ReCLAIM Project: Exploring Italian Slurs Reappropriation with Large Language Models

Marco Cuccarini[1,2,†], Lia Draetta[3,†], Chiara Ferrando[3,†], Liam James[4,†] and Viviana Patti[3]

[1]Department of Biology, University of Naples Federico II
[2]Department of Mathematics and Computer Science, University of Perugia
[3]Department of Computer Science, University of Turin
[4]DISI, University of Bologna

Abstract

Recently, social networks have become the primary means of communication for many people, leading computational linguistics researchers to focus on the language used on these platforms. As online interactions grow, recognizing and preventing offensive messages targeting various groups has become urgent. However, finding a balance between detecting hate speech and preserving free expression while promoting inclusive language is challenging. Previous studies have highlighted the risks of automated analysis misinterpreting context, which can lead to the censorship of marginalized groups. Our study is the first to explore the reappropriative use of slurs in Italian by leveraging Large Language Models (LLMs) with a zero-shot approach. We revised annotations of an existing Italian homotransphobic dataset, developed new guidelines, and designed various prompts to address the LLMs task. Our findings illustrate the difficulty of this challenge and provide preliminary results on using LLMs for such a language specific task.

Warning: This paper contains examples of explicitly offensive content.
Our positionality: This paper is situated in Italy in 2024 and is authored by researchers specializing in Natural Language Processing (NLP). Beyond our academic work, we are sensitive to *anti-hate speech* issues. Our backgrounds fields are theoretical linguistics, computer science and NLP.

Keywords

Semantic requalification process, Homostransphobia detection, Slurs, Natural Language Processing, Large Language Models

1. Introduction

In recent years, social networks have become the primary means of communication for most people. With the daily growth of online interactions, it has become urgent to recognize and prevent the spread of offensive messages against different target groups based on gender, sex, sexual orientation, race, religion, language, or political orientation. Moreover, categorizing hate speech with clear-cut boundaries is overly simplistic, as it includes various forms of abusive language that imply disrespect and hostility. A recent challenge is finding a balance between detecting hate speech and preserving the free spread of ideas and opinions on the web, while promoting inclusive and fair language. Thiago et al. (2021) [1] highlighted how automated analysis can misinterpret context, risking the censorship of marginalized groups languages, such as those of the LGBT+ community. Another study by Pamungkas and colleagues (2020) [2, 3] emphasized the importance of considering context in Natural Language Processing (NLP) tasks to avoid misinterpretations of word meanings, noting that the same swear word can be used both abusively and non-abusively. An example of this phenomenon is the semantic reappropriation, a practice in which terms historically used as slurs against a specific target group lose their offensive intent in certain contexts, by expressing a sense of belonging and solidarity within the group members [4]. Although community visibility and the use of specific slang have been approached for years, to our knowledge only some hate speech studies specifically addressed slurs, and few focused on slurs semantic reappropriation [5]. Nowadays, recognizing this kind of semantic shift through NLP tools is crucial to avoid the risk of removing not abusive speech in online contents, which could paradoxically harm marginalized users [6, 7].

Our study is the first with the aim of investigating reappropriative use of slurs in Italian, highlighting the need to take a step ahead from the existing abusive language detection models. Having in mind the capability of LLMs in classification task, we leveraged a LLM with a zero-shot approach in order to recognize the presence of reappropriative uses in our dataset.

This study makes the following contributions:

- We partially revised the original annotation previously conducted on the HODI dataset (Homo-

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

[†]These authors contributed equally.

marco.cuccarini@unina.it (M. Cuccarini); lia.draetta@unito.it (L. Draetta); chiara.ferrando@unito.it (C. Ferrando); liam.james2@unibo.it (L. James); viviana.patti@unito.it (V. Patti)

transphobic Dataset in Italian)[1] [8], by developing new annotation guidelines.
- We used a LLM specifically fine-tuned on Italian language by leveraging prompt engineering.
- From a linguistic point of view, we showed why certain features of the Italian language make this task particularly challenging.

