Selected Papers of #AoIR2024:
The 25th Annual Conference of the
Association of Internet Researchers
Sheffield, UK / 30 Oct - 2 Nov 2024

Panel

TRANSFORMATIVE TOOLS, EMERGING CHALLENGES: EMPIRICAL AND PRACTICAL EXPERIENCES WITH LARGE LANGUAGE MODELS FOR TEXT CLASSIFICATION AND ANNOTATION IN COMMUNICATION STUDIES

Tariq Choucair
Queensland University of Technology

Ahrabhi Kathirgamalingam
University of Vienna

Fabienne Lind
University of Vienna

Jana Bernhard-Harrer
University of Vienna

Hajo G. Boomgaarden
University of Vienna

Bruna Silveira de Oliveira
Federal University of Minas Gerais

Rousiley Maia
Federal University of Minas Gerais

Sebastian Svegaard
Queensland University of Technology

Suggested Citation (APA): Choucair, T., Kathirgamalingam, A., Lind, F., Bernhard-Harrer, J., Boomgaarden, H. G., Silveira de Oliveira, B., Maia, R., Svegaard, S., O'Connor Farfan, K., Esau, K., Lubicz-Zaorski, C., Vodden, L., Bruns, A., Meyer, H., Puschmann, C., Brüggemann, M., Giglietto, F., Rossi, L., Righetti, N., & Marino, G. (2024, October). *Transformative Tools, Emerging Challenges: Empirical and Practical Experiences with Large Language Models for Text Classification and Annotation in Communication Studies*. Panel presented at AoIR2024: The 25th Annual Conference of the Association of Internet Researchers. Sheffield, UK: AoIR. Retrieved from http://spir.aoir.org.

Kate O'Connor Farfan
Queensland University of Technology

Katharina Esau
Queensland University of Technology

Carly Lubicz-Zaorski
Queensland University of Technology

Laura Vodden
Queensland University of Technology

Axel Bruns
Queensland University of Technology

Hendrik Meyer
University of Hamburg

Cornelius Puschmann
University of Bremen

Michael Brüggemann
University of Hamburg

Fabio Giglietto
University of Urbino

Luca Rossi
IT University of Copenhagen

Nicola Righetti
University of Urbino

Giada Marino
University of Urbino

Panel Introduction

Recent advancements in Large Language Models (LLMs) have opened significant research opportunities within the field of media and communication studies. LLMs offer the capacity to conduct large-scale content classification and annotation with low computational expertise and reduced manual coding efforts, potentially allowing more possibilities for researchers in social sciences to explore understudied topics (Bail, 2023; Chang et al., 2024). Because of its functioning and vast training in distinct domains and languages, LLMS also potentially unlocks more generalizable, complex, and diverse analyses across various communication materials compared to previous computational tools and approaches (Chang et al., 2024). These materials encompass a wide spectrum, ranging from journalistic content to the digital discourse of political actors and social

media conversation threads. At the same time, LLMs also raise important concerns with potential biases, data privacy, models' transparency, environmental impact, and power imbalances (Jameel et al., 2020; Fecher et al., 2023). Despite the increasing discussion around these models, there is a clear need for more dialogue that bridges empirical research and in-depth elaborations specifically for media and communication scholars (Gil de Zúñiga et al., 2024; Guzman and Lewis, 2020).

Our panel assembles a collection of diverse studies that harness LLMs to tackle text classification and annotation tasks related to media and communication problems, issues, and topics. These research papers engage in an exploration of: (a) pipeline structuring: diverse methodologies for structuring effective pipelines tailored to this form of analysis; (b) tools and models comparison: comparisons of the various LLMs tools and models available for text classification and annotation, highlighting their strengths and weaknesses; (c) optimal variables and tasks: identifying the variables and tasks where LLMs demonstrates exceptional performance and reliability; (d) limitations: discussions on the existing limitations of these tools, including limitations related to specific tasks, variables, languages and data formats; (e) prompt development: strategies for developing, adapting and adjusting prompts that allows better results for specific tasks; and (e) ethical and political dimensions: an examination of the ethical and political considerations inherent in the deployment of LLMs in communication research.

Ahrabhi Kathirgamalingam and colleagues (Paper 1) examines how biases in coding decisions arise from both human coders and Large Language Models (LLMs) when analyzing racism in news media. They investigate the influence of human coders' lived experiences and awareness of discrimination on their annotations, the role of persona assignments in shaping LLM coding biases and compares the biases of human coders and LLMs to understand differences in annotation decisions and the effect of text properties. Findings reveal systematic variation in human annotations linked to lived experience and awareness, as well as significant impacts of persona assignments on LLM outputs. The study emphasizes the importance of accepting systematic disagreement in annotations and offers recommendations to enhance the validity of both manual and automated analyses of constructs of marginalization, aiming to improve discrimination research and inform policies for equity.

Bruna Silveira de Oliveira and colleagues (Paper 2) investigate the potential of Large Language Models (LLMs) to analyze Brazilian masculinist ("manosphere") podcasts, a digital media space used by extremist groups to propagate misogynist ideologies. Drawing from a sample of 2,490 episodes, the research focuses on the interplay between legitimacy, intolerance, and recognition within these podcasts. The methodology includes automated transcription, speaker diarization, and the creation of a nuanced codebook covering variables like "presence of intolerance" and "object of intolerance." A reliability test validated the codebook, and LLMs were iteratively refined through prompt adjustments to improve agreement with human coders. Preliminary findings highlight the need for detailed prompts to enable LLMs to capture both explicit and implicit manifestations of intolerance, such as the denial or minimization of oppression. The study then shows the importance of combining human expertise and machine automation for analyzing sensitive extremist narratives - offering scalable

methods while mitigating the mental health risks for researchers exposed to toxic content.

Myself (Tariq Choucair) and colleagues (Paper 3) focused on stance detection across languages and platforms. By leveraging LLMs' training across diverse linguistic and contextual domains, the research evaluates their ability to generalize without language-specific training data. Two case studies are conducted: one analyzing multilingual election campaigns (in Brazilian Portuguese, Australian English, Danish, and Peruvian Spanish) and another examining platform-specific text variations in Australian Voice to Parliament referendum discussions across Facebook, Instagram, Twitter, and YouTube. Preliminary findings reveal that fine-tuning LLMs significantly improve their performance in the source language (English), with models achieving notable F1 score improvements. Cross-lingual tests show that larger models like GPT-4o maintain strong adaptability (e.g., high F1 scores in Portuguese), whereas smaller models like Mistral 7b underperform in cross-lingual contexts. However, smaller models demonstrate competitive performance in specific tasks within the source language post-fine-tuning, indicating that model size is not always the determinant of success.

