Word-Level Interpretation of Chatgpt Detector Based on Classification Contribution

Dekun Chen *

School of Data Science, Chinese University of Hong Kong, Shenzhen, China, 518172

* Corresponding Author Email: 120090336@link.cuhk.edu.cn

Abstract. The ChatGPT detector is considered a necessary task to standardize the use of ChatGPT. Difficulty interpreting the test process and results is a common problem with LLM. Most existing interpreters focus on attention visualization and rarely consider the classification process. This study presents a method to show the contribution of words to model predictions. Specifically, this study considers information from classification weight vectors, semantic vectors, and embedded input word vectors for a more complete interpretation of detector LLM. Three word-level attributes (word length, part of speech and word meaning) are compared with the conclusions of existing literatures to verify our method. Visual samples and analysis process can be found at https://github.com/salixc/WCC-DekunChen.

Keywords: ChatGPT Detector, Word-Level Interpretation, Classification Contribution.

1. Introduction

ChatGPT has gained significant attention in recent years, leading to a surge in research on this topic. Although ChatGPT can help make decisions, it is more significant to clarify the responsibility, especially in health care and finance. ChatGPT detector, built for verifying whether AI or humans generate the text, was regarded as one of the essential tasks. So far, many different detectors have been developed with very high prediction accuracy on different data sets. However, most detectors are black boxes for users, and only assertion is usually insufficient. Therefore, the visualized result of interpretation is of great importance. It would help both detector users and experts to better understand the captured features and judgement reasons of the detector model. In this work, we focus on the contribution of words to the results of the detector classification process. Since most detectors were transformer-based models, much research concentrated on attention visualization. Bertviz [1], a multi-scale attention visualization tool, could be used to explore the degree of the model's attention to tokens. Moreover, the more powerful T3 -Vis [2] could help experts more comprehensively analyze transformer-based model. However, they only showed transformer in detail and did not take the classification process into account. So that it was difficult to display the focus on the two categories. In addition, there were other detection approaches based on statistics. GLTR [3] was a visual tool that helped humans detect generated text. But it was less effective than the large language models predicted.

To solve the above issues, we propose a visual tool to interpret the classification process by word-level contribution. As shown in Figure 1, weight vectors from the classification process were considered, together with the output semantic vector. By calculating the similarity with input embedding vectors, the word contribution to classification could be visualized. In our model, we take advantage of the information from both transformer and classification process. So that it has a more accurate interpretation of the integrated language model (Figure 1).

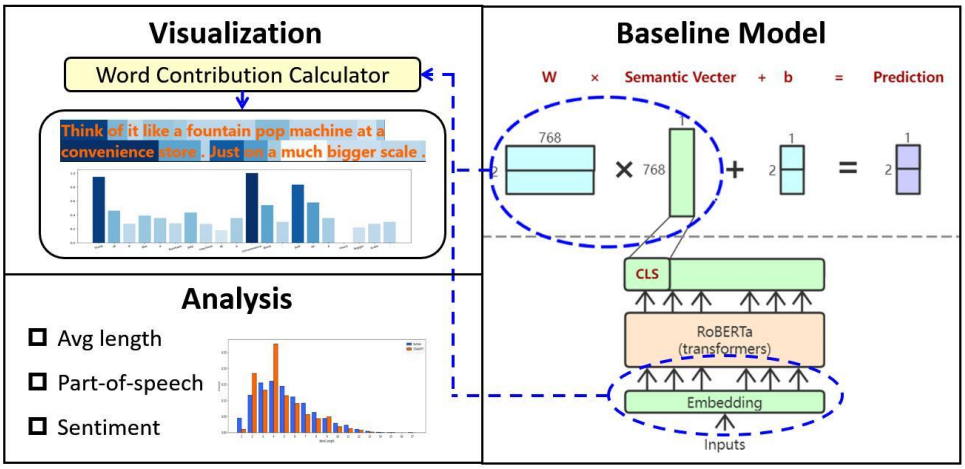

Figure 1. Based on the integrated RoBERTa classification model, the semantic vector, weight vectors, and embedded input are used in word contribution calculator to produce the visual results of word contributions behind the model prediction in two forms. Then, we apply our method on test set to discuss the correctness by statistical results from three perspectives.

