Language is Scary when Over-Analyzed: Unpacking Implied Misogynistic Reasoning with Argumentation Theory-Driven Prompts

Anonymous ACL submission

Abstract

We propose misogyny detection as an Argumentative Reasoning task and we investigate the capacity of large language models (LLMs) to understand the implicit reasoning used to convey misogyny in both Italian and English. The central aim is to generate the missing reasoning link between a message and the implied meanings encoding the misogyny. Our study uses argumentation theory as a foundation to form a collection of prompts in both zero-shot and few-shot settings. These prompts integrate different techniques, including chain-of-thought reasoning and augmented knowledge. Our findings show that LLMs fall short on reasoning capabilities about misogynistic comments and that they mostly rely on their implicit knowledge derived from internalized common stereotypes about women to generate implied assumptions, rather than on inductive reasoning.

1 Introduction

According to the 7^{th} Monitoring Round of the EU Code of Conduct on Countering Illegal Hate Speech Online,[1] Social Media are slowing down the removal of hateful content within 24 hours, dropping to 64% from 81% in 2021. The prevalence of hate speech phenomena has become a factor of polarization and pollution of the online sphere, creating hostile environments that perpetuate stereotypes and social injustice.

Previous work on hate speech detection from the NLP community has contributed to definitions (Fortuna et al., 2020; Pachinger et al., 2023; Korre et al., 2023), datasets (Chiril et al., 2020; Pamungkas et al., 2020; Guest et al., 2021; Zeinert et al., 2021), and systems (Caselli et al., 2021a; Lees et al., 2022). However, most of these contributions have focused (more or less consciously) on explicit forms of hate. Recently, there has been an increasing interest in the study of implicit realization of hate speech phenomena (Caselli et al., 2020; Wiegand et al., 2021; Ocampo et al., 2023).

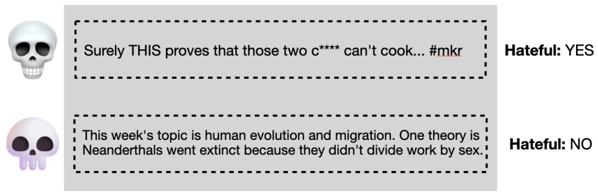

Figure 1: Results from `bert-hateXplain` model for explicit (💀) *vs.* implicit (💀) misogynous messages.

Implicit hate speech is more elusive, difficult to detect, and often hidden under apparently innocuous language or indirect references. These subtleties present a significant challenge for automatic detection because they rely on underlying assumptions that are not explicitly stated. As illustrated in Figure 1, the `bert-hateXplain` model[2] correctly mark as hateful the explicit message (💀), but it fails with the implicit ones (💀). To correctly spot the implicit message, the system would have to identify at least the implied assumptions that assume that "*women aren't as capable as men.*" and "*women should be told what to do*".[3]

In this contribution we investigate the abilities of large language models (LLMs) to correctly identify implicit hateful messages expressing misogyny in English and Italian. In particular, we explore how prompts informed by Toulmin's Argumentation Theory (Toulmin et al., 1979) are effective in reconstructing the *warrant* needed to make the content of the messages explicit and thus facilitate their identification as hateful messages (Kim et al., 2022). By prompting LLMs to generate such warrants, we further investigate whether the generated

[1] https://bit.ly/3yIRYWg

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

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

texts are comparable to those of human annotators, thus offering a fast and reliable solution to enrich hateful datasets with explanations and contributing to improve the generalization abilities of trained tools. We summarize our contributions as follows:

- We present a novel formulation of implicit misogyny detection as an Argumentative Reasoning task, centered on reconstructing implicit assumptions in misogynous texts (§3).
- We introduce the first dataset for implicit misogyny detection in Italian (§4).[4]
- We carry out an extensive set of experiments with two state-of-the-art instruction-tuned LLMs (`Llama3-8B` and `Mistral-7B-v02`) on English and Italian datasets (§5).
- We conduct an in-depth qualitative analysis of the automatically generated implicit assumptions against 300 human-generated ones (§6).

2 Related Work

Hate speech detection is a widely studied research area, covering different targets and linguistic aspects. We discuss literature work on implicit misogyny detection with particular attention to contributions in reconstructing implicit content.