Hendrik Meyer and colleagues (Paper 4) use LLMs for detect stances in journalistic coverage of climate protests, focusing on the movements "Last Generation" (LG) and "Fridays for Future" (FFF) in Germany, which employ differing protest strategies. Using a dataset of ~12,000 German-language news articles, the research evaluates the validity of zero-shot LLM classifications (e.g., GPT-4) in identifying stances, compares their performance to human coders, and explores how media portrayals differ for the two movements. Preliminary findings reveal that LLMs achieve strong alignment with human coders. Ethical considerations, including the challenges of detecting stances and the accessibility of larger versus smaller models, are discussed as the researchers develop a stance classifier using a smaller model, marking progress in scalable, nuanced and ethical analysis of politically charged media coverage.

Fabio Giglietto and colleagues (Paper 5) concludes the panel by reflecting on their investigation of the role of social media, particularly Facebook, in shaping exposure to and engagement with political news during the 2018 and 2022 Italian elections, employing large language models (LLMs) to address key methodological challenges. It examines 84,874 URLs shared during these periods, categorizing them into political and non-political content through a fine-tuned binary classifier. Using text embeddings and k-means clustering, the study groups and labels political URLs, with human coders assessing cluster validity. Giglietto and colleagues call attention to three primary challenges of LLM-based methodologies: the Swiss Army Knife Dilemma (balancing general-purpose flexibility with task-specific validation), the Granularity Spectrum Problem (managing the variability in clustering specificity), and the Expertise Paradox (reconciling LLM and human coder competencies).

This panel puts together valuable efforts of different research groups across the world to not only use, but also reflect on the use of LLMs in Communication studies. They show important avenues for the field to think about different approaches to validity, ethics and truthful cooperation between humans and computational models without erasing the challenges and disagreements.

References

Bail, C. A. (2023). Can Generative AI Improve Social Science? SocArXiv Papers https://doi.org/10.31235/osf.io/rwtzs

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology. https://doi.org/10.1145/3641289

Fecher, B., Hebing, M., Laufer, M., Pohle, J., & Sofsky, F. (2023). Friend or foe? Exploring the implications of large language models on the science system. AI & SOCIETY. https://doi.org/10.1007/s00146-023-01791-1

Gil de Zúñiga, H., Goyanes, M., & Durotoye, T. (2024). A Scholarly Definition of Artificial Intelligence (AI): Advancing AI as a Conceptual Framework in Communication Research. Political Communication, 41(2), 317–334. https://doi.org/10.1080/10584609.2023.2290497

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A Human–Machine Communication research agenda. New Media & Society, 22(1), 70–86. https://doi.org/10.1177/1461444819858691

Jameel, T., Ali, R., & Toheed, I. (2020). Ethics of Artificial Intelligence: Research Challenges and Potential Solutions. 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 1–6. https://doi.org/10.1109/iCoMET48670.2020.9073911

CODING RACISM IN MEDIA: BIASES IN CODING DECISIONS BY HUMAN CODERS AND LLMS

Ahrabhi Kathirgamalingam
University of Vienna

Fabienne Lind
University of Vienna

Jana Bernhard-Harrer
University of Vienna

Hajo G. Boomgaarden
University of Vienna

Part of social reality, and therefore a central subject of social scientific measurement, are constructs of marginalization, such as forms of discrimination, intersections of such, hate speech, incivility, and other types of communication potentially affecting marginalized groups. For these constructs of marginalization critically related to social justice, robust and valid measurement holds the power to provide evidence and explanations that can inform policy and promote equity (Scharrer & Ramasubramanian, 2021). To study such constructs, researchers often rely on manual quantitative content analysis. Besides, automated content analysis is more and more frequently used. Implementing such an algorithmic approach, however, often requires a larger number of manually annotated data.

The quality of human coding in terms of reliability and validity has long been the focus of methodological debates. Previous research indicated that next to the type of construct (e.g., Niemann-Lenz et al., 2023; Potter and Levine-Donnerstein, 1999) and the coder training and instructions (e.g., Lind et al., 2017), it is especially various coder-level characteristics (e.g., Niemann-Lenz et al., 2023, Peter et al., 2002) that are major influencing factors for the agreement or disagreement of coders. In this project, we add to this literature by investigating how coder-level characteristics impact the coding of racism in news media texts. We inspect more specifically how characteristics like coders' lived experience and awareness to discrimination impact coders' decisions on whether a news text is considered to include racism or not (RQ1).

As generative Large Language Models (LLMs) are increasingly being introduced to assist in content analysis tasks and are reported to be competitive with student and crowdworkers (Gilardi et al., 2023), we additionally investigate LLM coding performance specifically for coding racism in news media texts. More and more studies not only discuss the potential of LLM-based coding but also sources for variation when prompting the model repeatedly to code the same text (e.g., Reiss, 2023). Furthermore, issues such as selection and labeling bias (Hovy & Prabhumoye, 2021) also play a role. Assigning personas by enriching prompts with characteristics of human (e.g., political leaning) is increasingly studied (e.g., Beck et al., 2024; Deshpande et al., 2023; Gupta et al., 2024) and offers an interesting avenue for investigating coder bias. Relying on

this technique, we ask how persona assignments, compared to default models, influence variations in annotation decisions by LLMs for constructs of marginalization (RQ2).

Since LLMs currently are still primarily considered as an augmentation to human coders and to account for potential text property-driven variation, thirdly (RQ3), we ask how annotation decisions by LLMs and human coders compare regarding the presence and extent of coder bias in the detection of racism and how text properties might explain differences.

To investigate RQ1 and RQ2, we conduct two studies and compare their results in a third part for RQ3. To explore the sources of disagreement in human coding of racism (Study 1), we conducted a survey with 164 paid crowdworkers and 360 coding tasks. Before explaining the coding task to the participants, they were asked to complete a pre-questionnaire. For measuring coder-level characteristics, we included items to assess if coders are negatively affected by racism, their political attitudes, awareness of racism, and prior coding experience. Control variables measured were age, gender, education background and migration background. The pre-questionnaire was followed by a coder training with definitions and examples of racism in news media. The crowdworkers were asked to code 15 short paragraphs as racist or not. The short paragraphs were randomly selected from German mainstream (Bild, TAZ, Welt) and online far-right alternative news media (Junge Freiheit, PI News, zuerst). The selected outlets in each category are well-known and have a high reach in Germany. Further, the randomly selected paragraphs each contain a cue representing a potential target group of racism (e.g., ‘migrant’).

