Assistance and bias checking for writing crime news

Eleonora Calò[1], Loredana Caruccio[1] and Grazia Margarella[1,*]

[1]Department of Computer Science, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy

Abstract

In recent years, increasing attention has been paid to the way journalists report on crime-related news. During the writing process, biases related to gender, race, or other factors may be introduced, often resulting in an overly dramatized narrative. Such biases can elicit conflicting emotional responses from readers and undermine the objectivity and reliability of the information presented. To address this issue, we introduce NEMESI, a plug-in designed to detect the presence of bias within news articles during their drafting. Beyond detection, NEMESI also provides suggestions for mitigating identified biases, thereby encouraging more balanced and impartial reporting. NEMESI also offers text generation support features and guarantees the proper management of changes through a timeline of modifications, facilitating a more informed and transparent writing process. The plug-in presented in this study aims to serve as a resource for journalists engaged in crime reporting, fostering a more neutral and fact-based approach that respects both the truth of the events and the individuals involved.

Keywords

Writing assistance, Text generation, Bias mitigation, User interaction

1. Introduction

Crime news is a genre of journalism focusing on events related to crime, violence, accidents, and other incidents that undermine public safety. Due to its nature, this type of news often attracts significant public interest, but it also carries substantial ethical responsibilities for journalists. Crime narratives can influence perceptions of readers, changing their view on safety and society. For this reason, presenting facts in an overly dramatic or repetitive way can have negative consequences. For example, it can increase the reader's fear or have the opposite effect and desensitize them to violence. These influences can have a social impact, influencing public discussion and fueling stereotypes and prejudices.

With the current dissemination of news through digital media, the risk of biased reporting is greater. The pressure to publish quickly and capture attention may cause some media sources to emphasize bloody details or use unsuitable language, thereby compromising the quality of the provided information.

In crime news articles, Italian journalists often tend to dramatize or sensationalize the description of events, even in serious cases such as murders. In the upper left corner of Figure 1, there is a newspaper article recounting a gendered murder, in which the reporter focuses on irrelevant details of the victim, such as her clothing: *wore jeans and a black T-shirt.* The center of Figure 1 illustrates a case of bias related to the geographic origin of the attacker. The reporter writes: *21-year-old non-EU national*, referring to the individual accused of molesting a 14-year-old girl. Finally, the right-hand side of Figure 1 highlights the use of narrative bias and romanticized language, as in the sentences: *He, meanwhile, sees his castle of lies collapsing* and *the rival, meanwhile, becomes almost a friend.* Specifically, the first two examples fall under demographic bias: a gender bias and an ethnic bias, respectively. Instead, the last example represents a narrative bias, in which news reporting takes on the tones of a literary or cinematic tale.

To encourage clearer, unbiased, and more responsible storytelling in crime reporting, text generation technologies based on Natural Language Processing (NLP) could be a valuable tool for supporting journalists. The usage of models capable of generating text could assist during the article writing processes by suggesting neutral, cohesive, and consistent language formulations. Another important aspect is the

1st Workshop on supporting CRIme reSolution Through Artificial INtelligence (CRISTAIN), held in conjunction with CHITALY 2025, Salerno, Italy, October 6–10

*Corresponding author.

†All authors contributed equally.

ecalo@unisa.it (E. Calò); lcaruccio@unisa.it (L. Caruccio); gmargarella@unisa.it (G. Margarella)

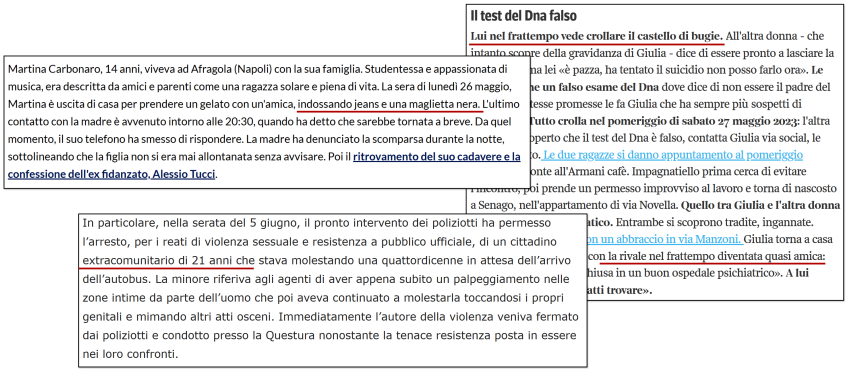


Figure 1: Example of biases in Italian crime news.

identification and mitigation of linguistic and narrative biases to avoid misleading, sensationalist, or stereotype-based formulations that could compromise the quality and impartiality of information.

