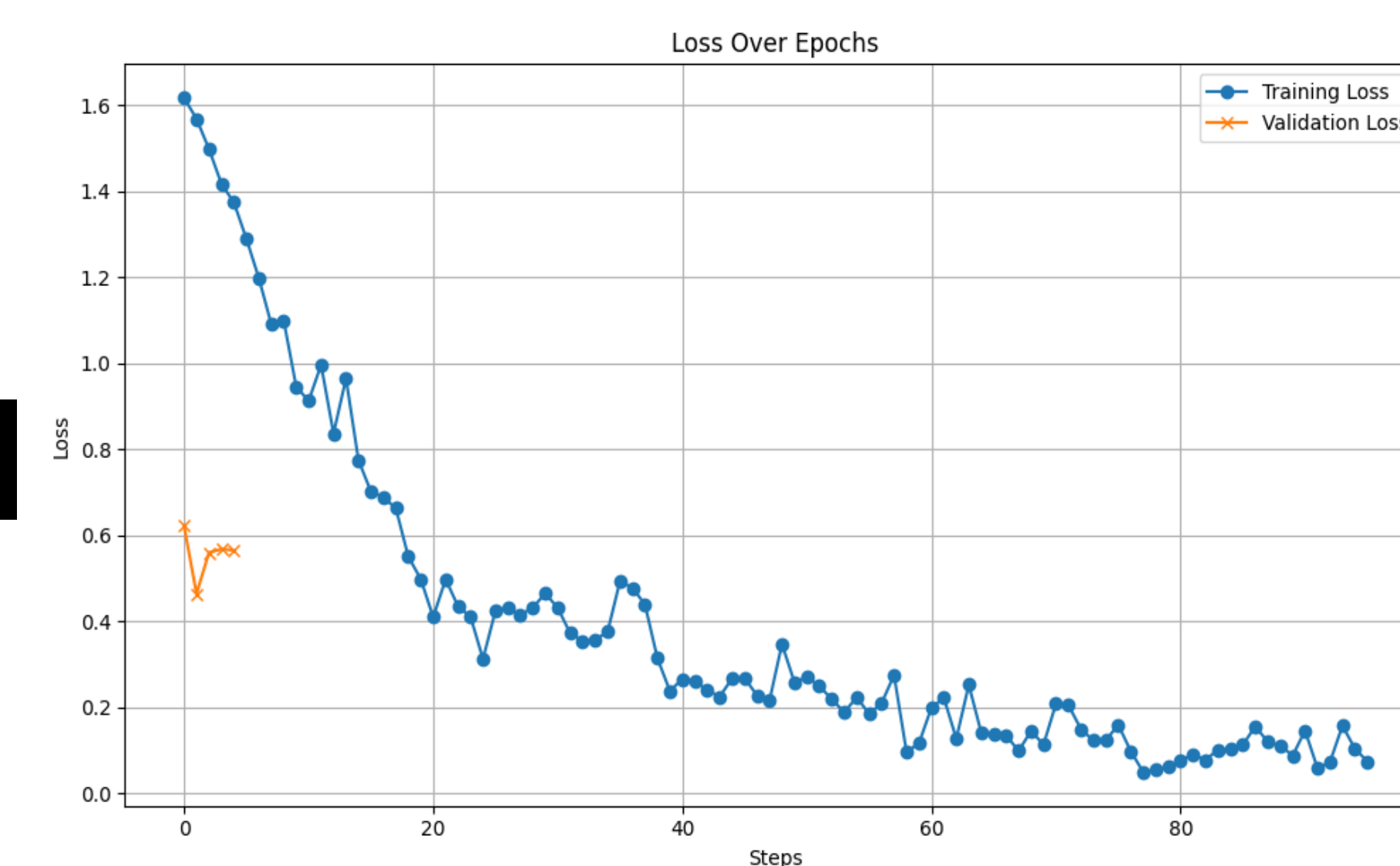
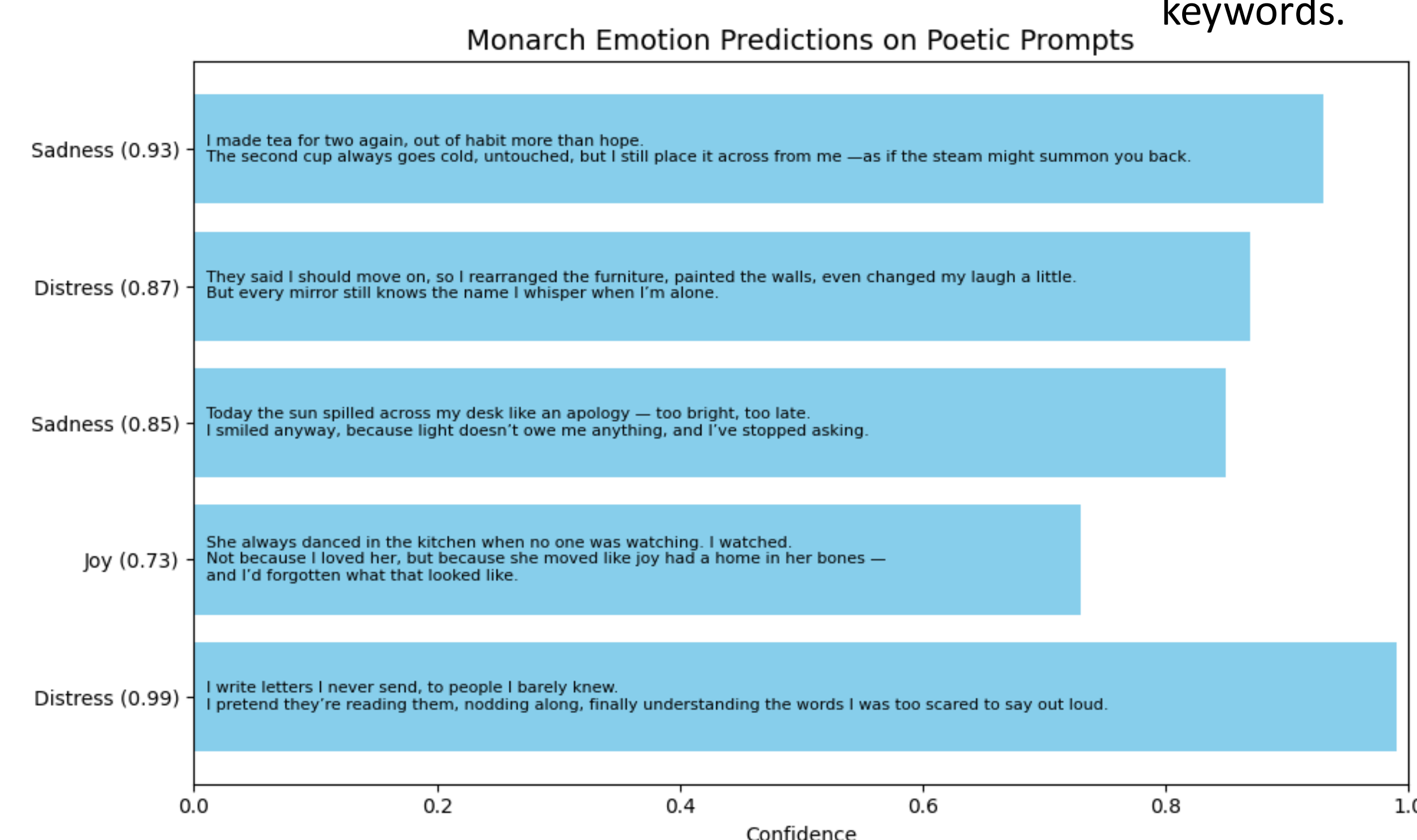Evaluation Metrics:
- Overall Accuracy: **87%**