For study 2, we use the insights from our human coding to design persona-based prompt experiments to explore biases in the annotations produced by LLMs systematically. We include two LLMs (GPT-3.5 and GPT-4o) to evaluate currently popular models. Based on the insights of our human coding, we create distinct personas based on combinations of the characteristics introduced in our human coding survey (experience, awareness, age and education) that are set as system parameters to the LLMs prior to coding. We prompt with the exact details provided to the human in our survey and gather five annotation decision generation per persona for the above-described 360 news media paragraphs. Thus, we enable the exploration of the alignment of LLM coding decisions with assigned persona profiles and differences between the LLMs.

Lastly, we compare human and persona-based LLM coding decisions from Study 1 and 2 by examining overlaps in annotation decisions and qualitatively exploring text properties that led to significant deviations between human and LLM annotations.

Our findings underscore the necessity of giving careful consideration when selecting coders, whether human or AI, to capture and analyze constructs of marginalization. We find that being affected by or being aware of marginalization causes systematic variation in human annotations, while persona assignment significantly impacts LLM outputs. As a key takeaway from our study, we argue to "agree to disagree", meaning to accept and

even intentionally introduce *valid* and systematic disagreement or variation in annotation decisions.

By offering more specific recommendations, we seek to strengthen the integrity and validity of manual and automated content analysis of constructs of marginalization. With improved measurement, communication research can better identify patterns of discrimination and inform policy toward a more equitable society.

References

Beck, T., Schuff, H., Lauscher, A., & Gurevych, I. (2024). Deconstructing the Effect of Sociodemographic Prompting. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 2589–2615.

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & Narasimhan, K. (2023). Toxicity in chatgpt: Analyzing persona-assigned language models. arXiv preprint arXiv:2304.05335.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). Chatgpt outperforms crowd-workers for text-annotation tasks. arXiv preprint arXiv:2303.15056.

Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., & Khot, T. (2024, January). Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. arXiv preprint arXiv:2311.04892.

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. Language and Linguistics Compass, 15 (8), e12432. https://doi.org/10.1111/lnc3.12432

Lind, F., Gruber, M., & Boomgaarden, H. G. (2017). Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. Communication methods and measures, 11(3), 191-209.

Niemann-Lenz, J., Dittrich, A., & Scheper, J. (2023). Coding quality in manual content analysis: An exploration of coder characteristics and category types for crowdworkers and student coders. SCM Studies in Communication and Media, 12(4), 327-353.

Peter, J., & Lauf, E. (2002). Reliability in cross-national content analysis. Journalism & mass communication quarterly, 79(4), 815-832.

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis.

Reiss, M. V. (2023). Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. arXiv preprint arXiv:2304.11085.

Scharrer, E., & Ramasubramanian, S. (2021). Quantitative research methods in

communication: The power of numbers for social justice. Routledge.

ANALYZING EXTREMISM IN PODCASTS: AN LLM-ASSISTED INVESTIGATION OF THE BRAZILIAN MANOSPHERE

Bruna Silveira de Oliveira
Federal University of Minas Gerais

Tariq Choucair
Queensland University of Technology

Rousiley Maia
Federal University of Minas Gerais

Introduction

Extremist groups often construct their identities by disseminating their narratives in media environments (Berger, 2018; Bolet & Foos, 2023; Peterka-Benton & Benton, 2023), and podcasts are a crucial format to do so. Podcasting is one of the leading digital media practices of the 21st century (Llinares et al., 2018), although it has yet to be studied compared to other communicative practices. In this work, we analyze Brazilian masculinist (or Brazilian manosphere) podcasts, arguing that the claims of extremist groups can be analyzed computationally with varied and complex variables with the assistance of a large language model (LLM). The so-called manosphere is an extremist set of misogynist movements that originate and operate on the internet to address issues aligned with masculinity (Nagle, 2017; Marwick & Caplan, 2018; Ribeiro et al., 2020; Tranchese & Sugiura, 2021; Vilaça & D'Andréa, 2021; Thorburn, 2023). We started from 34,060 podcast episodes from Deezer, Google Podcasts, Soundcloud, Spotify, YouTube, Apple Podcasts, Listen to Notes, Bit Chute, MGTOW TV, and Castbox and used a proportional stratified random sample of 2,490 episodes.

In general terms, the ideology of the manosphere is based on the belief that contemporary society gives women too much power, coupled with biologically essentialist interpretations and pseudo-scientific concepts from evolutionary psychology about relationship patterns, especially heterosexual ones (Thorburn, 2023). The feeling of victimization and persecution is constantly evoked (Barcellona, 2022; Ging, 2019; Marwick & Caplan, 2018; Tranchese & Sugiura, 2021; Vilaça & d'Andréa, 2021). Masculinist groups believe that there is a conspiracy involving the government, educational institutions, the judiciary and even the church to protect women and attack men. They also claim that this dynamic is a kind of social engineering. Our research concerns the triad between legitimacy, intolerance and recognition and seeks to investigate how the demands of Brazilian masculinist groups, despite being intolerant, fit into the search for legitimacy and can lead to false struggles for recognition.

Justification, Research Questions and Methods

Content analysis requires data preparation and an accurate coding process for possible methodological replicability (Bardin, 2016; Krippendorff, 2018; Maia, 2023; Neuendorf, 2002; Sampaio & Lycarião, 2021). However, empirical research can be challenging to

carry out efficiently due to the speed at which content is produced, especially with the rise of extremism on social media platforms (Bolet & Foss, 2023; Conway, 2016; Gaikwad et al., 2021; Leitch & Pickering, 2022). Because of its automation potential, using LLMs can facilitate the process. Summing with the scale-up, another advantage of using LLMs for content analysis in extremism studies is the tentative reduction of the impact on the people who analyze data on online extremism (Pearson et al., 2023).

Considering these potentials and the discussion about the LLM's ability to identify nuances in pronouncements (Chew et al., 2023; He et al., 2023; von der Heyde, 2023), we ask:

RQ1: To what extent can LLMs be used to identify detailed variables of intolerance and perception of harm related to struggles for recognition?

RQ2: What errors, limitations and difficulties were encountered in the coding process?

Our study is divided into five phases. The first was the automated transcription of the episodes, using Whisper. Beyond the transcription, we also diarized the episodes to differ between speakers. The second was constructing the codebook, which has seventeen categories, including, for instance, “presence of intolerance,” “object of intolerance,” “type of intolerance,” and “what audience is the episode aimed at.” The third phase, the reliability test to validate the codebook, was conducted between two human coders, with sufficient agreement in all variables. The fourth phase was the code validation between humans and LLM. To make this comparison, we coded a sample of 50 podcast episodes, and then the specific GPT 4 model coded the same sample. The method was to adjust the prompt until the model could present a higher agreement with the human coders. After dozens of rounds, we reached sufficient agreement on all variables. The fifth and final stage is the application to the 2,490 episodes.

