Fellow	Nicolás Benjamín Ocampo
Host Organisation	Centrum Wiskunde & Informatica (CWI)
Scientific coordinator	Laura Hollink

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

During my fellowship, I worked under the supervision of Davide Ceolin, Senior Researcher and research group leader of the Human-Centered Data Analytics (HCDA) at CWI. We also collaborated with Tommaso Caselli, Senior Assistant Professor at the University of Groningen, member of the GroNLP group.

My research started by identifying representative data at the intersection of hate speech and misinformation. In the initial phase, I analyzed datasets from both domains to identify their overlap. On one hand, current hate speech datasets either contain short, explicit messages that rely on slurs rather than meaningful claims, or contain implicit hate, consisting of very brief statements lacking supporting reasoning. On the other hand, misinformation datasets are designed to focus on claims relevant to the public that allow for reasoning to determine their veracity, but rarely include hateful content, highlighting a gap between the two areas.

To address this gap, we created WSF-ARG+, a hate speech dataset whose messages rely on one or several claims and represent real user behavior from Stormfront, the White Supremacy Forum (WSF) [1]. Claims in all these messages were identified and annotated for check-worthiness. Check-worthiness is defined by (i.) the presence of a verifiable claim and (ii.) its potential public relevance or impact. Check-worthy claims are those likely to attract attention or influence public discourse. To annotate for check-worthiness the claims of WSF-ARG+, we proposed and validated an LLM-in-the-loop framework to ensure high-quality annotations while mitigating known limitations when using LLM-as-annotators [2]. This resulted in:

- A validated LLM-in-the-loop framework that reduces human effort while achieving full human-level quality and addressing LLM-hacking limitations.
- WSF-ARG+, a hate speech dataset including claim-level check-worthiness annotations from both LLM-in-the-loop (gold) and fully human (platinum) labeling.

Investigating the intersection between hate speech and check-worthiness in WSF-ARG+, we show that messages containing check-worthy claims are more harmful and hateful and that incorporating check-worthiness labels improves hate speech detection up to 0.213 macro-F1 and to 0.154 macro-F1 on average for large models.

This work has been submitted and is under review (preprint: https://arxiv.org/pdf/2603.25269). WSF-ARG+ is publicly available under Creative Commons Attribution-ShareAlike 3.0 License.

From these findings, we proceeded to work in two directions: First, we aim to determine whether check-worthy claims in WSF-ARG+ have already been fact-checked using the Google Fact Checking API (https://developers.google.com/fact-check/tools/api), in order to assign veracity labels. These fact-checked claims will serve to generate counterspeech with targeted “attacking strategies,” focusing on true, false, and check-worthy claims, evaluated both automatically and by humans. This work continued

during the second phase of my ERCIM fellowship, in collaboration with Stefano Cresci and Marinella Petrocchi (IIT-CNR), whom I met during the Research Exchange Program (REP) aiming to be submitted by the end of May, 2026 through the ARR peer-reviewing system (https://aclrollingreview.org/).

Second, instead of choosing one model to assess whether a claim is check-worthy or not, we apply clustering to analyze the relationship between several LLMs' check-worthiness predictions and platinum annotations. This work is being prepared for submission to a Q1 journal and is expected to be submitted by the end of April 2026.

Finally, the ERCIM fellowship led to continued collaboration with Davide Ceolin and Tommaso Caselli, resulting in a one-year postdoctoral position at CWI within the HCDA group to further develop this line of research.

---

[1] Helena Bonaldi, Greta Damo, Nicolás Benjamín Ocampo, Elena Cabrio, Serena Villata, and Marco Guerini. 2024. *Is safer better? the impact of guardrails on the argumentative strength of LLMs in hate speech countering*. In Proceedings of EMNLP, pages 3446–3463. ACL

[2] Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. *Large language model hacking: Quantifying the hidden risks of using llms for text annotation*. Preprint, arXiv:2509.08825.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

Published Papers:

- Damo, G., & **Ocampo, N. B.** (2026). *HateItOff at MultiPRIDE: Linguistic and sentiment cues in reclaimed LGBTQ+ slur detection*. In *EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*. CEUR-WS. https://apa.dipsco.unitn.it/evalita2026/53.pdf
- **Ocampo, N. B.,** Cabrio E., Villata S. (2025). *From Hidden to Harmful: Connecting Implicit and Explicit Hate Through Implied Statements.* The 24th IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology.

Preprint (under review):

- ***Ocampo, N. B****., Caselli, T., & Ceolin, D. (2026). When hate meets facts: LLMs-in-the-loop for check-worthiness detection in hate speech. arXiv. https://arxiv.org/abs/2603.25269*

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

Presented Seminars:

- *(2025, June 6). Where lies meet hate: Combating harmful misinformation online.* 3IA Doctoral & Postdoctoral Seminar Series, Inria Côte d'Azur, France.

- *(2026, March 3). When hate meets facts: LLMs-in-the-loop for check-worthiness detection in hate speech.* CNR Pisa, Italy.
- *(2026, March 13). When hate meets facts: LLMs-in-the-loop for check-worthiness detection in hate speech.* Inria Côte d'Azur, France.

Workshop Presentations:

- *(2026, February 26-27). HateItOff at MultiPRIDE: Linguistic and sentiment cues in reclaimed LGBTQ+ slur detection.* EVALITA 2026 Conference, Bari, Italy.

Attended Conferences:

- *(2025, November 15-18).* The 24th IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology.

Attended Symposium:

- *(2025, May 9). Truth is in the eyes of the machines – Symposium.* Centrum Wiskunde & Informatica. Amsterdam, Netherlands.

IV – RESEARCH EXCHANGE PROGRAMME (REP)

From 2–6 June 2025, I undertook a research visit to the MARIANNE team at INRIA Côte d'Azur. The main objectives of this mobility were to present the initial stages of my research and to strengthen the collaboration between the HCDA group at CWI and the MARIANNE team. During this visit, I met Serena Villata, head of the MARIANNE group (https://team.inria.fr/marianne/) and Research Director at the 3IA Côte d'Azur institute. She provided me with the opportunity to present my work as part of the Doctoral and Postdoctoral Seminar series (https://3ia.univ-cotedazur.eu/research/doctoral-postdoctoral-seminar-1). This visit was highly valuable, as it allowed me to receive constructive feedback on the first two months of my ERCIM fellowship.

From 2–6 March 2025, I also carried out a second research exchange, visiting two teams at the *"Istituto di Informatica e Telematica"* at CNR in Pisa, Italy. I collaborated with Stefano Cresci from the *"Cyber Intelligence"* group (https://cyb.iit.cnr.it/) and Marinella Petrocchi from the *"Trust, Security and Privacy"* group (https://tsp.iit.cnr.it/). Marinella's research focuses on countering online misinformation at the intersection of cybersecurity and data science, while Stefano has extensive experience in developing content moderation strategies for hate speech, misinformation, and propaganda on social media. During the visit, I presented my work to both teams and received highly valuable feedback from experts working on Misinformation and Hate Speech. The primary objective of this exchange was to establish a collaboration on counterspeech generation towards hate and to jointly produce a research paper.