# Beyond Headlines: A Corpus of Femicides News Coverage in Italian Newspapers

Eleonora Cappuccio[1,2,3,*,†], Benedetta Muscato[1,4,†], Laura Pollacci[1,2],
Marta Marchiori Manerba[1,2], Clara Punzi[1,4], Chandana Sree Mala[1,4], Margherita Lalli[4],
Gizem Gezici[3], Michela Natilli[2] and Fosca Giannotti[4]

[1]*Università di Pisa, Pisa, Italy*
[2]*ISTI-CNR, Pisa, Italy*
[3]*Università degli Studi di Bari Aldo Moro, Bari*
[4]*Scuola Normale Superiore, Pisa, Italy*

**Abstract**

How newspapers cover news significantly impacts how facts are understood, perceived, and processed by the public. This is especially crucial when serious crimes are reported, e.g., in the case of femicides, where the description of the perpetrator and the victim builds a strong, often polarized opinion of this severe societal issue. This paper presents FMNews, a new dataset of articles reporting femicides extracted from Italian newspapers. Our core contribution aims to promote the development of a deeper framing and awareness of the phenomenon through an original resource available and accessible to the research community, facilitating further analyses on the topic. The paper also provides a preliminary study of the resulting collection through several example use cases and scenarios.

**Keywords**

Italian Dataset, Newspapers, Information Extraction, Information Retrieval, AI for Social Good, Femicides

## 1. Introduction

How newspapers and journalists present news plays a crucial role in shaping public understanding and perception of information. This is especially important when reporting serious crimes, such as femicides, where descriptions of the perpetrator and victim can create polarized opinions influencing readers' perceptions and interpretations of the event. According to Bouzerdan and Whitten-Woodring [1], news media often report incidents of women's homicides in a sensationalised manner, treating these crimes as isolated events rather than situating them within the bigger framework of violence against women. This narrative defies the global demands of human rights organisations to acknowledge and address this phenomenon as demanded by its intricate dynamics. Numerous countries have followed such recommendations only partially through the formal adoption of specific terminology such as *femicide* and *feminicide* in legal frameworks and public discourse. The two terms have related but distinct nuances of meaning. *Femicide*, a criminological concept initially coined in English by the feminist criminologist Diana H. Russell [2], denotes the murder of women by males due to their gender. Successively, the term femicide, translated in Castillian as *femicidio* or *feminicide* by the anthropologist Marcela Lagarde to attract political attention on the dire situation faced by women in Mexico [3], has gained global traction with varying interpretations, yet consistently denotes a patriarchal impetus behind homicides and other forms of male violence against women, primarily emphasising the sociological dimensions of abuse and the socio-political ramifications of the phenomenon. In the Italian language, the term *femminicidio* has been almost exclusively adopted, as evidenced by a Google Trends analysis comparing the search terms "femicidio" and "femminicidio" to queries regarding "femicide"[1].

An analysis of the phenomenon of femicide in the Italian context and, in particular, a linguistic investigation of it, are particularly relevant. Feminicide, a term used by the feminist movement in Italy since 2005, gained prominence in the media in 2011, especially thanks to the works of Barbara Spinelli [4]. The CEDAW Committee[2], based on data from the Shadow Report on the Implementation of CEDAW in Italy, addressed recommendations to the Italian government on feminicide in its Concluding Observations. This was the first time the committee addressed a European state on feminicide, a category previously reserved for warnings to Central

*CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy*

[*] Corresponding author.

[†] These authors contributed equally.

eleonora.cappuccio@phd.unipi.it (E. Cappuccio); benedetta.muscato@sns.it (B. Muscato)

[1] The conducted analysis included news web searches in Italy since 2022, i.e., since when the service implemented an enhanced data collection methodology.

[2] Committee on the Elimination of Discrimination Against Women.

American countries. The challenges in accurately contextualising feminicide in Italy also stem from a prolonged absence of official data, resulting in sensationalism and the perception of a dramatic rise in the crime. This may induce an emergency narrative that obscures the inherent structural dimensions of the phenomenon, thereby undermining the very essence of the term [5]. Media interpretations are essential for shaping a shared understanding across a vast audience, such as a whole country; hence, the examination of media discourse emerges as a significant analytical instrument on top of statistical evaluation of femicide data to understand the achievements and directions of state intervention towards the substantial granting of women's right to life [6].

In this regard, Aldrete and Fernández-Ardèvol [7] showed that there is a large body of empirical studies on femicide discourse across different socio-cultural contexts, which often justify the perpetrator's actions. Given the complexity of the phenomenon, a comprehensive investigation could be achieved by integrating media analysis with external data, such as demographics and current events, blending together researchers from different fields like computer science, social sciences, and complex systems science. The lack of accessible and relevant data specific to socio-culturally context where femicide is notably prevalent, such as in Italy, makes the task particularly challenging [8].

This paper presents FMNews, a new dataset of articles reporting femicides extracted from Italian newspapers[3]. We conduct a preliminary analysis of the resulting collection through several example use cases and scenarios. The primary contribution is to deepen understanding and awareness of femicide from a socio-technical perspective. We seek to examine how prominent Italian news sources report on the issue in connection to the shaping of public perception, while also offering an innovative and accessible resource to facilitate future investigation within the research community. Furthermore, this study was designed to enable a multifaceted investigation covering the following three dimensions:

- **Geographical**, with the aim to explore potential variations in framing between local and national media outlets. Indeed, previous research has shown that Italian local daily newspaper often suppress the agency of the perpetrator, portraying the events as mere occurrences [9]. We selecting newspapers reporting news at both the national[4] and local[5] level, with local editions spanning across the whole Italian territory.
- **Political**, which was granted by choosing national newspaper with varying political leanings.
- **Temporal**, where the time frame of national newspapers extends from November 2009 to February 2024, whilst that of the local ones ranges from November 2010 to February 2024[6].