This paper is structured as follows: in the Section 2 we review the most significant related work on hate speech detection and zero-shot approaches leveraging LLMs. In the Section 3 we describe our methodology for the dataset creation and the implementation of zero-shot tasks. In Sections 3 and 5 we respectively report results, analysis and main limitations of this work. Finally, in the last Section 6 we draw conclusions of the current research.

2. Related work

As presented above, hate speech is a challenging task, due to magnitude of the phenomenon and the difficulties of defining clear boundaries. Some recent developments in AI underlined the challenge of building corpora and models to automatically detect the abusive (or not abusive) nature of slurs in social media texts. Pamungkas' et al. (2020) [2] research focused on the use of swear words in English and aimed at differentiate between offensive and non-offensive occurrences of slurs. A Twitter English corpus, SWAD (Swear Words Abusiveness Dataset), was developed by manually annotating the abusive charge at the word level and models were trained to automatically predict abusiveness.

Over the last decade, most studies approached the hate speech detection in terms of binary classification [9]. For instance, Plaza et al. (2023) [10] examines this task by comparing the performances of an encoder-decoder model with several BERT-based models in both zero-shot learning and fine-tuning scenarios. The findings show that BERT-based models perform poorly in zero-shot learning, while the others, even without additional training, achieves results comparable to fine-tuned models.

Nowadays, research indicates that hate speech changes depending on the target groups [9]. Detecting homotransphobic hate speech (i.e. a specific abusive language addressed to LGBT+ community) has emerged as a critical research area, with various scholars proposing solutions in different languages such as English [11] and Italian [8, 12].

However, only few studies focused on the detection of slurs that have undergone a semantic reappropriation process. Zsisku and colleagues (2024) [5] approached the task by collecting the Reclaimed Hate Speech Dataset (RHSD), the first hate speech dataset dedicated at investigating the use of reclaimed slur terms, and by fine-tuning a baseline model which resulted in the Reclaimed Hate Speech (RHS) model.

As far as the Italian language is concerned, slurs recently became a significant topic from a linguistic and philosophical point of view, but there are not studies focusing on slurs reappropriation detection task. Philosophy of language studies highlighted that a key area of interest is slurs echoic uses, where target communities reappropriated derogatory terms to express pride, solidarity, or use them as tools for political and social activism [13, 4]. Nossem (2019) [14] observed a productive role in creating localized versions of *queer* by reappropriating and redefining existing local alternative terms, specifically *frocio* and *frocia*, *femminiellə*, and *ricchione*. At this point, it should be noted that Italian, differently from English language, lacks terms like *queer*, which bring with them such a long socio-cultural and historical background. The semantic requalification process of homotransphobic slurs is at its first steps and consists of a challenging task that has not yet been investigated in computational domains with LLMs.

3. Methodology

3.1. Dataset creation

To our knowledge, there are no available annotated datasets in the Italian language focusing on the phenomenon of slurs semantic reappropriation. To address the issue of limited data, in this preliminary research we utilized the HODI dataset [8], which contains 6000 Italian Twitter messages collected by using a set of 21 keywords (i.e., *gay, pride, lesbica, frocio*). The dataset is a collection of sentences directed against LGBT+ community who are target of homotransphobia. Our argument is that in such a corpus it is possible to find slurs used in both abusive and reappropriative contexts. With the aim of collecting messages suitable for our study, we filtered the HODI dataset by selecting tweets that contain at least one denigratory term, by adopting a two-fold strategy. To select the homotransphobic swear words, we used the HurtLex lexicon[2] [15], a multilingual lexicon containing an organized list of denigratory terms divided into 17 categories (i.e. negative stereotypes, ethnic slurs, moral and behavioral defects, words related to homosexuality). From HurtLex, we selected only the words categorized as homotransphobic, then we further narrowed the list to those that satisfy the slur definition[3] provided by Bianchi

[1]The HODI dataset was created for a shared task focused on identifying homotransphobia in Italian tweets.

