Προτεινόμενο Θέμα Διπλωματικής Εργασίας

[Καθ. Ευστάθιος Σταματάτος, Πανεπιστήμιο Αιγαίου]

Machine Generated Text Detection and Attribution

Abstract:

Recently, significant advances in generative language models (e.g., ChatGPT) have increased the users of such tools and facilitated the generation of large volumes of text for a wide variety of applications. Despite the great potential of those models to improve the performance of AI systems in several tasks, there are certain risks of abusing that technology. The development of tools that can automatically detect machine generated text could serve as a key countermeasure for mitigating the risks of abuse of powerful modern language models[1][2].

This thesis aims at studying the state-of-the-art in machine generated text detection and developing new, efficient, and effective methods that can distinguish machine generated text from human generated text. In addition, in case of machine generated texts, the specific model that was used to obtain it should be identified[3][4].

The deliverables include:

- A comprehensive survey of state-of-the-art in machine generated text detection
- A novel method able to detect machine generated text and attribute it to a certain language model. An experimental study demonstrating the efficiency and effectiveness of the new method.
- Final MSc Thesis document

Required background and skills: Python programming, Machine learning, Natural language processing
Other requirements: All documents must be written in English.

[1] Crothers, E., et al. (2022). Machine Generated Text: A Comprehensive Survey of Threat Models and Detection Methods, https://arxiv.org/abs/2210.07321

[2] Hoang-Quoc Nguyen-Son, et al. (2017). Identifying Computer-generated Text Using Statistical Analysis. In *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*

[3] Adaku Uchendu, et al. (2020). Authorship Attribution for Neural Text Generation. In *Proc. of Empirical Natural Language Processing Conference*.

[4] Shaoor Munir, et al. (2021). Through the Looking Glass: Learning to Attribute Synthetic Text Generated by Language Models. In *Proc. of Conference of the European Chapter of ACL*.