# Exploring YouTube Comments Reacting to Femicide News in Italian

Chiara Ferrando[1,*,†], Marco Madeddu[1,*,†], Beatrice Antola[2], Sveva Silvia Pasini[3], Giulia Telari[3], Mirko Lai[4] and Viviana Patti[1]

[1]*Università di Torino, Italy*

[2]*Università di Padova, Italy*

[3]*Università di Pavia, Italy*

[4]*Università del Piemonte Orientale, Italy*

### Abstract

In recent years, the Gender Based Violence (GBV) has become an important issue in modern society and a central topic in different research areas due to its alarming spread. Several Natural Language Processing (NLP) studies, concerning Hate Speech directed against women, have focused on misogynistic behaviours, slurs or incel communities. The main contribution of our work is the creation of the first dataset on social media comments to GBV, in particular to a femicide event. Our dataset, named GBV-Maltesi, contains 2,934 YouTube comments annotated following a new schema that we developed in order to study GBV and misogyny with an intersectional approach. During the experimental phase, we trained models on different corpora for binary misogyny detection and found that datasets that mostly include explicit expressions of misogyny are an easier challenge, compared to more implicit forms of misogyny contained in GBV-Maltesi.

*Warning*: This paper contains examples of offensive content.

### Keywords

Hate Speech, Misoginy Detection, Femicide, Social media, News, Responsibility framing

## 1. Introduction

Nowadays, the term **Gender Based Violence (GBV)** is used to identify all forms of abuse based on gender hatred and sexist discrimination [1]. Scholars in social science have defined as "rape culture" the society that normalizes sexist behaviours: from more common occurrences like victim blaming, slut shaming and gender pay gap to the apex of violence with femicide [2]. While general violent crimes decreased over time, GBV did not, alarming various bodies in modern society[1]. A report from the EU commission[2] states that 31%, 5% and 43% of European women suffered respectively from physical, sexual and psychological violence. Regarding the Internet sphere, a survey found that 73% of women journalists experienced online violence (threats, belittling, shaming,...) [3]. These

statistics become even more alarming when we consider studies that show the correlation between misogynistic online posts and GBV [4].

Like other countries, Italy is affected by GBV, with the national observatory managed by the "Non Una di Meno" association reporting 117 femicides in 2022, 120 in 2023 and more than 40 until June 2024[3].

Several studies about Hate Speech (HS) directed towards women often focus on developing taxonomies [5] rather than investigating low resource subjects in computational linguistics like GBV. These works often gather corpora by keyword search of gender slurs [6], retrieving comments left on misogynistic spaces like incel blogs [5, 7] or considering messages directed towards popular women figures highly debated on social media [8].

As GBV is a broad topic, we want to clarify that we focus on GBV in Western societies, particularly in Italy. The main goal of this project is to show what is the current perception of femicides expressed through comments on social media, focusing on the specific case of Carol Maltesi. We chose this femicide because the victim was a sex worker, meaning that she presented an intersectional trait, and it was a popular case in the media, enabling us to select enough material for the study. Further, we want to highlight how the socio-demographic characteristics of the victims determine the way they are described and how this influences the perception of the news. For instance, victim's features such as age, job, origin, skin

---

✉ chiara.ferrando@unito.it (C. Ferrando);
marco.madeddu@unito.it (M. Madeddu);
beatrice.antola@studenti.unipd.it (B. Antola);
svevasilvia.pasini01@universitadipavia.it (S. S. Pasini);
giulia.telari01@universitadipavia.it (G. Telari); mirko.lai@uniupo.it (M. Lai); viviana.patti@unito.it (V. Patti)

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

[2]https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/gender-equality/gender-based-violence/what-gender-based-violence_en

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

color, nationality, religion have different weight and determine the lesser or greater spread of the news [9]. To overcome the cited issues in current literature, in this research we considered the phenomenon by focusing on users' reactions in social media to news about femicides. We collected YouTube comments in response to videos talking about a specific case. In order to overcome the constraints of traditional sentiment analysis schemas, we annotated the data following a new semantic grid that can be used as a standard for comments regarding GBV.

In the experimental phase of this work, we created models based on different Italian misogyny datasets (including ours). The goal of such experiments is to analyze the different features of these corpora and what forms of misogyny are harder to detect. We performed both a quantitative and qualitative analysis of the results.