Preliminary Results

Refining our use of LLMs to assist the analysis of Brazilian masculinist podcasts included an iterative process of prompt adjustment. The complexity of the concepts under study and their respective variables we sought to analyze had significant challenges. These variables spanned a broad spectrum, from direct manifestations of intolerance (e.g., discursive violence towards specific sub-groups like older women, trans women, and solo mothers) to more nuanced expressions, such as the denial of oppression or violence against women. To show this, we present the process of one of the variables, “presence of intolerance.”

Our initial prompt in Portuguese, translated to English, was: "This transcription is from a single podcast episode. Analyze the content to identify any manifestations of intolerance." However, this needed to adequately capture the cases manual coders identified. It was too broad and failed to capture the subtleties of intolerance, particularly in cases where oppression or violence was denied or minimized.

Recognizing this limitation, we started a qualitative refinement process for a prompt that would better guide the LLM in identifying explicit and implicit forms of intolerance. The

revised prompt, translated to English, reads: "Please analyze the following transcription of a podcast episode with special attention to any manifestations of intolerance against women. This includes, but is not limited to, foul language directed at women, harmful gender stereotypes, denial or minimization of the oppression suffered by women, personal attacks based on gender, and incitations to violence against women. We consider manifestations of intolerance to include not only explicit statements but also subtle insinuations, derogatory jokes, perpetuation of negative stereotypes, and disqualification of women's experiences or feelings. Be detailed in your analysis and justify your answer with specific examples from the transcription when applicable. Please identify any form of intolerance against women in the transcription, with specific examples and justification."

This transformation of the prompts from a generic request for intolerance identification to a detailed, nuanced directive had significant implications. It required the model to engage in a deeper, more critical analysis of the content, considering the explicit and implicit ways in which intolerance could be manifested. The detailed prompt encouraged the model to discern subtle insinuations of intolerance and to justify its conclusions with specific examples from the text, enhancing the reliability and depth of the analysis. The evolution of our prompts showed the importance of qualitative, iterative processes in using LLMs for communication research, especially if dealing with sensitive and complex concepts like intolerance and extremist groups.

This work implies five conclusions remarks: i) the importance of podcast analysis – it is a cultural media phenomenon characterized by the closeness between the speaker and the audience, presented in an appealing and easily accessible format; ii) the use of LLMs to analyze claims by extremist groups reduces the impact on researchers' mental health, who would otherwise be exposed to more toxic data without this automated process. iii) the possibility of large-scale analysis; iv) the interpretative flexibility provided by the analytical process, which allows us to carry out more elaborate qualitative analyses; v) and finally, the collaboration between humans and machines is crucial in this methodological process. Human input was essential for calibration and adjustments at every operational stage. In other words, automating processes do not eliminate the need for human labor.

References

Bächtiger, A., Dryzek, J. S., Mansbridge, J., & Warren, M. E. (Eds.). (2018). The Oxford Handbook of Deliberative Democracy. Oxford University Press.

Barcellona, M. (2022). Incel violence as a new terrorism threat: A brief investigation between Alt-Right and Manosphere dimensions. Sortuz: Oñati Journal of Emergent Socio-Legal Studies, 11(2), 170-186.

Bardin, L.(2016) Análise de Conteúdo. São Paulo: Edições 70, 2016.
Berger, J. M. (2018). Extremism. MIT Press.

Bolet, D., & Foos, F. (2023). Media platforming and the normalisation of extreme right views. URPP Equality of Opportunity Discussion Paper Series No.22 https://doi.org/10.31235/osf.io/urhxy

Chambers, S. (2017). Balancing epistemic quality and equal participation in a system approach to deliberative democracy. Social Epistemology, 31(3), 266–276.

Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). LLM-assisted content analysis: Using large language models to support deductive coding. arXiv preprint arXiv:2306.14924.

Conway, M. (2016). Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research. Studies in Conflict & Terrorism, DOI: 10.1080/1057610X.2016.1157408

Gaikwad, M., Ahirrao, S., Phansalkar, S., & Kotecha, K. (2021). Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. Ieee Access, 9, 48364-48404.

Ging, D. (2019). Alphas, betas, and incels: Theorizing the masculinities of the manosphere. Men and masculinities, 22(4), 638–657.

Habermas, J. (2006). Political communication in media society: Does democracy still enjoy an epistemic dimension? The impact of normative theory on empirical research. Communication theory, 16(4), 411–426.

He, Z., Guo, S., Rao, A., & Lerman, K. (2023). Inducing political bias allows language models anticipate partisan reactions to controversies. arXiv preprint arXiv:2311.09687.

Krippendorff, K. (2018). Content analysis: An introduction to its methodology. Sage publications.

Leitch, S., & Pickering, P. (2022). Rethinking social media and extremism (p. 194). ANU Press.

Maia, R. C. M. (Org.). (2023). Métodos de pesquisa em comunicação política. Salvador: Edufba.

Marwick, A. E., & Caplan, R. (2018). Drinking male tears: Language, the manosphere, and networked harassment. Feminist media studies, 18(4), 543–559.

Nagle, A. (2017). Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right. John Hunt Publishing.

Neuendorf, K. A. (2002). The Content Analysis Guidebook Thousand Oaks: Sage. Pearson, E., Whittaker, J., Baaken, T., Zeiger, S., Atamuradova, F., & Conway, M. (2023). Online extremism and terrorism researchers' security, safety, and resilience: findings from the field.

Peterka-Benton, D., & Benton, B. (2023). Online Radicalization Case Study of a Mass Shooting: the Payton Gendron Manifesto. Journal for Deradicalization.

Ribeiro, M. H., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., Long, S., & Zannettou, S. (2021, May). The evolution of the manosphere across the web. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 15, pp. 196–207).

Sampaio, R. C., & Lycarião, D. (2021). Análise de conteúdo categorial: manual de aplicação. Escola Nacional de Administração Pública (Enap).

Thorburn, J. (2023). Exiting the Manosphere. A Gendered Analysis of Radicalization, Diversion and Deradicalization Narratives from r/IncelExit and r/ExRedPill, Studies in Conflict & Terrorism, doi: 10.1080/1057610X.2023.2244192.

Tranchese, A., & Sugiura, L. (2021). “I don’t hate all women, just those stuck-up bitches”: How incels and mainstream pornography speak the same extreme language of misogyny. Violence Against Women, 27(14), 2709-273.

