

Beyond Slurs: A Multilingual Analysis of Misogyny Detection Across Dutch, English, and Spanish

Jirky Dumon, Aaron Maladry and Els Lefever

LT3, Language and Translation Technology Team

Ghent University, Belgium

firstname.lastname@ugent.be

Abstract

The growing volume of gender-based abusive content on social media platforms has increased the need for reliable automatic misogyny detection across languages. While multilingual language models achieve strong overall performance, less is known about how well they handle phenomena such as implicit hate, stereotypes, or humour. In this research, we present a multilingual study of misogyny detection in Dutch, English, and Spanish Twitter data using mmBERT. We introduce a newly collected and manually annotated Dutch dataset and re-annotate English and Spanish HatEval samples using a unified annotation framework. Beyond binary misogyny labels, tweets are annotated for implicitness, aggressiveness, stereotypes, humour, and target type, enabling a fine-grained analysis of model behaviour. Experimental results reveal substantial differences across languages, and show that overall classification performance masks important category-specific weaknesses. While aggressive content is generally detected reliably, implicit misogyny, stereotypes, and humour remain challenging across all three languages. Qualitative analysis further demonstrates that the model over-relies on lexical cues, and struggles with reported speech and stereotype-based misogyny expressed without overtly offensive language. These findings highlight the value of fine-grained multilingual analyses for understanding the limitations of contemporary misogyny-detection systems.

Trigger warning: This paper contains examples of offensive and stereotypical content.

1 Introduction

The scale of social-media communication severely limits the efficacy of purely manual content moderation. Online platforms host large volumes of abusive content, including gender-based harassment, misogyny, and hate speech, creating a growing

need for automated detection systems. Despite considerable advances in natural language processing, reliably identifying misogynistic content remains challenging because harmful messages can be expressed not only through explicit insults and threats, but also through stereotypes, humour, irony, and other indirect forms of abuse. Models that rely heavily on offensive vocabulary may therefore overlook harmful content that lacks overt lexical cues (Wiegand et al., 2021).

A second challenge concerns evaluation. Much existing research on abusive-language detection focuses on overall classification performance, often reported as aggregate precision, recall, and F1-scores. While these metrics provide useful benchmarks, they reveal little about the types of content on which models succeed or fail. In particular, misogyny may be communicated through implicit meaning, gender stereotypes, reported speech, or humour, all of which complicate both annotation and automatic detection. Understanding how models handle these phenomena is essential for assessing their robustness in real-world moderation settings.

Research on abusive-language detection furthermore remains heavily skewed towards English, even though online abuse is a multilingual phenomenon (Abercrombie et al., 2023; Rottger et al., 2022). Although multilingual benchmarks such as HatEval (Basile et al., 2019) and IberEval (Fersini et al., 2018) have expanded research to additional languages including Spanish, Dutch remains comparatively under-explored. This imbalance matters because language-specific vocabulary, humour, stereotypes, contextual understanding, and pragmatic conventions affect both data collection and model behaviour. As a result, model performance may vary substantially across languages, even when the same multilingual architecture is used.

In this paper, we investigate multilingual misogyny detection in Dutch, English, and Spanish Twit-

ter data using mmBert, a multilingual encoder model pre-trained on over three trillion tokens across more than 1,800 languages (Marone et al., 2025). We contribute a newly collected and manually annotated Dutch dataset and re-annotate English and Spanish HatEval samples using a unified annotation framework. In addition to binary misogyny labels, tweets are annotated for implicitness, aggressiveness, stereotypes, humour, and target type, enabling a fine-grained evaluation of model behaviour across languages. Rather than focusing solely on overall classification performance, we examine which discursive categories are associated with model errors and whether these patterns differ across languages. More specifically, we address the following three research questions:

RQ1: To what extent can a multilingual language model detect misogynistic hate speech in Dutch, English, and Spanish tweets?

RQ2: How does misogyny-detection performance vary across discursive categories such as implicitness, aggressiveness, stereotypes, and humour?

RQ3: Which recurrent error patterns characterize multilingual misogyny detection across the three languages?

2 Related Work

2.1 Hate Speech and Misogyny detection

Early hate-speech detection systems relied on lexicons, rules, and traditional machine-learning approaches (MacAvaney et al., 2019; Basile et al., 2019; Bauwelinck et al., 2019). While effective for identifying explicit abuse, these methods often struggled with ambiguity, figurative language, and the evolving vocabulary of online discourse. The introduction of transformer-based models substantially improved contextual modelling and led to strong performance across a wide range of abusive-language detection tasks. As shown in HateBERT, systems for hate speech detection can benefit from continued pre-training on hateful social media data (Caselli et al., 2021a). More recently, multilingual encoder models have sought to extend these advantages across languages, with architectures such as mmBERT specifically targeting robust multilingual representations, including for lower-resourced languages (Marone et al., 2025).

