PejorativITy: Disambiguating Pejorative Epithets to Improve Misogyny Detection in Italian Tweets

Arianna Muti[1]✉, **Federico Ruggeri**[2], **Cagri Toraman**[3], **Lorenzo Musetti**[4], **Samuel Algherini**[4], **Silvia Ronchi**[4], **Gianmarco Saretto**[4], **Caterina Zapparoli**[4], **Alberto Barrón-Cedeño**[1]

[1]DIT, University of Bologna, Forlì, Italy
[2]DISI, University of Bologna, Bologna, Italy
[3]CEng, Middle East Technical University, Ankara, Turkey
[4]Expert AI, Modena, Italy

{arianna.muti2, federico.ruggeri6, a.barron}@unibo.it
ctoraman@ceng.metu.edu.tr
info@samuelalgherini.com
{lmusetti, sronchi, czapparoli, gsaretto}@expert.ai

Abstract

Misogyny is often expressed through figurative language. Some neutral words can assume a negative connotation when functioning as pejorative epithets. Disambiguating the meaning of such terms might help the detection of misogyny. In order to address such task, we present PejorativITy, a novel corpus of 1,200 manually annotated Italian tweets for pejorative language at the word level and misogyny at the sentence level. We evaluate the impact of injecting information about disambiguated words into a model targeting misogyny detection. In particular, we explore two different approaches for injection: concatenation of pejorative information and substitution of ambiguous words with univocal terms. Our experimental results, both on our corpus and on two popular benchmarks on Italian tweets, show that both approaches lead to a major classification improvement, indicating that word sense disambiguation is a promising preliminary step for misogyny detection. Furthermore, we investigate LLMs' understanding of pejorative epithets by means of contextual word embeddings analysis and prompting.

Keywords: Word sense disambiguation, Hate speech detection, Pejorative language

Disclaimer: This paper contains examples of offensive and explicit content.

1. Introduction

Pejorative language refers to a word or phrase that has negative connotations and is intended to disparage or belittle.[1] An inoffensive word becoming pejorative is a form of semantic drift known as pejoration; thus, pejorativity is context-dependent: pejorative words have one primary neutral meaning, and another negatively connotated meaning. The opposite is known as melioration, which is when a term begins as pejorative and eventually is adopted in a neutral sense, like in the case of slur reappropriation (Galinsky et al., 2013). Pejorative words are relevant in misogyny detection since many neutral words are used to address women in an offensive way, targeting either their physical aspect or their intelligence. We refer to such terms as **pejorative epithets**. Some examples in Italian are *balena* (whale/fat woman) and *gallina* (chicken/stupid). State-of-the-art models struggle to correctly classify misogyny when sentences contain such terms (Fersini et al., 2020). The occurrence of polysemic words with a pejorative connotation in the training set and a neutral connotation in the test set results in a great number of false positives (Muti and Barrón-Cedeño, 2020). For this reason, we introduce pejorative epithets disambiguation as a preliminary step to detect misogyny. Our goal is to assess whether the disambiguation of potentially pejorative epithets improves the detection of misogynistic language, while reducing the rate of false positives.

In this work, we aim to answer three research questions:

RQ1 Which epithets are used in misogynistic language in Italian?

RQ2 Can the disambiguation of such words decrease the error rate in misogyny detection?

RQ3 Can encoder-based language models and generative LLMs differentiate if a word in a tweet is pejorative or neutral based on its context?

To address **RQ1**, we compile a list of pejorative words used online to address women. We use such words to retrieve new tweets, and build PejorativITy, a novel corpus of Italian tweets, annotated at the

[1] https://www.merriam-webster.com/dictionary/pejorative

LREC-COLING 2024, pages 12700–12711

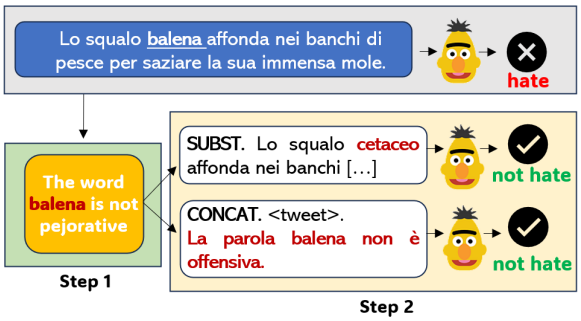

Figure 1: Our pipeline. Step 1: a model identifies the connotation of possibly pejorative epithets. Step 2: the identified connotation is used to enrich (CONCAT) and substitute (SUBST) part of the textual input for misogyny detection.

word level for pejorativity, and at the sentence level for misogyny.

To address **RQ2**, we fine-tune two BERT-based models: model$_{pej}$ to identify whether a word in the context of a tweet is pejorative or neutral, following Dinu et al. (2021), and model$_{mis}$ to detect misogyny. We use the output of model$_{pej}$ to inform model$_{mis}$ of whether the target word is pejorative within that context or not. Figure 1 represents our pipeline.

To address **RQ3**, we compare the cosine similarity between the contextualised word embeddings of a BERT-based model and their univocal corresponding words (**anchors**) before and after fine-tuning for pejorativity detection. Additionally, we prompt popular instruction-tuned LLMs to test their ability to disambiguate potentially pejorative words based on the context.

