Reliability and Toxicity Detection Tool in Digital Media

Robiert Sepúlveda-Torres[1,*], Alba Bonet-Jover[1], Sergio Espinosa-Zaragoza[1], Kateryna Mamchur[1], Isabel Cabrera-de-Castro[2], Alba M. Mármol-Romero[2], Estela Saquete[1], Patricio Martínez-Barco[1], María-Teresa Martín-Valdivia[2] and L. Alfonso Ureña[2]

[1]*Department of Software and Computing Systems, University of Alicante, Spain*
[2]*Computer Science Department, University of Jaén, Spain*

Abstract

The accelerated spread of information, the constant demand for user interaction, and the overabundance of online content has amplified disinformation and toxicity in digital media. This article presents a tool designed to detect the degree of reliability and toxicity in digital news articles and comments published on digital newspapers. It is the result of a proof of concept focused on developing a tool that, trained with language models and Artificial Intelligence (AI) agents, can generate an expert report capable of providing users with a detailed analysis of the potential presence of reliability or toxicity patterns in digital content. A study was carried out to assess both the usability and accessibility of the tool showing that more than half of the users (67.3%) were satisfied with the tool. This result shows that the tool presented contributes to creating a healthier digital environment and represents a step forward in the detection of disinformation and toxic language.

Keywords

Reliability, Toxicity, Natural Language Processing, Disinformation, AI Agents

1. Introduction

Building a healthier digital environment is now one of society's key needs and a significant challenge in research, particularly in the field of Natural Language Processing (NLP). The speed of communication, combined with the constant need for interaction, has led to the phenomenon of *infoxication* or, in other words, information overload or excess of information that we can found in various digital media [1], due to the vast amounts of information coming from each person's interaction with the Information and Communication Technologies (ICTs) and their accelerated growth [2].

This information glut has been further exacerbated by Artificial Intelligence Generated Content (AIGC) and its associated hallucinations, which makes it increasingly difficult to distinguish high-quality content from false or harmful information. As stated by Bandara, "hallucinated output from large language models (LLMs) can serve as a potent source of disinformation in online ecosystems" potentially "fueling conspiracy theories, fake news, and inflammatory content".

Within NLP, various research lines are actively working toward fostering a healthier digital space, including detection of violent language and hate speech [4, 5], disinformation and fake news [6], or toxicity [7, 8].

SEPLN 2025: 41st International Conference of the Spanish Society for Natural Language Processing, Zaragoza, Spain, 23-26 September 2025.

*Corresponding author.
rsepulveda@dlsi.ua.es (R. Sepúlveda-Torres); alba.bonet@dlsi.ua.es (A. Bonet-Jover); sergio.espinosa@ua.es (S. Espinosa-Zaragoza); kateryna.mamchur@ua.es (K. Mamchur); iccastro@ujaen.es (I. Cabrera-de-Castro); amarmol@ujaen.es (A. M. Mármol-Romero); stela@dlsi.ua.es (E. Saquete); patricio@dlsi.ua.es (P. Martínez-Barco); maite@ujaen.es (M. Martín-Valdivia); laurena@ujaen.es (L. A. Ureña)
https://github.com/rsepulveda911112/ (R. Sepúlveda-Torres); https://cvnet.cpd.ua.es/curriculum-breve/es/bonet-jover-alba/61154 (A. Bonet-Jover)
0000-0002-2784-2748 (R. Sepúlveda-Torres); 0000-0002-7172-0094 (A. Bonet-Jover); 0000-0001-6237-0291 (I. Cabrera-de-Castro); 0000-0001-7952-4541 (A. M. Mármol-Romero); 0000-0002-6001-5461 (E. Saquete); 0000-0003-4972-6083 (P. Martínez-Barco); 0000-0002-2874-0401 (M. Martín-Valdivia); 0000-0001-7540-4059 (L. A. Ureña)

