Empirical Methods in Natural Language Processing

Peking University, 2024

Homework 2: Due on Sunday, April 21 at 11:59 p.m.

Instructions

Please read these instructions to ensure you receive full credits on your homework.

- Submit your homework as a **zip** file through **Course**, which should include one report in PDF, your source code and prediction results on the test set.
- Any coding language is acceptable, but your code should be **your own**. **Do NOT** submit Jupyter or other notebooks, but the original source code only. We provide the code samples for evaluation in Python.
- You should write your report in **English** and submit it in PDF. We recommend using the official ACL style template for your report (https://github.com/acl-org/acl-style-files).
- Your code should be paired with a README file describing dependencies, code structures, etc.
- There is no need to submit the data you used and the model weights. Your grade will be based on the contents of the report and the source code.

Late submission policy

- Late homework will have 5% deducted from the final grade for each day late.
- **NO** submission will be accepted after April 28, a week after the due date. It is non-negotiable.

- Your submission time will be based on the time of your last submission to Course. Therefore, do NOT resubmit after midnight on the due date unless you are confident that the new submission is significantly better to overcompensate for the points lost.
- You can resubmit as much as you like, but each time you resubmit please be sure to upload all files you want graded!
- The number of points deducted will be rounded to the nearest integer.

Problem Description

In this homework, you will implement models for the Multidomain, Multimodel and Multilingual Machine-Generated Text Detection task (SemEval-2024 Task 8).

Large language models (LLMs) are becoming mainstream and easily accessible, ushering in an explosion of machine-generated content over various channels, such as news, social media, question-answering forums, educational, and even academic contexts. Recent LLMs, such as ChatGPT and GPT-4, generate remarkably fluent responses to a wide variety of user queries. The articulate nature of such generated texts makes LLMs attractive for replacing human labor in many scenarios. However, this has also resulted in concerns regarding their potential misuse, such as spreading misinformation and causing disruptions in the education system. Since humans perform only slightly better than chance when classifying machine-generated vs. human-written text, there is a need to develop automatic systems to identify machine-generated text with the goal of mitigating its potential misuse.

The original task contains three hierarchical subtasks:

Subtask A - Binary Human-Written vs. Machine-Generated Text Classification: Given a full text, determine whether it is human-written or machine-generated. **There are two tracks for subtask A**: monolingual (only

English sources) and multilingual.

Subtask B - Multi-Way Machine-Generated Text Classification: Given a full text, determine who generated it. It can be human-written or generated by a specific language model.

Subtask C - Human-Machine Mixed Text Detection: Given a mixed text, where the first part is human-written and the second part is machine-generated, determine the boundary, where the change occurs.

You will implement models on **subtasks A (just need to conduct experiment on monolingual track) &B** in this homework.

There is **no constraint** on the methods and data you use. You can implement your own model from scratch, finetune on pretrained language models, or use existing toolkits. **Please clearly describe the methods and data you use in the report.** Organizing your reports systematically is advisable, and for reference, you can structure your report as follows: begin your report with an abstract and an overarching introduction, followed by the introduction of your proposed method or model. Then in the experimental section, introduce the baseline models, present the experimental results of your model, and conduct a comprehensive analysis from various perspectives. Finally conclude the report with a conclusion and state your perspectives for the potential future development directions of this task. To support your argument, we highly recommend that you **utilize figures and tables more to demonstrate experiment results and analyses.**

NOTE: **We will not simply grade your homework based on the model performance, but mainly consider your completion of training, verifying and testing process during experiments. Therefore, it is not necessary to design overly complicated model structure.**

Data statistics

Subtask	**#Train**	**#Dev**	**#Test**
Subtask A (monolingual)	119,757	5,000	34,271
Subtask A (multilingual)	172,417	4,000	42,378
Subtask B	71,027	3,000	18,000
Subtask C	3,649	505	11,123

Useful links

Github (Including a detailed description of the task): https://github.com/mbzuai-nlp/SemEval2024-task8

Data link:

We have uploaded the data involved in this assignment to the PKU Netdisk, including train, validation and text datasets for three subtasks:

https://disk.pku.edu.cn/link/AA767A6AE905804946958E40B38DB2BA32

Related blogs and papers

1. A Comprehensive Guide to Understand and Implement Text Classification in Python (https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/)
2. Text classification with MLP (https://colab.research.google.com/drive/1WUy4G2SsoLelrZDkO2I0v9tHx9x27NJK?usp=sharing)
3. Text Classification with BERT (https://www.sabrepc.com/blog/Deep-Learning-and-AI/text-classification-with-bert)

4. A Survey on LLM-generated Text Detection: Necessity, Methods, and Future Directions (https://arxiv.org/abs/2310.14724)
5. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature (https://arxiv.org/abs/2301.11305)
6. RADAR: Robust AI-Text Detection via Adversarial Learning (NIPS 2023, https://arxiv.org/abs/2307.03838)

Contact

If you have any question about this homework, please email TA via lizhen63@pku.edu.cn .