Our contribution is threefold: (1) we release a corpus manually annotated for pejorativity at the word level and misogyny at the sentence level; (2) we develop a transformer-based model for detecting pejorative words, whose predictions are used to enhance the performance of the model targeting misogyny detection; and (3) we analyse the performance of SOTA generative models on pejorative epithets disambiguation.

To the best of our knowledge, this is the first work that proposes word sense disambiguation to linguistically inform computational models for misogyny detection.

The PejorativITy dataset and the code for all the experiments are available at https://github.com/arimuti/PejorativITy.

2. Related Work

Misogyny and sexism detection have been explored in different platforms, such as Gab and Reddit (Kirk et al., 2023; Guest et al., 2021), Twitter (Jha and Mamidi, 2017; Anzovino et al., 2018), and blogs (Breitfeller et al., 2019) in English; and in different languages, such as Spanish (Anzovino et al., 2018; Plaza et al., 2023), Arabic (Almanea and Poesio, 2022), and Turkish (Toraman et al., 2022). In Italian, the reference datasets for the identification of misogyny are the two compiled in the framework of the two editions of the Automatic Misogyny Identification shared task (AMI) (Fersini et al., 2018, 2020).

Our work takes inspiration from Dinu et al. (2021), who (a) explore pejorative language on social media for the first time; (b) build a multilingual lexicon of pejorative terms for English, Spanish, Italian, and Romanian; (c) release a dataset of tweets annotated for pejorative use; and (d) present an attempt to automatically disambiguate pejorative words in their dataset. Our contribution differs since, for the first time, the information about the pejorativity of a word is leveraged to inform the model for misogyny detection. Moreover, our pejorative lexicon contains words that are currently used on Twitter to address women in a misogynistic manner. Whereas Dinu et al.'s lexicon considers hate speech in general, most gender-based words are outdated or missing, and it does not focus on the sort of slang typically used online.

Another similar work is Pamungkas et al. (2023), who develop the Swear Words Abusiveness Dataset (SWAD), where abusive swearing in English tweets is manually annotated at the word level to address the task of predicting the abusiveness of a swear word based on its context. While their work focuses on spotting slurs when used in a neutral way (i.e. meliorations), our aim is to disambiguate neutral words used in an offensive way (i.e. pejoration). Moreover, Pamungkas et al. exclude highly ambiguous words when creating their target word lexicon, whereas we precisely focus on them.

3. Corpus Compilation

To provide an overview of which misogynous epithets are commonly used on Twitter in Italian (**RQ1**), we compile a novel corpus. The compilation involves two steps: the creation of a lexicon of polysemic words that can function as pejorative epithets for women, and the retrieval of tweets containing such words.

Lexicon. We collect our lexicon by selecting words from three distinct sources. (1) We ask ten Italian native speakers to provide a list of offensive words used online to address women. The speakers use social media on a daily basis and their age ranges between 27 and 39 years. (2) We retrieve the keywords used in the two Italian corpora for the Automatic Misogyny Identification (AMI) shared task (Fersini et al., 2018, 2020). 3) We consult the 'List of Dirty Naughty Obscene Bad

Word	Literal	Pejorative	Neutral anchor	Pejorative anchor
acida	acid/sour	peevish	aspra	intrattabile, stronza
asina	female donkey	stupid	ciuco	stupida
balena	whale/flash	fat woman	cetaceo, balenare	grassa
bambola	doll	girl (objectifying)	giocattolo	donna attraente
cagna	female dog	bitch	cane femmina, canide	donna di facili costumi, troia
cavalla	female horse	ugly/whore	equino	brutta, alta e grossa
civetta	owl	tease	volatile rapace	donna che cerca attenzioni
cesso	toilet	ugly	water, bagno, toilette	brutta
contadina	farmer	ignorant, illiterate	agricoltore femmina	donna ignorante
cortigiana	court lady	prostitute	dama di corte	prostituta
cozza	mussel	ugly/clingy	mollusco	donna brutta, appiccicosa
femminista	feminist	feminazi	femminista	polemica, fastidiosa
fogna	sewer	skanky	fognatura	schifosa, bocca
gallina	chicken	stupid	pennuto	stupida
grezza	raw	rude woman	non lavorato	rozza
lesbica	lesbian	lesbian (offensive)	donna a cui piacciono le donne	schifosa
lurida	dirty	skanky	sporca	promiscua, troia
maiala	sow	whore	maiale femmina	promiscua, troia
mucca	cow	bitch	bovide	stupida, troia
oca	goose	stupid girl	pennuto	stupida, pettegola
pecora	sheep	doormat	ovino	stupida
strega	witch	hag, unpleasant	maga	crudele
vacca	cow	whore	bovino	donna di facili costumi, troia
zingara	gipsy	shabby	gitana	trasandata

Table 1: Italian pejorative lexicon, their literal and pejorative translations in English, and their anchors.

Words'.[2] We only keep polysemic words whose primary meaning is neutral and that are frequently used on Twitter with both pejorative and neutral connotations. To ensure the quality of our vocabulary, we qualitatively verify that such words are used with both connotations by manually searching them on Twitter.[3]

Table 1 shows our lexicon of 24 words. For each word, we report the English translation of its literal and pejorative meaning, and their anchors in Italian. Anchor words refer to the unambiguous words used to define polysemic words. We call these words anchors because their meaning is univocal and does not change according to the context. For instance, the word *balena (whale)* is used to refer to either a sea mammal or an overweight woman. In contrast, the anchor words *cetaceo (cetacean)* and *grassa (fat)* only refer to the animal in the first case and to being overweight in the second case, at least as far as their use in Twitter is concerned.[4]

[2] https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/tree/master, consulted on January 2023.

