ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

TRADUZIONE, INTERPRETAZIONE E INTERCULTURALITA'

Ciclo 37

Settore Concorsuale: 10/L1 - LINGUE, LETTERATURE E CULTURE INGLESE E ANGLO - AMERICANA

Settore Scientifico Disciplinare: L-LIN/12 - LINGUA E TRADUZIONE - LINGUA INGLESE

HIDDEN IN PLAIN SIGHT: DETECTING MISOGYNY BENEATH AMBIGUITIES AND IMPLICIT BIAS IN LANGUAGE

Presentata da: Arianna Muti

Coordinatore Dottorato

Chiara Elefante

Supervisore

Luis Alberto Barron Cedeno

Co-supervisore

Beatrice Spallaccia

Esame finale anno 2025

Acknowledgments

Obtaining my PhD has profoundly impacted my life, and I wish to express my heartfelt gratitude to all those who played a significant role in making this journey possible and memorable, although it was one of the most difficult events of my entire life.

First and foremost, I would like to thank my supervisor, Alberto Barrón-Cedeño, for giving me the opportunity to pursue a PhD under his guidance, for his invaluable feedback, and for his unwavering support throughout this journey. His encouragement kept me going even in the most challenging moments. When we first met, my knowledge of NLP was limited, and I will always be grateful that he took a chance on me in this field. I thank him for believing in me when I did not. Without his trust and mentorship, I would not be where I am today. I want to thank Beatrice Spallaccia, my co-supervisor, it was a pleasure to get to know her and discuss about more theoretical aspects of misogyny and feminism. I wish we had more time to collaborate. I want to thank all the people in the DIT lab who gave me precious feedback and encouraged me to pursue this journey. In particular, I would like to thank my fellow PhD and Postdoc students from DIT and DISI (University of Bologna) for the deep and insightful conversations during our formal and informal meetings. I especially want to thank my friends and colleagues Katerina Korre and Federico Ruggeri who were great collaborators both in terms of ideas and practical help. I owe so much to you.

I owe a great debt of gratitude to the GroNLP lab, especially Tommaso Caselli, for his great supervision and patience when I was doing my research stay there. It was not an easy time of my life and all the people from the GroNLP lab made it bearable and joyful. I am also very thankful to them for the constructive and helpful feedback that I received during various talks and discussions. I grew so much during that time, both as a researcher and as a person.

I would like to express my appreciation to Lorenzo Musetti and all of his team, which gave me the opportunity to spend six months for an internship at the company Expert AI, allowing me to see the other side of research.

I want to express my deepest gratitude to my friends, whose support, encouragement, and laughter made the journey of writing this thesis not only bearable but truly meaningful. Your thoughtful conversations, late-night discussions, and constant belief in me helped me push through moments of doubt and exhaustion. Whether it was listening to my endless ramblings about the code that was not working, offering fresh perspectives, or simply reminding me to take a break, your presence made all the difference.

I particularly want to thank my parents and my sister for their unconditional love, patience, and unwavering support throughout this journey. Their encouragement, understanding, and belief in me gave me the strength to persevere through every challenge. No matter how difficult things became, they were always there to lift me up, reminding me of my abilities and keeping me grounded. I am deeply grateful for their presence every step of the way.

I also want to thank my conference buddies who helped me feel confident when facing the daunting experience of presenting my research in front of a room full of experts. Thanks for calming me down every time.

Last but not least I want to thank Debora Nozza and Franziska Pannach for their insightful feedback while reviewing my thesis. A double thanks to Debora who is giving me the amazing opportunity to conduct research at the MilaNLP lab, an opportunity I never imagined I would have. Her support and belief in me mean more than I can express, and I am truly grateful.

Abstract

This thesis investigates basic to advanced methods for detecting misogyny in social media, starting from the most explicit forms up to implicit and ambiguous expressions of misogyny across languages and within different platforms. With a focus on Italian and English, this research explores monolingual, cross-lingual, and multimodal approaches. Building upon transformer-based models and large language models (LLMs), the thesis examines the limitations of the former for misogyny detection, particularly regarding unintended bias with identity terms, ambiguities of pejorative language, and implications in language.

In addressing the complexities of pejorative language, this thesis conceptualizes pejorative epithet disambiguation as a preliminary step for misogyny detection in the form of a word sense disambiguation task. This includes the development of a new corpus for pejorative epithets in Italian: PejorativITy.

Then, this thesis delves into the complexities of implicit misogyny detection and explanation, examining how LLMs can contribute to understanding the underlying assumptions embedded in misogynistic language. In order to carry out the experiments, this thesis introduces the ImplicIT-Mis dataset, the first dataset specifically focused on implicit misogyny in Italian. Specifically, the experiments are designed to assess the potential of LLMs in recognizing and reconstructing implied meanings in misogynistic statements, which often require a nuanced understanding of social cues and common sense. This implicit language can manifest through backhanded compliments, stereotypes, irony, and other subtle forms, posing a challenge to traditional classification methods, which primarily rely on explicit markers of hate speech.

A unique aspect of the approach is the application of argumentation theory, which is used to decompose inferential processes behind misogynistic expressions. This approach aids in categorizing not only overt expressions of misogyny but also the implied and context-dependent forms that often evade

detection. Through carefully designed prompts and manual validation, the experiments reveal both the strengths and limitations of LLMs in reasoning-based detection tasks. Furthermore, I investigate the capacity of these models to identify social dynamics embedded within misogynistic language, which often remain hidden due to their implicitness. The findings underscore the potential of NLP models to identify and counteract misogyny. Nonetheless, challenges remain in creating context-aware and multicultural models that can reason about misogynistic language and that can adapt to the evolving landscape of online discourse.

Publications

1. Emanuele Moscato, **Arianna Muti**, and Debora Nozza. (2025). MNLP @ Multilingual Counterspeech Generation: Evaluating Translation and Background Knowledge Filtering. *Proceedings of the First Workshop on Multilingual Counterspeech Generation.* Association for Computational Linguistics: Abu Dhabi, UAE.

2. Katerina Korre, **Arianna Muti**, Federico Ruggeri, Alberto Barrón-Cedeño. (2025). Untangling Hate Speech Definitions: A Semantic Componential Analysis Across Cultures and Domains. *Findings of Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics.* Association for Computational Linguistics: Albuquerque, New Mexico.

3. Francisco Jáñez-Martino, Alberto Barrón-Cedeño, Rocío Alaiz Rodríguez, Víctor González-Castro, **Arianna Muti**. (2025). On persuasion in spam email: A multi-granularity text analysis. *Expert Systems with Applications.* Volume 265, 2025, 125767, ISSN 0957-4174.

4. **Arianna Muti**, Federico Ruggeri, Khalid Al-Khatib, Alberto Barrón-Cedeño, and Tommaso Caselli. (2024). Language is Scary when Over-Analyzed: Unpacking Implied Misogynistic Reasoning with Argumentation Theory-Driven Prompts. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21091–21107, Miami, Florida, USA. Association for Computational Linguistics.

5. **Arianna Muti**, Federico Ruggeri, Cagri Toraman, Alberto Barrón-Cedeño, Samuel Algherini, Lorenzo Musetti, Silvia Ronchi, Gianmarco Saretto, and Caterina Zapparoli. (2024). PejorativITy: Disambiguating Pejorative Epithets to Improve Misogyny Detection in Italian Tweets.

Proceedings of LREC 2024. ELRA and ICCL: Torino, Italia, pp. 12700-12711.

6. **Arianna Muti**. (2024). PejorativITy - In-Context Pejorative Language Disambiguation: A CALAMITA Challenge. *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy.

7. Katerina Korre, **Arianna Muti**, and Alberto Barrón-Cedeño. (2024). The Challenges of Creating a Parallel Multilingual Hate Speech Corpus: An Exploration. *Proceedings of LREC 2024*. ELRA and ICCL: Torino, Italia, pp. 15842-15853.

8. Francesco Antici, Federico Ruggeri, Andrea Galassi, Katerina Korre, **Arianna Muti**, Alessandra Bardi, Alice Fedotova, and Alberto Barrón-Cedeño. (2024). A Corpus for Sentence-Level Subjectivity Detection on English News Articles. *Proceedings of LREC 2024*. ELRA and ICCL: Torino, Italia, pp. 273-285.

9. Khalid Khatib, Sara Gemelli, Saskia Heisterborg, Pritha Majumdar, Gosse Minnema, **Arianna Muti**, and Noa Solissa. (2024). GroningenAnnotatesGaza at the FIGNEWS 2024 Shared Task: Analyzing Bias in Conflict Narratives. *Proceedings of The Second Arabic Natural Language Processing Conference*. Association for Computational Linguistics: Bangkok, Thailand, pp. 630-639.

10. Federico Ruggeri, Francesco Antici, Andrea Galassi, Aikaterini Korre, **Arianna Muti**, and Alberto Barrón-Cedeño. (2023). On the Definition of Prescriptive Annotation Guidelines for Language-Agnostic Subjectivity Detection. *Proceedings of Text2Story — Sixth Workshop on Narrative Extraction From Texts held in conjunction with the 45th European Conference on Information Retrieval (ECIR 2023).*

11. **Arianna Muti**, Francesco Fernicola, and Alberto Barrón-Cedeño. (2023). UniBoe's at SemEval-2023 Task 10: Model-Agnostic Strategies for the Improvement of Hate-Tuned and Generative Models in the Classification of Sexist Posts. *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics: Toronto, Canada, pp. 1138-1147.

12. Paolo Gajo, **Arianna Muti**, Katerina Korre, Silvia Bernardini, and Alberto Barrón-Cedeño. (2023). On the Identification and Forecasting of Hate Speech in Inceldom. *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing.* INCOMA Ltd.: Varna, Bulgaria, pp. 373-384.

13. Chiara Di Bonaventura, **Arianna Muti**, and Marco Antonio Stranisci. (2023). O-Dang at HODI and HaSpeeDe3: A Knowledge-Enhanced Approach to Homotransphobia and Hate Speech Detection in Italian. *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, vol. 3473., pp. 1-10. CEUR-WS.

14. **Arianna Muti** and Eleonora Mancini. 2023. Enriching Hate-Tuned Transformer-Based Embeddings with Emotions for the Categorization of Sexism. *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)*, vol. 3497, pp. 1012-1023. CEUR-WS.

15. Andrea Galassi, Federico Ruggeri, Alberto Barrón-Cedeño, Firoj Alam, Tommaso Caselli, Mucahid Kutlu, Julia Maria Struß, Francesco Antici, Maram Hasanain, Juliane Köhler, Katerina Korre, Folkert Leistra, **Arianna Muti**, Melanie Siegel, Mehmet Deniz Türkmen, Michael Wiegand, and Wajdi Zaghouani. (2023). *Overview of the CLEF-2023 CheckThat! Lab: Task 2 on Subjectivity in News Articles.* In *24th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF-WN 2023.* CEUR Workshop Proceedings, pp. 236-249.

16. **Arianna Muti** and Alberto Barrón-Cedeño. (2022). A Checkpoint on Multilingual Misogyny Identification. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop.* Association for Computational Linguistics: Dublin, Ireland, pp. 454-460.

17. **Arianna Muti**, Marta Marchiori Manerba, Katerina Korre, and Alberto Barrón-Cedeño. (2022). LeaningTower@LT-EDI-ACL2022: When Hope and Hate Collide. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion.* Association for Computational Linguistics: Dublin, Ireland, pp. 306-311.

18. **Arianna Muti**, Francesco Fernicola, and Alberto Barrón-Cedeño. (2022). Misogyny and Aggressiveness Tend to Come Together and Together We Address Them. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association: Marseille, France, pp. 4142-4148.

19. **Arianna Muti**, Katerina Korre, and Alberto Barrón-Cedeño. (2022). UniBO at SemEval-2022 Task 5: A Multimodal bi-Transformer Approach to the Binary and Fine-grained Identification of Misogyny in Memes. *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022).* Association for Computational Linguistics: Seattle, United States, pp. 663-672.

20. **Arianna Muti** and Alberto Barrón-Cedeño. (2020). UniBO@AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020).* CEUR Workshop Proceedings, CEUR.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In an era dominated by digital communication and social media, the widespread presence of misogyny online is a concerning phenomenon that makes online spaces unsafe for women, perpetuating stereotypes and social injustice. Even if social media were conceived to express personal opinions, share big events, follow people's updates, or just communicate with friends, an increasing number of users misuse it by engaging in trolling, cyberbullying, or by posting aggressive and misogynous contents (Samghabadi et al., 2020). According to Nocentini et al. (2010), these contents feature an imbalance of power, intention, repetition, anonymity and publicity. Anonymity, in particular, has allowed the amount of hateful posts to dramatically increase. As a consequence, social media platforms struggle to control *inappropriate* contents.[1] According to Vox (Italian Rights Observer), women are more targeted than ever on X, which was called Twitter at the time of the research. Although the overall number of hateful tweets started decreasing in 2020, the number of misogynous tweets increased significantly in the last four years: from 26% to 43%.[2]

Digital abuse can escalate to real-world violence. Fulper et al. (2014) and Blake et al. (2021) show that the amount of online misogyny is strongly correlated with events of rape and violence. Blake et al. (2021) identifies that

[1] `https://business.twitter.com/en/help/ads-policies/ads-content-policies/inappropriate-content.html`

[2] `http://www.voxdiritti.it/la-nuova-mappa-dellintolleranza-7/`

geolocated misogynous tweets co-occur with domestic and family violence in more than 400 areas across 47 American states. Some online communities celebrate violence against women, including physical violence, sexualized violence, online threats of violence, and the use of aggressive or violent language online. Incels (short for "involuntarily celibate") are one of these online communities, and they have been linked to the harassment, assault, and murder of women (Gosse et al., 2024).

Developing tools to automatically flag hateful language —specifically targeting women, who disproportionately experience bullying, abuse, and threats—is essential to fostering a safer online environment for them (Fallows, 2005). In particular, spotting harmful content is challenging when the text is misogynous only when considered within the context of the thread where it is posted. This is the case, for instance, of victim-blaming comments under news articles covering cases of rape or femicide, in which the responsibility for the abuse or the murder is implicitly placed on the victim. Consider the following comment in isolation: *It's always alcohol and drugs at the end of the story, I don't feel sorry.* On its own, it would not typically be considered misogynistic. Now, consider the context in which it was posted: a news article about a girl raped at a party. Creating automatic tools to detect such comments, so that they could be employed by social media companies for content removal, is timely to preserve the well-being of the victims of gender-based violence.

The need for detecting implicit misogyny becomes increasingly urgent as explicit hate speech often receives attention and condemnation, while subtler, more pervasive forms remain undetected, allowing biases to continue unchecked by content moderators. Large language models (LLMs) hold promise for tackling this issue, given their capabilities in solving language-related tasks. However, the effectiveness of LLMs in recognizing implicit misogynistic language across languages and different contexts remains underexplored. This thesis fills this gap by devising approaches specifically aimed at detecting and explaining misogyny across multiple languages and platforms, with a focus on ambiguous and implicit language. This allows to shed light on reasoning capabilities of such models.

1.2 Outline

The thesis is structured as follows.

Chapter 2 introduces the concept of misogyny and its presence in online spaces, while also diving into technological background. This chapter defines key terminology. It begins by offering a detailed exploration of misogyny, drawing distinctions between misogyny and sexism, and introducing implicit misogyny. Then, it addresses the technological side of the research, discussing the role of Transformers and LLMs in processing and analyzing language. These models serve as foundational tools for detecting misogyny across different languages and domains.

Chapter 3 presents a survey of misogyny detection within the NLP community, by providing an overview of annotated datasets that have been used in various tasks related to misogyny detection. These resources play a critical role in training models capable of identifying misogynistic content. Additionally, this chapter highlights the evolution of misogyny detection in terms of models, from early lexical approaches and the Transformers era up to the present, marked by the rise of large language models. Finally, it discusses the ongoing challenges in misogyny detection, with a focus on ambiguity and implicitness, the two key aspects of this thesis.

Chapter 4 addresses misogyny and sexism detection as classification tasks across different languages and platforms. The chapter details experiments on Italian misogynistic tweets and English posts from Gab and Reddit. It addresses unintended bias arising from identity terms, experiments with data augmentation, and tackles issues of cross-dataset generalization. It also explores cross-lingual approaches, with data in Italian, English, and Spanish to investigate model transferability, an important step towards multilingual misogyny detection frameworks. The chapter closes by expanding misogyny detection to multimodal data, specifically focusing on memes. It demonstrates that the multimodal model adopted struggles to detect misogyny when memes: *i)* rely on background knowledge or cultural references that the model lacks; *ii)* convey compliments about women (e.g., benevolent sexism); *iii)* contain identity terms or images of women, commonly associated with misogynistic content.

Chapter 5 introduces the concept of pejorative epithets—words or phrases with dual meanings that can be neutral or offensive depending on the context. This chapter features an innovative approach framing pejorative identification as a word sense disambiguation task. Experiments are carried out on the newly collected PejorativITy corpus, which includes Italian tweets annotated for pejorativity at the word level and misogyny at the sentence level. Results indicate that accurately disambiguating these terms enhances the overall misogyny detection model by reducing false positives. At the end, the ability of LLMs to correctly identify the meanings of pejorative epithets is investigated, showing lack of knowledge.

Chapter 6 investigates the capacity of LLMs to identify implicit misogyny, focusing on their ability to infer implied assumptions underlying subtle misogynistic statements. This chapter introduces the ImplicIT-Mis dataset, a novel Italian resource for implicit misogyny, and leverages additional datasets to analyze English-language instances. Through experiments, the chapter explores LLMs' ability to reconstruct implied assumptions in the form of free-text generation, and to identify social dynamics embedded in implicit misogyny through a multi-label classification. The results show that while LLMs can partially capture implicit misogynistic dynamics, they often struggle with culturally specific implications and lack the reasoning abilities to handle complex, indirect statements fully.

Chapter 7 summarizes the findings of the thesis and discusses potential future directions for research, highlighting the need for improved context-aware models that incorporate cultural and commonsense reasoning to advance implicit misogyny detection.

I would like to clarify that the mixed use of *I* and *we* in this thesis is intentional. I use *I* when referring to research and experiments I conducted independently, while *we* is used to acknowledge collaborative work carried out as part of a team. This distinction highlights the collective efforts behind certain aspects of this research.

Warning: this thesis discusses content that is sensitive or offensive in nature. Obfuscating offensive terms would hinder the clarity of linguistic analysis, preventing an accurate examination of misogynistic language in context. Direct representation is essential to preserve the authenticity of the data.

Chapter 2

Background

This chapter sets the basis for understanding the two core concepts of this thesis: the definition of misogyny and the computational tools used to detect it. This includes an examination of computational tools such as Transformer architectures and Large Language Models (LLMs), which represent state of the art in misogyny detection. Before delving into the technological aspects, it is essential to define misogyny and sexism from a gender- and feminist-studies perspective.

2.1 Misogyny: Terminology and Definitions

In recent years, the meaning and usage of the term **misogyny** have expanded far beyond the original definition; i.e., the hatred for women (Wrisley, 2023). In the context of this thesis, I follow and revise Lopes' definition of misogyny, referring to it 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 (Lopes, 2019; Barreto and Doyle, 2023). Misogyny occurs in the forms of male privilege, patriarchy, gender discrimination, sexual harassment, belittling of women, violence against women, and sexual objectification (Srivastava et al., 2017b). Patriarchal norms are understood as societal structures that uphold male authority, influence the power dynamics between men and women, and regulate women behavior by dominating, oppressing and exploiting them

Figure 2.1: The image represents a meme. The first image is from the Bojack Horseman TV series, season 3 episode 6. The one below is from Porta a Porta, an Italian political talk show.

(Walby, 1989). The meme in Fig. 2.1 represents an example of patriarchy in action in 2024: having men discussing women's reproductive rights. When misogyny is embedded in socially-accepted patriarchal norms, it is difficult for humans to recognize it, let alone for machines.

2.1.1 Misogyny as Hate Speech

Although there is not a universally recognized definition of hate speech, many researchers from different fields, ranging from sociology, psychology, linguistics (Spallaccia, 2020) to computer science (Zeinert et al., 2021) -

have acknowledged misogyny as a form of hate speech. Indeed, 43% of hate speech definitions, ranging from Wikipedia articles to research papers and legislation, include gender as component (Korre et al., 2024b). In particular, legal definitions have tended to dismiss gender as a feature of hate speech, especially in non-Western countries. The inclusion of gender in hate crime legislation has been heavily debated (Hagerlid, 2021). The most brought up reasons are that such hate crimes could be interpreted as interpersonal conflicts rather than acts of misogyny, gendered hate crime cases would overburden the justice system, since violence against women is so frequent in comparison to other forms of hate crime, and that the group of women is too large to be given protection under hate crime law (Hagerlid, 2021). The exclusion of women from hate speech definitions is not limited to legislation: some definitions consider hate speech as targeting "members of vulnerable minorities" or "with protected characteristics", by indicating race, religion, nationality, and ethnicity as the defining features of such groups.

In the context of this thesis, misogyny is treated as a form of hate speech, since the inclusion of gender is in line with hate speech based on other forms of prejudice or bias (Hagerlid, 2021).

2.1.2 Misogyny vs Sexism

There is an ongoing philosophical debate on whether sexism and misogyny are distinct concepts (Manne, 2017). Manne (2017) explores misogyny as a system that enforces patriarchal norms by punishing women who deviate from traditional gender roles, distinguishing it from sexism, which she defines as the ideological justification of patriarchy. Sexist ideology tends to discriminate between men and women, while misogyny differentiate between *good* women and *bad* ones. However, sexism and misogyny share a common purpose: to maintain a patriarchal social order. For Savigny (2020) instead, sexism is described as the expression of male superiority over women, and misogyny as a more violent expression of sexism that implies hate, sharing the view with Rodríguez-Sánchez et al. (2021) who treats misogyny as a subcategory of sexism. From a computational perspective, little distinction is made between the terms *sexism* and *misogyny* and usually researchers focus only on one of them. One exception is Parikh et al. (2021), who address both problems by first classifying sexism into 14 overlapping categories and then examining the efficacy of adapting the model for the detection (binary task) and classification (multi-class task) of misogynous tweets. According to them, sexism refers to discrimination based on one's gender that predominantly affects women, while misogyny implies hate or established prejudices against women. According to the *Ambivalent Sexism Theory* (Glick and Fiske, 2018), sexism includes both hostile and benevolent attitudes toward women. They argue that both forms of sexism function to maintain male dominance, aligning closely with the notion of misogyny as an expression of male hostility toward women. In this framework, misogyny is viewed as an extreme form of sexism, but fundamentally a part of the same system of discrimination. In the context of this thesis, I use the term misogyny rather than sexism, since my research applies only to women, while sexism should include nonbinary and gender non-conforming people as well. However, different definitions of misogyny and sexism will be given throughout the thesis. I will not discuss them here, as they neither reflect my conception of misogyny and sexism, nor come from sociology- or gender-studies, but from computer science and Natural Language Processing literature. The rationale is that during the collection and annotation process of misogynistic datasets that will be mentioned in the following chapters, different definitions of misogyny and sexism were followed.

2.1.3 Implicit Misogyny

Subtle misogyny represents an unequal and unfair treatment of women that is not recognized by many people because it is perceived to be normative and therefore does not appear unusual (Swim et al., 2004). Swim et al. (2004) make a distinction between covert and subtle sexism: subtle sexism is hidden but unlike covert sexism, subtle sexism is not intentionally harmful. In the context of this thesis, I group subtle and covert sexism into the term *implicit misogyny*, disregarding the intentionality component, which is hard to deduce from short social media posts.

Misogynous language is implicit when it shows coded or indirect language to disparage a woman on the basis of gender (ElSherief et al., 2021; Wiegand et al., 2021). Some linguistic devices through which implicit misogyny is expressed are:

Indirect sarcasm and humor (Wiegand et al., 2021; ElSherief et al., 2021).
E.g.: *I got divorced. I need to buy a dishwasher now* - implying that women's role is to wash dishes;

Euphemisms defined as a mild or indirect word or expression substituted for one considered to be too harsh or blunt (Wiegand et al., 2021). E.g.: *you inspire my inner serial killer* - a milder statement to indicate the will of killing;

Stereotypes (Wiegand et al., 2021). E.g.: *you're not that bad at driving for being a woman* - implying that women are typically not good at driving;

Call for action (Wiegand et al., 2021). E.g.: *she should be given 2000 volts* - 2000 volts stand for electric chair, implying the desire of the author for someone to kill the target woman;

Inferiority and dehumanizing language (Wiegand et al., 2021; ElSherief et al., 2021). E.g.: *you're such a baby doll* - although hidden behind a compliment, the target woman is compared to an inanimate object, a doll;

Inferences (Wiegand et al., 2021) Along with external world knowledge, they are required to understand the misogynistic nature of a statement. E.g.: *She looks like Moana and I guess they also share the same job* - Moana Pozzi was a famous Italian porn actress. One can infer that the author/speaker is implying in a derogatory way that the target woman is a sex worker.

Symbolic/metaphorical language (Wiegand et al., 2021). E.g.: *that toilet sucks* - toilet is a word for 'woman' in the Incel language. This linguistic phenomenon - the lexical shift of a word, understood only by a particular group of people, is called *dogwhistle* (Mendelsohn et al., 2023). This category includes also *pejorative language* (Dinu et al., 2021), which is context-sensitive, meaning that the same word might be neutral in one setting and pejorative in another.[1]

However, sometimes there are no linguistic indicators for implicit misogyny. This is the case, for instance, of the (un)acceptability of feminine job titles (Giusti and Iannàccaro, 2020), the dismissal of feminist causes, the men's rights movements united by their commitment to anti-feminism (Ging, 2019), and the traditional conservative values of a "natural order", sexual abstinence and "pro-life" statements (Giusti and Iannàccaro, 2020). Another aspect that

[1]Pejorative language will be thoroughly addressed in Chapter 5.

can be considered as a form of implicit misogyny is internalized sexism, i.e., women hating on other women or enacting sexist behaviors towards other women or themselves. This phenomenon is exemplified by the statement "I'm not like other girls" (Means, 2021). Such aspects will be further discussed in Chapter 6.

In the next section, I will introduce the most popular computational models used to detect misogyny.

2.2 Computational Background

To fully grasp how misogyny classification operates, it is essential to first examine the underlying computational models that enable such classification. Among the most advanced and widely used approaches in NLP there are Transformer-based architectures.

2.2.1 Transformers

Transformers are a type of deep learning model that have revolutionized the field of NLP. Introduced in 2017 (Vaswani et al., 2017), Transformers leverage a mechanism called self-attention, which allows the model to weigh the importance of different words in a sentence relative to each other. This self-attention mechanism enables transformers to handle long-range dependencies and capture the context of words more effectively than previous models. They have since become the foundation for many state-of-the-art NLP systems, including models like BERT (Devlin et al., 2019) and GPT (Kalyan, 2023), across a wide range of applications such as translation, summarization, and, of course, text classification.

Architecture The Transformer receives in input a piece of text and produces a transformed sequence as output. It has two main components: an encoder and a decoder block. Both the encoder and the decoder contain many layers, specifically six in the original Transformer model, meaning that the two blocks are composed of six encoders and six decoders. Each encoder has two components: a feed-forward neural network and a self-attention mechanism, while the decoder has an extra layer in between. That attention layer is responsible for helping the decoder to focus on relevant parts of the input sentence. Fig. 2.2 shows the architecture.

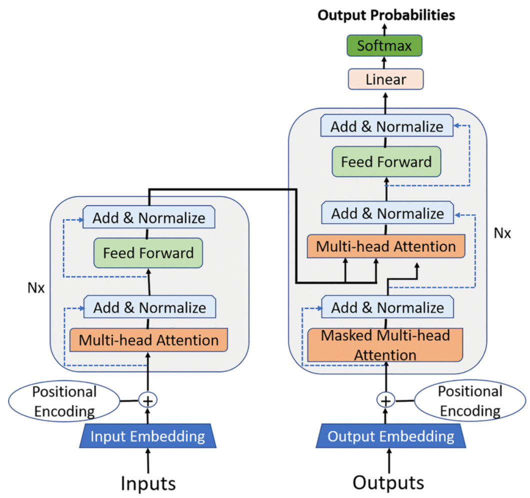


Figure 2.2: The Transformer model architecture.

To better understand their functioning, we can take as an example the language modeling task, which aims at predicting the next word, given a piece of text. If we feed the sequence "Call me by your" to the model, we expect the model to predict the word "name", referring to the movie "Call me by your name". The component responsible for this decision is the feed-forward neural network of the decoder. Predicting this word is not a challenging task, because Transformers-based models are usually trained on a very large amount of data and likely they have been exposed to this sequence of text before. However, there are more complex cases. If we have the sentence "The chicken did not cross the road because it", will the models guess what is the right sequence of words that come after? To do that, a model would need to understand what 'it' is referred to, because it could be the chicken, the road or another unknown entity. The component responsible for making this decision is the self-attention component. Therefore, an input text is first passed to the self-attention that considers relevant parts of the input, and then only those parts are fed to the feed-forward neural network for prediction.

2.2.2 BERT

BERT stands for Bidirectional Encoder Representation from Transformers and it is a task-independent language representation model, based on the Transformer architecture, introduced by Google AI researchers (Devlin et al., 2019). It is very popular because it has been shown to have reached state-of-the-art results for different NLP tasks such as text classification, question answering and named entity recognition, by performing a fine-tuning of the pre-trained model to the downstream task (i.e. the specific task that must be addressed) (Devlin et al., 2019). The advantage of BERT is that it can be fine-tuned by adding just one additional output layer to create state-of-the-art models for a wide range of tasks, without substantial task-specific architecture modifications.

Architecture BERT is an encoder-based model. The architecture is almost identical to that of Transformers (Vaswani et al., 2017), except for one final layer which is added in the fine-tuning phase. The parameters that can change according to the version are the number of encoder layers (L), the number of hidden units, which are the feed-forward networks (H), and the number of self-attention heads (A). `bert-base` is composed of L=12, H=768, A=12, while `bert-large` of L=24, H=1024, A=16. Just like the Transformer, BERT takes a sequence of words in input, which is passed through each layer of the encoder (self-attention first and then the output is sent to a feed-forward network), and then it is passed to the next encoder. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary.

Pre-Training and Fine-Tuning There are two steps involved in the use of BERT: pre-training and fine-tuning. For what concerns pre-training, BERT has been trained on a large unlabeled corpus, including the entire Wikipedia (2.5 million words) and Book Corpus (800 million words). BERT uses two training strategies: Masked Language Models and Next Sentence Prediction. For the first task, BERT uses a masking approach inspired by the Cloze task (Taylor, 1953). BERT randomly masks some input tokens within a sentence and then learns to predict the removed tokens based on the context. It is bidirectional because it makes use of Transformers that consider both the left and right context at once with respect to the hidden word to make the prediction upon. Moreover, it randomly replaces a word with another word and asks the model to predict the correct word in that position. For some tasks like question answering, it is important for the model to understand the relationship between two sentences, which is not directly captured by language modeling. That is why BERT has also been trained on the Next Sentence Prediction task: It receives two sentences as input, sentence A and sentence B and it learns to predict if sentence B follows sentence A in the original corpus. During training, 50% of the time sentence B is sequential to sentence A, while in the other case it is just a random sentence taken from the training set.

As for fine-tuning, BERT is first initialized with the pre-trained parameters, and all the parameters are fine-tuned using labeled data from the downstream tasks. The pre-trained architecture and the final task-specific architecture are very similar. The only thing that changes is that at the end, before the final output, one task-specific layer is added. For instance, in the case of text classification, a single-layer neural network (composed of a feed-forward NN and a softmax) will be added on top of the pre-trained BERT model, which is responsible to solve the actual classification task. For fine-tuning, most model's hyperparameters are the same as in pre-training, except for the batch size, learning rate, and number of training epochs.

2.2.3 Multilingual BERTs

mBERT (Devlin et al., 2019) has the same architecture as BERT, but it is trained on Wikipedia articles in multiple languages. Unlike many multilingual models that rely on language-specific tokenizers or parameters, mBERT operates using a shared vocabulary across languages, making it particularly effective for cross-lingual tasks. XLM-RoBERTa (Conneau et al., 2020) is another popular multilingual Transformer-based masked language model trained on one hundred languages jointly.

