Advancing Gender Equality in Media: Tackling Stereotypes and Biases with AI

Zhan Liu[1][0000−0003−3367−3204], Anne Darbellay[1], Nicole Glassey Balet[1][0000−0003−3268−8375], and Valérie Vuille[2]

[1] Media Innovation Lab
University of Applied Sciences and Arts Western Switzerland
HES-SO Valais-Wallis, Switzerland
{zhan.liu; anne.darbellay; nicole.glassey}@hevs.ch
[2] DécadréE, Switzerland
valerie.vuille@decadree.com

Abstract. This study presents an innovative approach to evaluating media representations of gender-based violence by integrating Natural Language Processing (NLP) techniques with the advanced capabilities of GPT-4, an Artificial Intelligence (AI)-based large language model. We developed a set of 27 expert-defined criteria to analyze a corpus of news articles, initially utilizing NLP methods for foundational text analysis. For more complex criteria, we employed GPT-4 and further enhanced its precision with fine-tuning. Our results indicate a significant increase in accuracy, achieving an overall 76% accuracy rate in content evaluation, which is 9 percentage points higher than using NLP alone. This research introduces a novel media content analysis framework and paves the way for future enhancements in automated journalism assessment and ethical reporting.

Keywords: gender-based violence · media content analysis · natural language processing · GPT-4 · artificial intelligence · automated evaluation.

1 Introduction

Gender-based violence, a pervasive global crisis deeply ingrained in gender inequality and unjust power dynamics, profoundly affects millions, especially women, girls, and the LGBTIQ+ community. According to DécadréE [1], this violence includes any act committed against a person's will, reflecting society's binary gender roles and unequal power relations. It encompasses threats, coercion, and can manifest as physical, emotional, psychosocial, or sexual violence, including deprivation of resources or access to services. Notably, women, minors, and those whose gender identity or sexual and affective orientation diverge from societal norms are predominantly affected. The World Health Organization's 2021 [4] data indicates that roughly one in three women have experienced such violence, either from an intimate partner or another individual. The multifaceted nature

of this violence, ranging from psychological abuse and physical harm to femicide, underscores the importance of media portrayal. Media representations can either perpetuate harmful stereotypes or contribute to a more nuanced and empathetic understanding. Thus, comprehending the complex dynamics of gender-based violence is vital as we explore how AI-driven analysis and intervention in media reporting can sensitively challenge biases and influence public perception and policy.

The media has been instrumental in bringing gender-based violence into public discourse, highlighting both high-profile cases and the daily struggles of survivors. However, this coverage is not without its flaws, as it often falls into the trap of perpetuating harmful stereotypes and biases. Women stepping forward with sexual assault allegations frequently face victim-blaming and skepticism, while male perpetrators sometimes receive sympathetic portrayals, especially if they are prominent figures. Additionally, the media's tendency to focus on sensationalist and extreme incidents can overshadow the broader social and cultural contributors to gender-based violence. This skewed portrayal underscores the urgency to rethink how gender-based violence is reported and to foster approaches that ensure balanced and equitable coverage.

In response to these issues, the Council of Europe [11] has urged member states to establish legal frameworks that uphold human dignity and prevent gender-based discrimination and violence. Media organizations are also encouraged to implement self-regulatory systems and ethical guidelines to support gender equality and eliminate discriminatory content. However, despite these initiatives, inconsistencies and biases in media language and imagery persist, undermining efforts to combat gender stereotypes and sexist violence. AI holds potential for transformative impact in this area. By integrating tools like machine learning, natural language processing, and sentiment analysis into media practices, AI can critically assess and adjust how gender-based violence is reported. These AI-driven tools can identify patterns of bias, assist in creating more gender-neutral content, and provide insights into diverse gender representations, promoting a more inclusive media landscape.