As research into hate speech detection has expanded, misogyny detection has emerged as a distinct research area with dedicated datasets and shared tasks for identifying sexism and misog-

yny in online communication (Fersini et al., 2018; Basile et al., 2019). Although often studied within the broader field of hate speech detection, misogyny extends beyond the mere expression of hostility towards women. Muti et al. (2025) argue that misogyny can be conveyed through both hostile and benevolent reactions directed at women perceived as violating patriarchal norms. Despite advances in modelling architectures, hate speech and misogyny detection remain a sociotechnical classification problem. Definitions of hate speech and misogyny vary across datasets and institutional settings and annotation decisions often depend on contextual and pragmatic factors, such as whether a message attacks women as a group, targets an individual woman, quotes sexist language critically, uses a slur playfully, or expresses gendered stereotypes indirectly. Consequently, strong overall classification performance does not necessarily imply a deeper understanding of misogynistic discourse.

2.2 Multilingualism and Under-resourced Languages

Despite ongoing efforts to broaden linguistic coverage, the distribution of abusive-language resources is highly uneven. In their systematic review of resources for automated identification of online gender-based violence, Abercrombie et al. (2023) identify continuing resource gaps beyond English. Similarly, Rottger et al. (2022) argue that extending hate-speech detection to under-resourced languages requires dedicated data collection rather than relying solely on cross-lingual transfer. One important strategy for reducing these resource gaps has been through shared tasks.

Basile et al. (2019) note that shared tasks have represented an important step towards multilingual evaluation, by providing benchmarks in a series of different languages, including Spanish, Italian, German, Mexican-Spanish and Hindi. Among these initiatives, AMI@IberEval and HatEval are particularly relevant to the English-Spanish setting considered in this work as they were the first to provide multilingual benchmarks enriched with fine-grained annotation dimensions beyond binary classification (Fersini et al., 2018; Basile et al., 2019). More recently, the EXIST (sEXism Identification in Social neTworks) initiative has broadened the study of online sexism through a series of shared tasks in English and Spanish. Beyond binary sexism identification, EXIST includes sexism categorisation, intention detection, and, more recently, the

detection of sexism in memes (Plaza et al., 2024).

Although multilingual benchmarks have significantly improved resource availability, important gaps remain. Dutch, in particular, continues to be comparatively under-represented in misogyny- and sexism-detection research, with the DALC-corpus (Caselli et al., 2021b) constituting one of the few dedicated resources currently available. Moreover, benchmark datasets often differ in collection strategies and annotation practices, complicating direct comparisons across languages. These challenges motivate the development of more comparable multilingual datasets and evaluation frameworks that reveal which types of content are responsible for errors.

2.3 Implicit Hate Speech, Context, and Stereotypes

A growing body of research has shown that harmful online content frequently extends beyond explicit insults and slurs, as perpetrators are aware of content moderation efforts and therefore increasingly resort to more implicit forms of hate speech to disseminate their ideas (MacAvaney et al., 2019). Such content may rely on stereotypes, sarcasm, humour, metaphors, or pragmatic or commonsense reasoning to convey hateful meaning (Van Hee et al., 2018; Maladry et al., 2024). Because the hateful intent of such messages can only be identified through an understanding of their underlying meaning and relevant contextual cues, many instances of abusive content remain difficult to detect (Ocampo et al., 2023). Several studies have shown that incorporating contextual information can positively affect both annotation outcomes and model predictions for content lacking explicit profanities. However, previous work does not consistently show that adding context improves automatic hate speech detection (Gao and Huang, 2017; Pavlopoulos et al., 2020; Menini et al., 2021).

Among the different mechanisms through which implicit hate speech is expressed, misogynistic stereotypes pose a related challenge (Cignarella et al., 2026). Misogyny is not limited to insults, but may be expressed through euphemisms, presuppositions and cultural references, with such subtle forms often being perceived as neutral and socially acceptable despite reproducing discriminatory beliefs about women without using overtly abusive words (Muti et al., 2025).

The complexity of implicit and context-dependent forms of misogyny has motivated inter-

est in explainable detection methods. Because such messages lack abusive markers, interpreting model decisions becomes particularly important. Mohammadi et al. (2024) propose a transparent pipeline for sexism detection that combines transformer architectures with explainability techniques such as SHAP. Their findings demonstrate the value of explainable approaches for identifying linguistic cues used by detection systems and for understanding why particular messages are classified as sexist.

These developments suggest that evaluation should move beyond aggregate binary classification metrics. Fine-grained annotations and discourse-specific evaluation categories can provide a more detailed understanding of model strengths and limitations. Building on this perspective, the present study investigates multilingual misogyny detection through a unified annotation scheme that explicitly captures implicitness, aggressiveness, stereotypes, humour, and target type, enabling both quantitative and qualitative analyses of model behaviour across languages.

