PolEval 2025

Łukasz Kobyliński[1], **Ryszard Staruch**[2], **Alina Wróblewska**[1], **and Maciej Ogrodniczuk**[1]

[1]Instutute of Computer Science PAS, Warsaw
[2]Adam Mickiewicz University, Poznań
l.kobylinski@ipipan.waw.pl, alina@ipipan.waw.pl
ryszard.staruch@amu.edu.pl, m.ogrodniczuk@ipipan.waw.pl

Abstract

PolEval is an annual shared-task evaluation campaign dedicated to advancing natural language processing for the Polish language. This paper presents an overview of PolEval 2025, the eighth edition of the campaign, which included three completed tasks covering machine-generated text detection, gender-inclusive language generation, and speech emotion recognition. The evaluation was conducted using standardized datasets and metrics via the AmuEval platform. PolEval 2025 attracted 15 teams and over 100 submissions, demonstrating continued engagement from the Polish NLP community. We describe the organization of the campaign, the evaluation setup, and the role of PolEval in fostering reproducible research and community-driven benchmarking.

1 Introduction

PolEval is an annual evaluation campaign, inspired by SemEval, dedicated to benchmarking and advancing natural language processing (NLP) tools for the Polish language. It provides a shared-task framework in which participating teams submit systems to solve a variety of pre-defined tasks using standardised datasets, with their performance evaluated under official protocols. Over the years PolEval has covered multiple NLP domains – from machine translation and named-entity linking to speech recognition, sentiment analysis, and more.

The first edition of PolEval was held in 2017 at the 8th Language & Technology Conference (LTC) in Poznań.[1] It did not have separate proceedings but 10 papers resulting from 2 tasks on part-of-speech tagging (Kobyliński and Ogrodniczuk, 2017) and sentiment analysis (Wawer and Ogrodniczuk, 2017) were included in the LTC proceedings (Vetulani and Paroubek, 2017). In the subsequent years, PolEval results were presented at two other conferences such as AI & NLP Day (Ogrodniczuk and Kobyliński, 2018, 2019, 2020, 2021) and FedCSIS (Kobyliński et al., 2023), and at the Natural Language Processing seminar[2] (Ogrodniczuk and Kobyliński, 2024) and its proceedings were usually published in a separate series.

2 PolEval 2025

In the 8th edition of the campaign – PolEval 2025 – four tasks were selected from the proposals submitted between March and May 2025 and subsequently announced publicly (see Section 2.2). Training data were released in August 2025, followed by the release of the test data in September 2025. An information and dissemination campaign was conducted on social media and at major AI-related events, including FedCSIS 2025 Conference in Cracow and AI & NLP Day jointly organized with Confitura in Warsaw. The evaluation was carried out using the PolEval benchmarking system (see Section 2.3), and the results were announced through the integrated leaderboards on November 17, 2025. The award ceremony and presentation of the winning solution took place at the Data Science Summit Conference 2025.

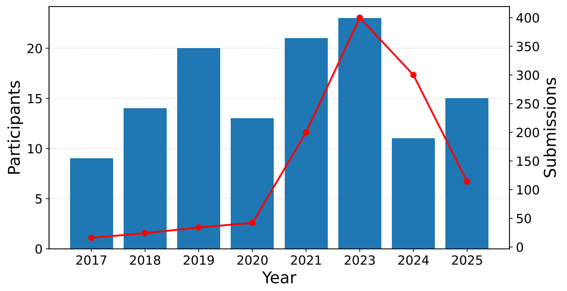

Figure 1: Yearly Numbers of Participants (left axis) and Submissions (right axis).

The 2025 edition of PolEval received over 100

[1] https://ltc.amu.edu.pl/

[2] https://zil.ipipan.waw.pl/seminar

Proceedings of the PolEval 2025 Workshop, pages 1–4
November 21, 2025 ©2025 Association for Computational Linguistics

submissions from 15 participating teams, which is comparable to previous years (see Figure 1).

2.1 Program Committee

General Chairs

Łukasz Kobyliński, ICS PAS,[3] Sages
Maciej Ogrodniczuk, ICS PAS

Evaluation Platform Administrator

Ryszard Staruch, AMU[4]

Publication Chair

Alina Wróblewska, ICS PAS

Publicity and Social Media Chair

Beata Milewicz, ICS PAS

Task Organisers

Iwona Christop, AMU
Maciej Czajka, AMU
Łukasz Kobyliński, ISC PAS
Piotr Przybyła, Universitat Pompeu Fabra, Barcelona
Jakub Strebeyko, University of Warsaw
Alina Wróblewska, ICS PAS
Aleksandra Zwierzchowska, ICS PAS

2.2 Tasks

Task 1: Śmigiel: Spotting Machine-Generated Text from Language Models for Polish
Śmigiel (Przybyła et al., 2025) is a shared evaluation task designed to benchmark the detection of AI-generated vs. human-written text in Polish. The task is formulated as a binary classification problem, and participants can choose among three subtasks: *unsupervised*, *constrained*, and *open*. The accompanying dataset (Strebeyko et al., 2025) includes human-written texts that are drawn from various sources – all distributed under open licenses, and represent diverse domains (reviews, literature, social media posts, Wikipedia articles, parliamentary transcripts, and news). It also includes corresponding machine-generated texts produced using a variety of open-source LLMs.

