Assessment of Text Classification Model

Rao Abdul Hannan, Yahan Yang

2024-10-23

1 Statement of the Problem

The AI research team has developed a random forest model to classify human-generated and AI-generated texts. The model performed quite well on the training data however, the performance decreased drastically on the test dataset. This paper investigates the question; what might be causing this difference and what might the team try as next steps?

2 Summary of Findings

First, we filtered the training dataset for human-generated and only GPT-4o generated text from the AI-models because it aligned well with our test dataset. We tokenized both the datasets i.e. splitting them into single words and compared the statistics, appended in Table 1. The training dataset contains $\approx 120,000$ tokens each for both human and GPT-4o generated texts, which for text analysis is almost negligible. Secondly, we evaluated the top 20 words in both the training and test datasets and while mostly similar, there were some differences as highlighted in Figure 1. Subsequently, we studied the Parts of Speech usage in the two datasets and found that verbs, pronouns and auxiliary verbs were more common in the training dataset but not in the test, suggesting major variations as emphasized in Figure 2. Moving forward, we assessed the keyness scores to understand the differences between human and AI-generated texts within both texts, shown in Figure 3. We found high keyness scores for punctuation such as ($ { } =) in the human-generated texts for test dataset. Upon further investigation, we discovered that very specific prompts were given in order to generate the texts in test dataset including the instructions, "formal academic and scientific writing voice. Use the first plural person form. Use active voice". This led to an increased usage of punctuation by the human authors since they are commonly used in academic writing. However the training dataset contains texts of academic, blog, fiction, news, spoken and technical & vocational material categories, and therefore did not have high keyness scores for punctuation. This finding motivated us to shift our analysis towards the Biber features, which the model actually uses to classify text. We performed principal component analysis (PCA) to determine how much variability in the data was explained by the Biber features. The resulting plot is appended as Figure 4 and clearly shows that the Biber features explain the variance in the training dataset quite well as evident by the spread of the Training Human and Training AI points. Meanwhile, the Test Human and Test AI points are clustered together which implies that 1) all the test data set texts are of the same nature i.e. academic and 2) the Biber features fail to explain the differences between the human and AI-generated texts in the test dataset. Since the AI research team's model is using these features to classify text, it is not performing well on the test dataset, as a result of being over-fitted on how the Biber features explain differences between the two text types within the training dataset.

3 Recommendations

Moving forward, we recommend that the AI research team take the following measures to improve the performance of the model:

- Increase the size of the training data set in an effort to include millions of tokens to train the model better
- Implement cross validation to avoid over-fitting of the model on the training data set

4 Appendix

Table 1: Token Counts

Dataset	Author Type	Tokens
Training	AI	126,062
Training	Human	115,033
Test	AI	48,275
Test	Human	48,140