In this paper, we present NEMESI, a visual plug-in designed to support Italian journalists in writing crime news articles. Although the system has not yet been implemented, the main goal is to offer journalists advanced language assistance tools, such as sentence generation, paraphrasing, summarization and expansion, while ensuring language neutrality and correctness. A key component of NEMESI is the bias detection module, designed to detect potentially biased elements in the text and to provide a contextual explanation that clarifies its criticality in journalism, offering more balanced and objective reformulation suggestions. The system is designed to act specifically on demographic biases and their main subcategories, i.e., gender, ethnic, religious, and socioeconomic biases. These features are essential to help journalists avoid the reproduction of stereotypes or the dissemination of discriminatory narratives, especially in relation to sensitive topics such as gender violence, homicides, or crimes involving foreign nationals.

2. Related work

This section examines the state of the art in the use of emerging technologies, with particular reference to the usage of Artificial Intelligence (AI) and NLP in the field of criminology and justice.

Processing and analyzing texts of the crime domain. Several studies proposed methodologies for analyzing criminal contexts based on unstructured data, highlighting the potential of AI and NLP technologies in identifying, preventing, and understanding criminal phenomena [1].

For example, in [2], the authors applied similarity measures and the TF-IDF technique to identify recurring patterns and reconstruct the chronological sequence of events. It demonstrated the possibility of identifying common patterns in similar cases. Researchers can also rely on alternative sources to analyze the impact of crimes all around the world. In [3], the authors conducted a sentiment analysis on tweets from seven cities (2014-2018), comparing them with official data. The analysis, conducted using an entropy tagger and Brown clustering, showed a good match with actual data, confirming the effectiveness of this approach in real-time monitoring. A further contribution is presented in [4], in which the authors combined social network analysis with a BiLSTM classifier optimized using an algorithm inspired by Spizella's swarm [5], which improves the convergence phase of learning. The evaluation metrics (accuracy, sensitivity, and specificity) showed increases of 0.5%, 1.16%, and 1.08%, respectively. To automatically extract spatio-temporal information from online news reports, a semantic approach that uses NLP techniques and a criminal domain ontology is proposed [6]. Relying on crime news in London, it provides the geographical distribution of crimes and allows to identify of the high-incidence areas. Similarly, in [7], the authors present a framework that integrates lexical, semantic, machine learning, and neural network techniques to extract criminal events from Italian articles, which are then represented in a knowledge graph.

Writing Assistance Technologies. *Writing Assistant Tools* can be defined as a product that aims to assist users with the production of written text. Examples are Grammarly[1], used for general purposes on the English language and the possibility to integrate different writing styles, and ProWritingAid[2], for story development and editing. Other tools have been tailored to domain-specific writing, including academic writing by Writefull[3] or legal writing by LexisNexis[4]. From early spell checkers to advanced AI-powered features, these tools have significantly evolved, transforming the way users structure, edit, and refine texts in disparate scenarios. The technologies underlying these tools can be classified into three categories [8]: i) **Pattern-matching style checkers** that rely on predefined rules and dictionaries to identify and correct errors in grammar, spelling, and punctuation, but are limited in understanding context; ii) **Language-model-driven rewriting tools** that provide an alternative to the traditional approaches, by allowing accuracy and reliability of tools that are capable to infer results from the context; iii) **Text generation tools** that are capable of creating texts from a user-provided prompt. They represent a paradigm shift from computer-as-editor to computer-as-author.

Overall, NLP and machine learning approaches have enabled the development of sophisticated writing assistants capable of offering real-time style generation. An example is the use of Retrieval-Augmented Generation (RAG), which allows generative models to access external knowledge bases to guarantee more accurate results. Examples are Panza [9] and PEARL [10], which are designed to have the user's historical documents as a knowledge base. Panza also focuses on privacy and security of user data, particularly critical in the case of email writing scenario, by run locally on the user's device. On the other hand, the PEARL's knowledge base is based on users' general-purpose documents, with a particular focus on the filtering and the selection of the most relevant chunks of text.

Writing assistant tools raise questions about authorship, originality, and over-reliance on algorithmic feedback. Such concerns have been inspected in [11] through a survey on the use of LLMs in writing. The study highlights that end-users are not sure in defining if a text developed with an LLM is definable of their intellectual property or not. Possible solutions to this problem include explainability, which could help the end-user to understand the actual contribution of the LLM.

