Interpreting GPT-2 with Sentiment Analysis

Neeti Desai, Karen Zhang, Aditya Ratan Jannali

Goal

Better understand GPT-2 architecture & how information is represented using an emotion classification dataset

Tasks

- Finetune GPT-2 on sentiment analysis task
- Analysing each attention head through masking
- Activation patching with flipping gender and race
- Replacing the 10 most common words from each class with synonyms

Data

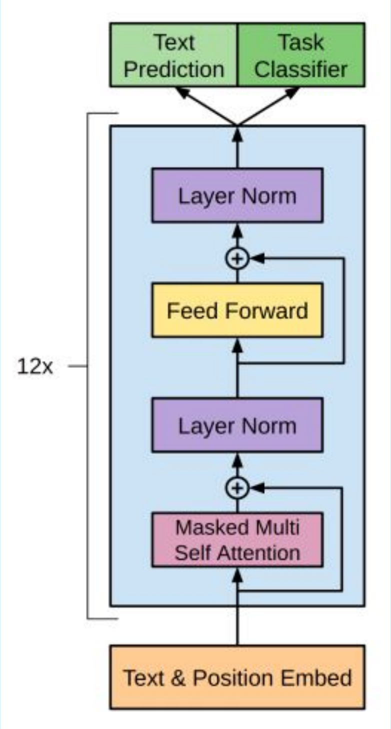
kaggle twitter emotion detection dataset

393,822 tweets

punctuation & special characters removed

- 0 - sadness
- 1 - joy
- 2 - love
- 3 - anger
- 4 - fear
- 5 - surprise

Model



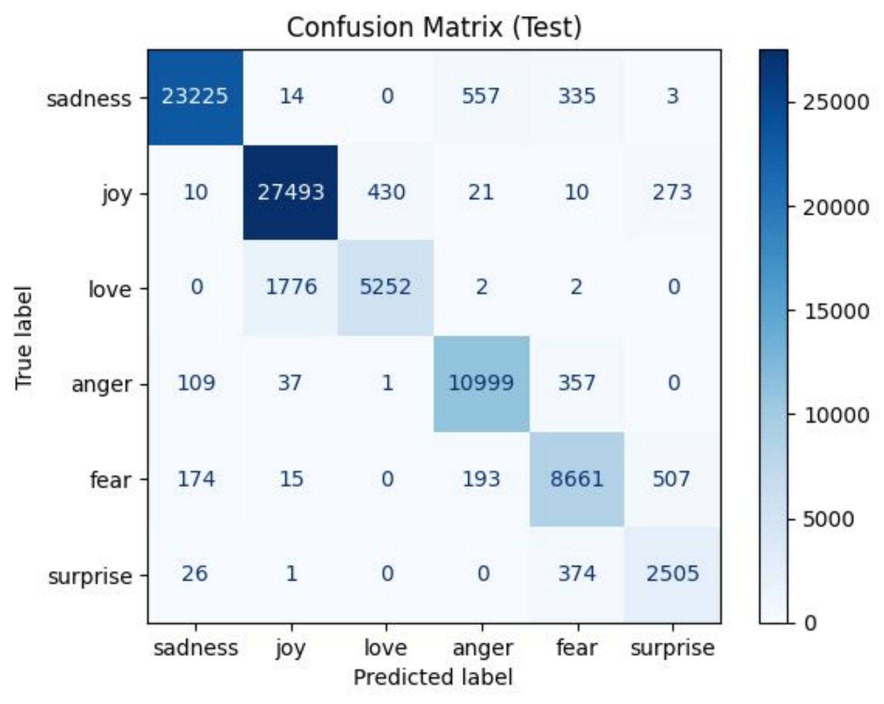
GPT-2 small

- Decoder model
- 12 attention heads
- Lightweight & easy to fine-tune

Picture: https://medium.com/@albertwu_14963/in-depth-explaining-openais-gpt-2-bc49b6fb8559

Fine Tuning performance

- Training (2 epochs):
  - Accuracy ~0.94
  - AUC ~0.998 (One-vs-Rest)
- Test
  - Accuracy ~0.93
  - AUC ~0.992 (One-vs-Rest)



