ASSESSING THE QUALITY OF AI METHODS FOR FAKE NEWS DETECTION

Supervisor(s) name(s): Vinicius Woloszyn
Contact email: woloszyn@tu-berlin.de
Minimum number of participants: 4
Maximum number of participants: 8
When and where the weekly meetings will take place:
Day of the week: Tue
Time from: 14:00
Time until: 15:00
Building: Virtual
Project language: English
Project Requirements (Desired Student Skills):

- Strong Knowledge of Python
- Knowledge with Machine Learning

Avaliation: Each group will write a paper (max 8 pages) reporting the findings, which will be later uploaded at https://arxiv.org/.

Project description: The spreading of disinformation throughout the web has become a critical problem for a democratic society. The dissemination of fake news has become a profitable business and a common practice among politicians and content producers. A recent study entitled 'Regulating disinformation with artificial intelligence', examines the trade-offs involved in using automated technology to limit the spread of disinformation online. Although AI and Natural Language Generation have evolved so much in the last decade, there are still few shortcomings that must be better understood for a stronger solution. The students will dive deeper into Natural Language Processing; therefore a strong knowledge of Python and AI is necessary.

1. **GROUP A (max 4 students): How vulnerable is fake news detection to attacks?**

There is a concern regarding the potential implications of using AI for the automatic moderation of content. One problem is that automatic moderation of content on social networks will accelerate a race where AI will be created to counter-attack AI.

Adversarial machine learning is an AI technique that attempts to fool models by exploiting vulnerabilities and compromising the results. For example, by changing particular words - e.g., "Barack" (Obama)-> "b4r4ck" - it is possible to mislead classifiers and overpass automatic detection filters.

Recent works (cite) show the state-of-art machine learning models are vulnerable to these attacks. This study will use state-of-art tools to attack and dive deep into understanding Fake News Detection vulnerabilities. The goal is to experiment with a state-of-the-art framework for adversarial attacks to discover and compute automatic fake news detection vulnerabilities to adversarial attacks.

2. **GROUP B (max 4 students) How fast models become outdated?**

Fake news is dynamic and has been created every day, and an AI trained with outdated knowledge cannot predict with the same performance current topics. One of the sources for the training of FN detection is Fact-checkers. However, Fact-checkers conduct their investigations after (*post hoc*) a FN is published and disseminated, therefore the training of AI is always slower than the generation of fake news.

The generalization of models is a critical issue in machine learning and refers to the model's ability to adapt appropriately to new, previously unseen data.

A. **Does artificial data improve the performance of fake news detection?**

Data augmentation is a technique used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data. It acts as a regularizer and helps increase classifiers' performance and reduce overfitting when training a machine learning model. Does it help to improve the performance of fake news detection?

B. **Does American Fake News help to detect European fake news?**

Transfer learning (TL) is a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related question. For example, the knowledge gained while learning to recognize fake news in English could be used when trying to identify fake news in German.

MAIN REFERENCES:

- Dive into Deep Learning, https://d2l.ai/d2l-en.pdf
- Speech and Language Processing, https://web.stanford.edu/~jurafsky/slp3/ed3book_dec302020.pdf

COMPLEMENTAR REFERENCES:

- https://github.com/QData/TextAttack
- https://github.com/flairNLP/flair
- https://huggingface.co/
- Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
- Woloszyn, Vinicius, et al. "Untrue. News: A New Search Engine For Fake Stories." arXiv preprint arXiv:2002.06585 (2020).
- Zhou, Zhixuan, et al. "Fake news detection via NLP is vulnerable to adversarial attacks." *arXiv preprint arXiv:1901.09657* (2019).
- Sinha, Abhishek, et al. "Negative Data Augmentation." *arXiv preprint arXiv:2102.05113* (2021).
- Morris, John, et al. "TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020