3 Data and Annotation

3.1 Dutch data collection

The Dutch dataset was constructed based on the nLT-Tweets corpus (Debaene et al., 2025), a pre-2023 collection of tweets obtained when the Twitter API was more accessible for research. To retrieve candidate tweets, we compiled Dutch profanity lists based on the Spanish and English HatEval lists (Basile et al., 2019), and expanded them with expressions commonly used to insult women online. After exploratory searches, the final retrieval lexicon contained 24 keywords and multiword expressions (see Appendix A). Whole-word matching was used to avoid false matches in neutral words containing offensive substrings.

The search returned 393,566 candidate tweets containing at least one keyword. We then sampled up to 100 tweets per keyword, resulting in a dataset of 1,920 Dutch tweets for manual annotation. This keyword-based strategy intentionally includes both clearly abusive tweets and non-abusive tweets containing profanity, as distinguishing misogyny from non-targeted or non-harmful uses of offensive language is central to the task.

3.2 English and Spanish data

For English and Spanish, we use samples from HatEval, the multilingual SemEval-2019 task on

Language	sampled	Relevant	Relevant $\cap$ HS
Dutch	1,920	1,363	926
English	1,200	827	483
Spanish	1,200	804	544

Table 1: Dataset overview: total number of samples per language, the number of tweets indicated as relevant (Relevant) and the number of tweets that are both relevant and labelled as hate speech. (Relevant $\cap$ HS)

detecting hate speech against immigrants and women (Basile et al., 2019). For each language, 1,200 tweets were sampled, balanced between 600 hate-speech and 600 non-hate-speech examples according to the original labels. The HatEval samples were re-annotated using the same scheme as the Dutch data. Tweets that did not meet the relevance criteria of the present study were excluded rather than automatically treating them as negative examples. This is important because an irrelevant tweet may still contain offensive language, but not be targeted at women. Treating such cases as ordinary negative examples could introduce ambiguity into the classification task. Table 1 gives an overview of the resulting number of instances per language.

3.3 Annotation scheme

All data were manually annotated by one annotator using Label Studio (Tkachenko et al., 2020-2022). Consistency checks included a second annotation pass, written guidelines (Dumon et al., 2026), and discussion of ambiguous cases with researchers experienced in hate speech detection. Figure 1 presents an overview of the annotation framework, which was inspired by previous work on fine-grained analysis of online harms (Nozza et al., 2023). Each tweet was first assessed for relevance. Tweets deemed relevant were subsequently annotated for the presence or absence of misogynistic hate speech. For hateful tweets, the annotator also recorded implicitness on a five-point Likert scale, aggressiveness on a five-point scale, and whether the target was an individual woman or women as a group. Independently of the hate-speech label, all relevant tweets were annotated for stereotypes and humour, with humour also represented on a five-point scale. The five-point scales were used to capture annotator uncertainty, an approach that has been shown to be effective for other subjective classification tasks, such as irony detection (Maladry et al., 2024). For model training and experiments, the five-point labels were merged into binary categories. We leave modelling directly on

the scalar labels to future work and use the scalar annotations only in the error analysis.

The annotation design supports analysis beyond a single binary label. Implicitness captures whether the harmful meaning is directly stated or inferred, while aggressiveness captures the intensity of the expression. The stereotype label records generalisations or gendered expectations, including domestic-role stereotypes. The humour label captures irony, sarcasm, and joking framings, which may complicate interpretation as interpreting a text as ironic or humorous can directly contradict its explicit meaning and potentially create a gap between intended and perceived meaning. These dimensions are not mutually exclusive: a tweet can be humorous, stereotypical, and misogynistic at the same time.

4 Experimental Setup and Results

We fine-tune mmBERT for binary misogyny detection in three different experimental setups. First, each language is evaluated with five-fold cross-validation: four folds are used for training and one for testing, rotating so that each tweet is evaluated once (setup *cv*). Second, the model is fine-tuned on the remaining English and Spanish HatEval misogyny data, excluding re-annotated tweets to prevent overlap between training and evaluation sets (setup *hateval*). Third, the model is first trained on the HatEval training data and then further fine-tuned with the language-specific cross-validation folds (setup *hate-eval+cv*).

Since all of the input texts are short tweets, the maximum length of the tokenizer was limited to 128 tokens, covering the complete length of all tweets in the corpus. The model was fine-tuned up to 5 epochs with default settings, including a learning rate of 5e-05, adamw as optimizer, and epsilon of 1e-08, and a batch size of 8. To avoid over-fitting, 10% of the training data was held out as a development set and early stopping was implemented, stopping the training if the macro F1-score did not improve on the dev set for 2 epochs. For the calculations of precision, recall and F1-score, macro-averaging was applied. Most models already reached their best performance after a few epochs of training, after which the development loss started to increase. The best-performing model (based on dev F1-score) was kept as the final model for each fold. The random seed was set to 666 to ensure reproducibility of the results.