2. Methodology

In this context, the SocialFairness project was born as a proof of concept aimed at improving the digital environment through Human Language Technologies (HLT). The goal of this initiative is to develop a tool that applies language models to assess the reliability and toxicity of information, with a focus on the Spanish language. These two dimensions have been thoroughly studied by the consortium members, who belong to the GPLSI research group from the University of Alicante (UA) and the SINAI research group from the University of Jaén (UJA). The SocialFairness project is divided into two subprojects, each focusing on a specific research module [9]:

- **SocialTrust** (PDC2022-133146-C22), led by UA, is focused on assessing trustworthiness and reliability in digital news following two approaches: i) the 5W1H, a journalistic technique that "clearly describes key information of news in an explicit manner" [10] and which consists of answering six key questions (who, what, when, where, why and how), and ii) the notion of reliability. Concerning the latter point, Mottola and Zhang et al. highlight that, some of the indicators that influence the reliability of a news item are the ambiguity of the information, the lack of data and sources, emotional-charged expressions or stylistic features, characteristics that are the subject of our analysis. This subproject has resulted in the development of a corpus of 9,034 news annotated with both reliability labels and 5W1H elements [13].
- **SocialTox** (PDC2022-133146-C21), led by UJA, is focused on assessing i) toxicity, defined by Salehabadi et al. as "rude, disrespectful, or unreasonable comment", which is studied based on the presence of patterns such as insults, threats or inappropriate language, and ii) constructiveness in digital news comments, in which we assess whether the comments provide relevant information or knowledge to the article, with particular attention to the exchange of objective knowledge, the presence of reasoned arguments, and supporting evidence. Within the framework of this subproject, another resource consisting of 4,011 annotated news comments, labeled in terms of toxicity and constructiveness, was developed [15].

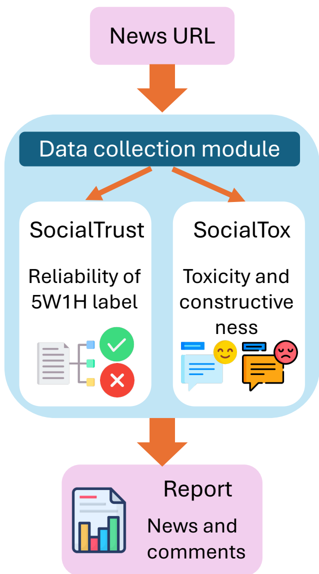

Figure 1: Functional diagram of the tool.

The research conducted across the modules culminated in a tool that generates expert reports on toxicity and reliability, helping users gauge the trustworthiness of news items and related comments.

For its generation, the two modules described in Figure 1 are taken into account. In the first module, for the analysis of the news item, an initial structuring of the news is carried out through the detection and extraction of its main content elements: the 5W1H. Then, a reliability value is assigned to each

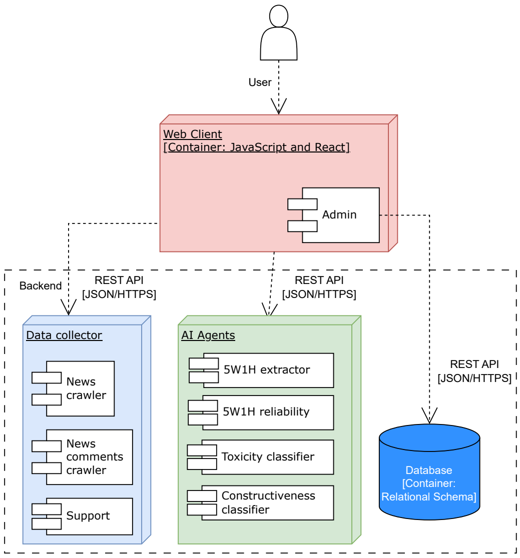

Figure 2: Microservices architecture diagram.

5W1H element. As for the second module, instead of carrying out an analysis of the news item text, the comments linked to it are analyzed and assigned a value of toxicity and constructiveness.

