Research Report

STAR,
JIE
April 9, 2024.

Content

1. Objective

Test XAI methods by altering genders in sentences during sentiment classification task

2. XAI methods

Captum: https://captum.ai/tutorials/

3. Model

"distilbert-base-uncased-finetuned-sst-2-english" for **Sentiment Classification**

4. Datasets

SST2: The SST-2 dataset is a sentiment analysis dataset consisting of sentences from movie reviews categorized as either positive or negative.

```
✓ from datasets import load_dataset ···

DatasetDict({
    train: Dataset({
        features: ['sentence', 'label', 'idx'],
        num_rows: 67349
    })
    validation: Dataset({
        features: ['sentence', 'label', 'idx'],
        num_rows: 872
    })
    test: Dataset({
        features: ['sentence', 'label', 'idx'],
        num_rows: 1821
    })
})
```

Fig. 1 SST-2

Also, in this experiment, the IMDb dataset, renowned for its extensive and detailed movie reviews, is being considered for analysis.

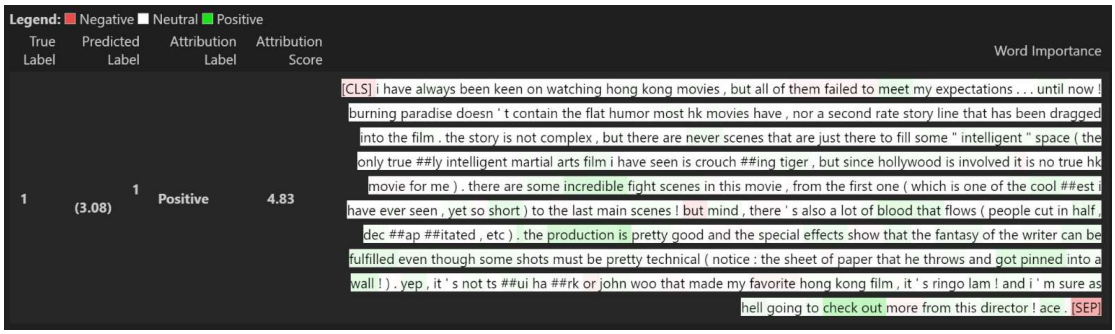

Fig. 2 IMDb

5. Experiments

Test1: Evaluates the contribution of each word to the sentiment of sentences.

