Investigating Explainable Methods for LLMs

Case Study: Emotionally Aware Chatbot

Presented by:

Domenico Pazienza, Leonardo Ercolani

Date: 12/12/2024

What is the project about?

Goal

Analyze explainable AI techniques to interpret the decisions made by the emotion detection model.

Key Features:

Analysis and Implementation of explainable methods

Fine-tuned conversational models for emotion detection

Why this project?

Existing LLMs often lack transparency

Regulatory Compliance:
Adhering with GDPR and European regulatory frameworks (AI Act)

Bridging the knowledge gap between human/computer interaction

Project Structure

1. Literature Review and Comparative Analysis of Explainable Methods for LLMs

2. Implementation of an Emotion Recognition Model

3. Model Explanation

4. Fine-Tuning and Evaluation

System Architecture

Input Handling:

Text input: Direct processing.

Voice input: Speech-to-text conversion

Emotion Analysis:

Emotion classifier pipeline (BERT)

Response Generation:

Generate an empathetic response based on the detected emotion

Output Delivery:

Response text returned to the user.

Datasets

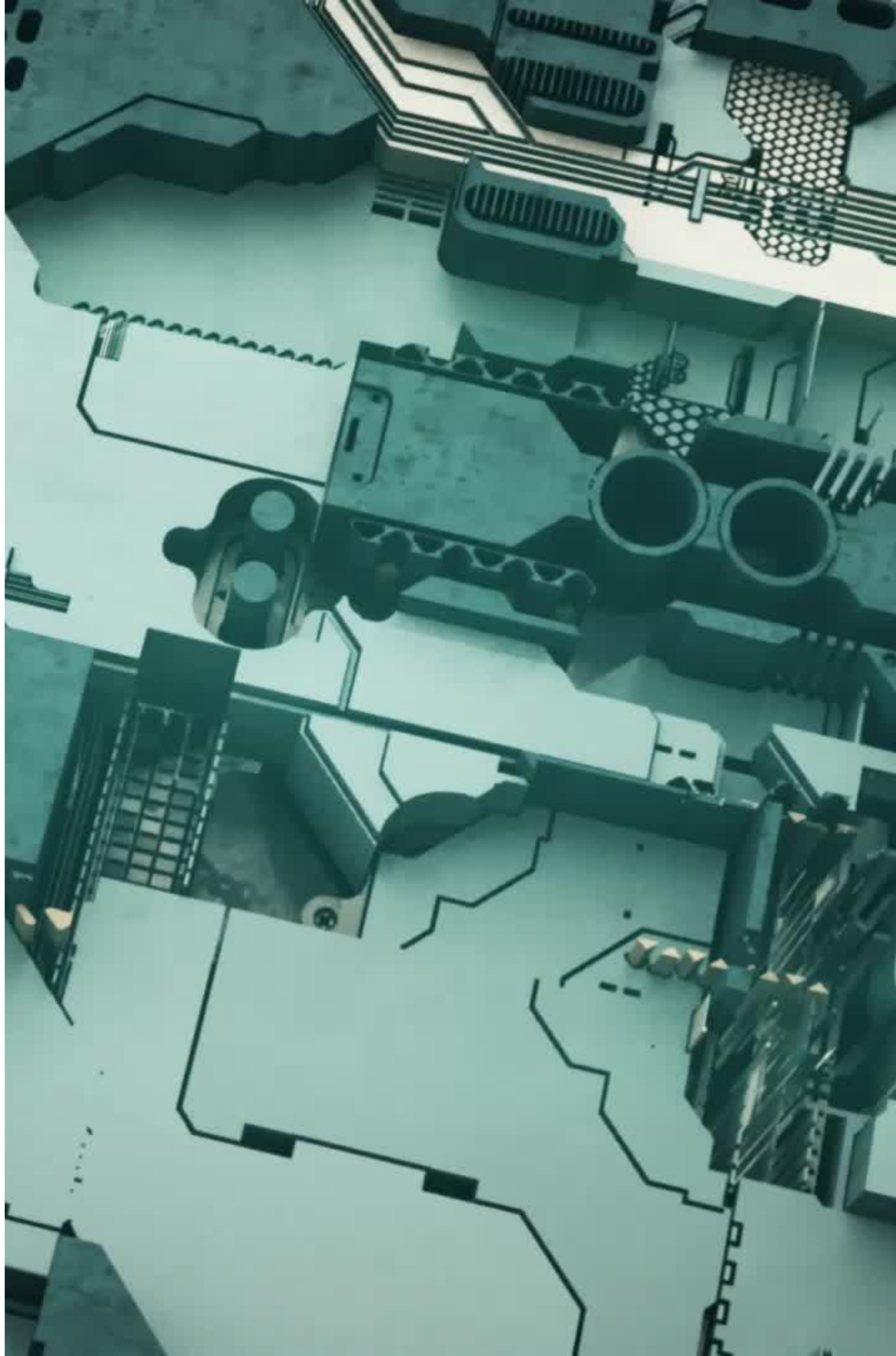
1. GoEmotions:
   1. Annotated with 27 emotions + neutral.
   2. Source: Curated from Reddit comments.
   3. Statistics:
      1. 58,000 annotated examples.
      2. Balanced across multiple emotion categories.
2. DailyDialog:
   1. Multi-turn conversational dataset.
   2. Annotated for emotions and dialogue acts.
   3. Statistics:
      1. 13,000 multi-turn dialogues.
      2. Includes annotations for intent and sentiment.