Ethical considerations when dealing with texts. The integration of ethical considerations in the design and deployment of writing assistants is becoming increasingly important, particularly for bias detection and mitigation, privacy, and potential misuse of generated content. Biases can manifest in various forms and can have significant implications for fairness, equity, and social justice. The definition of bias can be influenced by several aspects, including the social and economic background of the user defining it, the context in which it is identified, e.g., in a sarcasm context, or the type of model used in the detection process. A bias interferes with the fairness of a model, which can lead to discrimination and possible representational harms of minority groups [12]. In a typical scenario, biases are connected to the training data or the model parameter configuration [13]. In this case, the model can be trained to detect and mitigate biases in the text, e.g., by identifying and flagging potentially biased language or suggesting alternative formulations that are more neutral and inclusive. Biases can also be introduced in the interaction between the user and the model. For example, if a user provides biased feedback or prompts, the model may learn to generate biased content in response. To this end, it is essential to implement mechanisms that promote fairness and inclusion in the interaction process.

Addressing the issue of bias in journalism, particularly in relation to crime, politics, and social issues, represents a fundamental but critical scenario to consider. In [14], the authors propose a framework for automatic bias detection and correction, using advanced language models (GPT-4o, Gemini Pro, Llama 8B) on over $30,000$ articles published from 2013 to 2023. The approach involves paragraph-by-paragraph detection validated by human reviewers, followed by an iterative debiasing phase using GPT-4o Mini. Overall, the introduction of AI in the field of criminal justice, if not accompanied by solid ethical principles and transparent methodologies, can consolidate and amplify pre-existing inequalities.

[1] https://www.grammarly.com/
[2] https://prowritingaid.com/
[3] https://www.writefull.com/
[4] https://www.lexisnexis.com/en-us/products/lexis-plus-ai.page

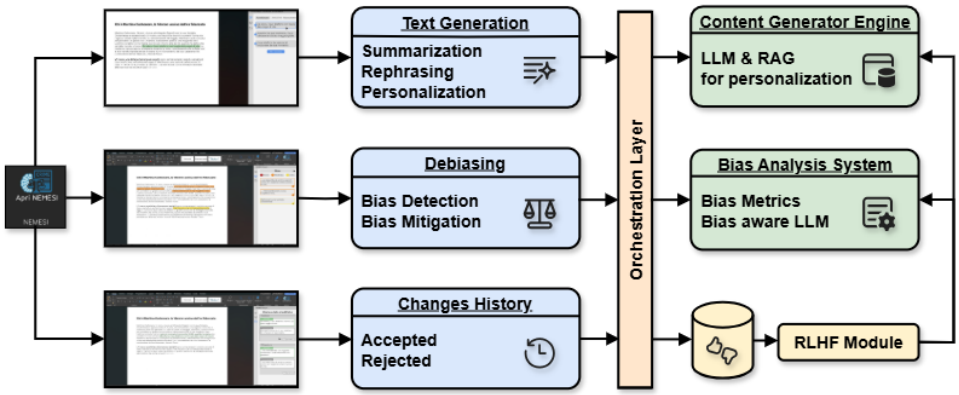


Figure 2: NEMESI proof-of-concept architecture.