## 2. Related Work

According to frame analysis, the ways in which newspapers cover news significantly impact how facts are understood, perceived, and processed by the public [10, 11]. Framing narratives means strategically including or omitting elements (such as problem definitions, explanations and evaluations) of a given situation in a communicative text [12, 13, 14]. This process aims to advocate for specific interpretations, assess moral responsibilities of individuals involved and propose solutions while also eliciting nuanced emotional responses from the audience, thereby affecting their perceptions and attitudes. It is worth noting that in the case of news articles, media framing can be seen as a demonstration of political power [10], influencing which actors or interests are involved shape narratives, often unnoticed by the audience [11]. The process of news framing becomes especially crucial when reporting serious crimes, such as femicides, as understanding femicide requires analyzing its evolution from both statistical and social perspectives, as discussed in the *Manifesto delle Giornaliste e dei Giornalisti per il Rispetto e la Parita' di Genere nell'Informazione*[7] (*Manifesto of Journalists for Respect and Gender Equality in News Reporting*, our translation).

The acknowledged impact of language on how readers perceive information has prompted researchers to explore how the language surrounding femicide has changed and how this influences individuals' responsibility perception [15], which can vary based on the way femicides are reported [1, 16, 9, 17]. Moreover, an initia-

[3] The choice of newspapers was dictated by the circulation volume released by Audipress, a company that collects data on the reading habits of daily and periodical press in Italy: https://audipress.it/quotidiani/.

[4] The selected national newspapers are the following: *Corriere della Sera, La Repubblica, La Stampa, Il Fatto Quotidiano, Il Giornale* and *Il Post.*

[5] The selected local newspapers are the local editions of the *CityNews* group, which cover the following cities: Agrigento, Ancona, Arezzo, Avellino, Bari, Bologna, Brescia, Brindisi, Caserta, Catania, Cesena, Chieti, Como, Ferrara, Firenze, Foggia, Forlì, Frosinone, Genova, Pescara, Piacenza, Latina, Lecce, Lecco, Livorno, Messina, Milano, Modena, Monza, Napoli, Novara, Padova, Palermo, Parma, Perugia, Pisa, Pordenone, Ravenna, Reggio, Rimini, Roma, Salerno, Sondrio, Terni, Torino, Trento, Treviso, Trieste, Udine, Venezia, Verona, Vicenza, Viterbo.

[6] In Fig. 3 in the Appendix, we report the distribution of articles across time.

[7] https://www.sindacatogiornalistiveneto.it/wp-content/uploads/2020/12/MANIFESTO-DI-VENEZIA.pdf.

tive by University of Bologna seeks to identify the main discursive features employed in discussions about femicide in public spaces, including media and legal speech[8].

Recognizing the significant role of linguistic expression in depicting incidents of gender-based violence, previous research has explored various NLP techniques. These studies aim to discern how NLP models can effectively predict and analyze human perception judgments concerning the sensitive issue of gender-based violence events. Following previous works on the impact of specific grammatical constructions and semantic frames [18] in describing the same event but with various nuances, Minnema et al. [19] introduced the first multilingual tool, based on Frame Semantics and Cognitive Linguistics, for detecting the focus or perspective depicted in an event, called *Socio Fillmore*. Furthermore, building on the linguistic analysis provided by Socio Fillmore, Minnema et al. [20] demonstrated that various linguistic choices trigger different perceptions of responsibility, which can be modeled automatically. As a result, their series of regression models revealed that these distinct linguistic choices significantly influence human perceptions of responsibility. Additionally, to promote awareness of perspective-based writing, Minnema et al. [21] introduced the novel task of *responsibility perspective transfer*. The task involves the automatic rewriting of descriptions of gender-based violence to alter the perceived level of blame attributed to the perpetrator. Both works leveraged one of the limited resources available for the Italian community, the `RAI Femicide Corpus`, a collection of 2.734 news articles covering 937 confirmed femicide cases in Italy happened between 2015 and 2017 [22]. Additional online resources, both official and unofficial, containing further statistics on the phenomenon of femicide in Italy are listed in the Appendix A.

# 3. FMNews Corpus

The main contribution brought by this paper is the production of two datasets derived from Italian newspapers: the `FMNews`[9] corpus. The corpus consists of the following components: `FMNews-Nat`, reporting data from national newspapers, and `FMNews-Loc`, which gathers articles from local newspapers in 53 Italian cities.

## 3.1. Data Extraction

Despite the heterogeneous HTML structures of the newspapers involved, it was feasible to generalise the data extraction process via the open source Python libraries `Selenium` [10] and `Beautiful Soup`[11]. Data scraping was performed in two subsequent phases. Firstly, a comprehensive list of article links was extracted by querying the internal search engine of the newspaper websites with the keywords `femminicidio`, `femminicidi`, `femminicida`: the first word stands for the Italian term "femicide", the second is its plural form, and the third indicates the "person who commits a femicide". The keywords were selected to concentrate our analysis on the media's representation and discourse surrounding this phenomenon. This choice intentionally excludes articles that discuss such crimes in general terms, allowing for a more focused examination of the femicide narratives. In the second phase, the web pages corresponding to such links were scraped to extract the text of the articles and other metadata to build the raw version of the dataset.