In the next sections, we describe: related work on hate speech and misogyny detection(Section 2), the annotation scheme and both a quantitative and qualitative analysis of the dataset (Section 3), and the results obtained in our experiments (Section 4). Lastly, we present some conclusions and delineate possible future developments (Section 5).

## 2. Related Work

In recent times, the creation and dissemination of hate speech are increasingly pervasive on online platforms, making social media a fertile ground for hateful discussions [10]. The escalation of offensive and abusive language, understood as content that discriminates a person or group on the basis of specific characteristics such as ethnicity, gender, sexual orientation, and more has aroused considerable interest in various fields. In fact, over the last decade, a large number of computational methods involving NLP and Machine Learning have been proposed for automatic online hate speech detection [11, 12]. Most of prior works have mainly considered hate speech as a classification task, by distinguishing between hate and non-hate speech. Hate speech takes on different nuances depending on the target groups at which it is directed, i.e. depending on the specific features that the target group have in common. Moreover, in some cases, these traits may intersect with each other, leading to different degrees of discrimination. This concept takes the name of intersectionality [13].

Among abusive languages, misogyny, considered as a specific offensive language against women, has become a contemporary research topic [14]. In automatic hate speech detection field, the Automatic Misogyny Identification (AMI) [15] series of shared tasks launched in EVALITA [6] and the SemEval-2019 HatEval challenge [16] have produced evaluation frameworks to identify misogynous tweets in English, Italian and Spanish [17].

Misogyny has become a pervasive phenomenon, widespread in very different spheres and expressed in both explicit and implicit forms [5, 18]. For this reason, even in online conversation about a dramatic act such as femicide, it is possible to find examples of veiled or explicit hostility towards the victims. The femicide phenomenon has been studied from different points of view. Several studies focused on GBV representation in Italian media [19, 20]. In 2020, Mandolini focused on the journalistic narratives of femicide in newspapers by means of a qualitative discourse analysis on two specific case studies [21]. The researcher attempted to describe changes in attitudes in the portrayal of femicide, focusing on discursive strategies that (directly or indirectly) blame the victim and implicitly excuse the perpetrator, referring to gender stereotypes and romantic love rhetoric.

Other studies focused on the responsibility framing in femicides news, by conducting an experiment where annotators rated excerpts from local newspapers on how much responsibility was given to the perpetrator [22]. As far as we know, there is only one line of work in NLP on GBV [23, 24, 25], which focuses on reader's perception of femicide news headlines and analyses the perception of responsibility attributed to victim and perpetrator; whereas, to our knowledge, there is no other study analysing social media reactions to GBV cases.

## 3. Dataset

### 3.1. Corpus Background

In a preliminary phase of our work, we conducted a research on the femicide case of Sara Di Pietrantonio[4], 22 years old, a white Italian student, from a wealthy family, murdered by her ex boyfriend on May 2016 [21]. In this preliminary research we set out to develop a corpus by collecting Twitter users' comments to femicide news on newspapers published online [5]. We created an annotation scheme for the data corpus consisting of two layers: the first focused on the dimensions of sentiment analysis and composed of three subtasks (subjectivity, polarity and irony), relevant for the detection of sentiment in social media [26]; the second focused on hate speech detection, including labels for misogyny, aggressiveness and its target. For more details on the annotation scheme and corpus description, please read below Appendix A.

Observing the results of the preliminary study, we discovered how the victim's characteristics influence the way newspapers present her femicide and users talk about it on social media. In fact, analyzing Di Pietrantonio's case, as she was a young, white, wealthy and Italian

---

[4]https://www.agi.it/cronaca/news/2019-09-11/sara_di_pietrantonio_processo_tappe-6170806/
[5]the dataset is available at https://github.com/madeddumarco/GBV-Maltesi

student, we found very few examples of misogyny and, in most cases, the aggressiveness was directed against the perpetrator. Furthermore, the scheme was not considered sufficiently suitable for bringing out important elements of femicide cases. In fact, the annotators expressed their difficulties caused by the scheme developed as it was deficient and too simplistic to recognise complex features of femicide events. In order to solve these issues, we decided to direct our efforts on another case study in which the victim exhibits intersectionality traits, which we assume may lead to more misogynistic content. In addition, we developed new schema and guidelines to have more accurate annotations specifically related to the femicide domain.