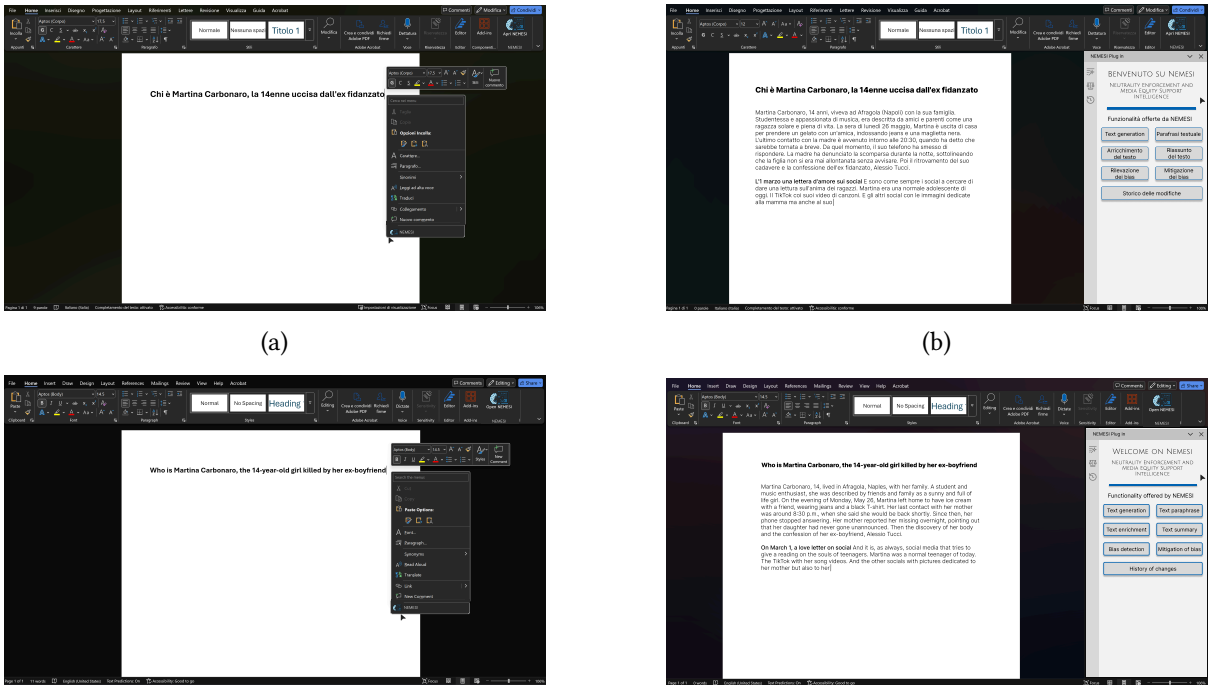
3. System design

This section presents an AI-powered assistant writer that combines personalized text generation, bias detection and mitigation, and feedback mechanisms to enhance journalists experience.

3.1. Proof-of-concept architecture

Figure 2 shows the proof-of-concept architecture of the proposed system, presenting the core components from the front to the back-end point of view. Then, implementation details are also provided.

Core Components. NEMESI is designed to assist journalists writing process while working with a familiar text editor, such as Microsoft Word, by providing a lightweight plug-in that integrates smoothly into the writing environment. The visual plugin is designed to allow the interaction between the user and the system through a simple interface, providing real-time suggestions on the text being written. The proposed proof-of-concept architecture is based on the interaction of three main components: the Content Generation Engine, the Bias Analysis System, and the Reinforcement Learning from Human Feedback (RLHF) Module. Concerning the first component, it is based on a fine-tuned LLM trained on balanced journalistic corpora. This allows to guarantee the generation of text and suggestions based on specific inputs from the crime news domain. In particular, this data is collected from the Italian journalistic domain, since the focus of the tool is on the Italian language. This module also incorporates the Retrieval-Augmented Generation (RAG) technology, which allows to generate of text based on the unique signature of the final user, i.e., the journalist. In order to achieve this, the system is designed to collect and store the user's previous articles and interactions with the system, allowing the model to adapt over time. Furthermore, it allows the user to rephrase text, enrich it with additional information, and summarize selected portions of the text. The **Bias Analysis System** employs an ensemble of specialized classifiers to identify various forms of bias. In this conceptualization stage, the proposed tool detects as main bias category the demographic bias. In particular, NEMESI analyzes how factors such as gender, race, political ideas, and socioeconomic status influence reporting, leading to unequal or disproportionate coverage of different demographic groups. The biases identified are listed by different levels of severity, ranging from severe to mild, and provide explanations for the detected criticalities. Another core feature of this module is the bias mitigation through language neutralization techniques, which suggest alternative phrasings while preserving the original meaning and intent of the content. A particular focus is given to euphemisms, inflammatory terminology, and culturally biased expressions. Finally, the **RLHF Module** is designed to continuously improve the system's performance and its ability to provide a fully tailored experience. Interactions of accepted and rejected suggestions provided by NEMESI are stored in a database of changes, and it is periodically sent to the RLHF module, allowing the continuous refinement of the bias detection algorithms and adaptation to evolving journalistic



(a) (b)

(c) (d)

Figure 3: NEMESI plug-in: activation of plug-in in Microsoft Word.

standards. This architecture is designed to be modular and extensible, allowing for future enhancements and integration with additional tools or data sources as needed. The front-end modules and interfaces interact through an Orchestration Layer with the previously described components. The Orchestration Layer ensures that the different components work together, managing the flow of data and requests between the front-end and back-end systems. This layer also handles user authentication, session management, and API requests.