Implicit Hate Speech Detection Hate Speech Detection is a popular research domain with more than 60 datasets covering distinct targets (e.g., women, LGBTIQ+ people, migrants) and forms of hate (e.g., sexism, racism, misogyny, homophobia)in 25 languages, according to the Hate Speech Dataset Catalogue.[5] In its early stages, but still predominant nowadays, research in this domain focused on developing datasets for detecting explicit cues of hate speech, like messages containing slurs or swear words (Jahan and Oussalah, 2023). However, hate speech is often implicit, characterized by the presence of code language phenomena such as sarcasm, irony, metaphors, circumlocutions, and obfuscated terms, among others (Waseem et al., 2017; Wiegand et al., 2021). For this reason, implicit hate speech detection has progressively gained 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). A relevant feature of these datasets is the presence of implied statements in free-text format, which contributes to explaining the content of hate speech messages. While the use of these statements has been shown to have a positive effect on classification performance (Kim et al., 2022, 2023), few efforts have been put in automatically generating such implied assumptions (ElSherief et al., 2021). As Yang et al. (2023) point out, current annotation schemes in this area present significant reasoning gaps between the claim and its implied meaning. Moreover, no effort has been made to evaluate widely adopted LLMs on their reasoning capabilities required to generate high-quality implied assumptions. To the best of our knowledge, our work is the first study to propose an empirical evaluation of LLMs for implicit misogyny detection and the generation of explanations for Italian and English. Available datasets targeting misogyny in Italian (Fersini et al., 2018, 2020) are highly biased toward explicit messages, with very few messages that qualify as implicit. To fill this gap, we have developed the first Italian dataset for this task, ImplicIT-Mis. In our work, we define misogyny 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), going beyond the simplistic definition of misogyny as hate against women.

Implied Assumptions Generation The implied assumptions instantiate statements that are presupposed by the implicit hate speech message. This can be seen as the elicitation of implicit knowledge, corresponding to new content semantically implied by the original message (Srikanth and Li, 2021; Zaninello and Magnini, 2023). Although limited, previous work on the generation of implied meanings —usually in the forms 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.,

[4] All data will be released via a Data Sharing Agreement.
[5] `https://hatespeechdata.com`

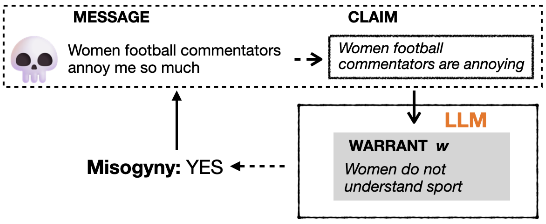

Figure 2: Example of a warrant (implicit logical connection) for an implicit misogynous message. Example and warrant are extracted from SBIC (Sap et al., 2020).

2020; Fang and Zhang, 2022; Yang et al., 2023). In this work, we follow the idea of using LLMs to identify the implied assumptions in the implicit messages, but rather than centering the reasoning process on identifying specific concepts, we formulate the problem as an Argumentative Reasoning task and apply Toulmin's Argumentation Theory (Toulmin, 1958).

3 Misogyny Detection as Argumentative Reasoning Understanding

The elusiveness of implicit hate speech is due to its ambiguity. Implicit messages could be understood as critiques, opinions, or statements (see Figure 1) rather than as hateful. Hate, in this case, is expressed by assuming social biases, stereotypes, and prejudices against a specific target, women in the case of misogyny. 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, the 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), we frame implicit misogyny detection as an AR task (Habernal et al., 2018a) based on the 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. More formally, let c be the claim associated to a given message and $W = \{w_1, \dots w_n\}$ be a set of possible warrants, i.e., logical statement(s) that support c. The model must generate an associated w and, based upon it, provide the requested classification: whether the message is misogynous or not.

Figure 2 graphically represents the approach described above. In this particular case, the generalization that women do not understand sport because it is stereotypically for men is what distinguishes a personal attack from a case of 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. To the best of our knowledge, our work is the first to assess LLMs on implicit misogyny through the lens of AR.

4 Data

This section introduces the datasets used in our experiments. For Italian, the newly created ImplicIT-Mis corpus (§ 4.1). For English, SBIC+, an extended version of the SOCIAL BIAS INFERENCE CORPUS (Sap et al., 2020) enriched with misogynous texts from IMPLICIT HATE CORPUS (ElSherief et al., 2021) (§ 4.2).