Fig. 1 Model Pipeline

Fig. 2 Training loss graph
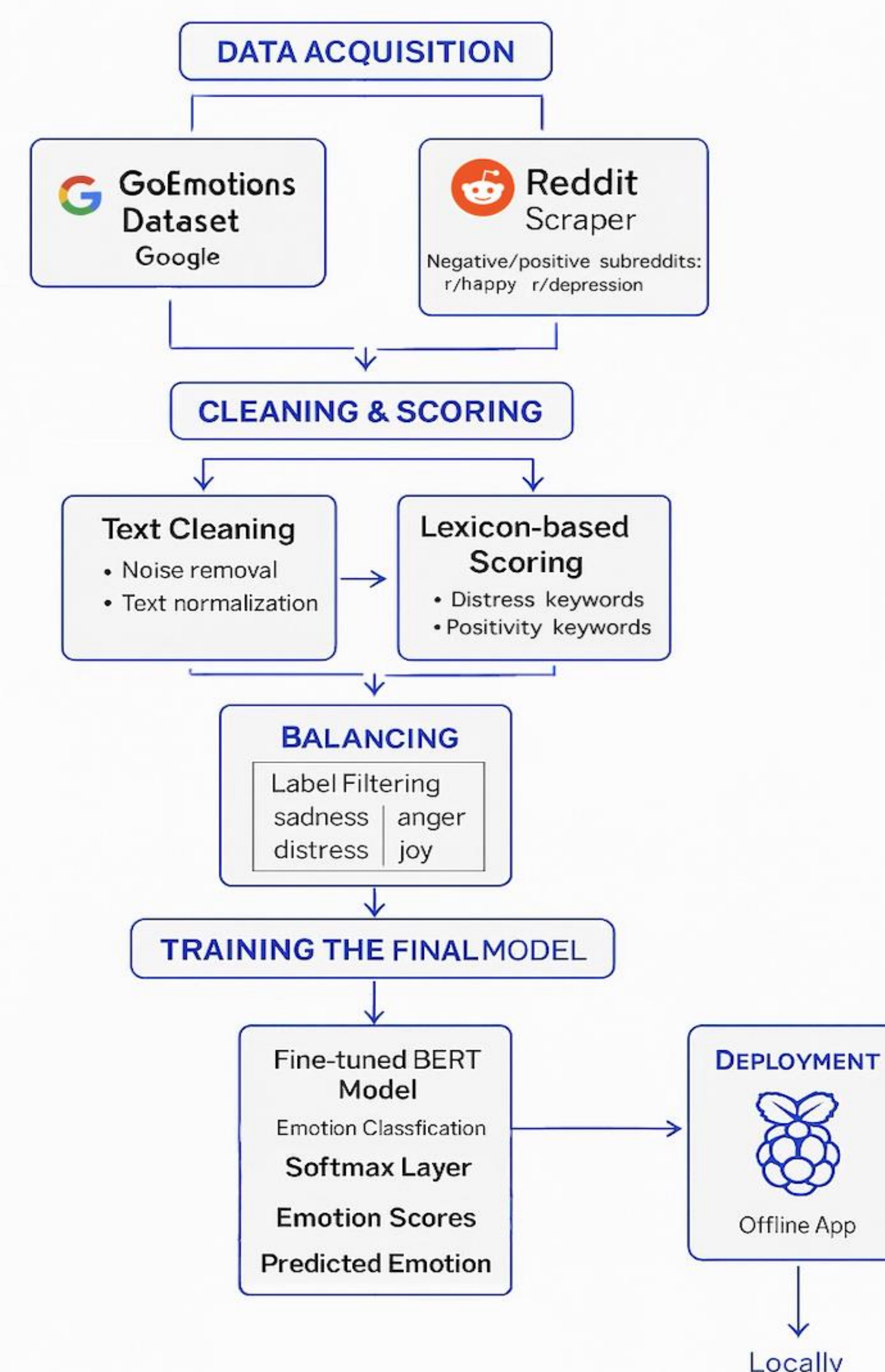
Emotion Recognition on Poetic Language:
- To test Monarch's robustness, we evaluated five poetic statements designed to obscure emotion through metaphor and ambiguity.
- Despite no poetic data was provided during training, Monarch still achieved a surprising **100%** accuracy, correctly identifying the dominant emotion in all five.
- Confidence ranged from **73%** to **99%**, even in the absence of explicit emotion keywords.

