DAGPap24: Detecting automatically generated scientific papers
Milestone 2
Adrian Dinu Urse

1. Overview

As AI technologies become increasingly capable of generating text that mimics human writing, there's a growing risk of fake or misleading scientific content being published. This can undermine the credibility of scientific communication and the trust that the public places in scientific findings. Ensuring the authenticity of scientific documents is crucial for maintaining the integrity of the scientific record.

This competition is a part of the shared task hosted within the 4th workshop on Scholarly Document Processing (SDP 2024), held in association with The 62nd Annual Meeting of the Association for Computational Linguistics ACL 2024.

The Detecting Automatically Generated Papers (DAGPap) competition aims to encourage the development of robust, reliable AI-generated scientific text detection systems, utilizing a diverse dataset and varied machine learning models in a number of scientific domains.

As a follow-up to DAGPap 2022 (Kaggle), which targeted sentence-level classification into human-written or ML-generated texts, we now extend the task to token-level classification of scientific texts into 4 classes:

- human-written,
- synonym-replace,
- chatGPT-generated,
- summarized

2. Dataset

The provided dataset within DAGPap 2024 is created from a large corpus of scientific papers. For this dataset, 25% of the text was modified using one of three techniques:

- Synonym replacement with NLTK;
- Summarization with one of Deep Learning models available via HuggingFace;
- Paraphrasing with ChatGPT

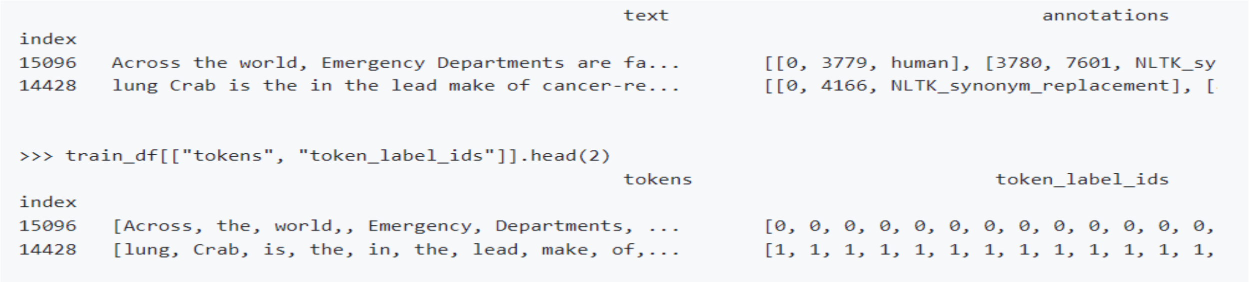
Each text has associated, a label list [[start_id, end_id, label_id]], where ids indicate the start and end token ids of a span, and label_id indicates its provenance ('human', 'NLTK_synonym_replacement', 'chatgpt', or 'summarized')

The training set has the following columns:

- text – a fragment from the article's full text;
- tokens – same as text, split by whitespaces;

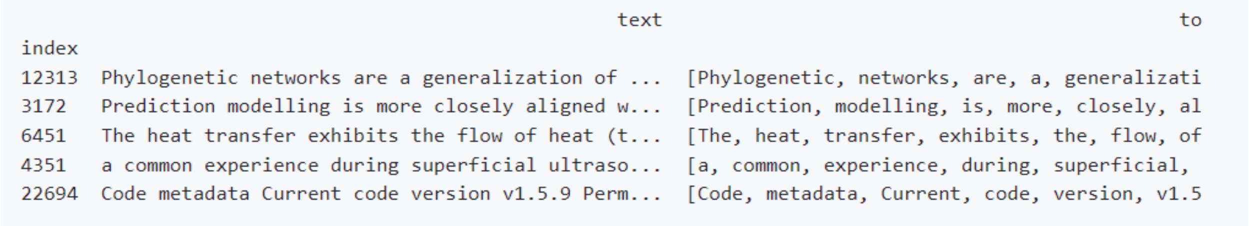
- annotations – a list of triples [[start_id, end_id, label_id]] where ids indicate start and end token ids of a span, and label_id indicates its provenance ('human', 'NLTK_synonym_replacement', 'chatgpt', or 'summarized');
- token_label_ids – a list mapping each token from tokens with a corresponding label id (from 0 to 3), according to annotations

