Overview of the DagPap22 Shared Task on Detecting Automatically Generated Scientific Papers

Yury Kashnitsky[1] **Drahomira Herrmannova**[1] **Anita de Waard**[1]
Georgios Tsatsaronis[1] **Catriona Fennell**[1] **Cyril Labbé**[2]
Elsevier, USA[1] Université Grenoble Alpes, France[2]
{d.herrmannova, a.dewaard, y.kashnitskiy
g.tsatsaronis, c.fennell}@elsevier.com,
cyril.labbe@imag.fr

Abstract

This paper provides an overview of the 2022 COLING Scholarly Document Processing workshop shared task on the detection of automatically generated scientific papers. We frame the detection problem as a binary classification task: given an excerpt of text, label it as either human-written or machine-generated. We shared a dataset containing excerpts from human-written papers as well as artificially generated content and suspicious documents collected by Elsevier publishing and editorial teams. As a test set, the participants were provided with a 5x larger corpus of openly accessible human-written as well as generated papers from the same scientific domains of documents. The shared task saw 180 submissions across 14 participating teams and resulted in two published technical reports. We discuss our findings from the shared task in this overview paper.

1 Introduction

There are increasing reports that research papers can be written by computers, which presents a series of concerns (e.g., see Cabanac et al. (2021)). For scientific publishers, the problem of automatic detection of generated scientific content provides a technical and ethical challenge. Technically, any detector of automatically generated content is hard to remain effective for long: e.g., if a new language or summarization model is developed to generate text, the detector no longer works (for more details see the paper by (Rosati, 2022)). In terms of ethics, it is important to distinguish malicious and benign scenarios of generated content appearing in submitted scientific manuscripts. It is possible that authors might resort to translation systems to aid their writing process, e.g. helping to translate some excerpts from their native language into English. However, there is increased evidence of fraudulent papers, partially or entirely artificially generated, that have passed the peer-review process and were published. Most notoriously, there has been an experiment called SCIgen[1] where an entire conference workshop was generated comprised of gibberish talks. See (Noorden, 2021) and (Labbé and Labbé, 2012) for more details on SCIgen's impact on science, SCIgen detectors, and other examples of gibberish papers lurking into scientific literature. Recently, "paper mills" (Else, 2021) have caught increased attention as the main source of potentially fabricated research content. In (Cabanac et al., 2021), the authors found traces of GPT2-generated content in scientific literature, along with "tortured phrases" appearing as a side effect of using generating models and paraphrasing tools like SpinBot[2].

Partly driven by this work, we have organised a competition to encourage the NLP community to detect automatically generated papers. This project is a collaboration between a publisher (Elsevier) and the research community to attempt a resolution through technical means. To build on the excellent detective work by the (Cabanac et al., 2021) team, excerpts from the papers in their paper were added as examples of "fake" text to the dataset in this competition.

2 Corpus creation

The data provided for this competition contains text excerpts from scientific papers and an indication of whether these texts are "fake" (probably generated) or "real", i.e. human-written. The data comes from both published and retracted Scopus papers with 5,327 records in the training set and

[1] https://pdos.csail.mit.edu/archive/scigen/

[2] https://spinbot.com

Proceedings of the Third Workshop on Scholarly Document Processing, pages 210–213
October 12–17, 2022.

21,310 records in the test set. Around 69% of all texts in both sets are "fake". The code reproducing some steps of the data generation process is publicly available (Kashnitsky, 2022).

The data comes from the following sources:

1. MICPRO retracted papers ("fake"). These are excerpts from a set of retracted papers of the "Microprocessors and microsystems" journal (MICPRO). Some of those are explored in (Cabanac et al., 2021) in the context of "tortured phrases";

2. Good MICPRO papers ("real"). Similar excerpts from earlier issues of the "Microprocessors and microsystems" journal;

3. Abstracts of papers related to UN's Sustainable Development Goals[3] ("real"). Sustainable Development Goals (SDGs) cover a wide range of topics, from poverty and hunger to climate action and clean energy;

4. Summarized SDG abstracts ("fake"). These texts were generated using "pszemraj/led-large-book-summary" model;

5. Summarized MICPRO abstracts ("fake"). The same model as above was applied to MICPRO abstracts;

6. Generated SDG abstracts (fake). These texts were generated using the "EleutherAI/gpt-neo-125M" model with the first sentence of the abstract being a prompt;

7. Generated MICPRO abstracts (fake). The same model as above was applied to MICPRO abstracts;

8. SDG abstracts paraphrased with Spinbot ("fake");

9. GPT-3 few-shot generated content with the first sentence of the abstract as a prompt ("fake").

We also experimented with back-translated content, e.g. when the original excerpt is translated to, say, German and then back to English. We found that modern translation systems are so advanced that the back-translated snippets look almost identical to the originals, hence we rejected the idea of including such content as "fake". Repeated back-translation, especially with under-represented languages (say, En -> Swahili -> Korean -> En) might introduce some artefacts and help the back-translated snippets look "more fake", but we didn't conduct such experiments.

