Detecting Generated Scientific Papers using an Ensemble of Transformer Models

Anna Glazkova
University of Tyumen
Tyumen, Russia
a.v.glazkova@utmn.ru

Maksim Glazkov
Voctiv RnD d.o.o. Beograd
Belgrade, Serbia
my.eye.off@gmail.com

Abstract

The paper describes neural models developed for the DAGPap22 shared task hosted at the Third Workshop on Scholarly Document Processing. This shared task targets the automatic detection of generated scientific papers. Our work focuses on comparing different transformer-based models as well as using additional datasets and techniques to deal with imbalanced classes. As a final submission, we utilized an ensemble of SciBERT, RoBERTa, and DeBERTa fine-tuned using random oversampling technique. Our model achieved 99.24% in terms of F1-score. The official evaluation results have put our system at the third place.

1 Introduction

State-of-the-art natural language processing (NLP) tools generate high-quality texts that could hardly be distinguished from human-written texts. This represents a remarkable achievement in modern science, but raises challenges in terms of detecting machine-generated texts. Detection of automatically generated texts is crucial for many NLP tasks, in particular, for prevention of spreading fake scientific publications and citations (Else et al., 2021). Here we focus on the task of detecting automatically generated scientific excerpts as a part of the Third Workshop on Scholarly Document Processing shared tasks. The source code that we used for fine-tuning our models as well as additional data generated by us are freely available[1].

The work is based on the participation of our team in the DAGPap22 shared task. The objective of the task is to detect automatically generated papers in terms of a binary classification task. This task is challenging due to the developing models for text generation and wide spreading of untruthful content on the internet. To date, language models for generating texts are widely used in the scientific domain, for example for producing long and short summaries (Gharebagh et al., 2020; Cachola et al., 2020; Takeshita et al., 2022), citation texts (Xing et al., 2020; Ge et al., 2021), keyphrases (Glazkova and Morozov, 2022; Chowdhury et al., 2022), peer reviews (Yuan et al., 2021). The scientific community has held several machine learning competitions to identify machine-generated texts in different domains (Uchendu et al., 2021; Shamardina et al., 2022).

The paper is organized as follows. We provide the dataset and task description in Section 2. In Section 3, we describe our experiments during the development phase and report the official results. Section 4 concludes this paper.

2 Task Overview

2.1 Task Definition

The objective of the task is to identify whether a text is automatically generated. Therefore, the task represents a binary classification problem, the purpose of which is to split the given texts into two mutually exclusive classes. Formally, the problem is described as follows.

- **Input.** Given a scientific excerpt.
- **Output.** One of two different labels, such as "human-written" or "machine-generated".

2.2 Data

The original training set contains 5350 excerpts from a scientific papers, among which 1686 are human-written and 3664 are machine-generated. The test set includes 21403 excerpts. The text corpus is based on the work by Cabanac et al. (2021), as well as fragments collected by Elsevier publishing and editorial teams. The statistics is presented

[1] https://github.com/oldaandozerskaya/DAGPap22

in Table 1[2]. Table 2 contains some examples of automatically generated texts.

Characteristic	Train	Test
Avg number of words	157.4	158.37
Min number of words	51	51
Max number of words	1895	1784
Avg number of sentences	5.8	5.75
Min number of sentences	1	1
Max number of sentences	63	68

Table 1: Data statistics.

ID	Excerpt
23	Electronic nose or machine olfaction are systems used for detection and identification of odorous compounds and gas mixtures Electronic nose or machine olfaction are systems used for detection and identification of odorous compounds and gas mixtures. Olfactors, e.g. motorbikes, are used for odor detection. These devices do not detect volatile agents or gas mixtures, and cannot be used for quantitative odor determination.
55	For the low price of coal and ineffective environmental management in mining area, China is in the dilemma of the increasing coal demand and the serious environmental issues in mining area For the low price of coal and ineffective environmental management in mining area, China is in the dilemma of the increasing coal demand and the serious environmental issues in mining area.
242	The motivation behind this paper is to answer analysis of the past portrayals of Sandler and Smith of the numeraire in an intertemporal investigation of Pareto effectiveness conditions. This reevaluation recommends that the job of the numeraire is demonstrated to be less obvious than Cabe infers. In addition, the examination shows that the prior ends are not critically subject to the numeraire presumption.

Table 2: Examples of generated texts from the official training set.