4.1 The ImplicIT-Mis Corpus

ImplicIT-Mis is a new manually collected and curated dataset for implicit misogyny detection in Italian. It consists of 1,120 Facebook comments as direct replies to either women-related news articles or posts on public pages of communities known to tolerate misogyny. An in-domain expert, who has been the target of misogyny, conducted the manual collection.[6] 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 phe-

[6]The annotator is also an author of this paper.

nomena (Caselli et al., 2021b; Abercrombie et al., 2023). For each comment, we keep source (either a newspaper or a Facebook page) and its context of occurrence (the news article or the main post). All instances in ImplicIT-Mis are misogynistic.

The collection period ran from November 2023 to January 2024. We selected 15 Facebook pages of news outlets covering the whole political spectrum as well as different levels of public outreach (national *vs.* local audiences), and 8 community pages. 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 Messagero*). The full overview is in Appendix A. On average, each comment is 19 tokens long, with the longest having 392 tokens 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.

ImpliciIT-Mis is enriched with one annotation layer targeting the implied assumptions, as defined in §2. A subset of 150 messages was annotated by three Italian native speakers who are master students in NLP. Each annotator has worked on 50 different messages. On average, annotators took 2 hours to complete the task. The annotation guidelines for the generation of the implied assumptions are in Appendix B. We evaluated the annotators' implied assumptions against those of an expert (a Master student in gender studies and criminology). We used 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 insight into how similar the human written implied assumptions are. We have obtained a BLEU score of 0.437 and an F1-BERTScore of 0.685 by combining all annotations. As the scores indicated, 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 generated implied assumptions have been retained as valid, meaning that for 150 messages, we have a total of 225 implied assumptions.

4.2 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 like 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. We retain all messages whose annotation for the target group was "women" or "feminists" and were labelled as hateful. We further cleaned the data from instances labeled as targeting women but were actually targeting other categories, like gay males. We also filtered out all texts containing explicit identity-related slurs to keep only implicit instances. For each message, we also retained all associated "target stereotype" which correspond to the warrants.

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 the associated implied statement (i.e., the warrant in our framework). We have selected only tweets whose target demographic group was "women".

5 Experimental Setup

Our main goal is to evaluate the abilities of models to generate the implied assumptions for implicit misogynous messages. By doing so, we can also evaluate the implicit knowledge of LLMs, for instance, named entities or events mentioned in texts. If they are not known, it would be impossible to understand the misogynistic nature of such texts.

Each batch of experiments aims to address two tasks: (i) the generation of the implied assumptions or warrants and (ii) the classification of the messages as misogynous or not. Regarding (i), we experiment with two prompting strategies: instructing the model to reconstruct the implied assumptions (**Assumption**) and the implicit claim c and related warrants W (**Toulmin**). We 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 we experiment with both.

We experiment with two state-of-the-art LLMs: `Llama3-8B` and `Mistral-7B-v02`.[7] For both, we select their instruction-tuned version. During preliminary experiments with 50 instances, we also tested Italian-specific LLMs, namely `LlaMantino`, `Fauno`, and `Camoscio`. They were all unable to generate valid implied assumptions, so we discarded them. We consider the following baselines: (i) fine-tuned encoder-based models; (ii) zero-shot classification with LLMs; and (iii) few-shot classification with LLMs without generating explanations.

`Llama3-8B` The Llama3 series has several improvements over preceding versions, including a better tokenizer with a vocabulary of $128k$ tokens, extended training on $15T$ tokens, and grouped query attention for efficiency. Around 5% of the pre-training data concerns more than 30 languages, including Italian. All Llama3 models have undergone safety fine-tuning for safeguarding the generation process over harmful content. This could trigger instances of over-safety, with the model being unable to follow the instructions and thus failing to provide a valid answer for our task.

`Mistral-7B-v02` A competitive fine-tuned version of Llama2 using group-query attention, developed by MistralAI (Jiang et al., 2023). In particular, the 7B version has been reported to obtain better performances when compared to `Llama2-7B` and `Llama2-13B`. While details about the fine-tuning data are lacking, in our experiments, we observe that the model is responsive to Italian prompts. The instruct-based versions of the models do not present any moderation mechanism. We thus expect this model to avoid over-safety and always return an implied statement and a classification value.

5.1 Prompting Techniques

Among recent prompting techniques, we selected **Chain-of-Thought** (CoT) and **Knowledge Augmentation**. CoT was 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 our goal of generating implicit components of arguments (implicit warrants) and support the construction of encoded warrant blocks. 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 can be found in Appendix C and D. 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 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 in our specific task.

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 we inject into the prompt to inform the classification.