In this study, we developed an automated, artificial intelligence-based content evaluation system, designed to streamline and enhance the analysis of gender-based violence in media reports. Collaborating with large language models from OpenAI[3], our system utilizes sophisticated natural language processing techniques to automatically analyze and evaluate media coverage from various sources. The integration of comprehensive processes – including data collection, preprocessing, feature engineering, model design, and the establishment of specific evaluation criteria – enables our system to function as a highly efficient decision-making aid. Through rigorous evaluating across 27 different gender equality evaluation criteria, our model demonstrated an average accuracy of 76%, with 8 criteria having peak accuracy above 90%. This research has garnered recognition from field experts for its significant contributions to enhancing

[3] https://openai.com/

work efficiency and promoting gender equality in news content, demonstrating the substantial potential of AI in transforming media practices.

The paper is structured to succinctly present our research and findings. Section 2 provides a literature review, framing our work within existing research. Section 3 details the dataset description, including data collection and preparation. Section 4 outlines our methodology, evaluation criteria and implementation process, while Section 5 presents experimental setup and results, demonstrating our system's effectiveness. The conclusion summarizes our contributions and proposes directions for future research.

2 Related Work

Extensive research has scrutinized the media's portrayal of gender-based violence, revealing a dual narrative. On one hand, studies ([10], [8], [9]) criticize media coverage for often sensationalizing extreme cases, perpetuating harmful stereotypes, engaging in victim-blaming, and sympathetically portraying perpetrators, thereby distorting the realities of gender-based violence. On the other hand, the media is acknowledged as a pivotal platform for raising awareness and advocating for societal change, particularly when adopting ethical standards and self-regulatory codes ([3], [2], [12]). DecadréE [1] underscores the complexity of this issue, attributing problematic media representations to a blend of individual, structural, and systemic factors, and emphasizes the need for a nuanced approach in addressing these narratives to foster fairer and more consistent reporting on gender-based violence.

The use of innovative technologies, specifically AI and NLP, in media content analysis is a rapidly growing field of study. Recent literature ([15], [14], [7]) indicated that machine learning and AI can be powerful tools for analyzing large volumes of media content, identifying patterns, and evaluating the quality of reporting. Such techniques can automatically categorize media content, detect biases, and evaluate how well the media adheres to ethical standards and practices, thus providing a scalable solution for media content evaluation.

In the context of gender-based violence, AI and NLP could potentially be used to identify problematic representations and language use in media coverage, thereby aiding in the promotion of gender equality and combating sexist violence. Existing research has shown that AI can be trained to understand complex language patterns and detect subtle biases that may be challenging for humans to identify. This could be particularly beneficial for understanding how gender-based violence is framed and discussed in the media. However, most of the current research ([19], [5]) focused on the classification of news content and does not address the evaluation of gender-based violence content from different media.

Moreover, some studies highlight the impact of AI-based solutions on journalism and content creation ([17], [20], [6]). These solutions could provide insights into journalistic practices and raise awareness about the importance of fair and equitable reporting. This has implications for improving media practices, pro-

moting gender equality, and ultimately contributing to the fight against gender-based violence.

Nevertheless, as with any technological solution, there are challenges and ethical considerations associated with using AI and NLP for media content analysis. Concerns around algorithmic bias, transparency, and the implications for freedom of speech and journalistic autonomy are prevalent in the literature. Therefore, we need to be careful about the design and implementation of such systems to ensure they are used responsibly and ethically.

3 Dataset description

This study utilizes a specifically curated dataset provided by research partner, drawn from a comprehensive Swiss media archive containing news articles across a wide range of sources and covering a substantial temporal range. To construct a dataset specifically relevant to our study of gender-based violence, we employed a keyword-based filtering approach. This process involved selecting articles from Swiss French-language media sources using keywords indicative of various facets of gender-based violence, including "violence against women", "sexual harassment", "street harassment", "rape", "marital drama", "family drama", "domestic violence", "mutilation", "femicide", and "sexual coercion". We designed these selection criteria to ensure the retrieval of articles directly pertinent to our research theme.