[3] Due to their exclusive neutral or negative connotation on Twitter, the following words are discarded: *barile, banco, botte, barbona, facile, gatta morta, passeggiatrice, porca, principessa, privilegiata, psicopatica, scrofa, somara, travestita*.

[4] In this case, the word *balena* has a third anchor word, from the verb *balenare*, which means 'to flash'.

Tweets. We use Twarc[5] to retrieve tweets from December 2022 to February 2023 containing words in our lexicon. We select 50 tweets for each word in our lexicon, resulting in 1,200 tweets. We keep a balance of pejorative and neutral use of lexicon words, although an equal distribution for each word could not be guaranteed.

4. Data Annotation

We recruit six annotators with a background in linguistics, gender studies, cognitive sciences, and NLP to label our corpus for pejorative word disambiguation and misogyny detection.

We first devise a pilot annotation study to explore the complexity of the task and observe differences in how male and female annotators perceive pejorative connotations. For this purpose, we follow a descriptive annotation paradigm (Röttger et al., 2022), which encourages annotator subjectivity by not providing guidelines. We split the annotators into two groups and assign 50 tweets each for labeling. Each group is composed of two women and one man with ages ranging between 27 and 39 years old.

We use Krippendorff's alpha (Krippendorff, 2011) to measure the inter-annotator agreement (IAA). The IAA of the first group is *moderate* for both pejorativity (0.48) and misogyny (0.50), whereas the IAA of the second group is *fair* for pejorativity (0.33) and *moderate* for misogyny (0.50). We observe that, in terms of gender differences, men tend to consider

[5] https://twarc-project.readthedocs.io

ID	Tweet	Translation
70019	*Non voglio una cagna un cane ce l'ho giaaaa*	*I don't want a female dog/bitch, I have a dog already.*
30021	*Wow sei una bambola!*	*Wow you're a beautiful girl/doll!*
10010	*Xchè avrà dato una risposta acida a lui*	*Because he/she will have given him a sharp answer*
61209	*Ma come fai a dire che sei una balena sei bellissima*	*How can you say you're a whale/fat, you're beautiful*

Table 2: Examples of tweets with potentially pejorative words (underlined).

sexual objectifying compliments as non-pejorative. Based on annotators' feedback, we identify five major areas of disagreement:

Lack of context. Some tweets are very short, lacking enough context to understand the intention of the author. We decide to label such tweets as neutral. Consider tweet *70019* in Table 2.
Although it is likely that the author uses humour to address a woman as a *cagna* (bitch), the context does not allow for a clear interpretation: it is possible that the author does not want another (female) dog, because he has already one.

Objectifying compliments. Some tweets are intended to compliment women, by means of objectification. Thus, we label them as pejorative. In the tweet *30021* in Table 2, the term *bambola* is used as a compliment, but it is objectifying and, therefore, should be considered pejorative.

Pejorative epithets towards objects. Some words are used pejoratively towards inanimate objects, therefore, they should be labeled as neutral. In the tweet *10010* in Table 2, the term *acida* refers to an inanimate thing (an answer), although the term is used pejoratively.

Pejorative epithets towards men. Words that are used pejoratively against men should be labeled as pejorative, so that the corpus can be used for the general task of pejorativity detection regardless of the auxiliary task.

Reported Speech. Some tweets contain pejorative epithets, although the intention is not harmful, because they are contained in reported speech. We label them as pejorative, since the annotation refers to the word, not to the whole sentence. Consider tweet *61209* in Table 2: the word *balena* is pejorative, but it is used in a positive way by means of negation.

We devise a second pilot annotation, getting closer to a prescriptive annotation paradigm (Röttger et al., 2022), by providing the above guidelines to the annotators. We select the top 50 tweets that caused more debate during the first study annotation phase. The IAA computed on all six annotators is 0.53 (*moderate*), denoting an improvement over the first pilot study.

Class	Training	Test	Total
Misogynous	369	28	397
Pejorative	363	28	391
Not pejorative	6	–	6
Non-misogynous	735	68	803
Pejorative	172	18	190
Not pejorative	563	50	613

Table 3: Statistics of the PejorativITy corpus. The same tweets are annotated for misogyny and pejorativity, for a total of 1,200 instances. For both the misogynous and the non-misogynous tweets, we report how many contain a pejorative word and how many do not.

PejorativITy. After the pilot studies, we annotate our collected corpus of 1,200 tweets. Only one person carries out the whole annotation process. We select the annotator with the most interdisciplinary background, who is an expert in gender studies, linguistics and NLP, who has been a target of misogyny. This setting is considered among the best practices for the annotation of phenomena like pejorative epithets and misogyny (Abercrombie et al., 2023).
Table 3 shows the statistics of our corpus. The Pearson correlation between misogyny and pejorativity labels is 0.70, which is in line with our expectations. The tweets for which misogyny and pejorativity are not aligned are mainly reported speech or men-related offensive language. It is worth noting that some sentences might not be considered misogynous, as they do not express hate towards women. However, they might be considered sexist. For instance, the sentence *"che bella bambola ciao tesoro"*[6] does not express hate but perpetuates the objectification of women by addressing the target of the tweet as a doll, falling into the category of benevolent sexism (Gothreau et al., 2022).