2.2.3.1 BERT for Italian: AlBERTo

Since mBERT has proven to be worse than a single-language model, as reported by the BERT documentation, language-specific models are born. For Italian, AlBERTo is the perfect fit for its focus on the language used on social networks, specifically on Twitter (Polignano et al., 2019). It has been trained on 200 million tweets randomly sampled from the TWITA corpus (Basile and

Nissim, 2013). The presence of mentions, uncommon terms, emojis, links, and hashtags, were all included in the text spans used for pre-training the model. The architecture is the same as BERT's. The difference is that it has been trained only on masked learning task, discarding the Next Sentence Prediction. The reason for that is that in the case of tweets, there is no flow of tweets as it happens in a dialog. For this reason, AlBERTo is not suitable for the task of question answering, but it has shown to perform very well on classification and prediction tasks. For a more-detailed list of BERT-based monolingual models on a variety of languages please refer to Nozza et al. (2020).

Although Transformers-based models were, until recently, the most popular architecture to perform classification tasks, the advent of Large Language Models brought another pradigm shift.

2.2.4 Large Language Models

In the context of this thesis, by Large Language Models (LLMs) I refer to decoder-based models such as GPT. GPT uses an autoregressive language modeling objective. It is trained to predict the next word in a sequence, making it more specialized for text generation tasks. LLMs have introduced a paradigm change in NLP. Their availability has been accompanied by claims about "emerging abilities" (Wei et al., 2022). Recent work has proposed distinguishing the acquisition of competencies in LLMs either as *abilities*; i.e. the capacity to solve a task absent in smaller models as an effect of the size of the models themselves, and *techniques*; i.e. the beneficial effect of different prompting methods that are ineffective in smaller models (Lu et al., 2023). The experiments that have been conducted —using zero-shot settings on multiple tasks— show that an ability such as reasoning is an effect of prompting techniques (e.g., instruction-tuning or in-context learning), rather than an emergent ability.

Next, I will present the LLMs that will be mentioned throughout the thesis, especially in Chapters 5 and 6.

LlaMa (Touvron et al., 2023b) `Llama-2-7b-chat-hf` and `Llama-3.1-8B-Instruct` are optimized versions of the original LlaMa model developed by Meta (Touvron et al., 2023a). According to the documentation, the vast majority of the training materials are in English, although some instances of other languages, including Italian data, are attested at training time (0.11% for Italian). Both models have undergone a phase of safety fine-tuning to prevent the generation of harmful content. This can hinder hate speech detection by causing refusals to engage with sensitive language.

Mistral (Jiang et al., 2023) (`Mixtral-8x7B-Instruct-v0.1` and `Mistral-7B-Instruct-v0.2`): This is a family of LLMs based on LlaMa-2. The models have been trained using different attention mechanisms, such as group-query attention and sliding-window attention (excluded from v0.2). The 7B versions of these models have been reported to obtain better performances when compared to LlaMa-2-7B and LlaMa-2-13B. Details about the dataset used to generate the models are lacking.

Tower (Alves et al., 2024) (`TowerInstruct-7B-v0`): This is a multilingual model based on LlaMa-2. Multilinguality is achieved by an initial phase of further pre-training of LlaMa-2 with a multilingual corpus (20 billion tokens over 10 languages, including monolingual and parallel data). After this phase, the model has been fine-tuned on a dedicated dataset for translation-related tasks, including paraphrase generation and named-entity recognition.

Falcon (Almazrouei et al., 2023) (`falcon-7b-instruct`): This autoregressive model is part of a family of LLMs which have been pre-trained on a high quality Web corpus, the RefinedWeb dataset (Penedo et al., 2023). The model has been further fine-tuned using 250M tokens mixture of instructions and chat-based datasets, mostly from Baize (Xu et al., 2023), thus making the model easily usable for assistant-style tasks. The model has been trained mostly on English data, thus potentially underperforming for Italian. According to the documentation, the model has not undergone a safety fine-tuning step.

LLaMAntino (Basile et al., 2023) (`LLaMAntino-2-chat-7b-hf-UltraChat-ITA`): This was the most recent language-specific LLM for Italian at the time of writing this thesis. LLmaMantino is adapted from LlaMa-2. The dataset used in this phase is the `clean_mc4_ita`, a cleaned version of the Italian split of the multilingual Common Crawl's web crawl corpus (mC4) (Sarti and Nissim, 2022). For the chat-based functionality, the LLaMantino models have adopted a supervised fine-tuning training (SFTTraining) approach on a translated version of the UltraChat dataset (Ding et al., 2023), using more than 500k dialogues.

After having provided a definitional background of misogyny and a computational background of technology developed to detect it, in the next chapter I first present the datasets annotated for misogyny and sexism that will be mentioned throughout the thesis, then I provide an overview of how Transformers-based models have been employed for misogyny detection.

Chapter 3

Approaches and Datasets for Misogyny Detection

In this chapter, an overview of the past and current approaches to automatic misogyny identification in social media is provided. Misogynistic content has captured the attention of the NLP community in recent years, with an increasing amount of language resources (Anzovino et al., 2018; Chiril et al., 2020a; Pamungkas et al., 2020b; Guest et al., 2021; Zeinert et al., 2021) and dedicated shared tasks (Basile et al., 2019b; Fersini et al., 2018d, 2020c, 2018b; Mulki and Ghanem, 2021; Kirk et al., 2023; Plaza et al., 2023) over multiple languages. First, I introduce the datasets targeting misogyny and sexism, that will be used for experiments throughout the thesis. Then, I provide an overview of the NLP methods and models employed in misogyny detection from the early days to the present.

3.1 Annotated Datasets

In this section, I introduce the datasets compiled specifically for misogyny and sexism detection that will be employed to develop new methods throughout the thesis. Among such datasets, the great majority have been collected in the context of shared tasks. Although other datasets for misogyny, sexism and other gender-based violence phenomena exist, I report only those that are relevant for the purpose of this thesis. For a complete list, please refer to Abercrombie et al. (2023) and Zeinert et al. (2021).

Shared Tasks Shared tasks are competitions or challenges organized to evaluate and advance the development of computational models for NLP tasks. These tasks typically involve a well-defined problem in NLP, such as sentiment analysis (Barnes et al., 2022), machine translation (Semenov et al., 2023), argument mining (Liu et al., 2023b), or hate speech detection (Basile et al., 2019b), for which participants build computational models to address it. In a shared task, organizers provide a standardized dataset, typically divided into training, validation, and test sets, along with clear evaluation metrics. Participants use the training data to develop and fine-tune their models, which are then tested on the unseen data (test set) to measure the performance. The results are compared among the participants based on predefined metrics. The purpose of shared tasks is to establish a common benchmark to evaluate the performance of various approaches with respect to a task, helping researchers compare models on the same dataset under the same conditions.

3.1.1 Automatic Misogyny Identification

The first dataset for misogyny detection is developed in the context of the Automatic Misogyny Identification (AMI) shared task.[1] AMI is initiated in 2018 at IberEval (Fersini et al., 2018c), a series of shared tasks that focuses on evaluating NLP systems in the context of Iberian languages.

The IberEval edition of AMI adopts an oversimplified definition of misogyny: hate or prejudice against women (Fersini et al., 2018c). It addresses two languages, English and Spanish, and two subtasks: task A concerns misogyny identification as a binary classification problem, while task B proposes a multi-class setting to classify misogynous tweets into the following categories:

Discredit Attempts to undermine or damage the reputation of women. Example: *@melaniatrump stupid fucking bitch*

Stereotype Reinforcement of negative stereotypes about women. Example: *Women are good only into the kitchen..#makemeasandwich*

Objectification Reducing women to objects or focusing on physical appearance. Example: *Big tits and no brain..this is what I want as wife!*

[1]The first actual dataset is Hewitt et al. (2016), but it contains only tweets featuring the following keywords: cunt, slut, and bitch. I discard it for its limited scope.

Sexual Harassment Tweets containing sexual comments or harassment. Example: *I wanna touch your tits..you can't immagine what I can do on your body.*

Threats of Violence Direct or indirect threats aimed at women. Example: *Stupid bitch I'll put you down on the floor and I'll rape you! You should be scared!*

Dominance Content expressing male superiority or dominance over women. Example: *Women are inferior to men..so shut up please!*

Derailing Attempts to dismiss, undermine, or divert discussions about gender issues. Example: *@yesallwomen wearing a tiny skirt is "asking for it". Your teasing a (hard working, taxes paying) dog with a bone. That's cruel. #YesAllMen*

The dataset is collected by employing three approaches: downloading tweets using a set of representative keywords (bitch, whore, cunt); monitoring potential victims accounts, e.g., gamergate victims and public feminist women; downloading the history of identified misogynists, i.e., who explicitly declared hate against women on their Twitter profiles. Data were collected in 2017 and labeled manually by two annotators. A third one was employed in case of disagreement, followed by a majority voting approach to obtain the final label. The Spanish corpus is composed of 3307 tweets, while the English one is composed of 3251 tweets. More data were collected for the test set: 831 tweets for Spanish and 726 for English.

As for misogyny identification in Italian, the first edition of the AMI shared task took place in 2018 (Fersini et al., 2018a) in the context of EVALITA, a series of shared tasks for Italian language. Table 3.1 shows statistics for the three AMI editions of 2018. Task A is a binary misogyny identification task, while Task B aimed at recognizing whether a misogynous content is person-specific or generally addressed towards women as a group. Furthermore, participants were required to classify misogynistic tweets into the categories mentioned above.

The second edition of AMI was held in 2020, again in the context of EVALITA (Fersini et al., 2020b). This time, only Italian tweets were considered. The new dataset developed for this task is composed of 6000 manually annotated tweets, balanced between the misogynous and not misogynous

		train		test	
Lang	**Venue**	**NM**	**MIS**	**NM**	**MIS**
EN	IberEval 2018	1,683	1,568	443	283
EN	EVALITA 2018	2,215	1,785	540	460
ES	IberEval 2018	1,658	1,649	416	415
IT	EVALITA 2018	2,171	1,828	509	491

Table 3.1: Class distribution for the three corpora in English (en), Spanish (es) and Italian (it). NM stands for non-misogynous and MIS for misogynous.

Partition	**Misogynous**	**(Aggr.)**	**Other**	**Total**
train	2,337	(1,783)	2,663	5,000
test	500	(176)	500	1,000
overall	2,837	(1,959)	3,163	6,000

Table 3.2: Statistics of the AMI 2020 corpus for misogyny and aggressiveness identification (Fersini et al., 2020b).

ones (see Table 3.2). This task focuses on both misogyny and aggressiveness identification. My participation in the task (Muti and Barrón-Cedeño, 2020) obtained the best performance with a multi-class approach based on Transformers. It will be described in Section 4.1.

3.1.2 Explainable Detection of Online Sexism

A more fine-grained dataset is produced for The Explainable Detection of Online Sexism (EDOS) task (Kirk et al., 2023), presented at the 17th International Workshop on Semantic Evaluation (SemEval-2023). This dataset contains 20,000 posts extracted from the social platforms Gab and Reddit. They emphasize the need for explainability by classifying posts into granular categories based on the provided taxonomy. Fig. 3.1 shows the taxonomy and the hierarchical tasks, while Table 3.3 shows statistics for each class. While Task A addressed the binary classification of sexist posts, Task B (represented by Category) and Task C (represented by Vector) aim at a fine-grained classification of sexism. The dataset for Task A is skewed towards the negative class. As the numbers for Task B show, derogation is the most frequent type of sexism, followed by animosity, while prejudiced discussions and threats are the least frequent. These four classes are further divided for

Category	Vector	Definition	Example
1. Threats, plans to harm and incitement	**1.1** Threats of harm	Expressing intent, willingness or desire to harm an individual woman or group of women. This could include, but is not limited to: physical, sexual, emotional or privacy-based forms of harm.	*I'll kill any women that talks back to me*
	1.2 Incitement and encouragement of harm	Inciting or encouraging an individual, group, or general audience to harm a woman or group of women. It includes language where the author seeks to rationalise and/or justify harming women to another person.	*Raping her would put her in her place*
2. Derogation	**2.1** Descriptive attacks	Characterising or describing women in a derogatory manner. This could include, but not limited to: negative generalisations about women's abilities, appearance, sexual behaviour, intellect, character, or morals.	*Women's football is so shit, they're so slow and clumsy*
	2.2 Aggressive and emotive attacks	Expressing strong negative sentiment against women, such as disgust or hatred. This can be through direct description of the speaker's subjective emotions, baseless accusations, or the use of gendered slurs, gender-based profanities and gender-based insults.	*I hate women*
	2.3 Dehumanising attacks and overt sexual objectification	Derogating women by comparing them to non-human entities such as vermin, disease or refuse, or overtly reducing them to sexual objects.	*Women are pigs*
3. Animosity	**3.1** Causal use of gendered slurs, profanities and insults	Using gendered slurs, gender-based profanities and insults, but not to intentionally attack women. Only terms that traditionally describe women are in scope (e.g. 'b*tch', 'sl*t').	*Stop being such a little bitch*
	3.2 Immutable gender differences and gender stereotypes	Asserting immutable, natural or otherwise essential differences between men and women. In some cases, this could be in the form of using women's traits to attack men. Most sexist jokes will fall into this category.	*Men and women's brains are wired different bro, that's just how it is*
	3.3 Backhanded gendered compliments	Ostensibly complimenting women, but actually belittling or implying their inferiority. This could include, but is not limited to: reduction of women's value to their attractiveness or sexual desirability, or implication that women are innately frail, helpless or weak.	*Women are delicate flowers who need to be cherished*
	3.4 Condescending explanations or unwelcome advice	Offering unsolicited or patronising advice to women on topics and issues they know more about (known as 'mansplaining').	*My gf always complains about period pains but she just doesn't understand the medical science for eliminating them!*
4. Prejudiced Discussion	**4.1** Supporting mistreatment of individual women	Expressing support for mistreatment of women as individuals. Support can be shown by denying, understating, or seeking to justify such mistreatment.	*Women shouldnt show that much skin, it's their own fault if they get raped*
	4.2 Supporting systemic discrimination against women as a group	Expressing support for systemic discrimination of women as a group. Support can be shown by denying, understating, or seeking to justify such discrimination.	*The leadership of men in boardrooms is a necessary evil—corporations need to be efficiently run*

Figure 3.1: The taxonomy of sexism provided by the EDOS task organizers.

Task C, which zooms into different subtypes of sexism. This dataset shows high annotation quality: Kirk et al. (2023) recruited highly-trained annotators who all self-identify as women, to mitigate implicit biases in labelling and to evoke a participatory approach to AI where the communities primarily harmed from specific forms of content are included in the data development process. The taxonomy was constructed based on gender- and feminist-studies and was further refined with empirical entries from the dataset to merge or adjust the schema (Kirk et al., 2023).

This dataset allows to showcase how all top systems failed to detect sexism in the following implicit cases:

- Expressions that appear to be positive in sentiment but are actually backhanded gendered compliments (e.g., "STEM isn't hard, women can

		train	dev	test
	Task A			
	Sexist	3,398	486	970
	Not Sexist	10,602	1,514	3,030
	Task B			
1	Threats	310	44	89
2	Derogation	1,590	227	454
3	Animosity	1,165	167	333
4	Prejudiced Discussion	333	48	94
	Task C			
1.1	Threats of harm	56	8	16
1.2	Incitement and encouragement of harm	254	36	73
2.1	Descriptive attacks	717	102	205
2.2	Aggressive and emotive attacks	673	96	192
2.3	Dehumanizing attacks and overt sexual objectification	200	29	57
3.1	Casual use of gendered slurs, profanities, and insults	637	91	182
3.2	Immutable gender differences and gender stereotypes	417	60	119
3.3	Backhanded gendered compliments	64	9	18
3.4	Condescending explanations or unwelcome advice	47	7	14
4.1	Supporting mistreatment of individual women	75	11	21
4.2	Supporting systemic discrimination against women as a group	258	37	73

Table 3.3: Class distribution for the tasks A, B and C.

do it too!").

- Gender stereotypes presented in jokes (e.g., "I just bought a smart car. It doesn't turn on when the wife is in the driver's seat.").
- Slang usage (e.g., "Tha Brothas want dem PAWGs") or less common gendered insults (such as "hag" or "witch").
- Irony (e.g., "Either you worship women or you're a misogynist.").

This analysis highlights which aspects of sexist language are problematic to detect automatically. My participation in the EDOS shared task is reported in Chapter 4.2.

3.1.3 Multimedia Automatic Misogyny Identification

The Multimedia Automatic Misogyny Identification (MAMI) dataset is developed for a shared task on misogyny detection in a multimodal setting: memes (Fersini et al., 2022). Organizers proposed two tasks:

Task A A basic task about misogynous meme identification, where a meme should be categorized either as misogynous or not.

Task B An advanced task, where the type of misogyny should be recognized among potentially overlapping categories: stereotyping, shaming, objectification and violence.

According to the MAMI guidelines, a meme is misogynous when it conveys an offensive, sexist or hateful message (be it weak or strong, implicitly or explicitly) targeting a woman or a group of women. Four kinds of misogyny are considered for this task:

Shaming occurs when memes insult or offend women because of their physical aspect.

Stereotyping represents a fixed idea or set of characteristics; physically or ideological.

	train	test
Shaming	1,274	146
Stereotype	2,810	350
Objectification	2,202	348
Violence	953	153

Table 3.4: Number of instances per class.

Objectification represents a woman like an object through the over-appreciation of her physical appeal (sexual objectification) or by depicting her as an object (a human being without any value as a person).

Violence shows physical or verbal violence toward women.

The MAMI dataset contains 10,000 memes and it is completely balanced with respect to misogynous and non-misogynous memes. Table 3.4 shows statistics for the multi-label setting. Stereotype is the most predominant class (56% of the training set), followed by objectification (44%), shaming (25%) and violence (19%).

3.1.4 Microaggressions

The Tumblr website www.microaggressions.com curates a collection of self-reported accounts of microaggressions (MAs), which are defined as statements that subtly or unconsciously express a prejudiced attitude toward a member of a marginalized group (Breitfeller et al., 2019). This dataset was collected by asking people to fill out an online form with three questions: the story of the microaggression, the context, and how the microaggression made them feel. From this website, Breitfeller et al. (2019) collected 2,934 accounts of MAs (henceforth, posts). Most posts are manifestations of bias, targeting social groups frequently discriminated against, where gender is the most targeted component (1,314 posts), followed by race (1,278 posts) and sexuality (461 posts), among others.

3.2 Models for Misogyny Detection

Early Approaches Early research on misogyny and sexism detection is primarily based on manually crafted rules and keyword-based approaches. These methods focus on identifying specific slurs or derogatory phrases that are typically associated with misogynistic content. The rule-based systems, while easy to implement, are criticized for their rigidity and inability to generalize across different forms of sexist language (Waseem and Hovy, 2016; Frenda et al., 2019). The earliest work that makes use of this approach is published in 2016 (Hewitt et al., 2016). Misogynous tweets were collected by searching several terms typically used to attack women, and then labeled manually. This study shows that in English, most misogynist tweets contain the following words: slag, bitch, whore, cunt, ugly, skank, pig, hysterical, unfuckable, fuckstruggle, rape, hole, lesbian, hoe, slut. Among these terms, the author chose 'bitch', 'cunt' and 'slag' to extract tweets and then coded them accordingly to their misogynistic nature. Most of the tweets were actually misogynous, although the word 'bitch' was used more than the other ones, and often tweets containing this word were labeled as not misogynous because they did not insult women, in fact they were referring to some lyrics or used as casual slurs.

A purely lexical approach ignores the fact that words that are often offensive may be used in ways that do not necessarily disparage women, and apparently inoffensive words might be used to offend women in a metaphorical way, proving that this blacklist-based approach is not very effective.

The introduction of supervised machine learning algorithms marked a significant step forward in the detection of misogyny. Traditional machine learning algorithms such as Support Vector Machines, Logistic Regression and Naïve Bayes were among the first to be used (Fersini et al., 2018c,a; Anzovino et al., 2018; Frenda et al., 2019). These models required feature extraction techniques such as bag-of-words (BoW), term frequency-inverse document frequency (TF-IDF), and n-grams to encode the textual data into a numerical format that algorithms could process. While traditional machine-learning models take statistical and linguistic features as input, deep-learning models exploit word or sentence embedding. BERT (Devlin et al., 2019) for instance uses a sentence-level embedding, which revolutionized text classification tasks by preserving the uniqueness of the language used in hate speech contexts.

3.2.1 The BERT Era

The advent of BERT in 2019 revolutionized the whole NLP field and misogyny detection was no exception (see for instance Sun et al. (2019); Mozafari et al. (2019); Parikh et al. (2019); Pamungkas et al. (2020b)). Among relevant works outside the context of shared tasks, Jaki et al. (2019) analyze and automatically identify misogynistic language on Incels.me, a forum created by male supremacists to disparage women (now shutdown). Similarly, the work I carried out in Gajo et al. (2023) introduces a multilingual corpus for the analysis and identification of hate speech in the domain of Inceldom, built from Incel Web forums in English and Italian, including expert annotation at the post level for two kinds of hate speech: misogyny and racism. This resource paved the way for the development of mono- and cross-lingual models for (a) the identification of hateful (misogynous and racist) posts and (b) the forecasting of the amount of hateful responses that a post is likely to trigger. Fersini et al. (2020a) show that exploiting stilometry to profile users can lead to good discrimination of misogynous and not misogynous contents. Sharifirad et al. (2018) exploit ConceptNet and Wikidata to improve sexist tweet classification by augmenting and generating data. Chiril et al. (2020b) explore BERT contextualized word embeddings complemented with linguistic features in order to distinguish reported sexist acts from intended sexist messages.

Multilingual Approaches Few works are focused on the multilingual identification of misogyny. Basile and Rubagotti (2018) adopt a bleaching approach, i.e. transforming lexical strings into more abstract features (van der Goot et al., 2018), and test their model on Italian and English. They use a Support Vector Machine with n-gram features. In their work, they train on L1 and test on L1, train on L2 and test on L2, and they also train and test on both languages in combination. Pamungkas and Patti (2019) create bilingual misogynistic data in English, Italian, and Spanish with machine translation to train in a source language and predict in a target language. Pamungkas et al. (2020a) use multilingual Transformers to identify English, Spanish, and Italian misogynous tweets. They only train a model on one language and test it on a different one, without considering all language combinations. Nozza (2021) perform zero-shot cross-lingual experiments across Italian, English, and Spanish misogynous and racist tweets, finding that multilingual models are not able to capture common taboo language-specific expressions specifically regarding women, such as interjections in non-hateful contexts (e.g., *porca puttana*). Nozza et al. (2022) release HATE-ITA, a set of multi-language models trained on a large set of English and Italian datasets. HATE-ITA performs better than monolingual models and seems to adapt well also on language-specific slurs.

Multimodal Models and Approaches Multimodal hate speech has attracted the interest of the research community only recently. In 2019, Facebook AI launched the Hateful Memes Challenge (Kiela et al., 2021b), which consisted in identifying hate speech in memes: hateful vs not. It is constructed such that unimodal models struggle and only multimodal models can succeed: difficult examples ("benign confounders") are added to the dataset to make it hard to rely on unimodal signals. This is extremely challenging for machines, because they must combine the textual and the visual modalities and capture how the meaning changes when they are presented together.

The most successful approaches used both *early fusion* and *late fusion* (Kiela et al., 2021a), with the former achieving the best results. Those include VilBERT (Lu et al., 2019), VisualBERT (Li et al., 2019), MMF(Singh et al., 2020), MMBT (Kiela et al., 2020a) and CLIP (Radford et al., 2021).

In late fusion approaches, systems for each modality are trained independently. The scores produced by each model are joined during inference to produce the final prediction (Kiela et al., 2020b). In early fusion, the different modalities are combined at an early stage to learn one single classification model (Kiela et al., 2020b).

The first work on automatic detection of sexist memes can be found in Fersini et al. (2019). They explore unimodal and multimodal approaches both with late and early fusion to understand the contribution of textual and visual cues on the MEME dataset, a dataset containing 800 sexist and not sexist memes. From their work, it emerges that an unimodal textual model performs better than image-based ones. Concerning multimodality, late-fusion perform better. Moreover, participants to the Multimedia Automatic Misogyny Identification (MAMI) shared task (Fersini et al., 2022) propose different approaches to address multimodality in misogyny detection. The majority of the teams exploit pre-trained models, distinguished in text-based, where the most used ones are based on BERT, and image-based models, where the most adopted ones are based on VisualBERT. Among these systems, considered by 90% of the teams, half of them adopt an ensemble strategy to make the final predictions. Among the top-performing approaches, Zhou et al. (2022) has the most original. They base their system on ERNIE-ViL-Large (Yu et al., 2021), a model which incorporates structured knowledge obtained from scene graphs to learn joint representations of vision-language. Moreover, they implement mitigation strategies for *unintended bias*: they mask biased words and apply template substitutions to memes which are overfitting in unimodal information. Chen and Chou (2022) base their approach on a data-centric principle. They first use CLIP model Radford et al. (2021) to obtain coherent vision and language features and then use a logistic regression model to make binary predictions. Then, they expand available training data by manually marking more samples from the evaluation set.

3.2.2 The LLM Era

So far, misogyny detection has been treated as a text classification task, but the advent of decoder-based models is starting to shift the paradigm. Traditional machine learning and encoder-based models struggle to correctly classify misogyny when sentences contain figurative, ambiguous, sarcastic and implicit language (Muti and Barrón-Cedeño, 2020). This could be overcome by decoder-based LLMs, as they could rely on their implicit knowledge to grasp the meaning of such terms. However, employing decoder-based models for misogyny detection is underexplored. For instance, Morbidoni and Sarra (2023) show that zero-shot detection GPT capabilities- against human annotations- outperform supervised baselines on their Reddit evaluation dataset and that ensembling different prompts further improve the accuracy.

Sap et al. (2020) handle misogyny detection as a style transfer task, by asking LLMs to turn implicit instances of biased language into explicit statements, in order to ease classification. The advent of LLMs made implicit hate speech detection progressively gain momentum in recent years, and several efforts have been put into the development of datasets for this purpose (Sap et al., 2020; ElSherief et al., 2021; Hartvigsen et al., 2022; Ocampo et al., 2023). However, such datasets do not focus on misogyny specifically.

After having presented the datasets, approaches, and models used to address misogyny detection, the next section focuses on the challenges that are still open and underexplored and that I intend to address throughout the thesis.

3.3 Open Challenges

One of the primary challenges in misogyny detection is dealing with **ambiguous and pejorative language**. Words or phrases that may appear harmless or neutral in some contexts can be weaponized in others to demean or undermine women. For example, terms like "bossy" or "emotional" might be used in non-misogynistic ways but can also be used to belittle women. The same goes for pejorative slurs like *bitch* that carry misogynistic overtones but may be cloaked in humor or sarcasm, or used casually without the intention to harm. This aspect will be handled in Chapter 5.

Additionally, language evolves rapidly on social media, and new slang or **coded language** can emerge as a way to express misogyny without directly using offensive terms. Computational models struggle to keep pace with these linguistic shifts, especially since words can take on different connotations depending on regional, cultural, or even subcultural contexts. For example, online forums or specific subgroups may develop their own derogatory slang for women that outsiders (and detection systems) may not immediately recognize, like in the case of Incels (Gajo et al., 2023) or dogwhistles (Mendelsohn et al., 2023).

Much of the misogyny on social media is not overtly aggressive or explicit but instead implicit and subtle, in order to avoid content removal. Rather than using direct hate speech or offensive terms, individuals might rely on insinuation or using a language that communicates misogynistic attitudes without triggering traditional misogynistic words. Implicit misogyny can manifest through patronizing comments, sexist stereotypes, microaggressions or generalizations about women's capabilities or roles in society. Detecting this **implicit language** is particularly challenging for current systems, as it often requires a deep understanding of social cues and commonsense reasoning. For instance, a phrase like "women should stick to what they're good at" might seem innocuous but conceal the harmful stereotype of the role of women "belonging to the kitchen". Similarly, passive-aggressive remarks such as "fattela una risata", translated as "lighten up", can carry misogynistic undertones when used to trivialize valid complaints or discomfort, implying that the woman is overreacting. Yet such comments may not be spot by existing algorithms that focus on more explicit hate speech. This aspect will be discussed in Chapter 6.

Finally, online conversations often happen in bursts, with different users responding to each other in quick succession. Detecting misogyny in such exchanges requires tracking the **conversational thread** to identify if certain statements are part of a broader pattern of harassment or if they are isolated remarks that might otherwise seem benign. However, most of datasets and models, with few exceptions (Gajo et al., 2023; Vidgen et al., 2021) analyze individual posts or comments in isolation. The ImplicIT-Mis corpus I collected, presented in Chapter 6, considers this aspect by capturing not only the misogynistic comments, but also their context of occurrence, i.e., the main post under which comments are posted. However, experiments on how well models can incorporate and employ the context is left for future work.

Chapter 4

Misogyny and Sexism as Classification Tasks

In this chapter, I present the deep learning models developed for the identification of misogyny and sexism in Italian, English and Spanish in a cross-lingual and multimodal setting. All the approaches handle misogyny and sexism detection as a classification task. Section 4.1 presents the models developed to identify misogyny and aggressiveness in Italian on Twitter; Section 4.2 addresses the fine-grained classification of sexism in English on Gab and Reddit posts. I experiment with obfuscation of identity terms to reduce unintended bias and data augmentation to have a more balanced dataset, introducing the issue of cross-dataset generalization. Section 4.3 addresses the problem of identifying misogyny in tweets in mono- and multilingual settings in three languages: English, Italian and Spanish. I explore model variations considering single and multiple languages both in the pre-training of the transformer and in the training of the downstream task to explore the feasibility of detecting misogyny through a transfer learning approach across multiple languages. Finally, in Section 4.4, I address misogyny detection in memes.

4.1 Misogyny and Aggressiveness in Italian

In this section, I approach the problem of spotting whether a tweet in Italian is misogynous and, if it is, whether it is also aggressive. Since the evaluation of the approach is based on the AMI dataset (see Section 3.1.1), we refer to misogyny according to the definition provided for annotating the dataset. According to Fersini et al. (2020b), a tweet is considered misogynous if it expresses hating towards women in the form of insulting, sexual harassment, threats of violence, stereotype, objectification, and negation of male responsibility. Moreover, a tweet is considered aggressive if it, implicitly or explicitly, presents, incites, threatens, implies, suggests, or alludes to: attitudes, violent actions, hostility or commission of offenses against women; social isolation towards women for physical or psychological characteristics; justify or legitimize an aggressive action against women.

Existing systems struggle to identify the target of an aggressive statement, leading to a great number of false positives (Fersini et al., 2020b). Therefore, the proposed experiments and analysis help to shed light on why existing models struggle to recognize the target of aggressive instances.

This chapter is adapted from Muti et al. (2022a).

4.1.1 Data

To address this task, I use the dataset collected by Fersini et al. (2020c): AMI 2020 (see section 3.1.1). AMI 2020 is composed of $5,000$ tweets, manually labeled according to two classes, misogyny and aggressiveness. The first one defines whether a tweet has been flagged as misogynous (positive class) or not (negative class). If a tweet has been flagged as misogynous, it is further determined whether it is considered as aggressive (positive class) or not (negative class). Therefore, the task is hierarchical.

The training dataset is fairly balanced in terms of misogyny. It contains $2,337$ misogynous and $2,663$ non-misogynous instances. A total of $1,783$ of the former are also considered as aggressive, whereas only 554 are not. The test set is composed of $1,000$ tweets.

The supervised data is randomly split into training and validation sets: $4,700$ instances for the former and 300 for the latter.

4.1.2 Hypotheses

The objective is to find the best pipeline to detect aggressive and misogynous tweets. We draw three hypotheses.

H1: Training the aggressiveness model on the full training set, even if half of the instances have not been labeled according to aggressiveness and therefore are assumed non-aggressive, boosts the performance. The intuition is that, even if the extra instances are noisy, the more data, the better.

H2: Both tasks are interdependent and provide relevant information for each other, hence instantiating a cascaded model to solve one problem next to the other is better than addressing both independently.

H3: A multi-class model that discriminates the classes altogether (aggressive-misogynous vs misogynous vs non-misogynous) performs better than a cascaded model because it assesses both problems at once.

4.1.3 Experimental Setup

No pre-processing other than using the pre-trained AlBERTo tokenizer is performed. AlBERTo (Polignano et al., 2019) is fine-tuned on the downstream tasks. A softmax output layer is used with either two units for the binary tasks of misogynous vs not-misogynous and aggressive vs not aggressive, and three units for the multi-class setting. Independent losses are computed for misogyny and aggressiveness in the cascaded settings; one single categorical cross-entropy loss is computed for the multi-class setting.