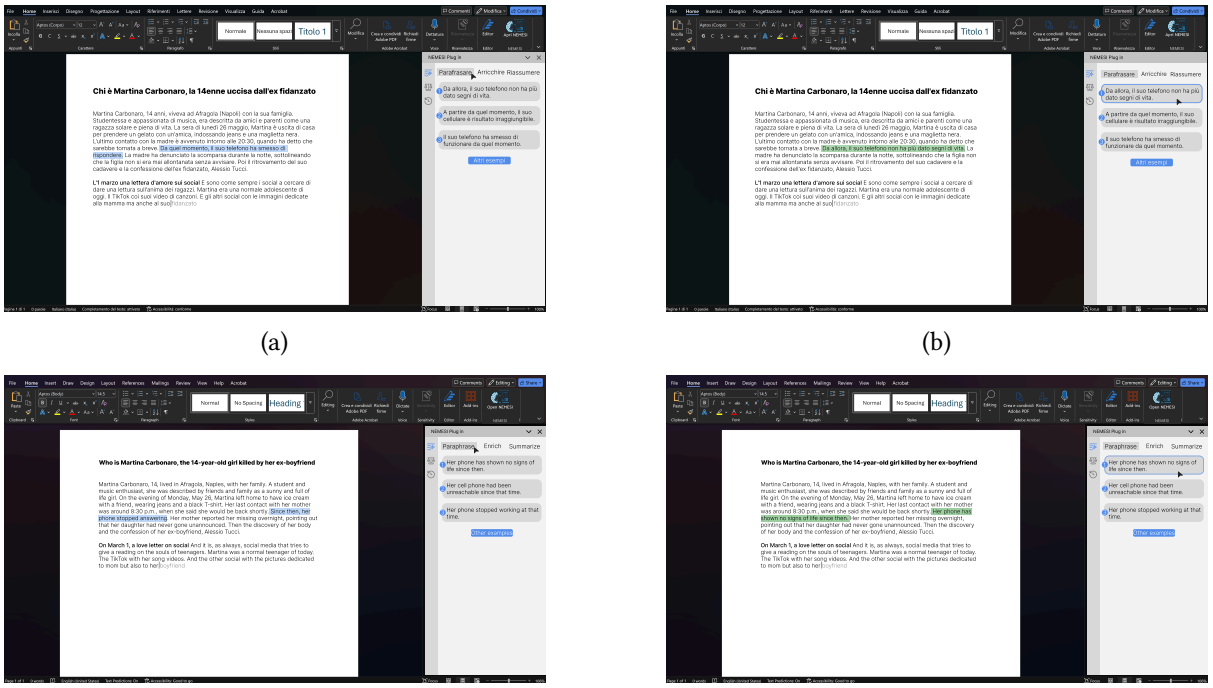
Implementation Framework. The system could be implemented as a text editor plugin. In this proposal, we considered NEMESI as a Microsoft Word plug-in, which can also be adapted to other platforms. Instead, the back-end could be built using Python with Flask or FastAPI for the Orchestration layer. The content generation engine would leverage a fine-tuned transformer model such as GPT-3.5 or Llama 2, trained on a balanced corpus of journalistic texts and working locally on the user device to be privacy-preserving. The bias analysis system would utilize a combination of pre-trained models and custom classifiers, possibly employing libraries like Hugging Face Transformers and Scikit-learn for model training and evaluation. The level of bias severity could be determined using specific metrics well known in the literature, such as the Bias Severity Score (BSS) accessible through Perspective API[5]. In the system implementation a special attention should be given to include content from various cultural, geographic, and ideological perspectives to ensure robust bias detection capabilities.

3.2. Functionalities of the tool

In the following, through the representation of appropriate mock-ups, we present how the user can interact with NEMESI for neutral editing of a crime news article. The plug-in is designed to work in the Italian language, so the examples are shown initially in Italian and then in English to ensure greater understanding by a wider audience.

In order to show the functionality of the plug-in, a real-life crime case from an event that occurred in

[5]https://www.perspectiveapi.com/



(a) (b) (c) (d)

Figure 4: Text generation functionalities.

Italy was considered. The article, published in the news outlet TGCOM24[6], was partially reproduced in the example for demonstration purposes.

Figure 3 shows the activation mode of NEMESI. Specifically, in Figure 3c (3a, resp.). The user has already entered the title of the article; right-clicking will open a drop-down menu, from which the NEMESI item can be selected to activate the plug-in. Once activated, a panel appears on the right side of the interface, as shown in Figure 3d (3b, resp.), which introduces the user to the use of the system. This initial screen presents an overview of the functionality offered by the plug-in to provide a clear orientation to what can be done with the tool. All the functionalities of NEMESI can be grouped in three main categories: **Text generation**, **Debiasing**, and **History of changes**.