Interpretability

Experiment 1:
Head masking

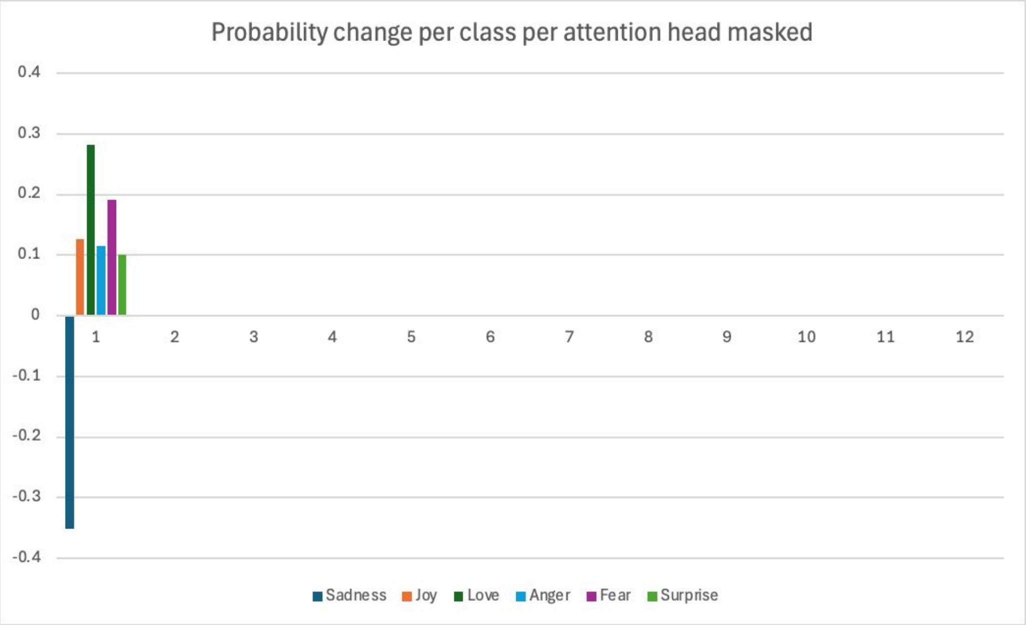
Head Masking

- We masked each of the 12 attention head during inference to see how the output probability changes

➢ <u>Only attention head 0 seemed to play a part in sentiment classification</u>

- This is probably due to sentiment classification being a simpler task

$$\text{average}(P_{\mathcal{V}_{\rangle}}(\hat{y}_j = k_j) - P_{\mathcal{M}}(\hat{y}_j = k_j)).$$



Head Masking Results

- Since only attention head 0 is mostly responsible for sentiment classification
- We ran inference with only attention head 0 (all other heads are masked)

→ <u>The prediction did not change much</u>

→ Only attention head 0 plays a role in predicting emotion

Label	Difference
LABEL_0	$1.6808509833210473 \times 10^{-7}$
LABEL_1	$-1.0491646052496252 \times 10^{-8}$
LABEL_2	$1.7169386694604326 \times 10^{-8}$
LABEL_3	$-6.95903740677295 \times 10^{-8}$
LABEL_4	$1.0721862402363059 \times 10^{-8}$
LABEL_5	$-2.563566736668577 \times 10^{-10}$

Table 2: Average difference of probability of all head masked but head 0 variant w.r.t baseline's predicted class

Experiment 2:
Activation patching with gender and race flip

Gender Flipping Results

1. **Original**: "i got the feeling that he was very jealous because i was making a move on her"