### 3.2. Data Collection

In this section we provide a description of the new dataset built and the methodology used.

As mentioned above, we focused our research on the femicide of Carol Maltesi[6], a 26 years old, white Italian woman, mother and online sex worker, who was brutally murdered in January 2022 by her ex partner, Davide Fontana, a 44 years old white Italian bank employee.

With the aim of collecting users' responses to femicide, we chose to collect comments using YouTube Data API, as it is freely available and allows us to easily access comments focused on specific news. The process of obtaining data followed several steps: first, we selected the 31 most popular YouTube videos based on number of views and comments. We chose videos about Maltesi femicide from different types of sources: national (mainly the Italian broadcaster RAI) and local news. The selection of videos is diachronic spanning from March 2022 to June 2023; this was done because the various media channels covered the story as it evolved starting from the discovery of the nameless body and ending with the sentence given to the perpetrator. Afterwards, we collected comments from all the videos selected. Due to the API policy, we were restricted to collect only first-level comments and at most 5 oldest responses to them. In total, we retrieved 3,821 comments.

### 3.3. Annotation Scheme

From the previous experience of the Di Pietrantonio corpus, we decided that a generic sentiment analysis schema proved to be too rigid to understand such a complex phenomenon. We created an annotation scheme and a new online platform to facilitate the raters work. We involved 5 annotators, 4 of them self-identified as women and 1 as a man, all interested in the topic and mostly coming from humanistic background. They were all students and voluntarily participated to the project. The annotation guidelines were decided with the annotators after a pilot study and a subsequent group discussion where the raters pointed out the main faults of the schema. Each annotator analyzed all the comments according to the following guidelines:

- *Non classifiable:* if the comment cannot be analysed because it is not written in Italian, because it consists only of emojis, because it is not comprehensible or not relevant to the topic (any comment that was marked as NC from at least 1 annotator was removed from the corpus);
- *Empathy:* whether, in the comment, there are expressions of empathy in support of the victim, her family or the event in general (i.e., condolences);
- *Misogyny:* whether, in the comment, there is a presence of discriminatory expression against women, including blaming, objectifying, discriminatory and sexist practices used towards them and their life choices. If misogyny is present, we asked annotators to indicate its target (group or individual) based on [16]. Moreover, we asked to specify if the expressed misogyny contained intersectionality traits and to select from a list what other dimensions were involved: age, religion, job, nationality, skin color, class, sexual orientation, gender, physical condition, educational background, language and culture;
- *Aggressiveness:* whether there is aggressiveness in the comment and to whom it is directed (allowing multiple choices): victim, perpetrator, social network (family, friends, colleagues), media, rape culture;
- *Responsibility:* if there is explicit attribution of responsibility for the murder in the text, state who is blamed (allowing multiple choices): victim, perpetrator, social network (family, friends, colleagues), media, rape culture;
- *Humor:* specify whether the text conveys humorous content through irony, sarcasm, word games or hyperbole;
- *Macabre:* specify whether there are macabre aspects detailing how the victim was killed;
- *Context:* indicate whether the context was helpful to better understand the meaning of the comments;
- *Notes:* free space for suggestions, observations or doubts.

### 3.4. Dataset Analysis

The dataset, GBV-Maltesi [7], is composed of 2,934 comments annotated on all categories by all annotators. We
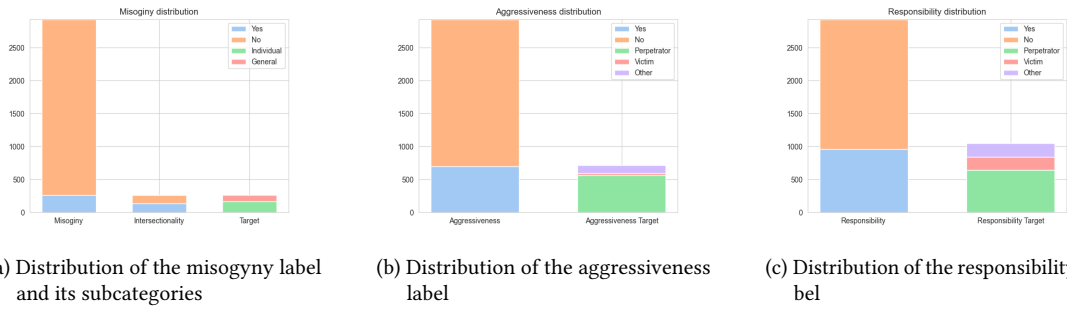
(a) Distribution of the misogyny label and its subcategories

(b) Distribution of the aggressiveness label

(c) Distribution of the responsibility label

**Figure 1:** Histograms for distributions of relevant labels

aggregated dimensions through majority voting. As our schema is composed by many different labels, we will focus only on the dimensions that we consider the most relevant, but all statistics can be found in Appendix C.