Our data compilation yielded a collection of 1,752 news articles from 19 distinct French-speaking media outlets in Switzerland for the year 2022, including prominent publications like "Le Courrier", "Le Temps", "RTS Info", "24 Heures", "Le Nouvelliste", "20 Minutes", "Le Matin", "Swissinfo", and "Blick". Prior to analysis, we performed several preprocessing steps on the dataset: standardizing the format by removing all HTML tags and excluding duplicates and incomplete articles, resulting in a refined subset of 1,046 articles. Each article was categorized with a title, summary, and main body. The titles had an average length of 11.5 words (SD = 2.92), with a range of 4 to 21 words. The body text of the articles was substantially longer, averaging 508.5 words (SD = 332.67), with the shortest at 15 words and the longest at 3,757 words.

A focused subset of 1,046 articles underwent detailed annotation based on 27 expert-defined criteria, which provided a nuanced framework for our analysis of gender-based violence representation in the media. This rigorous annotation was pivotal for a thorough examination of these portrayals within the Swiss French-language media landscape. It also established a benchmark for evaluating the performance of our AI-based content evaluation system. By measuring the AI system's outputs against the expert annotations, we could determine the system's accuracy and its capability to emulate expert analysis. This step was crucial in validating the AI system's utility for media content analysis, with a particular emphasis on identifying and assessing the nuanced depictions of gender-based violence.

4 Methods and Implementation of Content Evaluation System

In this section, we detail the methodologies and processes used in creating and applying our AI-based system for evaluating news media portrayals of gender-based violence. We outline the step-by-step technical and procedural aspects, from data collection to AI system evaluation. The following subsections will describe our workflow, evaluation criteria, and the systematic implementation of our solution, highlighting the methodological rigor and innovation at the core of our research.

4.1 Architecture

The architecture of our content evaluation system is underpinned by a three-tiered workflow (as shown in Figure 1), designed to streamline the process from data acquisition to the synthesis of actionable insights. Each phase is interlinked, forming a cohesive pipeline that underwrites the robustness of our analysis.

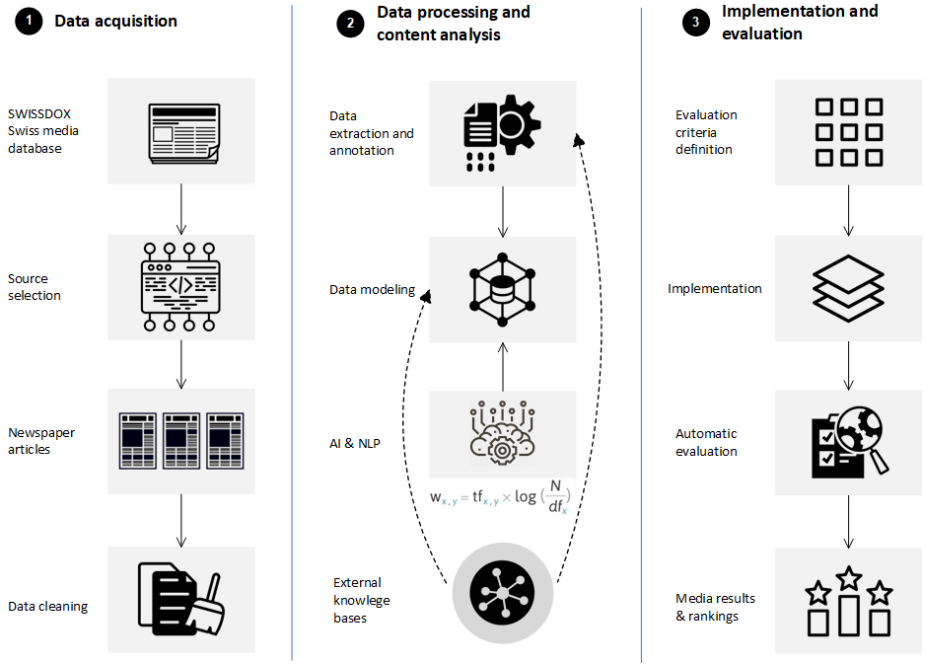

Fig. 1. Architecture of news content evaluation system

Our system's foundation was built upon a curated dataset provided by research partner, comprising Swiss French-language media articles. Sources were