3 Our Work

3.1 Models

Model	Value
Vocabulary (K)	
SciBERT	30
RoBERTa	50
DeBERTa	50
Backpone Parameteres (M)	
SciBERT	110
RoBERTa	355
DeBERTa	350
Hidden Size	
SciBERT	768
RoBERTa	1024
DeBERTa	1024
Layers	
SciBERT	12
RoBERTa	16
DeBERTa	24

Table 3: Hyperparameteres of the considered BERT-based models (SciBERT$_{base-cased}$, RoBERTa$_{large}$, and DeBERTa$_{large}$).

In this work, we used neural models based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) because they showed high results in the scientific domain (Glazkova, 2021; Pan et al., 2021; Zhu et al., 2021). We experimented with the following models, the overview of which is presented in Table 3:

- SciBERT$_{base-cased}$ (Beltagy et al., 2019), a BERT-based model that is pretrained on the texts of papers taken from Semantic Scholar.
- RoBERTa$_{large}$ (Liu et al., 2019), a modification of BERT that is pretrained using dynamic masking.
- DeBERTa$_{large}$ (He et al., 2020), a model that is pretrained using disentangled attention and enhanced mask decoder.

To evaluate our models during the development phase, we performed 3-fold cross-validation on the training set. The results were evaluated in terms of macro-averaged F1-score (F1), precision (P), and recall (R).

[2] The number of words and sentences was defined using NLTK (Bird, 2006)

Model	P	R	F1
SciBERT$_{128}$	96.19	94.69	95.38
SciBERT$_{256}$	97.58	96.49	96.99
SciBERT$_{512}$	97.84	97.16	97.49
RoBERTa	96.54	94.89	95.65
DeBERTa	97.35	97	97.17
SciBERT$_{512}$ + oversampling	**98.2**	97.92	**98.06**
SciBERT$_{512}$ + undersampling	97.07	95.42	96.15
SciBERT$_{512}$ + class weighting	98.05	97.81	97.93
RoBERTa + oversampling	96.92	96.5	96.7
RoBERTa + undersampling	95.55	92.83	93.89
RoBERTa + class weighting	96.62	96.49	96.56
DeBERTa + oversampling	97.51	96.61	97.04
DeBERTa + undersampling	95.62	93.04	94.13
SciBERT$_{512}$ + KP20K (BT) + oversampling	97.65	**98.18**	97.91
SciBERT$_{512}$ + KP20K (GPT-2) + oversampling	97.16	97.03	97.07
SciBERT$_{512}$ + original (BT) + oversampling	97.44	97.75	97.59
SciBERT$_{512}$ + original (GPT-2) + oversampling	97.56	98.15	97.84
RoBERTa + KP20K (BT) + oversampling	96.86	96.48	96.66
RoBERTa + KP20K (GPT-2) + oversampling	96.49	95.2	95.8
RoBERTa + original (BT) + oversampling	96.56	95.99	96.26
RoBERTa + original (GPT-2) + oversampling	96.12	96.12	96.12
DeBERTa + KP20K (BT) + oversampling	96.76	97.03	96.89
DeBERTa + KP20K (GPT-2) + oversampling	94.16	95.86	94.95
DeBERTa + original (BT) + oversampling	96.51	96.7	96.59
DeBERTa + original (GPT-2) + oversampling	96.58	96.94	96.76

Table 4: Results (%, development phase).

3.2 Experiments

We adopted pretrained models from HuggingFace (Wolf et al., 2020) and fine-tuned them using SimpleTransformers[3]. We fine-tuned each pre-trained language model for three epochs with the learning rate of 2e-5 using the AdamW optimizer (Loshchilov and Hutter, 2017). We set batch size to 16 and used the sliding window technique to prevent truncating longer sequences. We utilized the maximum sequence length equal to 128, 256, and 512 for SciBERT (SciBERT$_{128}$, SciBERT$_{256}$, and SciBERT$_{512}$ respectively) and 128 for RoBERTa and DeBERTa due to the limited computing resources. Similar to our previous work (Glazkova et al., 2021), we used raw texts as an input.

Since the corpus provided by the organizers is imbalanced, we explored several techniques to handle imbalanced data. Namely, we used a) random oversampling, b) random undersampling, c) class weighting, d) generating new data. Random oversampling and undersampling are implemented using the Imbalanced-learn library[4]. To generate new data, we experimented with the original corpus and the fragment of the KP20K dataset (Meng et al., 2017). KP20K is a large-scale scholarly papers dataset for keyphrase extraction containing 568K papers with their abstracts. To produce new machine-generated data, we utilized two techniques for text generation: a) Back Translation (BT)[5] through Googletrans[6], and b) zero-shot generation by prompting GPT-2 (Radford et al.) and specifying the maximum number of generated tokens equal to the number of tokens in the source text (see Figure 1 for example).