This work focuses on presenting the tool developed, detailing its objectives, the data and models used for training, and the results obtained. The article is structured as follows: Section 3 introduces the architecture and implementation; Section 4 presents the solution evaluation; Section 5 presents the conclusions and future work.

3. Architecture and implementation

According to the functional diagram in Figure 1, we identify the need for a data capture module to extract news texts and their associated comments. In a second stage, a specialized analysis must be conducted for each type of information. Since performing this analysis manually is unfeasible, Artificial Intelligence (AI) agents that emulate the behavior of domain experts are proposed. To address this challenge, we propose modules dedicated to evaluating the reliability of news articles, as well as measuring the toxicity and constructiveness of comments. Finally, all processed information is consolidated into a detailed report that presents the results obtained for each analyzed element.

The proposed solution adopts a microservices architecture. A microservice is a minimal functional software module that is independently developed and deployed, facilitating the design and development of a tailored solution for each system module [16]. Figure 2 presents a diagram illustrating the architecture of the tool. As can be seen, it consists of two clearly differentiated elements: the web client and the backend, responsible for providing the functionality.

3.1. Backend of the tool

The backend of this tool is entirely developed in Python. However, as shown in the diagram, it is not a monolithic system but follows a microservices architecture. Each of these microservices communicates with the web client (frontend) via the REST API protocol and is implemented using the FastAPI library.

3.1.1. Data collection

The data collection module's purpose is to extract the content of a news article and its comments. To achieve this, two crawlers have been designed: one responsible for retrieving the body of the news article and the other for collecting the comments. Although each crawler operates as an independent service, due to their similarity, both are deployed within the same microservice.

The first crawler obtained the HTML content via the Requests library and parsed it using BeautifulSoup library[1]. On the other hand, news comments are extracted using Python's Selenium library[2]. Unlike news, comments are usually dynamic content, so it is more convenient to use a high-level library such as Selenium. Implementing this functionality with the libraries mentioned above would be much more complex. The web crawlers are designed to work in 20 Spanish-language digital media. These media outlets were selected according to criteria including editorial relevance, social impact, geographical diversity, and technical feasibility for automated data collection.

3.1.2. Automated agents

The microservice integrating the AI agents comprises four distinct models, each specifically designed to address the tasks described in the Methodology. This microservice is deployed on a server with NVIDIA RTX 4090 GPU, which significantly accelerates the inference process.

The AI agents are built upon a language model fine-tuned for a specific task. To carry out this process, a dataset for each task was used for training and testing. In the case of comments, a corpus with 4,011 annotated examples was employed. Tables 1 and 2 present the training and test sets used to train the Toxicity and Constructiveness models.

Table 1
Distribution of annotated examples by toxicity class.

	Non-toxic	Partially toxic	Toxic
Training	1,052	1,243	748
Test	381	439	148
Total	1,433	1,682	896

Table 2
Distribution of annotated examples by constructiveness class.

	Constructive	Non-constructive
Training	997	2,046
Test	349	619
Total	1346	2,665

On the other hand, for the 5W1H task, a corpus with 9,034 5W1H labels and their respective reliability values was used. Tables 3 and 4 present the set used to train 5W1H extractor and 5W1H reliability models, respectively.

Table 3
Distribution of annotated examples by 5W1H label.

	What	Who	When	Where	Why	How
Training	2,711	1,843	778	801	238	563
Test	736	482	182	215	61	167
Total	3,447	2,325	960	1,016	299	730

[1]https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[2]https://www.selenium.dev/documentation/

Table 4
Distribution of annotated examples by Reliability class.

	Reliable	**Partially Reliable**	**Unreliable**
Training	4,765	1,276	893
Test	806	243	162
Total	5,571	1,519	1,055

To tackle these tasks, a wide range of approaches were explored, including both traditional machine learning algorithms and state-of-the-art generative models built on transformer architectures. The experiments were carried out using 4 NVIDIA A100 40GB GPUs.