[6]translation: what a beautiful doll (girl), hi darling

AMI-2018	Misogynous	Not	Total
Train	1,828	2,172	4,000
Test	512	488	1,000
AMI-2020	**Misogynous**	**Not**	**Total**
Train	2,337	2,663	5,000
Test	500	500	1,000

Table 4: Statistics of the AMI 2018 and 2020 corpora (Fersini et al., 2018, 2020).

5. Experiments

To understand the impact of disambiguating pejorative words for misogyny detection (**RQ2**), we experiment with AlBERTo (Polignano et al., 2019), a popular BERT-based model trained on $200\,M$ Italian tweets. In particular, we fine-tune AlBERTo on two downstream tasks: pejorative word disambiguation and misogyny detection.

For pejorativite word disambiguation, we evaluate AlBERTo only on our corpus. For misogyny detection, we also consider the two other benchmark datasets for Italian: AMI-2018 (Fersini et al., 2018) and AMI-2020 (Fersini et al., 2020). To the best of our knowledge, these are the only corpora that address misogyny detection on Italian tweets. Table 4 shows their statistics.

We formulate the disambiguation of pejorative words as a binary classification task, where a model classifies a word contained in a sentence as pejorative or neutral. Then, we use the information about the pejorativity of a word to enrich the input to the model responsible for the detection of misogyny. Since AMI-2018 and AMI-2020 are not annotated for pejorative word disambiguation, we use the model fine-tuned on our corpus to determine the connotation of ambiguous words.

Formally, we devise the following pipeline, where $w \in W$ is a word from our lexicon W of pejorative words:

(a) We train model$_{pej}$ that, given a tweet containing a word $w \in W$, predicts whether w is being used in a pejorative way.

(b) We enrich input tweets in all data partitions by injecting knowledge about the pejorativity of our lexicon words according to model$_{pej}$. We try two different approaches to modify the input data: *i)* we **concatenate** the information about the pejorativity of w at the end of the tweet or *ii)* we **substitute** the ambiguous w with its corresponding anchor word.

(c) We train model$_{mis}$ to detect misogyny with the enriched input tweets.

Our pipeline is meant to process any tweet. However, as a first step, we check whether it contains at least one $w \in W$. In our setup, when testing on the subset of AMI-2018 and AMI-2020 containing only pejorative words (epithets), thhat are recognized through string matching after lemmatization.

As hyper-parameters, we use the AdamW optimizer with $\epsilon = 1^{-8}$ (Loshchilov and Hutter, 2017). We fine-tune AlBERTo for 4 epochs with batch size 16. We report macro and per-class F_1-measure as standard metrics for binary classification tasks, averaged over three individual runs. All the experiments are run using Google Colab's GPU.

Approach	Macro	Mis.	Not
baseline	0.68	0.56	0.79
concatenation			
w/ gold	0.83	0.78	0.88
w/ predictions	0.75	0.68	0.82
substitution			
w/ gold	**0.87**	**0.82**	**0.92**
w/ predictions	0.77	0.69	0.84

Table 5: Macro and per-class F_1-score on PejorativITy concerning misogyny detection.

6. Results

Regarding pejorative word disambiguation, the fine-tuned AlBERTo model (model$_{pej}$) reaches a macro F_1-measure of 0.82 ± 0.03 on the PejorativITy test partition.

Table 5 shows the classification performance for misogyny detection on the PejorativITy test partition. We compare our fine-tuned AlBERTo model (*baseline*) against the alternatives that leverage pejorative word disambiguation. We evaluate the concatenation and substitution approaches using model$_{pej}$ (*w/ predictions*) and annotators' labels (*w/ gold*) since our corpus contains annotations for pejorative word disambiguation. The evaluation of our proposed approaches with gold labels defines an upper bound to our pipeline. We observe a notable improvement over the baseline model for concatenation ($+7$ absolute points) and substitution ($+9$ absolute points) when using model$_{pej}$ predictions. The improvement significantly increases when both approaches consider gold labels, with a maximum gain of $+19$ absolute points. These results reflect the effectiveness of our approach and corroborate our initial hypothesis on reducing the false positive rate for misogyny detection.

Table 6 shows the number of false positives in the three datasets, before and after the inclusion of pejorative information both by concatenation and substitution. The decrease of false positives is clear in AMI-2020 and in our PejorativITy test set. In AMI-2018, no decrease is observed. One of the reasons for this low impact is that AMI-2018

Dataset	Baseline	concat.	subst.
PejorativITy	25	16	21
AMI 2018	107	107	112
AMI 2020	127	126	121

Table 6: False positive rates comparison. In the PejorativITy the total number of instance is 96, while in AMI 2018 and 2020 is 1,000.

contains pejorative epithets only in 34 instances out of 1000 (compared to 192 in AMI-2020), therefore we did not expect our approach to have a huge impact on that dataset.