## 3.2. Data Cleaning

We implemented a supervised and semi-supervised data cleaning process, consisting of two phases, to prepare the data. In the first step, the same pipeline was applied to both `FMNews-Nat` and `FMNews-Loc`. We initially removed all duplicate articles from the collected data, i.e., those with identical texts (title and body), metadata (e.g., date), and source publication. Additionally, we converted the dates into the format of *yyyy-mm-dd* and removed articles where at least one of the following elements was missing: publication date, title, or body. Despite the removal of duplicates, certain articles had identical text bodies, albeit with minor variations primarily due to special character encoding (e.g., accents and apostrophes) or differences in web crawling (e.g., one article included the website menu or footer while the other did not). To address this issue, we implemented a method to identify and handle articles with identical or highly similar text bodies sharing the same title. In details, we first employed a TF-IDF[12] vectorizer to convert the raw text data into numerical vectors and then use them to compute the cosine similarities between all pairs of texts in the dataset. For more details on the parameters and thresholds employed, we refer to Appendix B. Finally, we utilized `Beautiful Soup` to remove any HTML tags that could have been mistakenly included in the article body during the collection phase.

The second step of the data cleaning process entailed supervised cleaning of the article texts and headlines. The article texts from national newspapers in `FMNews-Nat` displayed various noise patterns specific to each news media outlet. To address this issue, we manually created

[8] https://site.unibo.it/osservatorio-femminicidio/it.

[9] The collection can be accessed for research purposes by requesting it by email from the authors.

[10] https://selenium-python.readthedocs.io/.

[11] https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[12] Term Frequency-Inverse Document Frequency, in short TF-IDF, is a measure of the importance of a word to a document in a collection or corpus [23].

| Column | Description |
|---|---|
| Url | URL of the original newspaper article |
| Title | Title of the article |
| Text | Main section of the newspaper article |
| Newspaper | Name of the media outlet where the article was published. In FMNews-Loc, it reports the name of the city to which the local edition refers to. |
| Keyword | Keyword used to collect the article |
| Date | Publication date of the article in the format *yyyy-mm-dd* |

**Table 1**
Description of the FMNews Corpus.

| Dataset | Raw Data | Step I | Step II |
|---|---|---|---|
| FMNews-Nat | 12,790 | 7,511 | 7,443 |
| FMNews-Loc | 8,397 | 7,728 | 7,728 |

**Table 2**
Dimensions of the dataset in terms of number of articles from national news outlets (FMNews-Nat) and local newspaper editions (FMNews-Loc).

a list of replacements for each outlet, employing regular expressions for targeted removal of articles or specific sub-strings from article titles or bodies (we refer to Appendix B for additional details). In this stage, we also excluded articles whose text bodies did not contain information directly related to femicides, such as television programme listings or podcast episode agendas.

On the other hand, the articles from local newspapers in FMNews-Loc exhibited minimal noise within their text. Therefore, the data preparation phase focused on poorly encoded symbols and domain-specific substrings such as copyright indications and external contributions, e.g., government press releases. Unlike national newspapers, for journalistic publications, this ad-hoc cleaning did not result in data loss.

### 3.3. Final Dataset

Table 1 provides a detailed explanation of the data format for both datasets after the completion of the data preparation process. The number of entries for the two datasets is shown in Table 2. The table also shows the number of articles after two steps of data cleaning exemplified in B.

The analysis of FMNews-Nat after the last cleaning steps reveals the following summary statistics. The dataset covers a time span of 14 years, from November 2009 to February 2024. Regarding the distribution of articles across different newspapers in FMNews-Nat: *Il Fatto Quotidiano* has the largest number of articles, with a total of 2,861, followed by *La Repubblica* with 2,837 articles. *Corriere* is next, with a total of 968 articles. *La Stampa* has a more limited presence, with 292 articles. Il Post contributes 244 articles, and *Il Giornale* has the fewest entries in this set, with 241 articles. For FMNews-Loc, the time span after data cleaning ranges from November 2010 to February 2024.

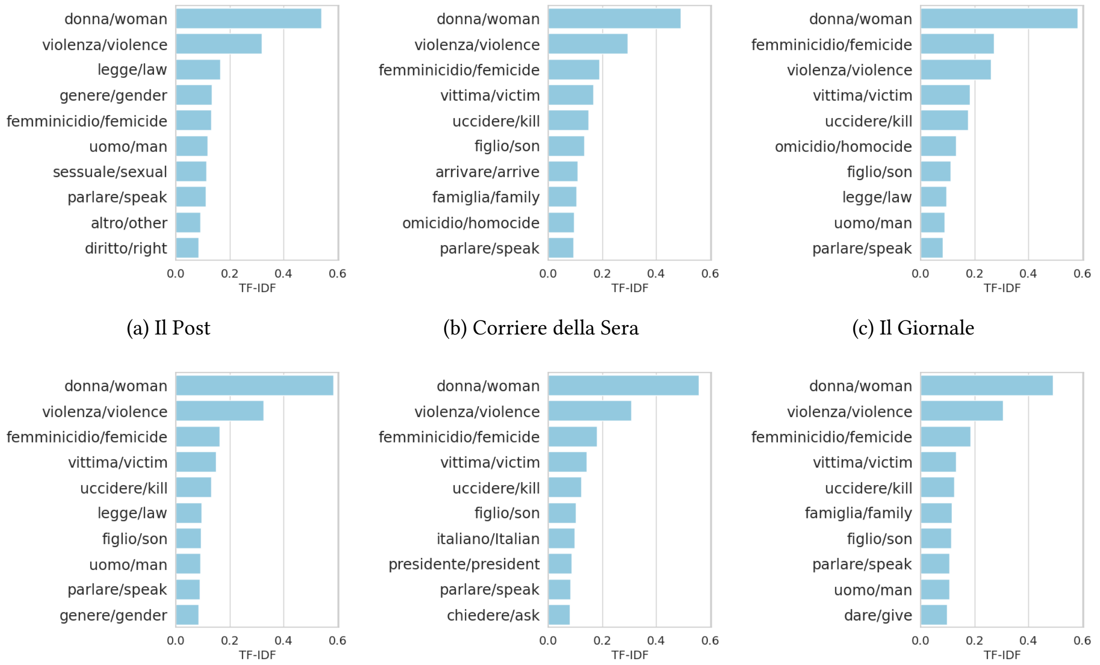
## 4. Use Cases and Scenarios

Since the two datasets share the same structure and we are interested in studying the phenomenon of femicide from both a national and local perspective, the analyses exemplified in the following were conducted on both datasets without distinction. After a textual analysis based on the tokenization, removal of stopwords, extraction of lemmas and a straightforward assessment of the lexical diversity (as detailed in the Appendix C), we approached a viable keyword extraction method to uncover relevant patterns in the documents.