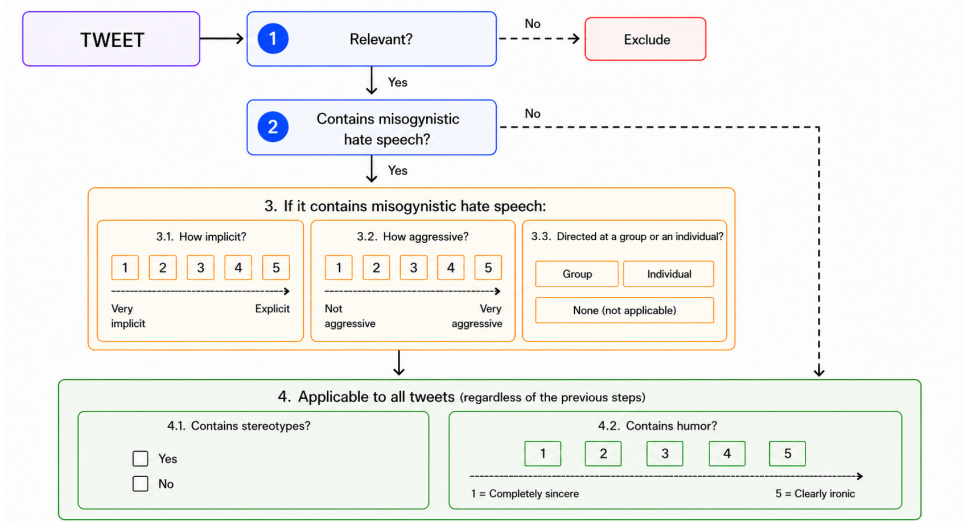


Figure 1: Overview of the annotation framework used for multilingual misogyny detection.

5 Results

5.1 Overall binary classification

Table 2 presents the binary misogyny-detection results, showing that even with a modest training set of about 800 samples, the models can attain decent performance. English achieves its best performance in the *hateval+cv* setting ($F_1 = 0.8651$), indicating that combining HatEval and language-specific training data is beneficial. Spanish obtains very similar results in both *hateval* and *hateval+cv*, suggesting that HatEval alone already provides a strong learning signal. Dutch shows a different pattern: the best performance is obtained with language-specific cross-validation ($F_1 = 0.7677$), while transfer from HatEval is substantially less effective. These results show that cross-lingual resources do not benefit all languages equally and highlight the importance of language-specific annotated data, particularly for Dutch.

5.2 Discursive subcategories

For further analysis, we focus on the predictions provided by the *hateval+cv* setting, which achieved strong results across all languages. Table 3 reports the frequency of four discursive subcategories and the proportion of instances in each category that were misclassified. Since these categories are not mutually exclusive, the table should not be interpreted as a partition of the data. Rather, it identifies phenomena that are associated with higher error rates. **Implicit tweets** produce substantial error rates in all three languages. Spanish has the highest proportion of misclassified implicit tweets

(31%), despite containing fewer implicit instances than Dutch and English. This suggests that the Dutch error patterns cannot be explained by implicitness alone. The distribution of implicitness also differs across languages: Dutch and English contain more implicit examples, whereas the Spanish data are more explicit and contain a higher proportion of aggressive tweets. **Aggressiveness** appears to be the least problematic category. Error rates remain between 6% and 8%, even though Spanish contains far more aggressive tweets than the other two languages. This suggests that overtly aggressive language is easier for the model to identify, likely because it often contains explicit lexical cues. **Stereotypes** are more difficult, especially in Dutch and Spanish. English shows an error rate of 12% for this category, compared with 24% for Dutch and 27% for Spanish. This supports the view that models struggle when misogyny is expressed through gendered expectations or normalised generalisations rather than direct insults. **Humour** is the clearest Dutch-specific difficulty. Dutch has 251 humorous instances, almost twice as many as English and substantially more than Spanish, with 35% of these instances being misclassified. English and Spanish also show elevated error rates for humorous tweets, but the Dutch rate is markedly higher. This indicates that irony and joking framings can obscure or invert literal meaning, requiring pragmatic interpretation that an isolated-tweet classifier may not reliably capture.

Language	Setting	Precision	Recall	F1	Accuracy
EN	hateval+cv	0.8665	0.8639	0.8651	0.8694
	hateval	0.8281	0.8357	0.8194	0.8198
	cv	0.8314	0.8356	0.8331	0.8368
ES	hateval+cv	0.8442	0.8476	0.8459	0.8644
	hateval	0.8379	0.8649	0.8476	0.8607
	cv	0.8326	0.8175	0.8242	0.8495
NL	hateval+cv	0.7281	0.6876	0.6992	0.7592
	hateval	0.5583	0.5494	0.4476	0.4508
	cv	0.7693	0.7664	0.7677	0.7751

Table 2: Binary misogyny-detection performance across languages and experimental settings. The *hateval+cv* setting combines HatEval fine-tuning with language-specific cross-validation, *hateval* uses only HatEval training data, and *cv* uses only language-specific cross-validation.