Table 7 shows the classification performance for misogyny detection on AMI-2018 and AMI-2020. To assess the impact of our pipeline on these corpora, we show the performance of the models both on the test instances that contain words in our lexicon (**epithets**) and on the whole corpora. In particular, we perform fuzzy string matching (Section 3) to filter tweets according to this criterion, resulting in 389 (355 train, 34 test) tweets for AMI-2018 and 605 (413 train, 192 test) tweets for AMI-2020 in the training and test set respectively. We observe an F_1-measure improvement of $+3$ absolute points in AMI-2018 and $+4$ absolute points in AMI-2020 with the concatenation approach. In contrast, the substitution strategy does not lead to any performance gain. A possible explanation is the quality of substituted anchors. We provide an example in the next section. Since AMI corpora mainly contain tweets with explicit misogyny, the limited number of retrieved samples is expected. For this reason, the observed gain on selected tweets does not impact the overall performance on the original test partition in both corpora (**whole**).

To sum up, our results suggest that the disambiguation of potentially pejorative words is helpful in addressing misogyny detection when targeting ambiguous examples.

6.1. Qualitative Error Analysis

We carry out a manual error analysis, by observing misclassified tweets in AMI-2020 epithets and our corpus for the task of misogyny detection. We compare misclassified tweets in the three settings: baseline, concatenation, and substitution.

Regarding the concatenation approach, most of the misclassifications occur when reported misogyny is concerned. The model struggles to recognise when a pejorative epithet is used in a reported speech to condemn a misogynistic attitude and not to address a potential target. It is worth noticing that if a pejorative connotation is predicted in reported speech, this does not imply that misogyny is predicted. Consider the following example:

Lei è acida perché non ha figli penso che darebbe fastidio a qualsiasi donna. Che schifo.[7]

AMI-2018	epithets			whole		
Approach	**Macro**	**Mis.**	**Not**	**Macro**	**Mis.**	**Not**
baseline	0.79	0.77	0.81	0.86	0.87	0.85
concatenation	**0.82**	0.81	0.83	**0.86**	0.88	0.85
substitution	0.79	0.79	0.80	0.86	0.87	0.84
AMI-2020	**epithets**			**whole**		
Approach	**Macro**	**Mis.**	**Not**	**Macro**	**Mis.**	**Not**
baseline	0.77	0.74	0.81	0.82	0.84	0.81
concatenation	**0.81**	0.77	0.84	**0.83**	0.84	0.82
substitution	0.77	0.73	0.81	0.82	0.84	0.81

Table 7: Macro and per-class F_1-measure on AMI-2018 and AMI-2020 concerning misogyny detection. We report metrics for each corpus (**whole**) and their subset containing words in our lexicon (**epithets**).

In this example, the author of the tweet criticises a reported misogynous sentence. Even if *acida* is correctly predicted as pejorative, the model still gets the correct prediction that the sentence is non-misogynous. Another observed pattern of misclassification is when the target of the pejorative epithet is a man. In this case, the tweet should not be considered misogynous, although it contains a pejorative word from our lexicon. This bias is introduced due to the annotation of pejorative epithets against men as pejorative. Overall, the overlap between tweets classified as containing pejorative words and those classified as misogynous is of 26 tweets in the PejorativITy test set (out of 96), 12 tweets out of 34 in the AMI2018_epithets, and 67 out of 192 in AMI2020_epithets. We highlight this aspect to show that model$_mis$ does not necessarily learn to classify misogyny according to model$_pej$'s outcome.

Regarding the substitution approach, we observe that a wrong pejorative prediction of lexicon words affects the prediction of misogyny. The following example:

Ma la balena con gli shorts cortissimi invece è vittima del patriarkato e può vestirsi come vuole?[8]

is correctly classified by the baseline model. A misclassification of the word *balena*,which model$_{pej}$ predicts as neutral, causes confusion in both enriched models.

7. Analysis of Contextualised Word Embeddings

To investigate the semantic knowledge of the AlBERTo pretrained language model (Polignano et al.,

[7] She's peevish because she doesn't have children I think it would bother all women. Disgusting.

[8] That whale/fat girl with very short pants is a victim of the patriarchy and can dress up as she wants?

Lexicon	Anchor	pretrained		Fine-tuned	
		Pejorative	Neutral	Pejorative	Neutral
acida	aspra	0.27 ± 0.12	0.27 ± 0.14	0.09 ± 0.12	0.29 ± 0.10
	intrattabile	0.28 ± 0.12	0.28 ± 0.14	0.28 ± 0.05	0.27 ± 0.07
	stronza	0.31 ± 0.14	0.31 ± 0.17	0.53 ± 0.12	0.23 ± 0.15
balena	balenare	0.26 ± 0.12	0.30 ± 0.10	0.19 ± 0.10	0.44 ± 0.08
	cetaceo	0.22 ± 0.12	0.26 ± 0.09	0.04 ± 0.10	0.36 ± 0.10
	grassa	0.19 ± 0.12	0.22 ± 0.09	0.29 ± 0.09	0.07 ± 0.07
cagna	canide	0.43 ± 0.15	0.29 ± 0.15	0.08 ± 0.05	0.25 ± 0.06
	donna di facili costumi	0.42 ± 0.13	0.27 ± 0.15	0.30 ± 0.04	0.21 ± 0.09
	troia	0.41 ± 0.16	0.26 ± 0.16	0.57 ± 0.08	0.21 ± 0.10
cesso	water	0.37 ± 0.14	0.37 ± 0.13	0.08 ± 0.06	0.26 ± 0.08
	bagno	0.39 ± 0.14	0.41 ± 0.13	0.07 ± 0.06	0.35 ± 0.10
	toilette	0.37 ± 0.13	0.39 ± 0.12	0.09 ± 0.05	0.30 ± 0.08
	brutta	0.39 ± 0.15	0.40 ± 0.13	0.43 ± 0.07	0.16 ± 0.09
lesbica	donna a cui piacciono le donne	0.40 ± 0.13	0.42 ± 0.16	0.28 ± 0.05	0.34 ± 0.09
	schifosa	0.32 ± 0.15	0.32 ± 0.17	0.30 ± 0.09	0.18 ± 0.06
vacca	bovino	0.31 ± 0.14	0.25 ± 0.12	0.10 ± 0.07	0.22 ± 0.07
	donna di facili costumi	0.35 ± 0.12	0.29 ± 0.12	0.27 ± 0.05	0.20 ± 0.08
	troia	0.35 ± 0.14	0.29 ± 0.13	0.50 ± 0.09	0.25 ± 0.14