Starting from misogyny, in Appendix C and in Figure 1a, we can see that 9.03% of cases are positive. This unbalance is typical of hate speech datasets [27] and we consider it surprisingly high if we take into account the tragic theme of GBV. It is very interesting that intersectionality represents over 50% of misogynous examples indicating how the personal traits of the victim affect the perception of the users commenting. Unsurprisingly, as the victim was a sex worker, 'work' is almost always the category chosen by the annotators. The target of misogyny was mostly individual, confirming the findings of SemEval-2019 Task 5 [16]. The annotators explained to us how the misogyny target was a difficult category to annotate as often comments used the victim as an example to offend the broader group of women and sex workers.

Aggressiveness is more present than misogyny in our dataset, with 24% positive examples mostly directed towards the perpetrator. Responsibility follows a similar trend with 32.89% positive examples most directed towards the perpetrator. Unlike aggressiveness, we can see a significant amount of comments holding the victim responsible (6.55%).

In Appendix B, we reported the inter-annotator agreement (IAA) scores for all dimensions. As our dataset is fully annotated by multiple people, the metric we chose is Fleiss' Kappa [28]. The metric has a possible range of [-1,1], with 1 indicating perfect agreement, and any value of $\kappa \leq 0$ indicates more disagreement between the annotators than expected by chance. We can see that most dimensions have a $\kappa$ in the [0.2, 0.7] range, indicating variable levels of agreement depending on the label. The dimensions with the highest agreement at 0.69 are empathy towards the event and aggressiveness towards the perpetrator. In fact, annotators explained to us that these two categories were the easiest phenomena to annotate

as they lacked ambiguity. On the other hand, we can see that aggressiveness towards the victim is much lower (0.28). In our discussions with the raters, it emerged how attacks towards the victim were harder to identify as they were more subtle leading to disagreement among annotators.

## 4. Experiments

We conducted experiments to validate our resource and to gain more insight into the difficulty of the misogyny detection task. The goal of this analysis is to understand how the presence of different forms of misogyny (implicit and explicit) affect the evaluation of modern classification models. We consider as explicit misogyny discourses that intentionally spread hate towards women mostly through slurs and other aggressive behaviors. Meanwhile, we intend implicit misogyny as more subtle and less conscious practices like victim blaming, slut shaming, de-responsibilization of the perpetrator and more. In addition to our corpus, we used 3 other datasets regarding the topic in Italian: AMI [6], PejorativITy [29] and Inters8 [8]. The former two have been mainly gathered by keyword search of sexist terms[8], meanwhile, Inters8 and our corpus are focused on more implicit forms of sexist hate directed towards a specific woman (i.e., Silvia Romano and Carol Maltesi). Details about all the datasets can be found in Appendix D.

To explore the potential bias of models towards explicit forms of misogyny, we created 4 different models for binary misogyny detection: BERT-Maltesi, BERT-AMI, BERT-PejorativITy and, BERT-Inters8 that were respectively trained on the GBV-Maltesi, AMI, PejorativITy and

---

[8]AMI is created following an hybrid approach selecting also comments from known misogynistic accounts and responses directed to feminist public figures. We conducted a qualitative analysis and we found that the misogyny contained is almost always explicit and depending on slurs. This lead us to place it in the keyword category.