6 Results

We 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, we 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), we report

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

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

Setting	Model	BERTScore EN	BERTScore IT	BLEU EN	BLEU IT
Assumption					
zero-shot	`Llama3-8B`	0.820	-	0.201	-
few-shot	`Llama3-8B`	**0.830**	-	0.744	-
	`Mistral-7B-v02`	0.823	**0.601**	0.361	0.240
Toulmin					
zero-shot	`Llama3-8B`	0.817	0.570	0.543	0.104
	`Mistral-7B-v02`	0.812	0.579	0.303	0.077
few-shot	`Llama3-8B`	0.817	0.570	**0.871**	0.261
	`Mistral-7B-v02`	0.813	**0.601**	0.396	**0.313**

Table 2: Automatic evaluation metrics for the best models generating implied assumptions/warrants (selection based on classification results).

the results using established automatic metrics (i.e. BERTScore and BLEU) as well as a manual validation on a subset of 300 messages (150 per language) (§ 6.2). The overall evaluation procedure we 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". We considered 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 E includes the results considering these cases as correct.

6.1 Classification Results

Table 1 summarizes the results for the classification task. With few exceptions - mostly related to `Mistral-7B-v02` - LLMs generally perform better than finetuned 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 (R=0.725). For English, on the other hand, the results are affected by the post-processing heuristics. Had we considered as correct the "refusal to answer cases", the best score for English would have resulted in `Llama3-8B` few-shot with implied assumption (R=0.913).

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, we observe a dramatic increase when switching from implied assumptions to Toulmin's warrant, with a performance gain of 24 points. On the contrary, on English, the warrant-based prompt falls behind.

6.2 Implied Assumptions and Warrants Generation

Table 2 gives an overview of the evaluation using BERTScore and BLEU for the best models for English and Italian. While for SBIC+ every message has an associated explanation, for ImplicIT-Mis, only 150 messages present the implied assumptions. 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, generates a response. In all the other cases, the model just answers the final question of whether it is misogynistic; therefore, we exclude them from the evaluation. We also exclude all the results that do not reach at least a recall of 0.3 due to their low quality, as confirmed by manual inspection. All BERTScores in English are around 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.

6.3 Manual Validation

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

Overall, we find that 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 leading to a wrong classification 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, we conclude that a correct explanation does not necessarily lead to a correct classification of misogyny, and this is always the case in the subsample we 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, requires much more reasoning to be understood.

We design a taxonomy to regroup all errors for both models. We identify seven kinds of common errors in warrant and implied assumption generation. Table 3 provides some examples. Notice that, although all error categories lead to wrong implied assumptions/warrants, we decide to keep a general "wrong inference" as a valid category for all the cases that do not fall under any other category or there is no evident reason.

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: *good girls don't get raped.*	
Generated: *Women have control over whether or not they get raped.*	
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 3: Error categories in warrant generation. For each category, we report an input example, the corresponding LLM generation, and the category's distribution in Italian and English evaluation samples.

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. We have observed a much more frequent occurrence in Italian - also because many messages use figurative or metaphorical expressions. As observed by Muti et al. (2024), misogyny in Italian is highly metaphorical, especially with references 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. (2024), 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 non-standard language, such as dialects and jargon expressions.

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 support the message. These errors occur in both languages, with a slight preference for 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 due to the model failing to identify the mentioned individuals (and their associated characteristics) that are used to trigger the misogynous content. In the example we report in Table 3, the model does not recognize (Moana) Pozzi as a famous porn actress, thus resulting in a warrant that fails to capture the insulting nature of the message (i.e., the target being promiscuous).

Wrong Inference This is the largest class of errors in both languages. We observe that wrong inferences are mainly driven by spurious correlations and the activation of implicit knowledge.

6.4 How Reasoning Impact Classification

We further examine how reconstructed implicit assumptions relate to classification predictions for misogyny detection.

We inspect 50 instances for each language that are correctly classified by our best baseline (few-shot `Llama3-8B`) and that are subsequently misclassified when the model is asked to reason about them. We 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"* [Today you are happy anorexic liar], the model responds that *"Anorexia is a negative condition. 3. 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 genres, 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 genres is detrimental for the classification.