Table 8: Average cosine similarity between lexicon word embeddings and both pejorative and neutral anchor word embeddings in pejorative and neutral samples. Embeddings extracted from both the pretrained and the fine-tuned AlBERTo model.

2019) about the pejorative epithets and to evaluate how fine-tuning affects its knowledge (**RQ3**), we extract and analyse the contextualised word embeddings of our lexicon words.

To extract these embeddings, we perform fuzzy string matching on input tweets to retrieve the tokenized text span corresponding to lexicon words. We use fuzzy string matching to address all representations of a lexicon word (e.g., *balena* and *balenare*). It is worth noticing that the retrieved text span may contain multiple tokens according to the employed tokenization process. In our scenario, the AlBERTo model employs the sentencepiece tokenizer (Kudo and Richardson, 2018), the common tokenization process for transformer models. For instance, the lexicon word *balena* is tokenized to the [`balen`, `##a`] text span. We then use these text spans to aggregate the corresponding word embeddings. We define the word embedding of a lexicon word as the average of the AlBERTo token embeddings in the retrieved text span. Considering *balena*, we define its word embedding by extracting the embeddings of `balen` and `##a` and computing their average.

We compute the average cosine similarity between lexicon words and their corresponding neutral and pejorative anchors. To carry out our analysis, we consider lexicon words from PejorativITy with several neutral and offensive anchors: *acida*, *balena*, *cagna*, *cesso*, *lesbica*, and *vacca*.

Table 8 reports the results on PejorativITy comparing the pretrained AlBERTo model and its fine-tuned version. The pretrained model does not discriminate between neutral and offensive anchors in pejorative and neutral samples. For instance, the average cosine similarity between *acida* and its pejorative anchor stronza is 0.31 in both class samples. In contrast, our fine-tuned AlBERTo model shows relevant discrepancies when considering lexicon word embeddings in pejorative and neutral samples. For instance, the similarity between *acida* and its neutral anchor aspra is 0.09 in pejorative samples and 0.29 in neutral ones. In contrast, the similarity between *acida* and its pejorative anchor stronza is significantly higher in pejorative samples (0.53), compared to neutral ones (0.23). This is an indicator that our fine-tuned model acquired knowledge about *acida* being a synonym of stronza when used in a pejorative way. We observe similar trends in all other selected lexicon words.

The average similarity with pejorative anchors in pejorative and neutral samples using pretrained embeddings is 0.34 and 0.29, respectively. In contrast, the similarities are 0.39 and 0.20 using fine-tuned embeddings. These results show pejorative anchors have higher similarity with lexicon words in pejorative samples. Likewise, neutral anchors with lexicon words in neutral samples: the average

similarity in pejorative and neutral samples using pretrained embedding is 0.33 and 0.34, respectively, while the similarities are 0.11 and 0.31 using fine-tuned embeddings.
To prevent a biased analysis where our fine-tuned model learns a representation of anchors similar to that of lexicon words, we compute each anchor frequency in our corpus. The computed frequency for each anchor is close to zero, thus excluding potential biases in our analysis.

8. Analysis of Large Language Models

We prompt open-source generative large language models (LLMs) to investigate their knowledge of pejorative words. Specifically, we aim to find out whether they are able to clarify the meaning of our lexicon words (**RQ3**).
We follow a zero-shot prompting approach without any training data for the downstream task (Liu et al., 2023). We design the following prompt that asks the meaning of a pejorative *{word}* in a given *{sentence}*. We prompt the LLMs for each test instance in our corpus. The prompt is given as follows:

```
[INST] Di seguito è riportata
un'istruzione che descrive un task.
Scrivete una risposta che completi
adeguatamente la richiesta.
### Istruzione:
Qual è il significato della parola
"{word}" in questa frase?
"{sentence}"[/INST]
### Risposta:
```

The translation in English would be:

```
[INST] Below there is an instruction
describing a task. Write a response
that completes the request appropri-
ately.
### Instruction:
What is the meaning of the word
"{word}" in this sentence?
"{sentence}"[/INST]
### Response:
```

We use three open-source LLMs for our analysis:

LlaMa: LlaMa is a decoder-based language model pretrained on publicly available data collections (Touvron et al., 2023a). Since LlaMa does not support Italian, we employ Camoscio[9], an Italian instruction-tuned LlaMa model.