Other parameters include using AdamW optimizer with $\epsilon = 1^{-8}$ (Loshchilov and Hutter, 2017) and greedily search for the optimal batch size and epoch number with a held-out strategy in ranges $[16, 32]$ and $[5, 8, 10, 15, 20]$ respectively. The best parameters are 16 batch size over 8 epochs. All the experiments are run using Google Colab's GPU.

I perform Experiments `sing A` and `sing B` to challenge **H1**. Both use the same architecture: two cascaded binary models, one for misogyny and one for aggressiveness, as shown in Figure 4.1. The difference lies in the training set for the aggressiveness model: in `sing A` both misogyny and aggressiveness models are trained on the whole dataset; in `sing B` the aggressiveness model is trained only on instances labeled as misogynous in the first place. Setting

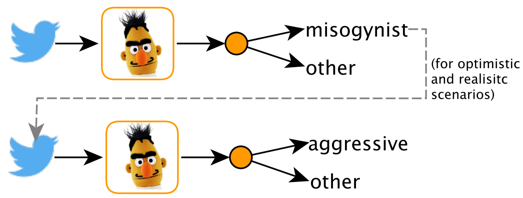


Figure 4.1: Cascaded architecture with two binary models (exps. `sing A` and `sing B`).

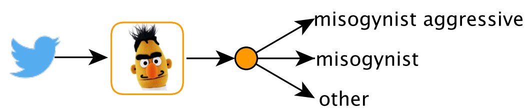


Figure 4.2: Multi-class architecture (exp. `multi`)

`sing B` aims at observing the behavior of the aggressiveness model when neglecting potentially noisy non-misogynous instances. These settings intend to mimic the corpus annotation procedure: identifying if a tweet is misogynous, and then, if true, whether it is also aggressive.

Three evaluation scenarios are considered for `sing A` and `sing B` to test **H2**: (*i*) naïve (baseline): the aggressiveness model predicts on all instances, regardless of the misogyny information; (*ii*) optimistic: only instances labeled as misogynous according to the gold standard are assessed for aggressiveness (i.e., assume a perfect misogyny classifier); and (*iii*) realistic: only instances identified as misogynous by the first classifier are classified by the aggressiveness model. For the last two scenarios, all non-misogynous instances are flagged as non-aggressive.

To challenge **H3**, model `multi` implements a multi-class model discriminating aggressive–misogynous vs misogynous vs non-misogynous at the same time. This is the same approach used to participate in AMI at Evalita 2020, for which our team Unibo ranked first (Muti and Barrón-Cedeño, 2020). Figure 4.2 represents its pipeline.

4.1.4 Results

In this section, the results of the computational experiments above mentioned are presented. Table 4.1 shows the F_1 score of the experiments.

When considering `sing A` vs `sing B` , as expected, the misogyny scores

exp	scenario	misogyny dev	misogyny test	aggressiveness dev	aggressiveness test	overall test
`sing A`	naïve	92.01	82.33	87.62	70.38	76.36
`sing A`	optim.	92.01	82.33	87.62	75.66	79.00
`sing A`	realistic	92.01	82.33	87.62	71.17	**76.75**
`sing B`	naïve	92.01	82.33	75.25	44.78	63.56
`sing B`	optim.	92.01	82.33	75.25	73.36	77.84
`sing B`	realistic	92.01	82.33	75.25	64.94	73.64
`multi`	–	87.59	82.48	84.76	68.61	75.54
unibo (us)	Muti and Barrón-Cedeño (2020)					74.38
jigsaw_1	Lees et al. (2020)					74.06
jigsaw_2	Lees et al. (2020)					73.80

Table 4.1: F_1 for `sing A` , `sing B` , `multi` and top Evalita 2020 models; three scenarios for aggressiveness detection for `sing A` and `sing B` .

remain the same since there is no difference in the training setup. Regardless of the scenario (used for H2 next), the aggressiveness model performs much better when `sing A` learns from the full dataset, despite the noise. Therefore, I consider hypothesis H1 to be true: using the full (partially-noisy) training material allows for better generalization.

Focusing on the three evaluation scenarios, regardless of the quality of the misogyny classifier, the aggressiveness one benefits from the filtered input of the optimistic and realistic scenarios. It is worth noticing that the performance shift is much smaller for the optimistic scenario (1.70 points absolute), in which 500 instances are simply assumed as non-aggressive. The drop in the naïve scenario is much bigger: 25.60 points absolute. Indeed, when comparing the three scenarios for both experiments, the naïve one —the only one which does not cascade the aggressiveness decision after that of misogyny—, consistently obtains the worst performance. This is a reflection of the veracity of H2: the better the input produced by the misogynous model (assumed to be perfect in the optimistic scenario), the more accurate the prediction of its aggressiveness.

Now focus on `multi` . Considering the information on aggressiveness helps to improve the decisions on the misogyny class, lifting it by 0.15 points absolute. Nevertheless, this comes at the cost of a degradation in the prediction of aggressiveness, which drops by 3.86 points, causing the `multi` architecture to run short with an overall F_1 = 75.54 (1.21 lower than `sing A` realistic).

Thus, modeling the problem in a cascaded rather than a multi-class fashion is better, refuting H3. Nevertheless, both architectures could be combined: a cascaded model departing from the enhanced misogynous decision from `multi` and ending in `sing A` for the aggressiveness decision results in $F_1 = 79.07$ in the optimistic and 76.83 in the realistic scenario. For comparison, the bottom of Table 4.1 shows the top Evalita 2020 models (Fersini et al., 2020b). The realistic `sing A` model outperforms all Evalita systems, including ours, reaching state-of-the-art performance in the task of misogyny detection in Italian.

4.1.5 Error Analysis

I perform a manual analysis on the non-misogynous instances judged as aggressive in the naïve scenarios of `sing A` and `sing B` , which flag on average 11 and 258 non-misogynous instances as aggressive, respectively. While in the optimistic and realistic models they are assumed as non-aggressive, in the naïve model they are subject to the model decision.

Models based on `sing B` rely on less negative instances, leading them to produce more false positives. In 80% of the cases, the false positive instances predicted as aggressive are indeed aggressive, but not according to the gold standard, because they are not women-oriented, while the final aim of this task is to spot aggressiveness targeting women. This is the case of instances 2 and 3 in Table 4.2, which shows examples of misclassified aggressive instances.

Words typically associated with aggression, such as *gola* (throat) and *schiaffi* (slaps) appear frequently in these instances (see instances 1 and 3 in Table 4.2). In both the training and test sets, *gola* is often the object of a violent act related to a sexual assault. All other instances contain swear words typical of aggressive language, in these cases used in non-aggressive circumstances and therefore misclassified as aggressive (see instance 4 in Table 4.2 which contains the casual use of the slur *bitch*). Kurrek et al. (2020) and Holgate et al. (2018) show that, in hate speech identification, the presence of swear words often leads to false positives when they occur in non-abusive contexts. They are often misclassified even if their function is not harmful, as they serve to intensify emotions and sarcasm (Pamungkas et al., 2020a).

Some neutral instances are misclassified because they contain words that are prone to misinterpretation, as they normally occur in women-oriented aggressive texts (*balena: whale/fat woman*; *scopo: aim/to fuck*). The word

	tweet	**mis**		**aggr**	
		actual	pred	actual	pred
1	mamma tranquilla che non sono l'unica 20enne che non sa cosa cazzo deve fare della sua vita chill che è già bello che non mi sia **sparata in gola** anni fa [*chill mom I'm not the only 20-something who doesn't know what the fuck to do with her life just be thankful that I haven't* ***shot myself in the throat*** *years ago*]	no	no	no	yes
2	@Nigagalsen @matteosalvinimi un follower dal 2016. . . e ti permetti di aprire quella **lurida** fogna di bocca che hai. [*@Nigagalsen @matteosalvinimi following since 2016. . . and you dare open that* ***filthy*** *ass mouth of yours.*]	no	no	no	yes
3	Comunque sti uagliuncelli del cazzo che mettono la musica sotto la finestra della camera in cui dorme mio padre (tornato stanchissimo da lavoro come sempre) **li prenderei a schiaffi** uno a uno [***I really want to smack*** *each one of these fucking kids putting on music right under the window of the room where my father's sleeping (exhausted after his day's work, as usual)*]	no	no	no	yes
4	[...] all'incirca mi ricordo qualche riga dei suoi testi ...."Che bello il lavaggio del cervello" Bum splash , la testa fa crash, **puttana** questo si chiama flash [*[...] I remember a few lyrics........"It's so nice to be brainwashed" Boom, splash! The head crashes!* ***Bitch*** *this is a flash*]	no	no	no	yes

Table 4.2: Instances from the test partition, their actual class and the one predicted by the naïve models.

scopo occurs 42 times in the training set, solely in instances labeled as misogynous-aggressive, it is therefore no surprise that our model classifies instances including that word as aggressive. However, this does not apply to *balena* which is equally used pejoratively as an insult and as a reference to the animal. The issue of polysemic words will be thoroughly addressed in Chapter 5.

In addition, 300 random instances misclassified by `multi` are analyzed. Again, the focus in on aggressiveness, because this is where the model struggles the most. Similar patterns of errors are observed. The most common mistakes are men-targeted aggressive tweets and polysemic words with a neutral meaning and a misogynous one.

4.1.6 Summary

This section presented a number of architectures for the tasks of misogyny and aggressiveness identification in Italian tweets. The experiments showed that addressing the two problems together —through two cascaded binary models— results in the best performance, and that the aggressiveness model benefits from the filtered output of the misogyny model. Moreover, training the aggressiveness model on the whole dataset turns out to be better than training only on the misogynous partition. The top model obtains $F_1 = 76.75$, reaching state-of-the-art performance. However, our models struggle the most when identifying aggressiveness, showing that it is hard to identify the target gender of an aggressive statement in short tweets, and that polysemic words used pejoratively confuse the classifiers.

Having examined the identification of misogynistic language in Italian tweets, the following section broadens the scope to the identification of sexist language in English, within Gab and Reddit posts. This shift allows for a preliminary exploration on how misogyny manifests across languages and platforms. Additionally, the next section aims to address unintended biases related to identity terms (Nozza et al., 2019), which can arise when models overly associate specific identity terms with hateful content. Standard transformers are compared to hate-tuned transformers to obtain an improvement in the performance. The evaluation of the proposed framework is carried out on the EDOS dataset (see Section 3).

4.2 Explainable Sexism Detection in English

This section addresses the detection of fine-grained categories of sexism on English posts from Reddit and Gab. This section is adapted from Muti et al. (2023), which describes my participation in the Explainable Detection of Online Sexism (EDOS) shared task. The shared task comprises three sub-tasks, which are all addressed.

Task A - Binary Sexism Detection: systems have to predict whether a post is sexist or not.

Task B - Category of Sexism: if a post is sexist, systems have to predict one of four mutually exclusive categories: (1) threats, (2) derogation, (3) animosity, or (4) prejudiced discussions.

Task C - Fine-grained Vector of Sexism: if a post is sexist, systems have to predict one among 11 mutually exclusive subcategories, e.g., threats of harm, descriptive attacks (see Table 3.3, bottom). This task is hierarchical to Task B.

For Task A and B, we employ BERT-based models (Devlin et al., 2019) and RoBERTa-based models (Zhuang et al., 2021) fine-tuned on hateful data. Two model-agnostic techniques are tested: obfuscation of identity terms and data augmentation with and without active learning. These strategies apply only to the model inputs and not to the internal model structure.

The first strategy involves masking gender-identifying information such as names and pronouns and aims at reducing *unintended bias*: the correlation between identity terms and the hateful class (Nozza et al., 2019; Dixon et al., 2018). The second strategy uses feedback from the model to iteratively select new training examples that positively influence the performance on the validation set.

For Task C, we employ a single hate-tuned model, RoBERTa-hate, and, given the poor performance, we explore the potential of generative models in such a fine-grained text classification setting via in-context learning.

4.2.1 Data

The proposed approach is evaluated on the dataset by Kirk et al. (2023), constructed for the EDOS shared task (see Chapter 3). Table 3.3 in Chapter

3 shows the class statistics for Tasks A, B and C. The dataset for Task A is skewed towards the negative class. For what concerns Task B, derogation is the most frequent type of sexism, followed by animosity, while prejudiced discussions and threats are the least frequent.

In addition to labeled data, we use 2M unlabeled posts provided by Kirk et al. (2023) —1M from Gab and 1M from Reddit—, which are used to augment the training set through active learning (Hino, 2020).

4.2.2 Models Description

We experiment with hate-tuned Transformer-based models and generative models. We compare the former with their original counterpart: BERT (Devlin et al., 2019) and RoBERTa (Zhuang et al., 2021). All hate-tuned models are fine-tuned on our downstream task. We perform a minimum parameter selection tuning on the validation set (10% of the training set). We select the highest performing learning rate $\in [1\text{e}{-5}, 2\text{e}{-5}, 1\text{e}{-2}]$; batch size $\in [4, 8, 16, 32]$; epochs in range $[1-10]$. Table 4.3 shows the best parameters for each model.

model	**task**	**lr**	**bs**	**epoch**
roBERTa-base	A	1e-5	16	4
roBERTa-base	B	1e-5	8	5
bert-base-uncased	A, B	2e-5	16	4
roberta-hate	A, B	1e-5	16	5
roberta-hate	C	1e-5	16	6
HateXplain	A	2e-5	16	8
HateXplain	B	2e-5	16	5
HateBert	A, B	2e-5	16	5

Table 4.3: Best hyperparameters per model and task, as fine-tuned on the development set (lr=learning rate, bs=batch size).

4.2.2.1 Hate-Tuned Encoder Models

We experiment with:

twitter-roberta-base-hate (Barbieri et al., 2020): a RoBERTa-base model trained on 58M tweets and fine-tuned for hate speech detection with the

TweetEval benchmark (Basile et al., 2019a).
hateXplain (Mathew et al., 2021): a BERT model trained on Twitter and Gab hateful posts. Each post has three levels of annotation: a multi-class labeling —hate, offensive or normal—, the target community and the rationales (i.e. the span of the post on which the labelling decision is based).
hateBERT (Caselli et al., 2021a): a re-trained BERT model for abusive language detection, trained on RAL-E, a large-scale dataset of Reddit comments in English from communities banned for being offensive, abusive, or hateful.

4.2.2.2 Generative Models

Although the GPT-3 family currently represents the de-facto standard for generative models, it is not open source and is only accessible through its dedicated API, which not only limits the possibilities for fine-tuning, but also bills per token, making it notably expensive (Webson and Pavlick, 2022).

For this reason, we opt for GPT-Neo (Wolf et al., 2020; Black et al., 2021; Gao et al., 2020), a transformer model developed starting from EleutherAI's GPT-3 architecture replica and trained on the Pile, a large scale curated dataset for language generation. We experiment using both the 1.3B parameter model (`EleutherAI/gpt-neo-1.3B`) and the 2.7B parameter model (`EleutherAI/gpt-neo-2.7B`).

4.2.3 Experiments and Results

In this section, I present the experiments performed for each of the tasks, along with the results on the development and test set.

4.2.3.1 Task A

Table 4.4 shows the results for Task A comparing standard encoder-based models with hate-tuned models. Among the standard models, BERT outperforms RoBERTa by 0.03 points. However, RoBERTa runs short compared to its hate-tuned version, RoBERTa-hate. The system submitted to the EDOS shared task (roberta-hate) differs from the top-performing one only by 0.04 points.

In the next paragraph, I present the strategies adopted to address unintended bias and data augmentation, along with the results.

model	strategy	dev	test
roBERTa-base	–	0.813	–
bert-base-unc.	–	0.781	–
HateXplain	–	0.791	–
HateBert	–	0.839	–
roberta-hate	–	**0.845**	0.835
roberta-hate	obfuscation	0.840	–
roberta-hate	data aug. (ext.)	0.820	–
roberta-hate	data aug. (int.)	0.830	–

Table 4.4: Macro F_1 for Task A. Strategy *data aug. (ext.)* refers to adding the positive instances from external datasets from similar tasks; *data aug. (int.)* refers to data produced via active learning.

Obfuscation of Identity Terms From the observation of previous tasks on misogyny detection (Fersini et al., 2020b; Nozza et al., 2019; Muti and Barrón-Cedeño, 2020), and from a preliminary error analysis on the validation set, I noticed that identity terms might lead to biased model decisions. Identity terms tend to be associated with the positive class due to their high co-occurrence. To reduce those spurious correlations, we obfuscate all identity terms in the training set. Specifically, we replace all instances of identity terms (woman, girls, female, etc.) with a generic placeholder token; e.g., [THEM] for plural and [IT] for singular forms.[1] We then train our best-performing model — `roberta-hate` — on this obfuscated dataset and evaluate its performance on the un-obfuscated dev set. Table 4.4 shows the results. The performance drops by 0.005 compared to our best model, i.e. `roberta-hate`. Figure 4.5 shows the confusion matrix for these two models. We manage to decrease the false positive rates, by limiting the spurious correlations with identity terms, at the expenses of an increase in false negatives. Given the unsatisfactory results, we discard this approach in the next steps.

Data Augmentation with External Resources Since the dataset is heavily imbalanced, we exploit the following task-related datasets annotated for misogyny or sexism to increase the size of the training set:

[1]For example, the sentence `I hate women` would be transformed into `I hate [THEM]`.

	RoBERTa-hate	
	Positive	Negative
Positive	1412	102
Negative	123	363

	R-h + Obfuscation	
	Positive	Negative
Positive	1421	93
Negative	135	351

Table 4.5: roberta-hate and roberta-hate+obfuscation

- **SBIC** (Sap et al., 2020) 150k social media posts with implied bias and offensiveness. The data comes from Reddit, Twitter and hate websites, such as Gab and Stormfront. We select those targeting women (3.7k posts), aiming to make the model more sensitive to implicit sexist statements.

- **AMI** (Fersini et al., 2018a; Anzovino et al., 2018) 4.4k misogynous tweets of the two editions of Automatic Misogyny Identification targeting the English language.

- **The 'Call me sexist but' Dataset** (Samory, 2021) 2.1k sexist tweets collected by querying the phrase 'call me sexist but', which were subsequently removed, leaving only the remaining text. This dataset contains in addition 1.1k hostile sexist instances from Waseem and Hovy (2016) and 821 instances of benevolent sexism from Jha and Mamidi (2017).

- **Microaggressions** (Breitfeller et al., 2019) 1.3k gender-based posts from *microaggressions.com*, which collects self-reported microaggression episodes. More details can be found in Chapter 3.

- **Incels.is** 1.1k posts that Gajo et al. (2023) bootstrapped from the *Incels.is* forum, annotated for misogyny.

- **Implicit Hate** (ElSherief et al., 2021) 6.4k implicitly hateful tweets, annotated for the target (e.g., race, religion, gender). We select the 65 posts targeting women.

dataset	dev	dev_sampled
SBIC	0.78	0.80
AMI	0.80	0.82
Call me sexist but	0.81	0.81
Microaggressions	0.82	0.82
Incels.is	0.81	0.81
Implicit Hate	0.82	–

Table 4.6: Macro F_1 score on the development test for Task A showing the impact of each dataset in the data augmentation process. All models are trained with roberta-hate, with the original training set plus the instances of each dataset.

We add in bulk only the positive instances to make the dataset more balanced. Table 4.4 reports the results of this experiment, showing a drop of 0.02 points compared to our best model.

We also include an ablation study in which we add one dataset at a time at the original training set, including both positive and negative instances from the external datasets. Results are shown in Table 4.6 column dev: we see no improvements in the performance with respect to our best model, i.e. roberta-hate. To better understand the impact of each dataset individually, we train a model with the original training set plus the same number of instances across all external datasets. We select the dataset with the least number of instances (Incels.is - 1.1k) and we select 1.1k instances from all datasets to be added to the original training set.[2] Column dev_sampled in Table 4.6 shows the results. The outcome changes only when adding data from SBIC and AMI, observing an improvement of 0.2 points in both cases. Limiting the number of external instances leads to an improvement in the performance, confirming that more data is not always the better in this cross-domain setting (Twitter, Reddit, blogs). This finding is in line with current research, claiming that hate speech detection models show low generalizability across datasets (Yin and Zubiaga, 2021).

Data Augmentation with Provided Resources Since the previous technique did not yield positive results, we employ data augmentation using the unlabeled data provided. We use the following approach. Let D_l be

[2]We discarded Implicit Hate because it counts only 65 instances.

model	strategy	dev	test
roBERTa-base	–	0.614	–
bert-base-uncased	–	0.570	–
roberta-hate	–	**0.638**	0.58
HateXplain	–	0.578	–
HateBert	–	0.606	–

Table 4.7: Macro F_1 for Task B. Only the top-performing model has been run on the test set.

our labeled training set, D_u the unlabelled dataset, and m_r our best baseline (`roberta-hate`): (*i*) Train m_r on D_l. (*ii*) Predict the instances in D_u with m_r. (*iii*) Rank the instances in D_u according to the confidence of the prediction score returned by m_r. (*iv*) Add iteratively the top-k instances in D_u as silver data to D_l. We repeat until the performance on the validation set improves and re-train a new model on our newly-originated training set at the end.

We set $k = 200$ and we manage to add $1k$ instances to the original dataset (after four iterations the performance has stopped improving).

As Table 4.4 shows, our model does not benefit from additional data through active learning. Active learning may prioritize samples that, being similar to those who positively affects the performance, do not provide new or informative linguistic variations, thus failing to significantly expand the model's understanding of sexist language. Still, using data from the same domain (internal augmentation) results in better performance than using out-domain data (external augmentation). Since data augmentation, either task-related, or labeled via active learning, do not yield the performance, we do not consider such strategies when addressing the other two tasks.

4.2.3.2 Task B

For Task B, we experiment with the same models employed for Task A. Table 4.7 shows the results. The top-performing model on the dev set is confirmed to be roberta-hate, therefore we use it to predict the test set of the shared task. However, in the test set, the performance drops by 0.06 points.

4.2.3.3 Task C

Starting from the best model for Task B, `roberta-hate`, we employ it for Task C as well. We develop two training strategies, once from scratch and once in a cascaded setting, following the broader category assigned by the Task B model. In the first approach (*all_categories*), the model has access to all eleven categories, whereas in the second approach (*subcategories*), we use four classifiers, one for each class predicted by the model for Task B. Therefore, if an instance has been predicted as '1. Threats' for Task B, the Task C model may only choose among the subcategories of that same class, namely '1.1 Threats of Harm' and '1.2 Incitement of Harm'. Table 4.8 shows the results. The model does not benefit from the pre-categorization of Task B, due to the noisy input, resulting in a 0.02 performance drop. To confirm that the errors are propagated by the imperfect Task B model, we perform an additional set of experiments on the test set with a hypothetical perfect classifier for Task B (by using gold labels) instead of relying on the predictions of the previous model. Using the same settings, the performance significantly increases by 0.25, thus confirming that the noisy input was swaying the model and that using separate classifiers for each subclass leads to increased accuracy for the predictions.

Given the low scores obtained and the significant amount of subcategories for Task C, we attempt to approach the task using generative models to leverage their high ability of contextual understanding. We employ prompt techniques to generate the predictions, by including examples and labels extracted from the main dataset (few-shot setting). Additionally, we attempt to explore the effect of different prompting scenarios on the performance of generative models, following the same settings previously used with `roberta-hate`. We also experiment with including either one or two examples per category for the *subcategories* setting to understand whether providing more in-context data improves model performance. Appendix A shows the prompt structure.

The temperature value for the GPT-Neo models is set empirically. In our case, being restricted among 11 classes meant the model should not be too creative, but setting it too low might confuse it. A range of $[0.5, 1]$ is commonly adopted for generative tasks, but we need to adapt it for classification tasks, which are more restrictive. We have first experimented with a value of 0.1 on a small set of prompts, which did not produce meaningful predictions. By increasing it to 0.3 we managed to obtain sensible outputs, we thus set the temperature to 0.3 for all of our experiments. Table 4.8 summarizes the

model	**strategy**	**dev**	**test**
GPT-Neo1.3	all categories	0.048	0.093
GPT-Neo1.3	subcategories	0.120	-
GPT-Neo2.7	all categories	0.025	-
GPT-Neo2.7	subcategories	0.120	-
roberta-hate	all categories	0.332	0.315
roberta-hate	subcategories	0.316	-
GPT-Neo1.3	subcategories+	-	0.180
roberta-hate	subcategories+	-	0.580

Table 4.8: Macro F_1 for Task C. The submitted system to the shared task (roberta-hate) differs from the top-performing system by 0.24 points. Rows marked with (+) start from a perfect Task B classifier, which is in line with the top-performing system.

results obtained using the different prompt settings. The only results we show for the *subcategories* setting are obtained using the prompt containing two examples, because it emerged that by providing a single example the model would often generate additional random categories, such as *'1.3'*, despite including the complete list of categories within the prompt itself. While the results on both *subcategories* settings might look promising, the results are actually misleading because the output prediction is always the last category provided in the examples, mimicking a repetitive pattern rather than actually generating a meaningful prediction. Only the *all_categories* prompt appears to be able to generate actual predictions and we experiment with it on the test set as well. The results are on par with those from random predictions, suggesting that such a fine-grained classification is difficult to predict using in-context learning with generative models. A better exploration of fine-grained classification of misogyny with more recent LLMs will be discussed in Chapter 6.

4.2.3.4 Error Analysis

I conduct an error analysis for Task A to understand patterns of misclassification.[3] I manually review all misclassified instances and extract frequently occurring words. In false positive cases in Task A, I observe that identity

[3]Task B and Task C are not included due to the high variability of the results.

terms (e.g., *women*) and sexually-explicit terms (e.g.,*rape* and *fuck*) appear without the intent to cause harm, such as in reports of sexist acts or the casual use of slurs. Since such terms are frequently used in sexist instances, the model struggles to differentiate between their harmful and neutral uses, leading to misclassification. This suggests that intentionality must be considered when discriminating between actually sexist and reported sexist acts, as stressed by Chiril et al. (2020b). The obfuscation of identity terms led to a decrease in false positive rates, at the expenses of a lower recall, resulting in more false negatives. This is expected, as for instance the sentence 'I hate women' would be likely identified by a standard model, but the sentence 'I hate [THEM]' would likely not be classified as sexist with our approach.

4.2.4 Summary

I addressed the task of a granular identification of sexism on English social media posts, in the context of the EDOS shared task. Our team experimented with hate-tuned Transformer-based models for Task A, B and C, and generative models for Task B and C. For Task A, we adopted model-agnostic strategies such as the obfuscation of identity terms and data augmentation, with and without active learning. For all three tasks, our best model is always a vanilla `roberta-hate`. For Task A, the model does not benefit from additional data, neither task-related annotated data, nor silver data produced via active learning. The obfuscation of identity terms does not negatively impact the performance, but does not help either, although we manage to decrease the false positive rates, at the expense of a lower recall. For Task C, deep learning models, be it encoder-based or decoder-based, do not have enough samples or internalized knowledge to learn from.

4.3 Cross-lingual Misogyny Detection

This section addresses the challenge of identifying misogyny in multiple languages, namely English, Italian and Spanish. It also represents a first step towards investigating the specificity of misogyny with respect to language and culture.

To address this novel research question, two contrasting hypotheses are tested:

H1 More data boosts the model performance, even if it is in a different language; therefore, the prediction of misogyny benefits from adding training material in diverse languages.

H2 misogyny is language-specific and therefore a monolingual model performs better, even if it is trained on smaller data.

Experiments are based on the use of: (a) data in each of the languages in isolation; or (b) data in various languages in conjunction, through the training of a single multilingual model.[4] I exploit monolingual transformers (BERT (Devlin et al., 2019)) for three languages — English, Italian, and Spanish — and one multilingual transformer (Multilingual BERT (Devlin et al., 2019)). I perform a thorough exploration combining different settings, which include training monolingual transformers with monolingual data, multilingual transformers with monolingual data, and multilingual transformers with multilingual data. Research presented in this section is adapted from Muti and Barrón-Cedeño (2022).

4.3.1 Data

I consider misogyny datasets in three languages, released under two editions of the AMI shared task: AMI at IberEval 2018 (Anzovino et al., 2018) and AMI at EVALITA 2018 (Fersini et al., 2018a). Details can be found in Section 3.1.1. I address the binary problem of whether a tweet is misogynous or not. Table 3.1 shows statistics for the three corpora.

I stick to the evaluation metric of the AMI shared task: the F_1 measure.

4.3.2 Models Description

The models to identify misogynous tweets are built on different variations of BERT (Devlin et al., 2019). In the monolingual settings, I use bert-base-uncased for English (Devlin et al., 2019), bert-base-spanish-wwm-uncased for Spanish (Cañete et al., 2020) and AlBERTo for Italian (Polignano et al., 2019). For the multilingual settings, I use multilingual BERT (mBERT)

[4] These settings avoid resorting to machine translation because the jargon used to convey hateful messages tends to produce faulty target texts, causing the classifiers to struggle (Casula and Tonelli, 2020; Pamungkas and Patti, 2019).

train	**en**	**es**	**it**	**all**
BERT en	**0.71**	–	–	–
BERT es	–	**0.85**	–	–
BERT it	–	–	**0.87**	–
mBERT en	0.65	0.14	0.17	–
mBERT es	0.62	0.81	0.50	–
mBERT it	0.47	0.63	0.87	–
mBERT en-es	0.67	0.83	–	0.75
mBERT en–it	0.66	–	0.86	0.77
mBERT es–it	–	0.80	0.86	0.84
mBERT en–es–it	0.68	0.82	0.86	0.78
best-AMI	0.70	0.81	0.84	–

Table 4.9: F_1 performance for the different language combinations. Best AMI shared task models shown at the bottom for comparison.

(Devlin et al., 2019). I also apply mBERT in monolingual settings, to observe its behaviour in zero-shot classification across languages. A description of the models can be found in Section 3.2.

The output layer is a soft-max with two units. I use the categorical cross-entropy loss function and the AdamW optimizer with a learning rate of 1-8 (Loshchilov and Hutter, 2017), batch size of 16 and 4 training epochs.

4.3.3 Experiments and Results

The objective is to assess whether and to what extent considering training material in diverse languages benefits in the prediction of misogyny in multiple languages. Experiments aim to test hypotheses H1 and H2, heading toward investigating the way in which misogyny is expressed in different languages. Even if the impact of shared vocabulary in multilingual settings remains unclear (Liu et al., 2020), I explore the feasibility of using multilingual embeddings to produce zero-shot classifications across languages —training on L_1 to predict on L_2— and as a data augmentation technique —training on L_1+L_2 to predict on L_1.

Ten models are trained considering all combinations of data in English (en), Spanish (es) and Italian (it): (*i*) one BERT model per language, (*ii*) one mBERT model per language, (*iii*) one mBERT model per each language pair, and (*iv*) one mBERT model with all three languages. Table 4.9 shows the results when predicting on data in each language and all together. The scores under columns en, es and it are comparable, whereas those under all are not, because the testing sets are different.

The monolingual BERT models consistently perform the best, improving over the best AMI approaches. There is a performance drop when monolingual models are trained on top of mBERT, with the model trained on English achieving the poorest performance: as low as F_1=0.14 and 0.17 when tested on Spanish and Italian and six points lower on English than the monolingual BERT alternative. The results suggest that this transfer learning approach is not suitable for languages which are relatively far from each other, e.g., a Romance and a Germanic one. Considering a second language during training improves the predictions of the mBERT models (*i*) on English in all three cases, (*ii*) on Spanish with pair en–es, but (*iii*) not on Italian. Indeed, combining English and Spanish produces better results for both languages than when combining either with Italian. Considering all three languages results in mixed effects. It has the best mBERT performance on English, but runs short by one point with respect to the pairwise combinations on the other two languages. The best performance on all three languages together is obtained when neglecting the training data in English: F_1=0.84.

These results confirm H1 only partially. On the one hand, monolingual models built on top of a monolingual BERT performs the best. On the other hand, considering multilingual training data with a multilingual BERT improves over considering monolingual data alone.

I performed an additional experiment to verify that the performance shifts are not caused by the increase in the volume of training data, rather than the inclusion of another language. I trained a bilingual English–Italian model considering only 2,000 instances per language (conforming to the volume of the monolingual datasets). The performance on the English test set drops from F_1=0.65 to 0.54,, while on Italian drops from 0.87 to 0.85.