**Keyword Extraction** According to Firoozeh et al. [24], specific criteria must be met for keywords to meet eligibility standards. In our case study, we emphasize the importance of keywords that show *representativity* and *exhaustivity*, aiming for terms that capture significant rather than marginal aspects of the subject matter. To assess the significance of words within our collection of documents, a standard approach involves the Term Frequency - Inverse Document Frequency (TF-IDF).

For a deeper analysis, we calculate TF-IDF for each news outlet. We utilize Spacy's Italian pipeline to preprocess texts by tokenizing, lemmatizing, and selecting only lemmas that are full words from specific part-of-speech classes (nouns, adjectives, verbs). By focusing only on content lemmas and excluding function words (like articles and prepositions), we eliminate noise and improve accuracy in analyzing relationships between documents and word relevance. The lists of lemmas do not include words containing numbers or Italian stopwords obtained from Nltk and Spacy, with additional crawling-dependent stopwords such as "it," "https," "min," and the names of months. Also, we preserve multi-word expressions identified by the lemmatizer by concatenating them to treat them as unique words during TF-IDF calculation. Articles are then grouped by news outlet, each acting as a single document for the TF-IDF computation. We use the TF-IDF Vectorizer from the scikit-learn[13] library to transform the lemmatized tokens into numerical features that reflect their importance within the text.

[13] https://scikit-learn.org/stable/index.html.



(a) Il Post

(b) Corriere della Sera

(c) Il Giornale

(d) Fatto Quotidiano

(e) La Repubblica

(f) La Stampa

**Figure 1:** Top 10 keywords in descending order for each news outlet `FMNews-Nat`.

Thus, TF-IDF measures the significance of terms concerning the news outlets. Fig. 1 illustrates the most relevant keywords extracted from `FMNews-Nat` by news outlet. As expected, terms like "woman," "violence," and "kill" (along with "femicide") are central to the narrative of femicide and are common across all outlets. Other keywords vary in relevance among multiple outlets; for example, "son" appears in all outlets except Il Post. Specific keywords are unique to one or two outlets: "gender," "right," and "sexual" appear only in Il Post; "family" is relevant in Corriere della Sera and La Stampa; and "man" is found in Il Post and Il Giornale. Due to the number of local news outlets in `FMNews-Loc` (50), Fig. 2 shows the top 20 keywords with the highest average TF-IDF, calculated as the mean of the TF-IDF values of the terms with respect to the news outlets. As expected, the highest ranks are occupied by the same relevant keywords found in national news outlets, such as "woman," "violence," "victim," and "femicide". Additionally, some keywords relevant to specific national news outlets show high relevance for local media, although with lower average TF-IDFs, such as "gender". Conversely, the distribution reveals previously unseen keywords, such as "young," "school," and "association".

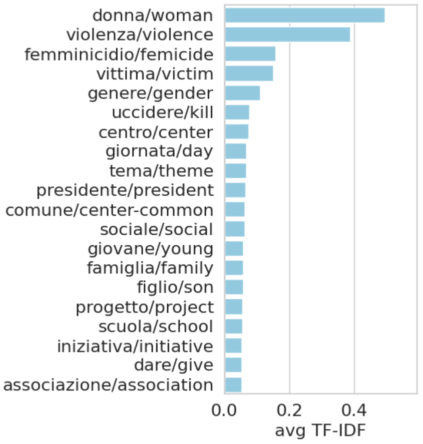


**Figure 2:** Top 20 Keywords by average TF-IDF in `FMNews-Loc`.

**Semantic Vector Extraction** For an additional layer of analysis, we chose to train a word embedding model to explore semantic relationships among words. This model represents words as continuous space vectors, where the proximity of vectors indicates the semantic similarity between the words they represent: closer vectors

**Table 3**
Most similar word embeddings to

(a) "uccidere" (to kill) in `FMNews-Nat`

| Word | Similarity score |
|---|---|
| *ammazzare* (to murder) | 0.77 |
| *ucciderla* (to kill - her) | 0.71 |
| *ammazzato* (murdered - him) | 0.66 |
| *ucciso* (killed - him) | 0.66 |
| *suicidarsi* (to commit suicide) | 0.63 |
| *strangolato* (strangled - him) | 0.62 |
| *furia* (fury) | 0.60 |
| *fucile* (rifle) | 0.59 |
| *sparare* (to shoot) | 0.59 |
| *accoltellato* (stabbed - him) | 0.59 |

(b) "vittima" (victim) in `FMNews-Loc`

| Word | Similarity score |
|---|---|
| *ragazza* (girl) | 0.69 |
| *giovane* (young) | 0.69 |
| *donna* (woman) | 0.67 |
| *madre* (mother) | 0.67 |
| *figlia* (daughter) | 0.64 |
| *scomparsa* (disappearance) | 0.62 |
| *uccisa* (killed - her) | 0.62 |
| 26*enne* (26 years old) | 0.61 |
| *massacrata* (massacred - her) | 0.59 |
| *povera* (poor) | 0.59 |

correspond to words with more similar meanings. We employed `Word2Vec` (W2V) [25], which operates by mapping words to high-dimensional vectors within a given vocabulary. This mapping is designed to represent semantic relationships between words in the vectorial space. W2V has been implemented through `Gensim`[14], a powerful tool set for NLP tasks. A key parameter in W2V is the "window", i.e., the number of context words to be considered, which we defined as 10 to consider a contextual window that extends neither too far nor too close to the current word, thereby striking a balance between contextual relevance and computational efficiency. To discover the semantic associations within our dataset, we leveraged the "most similar" method from `Gensim`, which computes the cosine similarity between word vectors to identify words with the closest semantic proximity. For both datasets the size of the training embeddings for the W2D model is fixed to 100 while the vocabulary size change accordingly to the dataset, in `FMNews-Nat` is 6809, in `FMNews-Loc` is 6064.

