Detect Human or Machine Text

Jany-Gabriel Ispas, Radu-Tudor Vrînceanu and Florin Brad*
University of Bucharest, Romania
*Bitdefender, Romania

jany-gabriel.ispas@s.unibuc.ro, radu-tudor.vrinceanu@s.unibuc.ro, fbrad@bitdefender.com

1. Introduction

1. In the age of advanced text generation technologies, distinguishing between human-written and machine-generated text has become increasingly crucial. This paper proposes a method to classify text content as either human or machine-written using transformer-based models, with a particular focus on BERT-like architectures. Our approach leverages cross-domain and cross-generator datasets to enhance the robustness and generalizability of the model.

2. A crucial aspect of this task was the number of training and evaluation procedures needed to be run for our proposed BERT-like models (BERT and DeBERTa) in order to compare the results obtained.

2. Dataset and Preprocessing

Dataset Description: The dataset we have used for our model training and also for this benchmark was the M4 dataset [1]. Below we present some stats about the dataset that we constructed from the orignial one to train our models.

Dataset name	No. Machine-Text samples	No. Human-Text samples	Total
arXiv abstract, ChatGPT	3000	3000	6000
Reddit ELI5, ChatGPT	3000	3000	6000
Wikipedia, ChatGPT	2995	3000	5995
Reddit ELI5, Cohere	3000	3000	6000
Reddit ELI5, Davinci003	3000	3000	6000

Data preprocessing: We used the WordPiece tokenizer for the content that was given to the BERT model and Byte Pair Encoding tokenization for content that was given to DeBERTa.

3. Models

In this section we present our approaches used in the final form for realising the benchmark:

Baseline: We trained a classical BERT[2] model with an MLP having 3 hidden layers that was attached used for the classification result.
Final result: We have done a transfer learning on a classical BERT pretrained model and DeBERTa[3] pretrained model without attaching our propper MLP with the following parameters: (5e-5 learning rate, 0.01 weight decay, 16 batch size, 2 epochs for DeBERTa and 3 epochs for Bert, Adam optimizer).

4. Results

In this section we want to offer some results which are based on the following metrics: Accuracy, Precision, Recall, F1 score and ROC-AUC score. We denote Reddit ELI5 ChatGPT with [0], Wikipedia ChatGPT [1], arXiv Abstract ChatGPT [2] for the datasets used in the Cross-Domain benchmark.

Model	Acc-[0]	Prec-[0]	F1-[0]	Rec-[0]	AUC-ROC-[0]
BERT-[0]	0.9942	1.0000	0.9940	0.9881	0.9941
BERT-[1]	0.6042	0.5589	0.7140	0.9883	0.6042
BERT-[2]	0.5100	1.0000	0.0068	0.0034	0.5017
DeBERTa-[0]	0.9792	0.9607	0.9791	0.9983	0.9796
DeBERTa-[1]	0.4600	0.2317	0.0554	0.0315	0.4629
DeBERTa-[2]	0.5067	0.4996	0.6663	1.0000	0.5140

Model	Acc-[1]	Prec-[1]	F1-[1]	Rec-[1]	AUC-ROC-[1]
BERT-[0]	0.9841	0.9983	0.9838	0.9697	0.9840
BERT-[1]	0.9858	0.9788	0.9934	0.9861	0.9858
BERT-[2]	0.4942	1.0000	0.1416	0.0762	0.5381
DeBERTa-[0]	0.6895	0.6280	0.7701	0.9952	0.6751
DeBERTa-[1]	0.9558	0.9221	0.9580	0.9967	0.9553
DeBERTa-[2]	0.4833	0.4815	0.6497	0.9983	0.5031

Model	Acc-[2]	Prec-[2]	F1-[2]	Rec-[2]	AUC-ROC-[2]
BERT-[0]	0.9750	0.9754	0.9754	0.9754	0.9750
BERT-[1]	0.5058	0.5029	0.6693	1.0000	0.5058
BERT-[2]	0.9933	0.9983	0.9931	0.9880	0.9932
DeBERTa-[0]	0.9783	0.9909	0.9768	0.9630	0.9776
DeBERTa-[1]	0.6050	1.0000	0.3507	0.2126	0.6063
DeBERTa-[2]	1.0000	1.0000	1.0000	1.0000	1.0000

We denote Reddit ELI5 ChatGPT with [3], Reddit ELI5 Cohere [4], Reddit ELI5 Davinci003 [5] for the datasets used in the Cross-Generator benchmark.