These results play in favor of H2: with the same amount of training data, the models do not benefit from data in other languages. Although these experiments are not enough to prove it, the results hint that misogyny is language-specific, meaning that it is conveyed through language-specific expressions which vary across languages and cultures.

it	FN	FP	en	FN	FP	es	FN	FP
bel	1	17	hysterical	27	20	puta	16	25
tette	0	8	woman	16	33	polla	3	13
culo	0	12	women	12	35	cállate	0	6
culona	3	12	fuck	9	27	acoso	7	5
porca	6	0	pussy	5	23	callate	2	4
figa	0	8	rape	3	27	madre	3	3
cazzo	3	4	bitch	4	22	mujer	6	5

Table 4.10: The most common words (sorted by inverse frequency) with the number of false positives (FP) and negatives (FN) in which they occur in the monolingual settings.

4.3.4 Analogies and Discrepancies of Mono- and Multilingual Models

A manual error analysis is conducted to assess how and which kind of errors are transferred from the mono- to the multilingual setting. This analysis has the objective to answer two questions. Question Q1 allows observing the behavior of the multilingual model with respect to the monolingual ones. Question Q2 helps to identify the words that are most likely responsible for the misclassification in the three languages.

Q1 Which instances are classified differently by the monolingual and the multilingual model?

The number of false positives and false negatives behave similarly in all languages. I analyze the instances that the monolingual model (BERT en) classified correctly and the multilingual one (mBERT en-es-it) got wrong. The number of found instances is 122, with 51 false negative (FN) and 71 false positive (FP). Among the FN, the five most common lexical words are *hysterical*, *woman*, *skank*, *women* and *ass*. Among the FP, the words *rape* and *women* are very present, followed by *fucking*, *fuck* and *shut*. False negatives instances are more lexically diverse.

Another point of analysis is to observe the intersection of misogynous tweets between the two models. The mono and multilingual model judged 543 and 541 tweets as misogynous. The intersection is of 438 instances, with 307 being correctly identified. Therefore, the majority of misogynist instances are detected by both models. This hints that there is no big difference between

it	FN	FP	en	FN	FP	es	FN	FP
culo	2	16	hysterical	28	20	puta	24	25
bel	2	20	woman	19	34	polla	2	25
figa	2	11	women	28	31	cállate	6	5
cazzo	0	7	fuck	3	37	callate	1	8
troia	7	3	rape	4	39	madre	4	3
tette	3	3	fucking	4	29	acoso	5	9
culona	1	12	bitch	3	26	escoria	1	7

Table 4.11: The most common words (sorted by inverse frequency) with the number of false positives and negatives in which they occur in the multilingual settings.

the models.

Q2 Which words are most present in instances misclassified by both mono and multilingual models?

The analysis begins by manually observing instances misclassified by both models, which are 70 FNs and 131 FPs. Table 4.10 shows the most frequent words in misclassified tweets in the monolingual and Table 4.11 in the multilingual settings. Table 4.12 shows the translations of the Spanish and Italian words. No significant differences are observed across datasets of the same language, but there are big differences in how misogyny is expressed. In Italian, most words are related to the physical appearance of a woman, linked to sexual objectification. Italian language shows more linguistic creativity. Whereas English contains more insults, Spanish is more aggressive. This contributes to prove that misogyny tends to take different forms across languages.

For English, the most frequent words are the same in both settings: *hysterical*, *woman*, *women*, *fuck* and *rape*. The fact that *woman* and *women* lie in the second and third positions might indicate an unintended identity-term bias (Fersini et al., 2020b), for which the model learnt that *woman* occurs in misogynistic contexts. In both cases, the words *hysterical*, *rape* and *kitchen* (linked to women's stereotyped role) have a strong co-occurrence with the terms *women, woman*, therefore we can assume that these words trigger an error. The word *rape* is common in highly offensive contexts, making it a decisive feature for misogyny; it is frequently present in false positives.

es	en	it	en
acoso	harassment	bel	beautiful
callate	shut up	cazzo	dick
cállate	shut up	culo	ass
escoria	scum	culona	big ass
madre	mother	figa	pussy
mujer	woman	porca*	slut
polla	dick	tette	boobs
puta	whore	troia	whore

*in most of the cases it refers to the expression *porca puttana* ≈ holy shit.

Table 4.12: Translation of the most common words in both Spanish and Italian into English.

1	La ragazza che lavora nel negozio dove vado a fare sempre shopping mi ha detto che ho un bel culo :3333 *The girl working at the place where I always do shopping told me I have a nice ass.*
2	He said she said di Ashley Tisdale fa uscire il puttanone che è in me *He said she said by Ashley Tisdale brings out the bitch in me.*
3	ciao kikka buon pm quanto 6 figa e sexy [...] *hi kikka good evening you are so hot and sexy*
4	figa stai zitta che sono a casa da sola *oh don't tell me I'm home alone*

Table 4.13: Instances of tweets misclassified by both the monolingual and multilingual models (original Italian tweets followed by English translations).

For Spanish, words *puta*, *polla* and *cállate* are common for both settings. I focus the rest of the analysis on Italian, since it shows the biggest discrepancies. Table 4.13 shows examples. In both cases, *bel* always co-occur with *culo*. In FPs, it is commonly used by women to comment on themselves in a positive way, as in example 1. The same happens with the word *tette*, where in FP instances women usually complain about their breast size. These words tend to occur in offensive contexts and therefore are inclined to be classified as misogynist. Another interesting phenomenon that triggers FPs is the presence of slur reappropriation, i.e. women reclaiming certain negative terms (Felmlee et al., 2020), as in example 2 of Table 4.13. Another word that triggers FPs is *figa*, as it is typically used in hypersexualised contexts (example 3) but also in neutral way as a filler word in northern Italy (example 4).

4.3.5 Summary

In this section, I explored the contribution of adding multilingual training material in the automatic identification of misogynist tweets in three languages: English, Spanish and Italian. The models trained on monolingual data achieve state-of-the-art performance. The inclusion of data in one or two other languages impacts the performance negatively when compared to BERT models, but positively when compared to mBERT models. Multilingual models can be used as data augmentation technique — train on L_1+L_2 to predict on L_1, but they are not suitable for zero-shot classification across languages — train on L_1 to predict on L_2, hinting that misogyny might be language-specific. In Italian, most words are related to the physical appearance of a woman, linked to sexual objectification. Moreover, Italian language shows more linguistic creativity. Whereas English features more insults and slurs, Spanish uses a more aggressive language to convey misogyny.

4.4 Meme-ingful Harm: Misogyny in Digital Humor

In this section, I tackle two multimodal tasks aimed at detecting and categorizing misogyny in memes.

Task A A binary classification task that determines whether a given meme is misogynistic.

Task B A multi-label classification task that identifies the specific type(s) of misogyny expressed in a meme, selecting from four potentially overlapping categories: shaming, stereotype, violence, and objectification.

Memes are relatable acts of communication made of visual and textual artifacts, where often an image is superimposed with text with a humorous purpose (MacDonald and Wiens, 2022). To be fully understood, memes require context and real-world knowledge. They are often satirical, implying humour and sarcasm in a subtle way (Sharma and Pulabaigari, 2020). These factors cause the identification of phenomena in them —such as expressions of misogyny— difficult.

Humor does not always come as harmless fun and that is the case with misogynous memes. Such memes contribute to the establishment of a rape culture (Ridgeway, 2014), where violence and sexual harassment are tolerated, belittled, normalized, excused, and transformed into jokes. Therefore, developing automatic approaches to tackle misogyny has both technological and social value.

The task is more challenging than when dealing with text alone because, in general, both the textual and the visual channels play an indivisible role in conveying the desired message.[5]

The experiments carried out in this section aim at understanding if and to what extent a multimodal model outperforms two unimodal ones that address the problem separately. Since meme classification is a challenging task due to its multimodal nature, I shed some light on which component should weigh more in the decision process —text or image— by observing the impact of both modalities in the predictions. In addition, linguistic and visual elements that are potentially responsible for the misclassification are identified and discussed.

This chapter provides an account of the participation of our team Unibo in the MAMI shared task (see Section 3.1.4) and it is adapted from Muti et al. (2022b)).

4.4.1 Data

The MAMI dataset is described in detail in Table 3.4 in Chapter 3. However, I report a summary for readability. The MAMI dataset contains 10,000 memes, equally divided between misogynous and non-misogynous. Each misogynous meme is labeled in four non-mutually exlusive categories. Stereotyping is the most represented class, with $3.2k$ instances overall, followed by objectification ($2.2k$) and shaming ($1.2k$); violence is the least represented, with less than $1k$.

[5]This is different from other multimodal scenarios, such as visual question answering or image captioning, where one of the two modalities tends to be the dominant one (Zhu, 2020).

4.4.2 Models Description

The approach adopted to participate in this shared task is based on the multimodal bi-transformer model (MMBT) (Kiela et al., 2020a). MMBT fuses image and text embeddings in an early fashion. MMBT jointly fine-tunes unimodally pre-trained text and image encoders by projecting image embeddings to the text token space. Figure 4.3 represents the model architecture. MMBT combines two segments: segment 0 corresponds to the text, whereas segment 1 corresponds to the picture. They are fed together to use attention over both modalities at the same time. Each token is indexed according to its position from 0 to the maximum text length, which is set to 80. Each image representation is indexed from 0 to 640.

The original MMBT combines BERT (Devlin et al., 2019) and ResNet (He et al., 2016). Our approach considers other models. Textual embeddings are produced with *bert-base-uncased-hatexplain* (Mathew et al., 2020), a version of BERT trained on identifying hate speech. For the image embedding, CLIP was selected due to its superior performance across a wide range of multimodal tasks, including OCR, action recognition in videos, geo-localization, and various types of fine-grained object classification (Radford et al., 2021). CLIP is pre-trained on the task of predicting which caption should be tied together with a given image. In this way, it learns state-of-the-art image representations from scratch, enabling zero-shot transfer of the model to downstream tasks.

The two embeddings are fused through MMBT. In Task A, the sigmoid activation function is applied to the output layer, with a threshold set at 0.5 to distinguish between misogynous and non-misogynous instances. For Task B, a binary relevance approach (Zhang et al., 2017) is used, combining four separate binary classification models. Each classifier output also employs a sigmoid function. This method was chosen after observing improved performance when treating classes independently, as compared to a multi-class model that predicted misogynous, misogynous-aggressive, or none (Muti and Barrón-Cedeño, 2020). This approach enables the prediction of multiple, mutually non-exclusive classes.

A filtering heuristic is applied to refine the multi-label decisions in Task B: all four labels are set to non misogynous if an instance was not classified as misogynous by the Task A model.

Pre-processing Since CLIP requires square images, following Neskorozhenyi (2021) all memes were resized to a 288×288 pixel version. The

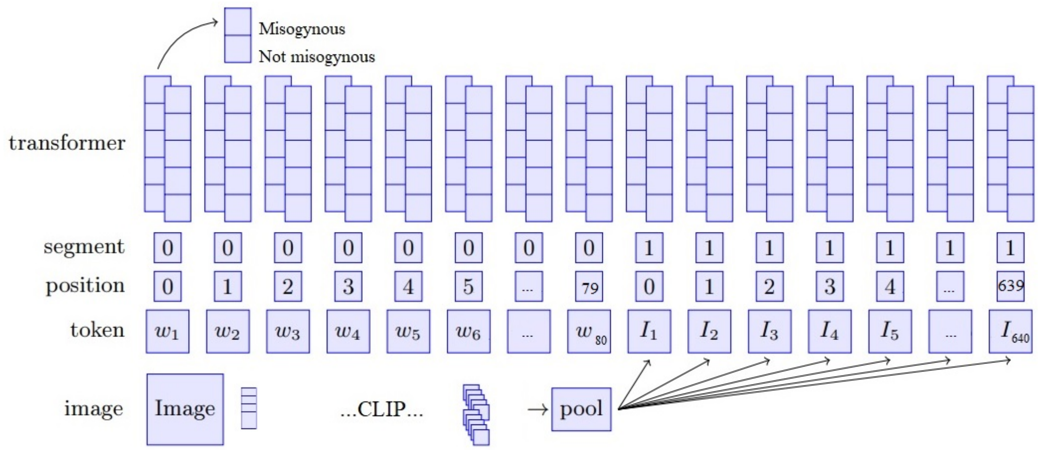


Figure 4.3: Representation of the MMBT model architecture combining CLIP and BERT; adapted from Kiela et al. (2020a).

memes come in different sizes and orientations, hence they had to be rescaled until the largest side reached 288 pixels respecting the aspect ratio. Then, to fill the empty space in the square, white pixels were added.

Then, the resized images are sliced into three equal parts: horizontally if the image orientation is landscape and vertically if it is portrait, to obtain both global and local image features. Four vectors are extracted for each image: a vector for each part encoding spatial information and one for the whole image. These operations are done by using the Pillow library (Clark, 2015).

No preprocessing is applied to the text, other than applying the BertTokenizer (Devlin et al., 2019).

4.4.3 Experimental Setup

The training set is shuffled to take 10% of the data for development, preserving the class distribution.

Three alternative models are trained to identify the best possible configuration. **Uni$_{\mathbf{txt}}$** is a BERT-based unimodal system that considers the text alone. **Uni$_{\mathbf{img}}$** is a CLIP-based unimodal system that considers the image alone. **Multi** is a multimodal system, fusing BERT and CLIP embeddings through MMBT.[6]

Hyperparameters For Multi, different numbers of learning epochs are tested, in range [3, 6]. The best validation performance is obtained after 5 epochs in Task A in the development set. The training is carried out over 5 epochs in Task B. For Uni_{txt} training is done over 4 epochs, while for Uni_{img} over 5 epochs. In all cases the model is saved only when an increase in the performance was obtained. Since the goal is also to assess how CLIP performs in making zero-shot predictions in this task, I experiment with CLIP without fine-tuning on the training set, to determine whether it could effectively detect misogyny without prior annotation. I test batch sizes of 16 and 32, with the former consistently yielding better results. The model is trained using a learning rate of 2e-4, the MADGRAD optimizer (Defazio and Jelassi, 2021), and a binary cross-entropy loss function.

The results reported in Section 4.4.4 are obtained with a model trained during 6 epochs for Task A and 5 epochs for Task B with a batch size of 16.

Evaluation metrics We stick to the official MAMI evaluation metrics: macro-averaged F_1-measure for the binary Task A and weighted-averaged F_1-measure for the multi-label Task B.

4.4.4 Results

In this section I present the results obtained for both Task A and Task B, providing insights into the effectiveness of the proposed approach.

4.4.4.1 Task A - Binary Classification

Table 4.14 shows the results for Task A, highlighting the performance of different models in identifying misogyny in memes. As expected, the multimodal approach Multi yields the best results, achieving the highest F_1 score of 0.727 after six training epochs. This underscores the importance of leveraging both textual and visual information for accurate classification. Considering the textual information alone runs short, with the best-performing Uni_{txt} model

[6] A variation of $\mathbf{Uni_{txt}}$ for Task A is explored, incorporating augmented training data from the tweets corpus of AMI at Evalita 2018 Fersini et al. (2018a). Since no improvement is observed in the model, the results are neglected.

model	**variation**	**macro F_1**
Multi_5	after 5 epochs	0.703
Multi_6	after 6 epochs	**0.727**
Uni_{txt}	bert-base-uncased	0.656
Uni_{txt}	bert-Hatexplain	0.569
Uni_{img}	CLIP	0.703
top		
Samsung		0.834

Table 4.14: Official macro-averaged F_1-measures for our submissions to Task A. The top submission is also reported.

reaching an F_1 score of 0.656 when trained on a *generic* BERT model. As expected, the zero-shot Uni_{img} model performs the worst. However, a proper fine-tuning of the Uni_{img} model turns into the second-best performance with F_1=0.703. The improvement of the Uni_{img} over the Uni_{txt} model by five points suggests that the visual information is captured better than the textual one. The reason might be that the text is too short and out of context to be captured effectively by BERT.

4.4.4.2 Task B - Multi-label Classification

Table 4.15 shows the results for Task B. In this case, one single Multi model is trained during 5 epochs. The difference between the two configurations lies in the filtering of the multi-label classification. With filtering I refer to the rule-based heuristic process where the model's predictions for Task B are constrained based on the outcome of Task A. The most successful multimodal model gets F_1=0.710, when filtered based on Task A's Multi_5 model. Filtering on the basis of Task A's Multi_6 model causes a performance drop of twelve points. This discrepancy can be attributed to differences in the number of misogynous instances predicted by the two models: Multi_5 identifies 678 instances as misogynous, whereas Multi_6 predicts 653. The latter's more conservative approach results in more predictions being blacked out, causing potentially correct decisions by the multi-label model to be disregarded, ultimately reducing overall performance. The text-alone approach, filtered with the corresponding Task A model, runs short by five points.

Table 4.16 zooms into the performance of Task B Multi model for each of the four classes. The model struggles the most when trying to spot stereotyping and shaming. This reflects the nature of misogyny. Stereotyping and shaming tend to be less explicit, and hence harder to spot —even for human beings. On the contrary, violence, which is the most explicit, is more likely to be identified. Stereotyping is the class that has been over-predicted the most (cf. Table 4.18).

model	**filtered on**	**weighted F_1**
Multi	Multi_5	**0.710**
Multi	Multi_6	0.588
Uni_{txt}	Uni_{txt} bert-base-uncased	0.660
top		
Samsung		0.731

Table 4.15: Official weighted F_1-measures for our three submissions to Task B. Column *masked with* specifies the model from Task A used to mask the output labels.

4.4.5 Qualitative Analysis

This section presents a qualitative analysis of the results to further examine the strengths and weaknesses of the proposed approach.

4.4.5.1 Analysis on Task A

To address the question of which component for detecting misogyny in multimodal settings is more important, we look at the distribution of the kind of errors made by the different models, as well as the overlapping instances among the four categories. Table 4.17 shows the relative frequencies. The number of false negatives is much lower than that of false positives across all models. Considering a practical application, false negatives could have a greater impact as they are the misogynous instances that could not be detected, and could therefore lead to harm. On the other hand, blocking instances that were not misogynous but were classified as such could be considered censorship.

	prec	**recall**	$\mathbf{F}_1$
Shaming	0.52	0.46	0.49
Stereotype	0.54	0.58	0.56
Objectification	0.69	0.66	0.67
Violence	0.73	0.48	0.58

Table 4.16: Per-class performance on the positive class for model Multi

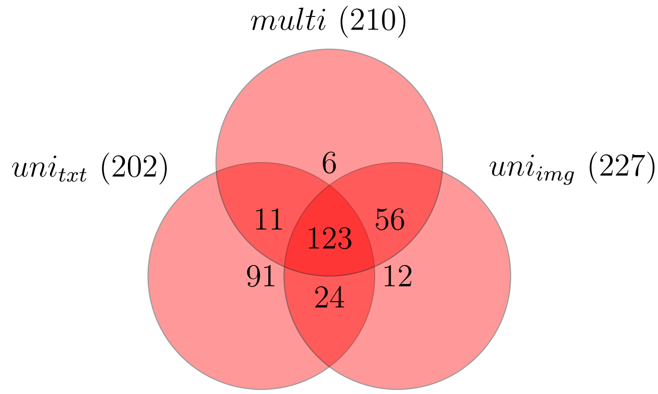
	$\mathbf{Uni_{txt}}$	$\mathbf{Uni_{img}}$	**Multi**
false positives	0.20	0.23	0.21
false negatives	0.14	0.06	0.06
true positives	0.36	0.44	0.44
true negatives	0.30	0.27	0.29

Table 4.17: Error analysis across all models for Task A showing relative frequencies.

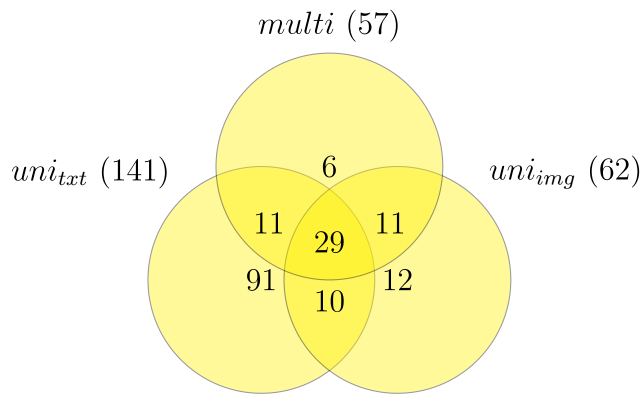
Table 4.17 shows the prediction analysis of the best runs for each modality. Looking at each model individually, $\mathrm{Uni_{txt}}$ has less false positives but more false negatives than the other two models. $\mathrm{Uni_{img}}$ has the highest number of false positives, and the same number of false negatives as the Multi model. This means that the textual model performs worse than the others in capturing misogyny, while the visual one tends to overpredict misogyny more than the other two models.

Figure 4.4 shows the intersections and differences in both false positives and false negatives by the three models. There are more false positive than false negative instances across all models, as observed in Table 4.17. Indeed, the number of common false positives by all models is almost 4 times as high as the number of common false negative values. This indicates that the models tend towards over-predicting misogyny. Taking into account the differences among the sets, $\mathrm{Uni_{txt}}$ accounts for the fewest false positive instances (Figure 4.4a), while it accounts for the most false negative instances (Figure 4.4b). Therefore, in this specific multimodal task, where we can be more lenient with false positives than false negatives, a textual model does not seem to be an optimal alternative.

Since the model does not allow for a great interpretability of the results, we manually inspect misclassified instances and select interesting cases for



(a) False positives.



(b) False negatives.

Figure 4.4: Venn diagrams representing the false positive and false negative errors by the three top Multi, Uni_{txt} and Uni_{img} models during the testing stage.

which we explore the potential reasons behind the errors. As Figure 4.4 shows, 132 instances are misclassified by all three models: 123 are false positives and 29 are false negatives. We observe the following trends after looking at the false negatives:

1. The level of misogyny is low or subjective, as the meme is not directly referred to women (e.g., Figure 4.5) or misogyny is expressed in a subtle way (e.g., Figure 4.6 implies the stereotype that women are complicated);

Figure 4.5: Instance 15846.

Figure 4.6: Instance 16132.

2. Real-world knowledge is required to understand the meme (Figure 4.7 can be better understood if we know Sarah Jessica Parker and the Twisted Sister band).

3. The stance of the text with respect to the image is relevant in order to convey the general meaning (see Figure 4.8);

Figure 4.7: Instance 17028.

Figure 4.8: Instance 16132.

Among the false positives, memes mostly contain:

1. Compliments, which are often associated to objectification (e.g. Fig. 4.9).

2. Images or phrases that often occur in misogynous contents (e.g., women in underwear, kitchen-related terms).

Figure 4.9: An example of false positive (instance 15094).

3. Identity terms (e.g., *wife, women, girls*), that tend to co-occur with misogynous contents in the training set.

Figure 4.10: An example of meme properly labeled by text models only (instance 15802).

We also perform an analysis on memes that have been correctly classified by only one model. Among the instances correctly classified only by the textual model, 11 are true positives and 56 are true negatives. The true positive cases predominantly rely on a strong textual component to convey misogyny, while the image either plays no significant role or serves merely to add an ironic tone to the text.(Fig. 4.10).

Among instances correctly classified only by the visual model, both true positives and true negatives are 11. Most true positives features an explicit visual component. For example, beaten women and texts justifying aggression or glorifying violence. Among instances correctly classified only by the multimodal model, 10 are true positive and 24 true negative. By observing the true positive instances, contrary to what is expected, misogyny is not always conveyed by the interaction of text and image, as in most of the cases the text is actually dominant.

4.4.5.2 Analysis on Task B

We perform a manual inspection focusing on the errors in predicting stereotyping. We observe a relatively large amount of compliments towards women, which tend to confuse the classifier. In particular, false negatives are often caused by the presence of *benevolent sexism* (Glick and Fiske, 1997), which shows a subjectively positive attitude towards women that conceals inferiority compared to men, and it is often disguised as a compliment. Figure 4.11 shows an example.

Figure 4.11: An example of benevolent sexism, a phenomenon that tends to confuse the classifier.

Another analysis focuses on the label overlapping, to determine whether the model effectively captures the intersection of different classes. To evaluate this, we compare the model's predictions to the gold labels in Table 4.18. The size of the intersection between stereotype and objectification is in the same order for gold and predictions: 152 vs 118. The intersection between cases of shaming and violence is practically null, which is well reflected in the model (2 vs 0). Less cases of both shaming and stereotyping than expected are identified (32 vs 20). The same applies to the combinations stereotyping–violence (40 vs 31) and objectification–violence (38 vs 23). The pair shaming–objectification tends to be overpredicted (25 vs 40).

4.4.6 Summary

In this section I presented the approach developed to participate in the Multimedia Automatic Misogyny Identification shared task. Our team addressed two problems: spotting whether a meme is misogynous and, if it is, what kind of misogyny it expresses. Unimodal models (text only and image only) are compared with a multimodal model based on Multimodal bi-Transformers. The image-only model performs better than the text-only one, suggesting that the visual information might be easier to capture than the textual one. The multimodal approach performs the best in both tasks. The errors come from more false positives than false negatives.

From the error analysis emerged that stereotyping and shaming are the most misclassified categories. This shows that more attention is needed on subtle and implicit forms of misogyny and sexism.

4.5 Final Remarks

This chapter delivered significant insights into misogyny detection across different languages and platforms. The major findings are summarized below.

- Although the fine-tuning of transformers-based models show promising results in the detection of misogyny in Italian and English social media posts, challenges remain. For what concerns Italian tweets, models are challenged especially by the polysemic words that carry both benign and pejorative meanings. For what concerns English, models tested on both Twitter, Reddit and Gab posts struggle with the distinction between offensive and casual slurs. Moreover, hate-tuned transformer models shows superiority over standard transformers. However, unintended biases due to identity terms persist, indicating the need for fine-tuning approaches that mitigate over-association between certain terms and misogynistic content.

	train	test	preds
Shaming	1,274	146	130
Stereotype	2,810	350	379
Objectification	2,202	348	334
Violence	953	153	102

Shaming	Stereotyping	Objectification	Violence	train	test	preds
■				400	24	32
	■			1,247	32	152
		■		992	37	96
			■	250	19	21
■	■			286	32	20
■		■		161	25	40
■			■	11	2	0
	■	■		412	152	118
	■		■	302	40	31
		■	■	116	38	23
■	■	■		301	45	32
■	■		■	55	3	1
■		■	■	12	5	0
	■	■	■	162	36	20
■	■	■	■	45	10	5
		Total		4,752*	500	591

* 248 of the misogynous memes lack type annotation.

Table 4.18: Number of instances per class for the multi-label Task B (top). Class distribution (bottom). Column **preds** shows the predictions of our best submitted model (Multi).

- After having explored the feasibility of using transfer learning across languages to detect misogyny, results show that multilingual models demonstrate variable effectiveness across languages, hardly reaching adequate performance, highlighting the importance of culturally contextual training data. Therefore, a zero-shot cross-lingual approach has not proved to be a solution for solving the lack of models and labeled corpora in non-English languages for misogyny detection. These findings are in line with Nozza (2021), despite using a different dataset to perform similar experiments. Certain misogynistic expressions are culturally embedded and may lose meaning when used across languages, limiting the effectiveness of multilingual models trained on diverse linguistic data.

- Both textual and multimodal models tend to over-predict misogyny in instances involving identity terms or images depicting them, such as *wife* or *girls*. Both models struggle with detecting benevolent sexism—statements that seem positive on the surface but imply an underlying gender bias. For example, compliments about a woman's appearance can be misclassified as misogynistic because they may implicitly objectify or stereotype women. Finally, memes require a deeper understanding of cultural context or commonsense knowledge to be understood, often lacking in models. For instance, a meme that uses humor based on cultural knowledge (such as a well-known female character stereotype) might be misclassified if the model cannot access this background information. However, this also happens in textual data, in the so-called *implicit misogyny*. This gap underscores a shared challenge in both models: handling implicit, culturally specific misogyny.

The next chapter addresses the issue of disambiguating polysemic words that carry both benign and pejorative meanings in order to improve misogyny detection. The focus is on Italian language, since the current chapter proved that this problem affects more Italian than English data, given the highly metaphorical usage of offensive words in Italian.

Chapter 5

What Do You Mean? Disambiguating Pejorative Language

In Chapter 2 I introduced the concept of pejorative language. In this chapter, I explore the challenges of understanding the meaning of *pejorative epithets* in Italian, offering a framework for disambiguating polysemic words that potentially cause harm. First, I introduce the concept of pejorative language and pejorative epithets. Then I present three research questions that this chapter proposes to address, followed by the experimental setup of the new proposed framework and finally the results, showing how the disambiguation of pejorative epithets positively affects the detection of misogyny.

This 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. This 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, the pejorative lexicon produced 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.

This chapter is adapted from Muti et al. (2024b).

5.1 Pejorative Language

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 one. 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), which refers to the process by which a marginalized group reclaims a derogatory term or slur that has historically been used to oppress or demean them. Pejorative words are relevant in misogyny detection since certain neutral words are used to address women in an offensive way, targeting either their physical aspect or their intellect. I refer to such terms as **pejorative epithets**. Some examples in Italian are *balena* (whale/fat woman) and *gallina* (hen/stupid). State-of-the-art models struggle to correctly classify misogyny when sentences contain such terms (Fersini et al., 2020b; Kirk et al., 2023). This is also confirmed by research carried out in Chapter 4: 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). The ambiguity of such terms is especially challenging for AI models, which struggle to grasp the connotation of such terms due to their non-standard and context-dependent usage. As a result, these models may miss or misinterpret the misogynistic undertones embedded within. For this reason, I introduce pejorative epithets disambiguation as a preliminary step to detect misogyny. With this new proposed framework, the aim is to answer three research questions:

RQ1 Which epithets are typically used in misogynistic language in Italian?

RQ2 Can the disambiguation of such words decrease the error rate in the task of 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**, I compile a list of pejorative words used online to address women. I use such words to retrieve tweets, and build PejorativITy, a novel

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

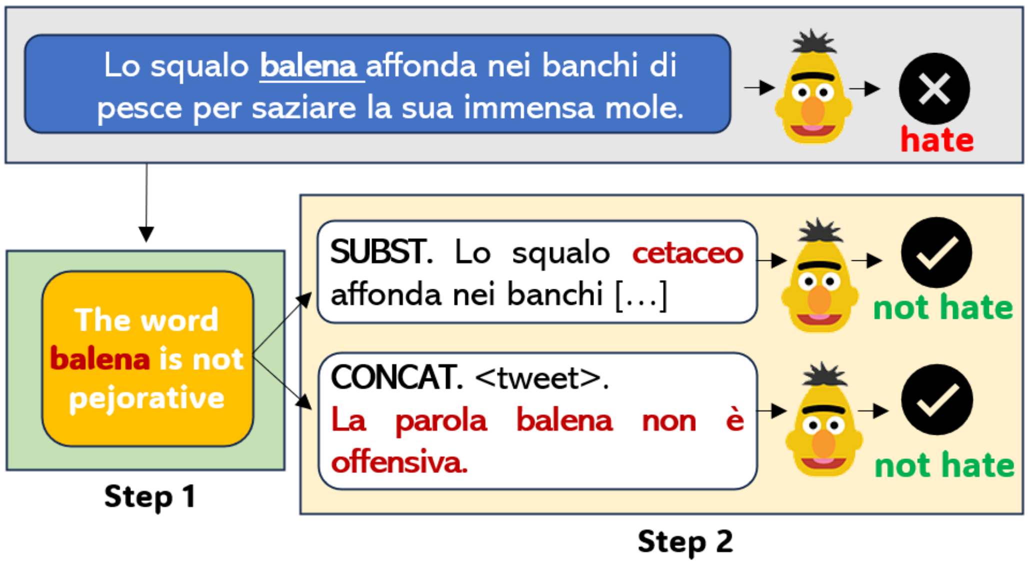


Figure 5.1: Our pipeline. Step 1: a model identifies the connotation of possibly pejorative epithets. Step 2: the identified connotation is used to enrich (CONCAT) and substitute (SUBST) part of the textual input for misogyny detection.

corpus of Italian tweets, annotated at the word level for pejorativity, and at the sentence level for misogyny.

To address **RQ2**, I 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. The output of model$_{pej}$ is used to inform model$_{mis}$ of whether the target word is pejorative within that context or not. Figure 5.1 represents the pipeline proposed.

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

5.2 Misogynistic Pejorative Epithets

To provide an overview of which misogynous epithets are commonly used on Twitter in Italian (**RQ1**), I 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. The lexicon is collected by selecting words from three distinct sources.

1. I 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. I retrieve the keywords used in the two Italian corpora for the Automatic Misogyny Identification (AMI) shared task (Fersini et al., 2018a, 2020b).