strategically selected using keywords related to gender-based violence, ensuring the dataset's relevance. Following selection, we conducted a rigorous data cleaning process to ensure high-quality information for subsequent analysis.

The second phase of our workflow involved processing and analyzing the prepared data. We began with data extraction and annotation, tagging content based on predefined criteria. Following this, data modeling structured the information for detailed analysis. We employed advanced AI and NLP technologies for a deeper understanding and interpretation of the corpus, integrating external knowledge bases to enhance and contextualize our analysis further.

In the final phase, we defined a comprehensive set of evaluation criteria, against which the media content was assessed. We then moved to the implementation stage, where a prototype of our content evaluation system was developed. This prototype was rigorously tested to ensure its efficacy in automatically evaluating gender-based violence content in news media. To ensure the validity of our system, we engaged in an automatic evaluation process, where the system's outputs were compared against benchmarks set by domain experts. Finally, we conducted results analysis and ranking, assessing the performance of various media outlets in their coverage of gender-based violence, and highlighting areas for potential improvement.

4.2 Evaluation criteria

Building upon our foundational methodology, which was informed by 27 expert-defined criteria from DécadréE [1], we embrace a broad spectrum of violence types, including psychological, economic, physical, and sexual violence, as outlined by [13]. This broad approach allows us to cover various manifestations of gender-based violence, such as insults, denigration, and economic deprivation. Nonetheless, despite the inclusion of "symbolic" violence in studies like [18], challenges in attributing specific perpetrators led us to focus our analysis on more directly identifiable forms of violence, adhering to our methodological strengths and the practical scope of our research.

The evaluation began by verifying the presence of fundamental metadata in each article, such as the publication date and the author's name, which are cornerstones of journalistic credibility. We assessed the diversity of media sources, specifically identifying each article's origin from our list of 19 French-speaking news outlets. This step was critical for understanding the landscape of the reporting and any potential media-specific biases.

We then analyzed the content's relevance, categorizing articles by their direct relation to gender-based violence, and classifying them by length to infer the detail level provided. The type of violence reported was identified, noting especially the mention of femicide, to determine the focus areas of media attention.

The language used within articles underwent meticulous examination. We looked for the presence of certain vocabularies that could either clarify or obfuscate the severity of the incidents. This included terms that could potentially romanticize the violence or attribute it to factors like "passion", which can be misleading and harmful.

The portrayal of individuals involved in reported incidents was another focal point. Descriptions of the victim's and perpetrator's behaviors were evaluated, alongside any mentions of the perpetrator's nationality or mental health, which are often laden with stereotypes. The inclusion of such details was noted for its potential to influence public perception and perpetuate stigmas.

Our criteria also extended to the analysis of articles for underlying themes that are pivotal in understanding gender-based violence. We checked for discussions on power dynamics, control mechanisms, and the escalation of violence. Furthermore, we considered whether articles connected individual incidents to wider societal issues, such as rape culture, and whether they employed statistical data to contextualize the violence within a broader societal framework.

Finally, we considered the articles' resourcefulness to the public, noting the inclusion of information beneficial to victims, like support hotlines or educational websites. The presence of expert interviews was also a key criterion, as it can lend authority and depth to the articles, providing readers with a more informed perspective on the subject matter.

Through the application of these expert-defined criteria, we aimed to provide a rigorous and nuanced assessment of news content, aiming to elevate the standard of reporting on gender-based violence and contribute to a more informed and empathetic public discourse.