As LLMs become more widely used in Polish, there is growing need for reliable methods to distinguish between human- vs. LLM-generated texts. Śmigiel seeks to support the development of robust machine-generated text detectors for Polish, with potential applications in journalism, academia, media verification and related fields.

Task 2: Gender-inclusive LLMs for Polish
PolEval 2025 Task 2 (Wróblewska, 2025) aims to support development of LLMs capable of generating grammatically correct and gender-aware Polish across a variety of contexts. The submitted LLMs are evaluated on two subtasks: *gender-inclusive proofreading*, which involves transforming Polish texts written in standard (typically masculine-centric) language into gender-inclusive counterparts, and *gender-sensitive translation*, which consists in translating between Polish and English in both directions while ensuring that the Polish output respects gender-inclusive language norms. Participants are provided with the official dataset, the Inclusive Polish Instruction Set (IPIS, Wróblewska and Żuk, 2025).

By promoting the generation of gender-inclusive Polish, this task seeks to mitigate masculine bias present both in conventional usage and in LLMs trained on standard corpora. In doing so, it contributes to broader social goals related of gender equality, e.g., elimination of sexism from language.

Task 3: Polish Language Document Layout Detection The task has been cancelled due to lack of participants.

Task 4: Polish Speech Emotion Recognition Challenge Task 4 (Christop and Czajka, 2025b) focuses on speech emotion recognition (SER) for the Polish language, where participants are required to develop a system that classifies speech recordings into one of six emotion categories: *anger*, *fear*, *happiness*, *sadness*, *surprise*, and *neutral*. The training data consists of multilingual emotional speech samples from seven non-Polish languages sourced from CAMEO (Christop and Czajka, 2025a). A separate Polish validation set is provided solely for model selection. The final evaluation is conducted on previously unseen Polish speech data. System performance is assessed primarily using the macro-averaged F1 score, with accuracy reported as a secondary metric.

The task addresses the challenge of cross-lingual SER in a low-resource setting, as Polish lacks large, publicly available emotional speech corpora. By restricting access to Polish training data, the shared task promotes the development of language-independent acoustic representations that generalise across languages. This setting reflects realistic deployment conditions and supports research on robust, transferable SER systems, with potential applications in human-computer interaction and

[3] Institute of Compute Science, Polish Academy of Sciences, Warsaw

[4] Adam Mickiewicz University, Poznań

assistive technologies for Polish speakers.

2.3 Evaluation system

PolEval 2025 shared tasks were hosted on a dedicated instance of the AmuEval evaluation platform (Jassem et al., 2024). The platform was chosen for its simplicity, allowing participants to concentrate on solving the tasks rather than dealing with the technical complexities often associated with submitting predictions on evaluation platforms.

Each challenge included one primary evaluation metric that determined the final leaderboard standings, along with up to two additional metrics to provide participants with further insight into the performance of their solutions. Some of these metrics required custom implementation, as they were not available in the platform's default metric set. This integration process was smooth and problem-free, demonstrating that the platform can be effectively customized to the needs of dedicated instances.

Datasets for each task were made available via public repositories on GitHub. The platform itself was used exclusively for the final evaluation of system outputs on held-out test sets. Each task featured two separate test subsets: Test-A and Test-B, organized as distinct challenges on the platform. Metric scores for Test-A were visible to participants during the competition, while scores for Test-B remained hidden until after the submission deadline. This setup was designed to prevent overfitting and ensure a fair comparison of final systems.

The positive experience with AmuEval during PolEval 2025 demonstrates its potential as a reliable foundation for future shared tasks and domain-specific evaluation campaigns.

3 Conclusion and future plans

PolEval has, over multiple editions, established itself as a stable and widely recognized evaluation campaign for Polish natural language processing. The 2025 edition confirmed the continued relevance of shared-task benchmarking for the Polish NLP community, while also highlighting the rapidly evolving landscape of artificial intelligence, in particular the growing dominance of large language models (LLMs).