LlaMa2: LlaMa2[10] is an optimized version of LlaMa by increasing context length from 2048 to 4096, and applying group-query attention (Touvron et al., 2023b). The majority of the training data of LlaMa2 is in English, but it still responds to Italian prompts due to a small amount of training data in Italian.

Mistral: MistralAI[11] is based on LlaMa2 but exhibits superior performance due to the employed attention mechanisms such as group-query attention and sliding-window attention (Jiang et al., 2023). The details of the corpora used in training are not given, yet it responds to Italian prompts.
For all models, we select the *7b* model version with 8-bit weights due to hardware constraints. We apply Beam Search for text generation with the following hyperparameters. The temperature is set to 0.2, the number of beams is set to 4 with a top-p value of 0.75, the output length is set to 300 tokens with a repetition penalty of 1.8. The analysis is conducted on 4 GPUs (NVIDIA GeForce RTX 2080 Ti).

8.1. Qualitative Analysis

We manually inspect all the responses generated by the three LLMs.

Mistral. We observe that Mistral is the best in disambiguating the connotations of lexicon words, especially when it comes to neutral senses. For instance, it correctly disambiguates when *balena* refers to the animal and when to the verb *balenare*. However, Mistral struggles when the term *balena* is used pejoratively. Consider Example 1 in Table 9. Mistral gets the idea that *balena* is used as a metaphor for something big, but it does not link its meaning to being overweight.
Mistral is remarkably good at capturing irony as well. Consider Example 2 in Table 9. While this example caused trouble to human annotators for the lack of context, Mistral is confident in identifying the pejorative connotation of the lexicon word *cagna*.
Although performing very well, Mistral struggles with reported speech, too. Consider Example 3 in Table 9. While Mistral correctly identifies the pejorative connotation, it fails to understand that the author of the tweet is condemning, not enforcing, a misogynistic statement.
Moreover, in some cases, Mistral makes up meanings. For instance, Mistral defines *cavalla* (horse / ugly and tall woman) as a *"a painful surprise"*, while it defines *cozza* (mussel / ugly, clingy) as *"impatiently waiting"*.[12] A possible explanation is that Mistral uses the semantics of the whole sentence to generate a definition of lexicon words. In some other cases, Mistral generates the opposite meaning. In Example 4 of Table 9, Mistral defines *cozza* as *"surprisingly beautiful"*.

[9] https://github.com/teelinsan/camoscio
[10] https://huggingface.co/meta-llama/Llama-2-7b-hf
[11] https://huggingface.co/mistralai
[12] In the sentence "Sta cozza non vedeva l'ora", translated as "That ugly girl couldn't wait"

	Tweet	Translation	Generated Disambiguation
1.	*Tra il non normalizzare e il dargli della balena ci sta l'oceano pacifico.*	*Between the non-normalisation and calling her whale there is the Pacific Ocean.*	*The meaning of the word balena in this sentence is to indicate something large, imposing, and difficult to overcome or face, similar to the Pacific Ocean.*
2.	*Non voglio una cagna un cane ce l'ho giaaaa*	I don't want a female dog/bitch, I have a dog already.	The meaning of the word cagna in this sentence is "ragazza" (girl). The word cagna is a slang term used to refer to a girl. In this context, the phrase means "I don't want a girl, I already have a dog."
3.	*Chi sei tu per dare cagna e zoccola alle persone?*	*Who are you to address people as bitch and whore?*	*In this case, the author of the sentence is calling the other user a bitch probably to express anger or contempt toward her.*
4.	*Madonna mia che cozza questa!!*	*What an ugly girl!*	The meaning of the word cozza in this phrase is to express surprise or amazement. Specifically, the phrase can be literally translated as "My God, how beautiful this is!".

Table 9: Examples of Mistral pejorative word disambiguation. Lexicon words are underlined.

Llama and Camoscio. Neither model shows an adequate performance in disambiguating lexicon words. In most cases, both models produce the following answer: *"the word {word} means {word}"*, which is not useful for disambiguation.

Our analysis suggests that off-the-shelf instruction-tuned LLMs have ample room for improvement concerning pejorative word disambiguation. A fine-tuning phase on the task could address the highlighted issues. However, we believe that a detailed analysis of instruction-tuned LLMs on our proposed pipeline deserves a separate study. We leave this analysis as future work.

9. Conclusions

We introduce pejorative word disambiguation as a preliminary step for misogyny detection to reduce the error rate of classification models on polysemic words that can serve as pejorative epithets. For this purpose, we build a lexicon of polysemic words with both pejorative and neutral connotations and use it to compile a novel corpus of 1,200 manually expert-annotated Italian tweets for pejorative word disambiguation and misogyny detection. We validate our pipeline by evaluating AlBERTo (Polignano et al., 2019) on our corpus and on two benchmark corpora in Italian: AMI-2018 (Fersini et al., 2018) and AMI-2020 (Fersini et al., 2020). We explore two approaches to inject pejorativity information: concatenation and substitution. Our results show that the disambiguation of potentially pejorative words leads to notable classification improvements in all testing scenarios. Furthermore, we analyse the word embedding representation of AlBERTo and show that the encoding of lexicon words is closer to their ground-truth connotation after fine-tuning. Lastly, we qualitatively analyse several off-the-shelf instruction-tuned LLMs on pejorative word disambiguation to evaluate their capabilities, showing that there is ample room for improvement.
Future research directions include the extension of our pipeline to automatically extract potentially pejorative words at the span level; the application of knowledge bases like ConceptNet (Speer et al., 2017) for pejorative word disambiguation; and the implementation of instruction-tuned LLMs in our pipeline. Moreover, we plan to expand this work towards other languages, and the cultures behind them. Our aim is to carry out a cross-cultural analysis on the differences in terms of pejorative terms for misogyny across cultures with different perspectives towards women rights and feminism.