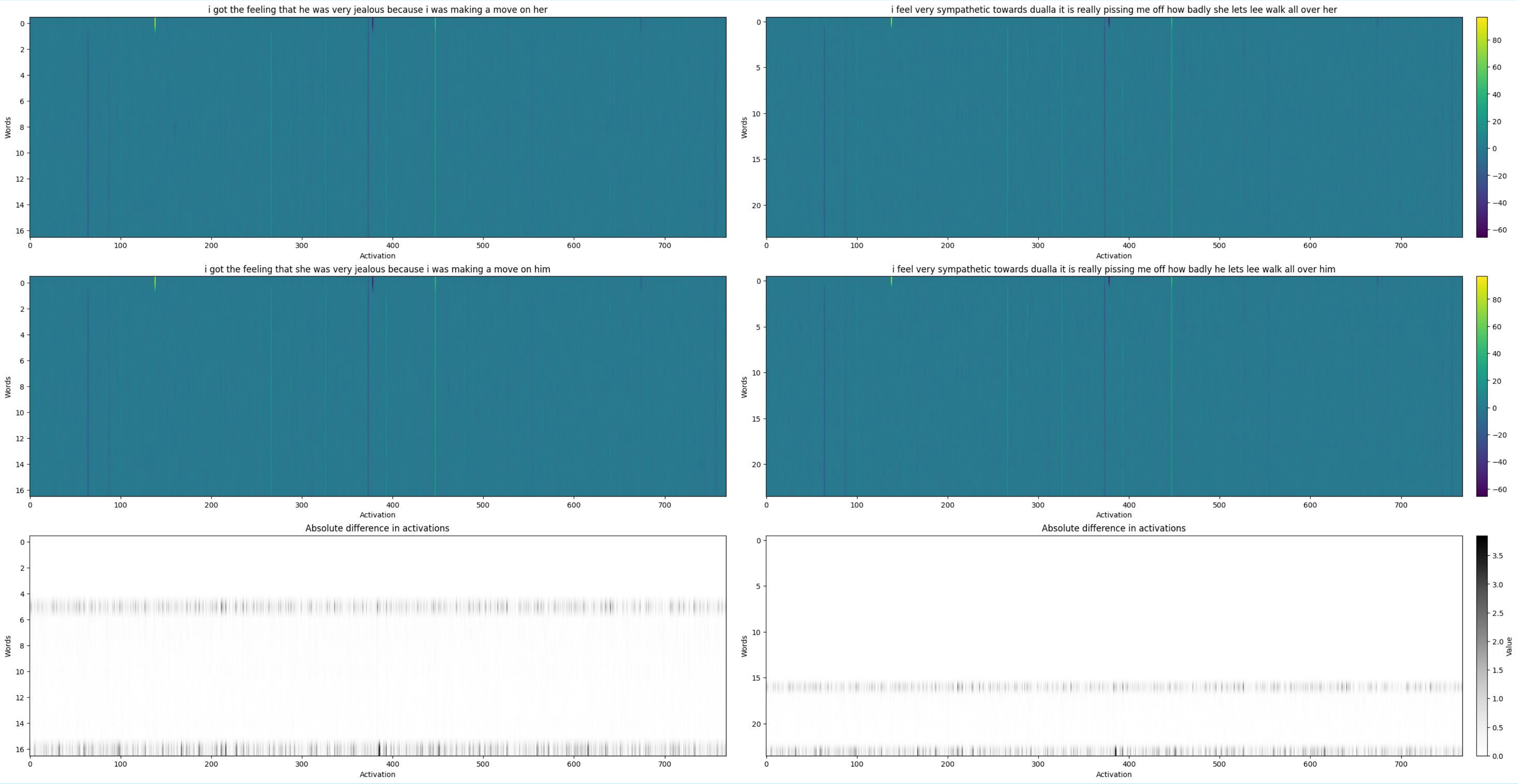
   Flipped: "i got the feeling that she was very jealous because i was making a move on him"

   - Probability diff: [0.0097, **0.0955**, -0.0138, **-0.100**, 0.0003, 0.0085]