2. Related Works

2.1. ChatGPT Detector

Many efforts have been made in the field of detectors. Various detection techniques include Zero-shot Detector based on probabilistic negative curvature [4], GPTZero based on perplexity [5], and LLM- based detectors [6,7]. Specifically, the baseline model in our work is Biyang Guo and his team's fine-tuned RoBERTa based detector [6], whose F1 score was up to 99%. The predicted label and its probability can be obtained by detecting a single text. Moreover, the research summarized four major differences between humans and ChatGPT, which will be helpful for the direction of our work.

2.2. Detector Visualization

Attention visualization has developed rapidly in recent years. BertViz was one of the early prominent works in this field [1]. It is very useful for analyzing the internal structure of BERT model by visualizing the multi-layer and multi-head attention mechanism. But it is less applicable to explaining the model's predictions. ExBERT was developed to help users identify the input words with the most impact on the predictions but was costly to calculate. T3 -Vis was a formal validation and interpretation toolkit to identify the causes of model prediction errors [8]. However, as mentioned above, the process of classification was not considered so that it is hard to illustrate the real extracted features of the predictions.

3. Approach

As shown in Figure 2, our work consists of two parts. The baseline model produces the probabilities of humans/ChatGPT for the input. And the word contribution calculator takes advantage of information from the baseline model to calculate the contribution scores of classifications for each input word.

3.1. Baseline Model

The baseline model is a fine-tuned RoBERTa classification model, which includes RoBERTa trans- formers and a fully connected layer classifier [9]. Transformers receive embedded input vectors,

and the semantic information is extracted into the [CLS] output vector [10], also called "Semantic vector", by attention mechanism. The classifier takes the semantic vector as the input. The weight vectors of the predictions of humans and ChatGPT are constantly updated by back-propagation.

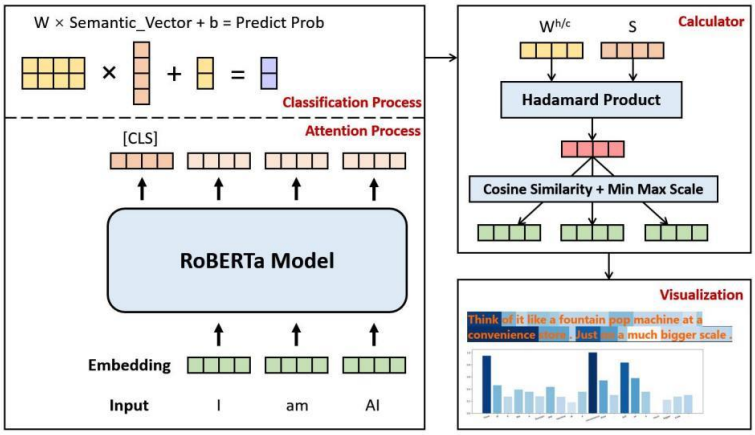

Figure 2. Detailed structure of baseline model and contribution calculator. We use Hadamard product to integrate the information of classification weight vectors and semantic vector. Then, we use cosine similarity to quantify the distance of token vectors.

3.2. Word Contribution Calculator

From the fully connected layer classifier, we obtain the semantic vector and weight vectors, representing by: (768 is the default dimension of embedding. And W^h for humans and W^c for ChatGPT).

$$S = [s_1 \quad s_2 \quad \dots \quad s_{768}] \tag{1}$$

$$W = \begin{bmatrix} w_1^h & w_2^h & \dots & w_{768}^h \\ w_1^c & w_2^c & \dots & w_{768}^c \end{bmatrix} \tag{2}$$

Take the Hadamard product of S with respect to W^h and W^c to get classification weighted vectors for humans and ChatGPT. Details can be found in Figure 3.

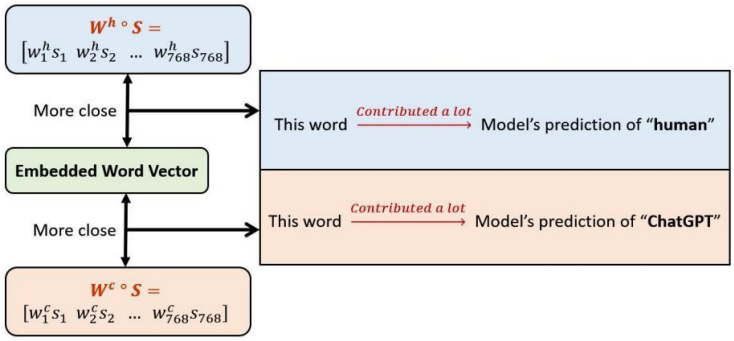

Figure 3. In the condition of the prediction "human", the closer the embedded input word vector and $W^h \circ S$ are, the higher the contribution of this word in the process of the prediction "human". Conversely, if the prediction is "ChatGPT", the closer the embedded input word vector and $W^c \circ S$ are, the higher the contribution of this word in the process of the prediction "ChatGPT".