The development and test sets have the following columns: text and tokens



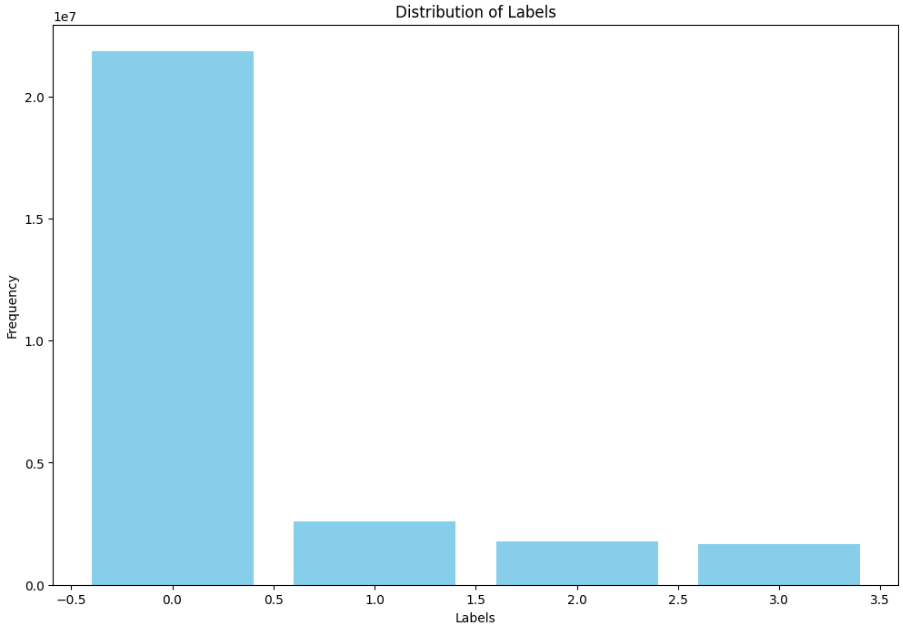
```
                                                        text                                   annotations
index
15096   Across the world, Emergency Departments are fa...      [[0, 3779, human], [3780, 7601, NLTK_sy
14428   lung Crab is the in the lead make of cancer-re...      [[0, 4166, NLTK_synonym_replacement], [


>>> train_df[["tokens", "token_label_ids"]].head(2)
                                                        tokens                             token_label_ids
index
15096   [Across, the, world,, Emergency, Departments, ...      [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
14428   [lung, Crab, is, the, in, the, lead, make, of,...      [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
```

Training Dataset



```
                                                      text                                  to
index
12313 Phylogenetic networks are a generalization of ...  [Phylogenetic, networks, are, a, generalizati
3172  Prediction modelling is more closely aligned w...  [Prediction, modelling, is, more, closely, al
6451  The heat transfer exhibits the flow of heat (t...  [The, heat, transfer, exhibits, the, flow, of
4351  a common experience during superficial ultraso...  [a, common, experience, during, superficial,
22694 Code metadata Current code version v1.5.9 Perm...  [Code, metadata, Current, code, version, v1.5
```

Dev/Test data



Distribution of labels on the training dataset

2. Current approach

To address the challenge of detecting AI-generated text in scientific papers, two solutions have been developed, each utilizing a transformer-based model The first solution employs SciBERT, a variant of the BERT model specifically optimized for scientific text. SciBERT is pre-trained on a large corpus of scientific papers, which enables it to capture the unique language and terminological nuances of scientific discourse. This model was fine-tuned for three epochs, adapting its pre-trained capabilities to more precisely perform the task of token level classification within the given dataset.

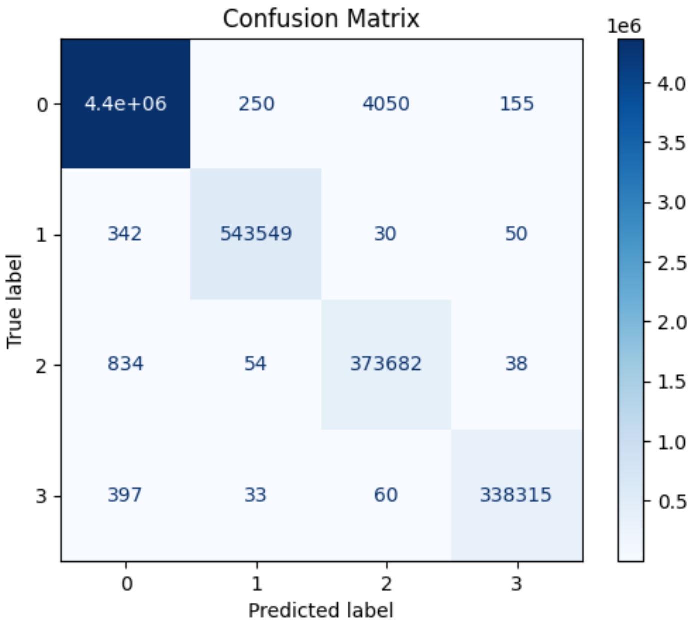
The second solution utilizes RoBERTa, a robustly optimized BERT architecture that has been shown to outperform the original BERT model on several NLP benchmarks. RoBERTa differs from BERT in several key areas: it is trained on a larger dataset and with more extensive training (longer and with bigger batches), and it removes the next-sentence prediction objective used in BERT, relying solely on masked language modeling for training. Like SciBERT, the RoBERTa model was fine-tuned for three epochs on the training dataset.