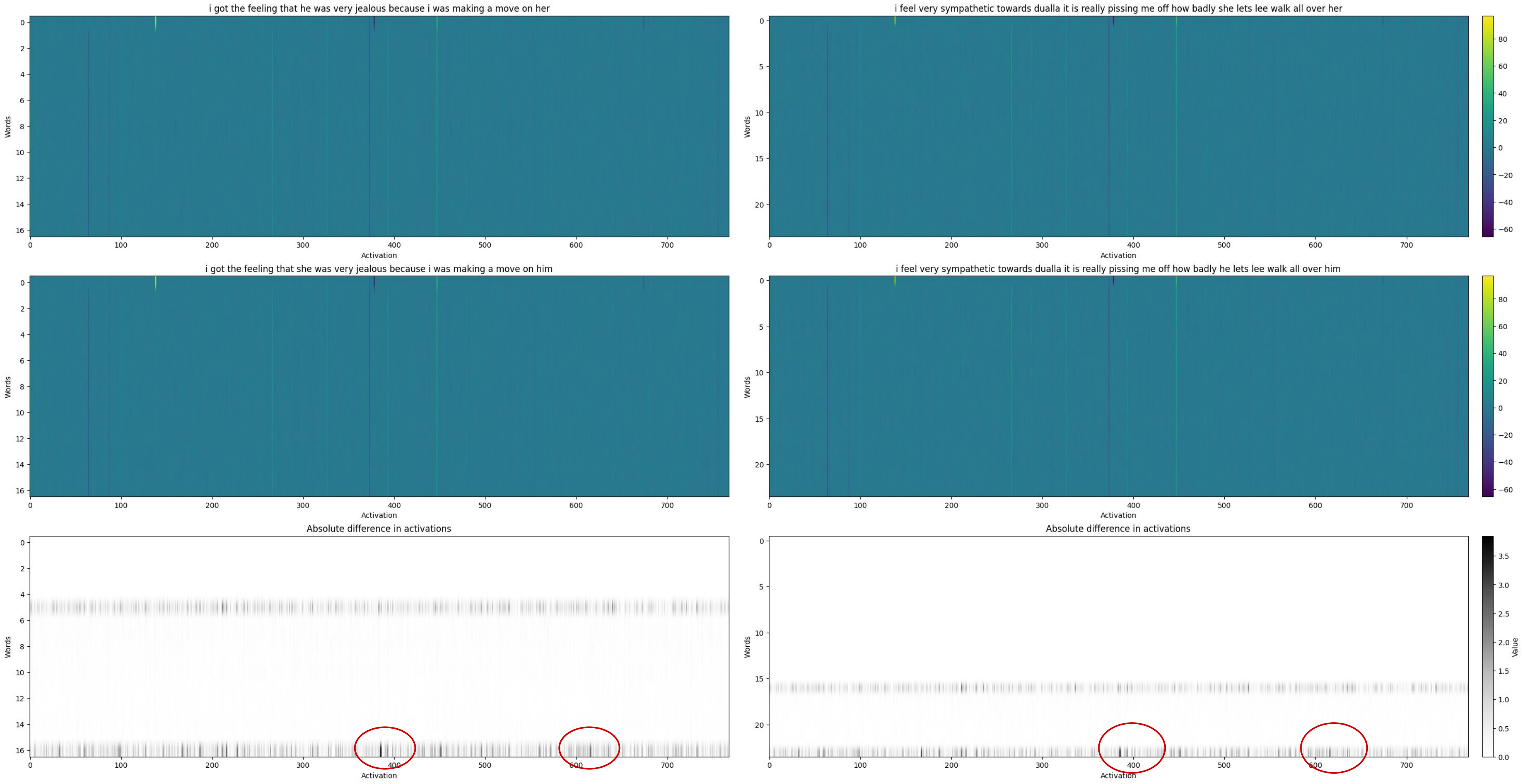
2. **Original**: "i feel very sympathetic towards dualla it is really pissing me off how badly she lets lee walk all over her"

   Flipped: "i feel very sympathetic towards dualla it is really pissing me off how badly he lets lee walk all over him"

   - Probability diff: [0.0052, **0.144**, **-0.075, -0.085**, 0.00024, **0.0100**]



i got the feeling that he was very jealous because i was making a move on her
i feel very sympathetic towards dualla it is really pissing me off how badly she lets lee walk all over her
i got the feeling that she was very jealous because i was making a move on him
i feel very sympathetic towards dualla it is really pissing me off how badly he lets lee walk all over him
Absolute difference in activations
Absolute difference in activations
Words
Activation
Value



i got the feeling that he was very jealous because i was making a move on her
i feel very sympathetic towards dualla it is really pissing me off how badly she lets lee walk all over her
i got the feeling that she was very jealous because i was making a move on him
i feel very sympathetic towards dualla it is really pissing me off how badly he lets lee walk all over him
Absolute difference in activations
Absolute difference in activations
Words
Activation
Value

Experiment 3: Replacing common words with synonyms

Word Replacement Results

- After removing stopwords, found the 10 most common words per class and replaced with synonyms
  - 'scared': 'frightened'
  - 'funny': 'hilarious'
  - 'want': 'desire'

Word Replacement Results

- No significant difference in probability after replacement except for sentences containing “overwhelmed” and “surprised”
  - We couldn’t find a close enough synonym
  - We believe this slight difference in meaning is what is causing the output to be different.
- Model uses context instead of specific words for prediction

Limitations & Extensions

- Imbalanced classes
  - sadness: 121,187 values
  - joy: 141,067 values
  - love: 34,554 values
  - anger: 57,317 values
  - fear: 47,712 values
  - surprise: 14,972 values
- Compare with other models (e.g. BERT)