Therefore, cosine similarity is calculated for each embedded vector and the classification weighted vector of the corresponding label to represent the classification contribution of each word: (Assume the embedded vector of i-th word is A_i.).

$$Contribution = \begin{cases} MinMax(Cos\_similarity(A_i, W^h \circ S)) & For\ prediction\ "humans" \\ MinMax(Cos\_similarity(A_i, W^c \circ S)) & For\ prediction\ "ChatGPT" \end{cases} \tag{3}$$

4. Experiments

4.1. Data

The dataset used for model training and analysis is HC3-English [6], which consists of nearly 25,000 questions, nearly 60,000 human responses and 27,000 ChatGPT responses. The Q&A data comes from wiki, Medical Dialog, and other fields. The dataset was divided into training set, verification set and test set according to the ratio of 7:1:2. In the evaluation and analysis part, 500 texts from both humans and ChatGPT respectively in the test set will be applied on our model to generate statistical results.

4.2. Evaluation method

The evaluation will be conducted from three word-level features: word length, part-of-speech, and sentiment. By comparing the statistical results of these three features applied on our method with conclusions of existing research, we can check the correctness of our method.

4.3. Experimental details

For each of the 500 pieces of text generated by humans and ChatGPT, each text was predicted by the model and the words contribution was calculated. Additionally, for a certain piece of text, words with contribution score greater than 0.75 were defined as "decision important" words and were extracted and used in statistical process.

To avoid the bias caused by frequent words like "the", we de-weighted all words and used the average contribution of the repeated "decision important" words to indicate its contribution. After that, we got 1023/1179 "decision important" words for humans/ChatGPT. For comparison, the top 1000 contribution of words for humans/ChatGPT were kept for further analysis.

4.4. Results

4.4.1 Average Word Length

The average length of 1000 decision important words for humans and ChatGPT is shown in Table 1. The total for each category is 1000. The average length of decision important words for humans and ChatGPT is shown in Figure 4.

Table 1. Average length of 1000 decision important words for humans and ChatGPT.

Word length	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Human	**4**	**18**	**41**	**105**	**135**	125	170	**148**	**113**	71	31	23	13	3	0	0	0	0
ChatGPT	1	10	31	83	126	**169**	**177**	142	93	**78**	**38**	**27**	13	**4**	**4**	**3**	0	**1**

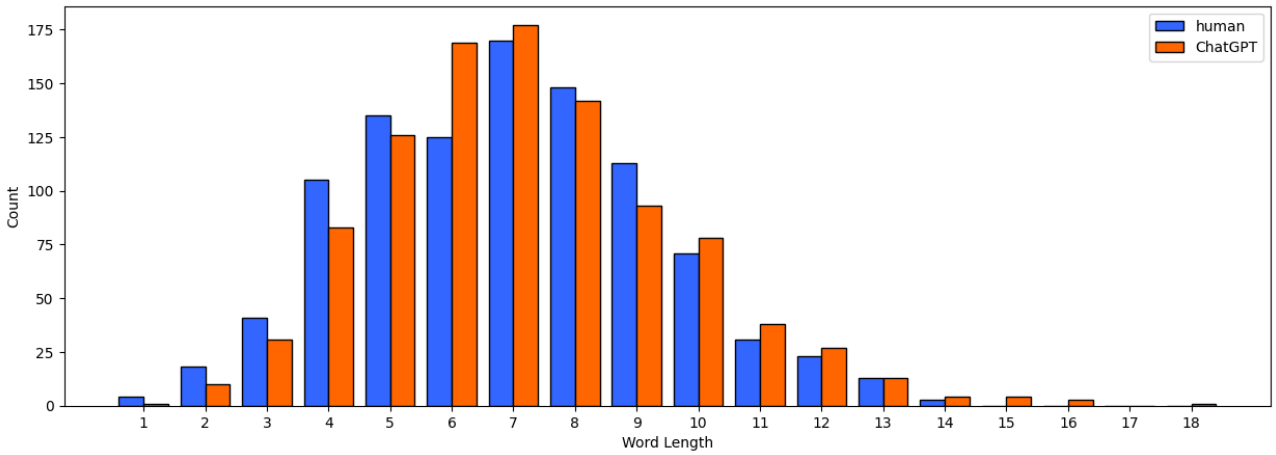

Figure 4. Average length of decision important words for humans and ChatGPT. The total count of each category is 1000.

