COMFOR at EXIST 2025: Support Vector Machines vs. Large Language Models in Sexism Detection

Notebook for EXIST at CLEF 2025

Fabio Fritzsche[1], Jenny Felser[1] and Michael Spranger[1,*]

[1]*Mittweida University of Applied Sciences, Technikumplatz 17, Mittweida, 09648, Germany*

Abstract

The increasing prevalence of sexist and misogynistic statements on social media poses a serious challenge. Due to the high volume of content, manual moderation is not feasible; therefore, automated detection systems are urgently needed. The EXIST task within CLEF 2025 is dedicated to the automatic identification of sexist content in social networks and the determination of the intention and subtype of sexism. This paper describes the COMFOR team's contribution to the first task of the competition, which focuses on tweets. A support vector machine (SVM) based on a comprehensive feature representation, including embeddings and lexical features, was used. For the third subtask, this classifier was used as the basis for a classifier chain. Additionally, the results of the first subtask were compared with those of a large language model (LLM) with an assigned persona. Our best models achieved an Information Contrast Measure (ICM) of 0.4928 (Subtask 1.1), −0.2203 (Subtask 1.2), and −0.4635 (Subtask 1.3) in the hard evaluation of the English test data.

Keywords

sexism detection, support vector machine, large language models, twitter

1. Introduction

Social networks are becoming increasingly important in today's society as a source of information, entertainment and even for exchanging opinions across geographical borders. Furthermore, a key advantage of online communication is the anonymity that allows people to express their opinions freely. However, this anonymity also has its downsides, as it can encourage the sharing of condescending or offensive content more frequently [1].

A particularly critical phenomenon is so-called hate speech, i.e. statements that attack individuals or groups based on characteristics attributed to the groups [2]. According to a 2022 report by the European Union, which examined offensive language and harassment on YouTube, Telegram, Reddit, and X, women were particularly affected by online hate compared to other social groups [3]. A study conducted by the Federal Working Group "Gegen Hass im Netz" showed, for example, that misogynistic online communication has increased from 2022 to 2023, and that women are targeted with disinhibited language that threatens and maligns them [4]. These findings demonstrate that social media can promote the spread of misogynistic views. Thus, users of social media specifically seek confirmation of their opinions and resort to sexist comments, often regardless of the discussion context [5].

To counteract the problem of online hate, including sexism, content moderators are tasked with monitoring communication and promptly removing harmful content if necessary [6]. However, given the sheer volume of content published daily, purely manual moderation appears too labour-intensive and time-consuming. Accordingly, there is an urgent need for methods to recognise sexist statements automatically.

The sEXism Identification in Social neTworks (EXIST) competition addresses this challenge by calling for the identification of content in text, image, and video data that either expresses critical or

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

ffritzsc@hs-mittweida.de (F. Fritzsche); felser@hs-mittweida.de (J. Felser); spranger@hs-mittweida.de (M. Spranger)

https://communication-forensics.de/ (M. Spranger)

0009-0008-7807-9018 (J. Felser); 0000-0002-6780-0841 (M. Spranger)

contemptuous views towards women or refers to such events [7]. It must be noted that the organisers equate misogyny with sexism, which is why this term is used throughout this document, even though sexism can be directed against any gender. The tasks were bilingual, in both English and Spanish. In this paper, we present our system for accomplishing the first task: detecting statements related to sexism in tweets, focusing on the English language. This task was divided into three subtasks:

1. a binary classification of whether a tweet is related to sexism
2. a multi-class classification of the intention of the author of sexist tweets
3. a multi-label classification to recognise all facets of a woman's life that are attacked in a sexist tweet

To accomplish these tasks, two different systems were compared with each other: On the one hand, a Support Vector Machine (SVM) was used, which was based on features such as embeddings, emojis and the frequency of sexist words. On the other hand, large language models (LLMs) were used as part of a few-shot prompting approach, in which the classification was carried out from the perspective of an assigned role (e.g., a male or female person). However, due to the long computing times of the LLMs, results could only be submitted for Subtask 1.1 in the competition using this approach.

The rest of this paper is organised as follows: After discussing the current state of the literature in section 2, the datasets used for this work and the methodology are briefly described in section 3. The results are then discussed in section 4. Finally, we conclude in section 5.

2. Related Work

Since this paper uses both traditional classification methods and advanced LLMs, the following section provides an overview of recent approaches to sexism detection, both in the field of traditional methods and modern language models, as well as related areas of application.