Looking ahead, one of the main challenges for future editions of PolEval will be to keep pace with these developments while preserving the campaign's core focus on the evaluation of Polish-language competence. As LLM-based systems increasingly achieve strong general performance across languages and tasks, there is a growing need for carefully designed, language-specific evaluations that test phenomena characteristic of Polish. Future PolEval tasks will therefore aim to balance openness to modern, large-scale models with evaluation settings that meaningfully differentiate systems based on their handling of Polish-specific linguistic and cultural properties.

Another important direction for the campaign is strengthening its educational role. We plan to further synchronize PolEval tasks and timelines with NLP- and AI-related university courses conducted in Poland, enabling students to participate as part of their coursework. Such integration has the potential to lower the entry barrier for new participants, foster practical skills in reproducible evaluation, and contribute to the training of the next generation of researchers and practitioners in Polish NLP.

Finally, increasing the visibility and impact of PolEval remains a key objective. Continued cooperation with national and international scientific conferences, workshops, and community events will help broaden the audience for the campaign, attract new participants, and facilitate the dissemination of results. By maintaining strong ties with both the research community and educational initiatives, PolEval aims to remain a relevant and adaptable platform for benchmarking Polish NLP in an era of rapidly advancing AI technologies.

Acknowledgments

We would like to express our sincere gratitude to the organisers of the PolEval 2025 evaluation campaign for their dedication and continuous efforts in supporting the Polish natural language processing community, particularly to the Institute of Computer Science, Polish Academy of Sciences (ICS PAS), for its leading organisational role. We also gratefully acknowledge the generous support of the campaign sponsors, especially Sages, whose contribution made the organisation of the event possible. We extend our sincere thanks to all participating teams and researchers for their enthusiasm, high-quality submissions, and valuable scientific contributions, which are the foundation of the success of the PolEval 2025 campaign. Finally, we thank the organisers of the associated scientific and community events – Data Science Summit, Confitura, the FedCSIS Conference, and the University of Warsaw – for providing the venues and infras-

tructure for dissemination, presentations, and the award ceremony.

References

Iwona Christop and Maciej Czajka. 2025a. Cameo: Collection of multilingual emotional speech corpora. *Preprint*, arXiv:2505.11051.

Iwona Christop and Maciej Czajka. 2025b. Polish Speech Emotion Recognition Challenge. In *Proceedings of the PolEval 2025 Workshop*.

Krzysztof Jassem, Andrzej Gajda, Mateusz Tylka, Ryszard Staruch, Grzegorz Lipiecki, and Szymon Bartanowicz. 2024. Amueval: A user-friendly educational platform for machine learning challenges. In *Proceedings of the Fortieth Information Systems Education Conference (ISECON 2024)*, pages 107–114, Virtual Conference, Chicago, IL, USA. Foundation for Information Technology Education. Virtual conference, October 19, 2024.

Łukasz Kobyliński and Maciej Ogrodniczuk. 2017. Results of the PolEval 2017 competition: Part-of-speech tagging shared task. In (Vetulani and Paroubek, 2017), pages 362–366.

Łukasz Kobyliński, Maciej Ogrodniczuk, Piotr Rybak, Piotr Przybyła, Piotr Pęzik, Agnieszka Mikołajczyk, Wojciech Janowski, Michał Marcińczuk, and Aleksander Smywiński-Pohl. 2023. PolEval 2022/23 challenge tasks and results. In *Proceedings of the 18th Conference on Computer Science and Intelligence Systems*, volume 35 of *Annals of Computer Science and Information Systems*, pages 1237–1244.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2018. *Proceedings of the PolEval 2018 Workshop*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2019. *Proceedings of the PolEval 2019 Workshop*. Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2020. *Proceedings of the PolEval 2020 Workshop*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2021. *Proceedings of the PolEval 2021 Workshop*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2024. *Proceedings of the PolEval 2024 Workshop*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Piotr Przybyła, Jakub Strebeyko, and Alina Wróblewska. 2025. PolEval 2025 Task 1 Śmigiel: Spotting Machine-Generated Text from LLMs for Polish. In *Proceedings of the PolEval 2025 Workshop*.

Jakub Strebeyko, Alina Wróblewska, and Piotr Przybyła. 2025. Śmigiel Dataset: Laying Foundations for Investigating Machine-Generated Text Detection in Polish. Unpublished.

Zygmunt Vetulani and Patrick Paroubek, editors. 2017. *Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu, Poznań, Poland.

Aleksander Wawer and Maciej Ogrodniczuk. 2017. Results of the PolEval 2017 competition: Sentiment Analysis shared task. In (Vetulani and Paroubek, 2017), pages 406–409.

Alina Wróblewska. 2025. PolEval 2025 Task 2: Gender-inclusive LLMs for Polish. In *Proceedings of the PolEval 2025 Workshop*.

Alina Wróblewska and Bartosz Żuk. 2025. Integrating gender inclusivity into large language models via instruction tuning. *Preprint*, arXiv:2508.18466.