In `FMNews-Nat`, the word "donna" (woman) yielded semantically related terms such as "vittima" (victim) and "prostituta" (whore). The term "femminicidio" (femicide) elicited associations like "violenza" (violence), "impressionante" (impressive), and "dramma" (drama). In Table 3a, the analysis of "uccidere" (to kill) encompasses related terms such as "ammazzare" (to murder), "ucciderla" (to kill her), "ammazzato" (murdered, masculine form), "ucciso" (killed, masculine form), "suicidarsi" (to commit suicide), and "strangolato" (strangled, masculine form). These terms may collectively pertain to the perpetrator's actions against the victim. Fig. 5 in the Appendix provides a comprehensive overview of word vectors closely associated with the previously extracted keywords, which were identified as the most significant in `FMNews-Nat`.

In Table 3b, the words correlated in meaning to "vittima" (victim) in `FMNews-Loc` are presented. As we would expect, nearly all terms are associated and highlight that the victim is a woman. In this regard, a drawback to consider is that the specific selection of the terms used for the data collection query may have hindered our analysis from uncovering insights about homicides committed against individuals who do not identify as woman or fit into the traditional gender binary. Indeed, the discussion around gender-based violence in Italy is still predominantly centred on women, while other genders remain significantly neglected[15].

## 5. Conclusion

In this contribution, we provided a novel dataset concerning the critical issue of femicide in Italy. Considering the absence of resources for conducting in-depth analyses on the subject, our intent was to bridge this gap and provide an original perspective for understanding and raising awareness about this severe phenomenon.

As suggested by Dobbe et al. [26], proposing a contribution within the Machine Learning domain responsibly and consciously means foremost acknowledging our own biases. In particular, we are referring to both the newspaper selection and choice of the terms used to extract the data, that certainly shaped the results (all design choices are justified in detail in Section 3). A future outlook concerns the investigation of how both victims and perpetrators are framed from a linguistic perspective. Further analyses could regard identifying temporal and geographical patterns arising from media attention manifested through the coverage of femicides and comparing the framing of these events with the political leaning of the respective newspapers.

[14] https://pypi.org/project/gensim/.

[15] As a matter of fact, there is no official collection of statistics regarding this specific kind of event. The only organisation that records the gender of the victims in its database is the *Observatory Femicides Lesbicides Transcides* managed by *Non una di meno*, the Italian section of movement *Ni una menos* (https://osservatorionazionale.nonunadimeno.net/).

## Acknowledgments

This work has been supported by the European Union under ERC-2018-ADG GA 834756 (XAI), by HumanE-AI-Net GA 952026, by the Partnership Extended PE00000013 - “FAIR - Future Artificial Intelligence Research” - Spoke 1 “Human-centered AI”, and by SoBigData.it that receives funding from European Union – NextGenerationEU – National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) – Project: “SoBigData.it – Strengthening the Italian RI for Social Mining and Big Data Analytics” – Prot. IR0000013 – Avviso n. 3264 del 28/12/2021.

## References

[1] C. Bouzerdan, J. Whitten-Woodring, Killings in context: An analysis of the news framing of femicide, Human Rights Review 19 (2018) 211–228.

[2] J. Radford, D. Russell, Femicide: The Politics of Woman Killing, Post-Contemporary Interventions, Twayne, 1992.

[3] M. M. L. y de los Ríos, Por la vida y la libertad de las mujeres: fin al feminicidio, Cámara de Diputados del Congreso de la Unión, LIX Legislatura, Comisión Especial para Conocer y Dar Seguimiento a las Investigaciones Relacionadas con los Feminicidios en la República Mexicana y a la Procuración de Justicia Vinculada, 2006.

[4] B. Spinelli, Femminicidio: dalla denuncia sociale al riconoscimento giuridico internazionale, Franco Angeli, 2008.

[5] B. Spinelli, L’italia rispetta la CEDAW? il femminicidio in italia alla luce delle raccomandazioni delle nazioni unite, in: I. Corti (Ed.), Universo femminile. La CEDAW tra diritto e politiche, eum edizioni università di Macerata, 2012.

[6] S. Abis, P. Orrù, et al., Il femminicidio nella stampa italiana: un’indagine linguistica, gender/sexuality/italy 3 (2016) 18–33.

[7] M. Aldrete, M. Fernández-Ardèvol, Framing femicide in the news, a paradoxical story: A comprehensive analysis of thematic and episodic frames, Crime, Media, Culture (2023) 17416590231199771.

[8] A. Forciniti, E. Zavarrone, Data quality and violence against women: The causes and actors of femicide, Social Indicators Research (2023) 1–25.

[9] C. Meluzzi, E. Pinelli, E. Valvason, C. Zanchi, Responsibility attribution in gender-based domestic violence: A study bridging corpus-assisted discourse analysis and readers’ perception, Journal of pragmatics 185 (2021) 73–92.

[10] R. M. Entman, Framing: Toward clarification of a fractured paradigm, Journal of Communication 43 (1993) 51–58. doi:10.1111/j.1460-2466.1993.tb01304.x.