| Model | Maltesi Test | | Intes8 Test | | PejorativITy Test | | AMI Test | |
|---|---|---|---|---|---|---|---|---|
| | F1 Macro | F1 1-Label | F1 Macro | F1 1-Label | F1 Macro | F1 1-Label | F1 Macro | F1 1-Label |
| BERT-Maltesi | **0.611** | **0.351** | 0.512 | 0.174 | 0.571 | 0.436 | 0.633 | 0.611 |
| BERT-Inters8 | 0.377 | 0.169 | **0.621** | **0.49** | 0.55 | 0.538 | 0.659 | 0.725 |
| BERT-PejorativITy | <u>0.528</u> | <u>0.226</u> | 0.483 | 0.128 | **0.67** | **0.604** | <u>0.675</u> | <u>0.732</u> |
| BERT-AMI | 0.494 | 0.155 | <u>0.59</u> | <u>0.299</u> | <u>0.654</u> | <u>0.601</u> | **0.877** | **0.886** |
| Average | 0.502 | 0.225 | 0.551 | 0.273 | 0.611 | 0.545 | 0.711 | 0.738 |

**Table 1**
Results for binary misogyny detection on all datasets

Inters8 datasets. The models were just trained on the comments and were not given any other extra-information such as video transcriptions. The only label we analyzed was **misogyny** and all datasets were divided in training, validation and, test sets following a 60%, 20% and, 20% split. We used the existing splits when provided in the papers[9], else, we randomly created them. All models are binary classifiers created by fine-tuning BERT [30], in particular we used the Italian version AlBERTo [31]. Due to the imbalanced nature of most corpora, the models were trained with a focal loss [32] setting the hyperparameter $\gamma = 2$. Models were trained for 5 epochs but, to avoid overfitting, we implemented an early stopping function which ends training after 2 epochs that report an increase in validation loss. We tested all models on their own test set and the other 3 corpora.

We want to underline that our goal is not to compare performance of the different models between each other as they have different number of training sets and positive examples. Rather, we intend to focus on how different test sets are more difficult compared to others which helps us understand what the current challenges in misogyny detection are.

In Table 1, we reported the positive label and the macro average f1-scores of all experiments. In addition, we also calculated the average scores for each test set. The best scores achieved on a certain test set are in bold, meanwhile, we underlined the best scores for cross-dataset testing. As expected, we can observe that all models had the highest score for their own set. Meanwhile, for cross-dataset testing, we can see that the models that tend to perform the best are BERT-PejorativITy and BERT-AMI. We suspect that this is caused by the dataset composition as their training sets present more positive examples compared to the others.

Interestingly, we can observe that certain models recorded higher scores on other test sets that were not their own. This mostly happens when focusing on BERT-Maltesi and BERT-Inters8, which record higher scores on AMI and PejorativITy. Even PejorativITy increseases its scores when tested on AMI. Observing the average scores for each test, we can see that Maltesi and Inters8 are the

most challenging datasets. This is especially true when observing the average f1-score on the positive label with the score being in the [0.2, 0.3] range, compared to much higher scores for PejorativITy and especially AMI. These trends indicate how misogyny detection is a much harder task when considering datasets that contain less explicit forms of hate (e.g., not gathered by keyword search of sexist slurs).

In addition, we conducted a qualitative analysis on the errors of the various classifiers. We found that for each test set most classifiers misclassified the same type of examples. Models almost never recognized texts which contained victim blaming and slut shaming in the GBV-Maltesi Dataset. The errors made on Inters8 mostly coincide with examples that are also racist and Islamophobic. The cases which proved to be more difficult in PejorativITy and AMI contain less explicit animal epithets like "cavalla" and nouns that refers to sex worker in a less explicit way like "cortigiana".

## 5. Conclusion and Future Works

In this paper, we presented GBV-Maltesi which is the first dataset regarding social reactions to GBV, in particular to a femicide case. The topic was chosen to shed light on the importance of having misogyny corpora that include forms of sexism that are more implicit and complicated to detect compared to the existing ones that focus on slurs and offensive terms. We also focused on the intersectionality aspects to better explore online hate. GBV-Maltesi is composed of 2,934 comments all annotated by 5 annotators and it is available at https://github.com/madeddumarco/GBV-Maltesi. In order to overcome limitations of generic semantic schema, the corpus has been annotated following a new schema specifically created for cases of GBV. In the experimental phase of our work, we created different misogyny binary classifiers and tested them in a cross-dataset way. We found that datasets gathered on keyword collection are easier benchmarks as the model showed bias towards slurs and not identifying more implicit cases of misogyny. This research on online discourse about GBV is not meant to be exhaustive, as several questions are still open.