Model	Acc-[3]	Prec-[3]	F1-[3]	Rec-[3]	AUC-ROC-[3]
BERT-[3]	0.9917	0.9862	0.9913	0.9965	0.9919
BERT-[4]	0.9800	0.9896	0.9794	0.9694	0.9798
BERT-[5]	0.9925	0.9848	0.9923	1.0000	0.9927
DeBERTa-[3]	0.9883	0.9824	0.9887	0.9951	0.9881
DeBERTa-[4]	0.4842	0.0000	0.0000	0.0000	0.5000
DeBERTa-[5]	0.9692	0.9554	0.9690	0.9830	0.9694

Model	Acc-[4]	Prec-[4]	F1-[4]	Rec-[4]	AUC-ROC-[4]
BERT-[3]	0.9600	0.9945	0.9573	0.9228	0.9590
BERT-[4]	0.9908	0.9837	0.9910	0.9983	0.9908
BERT-[5]	0.9750	0.9835	0.9755	0.9675	0.9752
DeBERTa-[3]	0.6342	0.9568	0.4464	0.2911	0.6388
DeBERTa-[4]	0.9992	1.0000	0.9992	0.9984	0.9992
DeBERTa-[5]	0.6758	0.8908	0.5215	0.3687	0.6635

Model	Acc-[5]	Prec-[5]	F1-[5]	Rec-[5]	AUC-ROC-[5]
BERT-[3]	0.9867	0.9983	0.9867	0.9754	0.9869
BERT-[4]	0.9508	0.9862	0.9510	0.9183	0.9522
BERT-[5]	0.9792	0.9603	0.9797	1.0000	0.9791
DeBERTa-[3]	0.9767	0.9847	0.9764	0.9683	0.9767
DeBERTa-[4]	0.5433	0.0000	0.0000	0.0000	0.5000
DeBERTa-[5]	0.9659	0.9386	0.9675	0.9984	0.9652

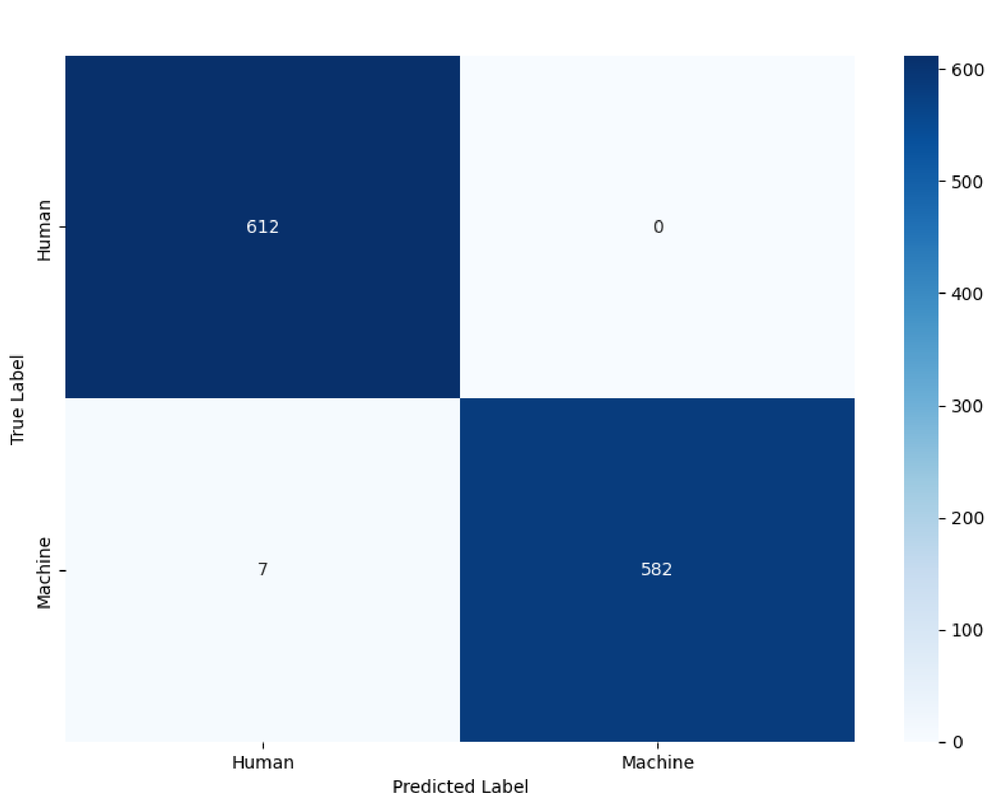
Detect Human or Machine Text

Jany-Gabriel Ispas, Radu-Tudor Vrînceanu and Florin Brad*
University of Bucharest, Romania
*Bitdefender, Romania

jany-gabriel.ispas@s.unibuc.ro, radu-tudor.vrinceanu@s.unibuc.ro, fbrad@bitdefender.com