Vilaça, G., & d'Andréa, C. (2021). Da manosphere à machosfera: Práticas (sub) culturais masculinistas em plataformas anonimizadas. Revista Eco-Pós, 24(2), 410-440.

von der Heyde, L., Haensch, A. C., & Wenz, A. (2023). Assessing Bias in LLM-Generated Synthetic Datasets: The Case of German Voter Behavior (No. 97r8s). Center for Open Science.

AUTOMATING POLITICAL TEXT CLASSIFICATION WITH LLMS: GENERALISATION ACROSS LANGUAGES AND PLATFORMS FOR STANCE DETECTION TOWARDS DIVERSE TARGET CATEGORIES

Tariq Choucair
Queensland University of Technology

Sebastian Svegaard
Queensland University of Technology

Kate O'Connor Farfan
Queensland University of Technology

Katharina Esau
Queensland University of Technology

Carly Lubicz-Zaorski
Queensland University of Technology

Laura Vodden
Queensland University of Technology

Axel Bruns
Queensland University of Technology

Supervised Machine Learning Gaps and LLMs Potential

In recent years, communication studies have increasingly used supervised machine learning (SML) algorithms to analyze online political messages at scale (e.g., Baden et al., 2020; Rytting et al., 2023). Compared to dictionary-based or unsupervised approaches, SML offers distinct advantages for studying the content of communication in specific contexts. It enables researchers to automatically apply categories that are anchored in existing theory and provides flexibility in capturing the complexity of textual data (Stromer-Galley & Rossini, 2023), as researchers can better control the variables included in the model. These strengths make supervised methods valuable for social science research in general and media and communication studies particularly. SML facilitates processing large text volumes based on comparatively small amounts of human-labelled training data. However, there are still challenges and limitations within the state of the art. In this study, we highlight two.

First, as SML approaches in media and communication studies are relatively new, there are specific tasks, types of content, and contexts to which they have not yet been applied. Specifically, there is a strong Anglocentric language bias, with the majority of studies analysing only English text data (Alslaity & Orji, 2024; Baden et al., 2022). Second, operationalization validity may suffer as scholars often rely on manual creation of training data sets as the source for the classification (Baden et al., 2022). While complex constructs arising from theory necessitate careful consideration and adaptation

to be operationalized in empirical research, machine learning algorithms in general rely strictly on manually labelled data to perform classification tasks, which can themselves be limited by the researchers' own biases. The characteristics of this data will dictate how the model will perform, but language can be articulated in countless ways, including, for instance, through irony, sarcasm, slang, colloquialisms and metaphors. SMLs identify patterns that correlate with provided classifications but may overlook valid linguistic variations. Generalization is compromised, as these methods do not "build upon an intuitive understanding of textual meaning" (Baden et al., 2022, p. 3). In summary, the model works based on training data – specific examples - and rather than the concept itself under investigation. New Large Language Models (LLMs) could address these issues.

LLMs, unlike supervised models reliant on task-specific training data, are trained in vast, varied text across domains and, importantly, languages. Despite a persistent English bias (Liang et al., 2023), training with other languages at least occurs to some extent and within the same model (i.e., the same model is trained on documents from different languages), making cross-language analysis more plausible than with previous tools and SML approaches. For example, an LLM trained on a vast corpus of political discourse across different languages could be used to analyse political messages in a new language, a task for which a supervised model would need coded training data in that language. Second, LLMs are trained using techniques like self-supervised learning, where they learn by interacting with and predicting patterns in unlabeled data. This approach allows them to develop a more complex and flexible interpretation of human language, moving beyond a reliance on manually labelled datasets. An advancement in this domain has been the transformer architecture (Vaswani et al.'s, 2017). Transformers employ a self-attention mechanism that enables models to weigh the importance of different parts of a sentence or sequence, thereby grasping contextually rich meanings embedded in the text. For instance, when analysing political messages to perform a stance detection, these models can discern not just specific keywords associated with each stance, but also the context and subtle issue positionality (e.g., against or in favour of an issue) potentially leading to more advanced interpretations of communication.

It's key to measure if and how LLMs improve generalization across languages by analysing diverse texts without needing language-specific training data. This possible flexibility extends to platform content types as well. For example, LLMs can potentially seamlessly adapt to the brevity and slang of Twitter, the multiple text fields structure of Facebook posts, or the specific tone of Youtube videos. Unlike other approaches that might require platform-specific trained models, LLMs can potentially generalize across both languages and platform formats.

Stance Detection towards Diverse Target Categories

Stance Detection is a crucial task within political text classification, where the goal is to identify the author's attitude (stance) towards a given target in a given message. In political communication, understanding whether a message supports, opposes, or is neutral towards a particular entity (such as a policy or candidate) is important for comprehending public opinion, polarisation, and discourse dynamics. This task,

however, poses significant challenges for traditional supervised machine learning (SML) models, particularly due to linguistic diversity and the need for nuanced contextual understanding. In this paper, we evaluate the performance of multiple LLMs on the target-stance detection task, examining their ability to generalize across languages and platforms, which is critical for scaling political communication research beyond Anglocentric and platform-specific confines.

Research Design

We conduct two case studies. The first explores the generalizability and the broader applicability of LLMs across different languages. We analyzed online election campaigns by political leaders from four different countries, each with a distinct language (Brazil, Australia, Denmark, and Peru). This comparison seeks to understand how well LLMs can adapt and perform in varying linguistic contexts. The second case study focuses on generalizing and ensuring broader applicability across different nature of posts (different types of posts within and across platforms). We analyse the discussions about the Voice to Parliament referendum in Australia in four major social media platforms: Facebook, Instagram, Twitter, and YouTube. By doing so, we assess the models' ability to interpret and classify political content that varies not only in format but also in the characteristics of communication that each platform may present.

For each study, we fine tune different LLMs with manual annotated data in the source category (i.e., English for the first study, and Facebook for the second study). We then test the fine-tuned model over expanded categories (Portuguese, Spanish, and Danish for the first study, and Twitter, Youtube, and Instagram for the second). The chosen models for this study are: gpt-4o, gpt-4o-mini, gpt-3.5-turbo, phi-3, mistral 7b. The selection criteria were to include state of the art models, as well as small models with good performance in related tasks, to evaluate both scalability and efficiency.