See **Figure 3** for full prompts and predictions.

These results suggest Monarch can infer emotional tone from subtle or figurative cues — not just direct sentiment — making it useful for analyzing expressive or reflective text.

Fig. 3: Model Predictions on Poetic Prompts (Top-1 Accuracy: 100%)

Conclusions

Monarch demonstrates that a fine-tuned NLP model can detect emotional cues like sadness, distress, worry, anger, and joy from cleaned, lexicon-applied Reddit data. Distress patterns frequently involve exhaustion, hopelessness, and disconnection. Balanced training improved accuracy and reduced confusion between overlapping emotions.

The model achieved high-confidence predictions and generalized well, even on poetic or ambiguous input — verifying that emotional signals in text can be interpreted offline, without relying on cloud services. Monarch supports the development of privacy-first NLP tools for emotional insight in education, research, and mental health contexts.
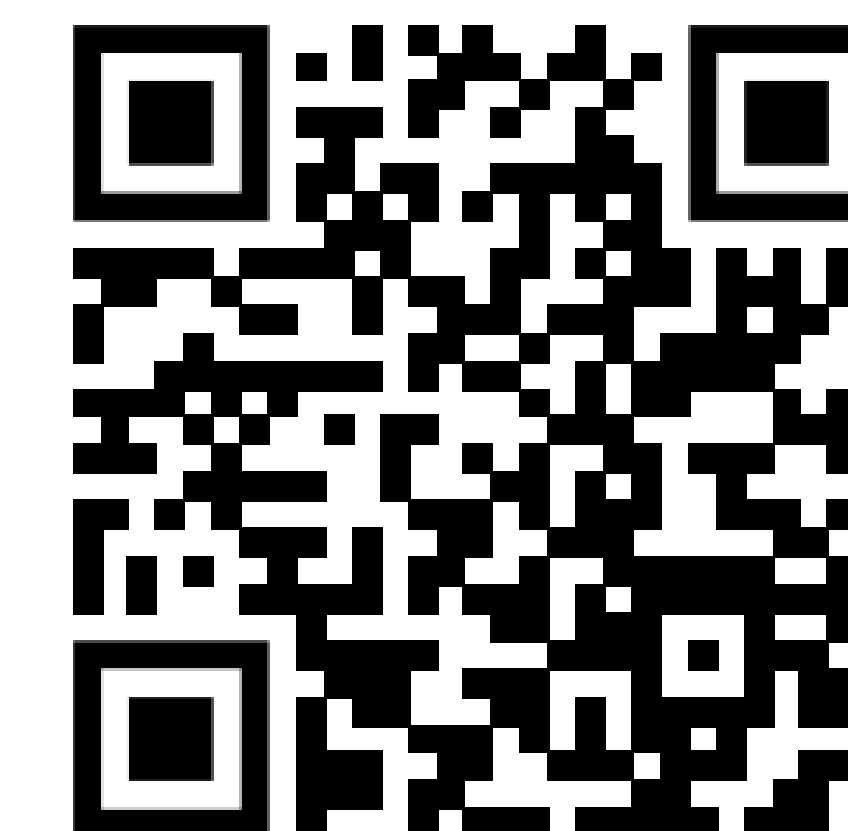
- We created this app not as a final product — but as a showcase of our research and the model's real-world potential. Whether embedded in journaling apps, used by school counselors, or powering offline analysis — Monarch shows that responsible emotional AI is possible.