[2]https://github.com/valeriobasile/hurtlex

[3]Bianchi (2014) [4] defines slurs "derogatory terms -such as 'nigger' and 'faggot'-targeting individuals and groups of individuals on the basis of race, nationality, religion, gender or sexual orientation. According to most scholars, slurs generally have a neutral counterpart,

Table 1
Examples of the target words in abusive context (Context 1) and semantic reappropriation context (Context 2)

Intention	Tweet	Translation
Abusive	Questo **frocio** con il tatuaggio del nome del moroso odio i gay.	This **fag** with the tattoo of his boyfriend's name I hate gays.
Not abusive	Io ero 6/7enne ed ero il **ricchione** alle elementari, all'oratorio, alle medie, al liceo e tutta la vita. E mi va bene così, c'è più colore in questo mondo 🌈	When I was 6/7 years old, I was the **gay** one in elementary school, at the youth center, in middle school, in high school, and all my life. And I'm okay with that, it adds more color to this world.

(2014,2015) [4, 16]. We chose to exclude words such as *gay, omosessuale, omofilo, pederasta*, and *diverso* because they are not strictly derogatory terms, hypothesizing that if words are not perceived as abusive, they cannot undergo a process of semantic reappropriation. After obtaining a list of 17 words, we filtered the HODI dataset by selecting only the tweets that contained at least one of the following target words: *anomalo, chiappa, frocio, invertito, travestiti, checca, deviato, culattone, finocchio, finocchi, finocchietto, sesso anale, frocia, ricchione, trans, troia, stesso sesso*. The resulting subset is a collection of 1742 tweets (see two examples in table 1).

3.2. Annotation guidelines

Establishing guidelines for such a subjective and previously unexplored topic has been challenging. Since the phenomenon lacks clear boundaries, we aimed to describe the task as clearly as possible. With this in mind, we based our guidelines on previous works in the field of the philosophy of language [4, 16, 13]. We asked three expert annotators to decide whether the target words in each tweet are used in a reappropriative context or not. Building on previously cited works, we defined reappropriation as the use of derogatory epithets by members of the target groups in a manner that is generally considered non-offensive. To better define the phenomenon we highlighted different contexts in which this linguistic behaviour could occur:

Friendly contexts – members of the target group use the derogatory terms in a non-offensive way in informal contexts.

- *Mamma mia raga come mi ha messa di buon umore il #LiguriaPride non mi sentivo così da un sacco grazie energia frocia* 😭😭❤️🌈🌈🌈
  [**English translation:** Mamma mia guys how the #LiguaPride has put me in such a good mood I haven't felt this way in a long time thanks FROCIA energy]

i.e. a non-derogatory correlate: 'Boche' and 'German', 'nigger' and 'African-American' or 'black', 'faggot' and 'homosexual'".

Political reappropriation contexts – target groups reclaim the use of derogatory epithets as a tool to emphasize a conscious and common political struggle.

- *Happy #PrideMonth e ricordatevi che l'orgoglio si celebra non solo quando andate a ballare nelle discoteche gay, ma anche quando si tratta di metterci la faccia e combattere per la causa perché altrimenti il ricchione lo state facendo solo col culo degli altri e non è carino* ❤️🌈
  [**English translation:** Happy #PrideMonth and remember that pride is celebrated non only when you go dancing in gay discos, but also when it comes to put your face out there and to fight for the cause because otherwise you are just being RICCHIONE on other people's ass and it is not nice]

Artistic contexts – artists reclaim derogatory epithets to subvert the dominant socio-cultural norms.

- *Poca gente che li guarda, c'è una checca che fa il tifo Se #LucioDalla avesse scritto #AnnaEMarco nel 2022 sarebbe stato accusato di omofobia, lui. Invece ha scritto una canzone immensa*
  [**English translation:** Few people look at them, there is a CHECCA cheering if #LucioDalla had written #AnnaEMarco in 2022 he would have been accused of homophobia. Instead he wrote a great song]

3.3. Zero-shot learning approach

After obtaining the described subset, we utilized zero-shot Learning (ZSL) with prompting to assess the model's ability to determine whether the target words are used in abusive or non-abusive context. Specifically, we employed the Qwen model [17], a multilingual decoded-only LLM pre-trained on Italian.