Preliminary Findings

The preliminary results for suggest that fine-tuning significantly improves the performance of LLMs in the source language (English), aligning them more closely with human benchmarks. For instance, GPT-4o initially outperformed others with F1 scores of 0.67, 0.63, and 0.53 for the classes "favour," "against," and "neither" respectively, but improved to 0.74, 0.78, and 0.67 with fine-tuning. Similarly, the smaller Mistral 7b model showed notable improvements post-tuning, achieving F1 scores of 0.7, 0.72, and 0.65. This shows that bigger models are not always needed. However, when tested in Portuguese (the only non-English language assessed so far), the fine-tuned GPT-4o demonstrated strong cross-lingual adaptability with F1 scores of 0.88, 0.84, and 0.89, whereas the smaller models suffered significantly, with Mistral 7b, for instance, scoring only 0.43, 0.42, and 0.38. This shows the challenges smaller models face in cross-lingual generalizability. These findings are still a work in progress, with analyses in other languages pending.

References

Alslaity, A., & Orji, R. (2024). Machine learning techniques for emotion detection and sentiment analysis: Current state, challenges, and future directions. Behaviour & Information Technology, 43(1), 139–164. https://doi.org/10.1080/0144929X.2022.2156387

Baden, C., Kligler-Vilenchik, N., & Yarchi, M. (2020). Hybrid Content Analysis: Toward a Strategy for the Theory-driven, Computer-assisted Classification of Large Text Corpora. Communication Methods and Measures, 14(3), 165–183. https://doi.org/10.1080/19312458.2020.1803247

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. Communication Methods and Measures, 16(1), 1–18. https://doi.org/10.1080/19312458.2021.2015574

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers (arXiv:2304.02819). arXiv. https://doi.org/10.48550/arXiv.2304.02819

Rytting, C. M., Sorensen, T., Argyle, L., Busby, E., Fulda, N., Gubler, J. & Wingate, D. (2023). Towards Coding Social Science Datasets with Language Models. arXiv. https://doi.org/10.48550/arXiv.2306.02177.

Stromer-Galley, J., & Rossini, P. (2023). Categorizing political campaign messages on social media using supervised machine learning. Journal of Information Technology & Politics, 0(0), 1–14. https://doi.org/10.1080/19331681.2023.2231436

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, Long Beach, CA, USA.

DECIPHERING COMPLEX STANCES IN (DISRUPTIVE) CLIMATE PROTEST COVERAGE: A COMPARISON OF HUMAN CODING AND LARGE LANGUAGE MODELS

Hendrik Meyer
University of Hamburg

Cornelius Puschmann
University of Bremen

Michael Brüggemann
University of Hamburg

Introduction & State of Research

Automatically detecting complex rhetorical structures, such as the stance towards politically charged issues or controversial actors, within media debates, still poses significant methodological challenges for communication research. The recent rapid evolution of large language models (LLMs), exemplified by advances in commercial resources such as Open AI's GPT-4, as well as in 'open-source' alternatives, such as Llama and Mistral, potentially accelerates the speed, lowers the cost and improves the quality of natural language inference (NLI), making zero-shot learning approaches a viable alternative to the manual coding of large samples (Laurer et al., 2023). Given the outsized importance of complex constructs such as stance for hotly contested political issues, these innovations are interesting for media and communication scholars.

One area of intense political contestation and controversy is climate change protests. The global climate emergency continues to escalate by the day, with natural disasters underscoring the need for broad and deep emission reductions, as emphasized by scientific and political actors and global activism. Despite the efforts by civil society, governments have not adequately prioritized mitigation and adaptation. It is, therefore, no surprise that climate protest movements have experienced significant growth in recent years.

Various strategies are employed by protesters to advance their respective objectives, with certain factions becoming more 'anti-normative' and 'disruptive' (Shuman et al., 2021), such as Extinction Rebellion or the 'Letzte Generation' ('The Last Generation') in Germany. The latter engages in civil disobedience by blocking streets or occupying buildings—actions deliberately intended to garner political influence through media attention.

One might reasonably expect that all these protest endeavors, demanding crucial climate policy measures, should generate increasing societal support. However, recent polls in Germany contradict this assumption (More in Common, 2023). Notably, the decline in populous support not only concerns disruptive forms of climate protest, but

extends to all forms of climate protests. Media representations of these movements play a crucial role in shaping societal perceptions (Goldenbaum & Thompson, 2020).

Previous studies have showcased climate movements' success in drawing mediated attention to climate change issues (Meyer et al., 2023), with climate change playing a key role in public debate for decades (Nerlich et al., 2010; Schäfer & Painter 2021). However, increased visibility of protest actions and climate change concerns could then result in backlashes with antagonists conquering climate change debates through "connective counter-action" (Meyer et al., 2024). Such debates are politicized and often polarized, translating into a strong engagement from strategic communicators (politicians, activists, scientists, etc.) who frequently insist on the exclusive validity of their own arguments (Brüggemann et al., 2020; Kaiser & Puschmann, 2017; Meyer et al., 2023). Thus, it is essential to examine whether various forms of protests spark divided stances and "Discursive Polarization" (Brüggemann & Meyer, 2023), which could, in turn, amplify media attention toward more disruptive protest actions (Garimella et al., 2017).

Therefore, our exploration of LLM capabilities aims to investigate the journalistic coverage surrounding two climate movement organizations associated with two very different degrees of disruptiveness in Germany: Fridays for Future (FFF) and the Last Generation (LG). These two civil society actors utilize markedly different strategies concerning how they organize, execute, and communicate their protests. While LG is perceived as more disruptive, potentially radical, and enjoys less approval from the German population (Göllert, 2023), FFF is viewed as a more established and accepted movement, as well as being more deeply embedded within civil society (Haunss & Sommer, 2020).

However, challenges arise when investigating contentious debates in modern media spheres, as they are complex and dynamically changing. Here, the recent evolution of LLMs and zero-shot learning bears great potential. First approaches to use these techniques in order to scrutinize politically-contested and potentially polarized debates show promising results and high validity, e.g. in incivility analyses (Matter et al. 2024) and other, more complex text-based analysis tasks such as stance detection (Liang et al. 2023; He et al. 2023; Lan et al. 2023).