4.4.2 Part-of-speech

The 17 parts of speech of 1000 decision important words for humans and chat systems are shown in Table 2. The total number of each category is 1000 humans and the speech of ChatGPT's decision important words is shown in Figure 5.

Table 2. 17 Parts of speech of 1000 decision important words for humans and ChatGPT.

Part-of-speech	**NOUN**	**VERB**	**ADJ**	**PROPN**	**ADV**	**PRON**	**NUM**	**PUNCT**	**ADP**
Human	371	**273**	**106**	**98**	**76**	**27**	22	7	7
ChatGPT	**529**	227	56	76	13	3	**27**	**44**	**12**
Part-of-speech	**AUX**	**INTJ**	**X**	**SCONJ**	**PART**	**CCONJ**	**SYM**	**DET**	
Human	**5**	**3**	**3**	1	**1**	0	0	0	
ChatGPT	1	2	1	**7**	0	**2**	0	0	

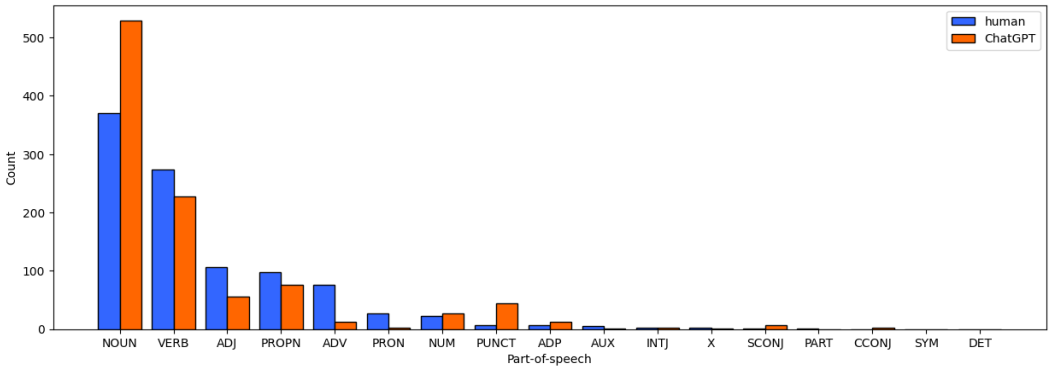

Figure 5. 17 Parts of speech of decision important words for humans and ChatGPT. The total count of each category is 1000.

4.4.3. Sentiment

Human and ChatGPT have different objectivity scores for decision important words, and the results are shown in Table 3. Human and ChatGPT have different objectivity scores for decision important words. If there are no matching words, the 0.0 category is hidden; If the number of words is large but the same, the 1.0 category is hidden. The result is shown in Figure 6.

Table 3. Different objectivity scores of decision important words for humans and ChatGPT.

Objectivity Score	0.125	0.250	0.375	0.500	0.625	0.750	0.875
Human	5	**17**	19	**35**	**28**	**40**	**91**
ChatGPT	**6**	13	**26**	28	26	35	58