We define the temperature of the model to be 1, a fair trade-off between randomness and determinism in the results, and a maximum sequence length of 2024. For inference, an A100 GPU provided by Google Colab was

Table 2
Inter-annotator agreement metrics

Fleiss' Kappa	0.57
Annotators	**Cohen's kappa**
Annotator 1 vs Annotator 2	0.559
Annotator 1 vs Annotator 3	0.528
Annotator 2 vs Annotator 3	0.617

used. The code is available on the following GitHub page[4].

As previously discussed, collecting a large-scale corpus for reappropriated language detection is challenging. To address the lack of data, we used a ZSL approach, prompting the model to recognize the presence of semantic requalification without providing additional information. This method evaluates the model's ability to generalize effectively with no training data, taking into account only information acquired during the LLM training phase.

Different studies [18, 19] showed that ZSL results are significantly influenced by the appropriateness and precision of the prompts used. Additionally, multiple researchers [19] proposed different methods to improve performances. Plaza-del-Arco et al. (2022) [18] demonstrated that one of the most critical factors is ensuring that the prompt fits well with the utilized corpus. Taking this into account, we designed four different prompts using the HODI sub-corpus with the reappropriation annotation as the gold standard, each including specific details about the task and the corpora. The first one is the most general - explaining only the task in few words - while the fourth is as precise as possible providing full list of target words (full prompts are provided in Appendix A).

4. Results

4.1. Annotation statistics

We calculated the annotator agreement firstly by using Fleiss' Kappa, obtaining 0.57, secondly through Cohen's Kappa between pairs of annotators (all metrics are displayed in table 2). The moderate agreement and metrics variability highlighted the task's difficulty and subjectivity. Despite the three annotators being experts on the topic, they encountered challenges in distinguishing the use of slurs.

The majority annotation indicates that out of a total of 1742 examples, only 168 were annotated as reappropriated.

To better understand annotators disagreements and collect challenging examples, we conducted an analysis on tweets labeled differently (some examples in Appendix B). We observed that out of a total of 217 tweets with annotation disagreement, 67 (30.88%) contained the word "frocia". This word likely caused confusion due to its unique history: unlike the other target words "frocia", feminine form of "frocio", originated in an already reappropriative context [5] [14]. In some cases, due to a lack of context, it was very difficult to understand the real communicative intent of tweets (i.e., *Sono ricchione. (senso andiamo)* - "I'm gay. (like, let's go)"). In other instances, it was challenging to determine whether the person who wrote the message is part of the LGBT+ community or not (*Oggi il mondo mi sta urlando contro che sono un ricchione colossale senza speranza ed io gli sto dando ragione* - "Today the world is shouting at me that I'm a colossal hopeless queer, and I'm agreeing with it"), assuming that only members of target community can use slurs in reappropriative sense. Finally, we also identified some noisy data in which target words have different meanings. For example, in the sentence *Il 4 è l'onomastico di checca frenzis ci ubriachiamo* ("On the 4th, it's Checca Frenzis' name day, so we're getting drunk") the term "checca"[6] is likely used as a diminutive of the Italian name "Francesca".

We also noticed that in some cases tweets labelled as reappropriative were also labelled as homotransphobic in the original annotation of HODI dataset. Due to this apparent contradiction, we conducted a qualitative linguistic analysis on this data. We realized that in four examples (*Oggi avrò di che parlare coi colleghi..un etero analfabeta che conquista l'attenzione di una checca alfabetizzata* 😎*, mi raccomando vai a fare la quarta dose che forse ti aiuta a dimenticarmi. Ciao* - "Today I'll have something to talk about with my colleagues... an illiterate straight guy who captures the attention of a literate queer. Make sure to get your fourth dose, maybe it'll help you forget about me. Bye"), it is unclear whether the writer is part of the LGBT+ community or not. In other words, it is uncertain if the users were using slurs to refer to themselves with reappropriative intent or to other persons in abusive term. In addition, in some of these examples, target words were used as part of figures of speech, mostly similes (*Fare come una checca* - "Behave as a faggot"). These expressions, highly lexicalized in Italian and often used as abusive idiomatic phrases, likely increased the difficulty in recognizing the correct usages.