3. I consult the 'List of Dirty Naughty Obscene Bad Words'.[2]

Only polysemic words whose primary meaning is neutral and that are frequently used on Twitter with both pejorative and neutral connotations are kept. To ensure the quality of the vocabulary, I verified that such words are used with both connotations by manually searching them on Twitter. Due to their exclusive neutral or negative connotation on Twitter, the following terms are discarded: *barile, banco, botte, barbona, facile, gatta morta, passeggiatrice, porca, principessa, privilegiata, psicopatica, scrofa, somara, travestita.*[3]

Table 5.1 shows the lexicon of 24 words.

For each word, the English translation of its literal and pejorative meaning, along with their anchors in Italian are reported. Anchor words refer to the unambiguous words used to define polysemic words. These words are called anchors because their meaning is univocal and does not change depending on 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*

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

[3]EN: barrel (fat), desk (stupid), barrel (fat), homeless (tramp), easy, cocktease, prostitute, sow (a woman who enjoys sex), princess (benevolent), priviledged, psycopath, sow, donkey (stupid), tranny (derogatory).

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

(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]

5.3 PejorativITy: Corpus Compilation

Twarc[5] is employed to retrieve tweets from December 2022 to February 2023 containing words in our lexicon. For each word in the lexicon, 50 tweets were selected, resulting in 1,200 tweets. During the selection process, I performed a preliminary annotation by keeping a balance between pejorative and neutral use of lexicon words. However, an equal distribution for each word could not

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

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

be guaranteed, since in the final annotation phase some labels were modified.

5.3.1 Data Annotation

Six annotators with a background in linguistics, gender studies, cognitive sciences, and NLP were recruited to label the corpus for pejorative word disambiguation and misogyny detection. [6]

I 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, I follow a descriptive annotation paradigm (Rottger et al., 2022), which encourages annotator subjectivity by not providing guidelines. Annotators are split into two groups and are assigned 50 tweets each for labeling. Each group is composed of two women and one man, with ages ranging between 27 and 39 years.

The inter-annotator agreement (IAA) is measured With Krippendorff's alpha (Krippendorff, 2011). Table 5.2 shows the results.

Paradigm	**Group**	**Pejorativity**	**Misogyny**
Descriptive	1	0.48	0.50
Descriptive	2	0.33	0.50
Prescriptive	together	0.53	0.53

Table 5.2: Krippendorff's alpha IAA among the three annotators for each group.

The IAA of the first group is *moderate* for both pejorativity and misogyny, whereas the IAA of the second group is *fair* for pejorativity and *moderate* for misogyny. In terms of gender differences, men tend to consider sexual objectifying compliments as non-pejorative. Based on annotators' feedback, during a discussion phase 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 6.6.

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

[6] The annotation was carried out during my internship at the company Expert.AI.

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 6.6, 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 6.6, 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 6.6: the word *balena* is pejorative, but it is used in a positive way by means of negation.

All the edge cases discussed above were then provided to the annotators in the form of guidelines, which were used to devise a second pilot annotation, getting closer to a prescriptive annotation paradigm (Rottger et al., 2022). We select the top 50 tweets that caused more debate during the first study annotation phase. The IAA computed on the new annotation among all six annotators is 0.53 (*moderate*) for both tasks, denoting an improvement over the first pilot study.

After the pilot studies, by following the decisions taken during the discussion phase on the edge cases, I annotated the whole corpus.

Table 5.3 shows the statistics. The Pearson correlation between misogyny and pejorativity labels is 0.70. The tweets for which misogyny and pejorativity are not aligned are mainly reported speech or men-targeted offensive language. It is worth noting that some sentences are implicitly misogynous since they are in line with the definition of misogyny presented in Chapter 2, however

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 5.3: Statistics of the PejorativITy corpus. The same tweets are annotated for misogyny and pejorativity, for a total of 1,200 instances.

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

they do not express any hate towards women. For instance, the sentence *"che bella bambola ciao tesoro"*[7] 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).

5.4 Experiments

In this section I introduce the experiments that prove the efficacy of the newly proposed framework: pejorative epithets disambiguation for misogyny detec-

[7]EN: 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 5.5: Statistics of the AMI 2018 and 2020 corpora (Fersini et al., 2018a, 2020b).

tion. The experimental setup foresees two computational tasks: word sense disambiguation for pejorative epithets disambiguation and text classification for misogyny detection.

5.4.1 Pejorative Epithets Disambiguation

Word Sense Disambiguation (WSD) tasks are typically treated as a multi-label classification problem for resolving the semantic interpretation of target words in context (Bevilacqua et al., 2021). For instance, the target word *mouse* can be associated to three distinct concepts: a small rodent, an hand-operated device of a PC, and a timid or quiet person (Barba et al., 2021). For this purpose, WSD is treated as a binary classification problem, since only two possible meanings (one neutral and one negatively connotated) are assigned to a target word that has the potential of being misogynistic. To understand the impact of disambiguating pejorative words for misogyny detection (**RQ2**), I experiment with AlBERTo (Polignano et al., 2019) (see Chapter 2 for details). In particular, AlBERTo is fine-tuned on two downstream tasks: pejorative word disambiguation and misogyny detection.

For pejorative word disambiguation, I evaluate AlBERTo only on the PejorativITy corpus. For misogyny detection, I also consider two other benchmark datasets for Italian: AMI-2018 (Fersini et al., 2018a) and AMI-2020 (Fersini et al., 2020b), which are described in Chapter 3. I report the data statistics for AMI in Table 5.5 for convenience.

I 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, I 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, I use the model fine-tuned on PejorativITy corpus to determine the connotation of ambiguous words.

5.4.2 The PejorativITy Framework

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

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

2. Enrich input tweets in all data partitions by injecting knowledge about the pejorativity of the lexicon words according to model$_{pej}$. Two different approaches are tested to modify the input data: *i)* **concatenate** the information about the pejorativity of w at the end of the tweet or *ii)* **substitute** the ambiguous w with its corresponding anchor word.

3. Train model$_{mis}$ to detect misogyny with the enriched input tweets.

The pipeline is meant to process any tweet. However, as a first step, I check whether it contains at least one $w \in W$. Therefore, as a first step I perform string matching after lemmatization.[8]

AlBERTo is fine-tuned for 4 epochs with batch size 16. I 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.

5.5 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.

[8]Lemmatization is an NLP technique that reduces words to their base or root form, the lemma. Unlike stemming, which may simply cut off suffixes, lemmatization uses vocabulary and morphological analysis to return the dictionary form of a word (Manning and Schütze, 1999).

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

Table 5.6 shows the classification performance for misogyny detection on the PejorativITy test partition. I compare the fine-tuned AlBERTo model (*baseline*) against the alternatives that leverage pejorative word disambiguation. I evaluate the concatenation and substitution approaches using model_{pej} (*w/ predictions*) and annotators' labels (*w/ gold*) since the PejorativITy corpus contains annotations for pejorative word disambiguation. The evaluation of the proposed approaches with gold labels defines an upper bound to the pipeline presented. A notable improvement can be observed 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.

In order to corroborate the initial hypothesis on reducing the false positive rate for misogyny detection, I compute the number of false positives in the three datasets, before and after the inclusion of pejorative information both by concatenation and substitution. Table 5.7 shows the results. The decrease of false positives is clear in AMI-2020 and in the PejorativITy test set. In AMI-2018, no decrease is observed. One of the reasons for this low impact is that AMI-2018 contains pejorative epithets only in 34 instances out of 1000 (compared to 192 in AMI-2020), therefore I did not expect the proposed approach to have a huge impact on that dataset.

Table 5.8 shows the classification performance for misogyny detection on AMI-2018 and AMI-2020. To assess the impact of the proposed pipeline on these corpora, the performance of the models both on the test instances that

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

Table 5.7: False positive rates comparison. In the PejorativITy the total number of instance is 96, while in AMI 2018 and 2020 is 1,000.

contain words in our lexicon (**epithets**) and on the whole corpora is shown. In particular, I perform fuzzy string matching 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. I 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. I provide an example in Section 5.5.1. 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**).

5.5.1 Qualitative Error Analysis

A manual error analysis is carried out by observing misclassified tweets in both AMI-2020 epithets and PejorativITy corpora for the task of misogyny detection. Misclassified tweets are compared in the three settings: baseline, concatenation, and substitution.

Concatenation Most of the misclassifications occur when reported misogyny is concerned. The model struggles to recognize 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:

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

Lei è acida perché non ha figli penso che darebbe fastidio a qualsiasi donna. Che schifo.[9]

In this example, the author of the tweet criticizes 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. 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. I highlight this aspect to show that model_*mis* does not necessarily learn to classify misogyny according to model_*pej*'s outcome.

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

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

Ma la balena con gli shorts cortissimi invece è vittima del patriarkato e può vestirsi come vuole?[10]

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.

5.5.2 Contextualized Word Embeddings

To investigate the semantic knowledge of the AlBERTo pre-trained language model (Polignano et al., 2019) about pejorative epithets and to evaluate how fine-tuning affects its knowledge (**RQ3**), I extract and analyze contextualized word embeddings of our lexicon words.

To extract these embeddings, I perform fuzzy string matching on input tweets to retrieve the tokenized text span corresponding to lexicon words. I 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. The AlBERTo model employs the sentencepiece tokenizer (Kudo and Richardson, 2018), a common tokenization process for transformer models. For instance, the lexicon word *balena* is tokenized to the [`balen`, `##a`] text span. I then use these text spans to aggregate the corresponding word embeddings. I define the word embedding of a lexicon word as the average of the AlBERTo token embeddings in the retrieved text span. Considering *balena*, I define its word embedding by extracting the embeddings of `balen` and `##a` and computing their average.

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

Table 5.9 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, the fine-tuned AlBERTo model shows relevant discrepancies

[10]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 5.9: 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.

when considering lexicon word embeddings in pejorative and in 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 the fine-tuned model acquired knowledge about *acida* being a synonym of stronza when used in a pejorative way. Similar trends are observed in all other selected lexicon words.

The average similarity of lexicon words with respect to 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. This means that pejorative words get closer to their pejorative anchors and further from their neutral anchors in the fine-tuned setting. Therefore, pejorative anchors have higher similarity with lexicon words in pejorative samples. Likewise, neutral anchors have higher similarities with lexicon words in neutral samples. The average similarity in pejorative samples using pretrained embedding is 0.33, while after fine-tuning is 0.11. This means that lexicon words get further from their pejorative anchors in neutral samples. To prevent a biased analysis where the fine-tuned model learns a representation of anchors similar to that of lexicon words, I compute each anchor frequency in the corpus. The computed frequency for each anchor is close to zero, thus excluding potential biases in the analysis.

5.6 Can LLMs do better?

Encoder-based models like BERT still struggle to correctly classify misogyny when sentences contain pejorative epithets, although an improvement is observed. This could be overcome by decoder-based LLMs, as they could rely on their implicit knowledge to grasp the meaning of such terms. Therefore, I replicate the pipeline by testing `Llama-3-8B-Instruct`, one of the most popular open-source LLM, in both tasks: epithets disamgiguation and misogyny detection. By asking models to determine whether a term is being used in a pejorative or non-pejorative sense, I 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. I expect models to struggle with this challenge, particularly in sentences with non-standard or regional varieties of Italian, which occur in the PejorativITy corpus.

5.6.1 Experimental Setup

Consider the sentence *Quella balena coi jeans non si può guardare*, translated as *Can't look at that whale with jeans*. First, the model is asked to identify whether the meaning of the target word (*balena (whale)* in our example) is pejorative or not. The model should rely on its internal knowledge accumulated during pre-training to understand whether the term *balena* 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. In the second step, the model is first informed with the decision of whether the

Model	Partition	Macro F1
AlBERTo	test	0.82
llama-3-8b-Instruct	test	0.58
llama-3-8b-Instruct	full corpus	0.64

Table 5.10: Results on pejorative epithets disambiguation.

target word is pejorative or not, and then asked to classify the input text as misogynous or not.

Prompts For pejorativity disambiguation, I design the following zero-shot prompt: `La parola {word} è utilizzata in modo peggiorativo nel testo in input? Rispondi solo Sì o NO.` [11]

For misogyny detection, I design two zero-shot prompts that ask whether a tweet is misogynous or not in order to evaluate two settings:

A The two tasks are independent, i.e. without having the decision for pejorativity to inform the decision for misogyny.

B The model is informed of the decision on pejorativity of the target word by injecting the information about pejorativity in the prompt for misogyny detection.

Prompt A: [12]

`Il testo in input è misogino? Rispondi solo Sì o NO.`

Prompt B: [13]

`Il testo in input è misogino? Rispondi solo Sì o NO.`
`Considera l'informazione sulla parola {word}:`
`{response from the first prompt}.`

In order to make the results comparable with AlBERTo, I also show the results for llama3 evaluated only on the test set partition (see Table 5.3).

[11] Is the word {word} used pejoratively in the input text? Answer only with YES or NO.

[12] Is the input text misogynous? Answer only YES or NO.

[13] Is the input text misogynous? Answer only YES or NO. Consider the information about the word {word}: {response from the pejoartivity prompt}

5.6.2 Results

Table 5.10 shows the results for pejorative word disambiguation. Llama3-8b-Instruct shows a score of 0.64 when evaluated on the full corpus, showing room for improvement in the prompt design. When evaluated on the test set partition, it falls behind of 0.24 points compared to AlBERTo. However, 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.* I excluded such cases from the evaluation. In the full corpus, of the 174 excluded instances, 123 were pejorative and 51 were not pejorative according to the gold standard. In the test partition, out of 13 excluded instances, 7 were pejorative and 6 non pejorative. Although the fine-tuned version of AlBERTo achieves a higher performance, llama aids in explainability by deliberately adding explanations of why it considers the target word to be pejorative or not.

Table 5.11 shows the performance regarding misogyny detection at sentence level.

Setting	**Model**	**Partition**	**Macro F1**
A	llama-3-8b-Instruct	full corpus	0.69
B	llama-3-8b-Instruct	full corpus	0.60
A	llama-3-8b-Instruct	test	0.55
B	llama-3-8b-Instruct	test	0.58
baseline	AlBERTo	test	0.68
concatenation	AlBERTo	test	0.75
substitution	AlBERTo	test	0.77

Table 5.11: Results on misogyny detection.

In Setting A, where the model is not informed of the output for pejorativity, Llama performs better compared to the previous task. However, the fact that all answers were provided in the misogyny task (unlike in the previous task) plays a role and does not necessarily imply that misogyny detection is an easier task than pejorativity disambiguation for llama. Setting A corresponds to the baseline setting when using AlBERTo. Llama3 still shows a lower performance, falling behind by 0.13 points.

In Setting B, the model is informed of the decision on pejorativity of the target word by injecting the information about pejorativity in the prompt.

The concatenation and substitution approaches with AlBERTo are more effective than injecting the information in the prompt with llama3. Moreover, when considering the full corpus, Llama3 does not benefit from the injection of knowledge about pejorative words, with a drop of 0.09 points; in the test partition there is a gain of 0.03 points. This could be due to the noisy response from the WSD task, including the refusal to answer, and wrong explanations of why the target word is used pejoratively or not. I explore this aspect in the next section.

5.6.3 Investigating Pejorative Word Comprehension with LLMs

I prompt open-source generative large language models (LLMs) to investigate their knowledge of pejorative words in Italian. Specifically, I aim to find out whether they are able to clarify the meaning of our lexicon words (**RQ3**).

I follow a zero-shot prompting approach without any training data for the downstream task (Liu et al., 2023a). I design the following prompt asking for the meaning of a pejorative {*word*} in a given {*sentence*}. I prompt the LLMs for each test instance in the 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 appropriately.
Instruction:
What is the meaning of the word {word} in this sentence?
{sentence}[/INST]
Response:

I use three open-source LLMs for the analysis: LlaMa2 (Touvron et al., 2023b), Mistral (Jiang et al., 2023), and Camoscio[14], which are presented in Chapter 2.[15]

[14]https://github.com/teelinsan/camoscio
[15]This research was carried out prior to the release of Llama3.

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

For all models, I select the *7b* model version with 8-bit weights due to hardware constraints. I 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. I manually inspect all the responses generated by the three LLMs.

Mistral. 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 5.12. 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 5.12. While this example causes trouble to human annotators for the lack of context, Mistral identifies the pejorative connotation of the lexicon word *cagna*. Although performing very well, Mistral struggles with reported speech, too. Consider Example 3 in Table 5.12. 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* in the sentence *Sta cozza non vedeva l'ora*, translated as *That ugly girl couldn't wait.* 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 5.12, Mistral defines *cozza* as *"surprisingly beautiful"*.

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. When they answer, Camoscio tends to answer with words from the same semantic field of the target word. For instance, when asked about the meaning of *balena* in pejorative instances it answers that *The meaning of the word balena is sea/dolphin.* However, there are few cases in which it recognizes the offensiveness of the target words, by saying *The meaning of the word x is offensive*, without actually specifying the meaning. Llama2 tends to answer with English translations, such as *Il significato della parola bambola in questa frase è doll.*[16] Llama2 struggles to identify the correct meaning of even common pejorative epithets like *cagna*, even when they co-occur with other slurs. Consider the sentence *Chi sei tu per dare cagna e zoccola alle persone?*[17] Llama responds that the meaning of the word *cagna* is *dog*. This raises questions regarding the multicultural elements embedded within multilingual models.

The analysis conducted suggests that off-the-shelf instruction-tuned LLMs have ample room for improvement concerning pejorative word disambiguation.

[16] EN: The meaning of the word doll (in italian bambola) is doll.
[17] EN: who are you to call them bitches and sluts?.

5.7 Final Remarks

In this chapter, I introduced pejorative word disambiguation as a preliminary step for misogyny detection with the aim of reducing the error rate of classification models on polysemic words that can serve as pejorative epithets.

For this purpose, I 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: PejorativITy. I validate the pipeline proposed by evaluating AlBERTo (Polignano et al., 2019) on the new corpus and on two benchmark corpora in Italian: AMI-2018 (Fersini et al., 2018a) and AMI-2020 (Fersini et al., 2020b). I explore two approaches to inject pejorativity information: concatenation and substitution. Results show that the disambiguation of potentially pejorative words leads to notable classification improvements in all testing scenarios. Furthermore, I analyze 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, I perform experiments with LLMs which show that they struggle more than encoder-based models to identify both pejorativity and misogyny. The qualitative analysis carried out with several off-the-shelf instruction-tuned LLMs shows that there is ample room for improvement and that although being multilingual, such models lack the cultural components to understand misogyny when is expressed through metaphors, which are culturally sensitive (Korre et al., 2024a). The use of metaphors and pejorative language can manifest itself in the form of implicit hate, which will be explored in the next chapter.

Chapter 6

LLMs for Implicit Misogyny Understanding

As discussed in Chapter 2, misogyny, as any other form of hate speech, can be either expressed explicitly or in a more veiled manner. Typically, the reference datasets for the identification of misogyny are produced by selecting misogynous keywords (Fersini et al., 2018c; Basile et al., 2019b). As a result, most implicit forms of misogyny are kept out and therefore are under-detected by classifiers, which often over-rely on identity terms and negative keywords (Hartvigsen et al., 2022; Yin and Zubiaga, 2022). Implicit instances of misogyny are harder to understand for humans too —potentially giving rise to disagreements in the annotation phase (Hartvigsen et al., 2022; Yin and Zubiaga, 2022). Early distinctions on the degree to which hateful content is expressed considered only a binary set (explicit *vs.* implicit), where explicitness is defined as unambiguous in its potential of being hateful (Waseem et al., 2017). In Section 2.1.3, I presented the features of implicit language in hate speech, such as sarcasm, figurative language and inferences, among others. These subtleties present a significant challenge for automatic detection because they rely on underlying assumptions that are not explicitly stated. As illustrated in Figure 6.1, the `bert-hateXplain` model[1] correctly marks as hateful the explicit message (💀), but it fails with the implicit one (💀). To correctly spot the implicit message, the system would have to identify at least the implied assumptions that "*women aren't as capable as men.*" and "*women should be told what to do*", underlying a centrality of gender distinction

[1] `https://huggingface.co/tum-nlp/bert-hateXplain`

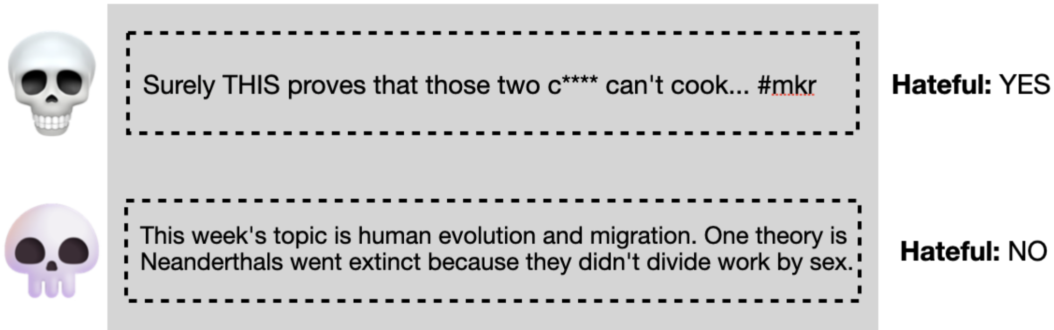


Figure 6.1: Results from `bert-hateXplain` model for explicit (💀) *vs* implicit (💀) misogynous messages.

and traditional women's role.[2]

In this chapter, I investigate the abilities of large language models (LLMs) to correctly identify implicit hateful messages expressing misogyny in both English and Italian. In order to do so, LLMs need to understand, and also reconstruct, the implied assumptions. The contribution of this chapter is threefold:

1. the introduction of the newly collected corpus for implicit misogyny in Italian Facebook comments: ImplicIT-Mis.

2. a novel framework for misogyny detection and explanation which makes use of Toulmin argumentation theory (Toulmin, 1958) in order to reconstruct implied assumptions.

3. a novel taxonomy of social dynamics occurring in implicit misogynistic statement, on which I evaluate an extensive sets of LLMs to detect such categories and spot the text span that triggers their decision, in order to control for random answers.

This chapter is adapted from Muti et al. (2024a), a work conducted while doing an internship at the University of Groningen supervised by Tommaso Caselli.

[2] Example and explanations extracted from Sap et al. (2020).

6.1 Understanding Implied Assumptions

Implied assumptions instantiate statements that are presupposed by implicit hate speech content. This can be seen as the elicitation of implicit knowledge, corresponding to content semantically implied by the original message (Srikanth and Li, 2021; Zaninello and Magnini, 2023). This process can be carried out either through free-text generation, by generating implied assumptions, or multi-label classification, by selecting the implied social dynamics entailed in the message. Although limited, previous work on the generation of implied meanings —usually in the form of explanations— has moved away from template-based methods (Zhang et al., 2014) to the application of encoder-decoder or decoder-only models (Saha et al., 2021; Xing et al., 2022; Cai et al., 2022). Generating explanations for implicit content poses multiple challenges concerning the quality of the generated texts, whose primary goal is to be reasonable and informative. Some approaches generate explanations by identifying pivotal concepts in texts and linking them through knowledge graphs (Ji et al., 2020). More recently, the underlying concepts are generated by directly querying LLMs (Talmor et al., 2020; Fang and Zhang, 2022; Yang et al., 2023). Hoyle et al. (2023) use LLMs to produce sets of propositions that are inferentially related to the implicit text to be observed, then validate the plausibility of the generated content via human judgments. Fei et al. (2023) elicit common-sense and reasoning ability from LLMs to infer the latent intent of an opinion for the task of sentiment analysis. In this work, I follow the idea of using LLMs to identify the implied assumptions in the implicit messages. I formulate the problem in two ways: (*i*) as an Argumentative Reasoning task by applying Toulmin's Argumentation Theory (Toulmin, 1958); (*ii*) through multiple-choice selection of social dynamics representing implied misogynistic behaviors, along with the identification of text spans referring to each social dynamic.

In order to evaluate the experiments, since data in Italian were lacking, I collect the first corpus for implicit misogynistic comments for Italian: ImplicIT-Mis; for English, I select misogynistic instances from the Social Bias Inference Corpus (SBIC) (Sap et al., 2020) and Implicit Hate Corpus (ElSherief et al., 2021), combining them into a single dataset, which I refer to as SBIC+.

6.2 ImplicIT-Mis: The First Italian Dataset on Implicit Misogyny

ImplicIT-Mis is a new manually collected and curated dataset for implicit misogyny detection in Italian. It consists of 1,120 Facebook comments posted as direct replies to either women-related news articles or posts on public pages of communities known to tolerate misogyny. As an in-domain expert, who has been the target of online misogyny, I personally conducted the manual collection. This is in line with a participatory approach to NLP where the communities primarily harmed by specific forms of content are included in the development of datasets addressing these phenomena (Caselli et al., 2021b; Abercrombie et al., 2023).

For each comment, I keep the source (either the name of the newspaper or the community Facebook page) and its context of occurrence (the news article or the main post). Note that all instances in ImplicIT-Mis are misogynistic, since during the collection phase I carefully selected only comments in which misogyny is conveyed through implied assumptions.

The collection period ran from November 2023 to January 2024. I selected 15 Facebook pages of news outlets covering a wide Italian political spectrum as well as different levels of public outreach (national *vs* local audiences), and 8 community pages. See Table 6.1 for a comprehensive list. ImplicIT-Mis is organized around 104 source posts; 70% of the 1,120 messages are comments to news articles from two national newspapers, *la Repubblica* and *Il Messaggero*. The full source overview is presented below. On average, each comment is 19 tokens long, with the longest having 392 and the shortest only one. An exploration of the top-20 keywords, based on TF-IDF, indicates a lack of slurs or taboo words, confirming the quality of our corpus for implicit misogyny (see Figure 6.2).

Annotations of ImplicIT-Mis have been carried out according to two levels: implied assumptions in the form of free text and implied social dynamics in the form of non-mutually exclusive categories.

6.2.1 Annotations for Implied Assumptions

A subset of 150 comments from ImpliciIT-Mis is enriched with one annotation layer targeting the implied assumptions. The sample is annotated by three

Source	Posts	Source	Posts	Source	Posts
National News		**Online news**		**FB Community**	
La Repubblica	411	Donna Fanpage	37	Caffeina Festival	65
Il Messaggero	378	Fanpage	33	Non sono bello	15
La Stampa	76	Huffington Post	6	ma spaccio	
TgCom24	20	TPI	4	La matita scarlatta	9
Libero	1	Il Post	1	Pastorizia never dies	9
Local news		Leggo	1	Stefano Valdegamberi	6
AnconaToday	20			I love Patriarcato 2	4
BolognaToday	9			La società femminista	4
Corriere Adriatico	2			L'uomo che bestemmiava	3
Palermolive.it	5			ai cavalli	

Table 6.1: List of sources, including newspapers and Facebook pages, with the total amount of extracted instances for the creation of the ImplicIT-Mis dataset.

Italian native speakers who are master students in NLP. Each annotator has worked on 50 different messages. On average, the task took each annotator two hours.

The annotation guidelines for the generation of the implied assumptions are in Appendix B. I evaluated the annotators' implied assumptions against those of an expert (a Master student in gender studies and criminology). I selected a subset of 75 sentences (25 from each annotator) and computed two metrics: BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020). These measures offer insights into how similar the human-written implied assumptions are among each other. BLEU score results in 0.437 and F1-BERTScore in 0.685 by combining all annotations. As the scores indicate, our pool of annotators tends to write the implied assumption adopting different surface forms, but with a similar semantic content, as suggested by the F1-BERTScore. Although implied assumptions have to be inferred, and therefore humans need to interpret the text, they tend to come to the same conclusions. In the final version of the data, all manually written implied assumptions have been retained as valid, meaning that for 150 messages, I have a total of 225 implied assumptions, with 75 instances having a double annotation.

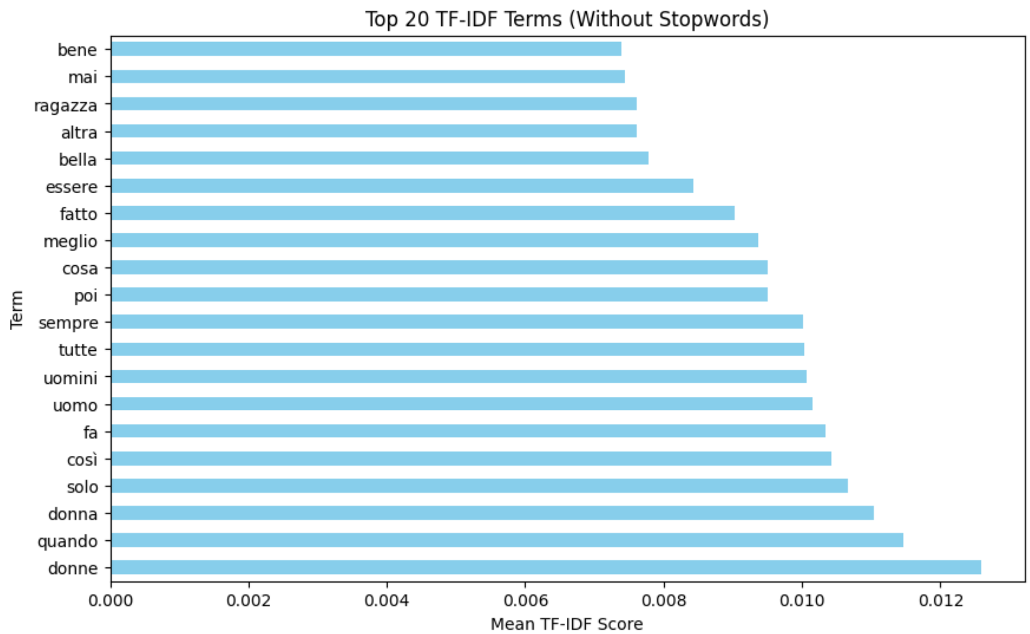


Figure 6.2: TF-IDF for ImplicIT-Mis.

6.2.2 Annotations for Social Dynamics

A subset of 50 instances from the ImpliciIT-Mis corpus has been enriched with an annotation layer that targets the underlying social dynamics, as well as the accompanying text span, for evaluation purposes. The subset has been annotated by three Italian linguists who are familiar with gender-based issues, including me. I followed a descriptive paradigm (Rottger et al., 2022), i.e., not providing guidelines to annotators, for two reasons: the self-explanatory category names did not require additional description, and I wanted to mimic the zero-shot prompt given to the LLMs for the auditing, which will be described in Section 6.5.

6.2.2.1 Social Dynamics

The developed taxonomy of social dynamics is grounded in feminist literature and gender studies (Wrisley, 2023; Ramati-Ziber et al., 2019; Srivastava et al., 2017a; Lopes, 2019; Kellie et al., 2019; Bergh and Brandt, 2023) and focuses on making explicit the underlying assumption(s) of misogyny rather than its linguistic manifestation. This is the main motivation for referring to them as "social dynamics": they describe the manifestation of underlying interactions, attitudes, and behaviors within groups of people (Bannester, 1969). This social dynamics taxonomy can be seen as a categorization of misogynistic implied assumptions. Table 6.2 reports the list of categories.

Social Dynamic Category	**Example**
1. derogatory treatment and belittling of emotions	*she's not depressed she just needs to get more*
2. man-dominated power structure	*women go to the club without their partners to cheat on them*
3. conservative limitations to women's freedom	*she wastes time on internet instead of being a good wife*
4. beauty standards expectations	*real beauty is something else, not this fake plastic*
5. mocking	*"ooh, my life is meaningless if I cannot show my tits"*
6. stereotyping and generalization	*women are always naked on social media*
7. whataboutism	*what about violence vs men?*
8. double standards	*it's unattractive when girls act like ghetto*
9. victim blaming	*she shouldn't have drunk so much*
10. aggressive and violent attitude	*she should be given 2,000 volts*
11. dismissal of feminism or neosexism	*patriarchy doesn't exist*
12. sexual objectification	*fresh meat*
13. centrality of gender distinction	*Born with Dick = Man; Born with Vag = Woman*
14. unfounded assumptions and prejudices	*Daily reminder that women are more racist than men.*

Table 6.2: Overview of the social dynamic categories and examples taken from our corpus.

