PejorativITy - In-Context Pejorative Language Disambiguation: A CALAMITA Challenge

Arianna Muti

University of Bologna - DIT

Abstract

Misogyny is often expressed through figurative language. Some neutral words can assume a negative connotation when functioning as pejorative epithets, and they can be used to express misogyny. Disambiguating the meaning of such terms might help the detection of misogyny. This challenge addresses a) the disambiguation of specific ambiguous words in a given context; b) the detection of misogyny in instances that contain such polysemic words. In particular, framed as a binary classification, our task is divided into two parts. In Task A, the model is asked to define if, given a tweet, the target word is used in pejorative or non-pejorative way. In Task B, the model is asked whether the whole tweet is misogynous or not.

Keywords

offensive language, pejorativity, misogyny

Warning: This paper contains offensive words.

1. Introduction and Motivation

This CALAMITA challenge [1] addresses the task of disambiguating pejorative language to detect forms of misogyny that are masked within ambiguous and context-dependent expressions. 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. In this challenge, our objective is to evaluate large language models (LLM) in Italian by focusing on the disambiguation of *pejorative epithets* used online to express misogyny. In this work, misogyny is defined as a property of social environments where women perceived as violating patriarchal norms are "kept down" through hostile or benevolent reactions coming from men, other women, and social structures [2, 3], in the form of sexual objectification, male privilege, gender discrimination, sexual harassment, belittling and violence [4].

An example of a pejorative epithet is *balena (whale)*, whose standard meaning refers to the sea mammal, but it is used offensively to address an overweight woman. Encoder-based models struggle to correctly classify misogyny when sentences contain such terms: 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 [5]. This could be overcome by decoder-based LLMs, as they could rely on their implicit knowledge to grasp the meaning of such terms. By asking models to determine whether a term is being used in a pejorative or non-pejorative sense, we challenge the LLMs' ability to comprehend semantic shifts in Italian. Moreover, asking whether a sentence containing that term is misogynous or not, enables us to comprehend to what extent LLMs understand misogyny, even when it is conveyed through figurative language. We expect models to struggle with this challenge, particularly in sentences with non-standard or regional varieties of Italian, which occur in our corpus.

2. Challenge: Description

We introduce pejorative language 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. Therefore, this challenge aims to address two tasks:

Task A Disambiguation of in-context polysemic words that can be used as pejorative epithets in misogynistic language;

Task B Misogyny detection at the sentence level.

Both tasks are conceived as binary classification tasks. Fig. 1 shows the pipeline for our tasks. Assume the sentence *Quella balena coi jeans non si può guardare*, translated as *Can't look at that whale with jeans*.

Task A: First, the model is asked to identify whether the meaning of the target word (*balena* in our example) is pejorative or not. The model should rely on its internal

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

arianna.muti2@unibo.it (A. Muti)

0000-0002-3387-6557 (A. Muti)

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

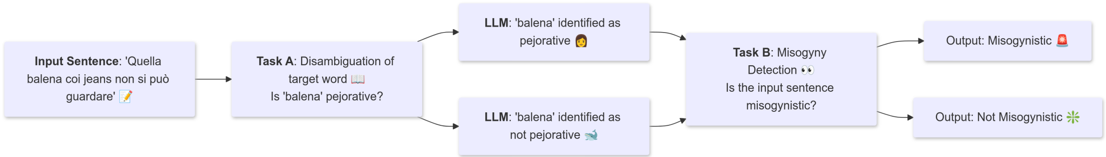
Input Sentence: 'Quella balena coi jeans non si può guardare'
Task A: Disambiguation of target word
Is 'balena' pejorative?
LLM: 'balena' identified as pejorative
LLM: 'balena' identified as not pejorative
Task B: Misogyny Detection
Is the input sentence misogynistic?
Output: Misogynistic
Output: Not Misogynistic

Figure 1: Visualization of our tasks.

knowledge accumulated during pre-training to understand whether the term *balena (whale)* refers to *woman* or *cetaceus*. Ideally, the model should exploit the context to perform the disambiguation, as the image of a whale with jeans is not plausible. That is why we encourage commonsense reasoning for this task.