Category	Language	Category instances	#errors	Error rate
Implicit	Dutch	178	43	24%
	English	149	30	20%
	Spanish	62	19	31%
Aggressive	Dutch	112	7	6%
	English	131	9	7%
	Spanish	475	37	8%
Stereotype	Dutch	220	52	24%
	English	170	20	12%
	Spanish	83	22	27%
Humorous	Dutch	251	87	35%
	English	127	26	20%
	Spanish	95	23	24%

Table 3: Category-level error analysis. Implicit tweets include scale values 1-3; aggressive tweets include scale values 4-5; humorous tweets include scale values 3-5.

5.3 Borderline implicitness and humour

The annotation scales also make it possible to isolate intermediate cases, where the annotator was less certain about the label. For **implicitness**, tweets assigned a score of three on the five-point scale are neither clearly implicit nor clearly explicit. Table 4 shows that these borderline cases are not uniformly the hardest instances. In Dutch and English, intermediate cases have lower error rates than the broader set of implicit tweets, whereas Spanish remains challenging. One possible explanation is that intermediate cases often combine indirect meaning with explicit lexical evidence. In Dutch and English, such evidence may make classification easier than in fully implicit tweets. However, the Spanish subset is small, so its 32% error rate should be interpreted with caution.

For **humour**, mid-scale cases follow a slightly different pattern. Dutch remains the most difficult language, while Spanish falls between Dutch and English. The error rates are close to those for the

Scale subset	Language	<i>n</i>	Errors	Rate
Implicit mid-scale	Dutch	99	20	20%
	English	74	11	15%
	Spanish	22	7	32%
Humour mid-scale	Dutch	107	30	28%
	English	76	13	17%
	Spanish	70	16	23%

Table 4: Reported errors for mid-scale cases on the implicitness and humour scales.

broader humour category, suggesting that humour is challenging not only for the clearly ironic tweets but also when the annotator perceived ambiguity. This is particularly relevant for moderation workflows, since uncertain cases are precisely those for which automatic predictions are least suitable as final decisions. The analysis of tweets assigned the central score on the five-point scale also illustrates a methodological advantage of graded annotation. Binary auxiliary labels would hide whether the model fails mainly on prototypical instances or on ambiguous ones. Graded labels do not eliminate subjectivity, but they make uncertainty visible and create a basis for constructing targeted challenge sets.

6 Qualitative Error Analysis

The quantitative results identify where errors are concentrated, while the qualitative analysis helps explain why these errors occur. The inspected examples reveal several recurring patterns. To avoid unnecessarily reproducing harmful content, we describe these patterns using lightly redacted examples.

6.1 Lexical over-reliance

The model appears to assign substantial weight to gendered insults and profanity. This is unsur-

prising, since such terms are often correlated with misogynistic abuse, and the collection strategy uses offensive keywords. However, this lexical sensitivity also produces false positives when the surrounding context changes the pragmatic function of the word. For example, a positive or playful message containing a term such as “bitch” may be labelled as hate speech even when it is used affectionately or among friends.¹

Example 1

Tweet: *Mijn zus wanneer ze hoorde dat ik afgestudeerd ben met grote onderscheiding: ‘bitch ge hebt de lat echt heel hoog gelegd voor mij nu’ @user thanks x*

In English, a benign message (Example 2) containing “hysterical” may trigger a prediction because the word is sometimes associated with gendered stereotypes.

Example 2

Tweet: *@user you guys nailed it. I’ve had 3 seperate fits of hysterical laughter and I didn’t even make it to the school yet.*

In Spanish, a message explicitly rejecting words (Example 3) such as “puta” or “zorra” can still be misclassified because the slurs remain present in the surface form.

Example 3

Tweet: *Ni perra , ni puta, ni zorra nada de esa mierda LA MUJER ES MUJER Y SE RESPETA ASI TENGA MIL NOVIOS Y SE DIVIERTA CON ELLOS COMO SE LE DE LA HPTA GANA :) (Y)*

This pattern echoes longstanding findings that profanity and hate speech are related but not equivalent (Malmasi and Zampieri, 2017). A moderation system that treats slurs as decisive evidence will over-moderate reclaimed, playful, critical, or quoted uses. This is not merely a technical inconvenience: false positives can suppress counter-speech and distort which users are disproportionately moderated.

6.2 Playful and self-directed insults

Some false positives involve jokes between friends or self-directed expressions. For instance, Example 4 may jokingly use an insult to refer to a friend while describing an everyday interaction (Example 4); and Example 5 may use playful profanity in a celebratory context; and Example 6 may combine an insult with an explicit expression of affection. Other tweets, like Example 7 use offensive terms self-referentially. Under the annotation guidelines,

such cases are not automatically labelled as misogynistic attacks, as they do not clearly intend to disparage another woman or women as a group.

Example 4

Tweet: *RT @user Ik vind @user echt een rotwijf, begint altijd te praten als ik nét ga slapen • Hahahaahah laaaat mij rotzak*

Example 5

Tweet: *Happy halloween! Hope everyone lives up to their full hoe potential tonight! I believe in you, you bust ass super skanks*

Example 6

Tweet: *@user TU ERES ESA AMIGA PUTA JAJAJAJ.*

Example 7

Tweet: *@user Um excuse you I am a twank, a twink and a skank*

These examples demonstrate the importance of pragmatic stance. The same token can function as harassment, teasing, self-deprecation, quotation, or reclamation. Isolated text classification provides only a limited view of these distinctions.