Research Questions

We illustrated that—in complex and dynamically changing media spheres on climate protests—LLMs could help to identify the distribution of oppositional stances within contested debates. Therefore, questions of validity of such methods for the context of climate protests arise:

RQ1*: How accurately do zero-shot classifications of Large Language Models discern varying degrees of support for climate protesters and their policy demands?*

After illustrating the validity of such novel methodological approaches, our work then aims to answer the following issue-specific research question:

***RQ2**: How do the journalistic debates surrounding the Last Generation and Fridays for Future vary regarding the degrees of support for climate protesters and their policy demands?*

Preliminary Results

We present a systematic approach to discerning stances towards the activism of (disruptive) climate protest movements, exemplified by "Last Generation", and "Fridays for Future", based on a press corpus of ~12,000 German-language articles from a broad range of news outlets, that each mention at least one movement. Relying on a combination of limited-scale human labeling and automated classification, we detail our coding instructions (for humans) and prompt refinements (for models), noting strengths and limitations of zero-shot stance detection, and leverage the ability of LLMs to provide detailed reasoning for their choices. Comparing five human coders with the capabilities of GPT-4 and subsequently fine-tuning two open source LLMs, preliminary results show good performance of LLMs for stance classification, underscoring both the technique's potential and its constraints: results show alignment in over 80% (Krippendorff's Alpha: 0.72) of the first sample of 300 coded articles, highlighting the technique's potential and limitations. These limitations, however, stem not only from the capacities of the models but also from the challenges of complex concepts, such as stances within protest frames, with LLMs prompting renewed questions about what constitutes a reliable concept definition.

The classification results indicate that, in general, a majority (62%) of outlets report on protest movements in a neutral manner. A smaller proportion (33%) adopts a more oppositional stance, while relatively few articles (5%) express support for the movements. This overall trend remains consistent across reports on the two different protest movements. However, the rates of neutral (67% versus 60%) and supportive (6% versus 4%) reports are higher, while texts expressing opposition (26% versus 35%) for the movements and their demands are lower, when articles refer to FFF as opposed to LG. This indicates a decrease in support for debates concerning more disruptive protests, while, simultaneously, coverage of these contentious events has increased. This raises normative questions about the productivity of potentially polarizing, protest-related counterpublics in instigating policy change (Meyer & Brüggemann, 2025).

Building on these initial methodological insights and going beyond protest debates, we leveraged the high reliability of the LLM codings to generate a large training dataset of 3,000 coded paragraphs, expressing stances across different climate change-related articles. Using this dataset, we trained a SetFit-based stance classifier. Although still in development and subject to ongoing improvement, the initial results indicate an accuracy of ~0.7 across different topical domains of climate change discourse. These domains include stances on climate protests and policies related to ecological transformations, such as the replacement of heating infrastructures or the implementation of a national speed limit. The classifier is publicly available on *HuggingFace* (see Puschmann, 2024), remains a work in progress, and will undergo continuous improvement.

References

Brüggemann, M., & Meyer, H. (2023). When debates break apart: Discursive polarization as a multi-dimensional divergence emerging in and through communication. Communication Theory, 33(2–3), 132–142. https://doi.org/10.1093/ct/qtad012

Goldenbaum, M., & Thompson, C. S. (2020). Fridays for Future im Spiegel der Medienöffentlichkeit. In S. Haunss & M. Sommer (Eds.), *Fridays for Future—Die Jugend gegen den Klimawandel* (pp. 181–204). transcript Verlag. https://doi.org/10.1515/9783839453476-009

Göllert, Lisa. (2023, January 19). Umfrage: Klimaschutz ja, radikaler Protest nein. https://www.ndr.de/ndrfragt/Umfrage-Letzte-Generation-geht-Mehrheit-zu-weit,ergebnisse1158.html

Haunss, S., & Sommer, M. (Eds.). (2020). *Fridays for Future: Die Jugend gegen den Klimawandel: Konturen der weltweiten Protestbewegung*. Transcript.

He, X., Bresson, X., Laurent, T., Perold, A., LeCun, Y., & Hooi, B. (2023). Harnessing Explanations: LLM-to-LM Interpreter for Enhanced Text-Attributed Graph Representation Learning. arXiv:2305.19523v3

Kaiser, J., & Puschmann, C. (2017). Alliance of antagonism: Counterpublics and polarization in online climate change communication. *Communication and the Public, 2*(4), 371– 387. https://doi.org/10.1177/2057047317732350

Lan, X., Gao, C., Jin, D., & Li, Y. (2023). Stance Detection with Collaborative Role-Infused LLM-Based Agents. arXiv:2310.10467

Laurer M, van Atteveldt W, Casas A, Welbers K. Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. Political Analysis. 2024;32(1):84-100. https://doi.org/10.1017/pan.2023.20

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D., Yang, X., Vodrahalli, K., He, S., Smith, D., Yin, Y., McFarland, D., & Zou, J. (2023). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. arXiv:2310.01783

Matter, D., Schirmer, M., Grinberg, N., & Pfeffer, J. (2024). Close to Human-Level Agreement: Tracing Journeys of Violent Speech in Incel Posts with GPT-4-Enhanced Annotations. arXiv:2401.02001

Meyer, H., Peach, A. K., Guenther, L., Kedar, H. E., & Brüggemann, M. (2023). Between Calls for Action and Narratives of Denial: Climate Change Attention Structures on Twitter. Media and Communication, 11(1), 278–292. https://doi.org/10.17645/mac.v11i1.6111

Meyer, H., Pröschel, L., & Brüggemann, M. (2024). From Disruptive Protests to Disrupted Networks? Analyzing Levels of Polarization in the German Twitter

Discourses around “Fridays for Future” and “The Last Generation”. https://doi.org/10.31219/osf.io/nd68z

Meyer, H., & Brüggemann, M. (2025): Discursive Polarization. In: Nai, A., Grömping, M., & Wirz, D. (Eds). *Elgar Encyclopedia of Political Communication*. Edward Elgar Publishing. Accepted version. https://doi.org/10.31219/osf.io/3x45p

More in Common. (2023). Wie schaut die deutsche Gesellschaft derzeit auf die Klimabewegung? https://www.moreincommon.de/klimabewegung/

Nerlich, B., Koteyko, N. and Brown, B. (2010). Theory and language of climate change communication. WIREs Clim Change, 1: 97-110. https://doi.org/10.1002/wcc.2

Puschmann, C. (2024): MiniLM-klimacoder_v0.1. *Hugging Face*. https://huggingface.co/cbpuschmann/MiniLM-klimacoder_v0.1

Schäfer, M. S., & Painter, J. (2021). Climate journalism in a changing media ecosystem: Assessing the production of climate change-related news around the world. WIREs Climate Change, 12, e675. https://doi.org/10.1002/wcc.675

UNDERSTANDING EXPOSURE TO AND ENGAGEMENT WITH POLITICAL NEWS ON FACEBOOK DURING THE 2018 AND 2022 ITALIAN ELECTIONS

Fabio Giglietto
University of Urbino

Luca Rossi
IT University of Copenhagen

Nicola Righetti
University of Urbino

Giada Marino
University of Urbino

The advent of social media, especially platforms like Facebook, has revolutionized the way news is consumed, a change starkly highlighted by the dissemination of misinformation and polarizing content, particularly after the 2016 US elections. This transformation has prompted the use of computational methods, including machine learning, to analyze political narratives on social media platforms. This research explores these dynamics within the Italian political media landscape, particularly focusing on the interplay between exposure to and engagement with political news stories on Facebook preceding the two most recent Italian general elections.