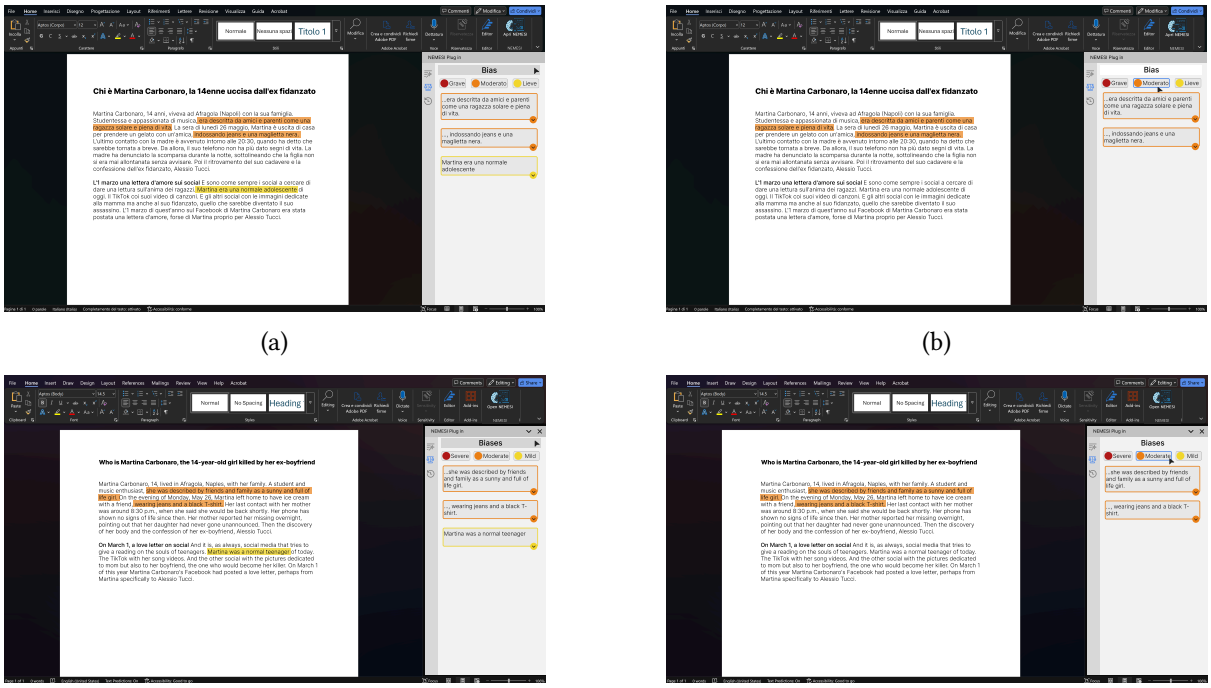
Text generation. The *Text generation* functionality, shown in Figure 4, is divided into four sub-features: direct text generation within the editor where the user writes the article, paraphrasing, summarization, and text enrichment. The provided example illustrates how to paraphrase a portion of text selected by the user.

In particular, in Figure 4c (or 4a, resp.), the written text is visible, where a light gray word appears immediately after the cursor. This element represents a suggestion automatically generated based on the context, designed to support the user during the writing phase. On the right side panel, the interface related to the three functionalities—paraphrasing, enrichment, and summarization—is displayed. In this example, the user selects a sentence within the text (highlighted in blue) and clicks the button dedicated to paraphrasing. The plug-in returns three possible reformulations of the selected sentence, with the option to generate additional suggestions by clicking the *More examples* button located at the bottom.

In Figure 4d (or 4b, resp.), the user selects the first paraphrase example. As a result, the sentence highlighted in blue in Figure 4c (or 4a, resp.), *Since then, her phone stopped answering.*, is replaced by the sentence highlighted in green in Figure 4d, *Her phone has shown no sign of life since then.*

The functionalities **Enrich** and **Summarize** follow a similar structure and operational flow.

[6] https://www.tgcom24.mediaset.it/



(a) (b) (c) (d)

Figure 5: Debiasing functionalities.

Debiasing. The *Debiasing* functionality is illustrated in Figure 5. It allows the user to quickly identify sentences in the article that exhibit potential biases, highlighting them with different colors based on the severity level: **Red** for severe cases, **Orange** for moderate, and **Yellow** for mild ones.

Figure 5c (or 5a, resp.) shows that by pressing the *Biases* button, the plug-in highlights in the text the sentences containing potential biases. These sentences are also displayed in the side panel, inside color-coded boxes. A consistent color scheme is used across both areas to ensure a clear visual correspondence between the main text and the side panel, thus facilitating the identification and analysis of biases.

The user can also filter the results based on the desired bias category. For example, by selecting the *Moderate* button, as shown in Figure 5d (or 5b, resp.), only the sentences marked as moderate biases will be displayed, both in the main text and the side panel.