6.3 Reported speech and counter-speech

Reported speech is another major source of ambiguity. Tweets may reproduce misogynistic language in order to describe an incident and condemn it (Example 8), or intensify the abuse (Example 9). A model that detects offensive wording without identifying the author’s stance may classify counter-speech as hate speech or fail to distinguish neutral reporting from endorsement.

Example 8

Tweet: *@user Telemarketer did same to me called me a whore and said he would rape me. Nice.*

Example 9

Tweet: *1) Lo mejor del día, escuchar como mi sobrino le dijo CALLATE PUTA a mi hermana*

This problem is especially relevant for content moderation. Removing a victim’s report of harassment while leaving the original abuse online can compound harm. Context-aware methods may help, but prior research has found mixed results when adding context to toxicity and abuse classifiers (Gao and Huang, 2017; Pavlopoulos et al., 2020; Menini et al., 2021). The key challenge is not simply adding more text; it is modelling who said what, to whom, and with what stance.

6.4 Stereotypes without overt insults

The inverse problem also occurs: when profanity is absent, the model may fail to detect misogyny expressed through stereotypes. Examples include jokes that associate women with kitchen work (Example 10) or statements suggesting that women

¹English translations of the Dutch and Spanish examples can be found in Appendix B.

belong in domestic roles (Example ??). Such tweets can be harmful even when they do not contain explicit slurs. Correctly classifying them requires recognising the gendered expectation being invoked and determining whether the message reproduces or critiques it.

Example 10

Tweet: *@user rape is no joke and something we do not support, the kitchen needs to be cleaned by the way.*

Example 11

Tweet: *user Movimiento SEMEN! A fregar todas! http*

This is particularly challenging because stereotypical messages may appear superficially polite, humorous, or commonplace. They are also culturally variable. In the analysed datasets, Dutch stereotypes often concern behaviour or appearance, Spanish examples frequently invoke domestic roles, and English examples more often take the form of jokes about competence or intelligence. These tendencies are descriptive rather than exhaustive, but highlight the importance of evaluating multilingual models with language-sensitive categories.

6.5 Humour and implicit meaning

In this paper, the term “humour” is used as an umbrella term covering different forms of humorous expression. This category combines several of the challenges discussed above. Irony, for example, may depend on a contradiction between the literal wording of a tweet and commonsense knowledge, a shared cultural script, or an implied stance (Van Hee et al., 2018; Maladry et al., 2024). A short tweet may therefore be impossible to interpret correctly without additional context. The high Dutch humour error rate suggests that this phenomenon deserves more attention in future Dutch-language data collection and modelling.

Importantly, humour should not be treated as a simple binary switch. A joke may be affectionate, self-directed, critical, ambiguous, or harmful. The five-point annotation scale is useful because it exposes borderline cases rather than collapsing them into a single category. The results also show that intermediate humour cases remain challenging, although their error rates differ only modestly from those of the broader humour category.

7 Discussion

7.1 Answering the research questions

The study demonstrates that mmBERT can learn effective misogyny-detection classifiers for Dutch,

English, and Spanish, although performance differs across languages and training configurations. Fine-grained evaluation reveals that aggregate performance masks substantial variation across discursive subcategories: aggressive content is detected relatively reliably, whereas implicit misogyny, stereotypes, and humour remain challenging. Qualitative analysis further shows that errors frequently stem from lexical over-reliance, reported speech, stereotype-based misogyny, and the contextual complexity of humour and irony.

7.2 Implications for dataset design

The Dutch dataset fills an important gap in misogyny-detection research, but also illustrates trade-offs in abuse-data collection. Keyword retrieval is efficient and yields many relevant candidates, yet it over-represents explicit profanity. This helps test whether a model can distinguish misogynistic abuse from non-misogynistic profanity, but may also encourage shortcut learning. Future Dutch data collection should therefore combine keyword retrieval with topic-based, conversation-level, and interaction-based sampling, while addressing privacy and ethical constraints.

The results also show the value of fine-grained annotation. A single binary label cannot reveal whether a model detects misogyny broadly or merely responds to slurs. Annotating implicitness, aggressiveness, stereotypes, humour, and target type makes it possible to identify category-specific weaknesses across languages, which is especially important for under-resourced languages.

The graded annotation scales add another advantage: they make annotator uncertainty visible. Although this study uses scalar labels only for error analysis, future work could model them directly through ordinal classification, probabilistic labels, or multi-task learning.