[4] https://github.com/marcocuccarini/ReCLAIMProject

[5] Nossem (2019) considers "frocia" as a calque of the English "queer" or "Alternatively, we could see it as a new concept which is specific to the Italian linguistic and cultural context, rather than an adaption or appropriation of the English "queer", i.e. some sort of a territorialised post-queer" [14].

[6] "Checca" as well as being a diminutive form of the Italian name "Francesca" is a colloquial and somewhat derogatory term in Italian used to refer to a gay man

Table 3
Zero-shot classification task results

Index	Weighted F1	Macro F1	Accuracy
1	0.64	0.43	0.55
2	0.73	0.49	0.66
3	0.66	0.45	0.57
4	**0.79**	**0.58**	**0.82**

4.2. LLM classification results

The results of the ZSL approach are detailed in Table 3. Notably, performances change among the prompts. The fourth prompt, which is the most specific, achieves the highest performance as it specifies all the target words considered during dataset construction. In contrast, the third one, focusing specifically on detecting homotransphobia by asking if the text intends to offend on the basis of sexual orientation or gender identity, has low performances. Among the four prompts, the first one ("Determine if the sentence contains semantic reappropriation; respond 'True' if it does and 'False' otherwise.") has the worst performances, likely due to the ambiguity of the expression "semantic reappropriation" for the model. Additionally, the model struggled to recognize the minority class (semantic requalification) because it is very complex for the model to recognize the context of the use of a slur, whether it is used to offend or not. This requires a deep understanding of the context and social dynamics, and it can also be a challenging task for humans.

To address this issue, balancing the information in the prompt by providing more details about semantic requalification could improve the model's overall performances. Therefore, we did not achieve very good performances, highlighting the importance of collecting new data and reviewing the computational approach.

5. Limitations and future works

The semantic requalification of slurs turned out to be a complex and time-consuming process in several aspects. Although the study has taken its first steps, some limitations must be acknowledged. Firstly, we realized that the HODI dataset [8] was not completely suitable for our purposes. Tweets had been collected for the homotransphobia detection aim and the difference of research goals did not provide us the right data to investigate the semantic requalification process of slurs. Secondly, a binary annotation proved to be limiting due to the difficulty of the task. The subjective evaluation of the annotators does not allow the problem to be simplified in terms of the presence or absence of semantic requalification process; therefore, a new scalar annotation scheme is probably required. Furthermore, the fact that only experienced young researches sensitive to LGBT+ issues were involved in the annotation task may have led to bias in the results.

As future work we plan to:

- create a new dataset and annotating it by following a perspectivist approach [7][20], i.e. by collecting different points of view from various social media, involving annotators with different backgrounds, in terms of age, origin, education, in/out target groups, and providing more context information during the annotation phase in order to better understand slurs' meanings and intents.
- through different LLMs, investigate which approach has better performances in recognising different uses of slurs, for instance by using ZSL approach between pairs of examples or defining few-shot with new suitable data.
- regarding ethical considerations, it is crucial to directly and actively involve the LGBT+ community. Gathering viewpoints and suggestions from those who experience daily oppression and denigration is essential not only to strengthen the research methodology but also to ensure its relevance and sensitivity to their lived experiences.