7 Conclusion

We proposed the task of implicit misogyny detection under an Argumentative Reasoning perspective, since to understand implicit statements, one needs to reconstruct the missing link (the warrant) between the claim and the assumption. Our work highlights the complexity of such a task, which paves the way for hate speech detection as a proxy task to probe the reasoning abilities of LLMs. Our prompt-based experiments show that LLMs fail 68% and 50% of the time in generating implied assumptions in Italian and English respectively. 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 skills. Our results 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 (implied assumptions/warrants) is wrong, due to lack of references, reasoning skills, or understanding of non-standard language. Indeed, verifying the validity of the generated text before injecting it in the prompt in a human-in-the-loop approach would be a next step to undertake. To conclude, our findings show that *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. Therefore, models for hate-related natural language inference tasks should be improved. One possible approach would be to inject external knowledge in the misogynous texts, in order to fill the gaps related to their lack of implicit knowledge. For instance, had the model known that Moana Pozzi was a porn actress, it would have probably inferred that when a person is compared to her, it is a derogatory way to address that woman.

Limitations

A limitation of our work is the integration of all generated knowledge (implied assumptions/warrants) and we do not evaluate them before using them to inform the classification task. This should be overcome with a human-in-the-loop approach that allows for the verification of the knowledge extracted by LLMs. We did not try to inject only the knowledge that led to a correct classification because of the low correlation between the generated implied statement and the class. Another limitation is that for what concerns Llama, many examples in English trigger the safeguard, therefore the scores for Llama might not be realistic.

8 Ethical Considerations

Improving LLMs abilities to understand the implied meaning of messages with sensitive content is a case of potential risks related to dual use. Although our work has focused on assessing LLMs abilities in generating implied assumptions/warrants, we see the benefits and the detrimental effects. On the one hand, improving LLMs abilities to understand the implied meaning of sensitive message can further be used to improve the generation of counter-speech and the development of assistive tools for experts in this area. At the same time, the process can be inverted: malevolent agents can feed models with implied assumptions and generate hateful messages. We are aware of this issue, and we think our work offers the community an opportunity to understand limitations of LLMs that have a not minor societal impact. In addition to this, our work indicates the need to adopt different safeguard methods that are able to capture the core meaning of a message and grounded in different cultures.

References

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

Manuela Barreto and David Matthew Doyle. 2023. Benevolent and hostile sexism in a shifting global context. *Nature reviews psychology*, 2(2):98–111.

Maria Becker, Katharina Korfhage, and Anette Frank. 2020. Implicit knowledge in argumentative texts: An annotated corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2316–2324, Marseille, France. European Language Resources Association.

Filip Boltužić and Jan Šnajder. 2016. Fill the gap! analyzing implicit premises between claims from online debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 124–133, Berlin, Germany. Association for Computational Linguistics.

Leonard Bongard, Lena Held, and Ivan Habernal. 2022. The legal argument reasoning task in civil procedure. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 194–207, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

ZeFeng Cai, Linlin Wang, Gerard de Melo, Fei Sun, and Liang He. 2022. Multi-scale distribution deep variational autoencoder for explanation generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 68–78, Dublin, Ireland. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021a. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. 2021b. Guiding principles for participatory design-inspired natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Aadit Trivedi, and Smaranda Muresan. 2021. Implicit premise generation with discourse-aware commonsense knowledge models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6247–6252, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. 2023. Exploring the potential of large language models in computational argumentation. *CoRR*, abs/2311.09022.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. An annotated corpus for sexism detection in French tweets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1397–1403, Marseille, France. European Language Resources Association.

Pieter Delobelle, Murilo Cunha, Eric Massip Cano, Jeroen Peperkamp, and Bettina Berendt. 2019. Computational ad hominem detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 203–209, Florence, Italy. Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yanbo Fang and Yongfeng Zhang. 2022. Data-efficient concept extraction from pre-trained language models for commonsense explanation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5883–5893, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, Oregon, USA. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, Paolo Rosso, et al. 2018. 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.

Elisabetta Fersini, Debora Nozza, Paolo Rosso, et al. 2020. 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.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Deniz Gorur, Antonio Rago, and Francesca Toni. 2024. Can large language models perform relation-based argument mining? *Preprint*, arXiv:2402.11243.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018a. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *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. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018b. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *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. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, and Minlie Huang. 2020. Generating commonsense explanation by extracting bridge concepts from reasoning paths. In *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. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2024. 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. Curran Associates Inc.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Youngwook Kim, Shinwoo Park, Youngsoo Namgoong, and Yo-Sub Han. 2023. ConPrompt: Pre-training a language model with machine-generated data for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10964–10980, Singapore. Association for Computational Linguistics.

Katerina Korre, Arianna Muti, and Alberto Barrón-Cedeño. 2024. The challenges of creating a parallel multilingual hate speech corpus: An exploration. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15842–15853, Torino, Italia. ELRA and ICCL.