7.3 Implications for modelling and evaluation

The error analysis suggests several directions for future work. First, evaluation should include targeted challenge sets with profanity-rich non-hate examples, counter-speech, reported speech, stereotypes without slurs, and culturally grounded humour. Such sets would help determine whether models learn robust patterns of misogyny or mainly rely on lexical shortcuts. Second, context-aware modelling should be explored carefully. Errors involving reported speech, humour, and stance suggest that isolated tweets often lack sufficient

information for reliable interpretation. However, simply adding more text is unlikely to be enough; models should account for conversational structure, speaker stance, quotation, target relations, and whether offensive language is endorsed, criticised, or reported. Third, explainability methods may help diagnose whether predictions are driven by slurs, target mentions, stereotypes, humour markers, or other cues. This is especially important in multilingual settings, where the same model may rely on different cues across languages (Vidgen et al., 2019).

More broadly, automatic misogyny detection should be framed as decision support rather than autonomous moderation. High recall can help prioritise human review, but false positives and false negatives carry different harms. Evaluation should therefore report not only aggregate scores, but also error types, affected discursive subcategories, language-specific weaknesses, and the moderation workflow in which the model would be used.

8 Conclusion and Future Work

This paper presented a multilingual analysis of misogynistic hate-speech detection in Dutch, English, and Spanish tweets using mmBERT. Its main contribution is a unified fine-grained evaluation framework that combines a newly collected Dutch dataset with re-annotated English and Spanish HatEval samples. Rather than reporting only binary performance, the study analysed how model errors relate to implicitness, aggressiveness, stereotypes, humour, and target type. The results show strong performance for English ($F_1 = 0.8651$) and Spanish ($F_1 = 0.8459$), but weaker transfer to Dutch ($F_1 = 0.6992$). Across languages, aggressive tweets are detected more reliably (error rates $<10\%$ across all languages), whereas implicit misogyny, stereotypes, and humour remain challenging, with Spanish implicit tweets (error rate = 31%) and Dutch humorous tweets (error rate = 35%) being especially error-prone.

The qualitative analysis shows that multilingual abuse detection is not solved by using a newer encoder alone. The model may over-rely on profanity, misread playful or self-directed insults, confuse reported speech with endorsement, and miss stereotype-based misogyny without overt slurs. Progress therefore requires larger and more diverse non-English datasets, multiple annotators, language-aware challenge sets, transparent report-

ing, and evaluation protocols that expose discourse-specific failure modes.

Limitations

This study has several limitations. First, the annotation was performed by one annotator. Although a second review pass, written guidelines, and consultation on ambiguous cases improved consistency, the study cannot report inter-annotator agreement. This is especially important for subjective phenomena such as humour, implicitness, and stereotype reproduction.

Second, the datasets are limited in size. Some category-language combinations contain fewer than 100 instances, which restricts the stability of fine-grained comparisons. The reported category error rates should therefore be interpreted as descriptive signals rather than definitive estimates.

Third, the Dutch dataset was collected using offensive keywords. This strategy provides a useful test of whether the model can distinguish misogyny from non-misogynistic profanity, but it also biases the sample toward lexically explicit content and may encourage shortcut learning. The English and Spanish samples originate from HatEval, where the collection strategy consisted of (1) monitoring potential victims of hate accounts, (2) downloading the history of identified haters, and (3) filtering Twitter streams using keywords (Basile et al., 2019). Consequently, collection procedures are not fully identical across languages.

Fourth, the analysis uses tweets in isolation. Conversational context, author information, and platform-level signals may change the interpretation of reported speech, humour, and implicit abuse. Adding such context could improve some decisions, but it also raises privacy, reproducibility, and modelling concerns.

Finally, the data come from Twitter/X and may not generalise to other platforms, time periods, user communities, or forms of online gender-based violence. The paper studies detection performance, not the effectiveness or fairness of a deployed moderation system.

Ethical Considerations

The data contain harmful and potentially distressing content. Researchers and annotators working with such material should be informed about the nature of the data and supported with appropriate safeguards. This paper avoids reproducing unnec-

essary abusive examples and uses redaction when discussing error patterns.

We will release the dataset in accordance with the Twitter/X Terms of Service and applicable privacy requirements. Since redistribution of full tweet text is restricted, the release will consist of tweet identifiers and annotation metadata rather than raw tweet content. Access to the anonymised data set is provided through a Data Usage Agreement, following practices used in similar shared-task and social-media dataset releases. To obtain access, researchers are required to register with a verified institutional email address and sign an agreement that limits use to the stated research purposes, prohibits re-identification attempts, and forbids redistribution of reconstructed tweet text.

Automatic misogyny detection can support moderation, but it can also cause harm through over-moderation, under-moderation, or unequal error rates across languages and communities. The reported scores should not be interpreted as evidence that the model is ready for autonomous deployment. Human oversight, appeals processes, and context-sensitive review remain necessary.