[3] https://sdgs.un.org/2030agenda

Source N	Source	Acc, %
4	summarized_sdg	100
5	summarized_micpro	99.9
8	spinbot_paraphrased	98.9
1	micpro_retracted	97
9	generated_gpt3	95.5
7	generated_micpro	87.3
6	generated_sdg	74
3	sdg_abstracts_original	57.4
2	micpro_original	57.3

Table 1: Validation accuracy split by data provenance type from Sec. 2. Model: logistic regression with Tf-Idf text representation.

3 Competition setup

3.1 Metric and data split

The metric chosen in the competition is average F1-score. We merged all data sources described in Sec. 2 (skipping only back-translated content as almost identical to the original), and performed a stratified 20/80 train-test split intentionally leaving a small train set. This resulted in 5327 training records and 21310 test records forming the datasets described on the competition page[4].

3.2 Baselines

As organizers, we provided 2 baselines: Tf-Idf & logistic regression[5] and fine-tuned SciBERT achieving 82% and 98.3% test set F1 score, respectively.

4 Experiments with data provenance

Given one of the simplest possible baseline models, namely, Tf-Idf & logistic regression, we explored model accuracy w.r.t. to data provenance, i.e. types of content described in Sec. 2.

Table 1 shows validation accuracy for the test set split by data provenance type, see Sec. 2 for details. The Tf-Idf & logistic regression model

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

[5] Kaggle Notebook: https://bit.ly/3dJR9m0

was trained with 5,327 training records (containing data from all 9 sources listed in Sec. 2), and then the predictions were evaluated separately for each data source, i.e. first for excerpts from retracted MICPRO papers, then for excerpts from good MICPRO papers, and so on, up to excerpts of text generated with GPT-3.

We see that summarized content was easily detected, probably due to peculiarities of the "pszemraj/led-large-book-summary" summarization model, e.g. most of the summaries are opened with "This paper is focused on..." or "In this paper, the authors ...". Likewise, SpinBot-generated content is easily detected, probably because SpinBot was found to introduce "tortured phrases" (Cabanac et al., 2021) and those can be spotted even with Tf-Idf. Somewhat surprisingly, the model had no problem with retracted MICPRO content.

The model had most trouble identifying original human-written content, a possible reason is that with all the generated content due to class imbalance (70% of the data is "fake"), it's easy to get false positives when a normal human-written text is easy to be confused with fake content.

5 Systems Overview

14 teams participated in the task this year, with a total of 180 submissions. Out of these, 11 teams managed to beat the publicly shared Tf-Idf & logreg baseline, and 5 teams managed to beat the fine-tuned SciBERT baseline which was not publicly shared. Three teams submitted peer-reviewed technical reports, of which two are published as part of the workshop proceedings. Both teams managed to achieve >99% test set F1-score.

In "Detecting Generated Scientific Papers using an Ensemble of Transformer Models" (Glazkova and Glazkov, 2022) the authors describe an ensemble of SciBERT, RoBERTa, and DeBERTa fine-tuned using random oversampling technique.

The winning team led by Domenic Rosati "SynSciPass: detecting appropriate uses of scientific text generation" (Rosati, 2022) generates a partially synthetic dataset similar to what we as competition organizers had done. Then Rosati shows that the models trained with the DAGPap22 generalize badly to a new data source. Ablations studies show that generalization to unseen text generation models might not be possible with current approaches. Rosati concludes that the results in his paper should make it clear that at this point machine generated text detectors should not be used in production because they do not perform well on distribution shifts and their performance on realistic full-text scientific manuscripts is currently unknown.

6 Discussion

It turned out that the task turned was very easy to solve, with winners' models hitting >99% of the test set F1 scores. Although this suggests that the task of detecting machine-generated content is easy, both work done at Elsevier and as reported by the team led by Rosati convinces us that we are far from developing a general detector of generated content. Each new model (say, GPT-4) for which we don't have training data poses a new challenge, and any detector is likely to fail at identifying content generated with such a model due to a data shift. In summary, the problem is far from being solved: at this point we can not rely on detectors of generated content to support our production systems. However, the DAGPap22 shared task did offer a step forward to explore this challenging problem, and we hope to work together with the community on resolving this pernicious issue.

References

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*.

Holly Else. 2021. 'tortured phrases' give away fabricated research papers. *Nature*.

Anna Glazkova and Maksim Glazkov. 2022. Detecting generated scientific papers using an ensemble of transformer models. In *Proceedings of the Third Workshop on Scholarly Document Processing*. Association for Computational Linguistics.

Yury Kashnitsky. 2022. Source code for the coling workshop competition "detecting automatically generated scientific papers". https://github.com/Yorko/fake-papers-competition-data.

Cyril Labbé and Dominique Labbé. 2012. Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science? *Scientometrics*, pages 10.1007/s11192–012–0781–y.

Richard Van Noorden. 2021. Hundreds of gibberish papers still lurk in the scientific literature. *Nature*.

Domenic Anthony Rosati. 2022. Synscipass: detecting appropriate uses of scientific text generation. In *Proceedings of the Third Workshop on Scholarly Document Processing*. Association for Computational Linguistics.