---

[9]PejorativITy provides a training and test split, but analyzing the code we found that the test set was used as a validation set so we decided to create a new one.

As future works, we intend to focus on how different framing of news can cause different online reactions, analyzing the differences between video transcripts of femicide news and the comments collected, in terms of words used, implicit references, attributions of guilt and descriptions of the people involved in the story. We also intend to gather more annotated corpora regarding femicides to explore how other characteristics of the victim (e.g., origin or skin color) and time of the murder differently influence the online reactions. In this regard, we intend to explore the question by investigating whether and how the discourse on misogyny changes depending on whether it is addressed to living or dead women (i.e., Giulia Cecchettin femicide and abusive discourse against her sister, Elena Cecchettin). Lastly, we would like to extend our research by following an intersectional approach, considering all the dimensions and characteristics that make up the identity of both victim and perpetrator. To conclude, we strongly advocate the importance of write the news correctly, as this has deep consequences on the readers' perception and the way they talk about it.

## Ethics Statement

The dataset was created in accordance with YouTube's Terms of Service. Considering the large number of users writing comments collected in the dataset, it was not possible to explicitly ask for their consent. No sensitive data are provided in the dataset and users' mentions have been anonymized to protect their privacy.

All the annotators involved in this research were free to participate without pressure or obligation. From the initial stages, they were aware of being free to leave at any time without negative consequences. During the annotation phase, we met several times to make sure that the topic did not disturb them psychologically or emotionally. We informed them to take their time, doing the annotation only when they felt like it and to contact us for support. This approach continued for all the research stages.

## Acknowledgements

## References

[1] M. L. Bonura, Che genere di violenza: conoscere e affrontare la violenza contro le donne, Edizioni Centro Studi Erickson, 2018.

[2] C. Vagnoli, Maledetta sfortuna, Rizzoli, 2021.

[3] J. Posetti, K. Bontcheva, D. Maynard, N. Aboulez, A. Lu, B. Gardiner, S. Torsner, J. Harrison, G. Daniels, F. Chawana, O. Douglas, A. Willis, F. Martin, L. Barcia, A. Jehangir, J. Price, G. Gober, J. Adams, N. Shabbir, The Chilling: A global study of online violence against women journalists, 2022.

[4] K. R. Blake, S. M. O'Dean, J. Lian, T. F. Denson, Misogynistic tweets correlate with violence against women, Psychological science 32 (2021) 315–325.

[5] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, H. Margetts, An expert annotated dataset for the detection of online misogyny, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 1336–1350. URL: https://aclanthology.org/2021.eacl-main.114. doi:10.18653/v1/2021.eacl-main.114.

[6] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami), in: EVALITA@CLiC-it, 2018. URL: https://api.semanticscholar.org/CorpusID:56483156.

[7] S. Gemelli, G. Minnema, Manosphrames: exploring an Italian incel community through the lens of NLP and frame semantics, in: P. Sommerauer, T. Caselli, M. Nissim, L. Remijnse, P. Vossen (Eds.), Proceedings of the First Workshop on Reference, Framing, and Perspective @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 28–39. URL: https://aclanthology.org/2024.rfp-1.4.

[8] I. Spada, M. Lai, V. Patti, Inters8: A corpus to study misogyny and intersectionality on twitter., in: CLiC-it, 2023.

[9] P. Lalli, L'amore non uccide. Femminicidio e discorso pubblico: cronaca, tribunali, politiche, Il Mulino, 2020.

[10] A. Tontodimamma, E. Nissi, A. Sarra, L. Fontanella, Thirty years of research into hate speech: topics of interest and their evolution, Scientometrics 126 (2021) 157–179.

[11] A. Ollagnier, E. Cabrio, S. Villata, Unsupervised fine-grained hate speech target community detection and characterisation on social media, Social Network Analysis and Mining 13 (2023) 58.

[12] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Lang. Resour. Evaluation 55 (2021) 477–523. URL:

https://doi.org/10.1007/s10579-020-09502-8. doi:10.1007/S10579-020-09502-8.

[13] K. W. Crenshaw, Mapping the margins: Intersectionality, identity politics, and violence against women of color, in: The public nature of private violence, Routledge, 2013, pp. 93–118.

[14] K. Manne, Down Girl: The Logic of Misogyny, Oxford University Press, 2018. URL: https://books.google.it/books?id=LqPoAQAACAAJ.