5. Analysis of results

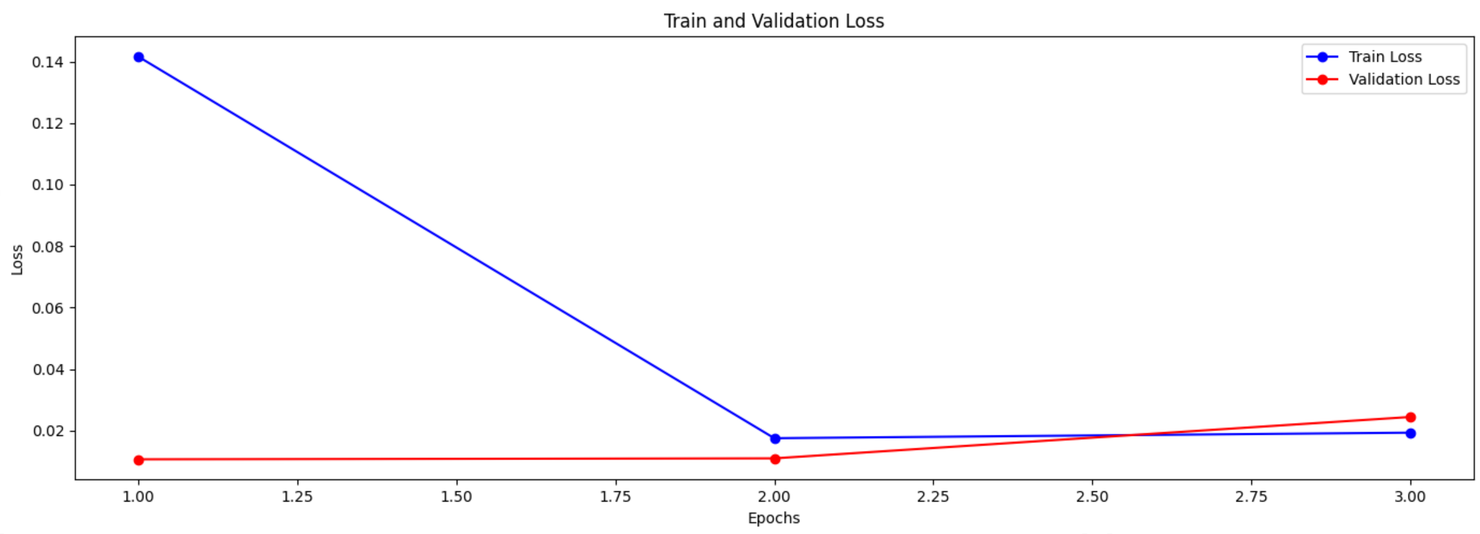
We can consider an example from the Cross-Domain challenge, we can choose BERT-[0], this was a classic BERT trained on Reddit ELI5 ChatGPT - [0] dataset, compared to Reddit ELI5 ChatGPT [0], Wikipedia ChatGPT [1], arXiv Abstract ChatGPT [2]. Our training, validation and test samples split were 70%, 10% and 20%. The dataset [0] has a ratio of 1:1 for every negative and positive class so we won't have any issues regarding imbalacend data.



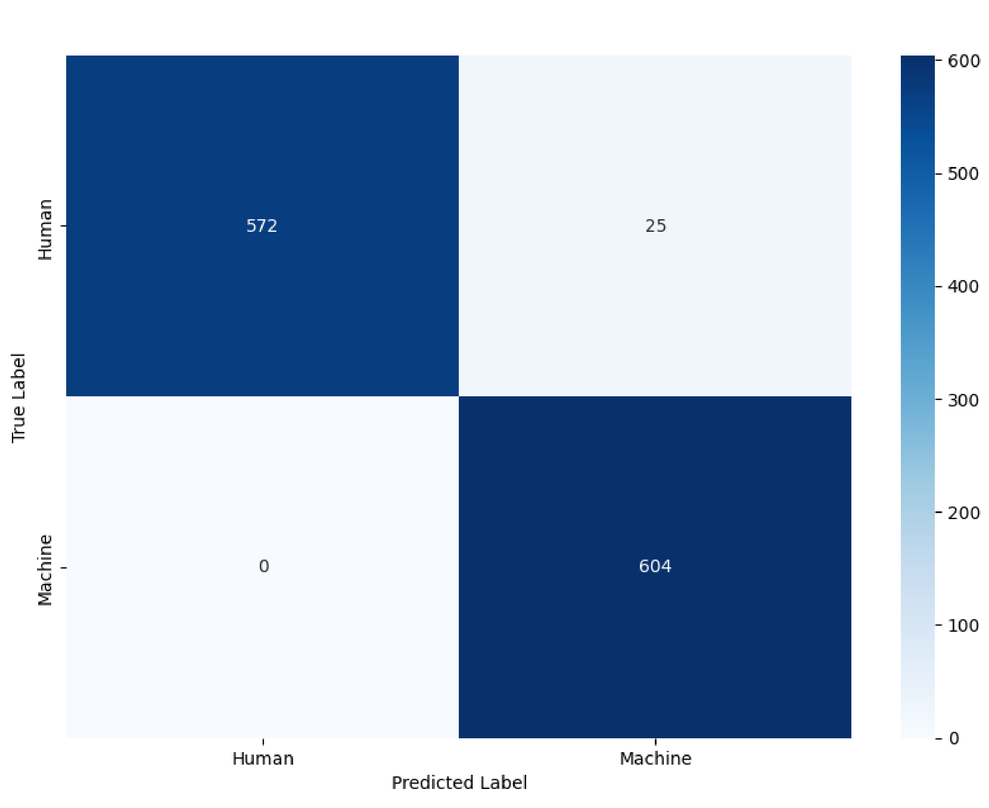
Wrong prediction examples:

Machine texts predicted as human

1. yes, it is possible for there to be masses that are close to the threshold of a black hole's escape velocity, but not quite there...
2. the short answer is that yes, it is possible to get sick from ingesting something contaminated by your own feces. while it is...



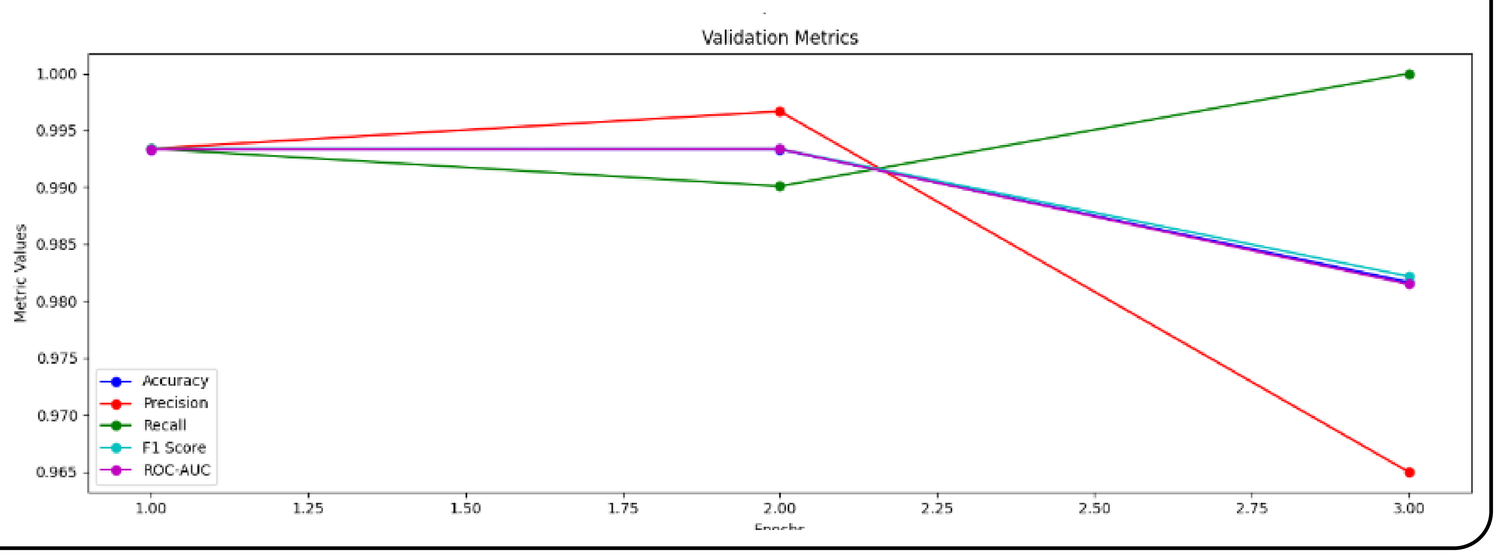
For the Cross-Generator challenge, we can choose BERT-[5], this was a classic BERT trained on Reddit ELI5 Davinci003 - [5] dataset. The Cross-Generator challenge has the same split ratio for training, validation and testing sample as the Cross-Domain challenge.



Wrong prediction examples:

Human texts predicted as machine

1. there are 3 ways to achieve backwards compatibility: 1) make the new console essentially a more powerful version of the old one. this is ...
2. to put it briefly, disease leading to a decreased population, economic distress, rebellions, civil wars, misleading numbers of roman troops...



6. Conclusion

This paper presents a robust approach to classifying text as human or machine-written using transformer-based models. By leveraging diverse datasets and fine-tuning BERT-like models, we achieve high accuracy and generalizability across domains and text generators. Our findings underscore the potential of transformer-based models in addressing the challenges posed by advanced text generation technologies.

References

[1] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[2] 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.

[3] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654, 2020.