The study tackles two significant challenges: the limited availability of exposure data from social media platforms for external research (Benkler, 2019) and the difficulties in efficiently aggregating news stories on specific topics (Bonikowski & Nelson, 2022), especially in non-English contexts. By utilizing data provided by Meta and employing Large Language Models (LLMs), this research offers new insights into the untapped potential of social media data, demonstrating how LLMs can streamline the analysis of online political discourse.

To address the limitations inherent in traditional human-coder-based approaches, especially for large datasets, supervised and unsupervised topic modeling techniques have emerged as critical computational alternatives for content analysis (Chen et al., 2023). Moreover, research has demonstrated the efficacy of transformer models like BERT and RoBERTa in detecting topics and emotions and identifying clickbait headlines (Adoma et al., 2020; Briskilal & Subalalitha, 2022; Rajapaksha et al., 2021; Reimers & Gurevych, 2019). However, these predominantly English-trained, off-the-shelf models encounter challenges when applied to low-resource languages, including Italian. To mitigate these challenges, scholars have turned to language-specific, fine-tuned models, such as AlBERTo, for the Italian language on social media (Polignano et al., 2019). Yet, these pre-trained, language-specific models are difficult to maintain and frequently necessitate additional fine-tuning to deliver reliable results in particular domains. Finally, recent benchmarks clearly point out that LLM-based text embedding

regularly overperforms BERT on a set of standardized tasks (Muennighoff et al., 2022; Setser et al., 2024).

The research methodology involved a multifaceted approach using LLMs to understand the dynamics of political discourse during the 2018 and 2022 Italian elections on Facebook. It began with identifying political links through fine-tuning OpenAI's Ada model with a dataset of URLs shared on Facebook around the election periods. We examined 65,183 URLs (shared from 23 December 2017 to 4 March 2018) and 19,691 URLs (shared from 21 July 2022 to 25 September 2022) from Meta's URL Shares Dataset. A team of Italian scholars has been involved in the training phase, manually classifying a sample of these URLs into political and non-political categories, resulting in a highly accurate binary classifier.

Further, the study acquired text-embedding-3-large for all identified political URLs using the OpenAI embeddings API (Giglietto, 2024), followed by a k-means cluster analysis to group the URLs. This process involved testing different types of embeddings and optimizing the number of clusters for effective analysis. The clusters were then automatically labeled using the GPT-4 model, with human coders assessing the quality of these labels.

This LLM-in-the-loop pipeline introduces unique validation challenges, necessitating a reevaluation of accuracy assessment strategies (Marino & Giglietto, 2024). In this study, we address three primary challenges:

- The Swiss Army Knife Dilemma.
- The Granularity Spectrum Problem.
- The Expertise Paradox.

The Swiss Army Knife Dilemma highlights the fact that LLMs are general-purpose tools (Burkhardt & Rieder, 2024). Similar to the shifting patterns of a kaleidoscope, each researcher's decisions—ranging from model selection to algorithm configuration—produce a distinct pipeline that requires dual validation: first, validating the chosen configuration, and second, demonstrating its superiority over alternative approaches. This dual-layered requirement creates a validation landscape significantly more complex than that of traditional single-purpose models.

The second challenge, the Granularity Spectrum Problem, emerges when LLMs are used to cluster political content. The clusters generated by LLMs can vary widely in specificity, encompassing broad themes like "Economic Policy" to highly specific narratives such as "Prime Minister's Tax Reform Speech." This variability complicates validation efforts, raising the question: how can accuracy be effectively assessed when content can be justifiably clustered at multiple levels of granularity?

Finally, the Expertise Paradox addresses the divergence between the structured knowledge of LLMs and that of human coders. Trained on vast datasets, LLMs may

exhibit competencies that surpass those of traditional human coders in certain domains, posing a challenge to conventional validation benchmarks.
The integration of LLMs in the analysis of political discourse on social media will be explored in light of these challenges, offering insights into the practical implications of adopting such tools.

References

Adoma, A. F., Henry, N.-M., & Chen, W. (2020). Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition. 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 117–121. https://doi.org/10.1109/ICCWAMTIP51612.2020.9317379

Benkler, Y. (2019). Cautionary notes on disinformation and the origins of distrust. Social Science Research Counci. https://doi.org/10.35650/md.2004.d.2019

Bonikowski, B., & Nelson, L. K. (2022). From Ends to Means: The Promise of Computational Text Analysis for Theoretically Driven Sociological Research. Sociological Methods & Research, 51(4), 1469–1483. https://doi.org/10.1177/00491241221123088

Briskilal, J., & Subalalitha, C. N. (2022). An ensemble model for classifying idioms and literal texts using BERT and RoBERTa. Information Processing & Management, 59(1), 102756. https://doi.org/10.1016/j.ipm.2021.102756

Burkhardt, S., & Rieder, B. (2024). Foundation models are platform models: Prompting and the political economy of AI. Big Data & Society, 11(2). https://doi.org/10.1177/20539517241247839

Giglietto, F. (2024). Evaluating Embedding Models for Clustering Italian Political News: A Comparative Study of Text-Embedding-3-Large and UmBERTo. https://osf.io/preprints/osf/2j9ed

Marino, G., & Giglietto, F. (2024). Integrating Large Language Models in Political Discourse Studies on Social Media: Challenges of Validating an LLMs-in-the-loop Pipeline. Sociologica, 18(2), 87–107. https://doi.org/10.6092/issn.1971-8853/19524

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). MTEB: Massive Text Embedding Benchmark. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2210.07316

Polignano, M., Basile, P., Degemmis, M., Semeraro, G., & Basile, V. (2019). AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets. Italian Conference on Computational Linguistics. https://iris.unito.it/handle/2318/1759767

Rajapaksha, P., Farahbakhsh, R., & Crespi, N. (2021). BERT, XLNet or RoBERTa: The Best Transfer Learning Model to Detect Clickbaits. IEEE Access, 9, 154704–154716. https://doi.org/10.1109/ACCESS.2021.3128742

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1908.10084

Setser, A., Lange, L., Weiss, K., & Barash, V. (2024). Content modeling in multi-platform multilingual social media data. Journal of Online Trust & Safety, 2(2). https://doi.org/10.54501/jots.v2i2.136