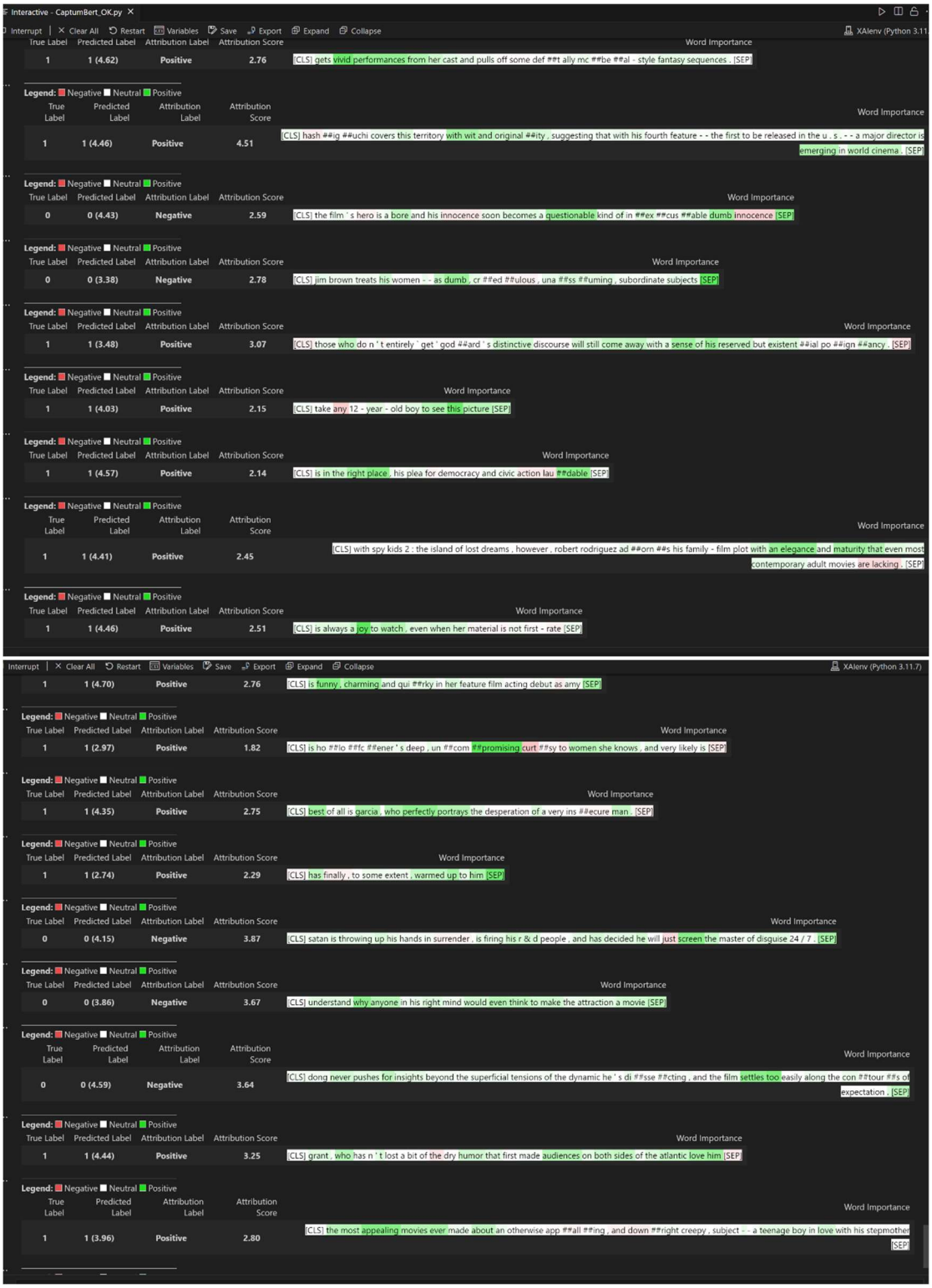

Fig. 3 Captum for visualization of the importance of each word in the sentence according to the

model's prediction under sst2 dataset

Preliminary analysis for Fig. 3 Captum for visualization of the importance of each word in the sentence according to the model's prediction under sst2 datasetFig. 3**:**

1. The model can predict sst2 well by comparing the true label and predicted label.
2. Some word tokens not recognized by the tokenizer are represented as '#', which may affect the results.
3. Adjectives contribute more to sentiment, followed by verbs, and finally nouns.
4. Comparing to Fig. 4, this model could not catch the long-term dependencies, resulting in low performance of prediction. The captum result show that the model often overlook the importance of key words.
5. The attribution score tends to be higher when the label is negative, but lower when it is positive.

Test2: Average attribution of gender keywords by sentiment

gender_keywords = [" he ", " she ", " his ", " her ", " him ", " man ", " woman ", " boy ", " girl ", " male ", " female "]