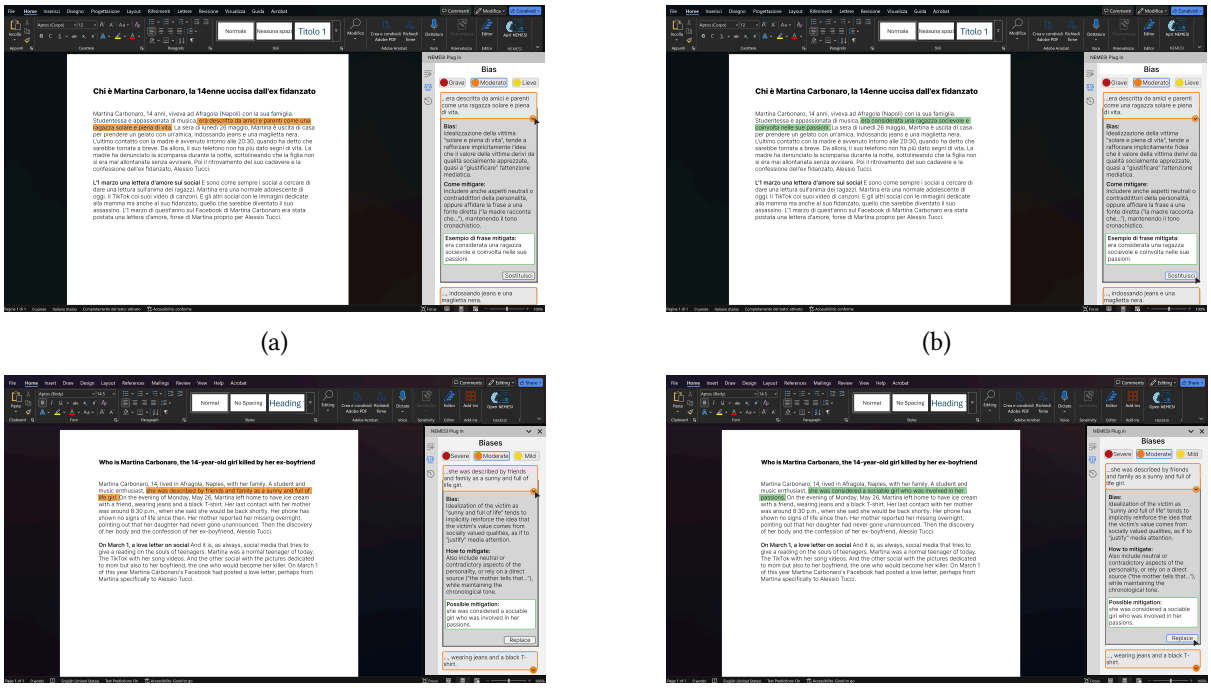
After identifying the biases, NEMESI allows for mitigating them. Figure 6 illustrates the main steps to modify a biased sentence by replacing it with a more neutral version.

In Figure 6c (or 6a, resp.), the user selects the box corresponding to the problematic sentence. The latter is highlighted in the text, while a sub-panel opens on the side explaining the detected bias and a mitigation proposal. A reformulation of the sentence is suggested, which can optionally be inserted into the text. As shown in Figure 6d (or 6b, resp.), by pressing the *Replace* button, the original sentence is replaced with the neutral version, highlighted in green to indicate the completed modification.

In Figure 6c, the initial sentence *she was described by friends and family as a sunny and full of life girl* is identified as an example of victim idealization, a bias that may imply that the person's value depends on socially appreciated qualities. The explanation in Figure 6d suggests that, in such cases, it is preferable to attribute the statement to a direct source and to include more neutral elements. As a result, the proposed reformulation is the sentence *Friends and family remember her as a girl with many interests, very attached to family and music.*

History of changes. The last functionality presented is *History of changes*, shown in Figure 7, which allows the user to view all the modifications made to the text. In the illustrated example, two previous modifications are displayed: a paraphrase and a mitigation.

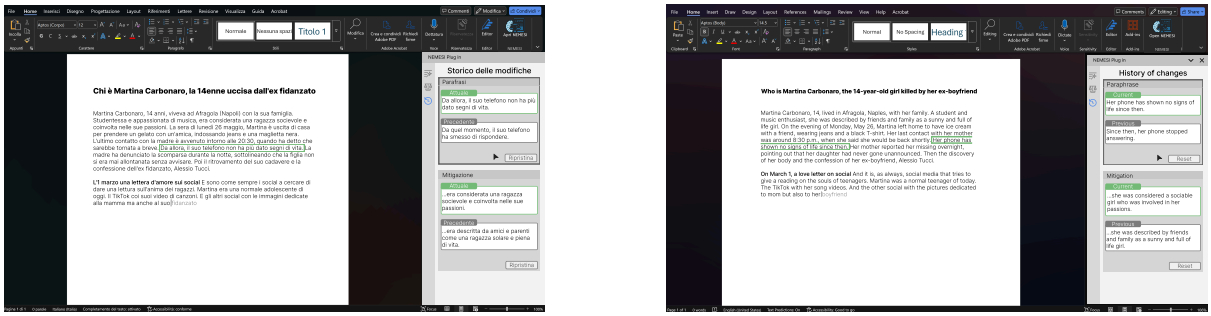
For each modification, the user can view both the current version of the sentence and the previous



(a) (b)

(c) (d)

Figure 6: Bias mitigation functionalities.



