Exploring Classifying Features for a RoBERTa-based Classifier through Visualizing the Attention and the Attribution Scores for each Token

{CSC3160} {Default} Project

Cheng Yang (Student ID, 120090195)
Fintech, Chinese University of Hong Kong, Shenzhen
120090195@link.cuhk.edu.cn

1 Introduction

- **Problem**: This project aims to figure out what features the RoBERTa-based classifier (1) captures to detect the ChatGPT-generated text.
- **Plan**: This project plans to explore the classifying features through visualizing the attention (2) and the attribution scores. The dataset is sourced from HC3-English with 80% for training, 20% for test and 10% for validation. The classifier is based on RoBERTa, referring to the baseline model of ChatGPT-Detector. I will investigate the relationship between the distribution of attribution scores and various variables (e.g. part of speech, token's frequency) to check the effect of a single token and visualize the attention to evaluate the relationship among tokens.

2 Milestone

- **Progress**: I have managed to train and save the classifier model. And I have visualized the attribution scores of each token for a test text (figure 1).
- **Challenges**: The cost of time for trying different visualization functions by using different metrics while training model is huge. Thus, I will first find the most appropriate method with a small amount of training sample and then corporate it with the whole training set.
- **Source Code**:
  Github repository: `https://github.com/SLPcourse/Final-Project/blob/main/Final_Project_RoBERTa_Visualization.ipynb`
  Colab link: `https://colab.research.google.com/drive/1-H4bj--vvcaZGc0_C9ml3Xr7nV7XEUv8?usp=sharing`
- **Paper Reading:** The current studies (3) conclude that human-written texts have a wider vocabulary but shorter length than ChatGPT, and ChatGPT uses more attributives, conjunctions and auxiliaries while keeping a neutral emotion.

References

[1] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, "How close is chatgpt to human experts? comparison corpus, evaluation, and detection," *arXiv preprint arXiv:2301.07597*, 2023.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[3] R. Tang, Y.-N. Chuang, and X. Hu, "The science of detecting llm-generated texts," *arXiv preprint arXiv:2303.07205*, 2023.

A Appendix (optional)

A.1 Team contribution percentage

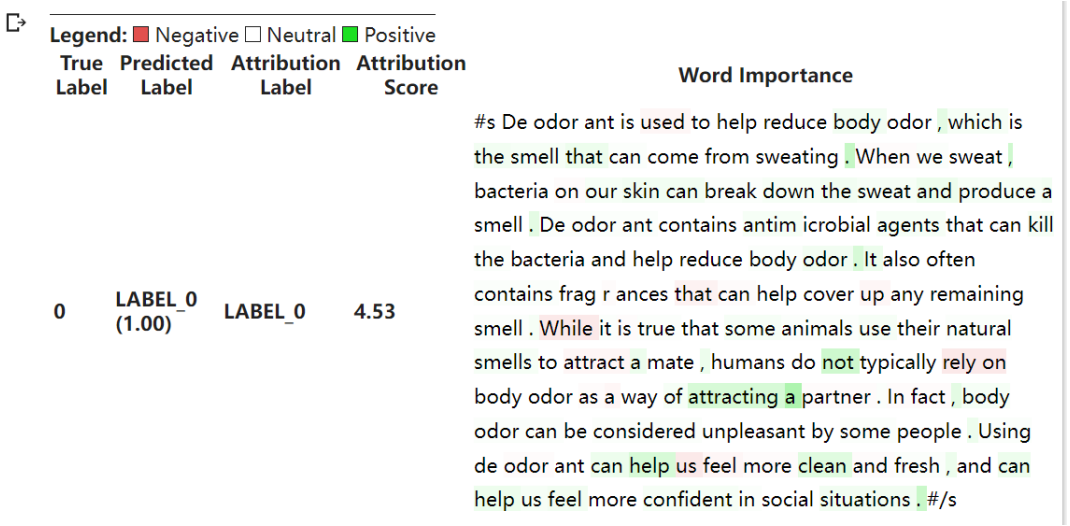
Not applicable

A.2 Related Work

The two main current methods for detecting ChatGPT-generated text are the black-box method and the white-box method. For external identity, black-box is preferred since white-box method needs high-level access permission for the language generating model. 'ChatGPT Detecor' applies a black-box method named RoBERTa to work on the task, an updated version of the BERT model. The transformers in the RoBERTa model involve some potential features that function well for detection. Some stuies point out that human-written texts have a wider vocabulary but shorter length than ChatGPT. And ChatGPT uses more attributives, conjunctions and auxiliaries while keeping a neutral emotion. We plan to visualize the attention and investigate the distribution of attribution scores to explore these features.

A.3 Workflow or/and Architecture

Figure 1: The visualization of attribution scores



The visualization of attribution scores adopt the package named 'transformers_interpret'. The attribution scores is calculated by integrating the gradients of the model's output with respect to the input tokens along a straight path from a baseline input to the actual input. To improve the credibility of the attribution scores, I will try other calculation methods such as perturbation-based methods to better reflect each token's contribution.

A.4 Experiment Design and Results

Experiment Step	**Tentative Results**
Train the RoBERTa-based classifier	Manage to train a classifier with accuracy rate greater than 99.5%.
Visualize the distribution of attribution	The absolute value of attribution scores of adjectives are higher than nouns.
Visualize the attention	
Show a detailed feature visualization for each input paragraph	
Conclude the findings regard to the classifying features	