Katerina Korre, John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, Ion Androutsopoulos, Lucas Dixon, and Alberto Barrón-cedeño. 2023. Harmful language datasets: An assessment of robustness. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 221–230, Toronto, Canada. Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Filipa Melo Lopes. 2019. Perpetuating the patriarchy: Misogyny and (post-)feminist backlash. *Philosophical Studies*, 176(9):2517–2538.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *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. Association for Computational Linguistics.

Eleonora Mancini, Federico Ruggeri, and Paolo Torroni. 2024. Multimodal fallacy classification in political debates. In *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. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

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. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12700–12711, Torino, Italia. ELRA and ICCL.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

Pia Pachinger, Allan Hanbury, Julia Neidhardt, and Anna Planitzer. 2023. Toward disambiguating the definitions of abusive, offensive, toxic, and uncivil comments. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 107–113, Dubrovnik, Croatia. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study. *Information Processing & Management*, 57(6):102360.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Paul Reisert, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2015. A computational approach for generating toulmin model argumentation. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 45–55, Denver, CO. Association for Computational Linguistics.

Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. 2021. ExplaGraphs: An explanation graph generation task for structured commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7740, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137, Online. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020. 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. Curran Associates Inc.

Junfeng Tian, Man Lan, and Yuanbin Wu. 2018. ECNU at SemEval-2018 task 12: An end-to-end attention-based neural network for the argument reasoning comprehension task. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1094–1098, New Orleans, Louisiana. Association for Computational Linguistics.

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

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

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. 2024. Argument quality assessment in the age of instruction-following large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1519–1538, Torino, Italia. ELRA and ICCL.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. 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. Curran Associates Inc.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.

Rui Xing, Shraey Bhatia, Timothy Baldwin, and Jey Han Lau. 2022. Automatic explanation generation for climate science claims. In *Proceedings of the 20th Annual Workshop of the Australasian Language Technology Association*, pages 122–129, Adelaide, Australia. Australasian Language Technology Association.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. HARE: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore. Association for Computational Linguistics.

Andrea Zaninello and Bernardo Magnini. 2023. A smashed glass cannot be full: Generation of commonsense explanations through prompt-based few-shot learning. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 18–29, Toronto, Canada. Association for Computational Linguistics.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*

Conference on Natural Language Processing (Volume 1: Long Papers), pages 3181–3197, Online. Association for Computational Linguistics.

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

Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. 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. Association for Computing Machinery.

Fangwei Zhu, Jifan Yu, Hailong Jin, Lei Hou, Juanzi Li, and Zhifang Sui. 2023. Learn to not link: Exploring NIL prediction in entity linking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10846–10860, Toronto, Canada. Association for Computational Linguistics.

A ImplicIT-Mis Sources

Table A shows statistics on the number of Facebook comments associated to each newspaper or Facebook community.

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

Table A: List of sources - newspapers and Facebook pages - with total amount of extracted instance for the creation of the ImplicIT-Mis dataset

B ImplicIT-Mis Annotation Guidelines for Implied Assumptions

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. 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 sexist because it implies that ...". Follow the examples. Le dovrebbero dare 2000 volts implica che la donna in questione dovrebbe essere uccisa. 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. 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.*

C English Prompts

Prompt	Prompt Details
Zero-shot	Is this text misogynous? Answer 1 for yes and 0 for no.
Few-shot	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:
Implied Assumption Zero-shot	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: <...>
Implied Assumption Few-shot	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:
Toulmin CoT Zero-Shot	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:
Toulmin CoT Few-Shot	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: English promts used in our experiments.

D Italian Prompts

Prompt	Prompt Details
Zero-shot	Questo testo è misogino? Rispondi solo con 1 (sì) o 0 (no)
Few-shot	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. Misogino: 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:
Implied Assumption Zero-shot	Quali sono gli assunti impliciti del testo {testo}? Sulla base degli assunti impliciti, il testo è misogino? Rispondere solo con SÌ o NO.
Implied Assumption Few-shot	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 donne. 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:
Toulmin CoT Zero-Shot	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: Misoginia:
Toulmin CoT Few-Shot	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 donne. 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 D: Italian prompts used in our experiments.

E Additional Classification Results

Table E 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 E: Overview of the results of the experiments on ImplicIT and SBIC+. Best results are in bold, while performance differences with respect to 1 are underlined. Answer considered valid with implied assumption/Toulmin's warrant only if the model generates the implied assumptions/warrants.