

AI Generated Text Detection

Rares Andrei Stefanoiu, Radu-Constantin, coordinator: Florin Brad*

University of Bucharest, Romania
*Bitdefender, Romania



1. Introduction

In today’s age AI permeates every level of society and as such AI generated content like text, images, videos is everywhere. In many situations, using AI generated code is considered a form of plagiarism, so tools for telling apart human content from machine generated one are in increasingly high demand. Quite ironically, this paper aims at presenting two AI models that identify the work of their machine "kin"

2. Dataset

In gathering the data we used some presented in the [1] to gather the human vs machine texts for a plethora of domains. The resulted giant dataset contains circa 100000 examples of machine and human code. Moreover, the set is well balanced the two classes representing 51% and 49% of the total.

4. Large Language Models

Our first approach was a traditional Bert base model, but after further researching the paper [1] we switched to ROBERTA. Additionally, at our coordinator’s advice we also used DEBERTA[2]. Both recurrent models were optimized using Adam with learning rate 2e-5 and trained for 3 epochs

6. Results

For the trained models, given the fact that we primarily analyse cross-domain behaviour, there are quite a few metrics we can use on the entire class of models trained with either architecture, that being the overall general Accuracy, Accuracy on Seen domain(the train domain) and the Accuracy on the Unseen Domains(all domains except the train one).

Model	Gen	Seen	Unseen
ROBERTA	0.75	0.99	0.69
DEBERTA	0.73	0.99	0.67

Further analysis of the results is provided on the right panel. Please note that this aggregation of data fails to capture the more relevant individual characteristics of each model and is thus by all means insufficient to choose one in favor of another.

8. Conclusions

The task was a complex one but we consider that we did a great job and worked united as a team we are. There is plenty of room for optimisation and further analysis. It was the first time for us when we used such a concept like cross-domain. This project taught us a lot of new pieces of information, as well as reinforcing the knowledge base gained throughout the semester.

References

[1] Petar Ivanov Jinyan Su Artem Shelmanov Akim Tsvigun Chenxi Whitehouse Osama Mohammed Afzal Tarek Mahmoud Toru Sasaki Thomas Arnold Alham Fikri Aji Nizar Habash Iryna Gurevych Preslav Nakov Yuxia Wang, Jonibek Mansurov. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. pages 5–9.

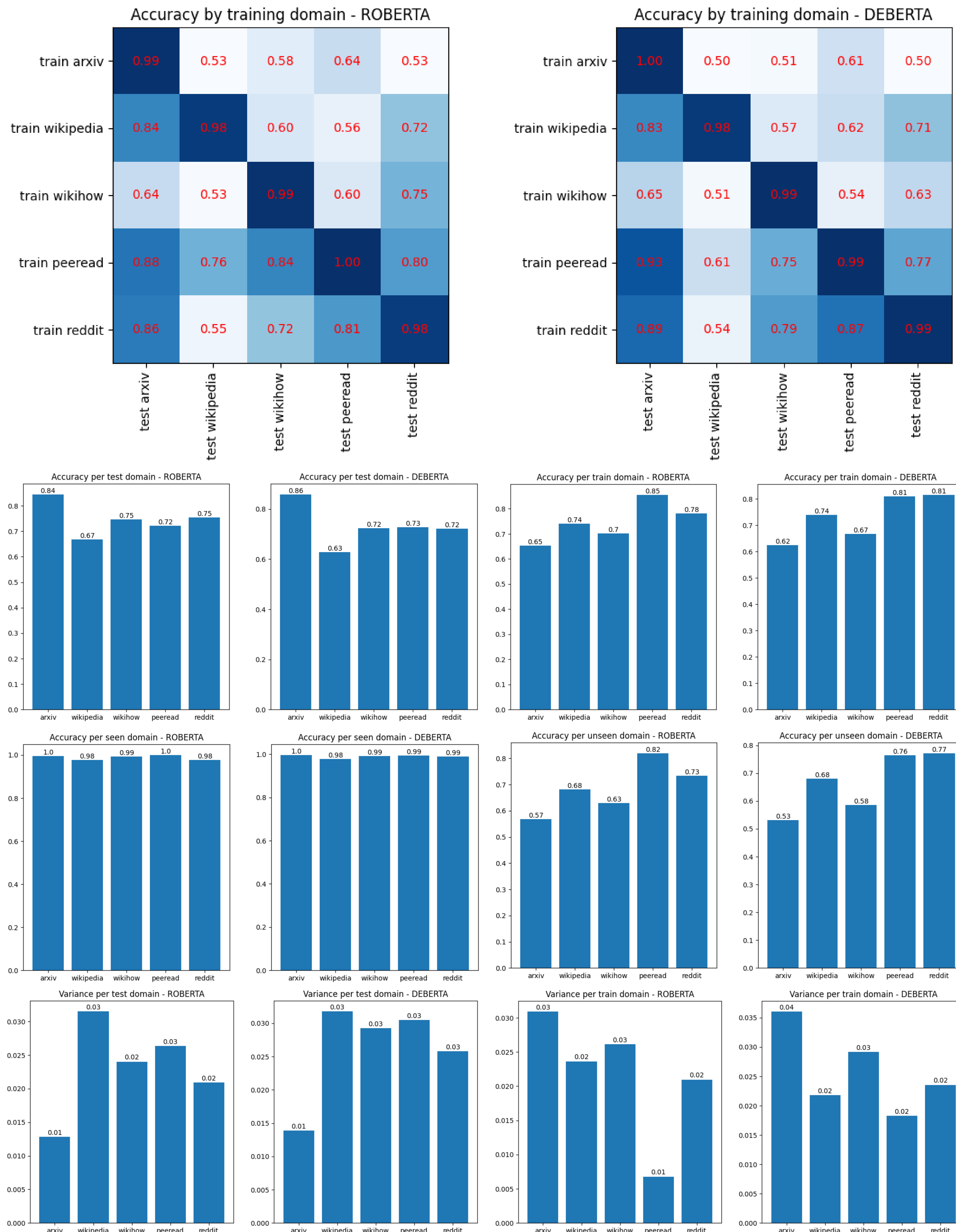
[2] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021.

Analysis

The performance is tracked both at classifier level to understand how powerful is it at generalisation on unseen types of data, as well as dataset level to determine how difficult is a type of content to be identified as either machine or human generated.

To do this we consider the resulted accuracy scores and extract important information from the matrices by applying means over alternating axes. Moreover, we also measure the variance because, according to ML philosophy a model with low variance and higher bias in preferable to one with low bias but high variance, since a higher variance is a signal that the model may not be reliable.

We will let the images speak for themselves



As it can be remarked, both classifiers have similar performances, although there are instances where one shines brighter than the other. As such, deciding which model to use in a real world scenario should be an act based on the context of the problem, namely the domain of use, rather than the absolute metrics from the results section