A detailed description of each category will follow. **1.** Demeaning or diminishing women's feelings or experiences, often by belittling their emotions; similar to 'Misogyny Derogation' in Guest et al. (2021). **2.** Situations where men have control or authority over women's decisions, reflecting a power imbalance (Jane, 2016). **3.** Conservative views which limits women's freedom, including criticism for not conforming to traditional roles or expectations and references to a "natural order", sexual abstinence and "pro-life" values (Siapera, 2019). **4.** Any expectations on beauty standards, including the rejection of self-defined expressions of beauty and appearance (Amundsen, 2019). **5.** Any ridiculing or humiliating expression based on jokes, sarcasm and irony. This category is seen alongside offensive terms used for other categories (Flick, 2020). **6.** Oversimplified beliefs about women that do not consider individual differences; it can be stereotypical with respect to the traditional gender role of women or made-up generalizations (Ging, 2019). **7.** A diversion tactic that shifts focus and derail conversations from issues affecting women by raising counterpoints about other issues, such as domestic violence against men, child custody, divorce and the feminization of education (Ging, 2019). **8.** Behaviors are judged differently based on gender, often detrimental to women. **9.** Usually in the context of sexual assault, when the victim is held responsible. **10.** Any threat or hostile behavior that is posed to women. **11.** Denying the existence of gender inequality, patriarchy. It includes the refusal of using gendered language (including feminine job titles) and personal attacks against feminists. It also includes positions explicitly against or mocking/denying gender equality or gender issues; statements that feminism is misandry or is compared to Nazism, and that men have to fight against their diminished power in society (Siapera, 2019). **12.** The reduction of a person to their physical attributes or sexual appeal, including men and women who perceive sexualized women as lacking certain human qualities such as mental capacity and moral status. This opposes to the view of self-enhancement and sexuality as empowering (Ging, 2019). **13.** Emphasizes binary views of gender identity based solely on biological sex, including what women and men are supposed to do and how should behave based on their biological sex. Disparaging heteronormativity (Fosbraey and Puckey, 2021) can also fit in this category. **14.** Any form of prejudice or assumption that has no evidence.

All the 14 categories have been identified in at least one example. The most frequent categories are "mocking" (15 instances), "unfounded assumptions or prejudices" (13 instances), "conservative view that limits women's freedom" (10), "derogatory treatment or belittling of emotions" (7), "beauty standards

Lang	A1-A2	A1-A3	A2-A3
IT	0.350	0.351	0.327
EN	0.201	0.188	0.120

Table 6.3: pair-wise Cohen's κ scores for the categories.

expectations" (7 instances).

Table 6.3 (first row) shows the pair-wise Cohen's κ scores for the categories. The scores represent minimal agreement, highlighting the complexity and subjectivity of such task. However, given the highly subjectivity of the task and the expertise of the annotators, cases of disagreements have been considered as different perspectives rather than errors, under the lenses of perspectivism (Cabitza et al., 2023) and human label variation (Plank, 2022).

6.2.2.2 Span Annotation

For each category labeled by annotators, I also ask to highlight the text span associated to each category. For instance, in the sentence *It seems to me that the other side of sexual harassment is what's not being talked about women using their bodies to get ahead.*, the category *whataboutism* should be linked to the span *the other side of sexual harassment is what's not being talked about*, while *women using their bodies to get ahead* should be linked to *stereotype and generalization*. The maximum quantity of text spans identified within a single category is three, observed in five instances; 20 instances comprise two text spans, while the rest features only one. In 22 instances, the text span corresponds to the entire text. This outcome is expected, since the misogynistic effect in implicit instances is conveyed throughout the entire text, as illustrated by the example *come tirare un salame nel corridoio*, literally translated as *like throwing a sausage in the corridor*, implying that a woman has had many sexual partners.

Given the perspectivist approach adopted, I aggregate all labels produced by annotators to obtain the human dataset. For the text span, in case of agreement on the category, I retain the longest overlapping span, otherwise each proposed span - associated with a category - is considered valid.

6.2.3 SBIC+

SBIC+ is a dataset of 2,409 messages for implicit misogyny in English obtained by merging together 2,344 messages from SBIC and 65 from the Implicit Hate Corpus (IHC).

The social bias inference corpus (SBIC) (Sap et al., 2020) consists of 150k structured annotations of social media posts for exploring the subtle ways in which language can reflect and perpetuate social biases and stereotypes. It covers over 34k implications about a thousand demographic groups. SBIC is primarily composed of social media posts collected from platforms such as Reddit and Gab, as well as websites known for hosting extreme views, such as Stormfront.

The structured annotation approach implies that different annotation layers are available to annotators according to their answers. The annotation scheme is based on social science literature on pragmatics and politeness. All messages whose annotation for the target group is "women" or "feminists" are retained and labeled as hateful. I further clean the data from instances which had been labeled as targeting women but were actually targeting other categories, like gay males. I also filter out all texts containing explicit identity-related slurs to keep only implicit instances. For each message, I retain all associated "target stereotype" which correspond to the warrant.

The Implicit Hate Corpus (IHC) (ElSherief et al., 2021) contains 6.4k implicitly hateful tweets, annotated for the target (e.g., race, religion, gender). The corpus comprises messages extracted from online hate groups and their followers on Twitter. Tweets were first annotated through crowdsourcing into explicit hate, implicit hate, or not hate. Subsequently, two rounds of expert annotators enriched all implicit messages with categories from a newly developed taxonomy of hate, for the target demographic group, and for the associated implied statement (i.e. the warrant in our framework). I select only tweets whose target demographic group is "women".

Whereas for the implied assumptions I could exploit existing annotations, the annotation for social dynamics is done from scratch. Similarly to the ImplicIT-Mis corpus, three experts annotate a subset of 50 messages for social dynamics and text spans. The low agreement (see Table 6.3 second row) could be due to the different cultural background of the annotators (Italian, Dutch and Jordanian - all proficient in English), which might result in a different perception of misogyny. The top five categories are: "mockery" (12 instances), "unfounded assumptions or prejudice" (11 instances), "centrality of gender distinction” (11 instances), “stereotyping and generalization” (10), and ”sexual objectification” (9).

Annotators identify on average a total of 73 text spans each. The maximum number of spans identified within a single instance is four, occurring in one instance. In four instances, three text spans are identified; in 19 instances, two text spans are identified; and in the remaining instances, only one text span per instance is identified. In 13 instances, the text span is represented by the whole text.

Both datasets are used to address two main research questions: whether LLMs can generate valid implied assumptions (or implicit warrants); and whether they can identify implied social dynamics by categorizing them across 14 possibly-overlapping categories, along with the text span which triggers the decision of the selected social dynamic(s).

6.3 Misogyny Detection as Argumentative Reasoning Understanding

Hate, in the case of implicit misogyny, is expressed by assuming social biases, stereotypes, and prejudices against women. The identification of these assumptions requires access to the reasoning process behind arguments and opinions.

Argumentative Reasoning (AR) offers a solution. AR relies on the notion of an argumentative model or scheme, i.e. a formal representation of arguments into intrinsic components and their underlying relations. It aims at explicating an argument through the identification of its constituent components and relations (Lawrence and Reed, 2019). For instance, Toulmin’s AR model organizes arguments into fundamental elements, such as claim, warrant and reason. AR models have been successfully applied in many NLP tasks, from Argument Mining (Stab and Gurevych, 2017; Habernal and Gurevych, 2017; Lauscher et al., 2018) to warrant and enthymeme reconstruction (Reisert et al., 2015; Boltužić and Šnajder, 2016; Habernal et al., 2018a; Tian et al., 2018; Chakrabarty et al., 2021; Bongard et al., 2022), argumentative scheme inference (Feng and Hirst, 2011), and fallacy recognition (Habernal et al., 2018b; Delobelle et al., 2019; Goffredo et al., 2022; Mancini et al., 2024).

Grounded on previous work on AR in user-generated content (Boltužić and Šnajder, 2016; Becker et al., 2020), I frame implicit misogyny detection

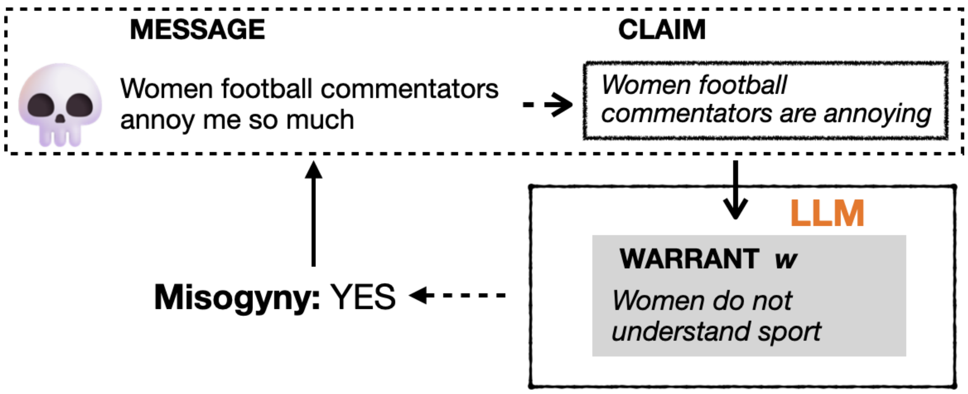


Figure 6.3: Example of a warrant (implicit logical connection) for an implicit misogynous message. Example and warrant are extracted from SBIC (Sap et al., 2020).

as an AR task (Habernal et al., 2018a) based on Toulmin's theory (Toulmin et al., 1979), with the aim of developing more robust detection tools by explicitly describing the underlying reasoning process in these messages. In order to classify a message, the model is first asked to generate a textual response corresponding to Toulmin's warrants. More formally, let c be the claim associated to a given message and $W = \{w_1, \ldots, w_n\}$ be a set of possible warrants, i.e. logical statement(s) that support c. Models must generate an associated w and, based upon it, classify the message as misogynous or not.

Figure 6.3 represents the approach described above. In the instance shown in the figure, the implied generalization that women do not understand sport because it is stereotypically for men is what distinguishes a personal preference from a case of misogyny, which falls into the category of gender-based double standard.

6.4 Can LLMs Generate Valid Implied Assumptions For Implicit Misogyny?

While there have been efforts on evaluating LLMs in argumentative tasks, such as quality assessment (Wachsmuth et al., 2024), component detection (Chen et al., 2023), and argumentative linking (Gorur et al., 2024), the capability of LLMs for implicit argumentative reasoning has yet to be explored. For the first time, LLMs are assessed on implicit misogyny through the lens of Argumentative Reasoning. By doing so, it is possible to evaluate the implicit knowledge of LLMs, such as named entities or events mentioned in texts, which, if not known, it would be impossible to understand the misogynistic nature of such texts.

Data As previously mentioned, two datasets are used to evaluate the proposed framework: the newly collected ImplicIT-Mis corpus for Italian, and SBIC+ for English.

6.4.1 Experimental Setup

Each batch of experiments is framed as in Section 6.3, i.e. the model generates the claim c and related implicit warrants W and, based on these, provides a classification label as misogynous or not (Toulmin setting). I also experiment with a different strategy by prompting the model to directly reconstruct the implied assumptions, without claim, and afterward classify the text (Assumption setting). I address these tasks both in a zero-shot and in a few-shot setting. While implied assumptions are generally broader than warrants, warrants specifically bridge the reasoning gap between claims and evidence. In our prompts, implied assumptions and warrants appear quite similar. Nevertheless, the use of these terminologies may significantly impact the model's behavior due to its sensitivity to prompt phrasing, therefore I experiment with both.

I experiment with two state-of-the-art LLMs, which are introduced in Section 2.2: `Llama3-8B` and `Mistral-7B-v02`.[3] For both, I select their instruction-tuned version. During preliminary experiments with 50 instances, I also tested Italian-specific LLMs, namely `LlaMantino`,[4] `Fauno`,[5] and `Camoscio`.[6] Since all three Italian-specific models are unable to generate valid implied assumptions, they are discarded. I consider the following baselines: (i) fine-

[3]Refer to `https://huggingface.co/meta-llama/Meta-Llama-3-8B` and `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2`

[4]`https://huggingface.co/swap-uniba/LLaMAntino-2-7b-hf-dolly-ITA`

[5]`https://huggingface.co/andreabac3/Open_Fauno-Italian-LLM-7bB`

[6]`https://huggingface.co/teelinsan/camoscio-7b-llama`

tuned encoder-based models, (ii) zero-shot classification with LLMs, and (iii) few-shot classification with LLMs without generating explanations.

6.4.2 Prompt Definition

Among recent prompting techniques, I select **Chain-of-Thought** (CoT) and **Knowledge Augmentation**. CoT has been chosen for its notable success in reasoning tasks (Lyu et al., 2023). On the other hand, Knowledge Augmentation has been observed to reduce hallucinations and enhance contextual depth in model prompts, facilitating the generation of sophisticated outputs beneficial for tasks requiring substantial domain knowledge and nuanced reasoning (Kang et al., 2024). Both techniques align with the goal of generating implicit components of arguments (implicit warrants) and support the construction of encoded warrant blocks. More in detail, CoT sequentially guides the model through a series of reasoning steps before arriving at a final answer or conclusion (Wei et al., 2024). By following this structured approach, CoT prompts allow for the identification of how the model's reasoning process influences its conclusions. This capability is particularly useful for reconstructing warrants that underlie the model's interpretations.

Knowledge-augmented prompting generates knowledge from an LLM and incorporates it as additional input for a task (Liu et al., 2022). In our task, the generated knowledge serves as either the implied assumption or the warrant, that I inject into the prompt to inform the classification.

To the best of our knowledge, these techniques have not been used yet for a computational argumentation task, which makes them worth investigating. The full list of prompts, both for English and Italian, can be found in Appendix C.

6.4.3 Results

I report two blocks of results: the first block focuses on **classification** of the messages. Since both the Italian and the English datasets contain only positive classes, I only report the Recall. The classification task offers an indirect evaluation on the goodness of the AR methods. The second block targets the **generation** of the implied assumptions/warrants. Considering the complexity and the pending issues related to the evaluation of automatically generated text (Chang et al., 2024), I report the results using established automatic metrics (BERTScore, BLEU, and ROUGE-L) as well as a manual validation on a subset of 300 messages (150 per language). The overall evaluation procedure I have devised allows us to assess both the performance of the models in detecting implicit misogyny and the alignment between LLMs and human annotators in generating reasoning-based explanations.

All answers from LLMs have undergone post-processing to evaluate them properly. Two main post-processing heuristics concern the treatment of the refusal to provide an answer (including the refusal to generate the warrants) and the "need of more context". I consider both cases as if the messages were marked as not misogynous. While `Llama3-8B` tends to return refusals to answers, mostly due to the safeguard layer, `Mistral-7B-v02` has a tendency towards indecisive answers requiring more context. `Llama3-8B` always provides an answer when applied to the Italian data. For completeness, Appendix D.1 includes the results considering these cases as correct.

Setting	**Model**	**ImplicIT-Mis**	**SBIC+**
fine-tuning	`bert-hateXplain`	–	0.342
	`ALBERTo`	0.380	–
zero-shot	`Llama3-8B`	0.588	0.609
	`Mistral-7B-v02`	0.050	0.319
few-shot	`Llama3-8B`	**0.738**	**0.719**
	`Mistral-7B-v02`	0.259	0.416
zero-shot Assumption	`Llama3-8B`	0.542	0.448
	`Mistral-7B-v02`	0.050	0.259
few-shot Assumption	`Llama3-8B`	0.480	0.616
	`Mistral-7B-v02`	0.461	<u>0.685</u>
zero-shot Toulmin	`Llama3-8B`	0.557	0.452
	`Mistral-7B-v02`	0.346	0.374
few-shot Toulmin	`Llama3-8B`	<u>0.725</u>	0.594
	`Mistral-7B-v02`	0.556	0.604

Table 6.4: Classification results on ImplicIT and SBIC+. Best results in bold; second best underlined.

Setting	Model	BERTScore		BLEU		Rouge-L	
		EN	**IT**	**EN**	**IT**	**EN**	**IT**
Assumption							
zero-shot	`Llama3`	0.820	-	0.201	-	0.040	-
few-shot	`Llama3`	**0.830**	-	0.744	-	0.085	-
	`Mistralv2`	0.823	**0.601**	0.361	0.240	0.099	0.062
Toulmin							
zero-shot	`Llama3`	0.817	0.570	0.543	0.104	0.046	0.025
	`Mistralv2`	0.812	0.579	0.303	0.077	0.055	0.026
few-shot	`Llama3`	0.817	0.570	**0.871**	0.261	0.060	0.028
	`Mistralv2`	0.813	**0.601**	0.396	**0.313**	0.088	0.048

Table 6.5: Automatic evaluation metrics for the best models generating implied assumptions/warrants (selection based on classification results).

6.4.3.1 Classification Results

Table 6.4 summarizes the results for the classification task. With few exceptions —mostly related to `Mistral-7B-v02`— LLMs generally perform better than fine-tuned models. All few-shot experiments outperform their zero-shot counterpart, and `Llama3-8B` consistently performs better than `Mistral-7B-v02`. The best results are obtained by `Llama3-8B` with few-shot and no generation of either the implied statements or the warrants. However, for Italian, the `Llama3-8B` with the Toulmin warrant in few-shot achieves very competitive results (recall=0.725). For English, the results are affected by the post-processing heuristics. Had I considered as correct the "refusal to answer cases", the best score for English would have resulted in `Llama3-8B` few-shot with implied assumption, with R=0.913 (see Appendix D.1).

In all zero-shot settings, the prompt based on Toulmin's warrant outperforms the prompt based on implied assumptions. In the few-shot settings, in ImplicIT-Mis I observe a dramatic increase when switching from implied assumptions to Toulmin's warrants, with a performance gain of 24 points. On the contrary, in English, the warrant-based prompt falls behind.

6.4.3.2 Implied Assumptions and Warrants Generation

Table 6.5 gives an overview of the evaluation using BERTScore, BLEU, and ROUGE-L for the English and Italian models. Whereas for SBIC+ every message has an associated explanation, for ImplicIT-Mis only 150 messages are enriched with the implied assumption. When `Llama3-8B` is asked to elaborate on the implied assumption in both zero- and few-shot settings, it does not follow the instruction, and only in 87 and 71 instances for Italian and English, respectively, it generates a response. In all the other cases, the model just answers to the final question of whether it is misogynistic; therefore, I exclude them from the evaluation. I also exclude all the results that do not reach a lower-bound recall of 0.3 due to their low quality, as confirmed by manual inspection. All BERTScores in English are in the range of 0.81-0.83, showing high similar content between the human-written texts and the answers generated by the models. Therefore, both the implied assumptions and the warrants are aligned with those written by humans. In Italian, the scores drop to 0.57-0.60. In terms of BLEU scores, the highest scores for English are produced by `Llama3-8B` few-shots with warrants, which shows an alignment with humans in terms of word choices. For Italian, the scores are much lower, probably because of many wrong translations and lack of Italian references, which cause wrong inferences. The low ROUGE-L scores indicate that the surface forms of sentences generated by humans and models are very different. This is confirmed by the manual inspection, from which I observe that the models' responses are more verbose.

6.4.4 Manual Validation

I further validate the generated implied assumptions and warrants by manually exploring a subset of 300 messages, 150 for each language. For ImplicIT-Mis, I use the manually annotated instances, while I randomly extract 150 instances for SBIC+. I focus only on the best models: `Llama3-8B-Instruct` few-shots warrant for ImplicIT-Mis and `Mistral-7B-Instruct-v02` few-shots implied assumptions for SBIC+.

The manual validation assesses the plausibility of warrants. When I say that a warrant is correct, it means that it is valid, although there are others that could possibly be valid as well, as our setting does not assume only one correct response. Overall, 35% of the generated warrants for ImplicIT-Mis are correct and 32% lead to a correct classification of the messages. For SBIC+, the percentage of valid implied assumptions leading to a correct classification is 50%, while correct implied assumptions are 52%. However, in Italian, all the correctly predicted examples were actually predicted for the wrong reasons, while in English, this happened 37% of the time. Therefore, I conclude that a correct explanation does not necessarily lead to a correct classification of misogyny, and this is always the case in the subsample I manually evaluated for Italian. This can be seen as an evidence that the model relies on their internalized knowledge and spurious correlations to address the task and shows no reasoning skills, since the Italian texts, being collected to address this task, require much more reasoning to be understood.

I design a taxonomy to regroup all errors for both models. I identify seven kinds of common errors in warrant and implied assumption generation. Table 6.6 provides some examples. Although all error categories lead to wrong implied assumptions/warrants, I decide to keep a general "wrong inference" as a valid category for all the cases that do not fall under any other category, lack an evident reason or shows wrong reasoning processes.

Sarcasm/Irony. This is a common error in English, due to the relatively high number of jokes in SBIC+. In these cases, the LLMs fail to capture the sarcastic/ironic intended meaning of the message and go for a more literal interpretation.

Metaphorical and Figurative Language. This category indicates a failure to interpret another level of non-literal meaning. I have observed a much more frequent occurrence in Italian —also because many messages use figurative or metaphorical expressions. As observed in Chapter 5, misogyny in Italian is highly metaphorical, especially in dehumanizing women comparing them to animals. In Italian, not identifying metaphors could also be attributable to translation errors since metaphors are cultural-dependent. This highlights the complexity of cross-lingual implicit HS detection, as also pointed out by Korre et al. (2024a), since the translation of a term often does not carry the same implications as in the source language.

Wrong Translations. This is a category of errors that applies only to Italian. It comprises errors due to wrong translations of messages or to a lack of understanding of regional languages and jargon expressions.

Sarcasm/Irony	**→ IT: 2; EN: 26**
Example: *my wife gets angry when i call her pet names apparently sandwich making machine is sexist…*	
Generated: *The text mentions a personal interaction between the speaker and their wife, and does not make a generalization about women*	
Metaphorical and Figurative Language	**→ IT: 14; EN: 1**
Example: *Put a stuffed animal between them and they fight over it*	
Generated: *Men argue easily*	
Wrong Translations	**→ IT: 10; EN: N/A**
Example: *la lagna di mezzanotte* [The midnight complain]	
Generated: *There is a belief in witches and the existence of a "witches' hour". […] The use of the term "witches" is likely a poetic or metaphorical way of describing challenges.*	
Opposite Intention	**→ IT: 12; EN: 9**
Example: *What will the lgbt community achieve together this month? Aids*	
Generated: *The LGBTQ+ community is capable of achieving great things together.*	
Denial of Misogyny	**→ IT: 3; EN: 4**
Example: *doesn't society's acceptance of makeup for women give them a gender imbalance and greater attractiveness in aggregate over men, therefore inequality?*	
Generated: *Women are judged and valued primarily based on their physical appearance.*	
Lack of Reference	**→ IT: 17; EN: 0**
Example: *Pozzi era meglio, non cantava ma… il lavoro era lo stesso.* [Pozzi was better, she didn't sing but… the job was the same]	
Generated: *Pozzi is a better singer than the other one*	
Wrong Inference	**→ IT: 42; EN: 34**
Example: *i can't decide if today is a good or bad day my ex got hit by a bus, but i lost my job as a bus driver*	
Generated: *The text does not provide any assumptions about women or gender. The text is about the speaker's personal experiences and emotions.*	

Table 6.6: Error categories in warrant generation. For each category, I report an input example, the corresponding LLM generation, and the category's distribution in Italian and English evaluation samples.

Opposite Intention. These errors could be considered an instance of LLM hallucinations Maynez et al. (2020). In these cases, the models completely misinterpret the message's content, resulting in generated implied assumptions that tend to contradict the message. These errors occur in both languages, with a slightly higher frequency in Italian.

Denial of Misogyny. This class of errors indicates a lack of connection between the generated implied assumptions and the answer of the model. In other words, the generated text is logical and correct as it clearly identifies the misogynous nature of the message. However, the model classifies the texts as non-misogynous.

Lack of Reference. This kind of error has been observed only in Italian —mostly because very few English data contains direct reference to individuals. The errors in these cases are mostly because the model lacks knowledge about the mentioned individuals (and their associated characteristics) that are used to trigger the misogynous content. In the example I report in Table 6.6, the model does not recognize (Moana) Pozzi as a famous porn actress, thus resulting in a warrant that fails to capture the derogatory nature of the message (i.e. the target being promiscuous). This can also occur as a lack of textual reference as in the example text *Escile* translated as *Show them*. In this case, them is referred to women breast. Mostly of those missing textual references are implied sexual references.

Wrong Inference. This is the largest class of errors in both languages. I observe that wrong inferences are mainly driven by spurious correlations and the activation of implicit knowledge. For instance, given the statement *their partner choices are quite questionable*, LlaMa3 derives the warrant that women are more emotional than men, which is a commonly known stereotype that the model has internalized and has nothing to do with the text itself. Those kinds of conclusions serve as evidence that LLMs lack reasoning skills and emergent abilities, and rely solely on keywords and implicit knowledge.

6.4.5 How Reasoning Impacts Classification

I further examine how reconstructed implied assumptions/warrants relate to classification predictions for misogyny detection.

I inspect 50 instances for each language that are correctly classified by the best baseline (few-shot `Llama3-8B`) and that are subsequently misclassified when the model is asked to reason about them. I observe the tendency to reduce a gender-related problem to the whole category of human beings, minimizing the misogynistic nature of the statement. For instance, in the Italian text *"Oggi sei felice anoressica bugiarda"*, translated as "Today you are happy anorexic liar", the model responds that *"Anorexia is a negative condition. Lying is a negative behavior. No, the text is not misogynist. Anorexia is a condition that affects both men and women."*. Although being true that it is a condition that affects both genders, in this case it targets a woman, which is clear in Italian by the gendered *"anoressica"*, instead of *"anoressico"* which would be used for men. Adding the statement that it affects both genders is detrimental to the classification.

6.5 Can LLMs Identify Social Dynamics in Implicit Misogyny?

In this section I aim at auditing the performance of LLMs in understanding the implied social dynamics in misogynistic statements in social media posts in Italian and English through a non-mutually exclusive multi-label task, rather than free-text generation. To validate the model response, I also prompt LLMs to report the text span that triggered the decision. This is done to limit LLMs' random choices. A side aim of this work is to test reasoning or emergent abilities of these models in the absence of additional fine-tuning or in-context learning. Recent work has proposed distinguishing the acquisition of competencies in LLMs either as *abilities*, i.e. the capacity to solve a task absent in smaller models as an effect of the size of the models themselves, and *techniques*, i.e. the beneficial effect of different prompting methods that are ineffective in smaller models (Lu et al., 2023). The experiments that Lu et al. (2023) have conducted —using zero-shot settings on multiple tasks— show that an ability such as reasoning is an effect of prompting techniques (e.g., instruction-tuning or in-context learning), rather than an emergent ability. In this section, I follow a similar experimental setting, where I investigate the behavior of LLMs in zero-shot settings when it comes to high-level tasks such as explaining the underlying societal assumptions of implicit misogynistic messages via multiple choice. The proposed framework can be considered as a new probing approach to assess the (non-)linguistic knowledge accumulated by LLMs during pre-training.

6.5.1 Data

I target both languages, Italian and English, by using ImplicIT-Mis and SBIC+. For the complete list and explanations of the social dynamics, please refer to Section 6.2.2.

6.5.2 Experimental Setup

For the experiments, I select six decoder-based LLMs following two criteria: they must be open-source and have the same number of parameters. Due to available computing infrastructure, I could experiment with models with a 7B-parameter size, except for Llama3 which has 8B. For each model, I select their instruction-tuned version. I follow a zero-shot prompting approach without further fine-tuning the models or providing in-context learning methods for the downstream task (Liu et al., 2023a). I use the following models: `Llama-2-7b-chat-hf`, `Llama-3-8b-Instruct`, `Mistral-7b-Instruct-v.02`, `falcon-7b-instruct`, `TowerInstruct-7B-v0`, `LLaMAntino-2-chat-7b-hf-UltraChat-ITA`. A detailed description of their main characteristics is presented in Section 2.2.

6.5.3 Prompt Definition

It is a known phenomenon that the specific format of a prompt may result in very different outcomes when applied to LLMs. To control for this, I investigate the behaviors of the models to follow prompt instructions. I experiment with about 100 prompts across the two languages. For the Italian model, I translate the prompts into Italian. Prompt variations mainly involve the use of synonyms, paraphrasing and descriptions of how the final output should be structured. I initially run preliminary experiments with an initial set of 50 instances in English and Italian, respectively, to analyze the output and decide on the final prompt format. For the prompt design, I took inspiration from existing work (Hromei et al., 2023; Hasanain et al., 2024; Lu et al., 2023).

When running LlaMa and Mistral on the Italian messages, I observe a tendency of these models to translate the Italian input or text spans into English. To limit this, I explicitly ask the models not to translate the message in the prompt. The final prompt instructs the models to select one or more of the 14 social dynamics as well as the corresponding text span. The English prompts and the Italian translation are reported in Appendix E.

The models prompted are LLaMa2, LlaMa3, Mistral (v1.0 and 2.0), Falcon, and Tower on both English and Italian data. LLaMAntino is prompted only on Italian data. When running LLaMa, Mistral, Falcon, and Tower on Italian data, the prompt is given in English. For all models, temperature is set to 0.7.

6.5.4 Results

A quantitative and qualitative analysis of the LLMs outputs allows for the analysis of *i*) whether the models are able to complete the task properly with respect to the instructions and structure of the output and *ii*) to what extent the models are able to provide a reasonable output; i.e., the social dynamic category(ies) and the corresponding text span(s).

6.5.4.1 Correctness of the output format

In the prompt(s), LLMs are asked to provide the output in a specific format; i.e., `[social dynamic:‘‘text span’’]`. Here, the extent to which the models completed this task is evaluated. When prompted on the English data, LLaMa2 has only 21% of correct output format. This is due to the model refusal to answer because of its over-safety. The second lowest model is Falcon but the delta with LlaMa2 is quite large (+38% of correctly formatted answers). The other models show good results: Tower has a correct output format in 68% of the cases, Mistral-v2 in 79%, Mistral-v1 in 84% and LlaMa3 in all instances.

6.5.4.2 Inter-model agreement

Before comparing the performance of the models against the manual annotations, I investigate to what extent models are in agreement with each other, both concerning the social dynamic categories and the text spans. For all cases where the instructed output format was not followed, post-processing is applied to avoid over-penalizing the models. Usually this involves deleting special characters, uniforming uppercasing or extracting the text span from a verbose output. In the case of multiple categories and only one text span, the text span is associated to all the categories.

The evaluation is run on the entire dataset for each language. For the **category**, I compute macro-F1 scores in a pairwise setting, assuming one model at the time to be the reference and all the others the predictions. Figure 6.4 shows the results. In general, I observe higher scores in Italian than in English. For Italian, Llama2 and Tower reach the highest agreement, while for English it is reached by Llama2 and Falcon.

I observe a consistent trend across all models, and especially Falcon, to prioritize the categories listed initially in the prompt (e.g., "victim blaming"). Top five categories for English are "victim blaming" (8,801 instances), "sexual objectification" (4,130 instances), "derogatory treatment or belittling of emotions" (3,017 times), "mockery" (2,759) and "aggressive and violent attitude" (2,735). For Italian, I observe a slightly different pattern, with "victim blaming" (3,571 instances), "derogatory treatment and belittling of emotions" (1,801), "sexual objectification" (1,679 instances), "centrality of gender distinction" (1,660 instances) and "expectations with respect to beauty standards" (1,653 instances). The distribution of the categories is similar in both languages, however, in English "mockery" and "aggressive and violent attitude" occur more frequently. Indeed, SBIC contains many jokes characterized by the format "*what is the difference between a woman and X?*" and explicit incitements to violence. In Italian, the categories that do not appear in the top-five for English are "centrality of gender distinction" and "expectations with respect to beauty standards". The rationale is that the Italian data contains messages from Facebook pages that explicitly mock feminism, triggering comments about the difference of men and women. Moreover, Italy can be considered conservative when it comes to heteronormativity, resulting in more prejudice against gender identity. Furthermore, reaction posts from news articles about famous women, whose physical appearance is heavily criticized, trigger comments about expectations of beauty standards.