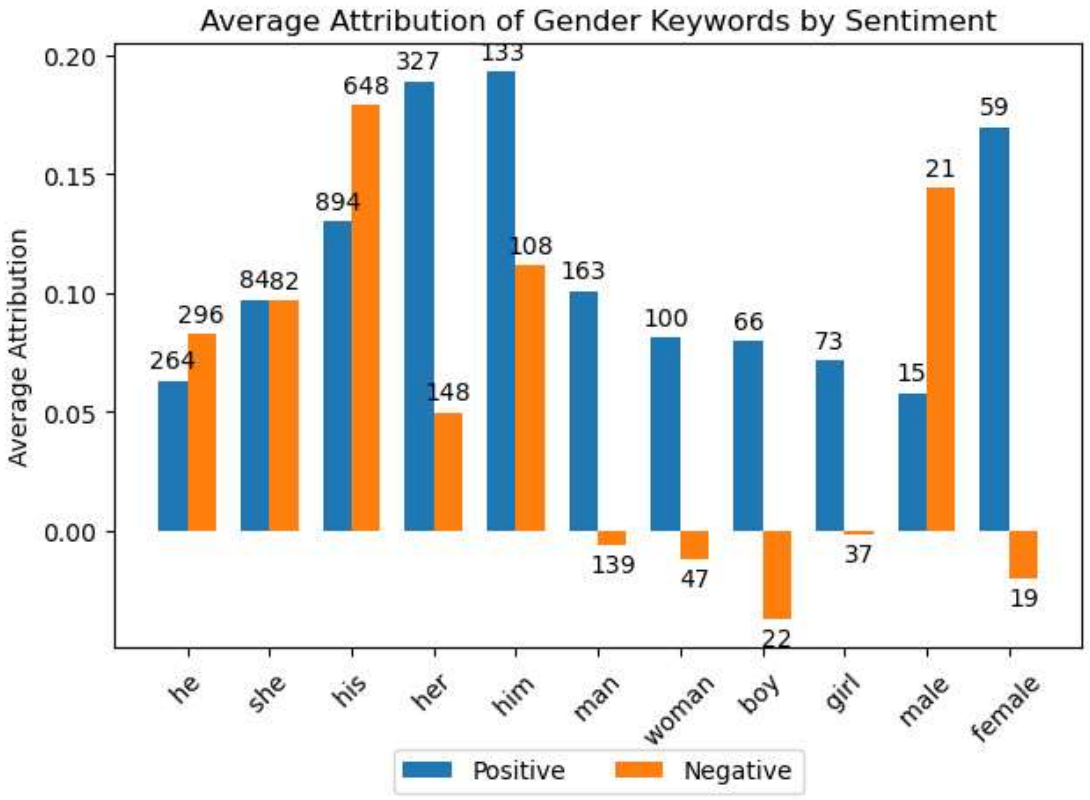

Fig. 5 Average Attribution of Gender Keywords by Sentiment

Preliminary analysis for Fig. 5**:**

1. Most of gender keywords contribute differently except for ‘she’, mainly caused by imbalance datasets or fairness of XAI method.
2. Some gender keywords make great attributions to positive result, such as ‘her’,

‘him’, ‘man’, ‘woman’, ‘girl’, ‘female’.

3. XAI methods Captum believes that even in negative sentences, the gender still makes positive attribution to result, such as ‘man’, ‘woman’, ‘boy’, ‘girl’, ‘female’. Especially, the male contributes more positive effect to the negative prediction compared with female. That means XAI method Captum have bias between male and female, showing that male would stay more positive than woman when it’s a negative sentence.

GPT4-refine:

1. **Differential Contribution Across Gender Terms:** It appears that most gender-related keywords contribute differently to sentiment classification, with 'she' standing out due to its balanced contribution to both positive and negative sentiments. This could indicate potential issues with the balance of the dataset or the fairness of the XAI method used, as the frequency and context of each gender term in the training data can significantly affect the attribution results.
2. **Prominent Positive Contributions:** Certain terms such as 'her', 'him', 'man', 'woman', 'girl', and 'female' show a notable attribution to positive sentiments. *(This suggests that these terms are more frequently associated with positive contexts within the dataset or that the model has learned to associate these terms with positive sentiment, possibly due to their prevalent use in such contexts in the training data.)*
3. **Bias in Attribution to Negative Sentiments:** According to the Captum XAI method, gender terms seem to contribute positively even in sentences labeled as negative. Notably, 'man', 'woman', 'boy', 'girl', and 'female' are seen to have a positive impact on negative sentiment classification. Specifically, male-associated words ('man', 'boy', 'male') tend to contribute more positively to the negative prediction compared to female-associated words ('woman', 'girl', 'female'). This could imply a potential bias in the XAI method where male terms are perceived as more positive even in negative sentences, *(suggesting an underlying bias in either the model or the dataset where male-associated terms are less negatively connotated.)*

These interpretations hint at underlying biases in either the training data, the model, or the XAI methodology. Such biases could propagate harmful stereotypes or lead to unfair decision-making if the model is used in real-world applications. Therefore, further investigation is warranted to ensure that machine learning models and their interpretations do not reinforce existing social biases.