4.3 Implementation process

Our content evaluation system represents an intricate fusion of NLP techniques and the nuanced understanding capabilities of AI-based large language models like ChatGPT. Initially, we employed foundational NLP methods - lemmatization, tokenization, POS tagging, and Named Entity Recognition - to structure the data efficiently. To quantify the significance of specific terms within our corpus, we integrated the TF-IDF weighting scheme, calculated using $W_{xy} = TF_{xy} \times \log\left(\frac{N}{df_x}\right)$. This formula, where TF_{xy} represents term frequency and $\log\left(\frac{N}{df_x}\right)$ signifies the inverse document frequency, helps in moderating a term's weight based on its occurrence, ensuring a balanced emphasis on contextually relevant terms.

Beyond these traditional methods, we addressed more complex evaluation criteria by employing large language models, recognizing their adeptness in processing extensive data and discerning subtle narrative nuances. Specifically, these models excelled in analyzing societal implications of gender-based violence and understanding intricate power dynamics between involved parties, tasks that extend beyond the scope of conventional NLP.

To further refine our system's performance, we fine-tuned the large language model (LLM), a process that significantly enhanced its predictive accuracy. Our fine-tuning process involved adjusting the LLM model's parameters specifically for the context of gender-based violence in media content. This was achieved by initially training the model on a broad dataset of general text to establish a

baseline understanding. Subsequently, we introduced a specialized dataset comprised of articles related to gender-based violence, annotated by experts with the nuances and complexities specific to this subject matter. The training process utilized a smaller learning rate to make subtle adjustments, ensuring the model can simulate natural human responses, to became more aligned with the thematic content of our research. We also implemented validation checks to monitor the model's performance and avoid overfitting, thereby maintaining its ability to generalize across different contexts while being adept at recognizing and analyzing the specific patterns of gender-based violence in media reporting.

Integrating this enhanced model with our robust NLP framework allowed us to construct a system of remarkable analytical depth. This synergy between fine-tuned AI models and rigorous NLP techniques enabled us to conduct a granular yet comprehensive examination of news content. Consequently, our system proved exceptionally adept at navigating the complex narrative landscape of gender-based violence, providing insights with unprecedented precision and depth. This advanced analytical capacity ensured a thorough and nuanced exploration of media reporting, setting a new standard for automated content analysis in the context of social and cultural research.

5 Experimental Setup and Results

In this section, we describe the environment used to perform our experiments and the accuracy results through our methodologies.

5.1 Environment setup

Our experimental environment leveraged the Python programming language, known for its versatility and robust library ecosystem in data science and machine learning. Central to our computational work was the IPython framework [16], which facilitated interactive scientific computing and dynamic code execution, crucial for our complex data analysis tasks.

For natural language processing, we employed Spacy[4], an open-source NLP library, which provided the necessary tools for efficient and in-depth text analysis. The core of our computational analysis hinged on the OpenAI API's GPT-4 large multimodal model[5], renowned for its superior performance in text generation and understanding. Initially, we performed TF-IDF and classification calculations using the Scikit-learn toolkit[6], but shifted to GPT-4 due to its enhanced accuracy in handling complex analytical tasks. In addition, we applied the Pandas[7] and Numpy[8] libraries for data processing and numerical analysis. These tools were instrumental in managing, processing, and analyzing our dataset, ensuring a streamlined and effective experimental setup.

[4] https://spacy.io/
[5] https://openai.com/gpt-4
[6] https://scikit-learn.org/
[7] https://pandas.pydata.org/
[8] https://numpy.org/

5.2 Experimental results

In our evaluation of media content against the 27 defined criteria, we initially utilized various NLP and classification methods to analyze the dataset. Our findings indicated that the results for 12 of these criteria were below our threshold for satisfaction. Consequently, we turned to advanced models, specifically GPT-4 and its fine-tuned counterpart, to reassess these 12 criteria. The comparative analysis revealed a marked improvement in accuracy when employing these AI models.