Figure 1: Top 20 Words

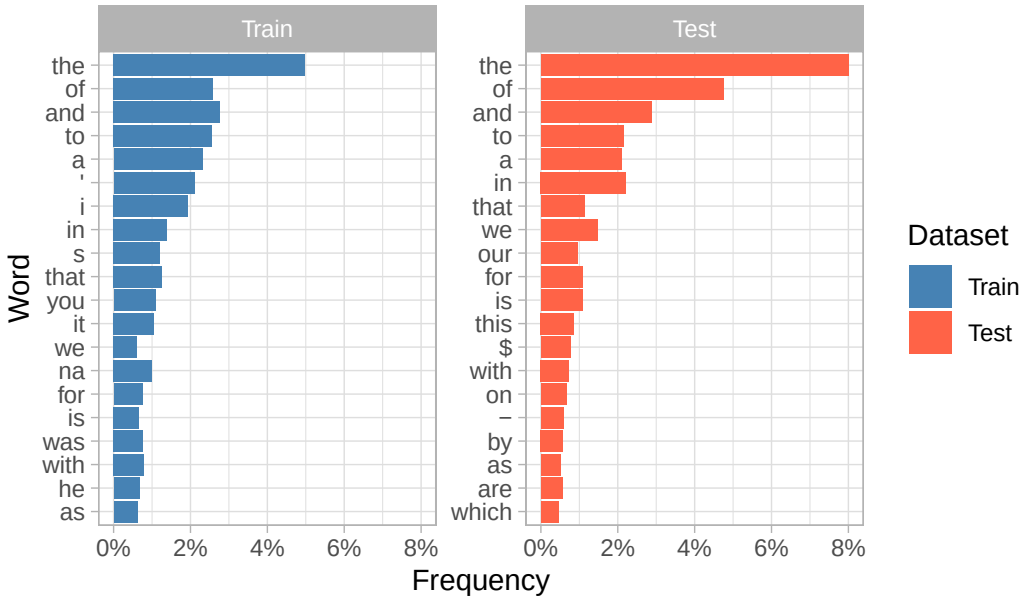

Figure 2: Parts of Speech Distribution

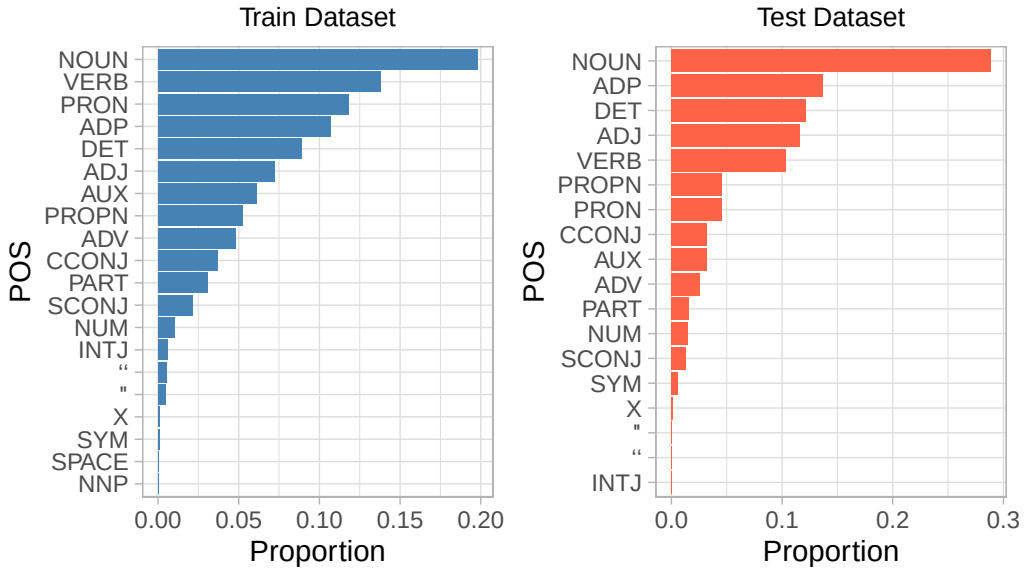

Figure 3: Keyness Analysis

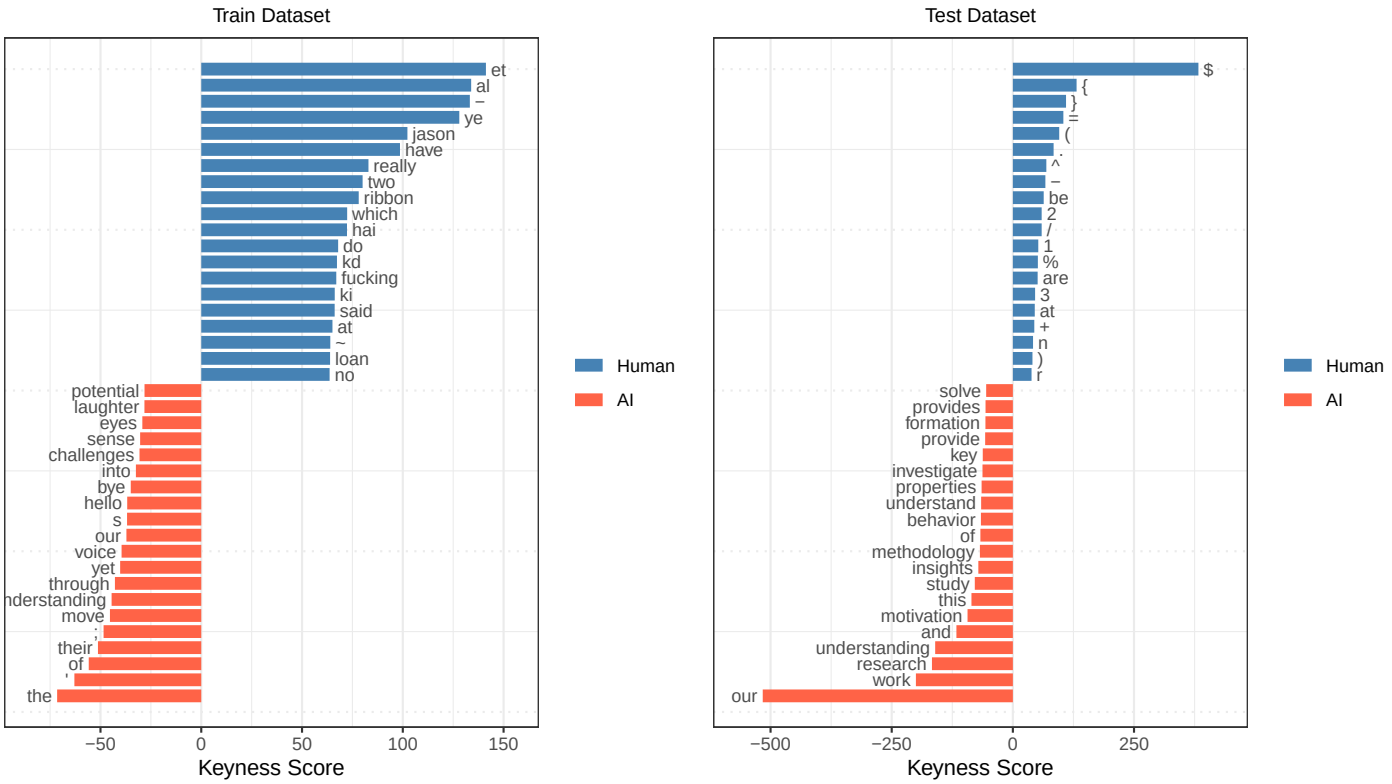

Figure 4: PCA of Biber Features: Training vs Test Data