[15] E. W. Pamungkas, A. T. Cignarella, V. Basile, V. Patti, et al., Automatic identification of misogyny in english and italian tweets at evalita 2018 with a multilingual hate lexicon, in: CEUR Workshop Proceedings, volume 2263, CEUR-WS, 2018, pp. 1–6.

[16] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, S. M. Mohammad (Eds.), Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://aclanthology.org/S19-2007. doi:10.18653/v1/S19-2007.

[17] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Inf. Process. Manag. 57 (2020) 102360. URL: https://doi.org/10.1016/j.ipm.2020.102360. doi:10.1016/J.IPM.2020.102360.

[18] P. Zeinert, N. Inie, L. Derczynski, Annotating online misogyny, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 3181–3197.

[19] F. Formato, Gender, discourse and ideology in Italian, Springer, 2019.

[20] L. Busso, C. R. Combei, O. Tordini, Narrating gender violence a corpus-based study on the representation of gender-based violence in italian media, in: Language, Gender and Hate Speech: A Multidisciplinary Approach, 2020.

[21] N. Mandolini, Femminicidio, prima e dopo. un'analisi qualitativa della copertura giornalistica dei casi stefania noce (2011) e sara di pietrantonio (2016), Problemi dell'informazione 45 (2020) 247–277.

[22] E. Pinelli, C. Zanchi, Gender-Based Violence in Italian Local Newspapers: How Argument Structure Constructions Can Diminish a Perpetrator's Responsibility, 2021, pp. 117–143. doi:10.1007/978-3-030-70091-1_6.

[23] G. Minnema, S. Gemelli, C. Zanchi, V. Patti, T. Caselli, M. Nissim, Frame semantics for social NLP in italian: Analyzing responsibility framing in femicide news reports, in: E. Fersini, M. Passarotti, V. Patti (Eds.), Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022, volume 3033 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: https://ceur-ws.org/Vol-3033/paper32.pdf.

[24] G. Minnema, S. Gemelli, C. Zanchi, T. Caselli, M. Nissim, Dead or murdered? predicting responsibility perception in femicide news reports, in: Y. He, H. Ji, S. Li, Y. Liu, C.-H. Chang (Eds.), 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.

[25] G. Minnema, H. Lai, B. Muscato, M. Nissim, Responsibility perspective transfer for Italian femicide news, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), 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. doi:10.18653/v1/2023.findings-acl.501.

[26] V. Basile, N. Novielli, D. Croce, F. Barbieri, M. Nissim, V. Patti, Sentiment polarity classification at evalita: Lessons learned and open challenges, IEEE Transactions on Affective Computing 12 (2021) 466–478.

[27] B. Vidgen, L. Derczynski, Directions in abusive language training data, a systematic review: Garbage in, garbage out, Plos one 15 (2020) e0243300.

[28] J. Fleiss, Measuring nominal scale agreement among many raters, Psychological Bulletin 76 (1971) 378–. doi:10.1037/h0031619.

[29] A. Muti, F. Ruggeri, C. Toraman, L. Musetti, S. Algherini, S. Ronchi, G. Saretto, C. Zapparoli, A. Barrón-Cedeño, Pejorativity: Disambiguating pejorative epithets to improve misogyny detection in italian tweets, arXiv preprint arXiv:2404.02681 (2024).

[30] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[31] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets, in: Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), volume 2481, CEUR, 2019. URL:

| Dimension | Yes % | No % |
|---|---|---|
| Subjectivity | 70.48% | 29.52% |
| Misogyny | 3.76% | 96.24% |
| Polarity-Negative | 51.89% | 48.11% |
| Polarity-Positive | 4.93% | 95.07% |
| Aggressiveness | 24.02% | 75.98% |
| Irony | 7.09% | 92.91% |
| Context | 81.48% | 18.52% |

**Table 2**
Distribution of the dimensions for the DiPietrantonio Dataset

https://www.scopus.com/inward/record.uri?
eid=2-s2.0-85074851349&partnerID=40&md5=
7abed946e06f76b3825ae5e294ffac14.

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[33] F. Barbieri, V. Basile, D. Croce, M. Nissim, N. Novielli, V. Patti, et al., Overview of the evalita 2016 sentiment polarity classification task, in: CEUR Workshop Proceedings, volume 1749, CEUR-WS, 2016.