Figure 2 illustrates some examples of the accuracy levels achieved by different methods. The vertical axis quantifies the accuracy, ranging from 0 to 1, while the horizontal axis categorizes the methods into NLP, GPT-4, and GPT-4 with fine-tuning. Notably, the fine-tuned GPT-4 model consistently outperformed the others, underscoring the value of model customization.

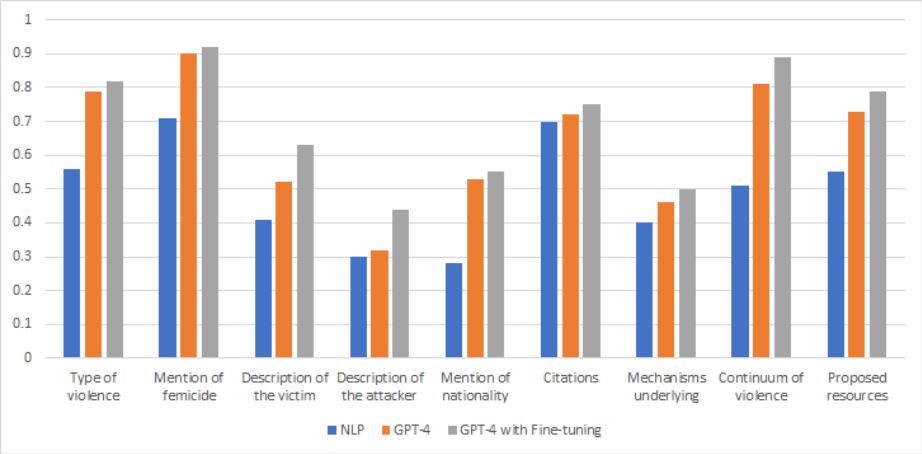

Fig. 2. Accuracy of criteria evaluation with different methods

When considering the entire set of 27 criteria, we observed a significant variance in accuracy rates, with the highest-performing criterion reaching 96% accuracy and the lowest at 44% by using the model of GPT-4 with fine-tuning. Our analysis through the GPT-4 model highlighted that article length was a determining factor in accuracy, longer articles tended to yield lower accuracy rates. Additionally, the style of writing influenced the results, a simpler narrative structure facilitated higher accuracy levels. Through the integration of GPT-4 models, we achieved an overall accuracy of 76% in our news article analysis, an improvement of nine percentage points over the methods relying solely on NLP. This enhancement demonstrates the effectiveness of incorporating AI-based large language models and fine-tuning techniques in the context of complex content analysis tasks.

6 Conclusion

In this study, we have demonstrated the efficacy of integrating advanced NLP methods with AI-driven large language models, particularly GPT-4, to evaluate media content related to gender-based violence. Our hybrid approach utilized traditional NLP techniques for initial text analysis, followed by the application of GPT-4 models, fine-tuned to our specific dataset, to handle more complex analytical tasks. The results signify a considerable improvement in accuracy, particularly in analyzing criteria that traditional NLP methods found challenging.

Our contributions through this research are twofold: First, we have provided a comprehensive set of evaluation criteria grounded in expert knowledge, tailored to assess the quality of media reporting on gender-based violence. Second, we have established a methodological framework that combines the best of NLP and AI technologies to create a powerful tool for content analysis. The success of our approach is reflected in the precision of our system, which outperformed standard NLP techniques by a significant margin.

For future research directions, expanding the system's capabilities to handle broader and more diverse datasets could offer insights into global media narratives. Enhancing the fine-tuning process with a variety of large language model architectures could improve accuracy. Additionally, broadening the evaluation criteria to address new themes in media, such as intersectionality within gender-based violence contexts, presents a valuable area for study. Ethical considerations around AI's use in media analysis and the potential for real-time feedback mechanisms for media outlets also warrant further exploration, aiming for transparency, fairness, and improved reporting standards.