6. Conclusion

This paper presents the first attempt to specifically address the detection of slur reappropriation in the Italian language. One of the reasons that motivated us to undertake this task is the need to ensure a safe linguistic environment on social networks without risking the censorship of individual freedom of expression. Since there was no existing dataset to explore homophobic slurs in the Italian language, we filtered a pre-existing homotransphobic dataset to build a subset containing only tweets with slurs occurrences, used both abusively and non-abusively. We then designed precise new guidelines and annotated the filtered subset, focusing on the presence of slur semantic reappropriation. With the newly annotated dataset, we approached a classification task using LLMs with zero-shot techniques. Leveraging the Qwen model [17], we proposed four different prompts. As suggested by previous literature, more specific prompts and those better suited to the dataset yielded better performance. In this work, we proposed an important and under-explored task through a two-fold contribution. On one hand, we highlighted the lack of data in the Italian language dealing with this phenomenon and the necessity of building

[7] https://pdai.info/

an up-to-date corpus that comprehensively includes multiple sources and semantic contexts. On the other hand, we demonstrated a possible approach by leveraging new state-of-the-art LLMs. Finally, it is important to have in mind that compared to English, Italian has a different history and cultural background, resulting in a much slower linguistic evolution. This makes establishing precise characteristics of this topic a challenging task due to the lack of solid foundational knowledge. In conclusion, we believe that bringing attention to the issue will lead to anti-discrimination activities, the creation of safer spaces in online communication, and the inclusion and acceptance of LGBT+ communities.

References

[1] D. O. Thiago, A. D. Marcelo, A. Gomes, Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online, Sexuality & culture 25 (2021) 700–732.

[2] E. W. Pamungkas, V. Basile, V. Patti, Do you really want to hurt me? predicting abusive swearing in social media, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 6237–6246. URL: https://aclanthology.org/2020.lrec-1.765.

[3] E. W. Pamungkas, V. Basile, V. Patti, Investigating the role of swear words in abusive language detection tasks, Lang. Resour. Evaluation 57 (2023) 155–188. URL: https://doi.org/10.1007/s10579-022-09582-8. doi:10.1007/S10579-022-09582-8.

[4] C. Bianchi, Slurs and appropriation: An echoic account, Journal of Pragmatics 66 (2014) 35–44.

[5] E. Zsisku, A. Zubiaga, H. Dubossarsky, Hate speech detection and reclaimed language: Mitigating false positives and compounded discrimination, in: Proceedings of the 16th ACM Web Science Conference, 2024, pp. 241–249.

[6] T. Gillespie, Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media, Yale University Press, 2018.

[7] N. Strossen, Hate: Why we should resist it with free speech, not censorship, Oxford University Press, 2018.

[8] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the first shared task on homotransphobia detection in italian, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023, volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3473/paper26.pdf.

[9] 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.

[10] F. M. Plaza-del arco, D. Nozza, D. Hovy, Respectful or toxic? using zero-shot learning with language models to detect hate speech, in: Y.-l. Chung, P. Röttger, 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. 60–68. URL: https://aclanthology.org/2023.woah-1.6. doi:10.18653/v1/2023.woah-1.6.

[11] S. Kumar, A. Nagar, A. Kumar, A. Singh, Hate speech detection: A survey, in: 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), IEEE, 2022, pp. 171–176.

[12] D. Locatelli, G. Damo, D. Nozza, A cross-lingual study of homotransphobia on twitter, in: Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), 2023, pp. 16–24.

[13] B. Cepollaro, et al., Linguaggio d'odio, in: Pragmatica Sperimentale, Società Editrice il Mulino spa, 2022, pp. 145–156.

[14] E. Nossem, Queer, frocia, femminielle, ricchione et al. - localizing 'queer' in the italian context, GSI: Gender, Sexuality, Italy 6 (2019) 1–27.

[15] E. Bassignana, V. Basile, V. Patti, Hurtlex: A multilingual lexicon of words to hurt, in: E. Cabrio, A. Mazzei, F. Tamburini (Eds.), Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018, volume 2253 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018. URL: https://ceur-ws.org/Vol-2253/paper49.pdf.

[16] C. Bianchi, Il lato oscuro delle parole: epiteti denigratori e riappropriazione, Sistemi intelligenti 27 (2015) 285–302.