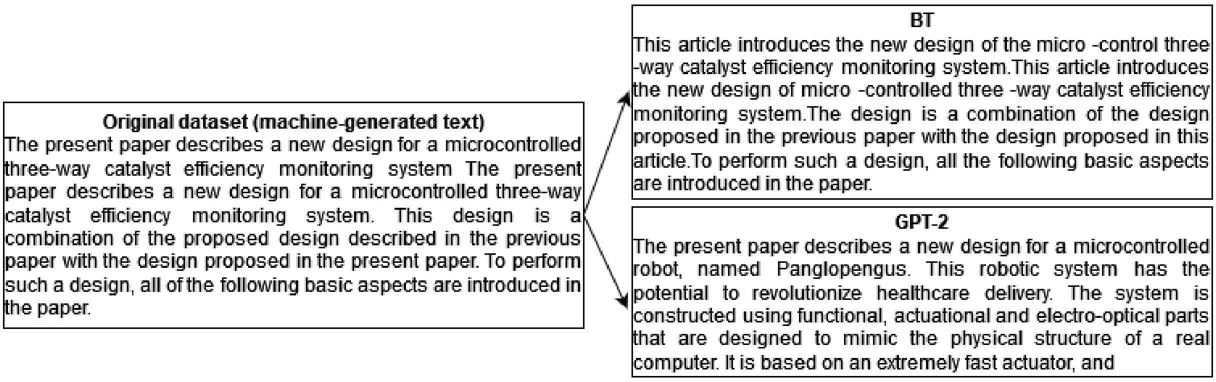
The results are presented in Table 4. In our experiments, the model fine-tuned on longer input sequences (SciBERT$_{512}$) performed better than other baselines despite the use of the sliding win-

[3] https://simpletransformers.ai

[4] https://imbalanced-learn.org

[5] https://github.com/hhhwwwuuu/BackTranslation

[6] https://py-googletrans.readthedocs.io/en/latest



Original dataset (machine-generated text)

The present paper describes a new design for a microcontrolled three-way catalyst efficiency monitoring system The present paper describes a new design for a microcontrolled three-way catalyst efficiency monitoring system. This design is a combination of the proposed design described in the previous paper with the design proposed in the present paper. To perform such a design, all of the following basic aspects are introduced in the paper.

BT

This article introduces the new design of the micro -control three -way catalyst efficiency monitoring system.This article introduces the new design of micro -controlled three -way catalyst efficiency monitoring system.The design is a combination of the design proposed in the previous paper with the design proposed in this article.To perform such a design, all the following basic aspects are introduced in the paper.

GPT-2

The present paper describes a new design for a microcontrolled robot, named Panglopengus. This robotic system has the potential to revolutionize healthcare delivery. The system is constructed using functional, actuational and electro-optical parts that are designed to mimic the physical structure of a real computer. It is based on an extremely fast actuator, and

Figure 1: Example of generating new data using BT and GPT-2.

dow technique. Due the processing of class imbalance, we found that oversampling and class weighting increase the performance of the models while undersampling produces lower results. Further, we experimented with using additional data. First, we made an attempt to add scientific abstracts from KP20K utilizing texts of 1000 random abstracts and 1000 texts generated by BT or GPT-2 and than perform oversampling. Second, we tried to produce new examples of machine-generated excerpts from the dataset provided by the organizers of the competition. We generated 1000 examples using BT and GPT-2, added them to the training set, and finally performed oversampling. The use of additional data showed no increase compared to the models fine-tuned with oversampled texts.

3.3 Results

During the evaluation phase, we experimented with the hard and soft voting ensembles of transformer-based models. The results were evaluated on the official test set. Our best submission is an ensemble of SciBERT, RoBERTa, and DeBERTa fine-tuned using random oversampling technique. The confusion matrix for this solution is presented in Figure 2. The ensembling of predictions was performed at two levels:

1. Model level, i. e. soft voting calculated for three models of the same type fine-tuned with different random seeds.

2. Ensemble level, i. e. hard voting for the labels produced by the models of different type.

Table 5 shows the comparison of our best solution to the official scores from the private leaderboard of the competition[7]. In this competition, only five models outperformed the baseline provided by the organizers. Our model achieved 99.24% of F1-score and ranked the third place of the leaderboard for this task.