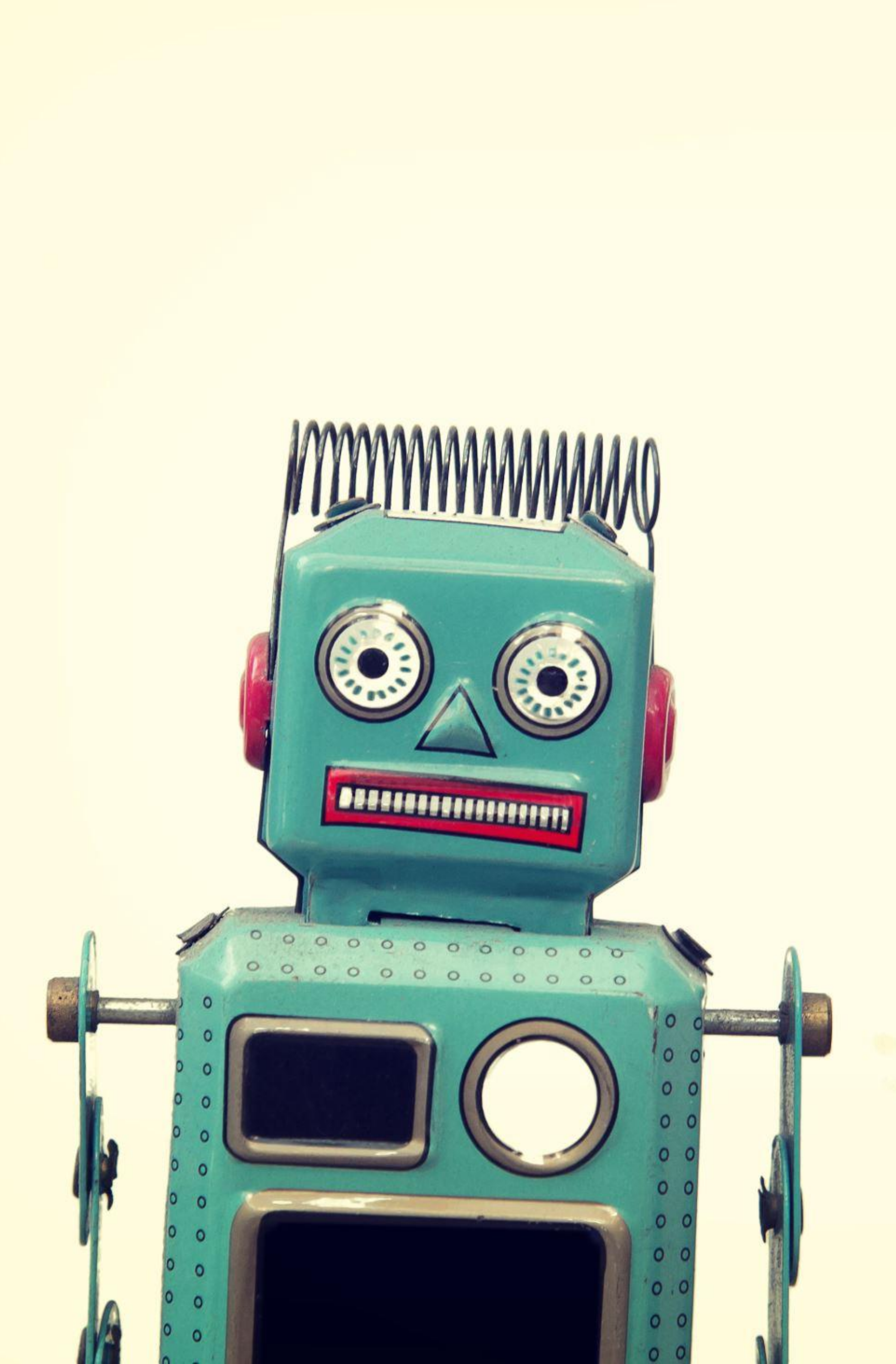
Model Choices

1. Emotion Detection:
    1. **Model**: BERT fine-tuned on GoEmotions.
    2. **Output**: Probabilities for each emotion.
2. Response Generation:
    1. **Model**: using DialoGPT or BlenderBot Fine-Tuning on DailyDialog.

Fine-Tuning Process

- Dataset: DailyDialog (context-response pairs).
- Steps:
  1. Tokenize and preprocess conversations.
  2. Train using transformers library.
  3. Evaluate and save the fine-tuned model.





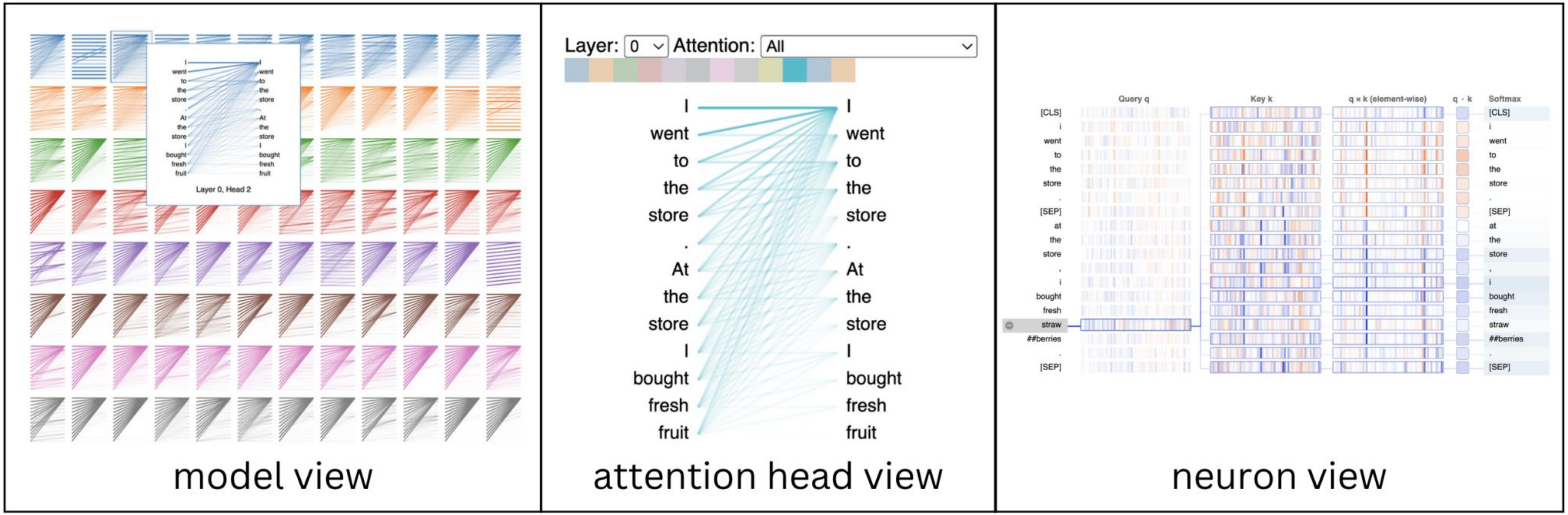
Live Demonstration

- Input examples:
    - "I passed my exam!"
        - Detected Emotion: Joy (95%).
        - Bot Response: "Congratulations! You worked hard for this!"
    - "I feel so sad today."
        - Detected Emotion: Sadness (85%).
        - Bot Response: "I'm sorry to hear that. Want to talk about it?"

Explanations

- Feature-Attribution Based Explanations
- Attention-Based Explanations
- Example-Based Explanations

BERTViz



model view

attention head view

neuron view

https://www.comet.com/site/blog/explainable-ai-for-transformers/

Key Challenges

Balancing response diversity and coherence.

Selecting the right method

Managing computational requirements for fine-tuning.

Interpreting ambiguous results

Thank you for your attention

Questions?

References:

BertViz: https://www.comet.com/site/blog/explainable-ai-for-transformers/

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., ... & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, *15*(2), 1-38.