Report on the outcomes of a Short-Term Scientific Mission[1]

Action number: CA21129

Grantee name: Dr Ljubisa Bojic

Details of the STSM

Title: Detecting Opinions, Values, and Authenticity in Multilingual AI-Generated Text

Start and end date: 22/09/2025 to 13/10/2025

Description of the work carried out during the STSM

From September 22 to October 13, 2025, I conducted a 22-day Short-Term Scientific Mission at the University of Tuzla in

Bosnia and Herzegovina, aligned with the objectives of COST Action CA21129, OPINION. I drew on my expertise in AI alignment, communication science, and bias analysis in machine-generated texts. This mission extended my prior work on sentiment analysis and large language models by exploring the authenticity of AI-generated fanfiction within the Harry Potter fandom, a domain rich with emotional and cultural depth. While OPINIONs capture the raw, real-world articulations of beliefs and attitudes in communicative acts, fiction extends this framework by weaving hypothetical narratives that probe the deeper societal, ethical, and imaginative ramifications of such viewpoints, thereby enriching theoretical models of opinion dynamics.

The research began with the creation of a specialized corpus tailored to the Harry Potter fandom. I collaborated with a team comprising Selma Veseljevic Jerkovic and Anela Mulahmetovic Ibrisimovic from the University of Tuzla, Mirjana Necak and Ana Lipij from the University of Belgrade, Vuk Vukovic from the University of Montenegro, and Anguelina Popova from the American University of Central Asia. Selma Veseljevic Jerkovic, Mirjana Necak, and Ana Lipij crafted a set of human-authored fanfiction stories to serve as training data for the AI models. These stories were based on four narrative prompts covering distinct Harry Potter eras, including the Marauders era, pre-Hogwarts, Hogwarts, and post-Second Wizarding War, as outlined in the guidelines. The prompts were designed to evoke narratives that preserved character authenticity, emotional resonance, and thematic fidelity to the Harry Potter canon while allowing for Alternate Universe variations. Using advanced language models such as GPT-4o and Grok, we generated approximately 400 fanfiction stories, each between 500 and 1000 words, in Bosnian, Serbian, and English, adhering to General or Teen and Up ratings per

[1] This report is submitted by the grantee to the Action MC for approval and for claiming payment of the awarded grant. The Grant Awarding Coordinator coordinates the evaluation of this report on behalf of the Action MC and instructs the GH for payment of the Grant.

the Archive of Our Own system. Selma Veseljevic Jerkovic, Mirjana Necak, and Ana Lipij wrote new human-authored stories based on the same prompts, which were combined with the AI-generated stories to form a corpus of 1200 stories.

The annotation phase focused on evaluating the authenticity of these texts, building on my previous work in discourse analysis and language model personality traits. Anela Mulahmetovic Ibrisimovic, Vuk Vukovic, Anguelina Popova, and I formed the annotation team, developing a framework to assess four key dimensions: consistency of character portrayal with canon or AU contexts, coherence of narrative structure within fandom-specific settings, authenticity of emotional expression, and fidelity to cultural and fandom-specific tones. We annotated a subset of 300 stories, comparing the AI-generated texts against the new stories written by Selma Veseljevic Jerkovic, Mirjana Necak, and Ana Lipij. Weekly calibration sessions resolved discrepancies, revealing that AI-generated stories often failed to capture subtle fandom elements, such as the Marauders' dynamic camaraderie or Severus Snape's complex motivations, which the human-authored stories portrayed with greater depth.

The progress meetings with the team ensured smooth collaboration, and we extended corpus generation by two days to address sourcing hurdles, without affecting the overall timeline. The partnership enriched my computational approach with the team's insights into regional linguistics and fan culture, laying a foundation for ongoing contributions to the OPINION network.

Description of the STSM main achievements and planned follow-up activities

This mission advanced the objectives of the OPINION Action by developing a robust corpus and analytical framework for assessing authenticity in AI-generated fanfiction, using the Harry Potter fandom as a case study. The research addressed critical gaps in AI's ability to produce authentic creative texts, contributing to the Action's mission of harmonizing approaches to multilingual and culturally sensitive text analysis. My expertise in AI ethics and futures studies helped with a forward-looking perspective, emphasizing the ethical implications for fan communities and creative industries.

A primary achievement was the creation of a high-quality corpus of 1200 Harry Potter fanfiction stories in Bosnian, Serbian, and English, comprising AI-generated texts and new human-authored stories by Selma Veseljevic Jerkovic, Mirjana Necak, and Ana Lipij. The corpus will be deposited in the OSF repository. This dataset captures variations in character authenticity, narrative style, and emotional resonance, revealing that AI-generated stories often lack the character dynamics and fandom-specific tropes that the human authors portrayed with finesse. This finding aligns with my earlier research on cultural biases in AI-generated content. The annotation framework, validated at 0.80 intercoder reliability (Krippendorff's alpha) by Anela Mulahmetovic Ibrisimovic, Vuk Vukovic, Anguelina Popova, and myself, provides a standardized tool for evaluating authenticity across character consistency, narrative coherence, emotional depth, and cultural fidelity. Shared during the mission's closing meeting, this framework is adaptable to other fandoms and low-resource languages, supporting the Action's goal of methodological interoperability.

The semi-automated pipeline represents a technical milestone, achieving 76 percent accuracy in detecting authenticity markers and identifying both universal signals, such as sentiment polarity, and fandom-specific elements, like the use of "Felix Felicis." Documented in a reproducible Jupyter notebook, this pipeline paves the way for scalable analysis and informs future model fine-tuning. The training of Anela Mulahmetovic Ibrisimovic in advanced annotation and pipeline deployment, alongside the contributions of Selma Veseljevic Jerkovic, Mirjana Necak, and Ana Lipij in story creation, enhanced regional expertise, aligning with OPINION's commitment to inclusivity and capacity building in the Western Balkans.

I will collaborate with Selma Veseljevic Jerkovic, Anela Mulahmetovic Ibrisimovic, Mirjana Necak, Ana Lipij, Vuk Vukovic, and Anguelina Popova to co-author a manuscript for submission in a scientific journal by February 2026. This open-access publication will detail the corpus, framework, and pipeline, highlighting AI's authenticity gaps in fanfiction and advancing OPINION's publication goals. Our further joint work will expand the corpus to include languages like Croatian and refine the pipeline toward 80 percent accuracy using transfer learning, drawing on my expertise in language model alignment. Bi-monthly virtual working groups through OPINION's Working Group 4 will oversee corpus expansion and framework adaptations for other fandoms. Integration with the TWoN Horizon project, where I serve as coordinator, will explore AI-generated fanfiction in social media contexts, potentially yielding shared datasets. These efforts will sustain the mission's impact, embedding its innovations into broader networks for ethical AI development in creative domains.