For the evaluation of the **text spans** I combine multiple evaluation measures, ranging from F1 score using character-overlap (Da San Martino et al., 2020) to the use of automatic measures such as BERTscore (Zhang et al., 2019) and (Sacre)BLEU (Papineni et al., 2002; Post, 2018). Similar to the category evaluation, models are compared using a pairwise approach. Considering the complexity of the task, I have adopted a relaxed evaluation scenario: text spans have been evaluated only on the categories that have been identified by both models, discarding all the instances where there is no agreement on the category. Detailed results are reported in Appendix F. Here

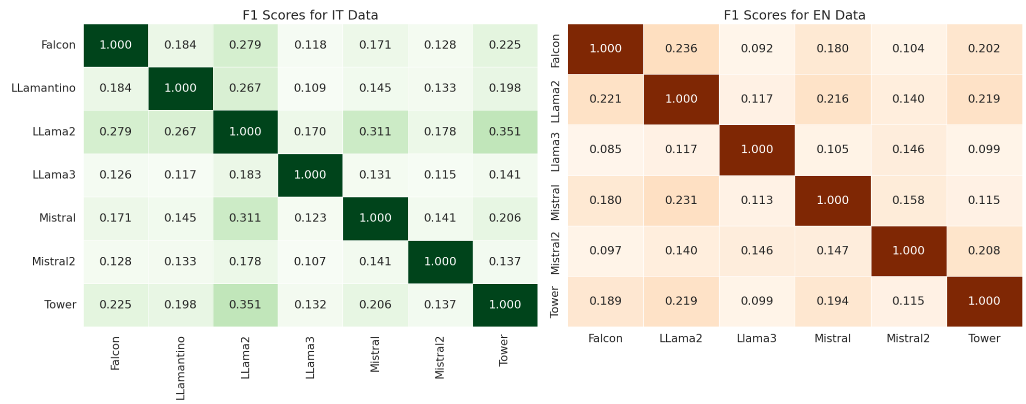


Figure 6.4: Inter-model agreement for Italian (left) and English (right). Columns represent the reference labels and rows the predictions.

I report a summary. For English, the top-performing model pairs according to Character Overlap, BERTScore F1, and BLEU are: Mistral-v1/Mistral-v2 for Character Overlap (0.330), Mistral-v2/LaMa2 for BERTScore F1 (0.868), Mistral-v2/LlaMa3 for BLEU (36.232). For Italian, the corresponding top-performing model pairs are Mistral-v1/Mistral-v2 with 0.192 for Character Overlap, 0.825 for BERTScore and Mistral-v1/LlaMa3 with 14.127 for BLEU. Regarding the character overlap, the scores are generally low for both English and Italian, with LLaMAntino consistently performing worst, despite being the only Italian-specific model. This is expected due to its verbose output, preventing it from reporting the exact words that occur in the original text. Instead, it tends to re-elaborate words and explaining why a certain category is identified. For BERTScore, the figures are reasonably high for both languages, with most scores above 0.800. This indicates that the models are generating text with contextual embeddings that are quite similar to the reference text. For BLEU, scores show more variation, but in any case the figures are relatively low indicating disagreement among the models on the extracted text spans, except for LlaMa3/Mistral-v1 and -v2 for English, where the scores are higher than 20. For Italian, the BLEU scores are generally lower, with the lowest being all the combinations with LLaMantino.

6.5.4.3 Evaluation against human annotation

I report the results of the evaluation against the manual annotations in Table 6.7 for the categories and the text span. As for the social dynamics, models reach higher performance with Italian data. Mistral-v1 achieves the best score in English (F1=0.181), while in Italian, LlaMa2 obtains 0.313. Concerning the text spans, on Italian data models obtain higher performance than in English. LLaMAntino consistently performs the worst, except for the BERTScore, in which it overcomes LlaMa2. This is due to the fact that since it tends to answer in a dialogue-fashion, instead of selecting a category and the exact text span, it re-elaborates the text by producing a verbose output, thus benefiting from similarity-based measures. Remarkably, excluding Falcon and LLaMAntino, the BLEU scores are better than those observed across models, suggesting that—although limited—models can capture problematic passages.

I observe a tendency of the models to select the category based on semantic similarity instead of the intended meaning of the sentence. For instance, the sentence *IT: non è* ***colpa*** *loro se sono stupide - EN: it's not their fault if they're stupid*, gets labeled as *victim blaming*, in Italian *colpevolizzazione della vittima*, most likely for the semantic association between *colpa* and *colpevolizzare*. However, in this case the correct label should be *stereotyping and generalization*, since the author is generalizing about the fact that women are stupid. Another aspect to consider is that not all errors are actually wrong. For instance, the text fragment *everybody would want to see that*, gets labeled as *stereotyping and generalization*. This is not wrong from a general point of view, since the sentence contains a social dynamic of generalization, but not towards women, and therefore should not be considered at all.

		Models						
Task	**Lang.**	**LLama3**	**LLama2**	**Mistral-v1**	**Mistral-v2**	**Tower**	**Falcon**	**LLaMAntino**
Social Dynamic	EN	0.126	0.00	**0.181**	0.143	0.116	0.060	-
	IT	0.217	**0.313**	0.199	0.213	0.213	0.182	0.144
Text span char. overlap	EN	**0.523**	0.000	0.409	0.466	0.173	0.016	-
	IT	0.481	**0.587**	0.472	0.403	0.278	0.218	0.178
Text span BERT-F1	EN	0.819	**0.883**	0.860	0.876	0.821	0.825	-
	IT	0.851	0.777	0.843	**0.864**	0.830	0.773	0.810
Text span BLEU	EN	29.749	**34.140**	16.500	26.260	13.960	1.411	-
	IT	18.875	10.026	16.108	**20.803**	11.856	2.702	1.211

Table 6.7: Evaluation against manual annotated data for English and Italian.

6.5.4.4 Models' errors.

In general, errors occur for both social dynamics and text spans. For the social dynamics, models tend to either distort the name of the categories (e.g., "whatsaboutism" instead of "whataboutism") or truncate the name of categories, up to inventing new categories. For English, a total of 91 made-up categories are found, among which "homophobia", "exclusionary attitudes towards feminism", and "sexual assault" could be plausible. The model that mostly generates new social dynamics is Mistral-v.1 (with 88 made-up categories). In Italian, the made-up categories are significantly fewer, only 10 in total. Regarding the text spans, models either tend to translate the original text, especially in models specifically trained for translation tasks like Tower, or they produce verbose answers by providing an explanation of why the sentence is misogynous. The model that tends to generate new texts more often—rather than extracting the text spans from the message—is Tower.

6.6 Final Remarks

I proposed the task of implicit misogyny detection under two different yet complementary subtasks: **implied assumptions generation under an Argumentative Reasoning perspective**; and **implied social dynamics identification as multi-label classification and text span identification**. This work establishes hate speech detection as a proxy task to evaluate the reasoning abilities of large language models (LLMs), as understanding implied statements—whether in free-text or multiple-choice form—requires reconstructing the missing link between a claim and its underlying meaning.

Prompt-based experiments show that LLMs fail 68% and 50% of the time in generating implied assumptions in Italian and English respectively. the The results are even lower when selecting social dynamics and text spans. Overall, I observe that models struggle to follow instructions (compliance with output format) and have limited abilities when it comes to associating the content of a message with the list of social dynamic categories. While this supports previous findings on recasting claims concerning emerging abilities of LLMs (Lu et al., 2023), it also indicates that LLMs have limited understanding of implied societal assumptions encoded in messages indicating the need of additional training/tuning.

The poor relationship between wrongly generated explanations and correctly predicted classes shows LLMs' over-reliance on their implicit knowledge and spurious correlations rather than reasoning skill. This is also confirmed when selecting multiple choice categories based on keywords that occur both in the text and in the selected category (e.g., *victim* in victim blaming), rather than grasping the underlying social dynamics. Findings are consistent with Zhu et al. (2023): prompting strategies that rely on implicit knowledge in LLMs often generate an incorrect classification when the generated knowledge is wrong due to - in our case - lack of references, reasoning skills, or understanding of non-standard language.

To conclude, *(i)* the performance of the classification task cannot be used as a proxy to guarantee the correctness of the implied assumption/warrant; *(ii)* LLMs do not have the necessary reasoning abilities in order to understand highly implicit misogynistic statements. For example, such statements might involve coded language, ambiguous phrasing, or sarcastic undertones, which require an understanding of pragmatics, societal context, and speaker intent—areas where LLMs are inherently limited. Additionally, implicit misogyny often intersects with stereotypes or allusions that demand world knowledge and the ability to infer meaning beyond the surface text. These limitations arise because LLMs primarily operate on statistical correlations within their training data, rather than possessing genuine conceptual understanding or critical reasoning skills.

Chapter 7

Conclusion and Future Work

This thesis has explored NLP methods to address misogyny that is "hidden in plain sight" within linguistic structures. From early approaches to more advanced methods, I developed computational models to detect and categorize misogyny based on different taxonomies, considering both monolingual (Italian and English) and cross-lingual contexts. I also explored approaches that combine visual and textual information to identify misogyny in a multimodal setting, specifically in memes.

Early experiments and analysis of the results presented in Chapter 4 showed that:

- Models to identify misogyny in English data struggle with the use of casual slurs, by producing a great number of false positives.
- Models to identify misogyny in Italian data struggle mostly with the use of pejorative language, whose ambiguity confuses both encoder-based and decoder-based models.
- All models tested for English and Italian data find it difficult to spot implicit misogyny.

Based on these findings, the focus of the rest of the thesis is shifted to two main aspects: pejorative language disambiguation for Italian and implicit misogyny explanation for both Italian and English, which are two core issues of state-of-the-art models for misogyny detection.

The first main contribution of this thesis is the development of two new Italian datasets for misogyny understanding and detection: PejorativITy (see

Chapter 5) and ImplicIT-Mis (see Chapter 6). Those datasets make up for the lack of resources for the Italian language when it comes to hate speech, especially misogyny.

The second main contribution is the novel framework proposed in Chapter 5 to address the issue of polysemic words that trigger errors in classifiers, especially in the Italian language. The framework foresees the disambiguation of misogynistic pejorative epithets as a preliminary step for misogyny detection. This step improves the system's ability to capture context-specific meanings and contributes to a more precise identification of misogyny. Results show that fine-tuning encoder-based systems perform better than prompting more recent decoder-based LLMs in both the disambiguation of pejorative epithets and the detection of misogyny. This suggests that LLMs lack significant knowledge about cultural aspects and linguistic expressions of Italian language. This raises questions about the cultural awareness of multilingual models.

The third main contribution is presented in Chapter 6, which presents the novel approach of addressing the task of implicit misogyny detection and explanation under the Argumentation Reasoning framework, which provides a structured basis for reasoning about the implications of misogynistic language. LLMs are asked to identify the misogynistic implications in Twitter posts and Facebook comments by (*i*) reconstructing the *warrant* of a statement through free-text generation or (*ii*) classifying implications into a newly developed taxonomy of social dynamics. Overall, experiments reveal that LLMs, while capable of basic misogynistic content recognition, often misinterpret or overlook implicit cues that require commonsense reasoning, such as real-world knowledge, sarcasm, and culturally sensitive metaphors specifically in Italian. This exposes the limits of LLMs' ability to identify misogyny in non-English languages.

7.1 Future Directions

There are several possible research directions that can be investigated to further improve and expand the contributions presented in this thesis.

Cross-lingual Pejorative Language Experiments performed to investigate LLMs' understanding of pejorative epithets in Italian show a low performance, as the non-standard meaning of such terms is not identified by the models. As future direction it would be interesting to investigate whether this finding is valid for other languages, and how the connotative shifts of pejorative epithets evolve over time and across platforms. By analyzing contextualized word embeddings of data containing the target words mentioned in Chapter 5, it would be possible to track how terms acquire misogynistic connotations within specific domains and compare these across time and languages. This also includes a study on the negative connotation of the words *feminist, feminists* across time, languages and domains.

Knowledge-Augmentation for Implicit Misogyny The experiments performed in Chapter 6 showed that hate-related natural language inference tasks are still a big challenge to address in NLP. Addressing implicit misogyny, especially in non-English languages, requires enhancing models with language-specific and culture-specific commonsense knowledge to identify aspects often lost in translation or lacking in the knowledge acquired by LLMs during pre-training. Future work could leverage the injection of cultural references and gender-based societal norms, which can be retrieved via encyclopedic knowledge, either in the model architecture in the pre-training phase, or in the prompt at inference phase. This would include knowledge tied to non-standard and regional expressions.

Personalized LLMs The way in which misogyny and hate speech in general are perceived could vary across languages and cultures. This is also visible in legislation: different countries consider different hate speech components to be prosecutable. For instance, more progressive countries like Denmark and The Netherlands consider hate targeting sexual orientation as prosecutable; countries from the Arab world rarely include sexual orientation and prioritize religion-based hate speech, whereas Indian subcontinent considers caste-related hate in its legislation. Ideally, systems should be able to personalize to a user's cultural background, as well as to their identity and values, going beyond standard demographic profiling. The system could also allow users to input their personal preferences or select predefined profiles (e.g., "sensitive to gender-based hate," "sensitive to religion-based hate") to refine the detection. This would make it adaptable not only to national or cultural standards but also to individual sensitivities, while still enforcing universal ethical boundaries. This would also allow for an exploration of whether national legislation for hate speech reflects individual values.

Bibliography

Gavin Abercrombie, Aiqi Jiang, Poppy Gerrard-abbott, Ioannis Konstas, and Verena Rieser. Resources for automated identification of online gender-based violence: A systematic review. In Yi-ling Chung, Paul R\"ottger, Debora Nozza, Zeerak Talat, and Aida Mostafazadeh Davani, editors, *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 170–186, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.woah-1.17. URL `https://aclanthology.org/2023.woah-1.17`.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. Tower: An open multilingual large language model for translation-related tasks. 2024.

Rikke Amundsen. *Cruel Intentions and Social Conventions: Locating the Shame in Revenge Porn*, pages 131–148. Springer International Publishing, Cham, 2019. ISBN 978-3-319-96226-9. doi: 10.1007/978-3-319-96226-9_7. URL `https://doi.org/10.1007/978-3-319-96226-9_7`.

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

E. Michael Bannester. Sociodynamics: An integrative theorem of power, authority, interfluence and love. *American Sociological Review*, 34(3): 374–393, 1969. ISSN 00031224. URL `http://www.jstor.org/stable/2092502`.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. ConSeC: Word sense disambiguation as continuous sense comprehension. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.112. URL `https://aclanthology.org/2021.emnlp-main.112`.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.148. URL `https://aclanthology.org/2020.findings-emnlp.148`.

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. SemEval 2022 task 10: Structured sentiment analysis. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.semeval-1.180. URL `https://aclanthology.org/2022.semeval-1.180`.

Manuela Barreto and David Matthew Doyle. Benevolent and hostile sexism in a shifting global context. *Nature reviews psychology*, 2(2):98–111, February 2023. ISSN 2731-0574. doi: https://doi.org/10.1038/s44159-022-00136-x.

Angelo Basile and Chiara Rubagotti. Crotonemilano for AMI at evalita2018. A performant, cross-lingual misogyny detection system. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of* *the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018. URL http://ceur-ws.org/Vol-2263/paper034.pdf.

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. Llamantino: Llama 2 models for effective text generation in italian language. *arXiv preprint arXiv:2312.09993*, 2023.

Valerio Basile and Malvina Nissim. Sentiment analysis on italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, 2013.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007. URL https://aclanthology.org/S19-2007.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019b. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007. URL https://aclanthology.org/S19-2007.

Maria Becker, Katharina Korfhage, and Anette Frank. Implicit knowledge in argumentative texts: An annotated corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios

Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2316–2324, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.282.

Robin Bergh and Mark J. Brandt. Generalized prejudice: Lessons about social power, ideological conflict, and levels of abstraction. *European Review of Social Psychology*, 34(1):92–126, 2023. doi: 10.1080/10463283.2022.2040140. URL https://doi.org/10.1080/10463283.2022.2040140.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. Recent trends in word sense disambiguation: A survey. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/593. URL https://doi.org/10.24963/ijcai.2021/593. Survey Track.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL https://doi.org/10.5281/zenodo.5297715. If you use this software, please cite it using these metadata.

Khandis R. Blake, Siobhan M. O'Dean, James Lian, and Thomas F. Denson. Misogynistic tweets correlate with violence against women. *Psychological Science*, 32(3):315–325, 2021. doi: 10.1177/0956797620968529. URL https://doi.org/10.1177/0956797620968529. PMID: 33593204.

Filip Boltužić and Jan Šnajder. Fill the gap! analyzing implicit premises between claims from online debates. In Chris Reed, editor, *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 124–133, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2815. URL https://aclanthology.org/W16-2815.

Leonard Bongard, Lena Held, and Ivan Habernal. The legal argument reasoning task in civil procedure. In Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, and Daniel Preoțiuc-Pietro, editors, *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 194–207, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nllp-1.17. URL `https://aclanthology.org/2022.nllp-1.17`.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1176. URL `https://aclanthology.org/D19-1176`.

Federico Cabitza, Andrea Campagner, and Valerio Basile. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868, June 2023. ISSN 2159-5399. doi: 10.1609/aaai.v37i6.25840. URL `http://dx.doi.org/10.1609/aaai.v37i6.25840`.

ZeFeng Cai, Linlin Wang, Gerard de Melo, Fei Sun, and Liang He. Multi-scale distribution deep variational autoencoder for explanation generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 68–78, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.7. URL `https://aclanthology.org/2022.findings-acl.7`.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. HateBERT: Retraining BERT for abusive language detection in English. In Aida Mostafazadeh Davani, Douwe Kiela, Mathias Lambert, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem, editors, *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.woah-1.3. URL `https://aclanthology.org/2021.woah-1.3`.

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. Guiding principles for participatory design-inspired natural language processing. In Anjalie Field, Shrimai Prabhumoye, Maarten Sap, Zhijing Jin, Jieyu Zhao, and Chris Brockett, editors, *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35, Online, August 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.nlp4posimpact-1.4. URL https://aclanthology.org/2021.nlp4posimpact-1.4.

Camilla Casula and Sara Tonelli. Hate speech detection with machine-translated data: The role of annotation scheme, class imbalance and undersampling. *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, 2020.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020.

Tuhin Chakrabarty, Aadit Trivedi, and Smaranda Muresan. Implicit premise generation with discourse-aware commonsense knowledge models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6247–6252, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.504. URL https://aclanthology.org/2021.emnlp-main.504.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), mar 2024. ISSN 2157-6904. doi: 10.1145/3641289. URL https://doi.org/10.1145/3641289.

Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. Exploring the potential of large language models in computational argumentation. *CoRR*, abs/2311.09022, 2023. doi: 10.48550/ARXIV.2311.09022. URL https://doi.org/10.48550/arXiv.2311.09022.

Lei Chen and Hou Wei Chou. RIT boston at SemEval-2022 task 5: Multimedia misogyny detection by using coherent visual and language features from CLIP model and data-centric AI principle. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, *Proceedings* *of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 636–641, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.semeval-1.87. URL `https://aclanthology.org/2022.semeval-1.87`.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. An annotated corpus for sexism detection in French tweets. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1397–1403, Marseille, France, May 2020a. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.175`.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. He said "who's gonna take care of your children when you are at ACL?": Reported sexist acts are not sexist. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4055–4066, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.373. URL `https://aclanthology.org/2020.acl-main.373`.

Alex Clark. Pillow (pil fork) documentation, 2015. URL `https://buildmedia.readthedocs.org/media/pdf/pillow/latest/pillow.pdf`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. 2020. URL `https://arxiv.org/abs/1911.02116`.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.186. URL `https://aclanthology.org/2020.semeval-1.186`.

Aaron Defazio and Samy Jelassi. Adaptivity without Compromise: A Momentumized, Adaptive, Dual Averaged Gradient Method for Stochastic Optimization. *arXiv*, 2021.

Pieter Delobelle, Murilo Cunha, Eric Massip Cano, Jeroen Peperkamp, and Bettina Berendt. Computational ad hominem detection. In Fernando Alva-Manchego, Eunsol Choi, and Daniel Khashabi, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 203–209, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-2028. URL `https://aclanthology.org/P19-2028`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.

Liviu P. Dinu, Ioan-Bogdan Iordache, Ana Sabina Uban, and Marcos Zampieri. A computational exploration of pejorative language in social media. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3493–3498, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.296. URL `https://aclanthology.org/2021.findings-emnlp.296`.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL `https://doi.org/10.1145/3278721.3278729`.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.29. URL `https://aclanthology.org/2021.emnlp-main.29`.

Deborah Fallows. How women and men use the internet. Technical report, Pew Internet & American Life Project, December 2005.

Yanbo Fang and Yongfeng Zhang. Data-efficient concept extraction from pre-trained language models for commonsense explanation generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5883–5893, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.433. URL `https://aclanthology.org/2022.findings-emnlp.433`.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1171–1182, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.101. URL `https://aclanthology.org/2023.acl-short.101`.

Diane Felmlee, Paulina Inara Rodis, and Amy Zhang. Sexist Slurs: Reinforcing Feminine Stereotypes Online. *Sex Roles*, 83(1):16–28, July 2020. ISSN 1573-2762. doi: 10.1007/s11199-019-01095-z. URL `https://doi.org/10.1007/s11199-019-01095-z`.

Vanessa Wei Feng and Graeme Hirst. Classifying arguments by scheme. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `https://aclanthology.org/P11-1099`.

Elisabetta Fersini, Debora Nozza, and Paolo 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*, pages 59–66. Torino: Accademia University Press, 2018a. doi: doi:10.4000/books.aaccademia.4497.

Elisabetta Fersini, Debora Nozza, Paolo Rosso, et al. Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of the 6th Evaluation Campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2018)*, volume 2263, pages 1–9. CEUR-WS, 2018b.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. Overview of the task on automatic misogyny identification at IberEval 2018. In *Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, Sevilla, Spain, 2018c.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, 2018d.

Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. Detecting sexist meme on the web: A study on textual and visual cues. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 226–231, Los Alamitos, CA, USA, sep 2019. IEEE Computer Society. doi: 10.1109/ACIIW.2019.8925199. URL `https://doi.ieeecomputersociety.org/10.1109/ACIIW.2019.8925199`.

Elisabetta Fersini, Debora Nozza, and Giulia Boifava. Profiling Italian misogynist: An empirical study. In Johanna Monti, Valerio Basile, Maria Pia Di Buono, Raffaele Manna, Antonio Pascucci, and Sara Tonelli, editors, *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*, pages 9–13, Marseille, France, May 2020a. European Language Resources Association (ELRA). ISBN 979-10-95546-49-8. URL `https://aclanthology.org/2020.restup-1.3`.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. AMI@EVALITA2020: Automatic misogyny identification. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*. CEUR, 2020b.

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

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. SemEval-2022 task 5: Multimedia automatic misogyny identification. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.semeval-1.74. URL `https://aclanthology.org/2022.semeval-1.74`.

Caterina Flick. *The Legal Framework on Hate Speech and the Internet Good Practices to Prevent and Counter the Spread of Illegal Hate Speech Online: Good Practices to Prevent and Counter the Spread of Illegal Hate Speech Online*. 12 2020. ISBN 978-88-6969-479-0. doi: 10.30687/978-88-6969-478-3/011.

Glenn Fosbraey and Nicola Puckey. *Misogyny, Toxic Masculinity, and Heteronormativity in Post-2000 Popular Music*. 01 2021. ISBN 978-3-030-65188-6. doi: 10.1007/978-3-030-65189-3.

Simona Frenda, Bilal Ghanem, Manuel Montes y Gómez, and Paolo Rosso. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *J. Intell. Fuzzy Syst.*, 36:4743–4752, 2019. URL `https://api.semanticscholar.org/CorpusID:156056029`.

Rachael Fulper, Giovanni Luca Ciampaglia, Emilio Ferrara, Yong-Yeol Ahn, Alessandro Flammini, Filippo Menczer, Kehontas Rowe, and Bryce Lewis. Misogynistic language on twitter and sexual violence. In *Proceedings of the acm web science workshop on computational approaches to social modeling (chasm)1–4.*, 2014.

Paolo Gajo, Arianna Muti, Katerina Korre, Silvia Bernardini, and Alberto Barrón-Cedeño. On the identification and forecasting of hate speech in inceldom. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 373–384, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria. URL `https://aclanthology.org/2023.ranlp-1.42`.

Adam D. Galinsky, Cynthia S. Wang, Jennifer A. Whitson, Eric M. Anicich, Kurt Hugenberg, and Galen V. Bodenhausen. The reappropriation of stigmatizing labels: The reciprocal relationship between power and self-labeling. *Psychological Science*, 24(10):2020–2029, 2013. ISSN 09567976, 14679280. URL `http://www.jstor.org/stable/24539396`.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Debbie Ging. *Bros v. Hos: Postfeminism, Anti-feminism and the Toxic Turn in Digital Gender Politics*, pages 45–67. Springer International Publishing, Cham, 2019. ISBN 978-3-319-96226-9. doi: 10.1007/978-3-319-96226-9_3. URL `https://doi.org/10.1007/978-3-319-96226-9_3`.

Giuliana Giusti and Gabriele Iannàccaro. *Language, Gender and Hate Speech A Multidisciplinary Approach.* Fondazione Università Ca' Foscari, dec 2020. doi: 10.30687/978-88-6969-478-3. URL `https://doi.org/10.30687%2F978-88-6969-478-3`.

Peter Glick and Susan T. Fiske. Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of Women Quarterly*, 21(1):119–135, 1997. ISSN 0361-6843. doi: 10.1111/j.1471-6402.1997.tb00104.x.

Peter Glick and Susan T. Fiske. *The ambivalent sexism Inventory: Differentiating hostile and benevolent sexism*, pages 116–160. Taylor and Francis, United States, January 2018. ISBN 9781138734333. doi: 10.4324/9781315187280. Publisher Copyright: © 2018 Susan T. Fiske.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. Fallacious argument classification in political debates. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/575. URL `https://doi.org/10.24963/ijcai.2022/575`. Main Track.

Deniz Gorur, Antonio Rago, and Francesca Toni. Can large language models perform relation-based argument mining?, 2024.

Meghan Gosse, Michael Halpin, and Finlay Maguire. Stochastic gender-based violence: How incels justify and encourage sexualized violence against women. *Violence Against Women*, 0(0):10778012241279816, 2024. doi: 10.1177/10778012241279816. URL `https://doi.org/10.1177/10778012241279816`. PMID: 39279344.

Claire Gothreau, Kevin Arceneaux, and Amanda Friesen. Hostile, Benevolent, Implicit: How Different Shades of Sexism Impact Gendered Policy Attitudes. 4, 2022. ISSN 2673-3145. doi: 10.3389/fpos.2022.817309. URL `https://www.frontiersin.org/articles/10.3389/fpos.2022.817309`.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. An expert annotated dataset for the detection of online misogyny. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.114. URL `https://aclanthology.org/2021.eacl-main.114`.

Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179, April 2017. doi: 10.1162/COLI_a_00276. URL `https://aclanthology.org/J17-1004`.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-1175. URL `https://aclanthology.org/N18-1175`.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana, June 2018b. Association for Computational Linguistics. doi: 10.18653/v1/N18-1036. URL `https://aclanthology.org/N18-1036`.

Mika Hagerlid. Swedish women's experiences of misogynistic hate crimes: The impact of victimization on fear of crime. *Feminist Criminology*, 16(4): 504–525, 2021. doi: 10.1177/1557085120957731. URL `https://doi.org/10.1177/1557085120957731`.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.234. URL `https://aclanthology.org/2022.acl-long.234`.

Maram Hasanain, Fatema Ahmed, and Firoj Alam. Large language models for propaganda span annotation. 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Sarah Hewitt, T. Tiropanis, and C. Bokhove. The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, page 333–335, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342087. doi: 10.1145/2908131.2908183. URL https://doi.org/10.1145/2908131.2908183.

Hideitsu Hino. Active learning: Problem settings and recent developments. *CoRR*, abs/2012.04225, 2020. URL https://arxiv.org/abs/2012.04225.

Eric Holgate, Isabel Cachola, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. Why swear? analyzing and inferring the intentions of vulgar expressions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4405–4414, Brussels, Belgium, October-November. Association for Computational Linguistics.*, 2018.

Alexander Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. Natural language decompositions of implicit content enable better text representations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13188–13214, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.815. URL https://aclanthology.org/2023.emnlp-main.815.

Claudiu Daniel Hromei, Danilo Croce, Valerio Basile, and Roberto Basili. Extremita at evalita 2023: Multi-task sustainable scaling to large language models at its extreme. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, Parma, Italy, September 2023. CEUR.org.

Sylvia Jaki, Tom De Smedt, Maja Gwóźdź, Rudresh Panchal, Alexander Rossa, and Guy De Pauw. Online hatred of women in the Incels.me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict*, 7(2):240–268, 2019. ISSN 2213-1272. doi: https://doi.org/10.1075/jlac.00026.jak. URL https://www.jbe-platform.com/content/journals/10.1075/jlac.00026.jak. Publisher: John Benjamins Type: Journal Article.

Emma Jane. *Misogyny Online: A Short (and Brutish) History*. 10 2016. ISBN 9781473916005. doi: 10.4135/9781473916029.

Akshita Jha and Radhika Mamidi. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In Dirk Hovy, Svitlana Volkova, David Bamman, David Jurgens, Brendan O'Connor, Oren Tsur, and A. Seza Doğruöz, editors, *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2902. URL https://aclanthology.org/W17-2902.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, and Minlie Huang. Generating commonsense explanation by extracting bridge concepts from reasoning paths. In Kam-Fai Wong, Kevin Knight, and Hua Wu, editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 248–257, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.aacl-main.28.

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. Mistral 7b, 2023.

Katikapalli Subramanyam Kalyan. A survey of gpt-3 family large language models including chatgpt and gpt-4. 2023. URL https://arxiv.org/abs/2310.12321.

Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

Dax J. Kellie, Khandis R. Blake, and Robert C. Brooks. What drives female objectification? an investigation of appearance-based interpersonal perceptions and the objectification of women. *PLoS One*, 14(8):e0221388, 2019. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6707629/.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide

Testuggine. Supervised multimodal bitransformers for classifying images and text, 2020a.

Douwe Kiela, Hamed Firooz, and Aravind Mohan. Hateful Memes Challenge and dataset for research on harmful multimodal content. *MetaAI*, 2020b. URL https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set/.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A. Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, Niklas Muennighoff, Riza Velioglu, Jewgeni Rose, Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, Helen Yannakoudakis, Vlad Sandulescu, Umut Ozertem, Patrick Pantel, Lucia Specia, and Devi Parikh. The hateful memes challenge: Competition report. In Hugo Jair Escalante and Katja Hofmann, editors, *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 344–360. PMLR, 06–12 Dec 2021a. URL https://proceedings.mlr.press/v133/kiela21a.html.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *arXiv preprint arXiv:2005.04790*, 2021b.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. SemEval-2023 task 10: Explainable detection of online sexism. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.305. URL https://aclanthology.org/2023.semeval-1.305.

Katerina Korre, Arianna Muti, and Alberto Barrón-Cedeño. The challenges of creating a parallel multilingual hate speech corpus: An exploration. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources* *and Evaluation (LREC-COLING 2024)*, pages 15842–15853, Torino, Italia, May 2024a. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1376.

Katerina Korre, Arianna Muti, Federico Ruggeri, and Alberto Barrón-Cedeño. Untangling hate speech definitions: A semantic componential analysis across cultures and domains, 2024b. URL https://arxiv.org/abs/2411.07417.

Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011. URL https://api.semanticscholar.org/CorpusID:59901023.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Eduardo Blanco and Wei Lu, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. *Proceedings of the Fourth Workshop on Online Abuse and Harms (pp. 138-149)*, arXiv:1503.06733, 2020.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. An argument-annotated corpus of scientific publications. In Noam Slonim and Ranit Aharonov, editors, *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5206. URL https://aclanthology.org/W18-5206.

John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818, December 2019. doi: 10.1162/coli_a_00364. URL https://aclanthology.org/J19-4006.

Alyssa Lees, Jeffrey Sorensen, and Ian Kivlichan. Jigsaw@AMI and HaSpeeDe2: Fine-Tuning a Pre-Trained Comment-Domain BERT Model. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*. CEUR.org, 2020.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv preprint arXiv:1908.03557*, 2019.

Chi-Liang Liu, Tsung-Yuan Hsu, Yung-Sung Chuang, and Hung yi Lee. What makes multilingual bert multilingual? *arXiv*, 2020.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. Generated knowledge prompting for commonsense reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.225. URL `https://aclanthology.org/2022.acl-long.225`.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), jan 2023a. ISSN 0360-0300. doi: 10.1145/3560815. URL `https://doi.org/10.1145/3560815`.

Zhexiong Liu, Mohamed Elaraby, Yang Zhong, and Diane Litman. Overview of ImageArg-2023: The first shared task in multimodal argument mining. In Milad Alshomary, Chung-Chi Chen, Smaranda Muresan, Joonsuk Park, and Julia Romberg, editors, *Proceedings of the 10th Workshop on Argument Mining*, pages 120–132, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.argmining-1.12. URL `https://aclanthology.org/2023.argmining-1.12`.