[11] J. James W.Tankard, The empirical approach to the study of media framing, in: S. D. Reese, J. Gandy, A. E. Grant (Eds.), Framing public life, Taylor & Francis, Philadelphia, PA, 2001.

[12] M. Edelman, Contestable categories and public opinion, Political Communication 10 (1993) 231–242. doi:10.1080/10584609.1993.9962981.

[13] D. Kahneman, A. Tversky, Choices, values, and frames., American Psychologist 39 (1984) 341–350. doi:10.1037/0003-066x.39.4.341.

[14] P. M. Sniderman, R. A. Brody, P. E. Tetlock, Cambridge studies in public opinion and political psychology: Reasoning and choice: Explorations in political psychology, Cambridge University Press, Cambridge, England, 1993.

[15] C. Corradi, C. Marcuello-Servós, S. Boira, S. Weil, Theories of femicide and their significance for social research, Current sociology 64 (2016) 975–995.

[16] J. Fairbairn, C. Boyd, Y. Jiwani, M. Dawson, Changing media representations of femicide as primary prevention, in: The Routledge International Handbook on Femicide and Feminicide, Routledge, 2023, pp. 554–564.

[17] E. Pinelli, C. Zanchi, Gender-based violence in italian local newspapers: How argument structure constructions can diminish a perpetrator’s responsibility, in: Discourse Processes between Reason and Emotion: A Post-disciplinary Perspective, Springer, 2021, pp. 117–143.

[18] G. Minnema, S. Gemelli, C. Zanchi, V. Patti, T. Caselli, M. Nissim, et al., Frame semantics for social nlp in italian: Analyzing responsibility framing in femicide news reports, in: CEUR WORKSHOP PROCEEDINGS, volume 3033, CEUR-WS, 2021, pp. 1–8.

[19] G. Minnema, S. Gemelli, C. Zanchi, T. Caselli, M. Nissim, Sociofillmore: a tool for discovering perspectives, arXiv preprint arXiv:2203.03438 (2022).

[20] G. Minnema, S. Gemelli, C. Zanchi, T. Caselli, M. Nissim, Dead or murdered? predicting responsibility perception in femicide news reports, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online only, 2022, pp. 1078–1090. URL: https://aclanthology.org/2022.aacl-main.79.

[21] G. Minnema, H. Lai, B. Muscato, M. Nissim, Responsibility perspective transfer for Italian femicide news, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for

Computational Linguistics, Toronto, Canada, 2023, pp. 7907–7918. URL: https://aclanthology.org/2023.findings-acl.501.

[22] M. Belluati, Femminicidio, Una lettura tra realtà e interpretazione. Biblioteca di testi e studi. Carocci (2021).

[23] A. Rajaraman, J. D. Ullman, Data mining, in: Mining of Massive Datasets, Cambridge University Press, Cambridge, 2011, pp. 1–17. doi:10.1017/CBO9781139058452.002.

[24] N. Firoozeh, A. Nazarenko, F. Alizon, B. Daille, Keyword extraction: Issues and methods, Natural Language Engineering 26 (2020) 259–291.

[25] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[26] R. Dobbe, S. Dean, T. K. Gilbert, N. Kohli, A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics, CoRR abs/1807.00553 (2018). URL: http://arxiv.org/abs/1807.00553. arXiv:1807.00553.

# A. Additional Resources

## Official Resources

Official statistics on femicide cases in Italy can be accessed through ISTAT[16] and the Ministry of the Interior through the Department of Public Security website[17]. In particular, ISTAT provides data on victims of voluntary homicide, divided by gender, from 1992 to 2020, without additional information. In contrast, the Department of Public Security offers more detailed data covering a limited time range, i.e., from 2002 to 2022: victims are categorized by their relationship to the murderer. These categories include: *Partner* (husband/wife, domestic partner, boyfriend/girlfriend), *Former partner* (former husband/wife, former domestic partner, former boyfriend/-girlfriend), *Other relative, Other acquaintance, Perpetrator unknown to the victim*, and *Perpetrator unidentified.*

## Unofficial Resources

Unofficial data and statistics regarding femicides in Italy are also available, typically compiled by non-governmental or grassroots organisations. One notable example is the open database[18] managed by the Italian activists of *Ni una menos*[19], an international feminist movement that campaigns against gender-based violence. Although it covers a shorter time frame, this database offers disaggregated and more detailed information than the official statistics. For example, in addition to the names of the victims, the collection also includes important characteristics such as the age and nationality of the individuals involved, the geographical dimension, and the gender of the victim, including non-binary framings. While not readily accessible, a combined examination of both official and non-official data is essential for a more thorough and comprehensive analysis of the issues of femicide in Italy.

# B. Data Preparation

We applied a supervised and semi-supervised cleaning phase divided into two steps to prepare the data. In the first step, the same pipeline was applied to both datasets, primarily aimed at removing duplicate articles, formatting metadata, and reducing data and metadata sparsity. The second step entailed supervised cleaning of the article texts and headlines. We observed different types of noise in the texts of the national newspapers compared

[16] https://www.istat.it/it/violenza-sulle-donne/il-fenomeno/omicidi-di-donne.

[17] https://www.interno.gov.it/it/stampa-e-comunicazione/dati-e-statistiche/omicidi-volontari-e-violenza-genere.

[18] https://osservatorionazionale.nonunadimeno.net/anno/.

[19] https://nonunadimeno.wordpress.com/.

