# Doctoral School of Information and Biomedical Technologies
# Polish Academy of Sciences (TIB PAN)

**SUBJECT: Adversarial Robustness of Neural Text Classifiers: Exploring Languages, Sources and Evaluation**

**SUPERVISORS: dr hab. Maciej Ogrodniczuk, dr inż. Piotr Przybyła; Polish Academy of Sciences**

**DESCRIPTION:**

Automatic text classifiers have shown their usefulness in many tasks related to assessing content credibility [1]. This is especially true for user-generated text in online platforms, such as social media, where automatic classifiers help to perform content moderation at scale [2]. Unfortunately, models implemented using large neural networks have been shown to be vulnerable to **adversarial examples**, i.e. malicious modification of the input data that preserve the original meaning, but elicit an erroneous decision from the filtering model [3].

Recent research has clearly demonstrated that this danger exists and, indeed, detectors of content such as fake news, rumours or propaganda can be confused through small modifications [4]. Further efforts have shown additional attack techniques, such as using reinforcement learning [5] or LLM generations [6]. However, the dynamic nature of the phenomenon and its social importance motivate expanding the research in at least three directions:

**1. Multilinguality**

Virtually all of the previous research in the area was performed for text in English. However, other languages are also carriers of misleading information – in fact, the weaknesses of automatic moderation for them has been associated with negative consequences [7]. Extending the previous experiments into further languages would require solving many problems, such as obtaining good-quality classification benchmarks (e.g. for credibility assessment), finding or building basic elements of the workflow that support the language (esp. LLMs) or adjusting existing attack methods to the properties of the target languages (e.g. homoglyphs for richer scriptures).

**2. AI sources of low-credibility text**

Previous work involves human-written content, both of the credible and non-credible category. However, modern information ecosystem includes a wealth of machine-generated text [8], especially in low-credibility genres. Including this source of text in the robustness analysis would require (1) detecting machine-generated text, a valid task in itself [9] with uncertain robustness, and (2) checking victims' vulnerability to machine-rewritten text.

**3. Semantic evaluation**

Finally, the evaluation of adversarial attacks plays a crucial role. In particular, if input text is modified so heavily that its meaning is not preserved, the attack loses its purpose, even if the classifier's decision is changed. Therefore, performing precise evaluation of the meaning change is a crucial step. Previous automatic metrics have been shown to have low alignment

with human judgement. Therefore, new techniques need to be explored, of which LLM-as-a-judge [11] can be a promising direction.

**REQUIREMENTS:**

- Master's degree in computer science, linguistics, mathematics or related domains (completed before the start of the programme),
- Theoretical knowledge and practical experience in natural language processing and machine learning,
- High fluency in English, both written and spoken,
- The ability to work independently, both coming up with scientific ideas and managing time and effort to put them into practice,
- Previous experience in research work will be welcome.

Candidate should contact the authors of the proposal before formal submission of documents.

**BIBLIOGRAPHY:**


[1] Ciampaglia, G. L., Mantzarlis, A., Maus, G., & Menczer, F. (2018). Research Challenges of Digital Misinformation: Toward a Trustworthy Web. *AI Magazine*, *39*(1), 65. https://doi.org/10.1609/aimag.v39i1.2783

[2] Singhal, M., Ling, C., Paudel, P., Thota, P., Kumarswamy, N., Stringhini, G., & Nilizadeh, S. (2022). SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice. *The 8th IEEE European Symposium on Security and Privacy (EuroS&P 2023)*. https://doi.org/10.48550/arxiv.2206.14855

[3] Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2020). Adversarial Attacks on Deep-learning Models in Natural Language Processing. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *11*(3). https://doi.org/10.1145/3374217

[4] Przybyła, P., Shvets, A., & Saggion, H. (2024). Verifying the robustness of automatic credibility assessment. *Natural Language Processing*, *31*(5), 1134–1162. https://doi.org/10.1017/nlp.2024.54

[5] Przybyła, P., McGill, E., & Saggion, H. (2024). Know Thine Enemy: Adaptive Attacks on Misinformation Detection Using Reinforcement Learning. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis* (pp. 125–140). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wassa-1.11

[6] Przybyła, P., McGill, E., & Saggion, H. (2025). Attacking Misinformation Detection Using Adversarial Examples Generated by Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 27614–27630). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1405

[7] Nourooz Pour, H. (2023). Transitional justice and online social platforms: Facebook and the Rohingya genocide, *International Journal of Law and Information Technology*, Volume 31, Issue 2, Pages 95–113, https://doi.org/10.1093/ijlit/eaad016

[8] Hanley, H. W. A., & Durumeric, Z. (2024). Machine-Made Media: Monitoring the Mobilization of Machine-Generated Articles on Misinformation and Mainstream News Websites. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1), 542–556. https://doi.org/10.1609/icwsm.v18i1.31333

[9] Ahmad, Z., Torres-Ruiz, M., Mahmood, A., Quintero, R., Ameer, I., and Bölücü, N. (2026) Human or Machine? A Survey on Machine-Generated Text Detection, in *IEEE Access*, vol. 14, pp. 34113-34136, doi: 10.1109/ACCESS.2026.3666781

[10] Przybyła, P., Wu, B., Shvets, A., Mu, Y., Sheang, K. C., Song, X., & Saggion, H. (2024). Overview of the CLEF-2024 CheckThat! Lab Task 6 on Robustness of Credibility Assessment with Adversarial Examples (InCrediblAE). In G. Faggioli, N. Ferro, P. Galuščáková, & A. García Seco de Herrera (Eds.), *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*. https://ceur-ws.org/Vol-3740/paper-28.pdf

[11] Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., Shu, K., Cheng, L. and Liu, H. (2025). From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.