[17] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, T. Zhu, Qwen technical report, CoRR abs/2309.16609 (2023). URL: https://doi.org/10.48550/arXiv.2309.16609. doi:10.48550/

ARXIV.2309.16609. arXiv:2309.16609.

[18] F. M. Plaza-del Arco, M.-T. Martín-Valdivia, R. Klinger, Natural language inference prompts for zero-shot emotion classification in text across corpora, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 6805–6817. URL: https://aclanthology.org/2022.coling-1.592.

[19] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35.

[20] S. Frenda, G. Abercrombie, V. Basile, A. Pedrani, R. Panizzon, A. T. Cignarella, C. Marco, D. Bernardi, Perspectivist approaches to natural language processing: A survey, Language Resources and Evaluation Journal (2024). doi:10.1007/s10579-024-09766-4.

A. Zero-shot learning prompts

Prompt Index	Text	English translation
1	Stabilisci se la frase contiene riappropriazione semantica, rispondi 'Vero' se lo contiene e 'Falso' altrimenti.	Determine if the sentence contains semantic reappropriation; respond 'True' if it contains it and 'False' otherwise.
2	Stabilisci se la frase contiene un linguaggio che non ha intenzione di offendere, Rispondi 'Vero' se lo contiene e 'Falso' altrimenti.	Determine if the sentence contains language that has not abusive intent. Respond 'True' if it does and 'False' otherwise
3	Stabilisci se la frase contiene un linguaggio che intende offendere delle persone per il loro orientamento sessuale e le loro identità di genere, rispondi 'Vero' se lo contiene e 'Falso' altrimenti.	Determine if the sentence contains language intended to offend people based on their sexual orientation or gender identity. Respond 'True' if it does and 'False' otherwise.
4	Stabilisci se nelle frasi proposte le seguenti parole "frocio, invertito, travestit*, checca, deviato, culattone, finocchio, finocchi, omosex, finocchietto, omosessuali, frocia, ricchione, trans, troia" sono utilizzate per offendere le persone per il loro orientamento sessuale e/o identità di genere. Rispondi "Vero" se c'è un intento offensivo, altrimenti "Falso".	Determine if the following words in the proposed sentences—'frocio, invertito, travestit*, checca, deviato, culattone, finocchio, finocchi, omosex, finocchietto, omosessuali, frocia, ricchione, trans, troia'—are used to offend people based on their sexual orientation and/or gender identity. Respond 'True' if there is an offensive intent, otherwise respond 'False'.

B. Annotation disagreement examples

Category	Tweets	Translation
Containing "frocia"	Ho la bocca bollente...Voglio una **frocia** per me. Sono in uni e non riesco a non essere una **frocia** oggi aiutooo. Quanto è **frocia** la amo vuole la mappa cartacea per girare i giardini [URL]	My mouth is burning hot...I want a **fag** for myself. I'm at university and I just can't stop being so **gay** today, help! How **gay** is she, I love her, she wants a paper map to explore the gardens.
Lack of context	User_*sono **ricchione**. (senso andiamo). Uomo, marito, padre e **ricchione**.	User_*I'm **gay**. (like, let's go). Man, husband, father, and **faggot**.
Unknown writer membership	La fisica è una cosa da etero, e infatti io sono mezzo **ricchione**. Oggi il mondo mi sta urlando contro che sono un **ricchione** colossale senza speranza ed io gli sto dando ragione. Sto per fare un tweet molto **ricchione**	Physics is a straight thing, and in fact, I'm half **gay**.. Today the world is screaming at me that I am a colossal hopeless **fag**, and I'm agreeing with it. I'm about to tweet something very **gay**.
Noisy	Il 4 è l'onomastico di **checca** frenzis ci ubriachiamo 👊🏻👊🏻. Io e **checca** a spasso con i marmocchi. io, **checca** e la nostra fissa per i supermercati [URL]	On the 4th it's **Checca** Frenzis' name day, let's get drunk. Me and the **checca** taking the kids for a walk. Me, **Checca**, and our obsession with supermarkets