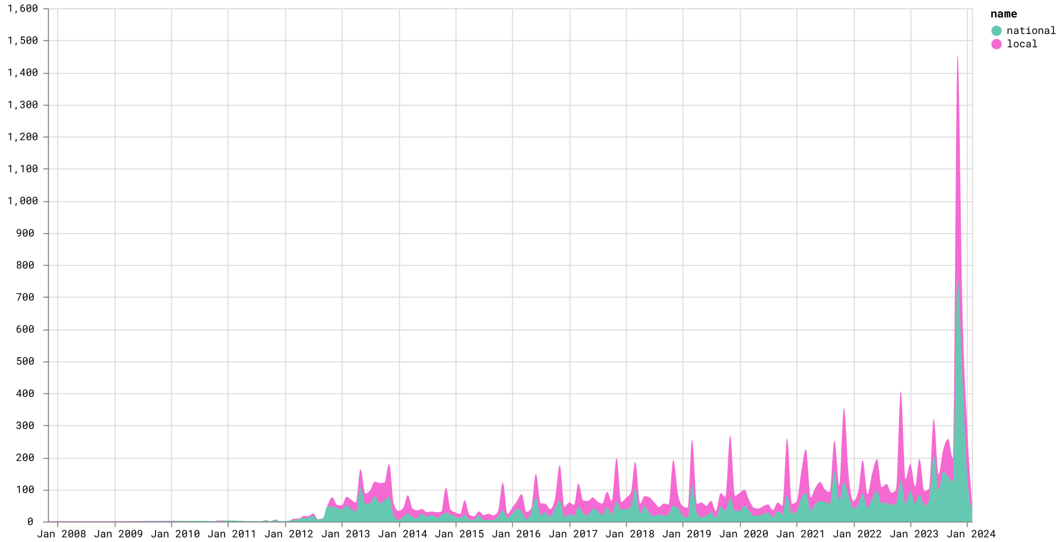


**Figure 3:** Number of articles throughout the years (2008-2024) for both `FMNews-Nat` and `FMNews-Loc`.

to the local ones. Hence, given that the two datasets are released and usable separately, we implemented a similar pipeline for both datasets, albeit customized for each.

## Data Preparation - Step I: Cleaning

We first removed all duplicate articles from the collected data (just under 12,800 articles from national newspapers and approximately 8,400 articles from local ones), i.e., those with identical texts (title and body), metadata (e.g., date), and source publication. Additionally, we converted the dates into the format of *yyyy-mm-dd* and removed articles where at least one of the following elements was missing: publication date, title, or body. Despite the removal of duplicates, some articles had identical text bodies, albeit with minor variations primarily due to special character encoding (e.g., accents and apostrophes) or differences in web crawling (e.g., one article included the website menu or footer while the other did not). To address this issue, we implemented a method to identify and handle articles with identical or highly similar text bodies, but only if they share the same title. The method relies on cosine similarity to determine whether two texts are the same. In particular, we first employed a TF-IDF vectorizer to convert the raw text data into numerical vectors. These vectors were then used to compute the cosine similarities between all pairs of texts in the dataset. Cosine similarity produces a value between 0 and 1, where 1 indicates identical texts and values closer to 0 indicate less similar texts. Since text preprocessing had not been performed yet and differences between text bodies could solely arise from symbols, we set a tolerance threshold of 0.89 to determine text equality. If two text bodies had a cosine similarity greater than 0.89, we considered them duplicates and retained only the first occurrence, removing the second found in the dataset. Finally, we utilized `Beautiful Soup` to remove any HTML tags that could have been mistakenly included in the article body during the collection phase. This step ensured that our text data was free from any undesired HTML tags before further processing or analysis.

## Data Preparation - Step II: `FMNews-Nat`

The article texts from national newspapers displayed various noise patterns specific to each news media outlet. To address this issue, we manually created a list of replacements for each outlet, employing regular expressions for targeted removal of articles or specific sub-strings from article titles or bodies. In particular, the body of articles from *Il Post, La Repubblica* and *Il Fatto Quotidiano* included parts of webpage menus and footers, as well as various types of news media outlet sponsorship, such as subscriptions, newsletter sign-ups, and agendas/lists of podcast episodes. On the other hand, articles from *Corriere della sera* included text substrings associated with the journalistic domain, such as headings containing the name of the correspondent, reporter, or photographer. We observed that the texts of the articles published by *Corriere della sera* often, but not always, follow a particular structure: "by Author_name Author_surname" (where <Author_name Author_surname> can be a natural person or abbreviations with one dot) or "Editorial team", followed by a city or "online", in either uppercase or lowercase. Occasionally, this structure is followed by another city, for instance, "Bologna Online Editorial Staff". Additionally, this "basic" structure may or may not be followed by "*inviato a* <City> <(Province)>", or "*inviata*", "*foto di* <Author_name Author_surname>". We generally excluded articles whose text bodies did not contain information directly related to femicides, such as television programme listings or podcast episode agendas. We retained the article whenever feasible, removing irrelevant substrings from the text bodies, such as menus and footers. The resulting `FMNews-Nat` dataset includes $7,443$ articles: in Fig. 4 we report the distribution of articles by media outlet.

## Data Preparation - Step II: FMNews-Loc

The articles from local newspapers exhibited minimal noise within their text. Therefore, the data preparation phase focused on poorly encoded symbols and domain-specific substrings such as copyright indications and external contributions, e.g., government press releases. Unlike national newspapers, for journalistic publications, this ad-hoc cleaning did not result in data loss . Therefore, the resulting `FMNews-Loc` dataset includes $7,728$ articles.