(a) (b)

Figure 7: History of changes functionalities.

one in the side panel. In the example, by selecting the modification related to the paraphrase, the corresponding sentence is highlighted in the text with a green outline. Within the side panel, each modification includes a *Reset* button, which, when pressed, allows the user to restore the original version of the sentence. This action is also recorded in the historical data as a new modification.

Consulting the edit history and choosing which version to maintain makes the system adaptable to the user's writing style, supporting a personalized experience based on the editorial preferences.

4. Conclusion

This paper presents the design of a writing support tool for Italian crime journalists. The plug-in is based on the integration of text generation technologies, supported by specific modules for the identification and mitigation of linguistic biases. This approach aims to promote more ethical, informed, and balanced storytelling, helping to reduce the spread of stereotypes through media. In addition, NEMESI aims to produce user-oriented texts, adapting to each author's editorial style. For future work, the goal is to develop the plug-in presented in this paper. It would also be useful, for evaluation purposes, to involve

professional journalists in a user study, in order to test its effectiveness in real-world contexts and gather valuable feedback to further improve the tool.

Acknowledgments

This study was funded by Ministero dell'Università e della Ricerca (MUR) of Italy in the context of the project denoted as BLOODSTAIN in the program PRIN 2022 (grant number D53D23008660006).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] P. Saravanan, J. Selvaprabu, L. Arun Raj, A. Abdul Azeez Khan, K. Javubar Sathick, Survey on crime analysis and prediction using data mining and machine learning techniques, in: Advances in Smart Grid Technology: Select Proceedings of PECCON 2019—Volume II, 2021, pp. 435–448.

[2] Z. Lwin Tun, D. Birks, Supporting crime script analyses of scams with natural language processing, Crime Science 12 (2023) 1.

[3] T. Vo, R. Sharma, R. Kumar, L. H. Son, B. T. Pham, D. Tien Bui, I. Priyadarshini, M. Sarkar, T. Le, Crime rate detection using social media of different crime locations and Twitter part-of-speech tagger with Brown clustering, Journal of Intelligent & Fuzzy Systems 38 (2020) 4287–4299.

[4] P. Mithoo, M. Kumar, Social network analysis for crime rate detection using Spizella swarm optimization based BiLSTM classifier, Knowledge-Based Systems 269 (2023) 110450.

[5] J. Xue, B. Shen, A novel swarm intelligence optimization approach: Sparrow search algorithm, Systems science & control engineering 8 (2020) 22–34.

[6] Y. Norouzi, Spatial, temporal, and semantic crime analysis using information extraction from online news, in: 2022 8th international conference on web research (ICWR), 2022, pp. 40–46.

[7] F. Rollo, L. Po, G. Bonisoli, et al., Online News Event Extraction for Crime Analysis, in: SEBD, 2022, pp. 223–230.

[8] R. Dale, J. Viethen, The automated writing assistance landscape in 2021, Natural Language Engineering 27 (2021) 511–518.

[9] A. Nicolicioiu, E. Iofinova, A. Jovanovic, E. Kurtic, M. Nikdan, A. Panferov, I. Markov, N. Shavit, D. Alistarh, Panza: Design and Analysis of a Fully-Local Personalized Text Writing Assistant, arXiv preprint arXiv:2407.10994 (2025).

[10] S. Mysore, Z. Lu, M. Wan, L. Yang, B. Sarrafzadeh, S. Menezes, T. Baghaee, E. B. Gonzalez, J. Neville, T. Safavi, Pearl: Personalizing Large Language Model Writing Assistants with Generation-Calibrated Retrievers, in: Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U), 2024, pp. 198–219.

[11] A. T. Wasi, M. R. Islam, R. Islam, Llms as writing assistants: Exploring perspectives on sense of ownership and reasoning, in: Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants, 2024, p. 38–42.

[12] R. Shelby, S. Rismani, K. Henne, A. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla-Akbari, J. Gallegos, A. Smart, E. Garcia, G. Virk, Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction, in: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 2023, p. 723–741.

[13] Y. Li, M. Du, R. Song, X. Wang, Y. Wang, A survey on fairness in large language models, arXiv preprint arXiv:2308.10149 (2023).

[14] C. W. Kuo, K. Chu, N. AlDahoul, H. Ibrahim, T. Rahwan, Y. Zaki, Neutralizing the Narrative: AI-Powered Debiasing of Online News Articles, arXiv preprint arXiv:2504.03520 (2025).