The comment classification models (Toxicity and Constructiveness) and the reliability detection model (5W1H Reliability) are based on the RoBERTa-BNE-base[3] model, an encoder model in Spanish. On the other hand, for 5W1H entity annotation (5W1H Extractor), a generative decoder-type model, specifically Llama 3.2 3B Instruct[4], was used. Below are the results of each AI Agent when predicting on the test partition of each task:

- **Toxicity classifier** [17]: 0.61.
- **Constructiveness classifier** [18]: 0.81.
- **5W1H extractor** [19]: 0.66.
- **5W1H reliability** [20]: 0.61.

We are currently working on improving the results achieved by each model. The hyperparameter configuration and the prompt used to train the LLaMA-based model can be found in the references associated with each model.

3.2. Web Client

The web client of this tool is implemented in React, a JavaScript library that facilitates the creation of interactive and reusable web components. This technology allows the development of web interfaces quickly and efficiently, optimizing the updating of data coming from the backend. In addition, its component-based architecture facilitates the modification and extension of web interfaces.

3.2.1. Administration module

News websites frequently update their HTML structure, which can render our crawlers obsolete due to their static code. To address this issue, this module includes tests for the microservice collecting news and comments from each newspaper. These results are stored in a relational database, using Supabase's Postgres manager.

4. Solution evaluation

The solution proposed through the created tool focuses on the generation of an expert assessment that shows the level of reliability and toxicity of the analyzed information. This tool is available at the following link: https://socialfairness.demos.gplsi.es.

When accessing the interface, the name and a brief description of the tool is displayed, along with the list of the 20 newspapers that have been chosen for the analysis of this proof of concept and that the tool is able to analyze. The tool also includes a couple of links that contain summarized user guidelines to help readers understand the main concepts of the analysis. In addition, we have also set up three

[3]https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne
[4]https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

predefined examples, which serve as an initial test to try out the tool more quickly before choosing your own examples to analyze.

Once the news link is entered, the tool first displays the reliability analysis through pie charts, colors and percentages associated with i) each 5W1H label (Figure 3) and ii) the reliability levels (Figure 4). We use three different colours to represent the reliability levels: green (reliable), yellow (partially reliable), and red (unreliable). Also, each 5W1H label is represented by a different color in the charts.

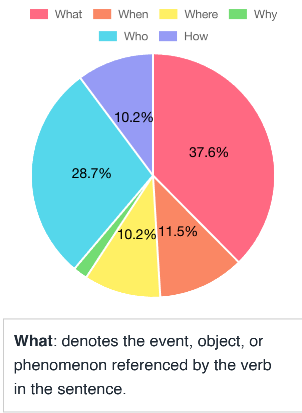

Figure 3: Pie chart with the percentage of 5W1H labels in a news item.

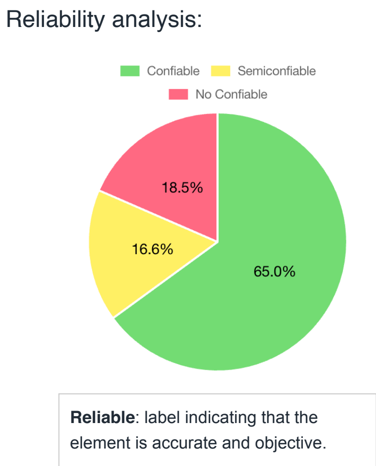

Figure 4: Pie chart showing the percentage of reliability labels of a news item.

In order to determine the reliability of a news article, thresholds are applied in order of priority. These thresholds have been defined through a comparative study between the global label assigned to the news item by expert annotators and anonymous evaluators and the percentages obtained for each label in each of the news items. As a result of this analysis, it was observed that:

- If more than 25% of the labels are unreliable, the news item is directly classified as **unreliable**.
- If this criterion is not met, it is then checked whether more than 60% of the labels are reliable; in that case, the news item is classified as **reliable**.
- If neither of the previous thresholds is met, the news item is considered **partially reliable**.