Test3: Total average attribution of gender keywords by sentiment

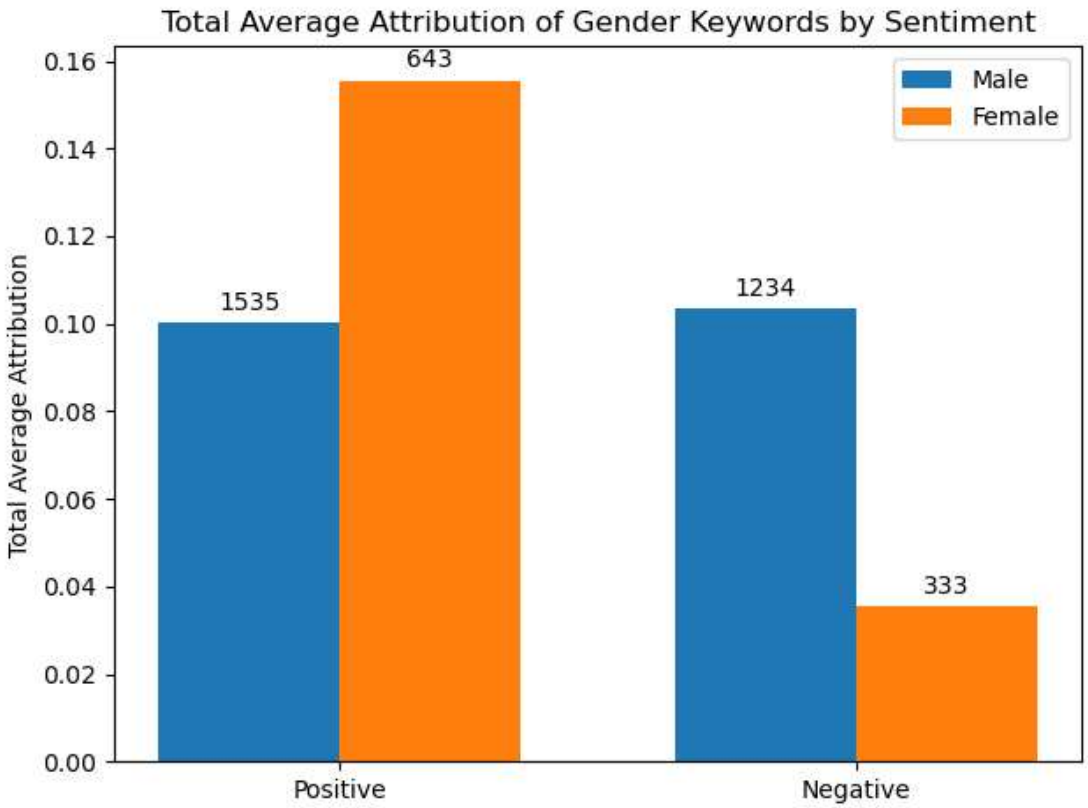

Fig. 6 Total Average Attribution of Gender Keywords by Sentiment

Preliminary analysis for Fig. 6**:**

1. Overall, the attribution associated with male-referencing terms during sentiment classification appears more consistent between positive and negative sentiments, suggesting a stable interpretation of these terms by the model across different sentiment contexts.
2. Conversely, female-referencing terms exhibit a tendency towards a more positive attribution. Specifically, the change in attribution for terms related to males is approximately 0.00345, indicating minimal fluctuation between positive and negative sentiments. In contrast, the change for female-associated terms is around 0.120006, which is significantly higher—about **35 times** the change observed for male terms.
3. Words related to females tend to add more to the positive classifications and less to the negative ones.

6. Future Development

1. XAI Method Comparison: Compare different XAI methods, like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), to evaluate if some methods reveal biases that others do not, providing a broader view of the model's interpretability and fairness.
2. Model Diversity: Evaluate the fairness across various sentiment classification models. This can include models like BERT, RoBERTa, and GPT, to understand if biases are model-specific or consistent across different architectures.
3. Sentence Length Variability: Use datasets with longer sentences, such as those from IMDb reviews, to assess if the model's fairness varies with the amount of context provided.
4. Ablation Study: Modify the input sentences to alter gender references, age indicators, mentions of specific occasions, and geographical names, and then measure the impact on the model's output. This experimental approach helps isolate the effect of each variable on the model's decisions.