Both models underwent a preprocessing step involving tokenization and the alignment of tokens to ensure that the input data was in a suitable format for model training.
Initial Results

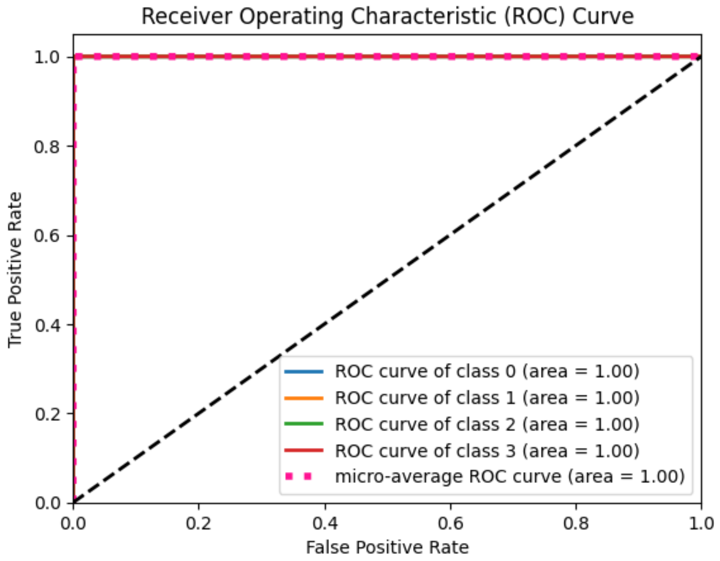
Model	train_loss	eval_f1	eval_loss	codabench_score
SciBERT	0.054	0.97	0.05	0.85
RoBERTa	**0.033**	**0.99**	**0.024**	**0.88**

In the evaluation of solutions for detecting AI-generated text in scientific papers, the RoBERTa model demonstrated superior performance over the SciBERT model. Specifically, RoBERTa achieved a CodaBench score of 0.88, outperforming SciBERT's score of 0.85.

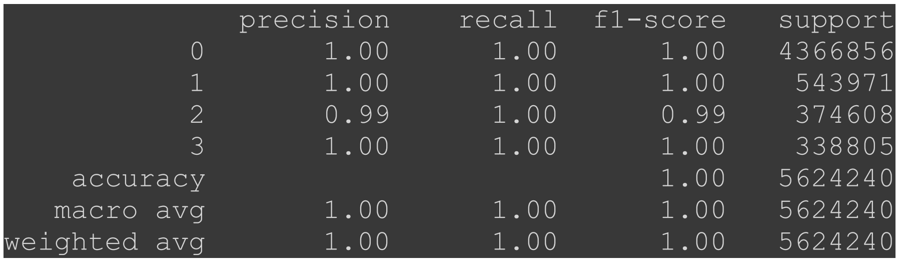
During the testing phase, I initially obtained a Codabench score of **0.88** on the test set using the baseline model. To improve the performance, I trained the RoBERTa model for **5** epochs and implemented a custom CrossEntropy Loss function that uses the inverse weights of the classes. This approach led to a significant reduction in training loss to **0.0233** and evaluation loss to 0.0035. Consequently, the f1-score improved substantially to **0.9977**.



Confusion Matrix



ROC Curve



```
              precision    recall  f1-score   support
           0       1.00      1.00      1.00   4366856
           1       1.00      1.00      1.00    543971
           2       0.99      1.00      0.99    374608
           3       1.00      1.00      1.00    338805
    accuracy                           1.00   5624240
   macro avg       1.00      1.00      1.00   5624240
weighted avg       1.00      1.00      1.00   5624240
```

4. Conclusions

Despite the promising results obtained during the training and validation phases, I was unable to test the latest improved RoBERTa model on the test set due to submission constraints. As a result, the final performance metrics on unseen data remain unverified. However, the substantial improvements in training and validation metrics suggest that the model is likely to perform well on the test set as well.

The solutions developed using SciBERT and RoBERTa have shown promising results in detecting AI-generated text in scientific papers. The RoBERTa model, in particular, demonstrated superior performance and was further improved by additional training and a custom loss function. While the final test set evaluation remains pending, the significant improvements in the F1-score and reduction in losses during training and validation phases indicate a strong potential for effective detection of AI-generated text in scientific literature.

5.References

SciBERT: A Pretrained Language Model for Scientific Text - https://arxiv.org/abs/1903.10676
RoBERTa: A Robustly Optimized BERT Pretraining Approach – https://arxiv.org/abs/1907.11692
NLRG at SemEval-2021 Task 5: Toxic Spans Detection Leveraging BERT-based Token Classification and Span Prediction Techniques - https://arxiv.org/pdf/2102.12254v2
https://huggingface.co/learn/nlp-course/chapter7/2