In conclusion, our research has not only advanced the current understanding and methodology of media content analysis, but has also laid the groundwork for future innovations that can further enhance the quality and ethical standards of journalism.

Acknowledgments

The research presented in this article was supported by a grant from the University of Applied Sciences and Arts of Western Switzerland (HES-SO) focused on the special topic of Gendered Innovation, under grant number 128298. We extend our sincere gratitude to our project partner DécadréE, for the invaluable support of the project.

References

1. Décadrée annual report: Media treatment of gender-based violence, https://decadree.com/wp-content/uploads/2020/09/rapport-2020.pdf
2. Journalism is an essential lever in the fight against gender-based violence. https://www.unicef.org/moldova/en/blog/journalism-essential-lever-fight-against-gender-based-violence, accessed: 2023-12-11

3. Media: A key to addressing and ending gender-based violence (gbv). https://catalystasconsulting.com/media-a-key-to-addressing-and-ending-gender-based-violence-gbv/, accessed: 2023-12-11
4. Violence against women. world health organization, https://www.who.int/news-room/fact-sheets/detail/violence-against-women
5. Abdulkareem, L.R., Karan, O.: Using ann to predict gender-based violence in iraq: How ai and data mining technologies revolutionized social networks to make a safer world. In: 2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). pp. 298–302. IEEE (2022)
6. Bailer, W., Thallinger, G., Krawarik, V., Schell, K., Ertelthalner, V.: Ai for the media industry: application potential and automation levels. In: International Conference on Multimedia Modeling. pp. 109–118. Springer (2022)
7. Bello, H.J., Palomar, N., Gallego, E., Navascués, L.J., Lozano, C.: Machine learning to study the impact of gender-based violence in the news media. arXiv preprint arXiv:2012.07490 (2020)
8. Berns, N.: Degendering the problem and gendering the blame: Political discourse on women and violence. Gender & society **15**(2), 262–281 (2001)
9. Buiten, D.: Silences stifling transformation: Misogyny and gender-based violence in the media. Agenda: Empowering women for gender equity (71), 114–121 (2007)
10. Cuklanz, L.: Representations of gendered violence in mainstream media. Questions de communication (35), 307–321 (2019)
11. de Europa, C.: Convention on preventing and combating violence against women and domestic violence (2011)
12. Jukic, E.: Research on media reporting on gender based violence against women in bosnia and herzegovina. Retrieved on March **25**, 2019 (2016)
13. Kelly, L.: The continuum of sexual violence. In: Women, violence and social control, pp. 46–60. Springer (1987)
14. de Lima-Santos, M.F., Ceron, W.: Artificial intelligence in news media: current perceptions and future outlook. Journalism and media **3**(1), 13–26 (2021)
15. Liu, Z., Balet, N.G.: Bringing big data into media: A decision-making model for targeting digital news content. In: 2022 IEEE/ACIS 7th International Conference on Big Data, Cloud Computing, and Data Science (BCD). pp. 218–223. IEEE (2022)
16. Pérez, F., Granger, B.E.: Ipython: a system for interactive scientific computing. Computing in science & engineering **9**(3), 21–29 (2007)
17. Pihlajarinne, T., Alen-Savikko, A.: Introduction to artificial intelligence, media and regulation. Artificial Intelligence and the Media (2022)
18. Sepulchre, S., Manon, T.: La représentation des violences sexistes et intrafamiliales dans la presse écrite belge francophone. Tech. rep. (2019)
19. Soldevilla, I., Flores, N.: Natural language processing through bert for identifying gender-based violence messages on social media. In: 2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE). pp. 204–208. IEEE (2021)
20. Tejedor, S., Vila, P.: Exo journalism: a conceptual approach to a hybrid formula between journalism and artificial intelligence. Journalism and media **2**(4), 830–840 (2021)