# A. Details about the Di Pietrantonio Dataset

The dataset GBV-DiPietrantonio is composed of 691 tweets fully annotated by 3 annotators, 2 of which self-identified as women and 1 as a man. The tweets were collected by gathering responses to news which covered the news of Di Pietrantonio femicide. The annotation scheme is composed of the slightly modified SENTIPOLC scheme[33] which consists of Subjectivity, Polarity (Positive, Negative) and Irony. In addition the semantic grid contained Misogyny, Aggressiveness and Target of Aggressiveness (towards Perpetrator, Victim, Other), Context, and Notes.

The statistics of the gold standard for the Di Pietrantonio dataset are in Table 2.

# B. Agreement of the Maltesi Dataset

Table 3 contains the agreement values calcolated with Fleiss' Kappa for all dimensions in the Maltesi dataset.

| Dimension | Fleiss' kappa |
|---|---|
| Misoginy | 0.56 |
| Target | 0.48 |
| Intersectionality | 0.32 |
| Aggressiveness | 0.53 |
| Agg. Perpetrator | 0.69 |
| Agg. Victim | 0.28 |
| Agg. Social Network | 0.23 |
| Agg. Media | 0.40 |
| Agg. Rape Culture | 0.10 |
| Responsibility | 0.21 |
| Resp. Perpatrator | 0.25 |
| Resp. Victim | 0.55 |
| Resp. Social Network | 0.13 |
| Resp. Media | 0.23 |
| Resp. Rape Culture | 0.19 |
| Empathy towards the event | 0.69 |
| Humor | 0.45 |
| Macabre | 0.49 |
| Context | -0.11 |

**Table 3**
Agreement of the Maltesi Dataset

| Dimension | Yes % | No % |
|---|---|---|
| Misoginy | 9.03% | 90.97% |
| Intersectionality | 4.63% | 95.36% |
| Aggressiveness | 24% | 76% |
| Agg. Perpetrator | 19.19% | 80.81% |
| Agg. Victim | 1.23% | 98.77% |
| Agg. Social Network | 0.88% | 99.11% |
| Agg. Media | 2.73% | 97.27% |
| Agg. Rape Culture | 0.41% | 99.59% |
| Responsibility | 32.89% | 67.11% |
| Resp. Perpetrator | 22.09% | 77.91% |
| Resp. Victim | 6.55% | 93.45% |
| Resp. Social Network | 2.11% | 97.89% |
| Resp. Media | 99.01% | 0.99% |
| Resp. Rape Culture | 4.06% | 95.94% |
| Empathy towards the event | 28.25% | 71.75% |
| Humor | 3.14% | 96.86% |
| Macabre | 3.27% | 96.72% |
| Context | 97.51% | 2.49% |

**Table 4**
Distribution of the binary dimensions of the Maltesi Dataset

# C. Distributions of the Maltesi Dataset

Table 4 contains the distribution of the binary labels in the Maltesi dataset. Table 5 contains the type of intersectionality and table 6 contains the type of misogyny target.

| Dimension | Percentage % |
|---|---|
| Work | 96.32% |
| Age | 0.73% |
| Work and Education | 0.73% |
| Work and Gender | 2.20% |

**Table 5**
Distribution of the values for the types of intersectionality selected

| Dimension | Percentage % |
|---|---|
| Individual | 63.40% |
| Grooup | 36.60% |

**Table 6**
Distribution of the values for the types of misogyny target selected

## D. Distributions of the Misogyny Dataset

Table 7 contains the details of the other existing misogyny datasets used in the experimental phase.

| Dataset | Topic | Num Examples | Num Pos. | Pos. % |
|---|---|---|---|---|
| Inters8 | Intersectional Hate focusing on Islamophobia in the case of hate towards Silvia Romano | 1,500 | 288 | 19.2% |
| AMI | Misogynistic slurs, attacks towards important figures who expressed support for women rights and posts from misogynistic account | 5,000 | 2,340 | 46.8% |
| Pejorativity | Words that can be used as misogynistic pejoratives in online discussion (e.g. Cavalla, cagna,...) | 1,200 | 397 | 33% |

**Table 7**
Distribution of the Italian misogyny Dataset