In addition to the levels and percentages, the text is displayed with annotations using the corresponding 5W1H content labels, classified according to their reliability. An example of an annotated news is shown in Figure 5.

Regarding the toxicity module, news comments are also presented with a double analysis: i) toxicity, where three colors (green, yellow and red) are associated with three different levels (not toxic, mildly

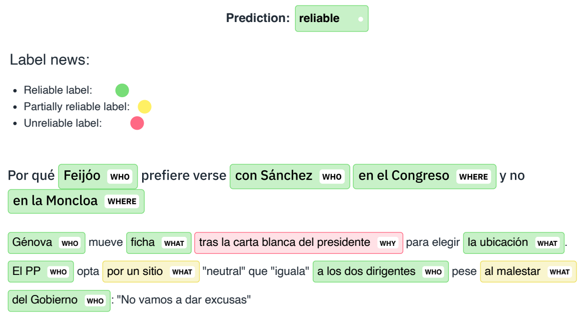

Figure 5: Example of news annotated with the 5W1H labels and their reliability.

toxic and toxic), and ii) constructiveness, represented by the symbols of a red cross (when there is no constructiveness) or a green check (in case there is constructiveness). An example of an annotated comment is shown in Figure 6.

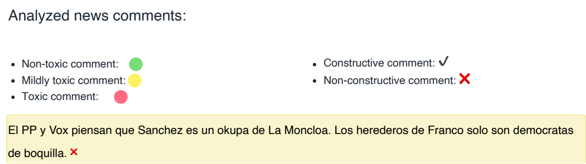

Figure 6: Example of comment annotated with the toxicity and constructiveness.

4.1. Usability and accessibility test

Once the development of the tool was completed, our objective was to identify possible usability and accessibility errors in the developed tool before its implementation. For usability evaluation, we adopted Jakob Nielsen's 10 heuristics as the evaluative approach [21].

For the evaluation, we designed a survey to allow users to determine whether each heuristic was met or not, as well as to provide suggestions for solving possible problems. In this process, we had the participation of 10 computer engineers who evaluated the usability of the system.

This survey identified that the main usability problems were related to help and documentation. In response, contextual help was incorporated into the annotated elements. Additionally, some error messages were found to be insufficiently clear, so they were rewritten to improve their understanding.

Table 5
Survey results on user satisfaction with the tool.

Satisfaction degree	Frequency	Percentage
Very satisfied	15	30.6
Satisfied	18	36.7
Somewhat satisfied	10	20.4
Not very satisfied	5	10.2
Not satisfied	1	2
Total	49	-

Regarding the usability evaluation, a survey was designed to ask about the ease of using the tool and analyzing news with it, as well as the understanding of the concepts used in the analyses. Additionally, participants were asked to analyze a news article and provide their opinion on it. Forty-nine users with diverse backgrounds (journalism, philology, AI) and of various ages participated.

Table 6
Survey results on the ease of use of the tool.

Usefulness degree	Frequency	Percentage
Totally agree	19	38.7
Somewhat agree	9	18.3
Neither agree nor disagree	12	24.4
Not very agree	3	6.1
Not agree	6	12.2
Total	49	-

After analyzing the users' responses regarding satisfaction and ease of use of the tool, the following results were obtained.

First, as can be seen in Table 5, more than half of the respondents (67.3%) are satisfied with the tool, while only 12.2% are not satisfied (or not very satisfied). After analyzing their comments, this was often due to the slow loading of the models, and therefore delays in analyzing the news, or the inability to view the news comments, issues that are already in the process of being improved.

Second, regarding the usefulness, Table 6 shows that 57% of the users consider the tool to be useful for the intended purpose, while the 18.3% do not share this view.

With accessibility in mind, both automated and manual assessments have been performed to ensure that our tool meets the criteria established in the Web Content Accessibility Guidelines (WCAG)[5].