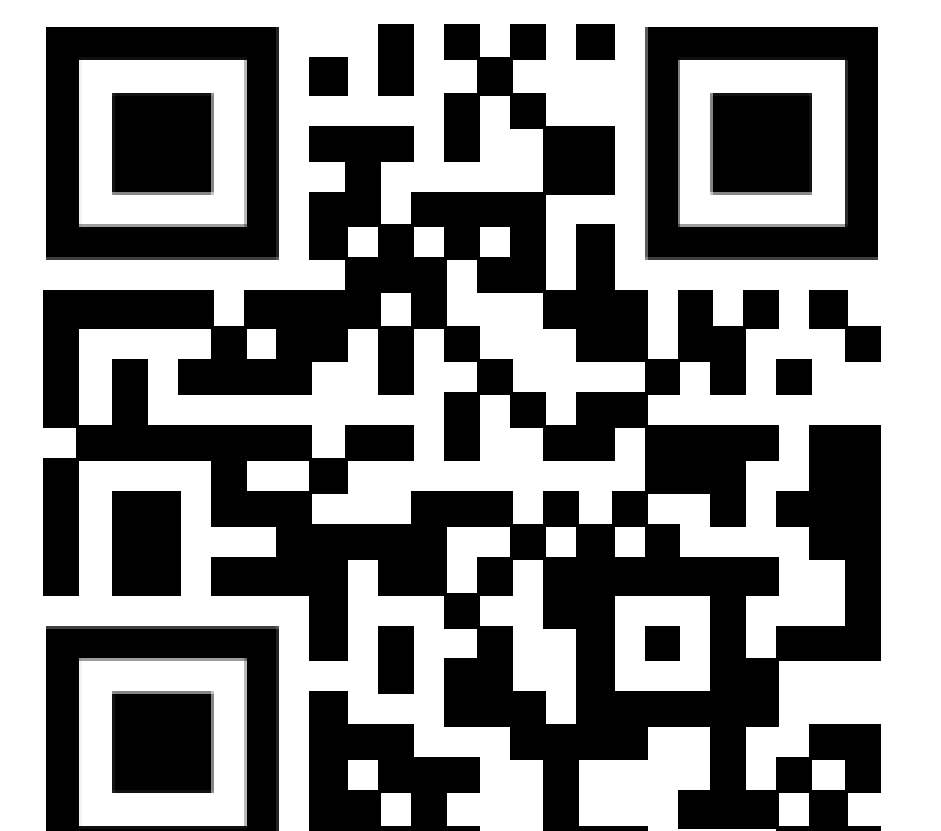

Acknowledgments

Contact & Project Links

Github

Monarch Download link

Contact information	Tyler Clanton	Derick Burbano-Ramon
LinkedIn	in/tylerclanton	In/derick-burbano-ramon
Email	tclanto2@students.kennesaw.edu	dburbano@students.kennesaw.edu

References

Alhuzali, H., Abdul-Mageed, M., & Elmadany, A. (2023). *Linguistic markers of depression and anxiety across the lifespan. Heliyon, 9*(11), e21435. https://doi.org/10.1016/j.heliyon.2023.e21435

Al-Mosaiwi, M., & Johnstone, T. (2018). *Linguistic markers of absolutism as predictors of depression and suicidal ideation. Clinical Psychological Science, 6*(4), 529–542. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10799169/

Mohammad, S. M. (n.d.). *NRC Emotion Lexicon*. https://saifmohammad.com/WebPages/lexicons.html

Mohammad, S. M. (n.d.). *WorryWords Dataset*. https://www.saifmohammad.com/worrywords.html

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Authors: Tyler Clanton, Derick Burbano-Ramon
Advisors: Jeff Adkisson