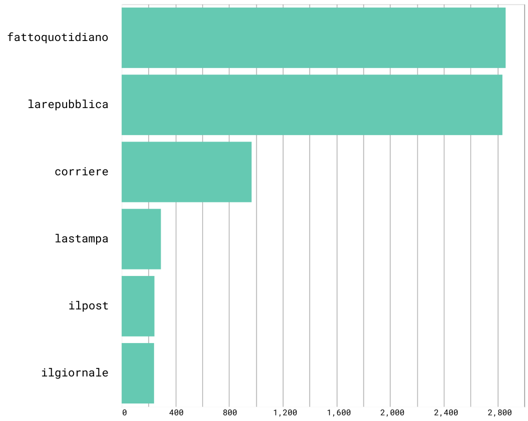


**Figure 4:** Final number of articles of `FMNews-Nat` extracted from the national newspapers.

# C. Textual Analysis

Although applying NLP models typically requires standardized and structured text, it is important to acknowledge that such preprocessing may result in the loss of some information. We believe it is important to keep track into texts of the elements we manipulate.

- **Emails and URLS.** Emails and URLs found within the body of the articles are replaced with a placeholder tag, such as "[[URL]]".
- **Uppercase words.** Words entirely in uppercase are not replaced or modified, as the text will be normalized in subsequent stages of the work, i.e., converted to lowercase. Uppercase words are extracted and saved for further analysis.
- **Punctuation, symbols, numbers.** Punctuation, symbols, and numbers are removed from the texts.
- **Stopwords.** We remove the stopwords included in the list provided by `NLTK` [20] and `Spacy`[21] libraries, along with a brief, manually compiled list of stopwords. This latter list includes domain-specific and context-related keywords, such as "Link Embed", "FOTO", "FOTOGRAMMA". It is important to note that the "ad hoc" stopwords were removed from the non-normalized text to mitigate the impact of stopwords removal. Indeed, during the analysis, we observed that some articles from national newspapers contained certain keywords entirely in uppercase to indicate elements attached to the article. Thus, we chose to compile the list of stopwords to be case-sensitive, aiming to avoid removing words within the body of the article.

After extracting the features from the raw texts, we proceeded with the following steps. First, we tokenized the body of articles using the `Spacy` library with the Italian module, selecting only words. Next, we extracted tokens that are not included in the stopwords. Then, we extracted the lemmas, again excluding stopwords. Finally, we further refined our selection by retaining from the tokens only words belonging to what is commonly referred to as "full" classes of speech, such as nouns, verbs, adjectives, and adverbs. This process of extracting "full" words aimed to focus our analysis on linguistically significant elements of the text. This approach allows us to study meaningful linguistic units, facilitating a more accurate understanding of the semantic content and structure of the text.

After tokenization, removal of stopwords, and extraction of lemmas, we computed the Type-Token Ratio (TTR) for the articles, a measure of the lexical diversity in a text. This is given by the proportion of unique words in a text, or "types", to the total number of words, or "tokens" and reads:

$$TTR = \frac{N_{\text{types}}}{N_{\text{tokens}}} \tag{1}$$

[20] https://www.nltk.org/.
[21] https://spacy.io/.

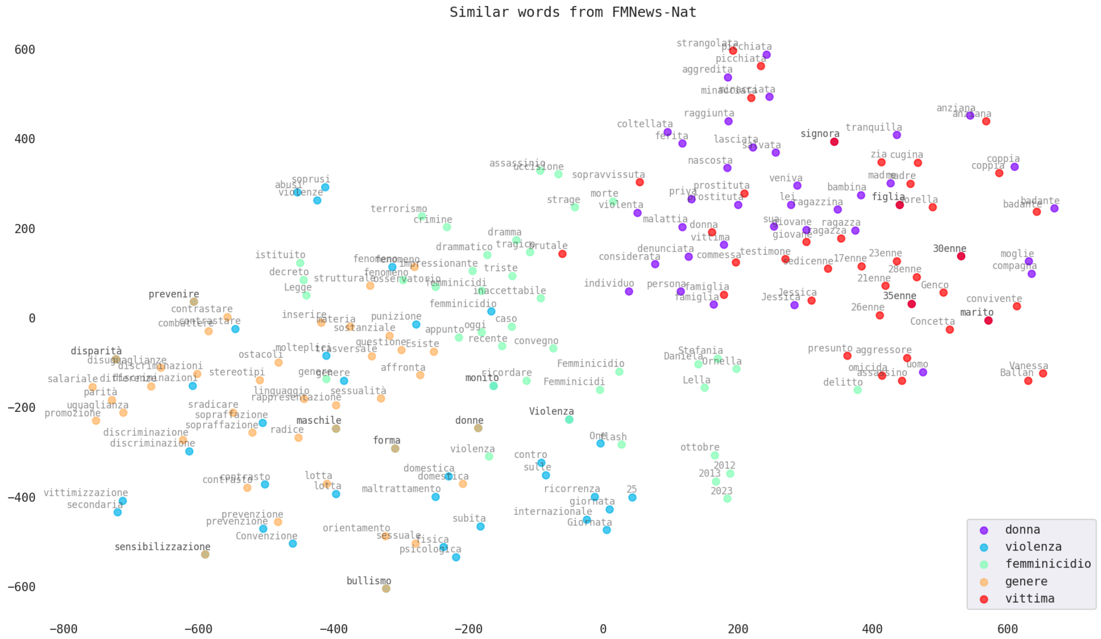


**Figure 5:** Similar word vectors in FMNews-Nat.

Where $N_{\text{types}}$ is the number of unique types and $N_{\text{tokens}}$ is the number of tokens in the text. TTR values range from 0 to 1, where a higher value indicates greater lexical variety, whereas a lower value implies more repetition of words in the text. This is a straightforward measure which nevertheless allows us to form an initial assessment of the lexical richness in the narrative surrounding femicides. The newspaper *Il Post*, along with *Il Fatto Quotidiano* and *La Repubblica*, exhibited a notable variation in terms of TTR. While FMNews-Nat shows variation in lexicon usage, FMNews-Loc exhibits a uniformity in language .