Task B: In the second step, the model is first informed with the decision of Task A, whether the target word is pejorative or not, and then asked to classify the input sentence as misogynous or not.

3. Data description

The compilation of our corpus 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 [6, 7].

(3) We consult the 'List of Dirty Naughty Obscene Bad 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]

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 maintain a balance of pejorative and neutral use of lexicon words, although an equal distribution for each word could not be guaranteed. We choose tweets as source of data for three reasons. First, Twitter is a prominent platform for expressing opinions, where language is varied, conversational, and often informal, which makes it suitable to analyze misogyny conveyed through figurative language. Second, at the time of data collection, Twitter API was public and free, which facilitated our data collection process. Third, the character limit on tweets encourages condensed language, limiting the context of expression. Choosing tweets allows us to challenge LLMs in disambiguating pejorative language for misogyny detection within the constraints of limited or lack of context.

3.1. Annotation Details

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

We first devise a pilot annotation study to explore the complexity of the task. For this purpose, we follow a descriptive annotation paradigm [8], 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

[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'.

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

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/tall/ungainly	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	dyke	donna a cui piacciono le donne	lesbica (offensivo)
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.

old. We use Krippendorff's alpha [9] 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 sexual objectifying compliments as non-pejorative. More details about the annotation process, including the discussion of edge cases, can be found in Muti et al. [10]. 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 misogyny [11].

3.2. Data format

Data are collected in an Excel file and published at https://github.com/arimuti/PejorativITy. Each row contains the ID of the tweet, the tweet, the target word, the annotation for pejorativity at word level and the annotation for misogyny at sentence level. Table 2 shows examples.

3.3. Detailed data statistics

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 are annotated as misogynous, although they do not express any form of hate towards women. However, they contain subtle sexist language, which we consider misogynous according to the definition provided in Section 1. 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 [12].

3.4. Prompt Design

We design two prompts to address the two task: pejorativity disambiguation at word-level and misogyny detection at sentence-level. We adopt a zero-shot approach, although participants are encouraged to experiment with different prompting techniques.

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

ID	Tweet	Pejorative	Misogyny
70019	*Non voglio una cagna un cane ce l'ho giaaaa*	1	1
10010	*Xchè avrà dato una risposta acida a lui*	0	0
61209	*Ma come fai a dire che sei una balena sei bellissima*	1	0

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

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.

Task A. We design the following zero-shot prompt that asks whether the target *{word}* in a given *{sentence}* is pejorative or not. We prompt the LLMs for each instance in our corpus. The prompt is given as follows:

```
   La parola {word} è utilizzata in modo peggiorativo
nel testo in input? Rispondi solo con 1 per sì, 0 per
no. Testo: {tweet}
```

Task B. We design two zero-shot prompts that ask whether a tweet is misogynous or not in order to evaluate two settings:

A We treat Task A and Task B as independent, i.e., without having the decision for Task A to inform the decision for Task B.

B We enrich the prompt for Task B with the output from the previous task;

Prompt A:

```
   Il testo in input è misogino? Rispondi solo con 1
per sì, 0 per no. Testo: {tweet}
```

Prompt B:

```
   Il testo in input è misogino? Rispondi solo con
1 per sì, 0 per no. Considera l'informazione sulla
parola {word}: {response from Task A} Testo: {tweet}
```

4. Evaluation of Baselines

In order to evaluate our tasks, we employ Macro F_1score. We select two models as baselines: the fine-tuned AlBERTo model [13] and llama-3-8b (both the standard and the Instruct version). The results are not comparable though, since llama is evaluated on the whole corpus, while AlBERTo on the partition of the test set (see Table 3).

Model	Macro F1
AlBERTo	0.82
llama-3-8b	0.00
llama-3-8b-Instruct	0.64

Table 4
Results on Task A. Scores are not comparable since the testing partitions are different.