The WAVE[6] and AChecker[7] tools were used to evaluate the accessibility of the project. Overall, it met most of the criteria set forth in WCAG 2.0. However, some problems were identified, including some elements that were not keyboard accessible (Criterion 2.1.1), problems with color contrast (Criterion 1.4.3), difficulties for people with color blindness due to certain colors (Criterion 1.4.1), inappropriate order of headings (Criterion 2.4.9) or incorrect language labels (Criterion 3.1.2). All these problems were corrected and, in addition, an accessibility statement was added in accordance with Implementing Decision (EU) 2018/1523.

5. Conclusions and future work

The developed tool aims to provide a valuable solution for both end users and journalists, helping them deliver accurate information to their readers. It seeks to generate an expert report based on different linguistic analyses on the reliability and toxicity of Spanish digital news content, in order to help readers to question the information and take into account relevant patterns when believing or disbelieving a news item.

An evaluation focusing on the tool's usability and accessibility indicated that 67.3% of users were satisfied with the tool, and that the main limitations were primarily related to the model's processing time during analysis, a process that is already being refined.

Moreover, the proposed microservices architecture enables the seamless modification and expansion of modules with flexibility and independence. We are continuously enhancing the accuracy of our AI agents by optimizing the trained models, allowing us to further refine the tool and provide more accurate reports.

[5]https://www.w3.org/WAI/WCAG22/quickref/
[6]http://wave.webaim.org/
[7]https://achecks.ca/achecker/

Acknowledgments

This research work is part of the R&D&I projects: COOLANG.CONSENSO/TRIVIAL (PID2021-122263OB-C21/PID2021-122263OB-C22) funded by MCIN/AEI/10.13039/501100011033/ and by ERDF A way of making Europe; SOCIALFAIRNESS.SOCIALTOX/SOCIALTRUST (PDC2022-133146-C21/PDC2022-133146-C21C22), funded by MCIN/AEI/10.13039/501100011033/ and by the European Union NextGenerationEU/PRTR; NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/021) funded by the Generalitat Valenciana, Spain.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly AI in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] W. A. C. Leyton, El problema de la infoxicación en el manejo de correos electrónicos corporativos, Tecnología Investigación y Academia 4 (2016) 136–141.

[2] J. A. G. Sanchez, P. A. A. Casas, Infoxicación e impacto en la sociedad: Infoxication and impact on society, Tecnología Investigación y Academia 10 (2022) 66–84.

[3] C. Bandara, Hallucination as disinformation: The role of llms in amplifying conspiracy theories and fake news, Journal of Applied Cybersecurity Analytics, Intelligence, and Decision-Making Systems 14 (2024) 65–76.

[4] T. Yallico Arias, J. Fabian, Automatic detection of levels of intimate partner violence against women with natural language processing using machine learning and deep learning techniques, in: Annual International Conference on Information Management and Big Data, Springer, 2021, pp. 189–205.

[5] Y. Ni, D. Barzman, A. Bachtel, M. Griffey, A. Osborn, M. Sorter, Finding warning markers: leveraging natural language processing and machine learning technologies to detect risk of school violence, International journal of medical informatics 139 (2020) 104137.

[6] C.-M. Tsai, Stylometric fake news detection based on natural language processing using named entity recognition: In-domain and cross-domain analysis, Electronics 12 (2023) 3676.

[7] M. Taleb, A. Hamza, M. Zouitni, N. Burmani, S. Lafkiar, N. En-Nahnahi, Detection of toxicity in social media based on natural language processing methods, in: 2022 international conference on intelligent systems and computer vision (ISCV), IEEE, 2022, pp. 1–7.

[8] A. Garlapati, N. Malisetty, G. Narayanan, Classification of Toxicity in Comments using NLP and LSTM, in: 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), volume 1, IEEE, 2022, pp. 16–21.

[9] L. A. Ureña López, M. T. Martín Valdivia, E. Saquete Boró, P. Martínez-Barco, et al., SocialFairness: Assessing Fairness in Digital Media, in: Proceedings of the XXX Workshop on Human-Computer Interaction (HCIX 2024), 2024, pp. 18–23. URL: https://ceur-ws.org/Vol-3729/.