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.
- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. [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. Association for Computational Linguistics.
- Nina Bauwelinck, Gilles Jacobs, Veronique Hoste, and Els Lefever. 2019. [Lt3 at semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter \(hateval\)](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 436–440.
- 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, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman, and Malvina Nissim. 2021b. [DALC: the Dutch abusive language corpus](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 54–66, Online. Association for Computational Linguistics.
- Alessandra Teresa Cignarella, Anastasia Giachanou, and Els Lefever. 2026. [A survey on stereotype detection in natural language processing](#). *ACM COMPUTING SURVEYS*, 58(5):33.
- Florian Debaene, Aaron Maladry, Pranaydeep Singh, Els Lefever, and Véronique Hoste. 2025. Unlocking domain knowledge: Model adaptation for non-normative dutch. *Computational Linguistics in the Netherlands Journal*, 14:279–306.
- Jirky Dumon, Aaron Maladry, and Els Lefever. 2026. Annotation guidelines for labelling misogyny in tweets.
- Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. [Overview of the Evalita 2018 Task on Automatic Misogyny Identification \(AMI\)](#), pages 59–66.
- Lei Gao and Ruihong Huang. 2017. [Detecting online hate speech using context aware models](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.
- Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. [Hate speech detection: Challenges and solutions](#). *PLOS ONE*, 14(8):e0221152.
- Aaron Maladry, Alessandra Teresa Cignarella, Els Lefever, Cynthia Van Hee, and Veronique Hoste. 2024. [Human and system perspectives on the expression of irony: An analysis of likelihood labels and rationales](#). In *Proceedings of LREC-COLING 2024*, pages 8372–8382.
- Shervin Malmasi and Marcos Zampieri. 2017. [Detecting hate speech in social media](#). In *Proceedings of RANLP 2017*, pages 467–472.
- Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. [mmbert: A modern multilingual encoder with annealed language learning](#). *arXiv preprint arXiv:2509.06888*.
- Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. 2021. [Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection](#). *arXiv preprint arXiv:2103.14916*.
- Hadi Mohammadi, Anastasia Giachanou, and Ayoub Bagheri. 2024. [A transparent pipeline for identifying sexism in social media: Combining explainability with model prediction](#). *Applied Sciences*, 14(19).

Arianna Muti, Chris Emmery, Debora Nozza, Alberto Barrón-Cedeño, and Tommaso Caselli. 2025. [The “r” in “woman” stands for rights. auditing LLMs in uncovering social dynamics in implicit misogyny.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5462–5479, Suzhou, China. Association for Computational Linguistics.

Debora Nozza, Alessandra Teresa Cignarella, Greta Damo, Tommaso Caselli, and Viviana Patti. 2023. [HODI at EVALITA 2023: Overview of the first shared task on homotransphobia detection in italian.](#) In *Proceedings of the 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2023)*, volume 3473, pages 267–278. CEUR Workshop Proceedings.

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.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. [Toxicity detection: Does context really matter?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305.

Laura Plaza, Jorge Carrillo-de Albornoz, Enrique Amigó, Julio Gonzalo, Roser Morante, Paolo Rosso, Damiano Spina, Berta Chulvi, Alba Maeso, and Víctor Ruiz. 2024. [Exist 2024: sexism identification in social networks and memes.](#) page 498–504, Berlin, Heidelberg. Springer-Verlag.

Paul Rottger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. [Data-efficient strategies for expanding hate speech detection into under-resourced languages.](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020–2022. [Label Studio: Data labeling software.](#) Open source software available from <https://github.com/heartexlabs/label-studio>.

Cynthia Van Hee, Els Lefever, and Veronique Hoste. 2018. [We usually don’t like going to the dentist: Using common sense to detect irony on twitter.](#) *Computational Linguistics*, 44(4):793–832.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. [Challenges and frontiers in abusive content detection.](#) In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

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.

A Dutch Retrieval Keywords

Table 5 lists the Dutch keywords and expressions used for data collection.

Keyword / expression
alle vrouwen zijn
je bent een hoer
slet
allemaal hoeren
je bent een slet
sletterig
allemaal sletten
kankerhoer
stoefhoer
bitch
kankerslet
teef
griet
kankerwif
trut
hoer
kuthoer
tyfushoer
hoerig
kutwif
vrouwen moeten
hysterisch
rotwif
wif

Table 5: Dutch retrieval keywords and expressions used for data collection.

B English Translations of Examples

Table 6 lists English translations of the Dutch and Spanish examples listed in the qualitative error analysis.

English example / translation
Example 1: My sister when she heard I graduated with great distinction: 'bitch you've set the bar really high for me now' @user thanks x
Example 3: Not a bitch, not a whore, not a slut, none of that crap. A woman is a woman, and she deserves respect, even if she has a thousand boyfriends. She can have fun with them however the hell she wants. :) (Y)
Example 4: RT @user I think @user is such a bitch, always starts talking just when I'm about to go to sleep. Hahahaahah, leaaaave me alone, you jerk.
Example 6: @user YOU'RE THAT ONE SLUTTY FRIEND HAHAHAH.
Example 9: The best part of my day was hearing my nephew tell my sister to SHUT UP, BITCH.
Example 11: @user SEMEN movement! Get cleaning! http

Table 6: English translations of the Spanish and Dutch examples provided in the qualitative error analysis.