Task A. Table 4 shows the results for pejorative word disambiguation. The fine-tuned AlBERTo model reaches a macro F_1-measure of 0.82 ± 0.03, as reported in [10]. When it comes to decoder-based models, llama3-8b-Instruct shows a lower score, with a difference of 0.18 points, showing room for improvement in the prompt design. However, those scores are not comparable as the testing partitions differ. Llama-3-8b fails to complete the task, since it only repeats the prompt without providing an answer. For this reason, we discard llama-3-8b in the next task. It should be noted that llama has undergone a safety tuning process, preventing the model from always providing an answer, responding *I cannot provide a response that condones hate speech.* We excluded such cases from the evaluation. Of the 174 excluded instances, 123 were pejorative and 51 were not pejorative according to the gold standard. Although the fine-tuned version of AlBERTo achieves a higher performance (in a smaller subset of instances), llama aids in explainability by deliberately adding explanations of why it considers the target word to be pejorative or not. We will explore the plausibility of such explanations in future work.

Task B. Table 5 shows the performance regarding misogyny detection at sentence level.

In Setting A, where the model is not informed of the output for Task A, AlBERTo scores are much lower compared to Task A, achieving 0.68 ± 0.03. Llama performs better in Task B compared to Task A, overcoming AlBERTo by just 0.01 point. However, the fact that all answers were provided in Task B (unlike in the previous

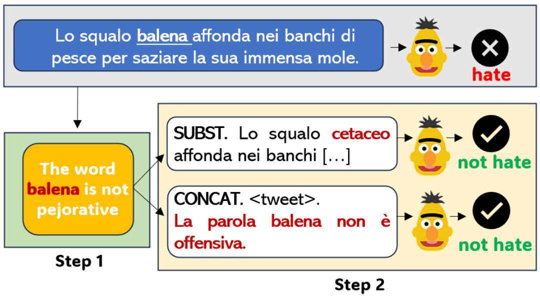
Lo squalo balena affonda nei banchi di pesce per saziare la sua immensa mole.
hate
The word balena is not pejorative
Step 1
SUBST. Lo squalo cetaceo affonda nei banchi [...]
not hate
CONCAT. <tweet>. La parola balena non è offensiva.
not hate
Step 2

Figure 2: Our pipeline for injecting information about pejorativity for Task B (setting B) in AlBERTo. 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.

task) plays a role and does not necessarily imply that misogyny detection is an easier task than pejorativity disambiguation for llama.

In Setting B, the model is informed of the decision on pejorativity of the target word. While for llama the information about pejorativity can be injected in the prompt, with AlBERTo we have adopted two approaches: *i)* we **concatenate** the information about the pejorativity of the target word at the end of the tweet or *ii)* we **substitute** the ambiguous word with its corresponding anchor word from our lexicon. Fig. 2 shows the pipeline. We observe a notable improvement over the baseline model for concatenation (+7 absolute points) and substitution (+9 absolute points) when using the predictions for Task A.

On the other hand, llama does not benefit from the injection of knowledge about pejorative words, with a drop of 0.09 points. This could be due to the noisy response from Task A, including the refusal to answer, and possible wrong explanations of why the target word is used pejoratively or not.

Setting	Model	Macro F1
A	AlBERTo	0.68
B_concat	AlBERTo	0.75
B_subst	AlBERTo	0.77
A	llama-3-8b-Instruct	0.69
B	llama-3-8b-Instruct	0.60

Table 5
Results on Task B. Scores are not comparable since the testing partitions are different.

5. Conclusion

We have presented a new challenge for CALAMITA: pejorative word disambiguation as a preliminary step for misogyny detection. We have designed two tasks as binary classification problems: A) pejorative language disambiguation at word level and B) misogyny detection at sentence level. Our preliminary experiments show that a Transformer-based fine-tuned model performs better than llama-3-8b-Instruct in detecting pejorative words, while llama-3-8b-Instruct performs slightly better than the Transformer-based model in misogyny detection. In the future, we plan to explore how the unrequested explanations provided by llama-3-8b-Instruct about the pejorativity of a target word impact the classification of misogynous sentences.