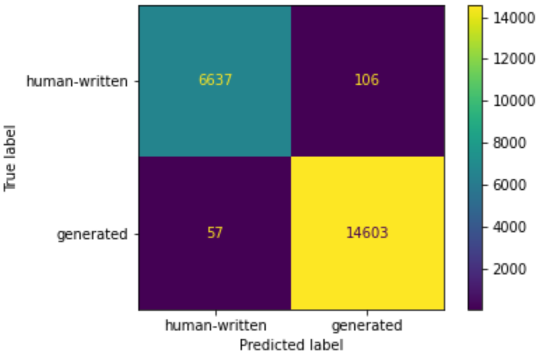


Figure 2: Confusion matrix for our model.

Run name	F1
Our solution	99.24
Stronger benchmark	98.32
Tf-Idf & logreg benchmark	82.04
Average scores	92.96

Table 5: Official results (%, private leaderboard).

4 Conclusion

In this work, we have explored the application of BERT-based models to the task of detecting machine-generated scientific texts. We have evaluated several techniques for handling imbalanced data and compared three models in a variety of settings. Our results on the test data showed that

[7] https://www.kaggle.com/competitions/detecting-generated-scientific-papers

the ensemble of different transformer-based models outperforms other our submissions and strong baselines. Moreover, our final model ranked third in this task.

A further study could explore the state-of-the art-in detecting automatically generated papers for other languages and multilingual corpora. Another future direction is to continue our experiments with generating new data to improve the classification performance.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP*. Association for Computational Linguistics.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Guillaume Cabanac, Cyril Labbé, and Alexander Magazinov. 2021. Tortured phrases: A dubious writing style emerging in science. evidence of critical issues affecting established journals. *arXiv preprint arXiv:2107.06751*.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777.

Md Faisal Mahbub Chowdhury, Gaetano Rossiello, Michael Glass, Nandana Mihindukulasooriya, and Alfio Gliozzo. 2022. Applying a generic sequence-to-sequence model for simple and effective keyphrase generation. *arXiv preprint arXiv:2201.05302*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Holly Else et al. 2021. 'tortured phrases' give away fabricated research papers. *Nature*, 596(7872):328–329.

Yubin Ge, Ly Dinh, Xiaofeng Liu, Jinsong Su, Ziyao Lu, Ante Wang, and Jana Diesner. 2021. BACO: A background knowledge-and content-based framework for citing sentence generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1466–1478.

Sajad Sotudeh Gharebagh, Arman Cohan, and Nazli Goharian. 2020. Guir@ longsumm 2020: Learning to generate long summaries from scientific documents. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 356–361.

Anna Glazkova. 2021. Identifying topics of scientific articles with bert-based approaches and topic modeling. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 98–105. Springer.

Anna Glazkova, Maksim Glazkov, and Timofey Trifonov. 2021. g2tmn at constraint@AAAI2021: exploiting CT-BERT and ensembling learning for COVID-19 fake news detection. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 116–127. Springer.

Anna Glazkova and Dmitry Morozov. 2022. Applying transformer-based text summarization for keyphrase generation. *arXiv preprint arXiv:2209.03791*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592.

Chunguang Pan, Bingyan Song, Shengguang Wang, and Zhipeng Luo. 2021. BERT-based acronym disambiguation with multiple training strategies. In *Scientific Document Understanding 2021*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Tatiana Shamardina, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the the RuATD shared task 2022 on artificial text detection in Russian. *arXiv preprint arXiv:2206.01583*.

Sotaro Takeshita, Tommaso Green, Niklas Friedrich, Kai Eckert, and Simone Paolo Ponzetto. 2022. X-scitldr: cross-lingual extreme summarization of scholarly documents. *arXiv preprint arXiv:2205.15051*.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. Turingbench: A benchmark environment for turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. 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.

Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. 2020. Automatic generation of citation texts in scholarly papers: A pilot study. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6181–6190. Association for Computational Linguistics.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. Can we automate scientific reviewing? *arXiv preprint arXiv:2102.00176*.

Danqing Zhu, Wangli Lin, Yang Zhang, Qiwei Zhong, Guanxiong Zeng, Weilin Wu, and Jiayu Tang. 2021. AT-BERT: Adversarial training BERT for acronym identification winning solution for SDU@ AAAI-21.