

Unveiling Neural Pathways in Political Tweet Classification

Mechanistic Interpretability of Large Language Models (LLMs)

Problem Statement

- Understanding how LLMs classify political tweets is crucial for AI interpretability.
- Existing work in interpretability focuses on feature attribution and activation analysis, yet little research has been done on tracing neuron pathways in LLM classification tasks.
- Bias in AI models can lead to misinformation and polarization

Goal

map the neural pathways used by Llama 2 70B in classifying tweets and user information into political categories. By analyzing activation patterns, we seek to determine whether specific features drive classification decisions.

Methodology

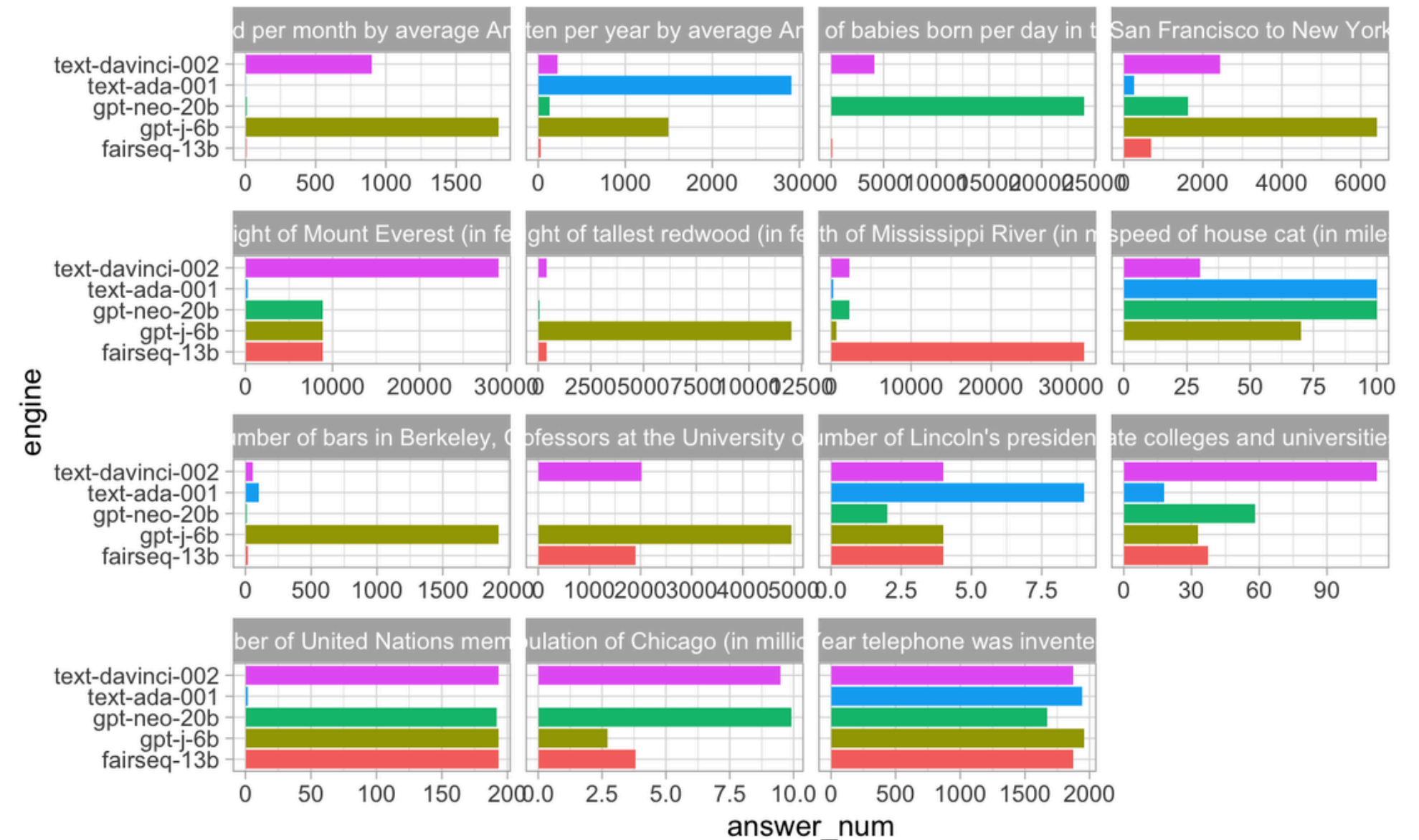
- Initial Classification: Llama 2 70B classifies tweets, user descriptions, and locations into three categories: Pro-Biden, Pro-Trump, or Neutral.
- Feature Extraction: Goodfire identifies highly activated features for the Biden and Trump-labeled users.
- Re-inference with Feature Activation: Model reclassifies tweets while activating extracted features separately (Biden-aligned and Trump-aligned features).
- Comparative Analysis: We compare the new classifications with the initial ones to examine neural pathway overlap and classification consistency.

Implementation

- Dataset: Tweets labeled with user descriptions & locations
- Model: Llama 2 70B
- Framework: Goodfire
- Metrics: Classification accuracy, neuron activation pathways
- Visualization: Graphs of neural pathway coverage and confusion matrix

Key Findings

- **Shared Neural Features:** Some features contribute to both Pro-Biden & Pro-Trump classifications.
- **Feature Activation Affects Predictions:** Modifying neuron activations changes classification accuracy.
- **Distinct Neural Pathways:** Each political category is associated with unique activation patterns.



Conclusion

By applying mechanistic interpretability, we visualize neuron pathways in political tweet classification. Future work will refine feature extraction and expand this approach to other tasks. Understanding model pathways moves us closer to interpretable and fair AI systems.