[10] H. Zhang, X. Chen, S. Ma, Dynamic news recommendation with hierarchical attention network, in: 2019 IEEE International Conference on Data Mining (ICDM), IEEE, 2019, pp. 1456–1461.

[11] S. Mottola, Las fake news como fenómeno social. análisis lingüístico y poder persuasivo de bulos en italiano y español, Discurso & Sociedad (2020) 683–706.

[12] A. X. Zhang, A. Ranganathan, S. E. Metz, S. Appling, C. M. Sehat, N. Gilmore, N. B. Adams, E. Vincent, J. Lee, M. Robbins, et al., A structured response to misinformation: Defining and annotating credibility indicators in news articles, in: Companion Proceedings of the The Web Conference 2018, 2018, pp. 603–612.

[13] R. Sepúlveda-Torres, A. Bonet-Jover, I. Diab, I. Guillén-Pacho, I. Cabrera-de Castro, C. Badenes-Olmedo, E. Saquete, M. T. Martín-Valdivia, P. Martínez-Barco, L. A. Ureña-López, Overview of FLARES at IberLEF 2024: Fine-grained Language-based Reliability Detection in Spanish News, Procesamiento del lenguaje natural 73 (2024) 369–379.

[14] N. Salehabadi, A. Groggel, M. Singhal, S. S. Roy, S. Nilizadeh, User engagement and the toxicity of tweets, arXiv preprint arXiv:2211.03856 (2022).

[15] A. Bonet-Jover, I. Cabrera-de Castro, K. Mamchur, R. Sepúlveda-Torres, M. T. Martín-Valdivia, L. A. Ureña-López, E. Saquete, P. Martínez-Barco, Socialtox dataset, 2025. URL: https://huggingface.co/datasets/gplsi/SocialTOX. doi:10.57967/hf/5666, accessed: 2025-05-28.

[16] I. Nadareishvili, R. Mitra, M. McLarty, M. Amundsen, Microservice architecture: aligning principles, practices, and culture, O'Reilly Media, Inc., 2016.

[17] A. M. Mármol-Romero, R. Sepúlveda-Torres, I. Cabrera-de Castro, A. Bonet-Jover, M. T. Martín-Valdivia, L. A. Ureña-López, E. Saquete, P. Martínez-Barco, Toxicity classifier fine-tuned from roberta-base-bne), 2025. URL: https://huggingface.co/gplsi/Toxicity_model, accessed: 2025-05-28.

[18] I. Cabrera-de Castro, A. M. Mármol-Romero, A. Bonet-Jover, R. Sepúlveda-Torres, M. T. Martín-Valdivia, L. A. Ureña-López, E. Saquete, P. Martínez-Barco, Constructive classifier fine-tuned from roberta-base-bne, 2025. URL: https://huggingface.co/gplsi/Constructive_model, accessed: 2025-05-28.

[19] R. Sepúlveda-Torres, A. Bonet-Jover, A. M. Mármol-Romero, C. de Castro, E. Saquete, P. Martínez-Barco, M. T. Martín-Valdivia, L. A. Ureña-López, 5W1H Extractor Fine-Tuned from Llama-3B-Instruct, 2025. URL: https://huggingface.co/gplsi/5W1H_Llama_3B, accessed: 2025-05-28.

[20] A. Bonet-Jover, R. Sepúlveda-Torres, I. Cabrera-de Castro, A. M. Mármol-Romero, E. Saquete, P. Martínez-Barco, M. T. Martín-Valdivia, L. A. Ureña-López, 5W1H reliability classifier Fine-Tuned from RoBERTa-base-bne, 2025. URL: https://huggingface.co/gplsi/reliability_5W1H, accessed: 2025-05-28.

[21] J. Nielsen, R. Molich, Heuristic evaluation of user interfaces, in: Proceedings of the SIGCHI conference on Human factors in computing systems, 1990, pp. 249–256.