10. Ethical Considerations

We use publicly available tweets to collect our corpus. All data collection adheres to Twitter's terms of service and privacy policies. As this research involves the analysis of publicly available tweets, we do not seek explicit consent from individual users. Nevertheless, we make every effort to protect the anonymity of all individuals mentioned or quoted in this work: Any reported example is carefully selected to avoid identifying specific users or victims.

11. Limitations

Language. In our study, we only focus on the Italian language. While this choice does not limit the applicability of our contributions, we are aware that including other languages could strengthen the impact of our results. We leave this extension as future work.

Corpus Although our lexicon covers a wide variety of words that can serve as pejorative epithets for women, it is not an exhaustive list, as we have discarded all the terms that are not polysemic and that are used only with one connotation (either positively or negatively) on Twitter.

Only 100 tweets are annotated by six annotators, while the remaining 1,100 are labelled by only one annotator. Although we select an expert with an interdisciplinary background in linguistics, gender studies and NLP to carry out all the annotations, their personal biases, opinions, or interpretations can lead to skewed or one-sided data.

Approaches. A limitation of our study concerns the substitution approach. First of all, some words have more than one neutral anchor words. This is the case of *balena*, which has two neutral anchors: *balenare* (to flash) and *cetaceo* (sea mammal). In neutral examples, we substitute *balena* with both anchors. This process may alter the semantic meaning of the tweet since only one anchor is suitable for substitution. Moreover, in some cases, we replace a lexicon word with anchors that do not have the same meaning. For instance, the neutral anchor of *acida* is *aspra* (*sour*). However, expressions like *sour beer* or *sour cream* do not have a valid anchor replacement. Therefore, replacing *aspra* with *acida* is not an appropriate substitution.

Models. We only employ AlBERTo to carry out our experiments. However, several other models might lead to different results. Therefore, our experiments are not sufficient to generalise the results of our analysis to all encoder-based models.

Prompting. The most popular generative models —GPT family— have not been included in this study, although they could have shown promising capacities in disambiguating the senses of our polysemic words. Nevertheless, we intentionally exclude them from this study, as we want to focus on open-source models only.

Acknowledgements

A. Muti's research is carried out under project "DL4AMI–Deep Learning models for Automatic Misogyny Identification", in the framework of Progetti di formazione per la ricerca: Big Data per una regione europea più ecologica, 'digitale e resiliente—Alma Mater Studiorum–Universita di 'Bologna, Ref. 2021-15854. The work of F. Ruggeri is supported by the European Union's Horizon Europe research and innovation programme under GA 101070000. We thank to Umitcan Sahin for his support during corpus compilation.

12. Bibliographical References

Gavin Abercrombie, Aiqi Jiang, Poppy Gerrard-abbott, Ioannis Konstas, and Verena Rieser. 2023. Resources for automated identification of online gender-based violence: A systematic review. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 170–186, Toronto, Canada. Association for Computational Linguistics.

Dina Almanea and Massimo Poesio. 2022. ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291, Marseille, France. European Language Resources Association.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Liviu P. Dinu, Ioan-Bogdan Iordache, Ana Sabina Uban, and Marcos Zampieri. 2021. A computational exploration of pejorative language in social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3493–3498, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adam D. Galinsky, Cynthia S. Wang, Jennifer A. Whitson, Eric M. Anicich, Kurt Hugenberg, and Galen V. Bodenhausen. 2013. The reappropriation of stigmatizing labels: The reciprocal relationship between power and self-labeling. *Psychological Science*, 24(10):2020–2029.

Claire Gothreau, Kevin Arceneaux, and Amanda Friesen. 2022. Hostile, Benevolent, Implicit: How Different Shades of Sexism Impact Gendered Policy Attitudes. *Frontiers in Political Science*, 4.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts.

2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Arianna Muti and Alberto Barrón-Cedeño. 2020. UniBO @ AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo. In *EVALITA Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop 12-13 December 2018, Naples*.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2023. Investigating the role of swear words in abusive language detection tasks. *Language Resources and Evaluation*, 57(1):155–188.

Laura Plaza, Jorge Carrillo-de Albornoz, Roser Morante, Julio Gonzalo, Enrique Amigó, Damiano Spina, and Paolo Rosso. 2023. Overview of exist 2023: sexism identification in social networks. In *Proceedings of ECIR'23*, pages 593–599.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. 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, Bari, Italy. CEUR.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B. Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 175–190. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Cagri Toraman, Furkan Şahinuç, and Eyup Halit Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

13. Language Resource References

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop 12-13 December 2018, Naples*, pages 59–66. Torino: Accademia University Press.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. Ami @ evalita2020: Automatic misogyny identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.