6. Limitations

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.

Moreover, 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.

Finally, our corpus is characterized by the presence of sarcasm, abbreviations, and non-standard varieties of Italian, which might make the semantics of our instances hard to be captured by current language models.

Another 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.

7. Ethical Issues

Our 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. However, the exposure to misogynistic content still poses a mental health risk for researchers and annotators.

Acknowledgments

Thanks to Federico Ruggeri, Cagri Toraman, Alberto Barrón-Cedeño, Lorenzo Musetti, Samuel Algherini, Silvia Ronchi, Gianmarco Saretto, and Caterina Zapparoli for helping with the brainstorming for the task design.

References

[1] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the Abilities of LAnguage Models in ITAlian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[2] F. M. Lopes, Perpetuating the patriarchy: Misogyny and (post-)feminist backlash, Philosophical Studies 176 (2019) 2517–2538. doi:10.1007/s11098-018-1138-z.

[3] M. Barreto, D. Doyle, Benevolent and hostile sexism in a shifting global context, Nature reviews psychology 2 (2023) 98–111. doi:https://doi.org/10.1038/s44159-022-00136-x.

[4] K. Srivastava, S. Chaudhury, P. S. Bhat, S. Sahu, Misogyny, feminism, and sexual harassment, Industrial Psychiatry Journal 26 (2017) 111–113. URL: https://journals.lww.com/inpj/fulltext/2017/26020/misogyny,_feminism,_and_sexual_harassment.1.aspx. doi:10.4103/ipj.ipj_32_18.

[5] A. Muti, A. Barrón-Cedeño, 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, 2020.

[6] E. Fersini, D. Nozza, P. Rosso, 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, Torino: Accademia University Press, 2018, pp. 59–66. doi:doi:10.4000/books.aaccademia.4497.

[7] E. Fersini, D. Nozza, P. Rosso, Ami @ evalita2020: Automatic misogyny identification, in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), CEUR.org, Online, 2020.

[8] P. Röttger, B. Vidgen, D. Hovy, J. B. Pierrehumbert, Two contrasting data annotation paradigms for subjective NLP tasks, in: M. Carpuat, M. de Marneffe, I. V. M. Ruíz (Eds.), 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, Association for Computational Linguistics, 2022, pp. 175–190. URL: https://doi.org/10.18653/v1/2022.naacl-main.13. doi:10.18653/v1/2022.naacl-main.13.

[9] K. Krippendorff, Computing krippendorff's alpha-reliability, 2011.

[10] A. Muti, F. Ruggeri, C. Toraman, A. Barrón-Cedeño, S. Algherini, L. Musetti, S. Ronchi, G. Saretto, C. Zapparoli, PejorativITy: Disambiguating pejorative epithets to improve misogyny detection in Italian tweets, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 12700–12711. URL: https://aclanthology.org/2024.lrec-main.1112.

[11] G. Abercrombie, A. Jiang, P. Gerrard-abbott, I. Konstas, V. Rieser, Resources for automated identification of online gender-based violence: A systematic review, in: Y.-l. Chung, P. R\"ottger, D. Nozza, Z. Talat, A. Mostafazadeh Davani (Eds.), The 7th Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 170–186. URL: https://aclanthology.org/2023.woah-1.17. doi:10.18653/v1/2023.woah-1.17.

[12] C. Gothreau, K. Arceneaux, A. Friesen, Hostile, Benevolent, Implicit: How Different Shades of Sexism Impact Gendered Policy Attitudes, Frontiers in Political Science 4 (2022). URL: https://www.frontiersin.org/articles/10.3389/fpos.2022.817309. doi:10.3389/fpos.2022.817309.

[13] 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, Bari, Italy, 2019. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074851349&partnerID=40&md5=7abed946e06f76b3825ae5e294ffac14.