Filipa Melo Lopes. Perpetuating the patriarchy: Misogyny and (post-)feminist backlash. *Philosophical Studies*, 176(9):2517–2538, 2019. doi: 10.1007/s11098-018-1138-z.

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. URL `http://arxiv.org/abs/1711.05101`.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *arXiv preprint arXiv:1908.02265*, 2019.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning? 2023.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.20. URL `https://aclanthology.org/2023.ijcnlp-main.20`.

Shana MacDonald and Brianna I. Wiens. Feminist memes: Digital communities, identity performance and resistance from the shadows. *Materializing Digital Futures: Touch, Movement, Sound and Vision*, page 123, 2022. Publisher: Bloomsbury Publishing USA.

Eleonora Mancini, Federico Ruggeri, and Paolo Torroni. Multimodal fallacy classification in political debates. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 170–178, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.eacl-short.16`.

Kate Manne. *Down Girl: The Logic of Misogyny*. Oxford University Press, 2017. ISBN 9780190604981.

Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*, 2020.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875, 2021.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL `https://aclanthology.org/2020.acl-main.173`.

Kira K. Means. "not like other girls": Implicit and explicit dimensions of internalized sexism and behavioral outcomes. 2021. URL `https://api.semanticscholar.org/CorpusID:235408551`.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.845. URL `https://aclanthology.org/2023.acl-long.845`.

Christian Morbidoni and Annalina Sarra. Can llms assist humans in assessing online misogyny? experiments with gpt-3.5. 2023.

M. Mozafari, R. Farahbakhsh, and N. Crespi. A bert-based transfer learning approach for hate speech detection in online social media. In H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, and L. M. Rocha, editors, *Complex Networks and Their Applications VIII: Proceedings of the Eighth International Conference on Complex Networks and Their Applications, COMPLEX NETWORKS 2019, Lisbon, Portugal, December 10-12, 2019*, volume 1 of *Studies in Computational Intelligence*, pages 928–940. Springer, 2019.

Hala Mulki and Bilal Ghanem. Armi at fire 2021: Overview of the first shared task on arabic misogyny identification. *FIRE (Working Notes)*, pages 820–830, 2021.

Arianna Muti and Alberto Barrón-Cedeño. UniBO@AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*. CEUR, 2020.

Arianna Muti and Alberto Barrón-Cedeño. A checkpoint on multilingual misogyny identification. In Samuel Louvan, Andrea Madotto, and Brielen Madureira, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 454–460, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-srw.37. URL `https://aclanthology.org/2022.acl-srw.37`.

Arianna Muti, Francesco Fernicola, and Alberto Barrón-Cedeño. Misogyny and aggressiveness tend to come together and together we address them. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4142–4148, Marseille, France, June 2022a. European Language Resources Association. URL `https://aclanthology.org/2022.lrec-1.440`.

Arianna Muti, Katerina Korre, and Alberto Barrón-Cedeño. UniBO at SemEval-2022 task 5: A multimodal bi-transformer approach to the binary and fine-grained identification of misogyny in memes. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 663–672, Seattle, United States, July 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.semeval-1.91. URL `https://aclanthology.org/2022.semeval-1.91`.

Arianna Muti, Francesco Fernicola, and Alberto Barrón-Cedeño. UniBoe's at SemEval-2023 task 10: Model-agnostic strategies for the improvement of hate-tuned and generative models in the classification of sexist posts. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings* *of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1138–1147, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.158. URL https://aclanthology.org/2023.semeval-1.158.

Arianna Muti, Federico Ruggeri, Khalid Al-Khatib, Alberto Barrón-Cedeño, and Tommaso Caselli. Language is scary when over-analyzed: Unpacking implied misogynistic reasoning with argumentation theory-driven prompts, 2024a. URL https://arxiv.org/abs/2409.02519.

Arianna Muti, Federico Ruggeri, Cagri Toraman, Alberto Barrón-Cedeño, Samuel Algherini, Lorenzo Musetti, Silvia Ronchi, Gianmarco Saretto, and Caterina Zapparoli. PejorativITy: Disambiguating pejorative epithets to improve misogyny detection in Italian tweets. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12700–12711, Torino, Italia, May 2024b. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1112.

Rostyslav Neskorozhenyi. How to get high score using MMBT and CLIP in Hateful Memes Competition, 2021. URL https://towardsdatascience.com/how-to-get-high-score-using-mmbt-and-clip-in-hateful-memes-\competition-90bfa65cb117.

Annalaura Nocentini, Juan Calmaestra, Anja Schultze-Krumbholz, Herbert Scheithauer, Rosario Ortega, and Ersilia Menesini. Cyberbullying: Labels, behaviours and definition in three european countries. *Australian Journal of Guidance and Counselling*, 20(2):129–142, 2010. doi: 10.1375/ajgc.20.2.129.

Debora Nozza. Exposing the limits of zero-shot cross-lingual hate speech detection. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.114. URL https://aclanthology.org/2021.acl-short.114.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 149–155, Thessaloniki, Greece, 2019.

Debora Nozza, Federico Bianchi, and Dirk Hovy. What the [mask]? making sense of language-specific bert models. 2020. URL https://arxiv.org/abs/2003.02912.

Debora Nozza, Federico Bianchi, and Giuseppe Attanasio. HATE-ITA: Hate speech detection in Italian social media text. In Kanika Narang, Aida Mostafazadeh Davani, Lambert Mathias, Bertie Vidgen, and Zeerak Talat, editors, *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 252–260, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.woah-1.24. URL https://aclanthology.org/2022.woah-1.24.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. An in-depth analysis of implicit and subtle hate speech messages. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.147. URL https://aclanthology.org/2023.eacl-main.147.

Endang W. Pamungkas, Valerio Basile, and Viviana Patti. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57, 2020a. doi: https://doi.org/10.1016/j.ipm.2020.102360. URL http://www.sciencedirect.com/science/article/pii/S0306457320308554.

Endang Wahyu Pamungkas and Viviana Patti. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-2051. URL https://aclanthology.org/P19-2051.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study. *Infor-*

mation Processing & Management, 57(6):102360, 2020b. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2020.102360. URL `https://www.sciencedirect.com/science/article/pii/S0306457320308554`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://aclanthology.org/P02-1040`.

P. Parikh, H. Abburi, P. Badjatiya, R. Krishnan, N. Chhaya, M. Gupta, and V. Varma. Multi-label categorization of accounts of sexism using a neural framework. *arXiv preprint arXiv:1910.04602*, 2019. URL `https://arxiv.org/abs/1910.04602`.

Pulkit Parikh, Harika Abburi, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. Categorizing sexism and misogyny through neural approaches. *ACM Trans. Web*, 15(4), jun 2021. ISSN 1559-1131. doi: 10.1145/3457189. URL `https://doi-org.ezproxy.unibo.it/10.1145/3457189`.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.731. URL `https://aclanthology.org/2022.emnlp-main.731`.

Laura Plaza, Jorge Carrillo-de Albornoz, Roser Morante, Julio Gonzalo, Enrique Amigó, Damiano Spina, and Paolo Rosso. Overview of exist 2023: sexism identification in social networks. In *Proceedings of ECIR'23*, pages 593–599, 2023. doi: 10.1007/978-3-031-28241-6_68.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio 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, Bari, Italy, 2019. CEUR. URL `https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074851349&partnerID=40&md5=7abed946e06f76b3825ae5e294ffac14`.

Matt Post. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL `https://aclanthology.org/W18-6319`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Leeat Ramati-Ziber, Nurit Shnabel, and Peter Glick. The beauty myth: Prescriptive beauty norms for women reflect hierarchy-enhancing motivations leading to discriminatory employment practices. *Journal of Personality and Social Psychology*, pages 1–27, 2019. doi: 10.1037/pspi0000209.

Paul Reisert, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. A computational approach for generating toulmin model argumentation. In Claire Cardie, editor, *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 45–55, Denver, CO, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-0507. URL `https://aclanthology.org/W15-0507`.

Shannon Ridgeway. 25 everyday examples of rape culture. *Everyday Feminism*, 2014. URL `https://everydayfeminism.com/2014/03/examples-of-rape-culture/`.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67(0):195–207, 2021. ISSN 1989-7553. URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6389.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. Two contrasting data annotation paradigms for subjective NLP tasks. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.13. URL https://aclanthology.org/2022.naacl-main.13.

Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. ExplaGraphs: An explanation graph generation task for structured commonsense reasoning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7740, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.609. URL https://aclanthology.org/2021.emnlp-main.609.

Niloofar S. Samghabadi, Parth Patwa, Srinivas Pykl, Prerana Mukherjee, Amitava Das, and Thamar Solorio. Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*, 2020.

Mattia Samory. The 'call me sexist but' dataset (cmsb). . Data File Version 1.0.0, https://doi.org/10.7802/2251, 2021.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.486. URL https://aclanthology.org/2020.acl-main.486.

Gabriele Sarti and Malvina Nissim. It5: Large-scale text-to-text pretraining for italian language understanding and generation. *ArXiv preprint 2203.03759*, mar 2022. URL `https://arxiv.org/abs/2203.03759`.

Heather Savigny. *Sexism and Misogyny*. John Wiley & Sons, Ltd, 2020. ISBN 9781119429128. doi: https://doi.org/10.1002/9781119429128.iegmc092. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119429128.iegmc092`.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. Findings of the WMT 2023 shared task on machine translation with terminologies. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.54. URL `https://aclanthology.org/2023.wmt-1.54`.

Sima Sharifirad, Borna Jafarpour, and Stan Matwin. Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. In Darja Fišer, Ruihong Huang, Vinodkumar Prabhakaran, Rob Voigt, Zeerak Waseem, and Jacqueline Wernimont, editors, *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 107–114, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5114. URL `https://aclanthology.org/W18-5114`.

Chhavi Sharma and Viswanath Pulabaigari. A Curious Case of Meme Detection: An Investigative Study. In *International Conference on Web Information Systems and Technologies (WEBIST)*, pages 327–338, 2020.

Eugenia Siapera. *Online Misogyny as Witch Hunt: Primitive Accumulation in the Age of Techno-capitalism*, pages 21–43. 07 2019. ISBN 978-3-319-96225-2. doi: 10.1007/978-3-319-96226-9_2.

Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. MMF: A multimodal framework for vision and language research. `https://github.com/facebookresearch/mmf`, 2020.

Beatrice Spallaccia. *It's a Man's World (Wide Web): A Critical Analysis of Online Misogyny and Hate Speech*. Bononia University Press, 2020. ISBN 9788869236143. URL `https://books.google.it/books?id=nQ_KzgEACAAJ`.

Neha Srikanth and Junyi Jessy Li. Elaborative simplification: Content addition and explanation generation in text simplification. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.455. URL `https://aclanthology.org/2021.findings-acl.455`.

Kalpana Srivastava, Suprakash Chaudhury, Pookala Bhat, and Samiksha Sahu. Misogyny, feminism, and sexual harassment. *Industrial Psychiatry Journal*, 26:111, 07 2017a. doi: 10.4103/ipj.ipj_32_18.

Kalpana Srivastava, Suprakash Chaudhury, P.S. Bhat, and Samiksha. Sahu. Misogyny, feminism, and sexual harassment. *Industrial psychiatry journal*, 26(2):111–113, 2017b. doi: https://doi.org/10.4103/ipj.ipj_32_18.

Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659, September 2017. doi: 10.1162/COLI_a_00295. URL `https://aclanthology.org/J17-3005`.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583, 2019. URL `http://arxiv.org/abs/1905.05583`.

Janet Swim, Robyn Mallett, and Charles Stangor. Understanding subtle sexism: Detection and use of sexist language. *Sex Roles*, 51:117–128, 01 2004. doi: 10.1023/B:SERS.0000037757.73192.06.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. Leap-of-thought: teaching pre-trained models to systematically reason over implicit knowledge. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Wilson L. Taylor. Cloze procedure: a new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433, 1953.

Junfeng Tian, Man Lan, and Yuanbin Wu. ECNU at SemEval-2018 task 12: An end-to-end attention-based neural network for the argument reasoning comprehension task. In Marianna Apidianaki, Saif M. Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat, editors, *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1094–1098, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1184. URL `https://aclanthology.org/S18-1184`.

Stephen Toulmin, Richard D Rieke, and Allan Janik. *An introduction to reasoning*. 1979.

Stephen E. Toulmin. *The Uses of Argument*. Cambridge University Press, 2 edition, 1958.

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. Llama: Open and efficient foundation language models, 2023a.

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. Llama 2: Open foundation and fine-tuned chat models, 2023b.

Rob van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank. Bleaching text: Abstract features for cross-lingual gender prediction. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2061. URL `https://aclanthology.org/P18-2061`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. Introducing CAD: the contextual abuse dataset. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.182. URL `https://aclanthology.org/2021.naacl-main.182`.

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. Argument quality assessment in the age of instruction-following large language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1519–1538, Torino, Italia, May 2024. ELRA and ICCL. URL `https://aclanthology.org/2024.lrec-main.135`.

Sylvia Walby. Theorising patriarchy. *Sociology*, 23(2):213–234, 1989. doi: 10.1177/0038038589023002004.

Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In Jacob Andreas, Eunsol Choi, and Angeliki Lazaridou, editors, *Proceedings of the NAACL Student*

Research Workshop, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-2013. URL `https://aclanthology.org/N16-2013`.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. In Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault, editors, *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3012. URL `https://aclanthology.org/W17-3012`.

Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.167. URL `https://aclanthology.org/2022.naacl-main.167`.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL `https://openreview.net/forum?id=yzkSU5zdwD`. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. Implicitly abusive language – what does it actually look like and why are we not getting there? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.48. URL `https://aclanthology.org/2021.naacl-main.48`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL `https://aclanthology.org/2020.emnlp-demos.6`.

Samantha Pinson Wrisley. Feminist theory and the problem of misogyny. *Feminist Theory*, 24(2):188–207, 2023. doi: 10.1177/14647001211039365. URL `https://doi.org/10.1177/14647001211039365`.

Rui Xing, Shraey Bhatia, Timothy Baldwin, and Jey Han Lau. Automatic explanation generation for climate science claims. In Pradeesh Parameswaran, Jennifer Biggs, and David Powers, editors, *Proceedings of the 20th Annual Workshop of the Australasian Language Technology Association*, pages 122–129, Adelaide, Australia, December 2022. Australasian Language Technology Association. URL `https://aclanthology.org/2022.alta-1.16`.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*, 2023.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. HARE: Explainable hate speech detection with step-by-step reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.365. URL `https://aclanthology.org/2023.findings-emnlp.365`.

Wenjie Yin and Arkaitz Zubiaga. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7, 2021.

Wenjie Yin and Arkaitz Zubiaga. Hidden behind the obvious: Misleading keywords and implicitly abusive language on social media. *Online Social Networks and Media*, 30:100210, 2022. ISSN 2468-6964. doi: https://doi.org/10.1016/j.osnem.2022.100210. URL `https://www.sciencedirect.com/science/article/pii/S2468696422000143`.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph, 2021. URL `https://arxiv.org/abs/2006.16934`.

Andrea Zaninello and Bernardo Magnini. A smashed glass cannot be full: Generation of commonsense explanations through prompt-based few-shot learning. In Bhavana Dalvi Mishra, Greg Durrett, Peter Jansen, Danilo Neves Ribeiro, and Jason Wei, editors, *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 18–29, Toronto, Canada, June 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.nlrse-1.3. URL `https://aclanthology.org/2023.nlrse-1.3`.

Philine Zeinert, Nanna Inie, and Leon Derczynski. Annotating online misogyny. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.247. URL `https://aclanthology.org/2021.acl-long.247`.

Min-Ling Zhang, Yukun Li, Xu-Ying Liu, and Xin Geng. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12: 191–202, 2017.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675, 2019. URL `http://arxiv.org/abs/1904.09675`.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, page 83–92, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450322577. doi: 10.1145/2600428.2609579. URL https://doi.org/10.1145/2600428.2609579.

Ziming Zhou, Han Zhao, Jingjing Dong, Ning Ding, Xiaolong Liu, and Kangli Zhang. DD-TIG at SemEval-2022 task 5: Investigating the relationships between multimodal and unimodal information in misogynous memes detection and classification. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 563–570, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.semeval-1.77. URL https://aclanthology.org/2022.semeval-1.77.

Fangwei Zhu, Jifan Yu, Hailong Jin, Lei Hou, Juanzi Li, and Zhifang Sui. Learn to not link: Exploring NIL prediction in entity linking. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10846–10860, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.690. URL https://aclanthology.org/2023.findings-acl.690.

Ron Zhu. Enhance Multimodal Transformer With External Label And In-Domain Pretrain: Hateful Meme Challenge Winning Solution. *CoRR*, abs/2012.08290, 2020. URL https://arxiv.org/abs/2012.08290.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In Sheng Li, Maosong

Sun, Yang Liu, Hua Wu, Kang Liu, Wanxiang Che, Shizhu He, and Gaoqi Rao, editors, *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China, August 2021. Chinese Information Processing Society of China. URL https://aclanthology.org/2021.ccl-1.108.

Appendix A

Prompts for EDOS Task C

This appendix contains the 3 prompts used for the Task C experiments of the EDOS shared task: the *all_categories* prompt (Figure A.1), the *subcategories* prompt with two examples (Figure A.2) and the *subcategories* prompt with one example (Figure A.3).

```
Predict the category for the last message based on the category
types in the examples.
Choose one among the following categories:
4.1 supporting mistreatment of individual women
2.3 dehumanising attacks & overt sexual objectification
2.2 aggressive and emotive attacks
1.2 incitement and encouragement of harm
4.2 supporting systemic discrimination against women as a group
1.1 threats of harm
3.1 casual use of gendered slurs, profanities, and insults
3.3 backhanded gendered compliments
3.4 condescending explanations or unwelcome advice
2.1 descriptive attacks
3.2 immutable gender differences and gender stereotypes

Examples:
###
Message: <message_1>
Category: <category_1>
###
Message: <message_2>
Category: <category_2>
###
...
###
Message: <message_to_predict>
Category:
```

Figure A.1: Example prompt for the *all_categories* setting using all available categories.

```
Predict the category for the last message based on the
category types in the examples.
Choose one among the following categories:

1.1 threats of harm
1.2 incitement and encouragement of harm

Examples:
###
Message: <message_1>
Category: <category_1>
###
Message: <message_2>
Category: <category_1>
###
Message: <message_3>
Category: <category_2>
###
Message: <message_4>
Category: <category_2>
###
Message: <message_to_predict>
Category:
```

Figure A.2: Example prompt for the *subcategories* setting using two examples per category.

```
Predict the category for the last message
based on the category types in the examples.
Choose one among the following categories:

1.1 threats of harm
1.2 incitement and encouragement of harm

Examples:
###
Message: <message_1>
Category: <category_1>
###
Message: <message_2>
Category: <category_2>
###
Message: <message_to_predict>
Category:
```

Figure A.3: Example prompt for the *subcategories* setting using one example per category.

Appendix B

ImplicIT-Mis Annotation Guidelines

Annotators were given the following guidelines for the identification of the implied statements in the ImplicIT-Mis dataset:

Read the following comments extracted from Facebook as a reaction to a news article or a post. The comments have been judged as being sexist or misogynistic by an expert. Your task is to provide an explanation for why the comment can be offensive towards women. The explanation must correspond to the implied assumption in the comment. To help you in this task, imagine that your explanations should complete the sentence: "This comment is misogynous/sexist because it implies that ..."

Follow the examples:
Le dovrebbero dare 2000 volts implica che la donna in questione dovrebbe essere uccisa sulla sedia elettrica.
Non dovremmo abbassare i nostri standard solo per assumere più donne implica che le donne sono meno qualificate degli uomini.
Quella mangia kinder bueno non meritava di vincere X-Factor implica che la donna in questione viene offesa in quanto grassa.
Bau bau implica che la donna in questione viene chiamata cagna.

Translated version:
She should be given 2000 volts implied that the targeted woman should be killed in the electric chair.
We shouldn't lower our standards just to hire more women implies that women are less qualified than men.
That Kinder Bueno eater did not deserve to win X-Factor implies that the targeted woman is called 'fat'.
Wof wof implies that the targeted woman is called 'cagna', translated as female dog, which is an epithet for b**ch.

Appendix C

Prompts for Generating Implied Assumptions

Table C.1 shows the prompts used in our experiments for English, while Table C.2 shows those for Italian.

Shot	Set.	Prompt Details
ZS	B.	Is this text misogynous? Answer 1 for yes and 0 for no.
FS	B.	Is the text misogynous? Answer only with YES or NO. Answer like in the examples. Text: we shouldn't lower our standards just to hire more women. Answer: YES Text: But since only women get pregnant, they get free birth control pills. Answer: NO Text: I can't stand women talking about football. Answer: YES Text: Lmao I'm down with women replacing God. The future is female Answer: NO Text: She must be on her period. Answer: YES Text: Dude, girls want sex as much as guys do. They have triggers, so do we. Answer: NO Text: {text} Answer:
ZS	A.	What are the implied assumptions of the text {text}? Based on the implied assumptions, is the text misogynous? Answer only with 1 for yes and 0 for no. Answer between the tags. Implied assumption: …Answer: …
FS	A.	What are the implied assumptions of the text {text}? Based on the implied assumptions, is the text misogynous? Answer only with YES or NO. Answer like in the examples. Text: we shouldn't lower our standards just to hire more women. Implied assumption: women are less qualified than men. Answer: yes. Text: Women shouldn't talk about football. Implied assumption: women cannot understand football, it's a male thing. Answer: YES. Text: She must be on her period. Implied Assumption: women are irrational or overly emotional during their periods. Answer: YES. Text: …Implied assumption: …Answer: …
ZS	T.	What is the claim and the implied warrant of the text {text}? Explain step by step. Based on the implied warrant, is the text misogynous? Answer only with 1 for yes and 0 for no. Warrant: …Misogyny: …
FS	T.	What are the claim and the implied warrant in the input text? Explain step by step. Based on the implied warrant, is the text misogynous? Answer only with YES or NO. Follow the example. Text: we shouldn't lower our standards just to hire more women. Claim: hiring more women would lower our standards. Warrant: women are less qualified than men. Answer: YES Text: I can't stand women talking about football. Claim: women should not talk about football. Warrant: football it's for men. Answer: YES Text: She must be on her period. Claim: She must be on her period. Warrant: women are irrational or overly emotional during their periods. Answer: YES. Text: {text} Claim: …Warrant: …Answer: …

Table C.1: English prompts used in our experiments (B. = Baseline, A. = Assumption, T. = Toulmin).

Shot	Set.	Prompt Details
ZS	B.	Questo testo è misogino? Rispondi solo con 1 (sì) o 0 (no)
FS	B.	Questo testo è misogino? Rispondi solo con 1 (sì) o 0 (no) come nell'esempio. Testo: quella mangia kinder bueno non meritava di vincere X-Factor. Risposta: 1. Testo: Lei è acida perché non ha figli penso che darebbe fastidio a qualsiasi donna. Che schifo. Risposta: 0 Testo: le donne che parlano di calcio non si possono sentire. Risposta: 1 Testo: Comunque le pringles più buone sono quelle alla panna acida e cipolla. Risposta: 0 Testo: Dopo aver visto (per sbaglio) 5 minuti di Uomini e Donne ho deciso che odierò le donne. Risposta: 1 Testo: pur abitando in culo alla balena, sono andata ad un bel po' di concerti e sono ricordi preziosi che conservo con affetto. Risposta: 0 Testo: {text}. Risposta: ...
ZS	A.	Quali sono gli assunti impliciti del testo {testo}? Sulla base degli assunti impliciti, il testo è misogino? Rispondere solo con SÌ o NO.
FS	A.	Quali sono gli assunti impliciti del testo {testo}? Sulla base dei presupposti impliciti, il testo è misogino? Rispondere solo con SÌ o NO. Seguire l'esempio. Testo: non dovremmo abbassare i nostri standard solo per assumere più done. Presupposto: le donne sono meno qualificate degli uomini. Risposta: SÌ. Testo: le donne che parlano di calcio non si possono sentire. Presupposto: le donne non capiscono niente di calcio, è una cosa da maschi. Risposta: SÌ. Testo: Dopo aver visto (per sbaglio) 5 minuti di Uomini e Donne ho deciso che odierò le donne. Presupposto: Le donne che vanno alla trasmissione Uomini e Donne sono stupide. Risposta: SÌ. Testo: {testo} Presupposto: ... Risposta: ...
ZS	T.	Quali sono il claim e il warrant implicito del testo? Spiegalo passo dopo passo. In base al warrant implicito, il testo è misogino? Rispondi solo con 1 per il sì e 0 per il no. Warrant: ... Risposta: ...
FS	T.	Quali sono il claim e il warrant implicito nel testo? Spiegalo passo per passo. In base al warrant implicito, il testo è misogino? Rispondere solo con SÌ o NO. Segui l'esempio. Testo: non dovremmo abbassare i nostri standard solo per assumere più done. Affermazione: assumere più donne abbasserebbe i nostri standard. Warrant: le donne sono meno qualificate degli uomini. Risposta: SÌ Testo: Non sopporto che le donne parlino di calcio. Affermazione: le donne non dovrebbero parlare di calcio. Warrant: il calcio è per gli uomini. Risposta: SÌ Testo: Deve avere il ciclo. Affermazione: deve avere le mestruazioni. Warrant: le donne sono irrazionali o eccessivamente emotive durante il ciclo mestruale. Risposta: SÌ Testo: {testo}. Affermazione: ... Warrant: ... Risposta: ...

Table C.2: Italian prompts used in our experiments.

Appendix D

Additional Classification Results

Table D.1 reports classification results when considering the refusal to answer due to model safeguard trigger to hateful content as misogynous. In particular, `Llama3-8B` is the only affected model in our experiments.

Exp. Setting	Model	ImplicIT-Mis	SBIC+
fine-tuning	`bert-hateXplain`	–	0.342
	`ALBERTo`	0.380	–
zero-shot	`Llama3-8B`	0.588	0.609
	`Mistral-7B-v02`	0.050	0.319
few-shot	`Llama3-8B`	**0.738**	<u>0.827</u>
	`Mistral-7B-v02`	0.259	0.416
zero-shot w. implied assumption	`Llama3-8B`	0.542	<u>0.891</u>
	`Mistral-7B-v02`	0.050	0.259
few-shot w. implied assumption	`Llama3-8B`	0.480	<u>**0.914**</u>
	`Mistral-7B-v02`	0.461	0.685
zero-shot Toulmin warrant	`Llama3-8B`	0.557	<u>0.643</u>
	`Mistral-7B-v02`	0.346	0.374
few-shot Toulmin warrant	`Llama3-8B`	0.725	<u>0.841</u>
	`Mistral-7B-v02`	0.556	0.604

Table D.1: Overview of the results of the experiments on ImplicIT and SBIC+. Best results are in bold, while performance differences with respect to 6.4 are underlined. Answer considered valid with implied assumption/Toulmin's warrant only if the model generates the implied assumptions/warrants.

Appendix E

Prompts for Social Dynamics Selection

Table E.1 shows the English and Italian prompts used for the identification of social dynamics and the accompanying text span.

Lang	**Prompt**
EN	Choose the social dynamics implied in the text provided between triple quotes. Report also the part of the text that triggered your choice. Do not provide further explanation. Choose the social dynamics from the options: victim blaming, derogatory treatment or belittling of emotions, male-dominated power structure, expectations with respect to beauty standards, conservative view that limits women's freedom, mockery, stereotyping and generalization, whataboutism, double standards, aggressive and violent attitude, dismissal of feminism or neo-sexism, sexual objectification, centrality of gender distinction, unfounded assumptions or prejudice. Answer in the format [social dynamic: "text span"]
IT	Scegli le dinamiche sociali implicite nel testo tra virgolette triple. Riporta la parte di testo responsabile della tua scelta. Scegli una o più tra le seguenti opzioni: colpevolizzazione della vittima, trattamento dispregiativo o sminuente delle emozioni, struttura di potere dominata dagli uomini, aspettative rispetto agli standard di bellezza, visione conservatrice che limita la libertà delle donne, derisione, stereotipi e generalizzazioni, benaltrismo, due pesi due misure, atteggiamento aggressivo e violento, rifiuto del femminismo o del neosessismo, oggettificazione sessuale, centralità della distinzione di genere, ipotesi infondate o pregiudizi Non fornire ulteriori spiegazioni. Rispondi nel formato [dinamica sociale: "parte di testo"]

Table E.1: Prompts for social dynamics identification.

Appendix F

Measuring Text Spans across Models

Table F.1 reports similarity scores across models of the spans identified by the models associated to predicted social dynamics on English and Italian datasets.

	English			Italian		
Model Pair	Ovl.	BERT-F1	BLEU	Ovl.	BERT-F1	BLEU
LlaMa2-Falcon	0.045	0.797	0.915	0.075	0.778	2.373
LlaMa2-LlaMa3	0.092	0.772	14.327	0.184	0.750	6.316
LlaMa2-Mistral	0.124	0.837	6.299	0.178	0.817	4.838
LlaMa2-Mistral2	0.034	0.850	9.442	0.068	0.821	6.101
LlaMa2-Tower	0.026	0.758	3.524	**0.277**	0.786	7.514
LlaMa2-LLaMAntino	-	-	-	0.0	0.797	1.859
LlaMa3-Falcon	0.078	0.661	2.269	0.067	0.599	3.660
LlaMa3-LlaMa2	0.072	0.772	13.641	0.079	0.750	9.891
LlaMa3-Mistral	0.066	0.800	26.127	0.056	0.767	**14.127**
LlaMa3-Mistral2	0.047	0.845	**36.232**	0.069	0.740	11.889
LlaMa3-Tower	0.091	0.700	15.741	0.067	0.734	10.053
LlaMa3-LLaMAntino	-	-	-	0.030	0.667	0.439
Mistral-v1-Falcon	0.073	0.808	1.924	0.093	0.771	1.930
Mistral-v1-Llama2	0.127	0.837	5.582	0.163	0.817	4.945
Mistral-v1-Llama3	0.154	0.800	24.741	0.131	0.767	10.968

Mistral-v1-Mistral-v2	**0.330**	**0.868**	18.076	0.192	**0.825**	8.458
Mistral-v1-Tower	0.230	0.783	7.874	0.242	0.800	7.101
Mistral-v1-LLaMAntino	-	-	-	0.0	0.803	0.202
Mistral-v2-Falcon	0.053	0.804	1.803	0.067	0.790	1.632
Mistral-v2-LlaMa2	0.039	0.850	8.492	0.25	0.821	6.238
Mistral-v2-LlaMa3	0.219	0.845	35.587	0.161	0.740	10.104
Mistral-v2-Mistral-v1	0.328	**0.868**	19.185	0.186	0.825	8.967
Mistral-v2-Tower	0.213	0.789	11.097	0.148	0.805	8.649
Mistral-v2-LLaMAntino	-	-	-	0.0	0.798	0.600
Falcon-Llama2	0.0	0.797	0.468	0.073	0.780	2.436
Falcon-Llama3	0.100	0.661	2.002	0.168	0.599	1.382
Falcon-Mistral-v1	0.012	0.808	1.801	0.94	0.77	1.796
Falcon-Mistral-v2	0.003	0.804	1.720	0.068	0.790	1.581
Falcon-Tower	0.013	0.745	2.211	0.109	0.784	4.025
Falcon-LLaMAntino	-	-	-	0.0	0.77	0.664
Tower-Llama2	0.026	0.758	2.834	0.277	0.786	6.978
Tower-Llama3	0.083	0.700	13.830	0.156	0.734	8.335
Tower-Mistral-v1	0.233	0.783	7.504	0.094	0.800	7.934
Tower-Mistral-v2	0.212	0.788	9.806	0.068	0.805	8.054
Tower-Falcon	0.065	0.744	2.110	0.109	0.784	3.791
Tower-LLaMAntino	-	-	-	0.0	0.788	0.419
LLaMAntino-Falcon	-	-	-	0.0	0.774	0.147
LLaMAntino-LlaMa2	-	-	-	0.0	0.796	0.174
LLaMAntino-LlaMa3	-	-	-	0.076	0.667	0.081
LLaMAntino-Mistral-v1	-	-	-	0.0	0.804	0.239
LLaMAntino-Mistral-v2	-	-	-	0.0	0.798	0.064
LLaMAntino-Tower	-	-	-	0.0	0.788	0.460

Table F.1: Span evaluation across models for English and Italian - scores reports F1 for character-level overlap, BERTScore F1, and BLEU. Best scores across all metrics are marked in bold.