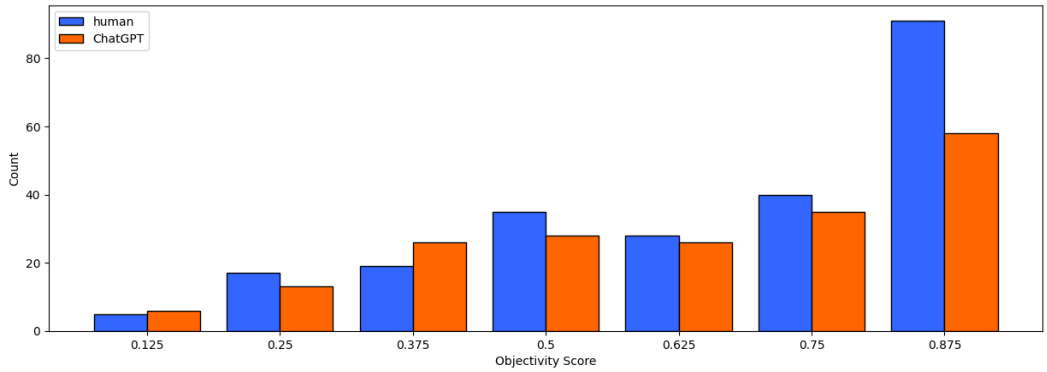

Figure 6. Different objectivity scores of decision important words for humans and ChatGPT. The 0.0 category is hidden for no matching words and the 1.0 category is hidden for large but same number of words.

5. Analysis

5.1. Average Word Length

Some research [11] revealed that humans tend to use shorter words to express the same meaning (e.g., using *math* rather than *mathematics*). Relatively, ChatGPT has no such tendency. As shown in Figure 4, among words of 1-5 length, the number of high-contribution words for prediction "human" is higher than that for ChatGPT. And for longer words, prediction "ChatGPT" is the majority. So, our method behaves well in this property.

5.2. Part-of-speech

According to Biyang Guo's research [6], ChatGPT tend to use less adverbs and punctuation but more other part-of-speech compared by humans, which shows the argumentation, informativeness and objectivity [12]. As shown in Table 2, the number of high-contribution nouns for prediction "ChatGPT" is nearly fifty percent higher than that of human. And much more adverbs were extracted for prediction "human". It indicates that the language model takes advantage of argument information to make the prediction of ChatGPT generated text.

However, the opposite results of verbs, adjectives, and punctuation show the weakness of our method. This may be led by the focus only on the word scale but ignoring the sentence structure.

5.3. Sentiment

Here we use the sentiment dictionary of WordNet [13], which provides a way to quantify the objectivity of a certain word by positive and negative scores. The details can be found in Appendix A. From Figure 6, we can see that most objectivity scores of less than 1.0 come from prediction "humans" and more objective words for prediction "ChatGPT". It suggests that words generated by ChatGPT have strong objectivity and that humans tend to use some emotional words. However, this distinction is not so obvious, whose reason may be that sentence structure is not taken into account in the sentiment labeling process.

6. Conclusions

In this work, we propose a method to interpret the word contribution of ChatGPT detector prediction. Specifically, we consider the information from classification process rather than only focusing on transformer attention. Our method can intuitively display the contribution of each word to the prediction result, with high calculation efficiency. The evaluation is conducted by comparing our statistical results with existing research from three perspectives. However, considering only in word-level may cause the distinction not to be obvious between two categories. So, for the future, it is necessary to take the sentence information into account within the process of word contribution calculating.

References

[1] Jesse Vig. A multiscale visualization of attention in the transformer model. arXiv preprint arXiv:1906.05714, 2019.

[2] Raymond Li, Wen Xiao, Lanjun Wang, Hyeju Jang, and Giuseppe Carenini. T3-vis: a vi- sual analytic framework for training and fine-tuning transformers in nlp. arXiv preprint arXiv:2108.13587, 2021.

[3] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. arXiv preprint arXiv:1906.04043, 2019.

[4] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. arXiv preprint arXiv:2301.11305, 2023.

[5] Alan Truly. Gptzero: how to use the chatgpt detection tool, February 2023. https://www.digitaltrends.com/computing/gptzero-how-to-detect-chatgpt-plagiarism/#dt-heading-how-does-gptzero-work, Last accessed on 2023-5- 1.

[6] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. arXiv preprint arXiv:2301.07597, 2023.

[7] OpenAI. Ai text classifier supported by openai, January 2023. https://beta.openai.com/ ai-text-classifier, Last accessed on 2023-5-2.

[8] Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. exbert: A visual analysis tool to explore learned representations in transformers models. arXiv preprint arXiv:1910.05276,2019.

[9] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the- art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[11] Zhenguang G Cai, David A Haslett, Xufeng Duan, Shuqi Wang, and Martin J Pickering. Does chatgpt resemble humans in language use? arXiv preprint arXiv:2303.08014, 2023.

[12] William Nagy and Dianna Townsend. Words as tools: Learning academic vocabulary as language acquisition. Reading research quarterly, 47(1):91– 108, 2012.

[13] Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, et al. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Lrec, volume 10, pages 2200– 2204, 2010.