2.1. Approaches based on Traditional Machine Learning

The automatic detection of sexist tweets and the determination of their intention and subtype have already been addressed in two previous shared tasks [8, 9], using the same dataset. However, only a few participants employed traditional machine learning approaches in these competitions, for example [10, 11]. Nevertheless, the potential of these approaches should not be underestimated. For example, during the GerMS-Detect task of GermEval 2024 [12] – a competition series focusing on the German language - Donabauer [13] demonstrated that traditional methods, such as XGBoost, can achieve better results in the binary detection of sexism and misogyny than transformer-based models, such as the Bidirectional Encoder Representations from Transformers (BERT) model [14]. This result illustrates that more powerful language models do not necessarily deliver better results than classic approaches. Therefore, one focus of this work was on analysing traditional methods, particularly taking into account different feature representations.

SVM was chosen as the classification algorithm. As demonstrated by the survey conducted by Abdollah Zadeh et al. [15], SVM is robust against outliers and particularly well-suited to high-dimensional data, a property that is particularly relevant in text classification due to the large number of potential features. However, the disadvantage of classic SVM is that it cannot handle overlapping classes. For Subtask 1.2, in which multiple classes exist, this problem can be solved by using separate binary classifiers for each class.

Subtask 1.3, however, represents a genuine multi-label problem that requires more complex solutions. Asti et al. [16] investigated various methods for multi-label classification, including the combination of classification algorithms such as SVM, Multinomial Naive Bayes (MNB) and Random Forest (RF) with transformation methods such as the binary relevance method [17], classifier chains [18] and the label powerset transformation [19]. The combination of SVM and classifier chains, as introduced by Read et al. [20] in particular, showed superior performance with identical features [16]. That means that

SVM can be used not only as a binary classifier for Subtask 1.1, but also, with appropriate extensions, for the more complex Subtasks 1.2 and 1.3.

A key element in modelling is the selection of suitable features. Fasoli et al. [21], for example, compiled a list of sexist swear words and their social acceptability. Pamungkas et al. [22] also utilised swear words from both formal and informal language, drawing on data from the NoSwearing website [23], among other sources.

In addition, binary, dictionary-based features can be defined, for example, by checking whether a tweet contains terms associated with the word "woman" [21]. A hate word lexicon based on Bassignana et al. [24] was also used, which is divided into 17 weightable subcategories.

In the field of lexical features, classic representations such as Bag of Words (BoW), Bag of Hashtags and Bag of Emojis were used [22]. In addition, newer approaches demonstrate that semantic representation using word embeddings can yield better results [25]. Asudani et al. [25] provided a comprehensive overview of various methods, including Word2Vec [26] and its further development Global Vectors for Word Representation (GloVe) [27]. GloVe not only takes local contexts into account but also global co-occurrence statistics, and is particularly well suited for this application thanks to pre-trained models on Twitter data.

In summary, the overview of related research indicates that SVM, in conjunction with a classifier chain and features such as embeddings, presents an intriguing approach and can contribute to the diversification of methods within shared tasks.

2.2. Approaches using Large Language Models

In contrast to traditional approaches, participants in the two previous EXIST competitions primarily employed smaller and larger language models [28, 29, 30]. In particular, variants of BERT were frequently used, such as RoBERTa [28] applied for instance by Mohammadi et al. [31], Multilingual BERT (MBert) [14] for classifying the English and Spanish tweets [32] and Twitter-specific models such as Twitter-RoBERTa [29] applied by Martinez et al. [33]. A common practice for detecting sexism in social media using BERT is fine-tuning for the classification task [30, 34, 35], which is typically resource-intensive [36].

Instead of BERT-based and other so-called encoder-only models, larger, mostly decoder-only models, for instance from the Generative Pre-trained Transformer (GPT) [37] or Large Language Model Meta AI (LLaMA) family [38], have been proven promising for binary sexism and misogyny detection of English social media posts [39, 40, 41] as well as for their categorisation into subforms [40, 42].

The work of Samani et al. [40] is interesting in that they investigated various strategies for binary and fine-grained classification of sexism – zero-shot prompting, supervised fine-tuning and Reinforcement Learning from Human Feedback (RLHF) – based on the dataset of the Explainable Detection of Online Sexism (EDOS) task [43]. They concluded that the open-source model LLaMA, in particular, proved to be effective using RLHF. However, another promising technique that the authors did not include in their study is few-shot prompting (i.e., in-context learning). In this simple approach, the model is shown a few examples of the task via prompting, without the need for larger annotated datasets or human feedback [44]. This approach has proven successful, for example, in detecting hate speech in low-resource languages [45, 46], which can be seen as a kind of generalisation of sexism detection and is therefore also relevant to the task at hand. Nevertheless, both in the recognition of hate speech [45, 47] and, more importantly, in the detection of sexism within the EXIST 2024 Task [30], the results of LLMs applied in the few-shot scenario were outperformed by fine-tuned BERT-based models.

One possible approach for improvement is to combine few-shot prompting with the specification of a persona (i.e., role) or a specific perspective from which the LLM should perform the classification [48, 42, 49]. Examples of this person-based prompting include assigning a political stance to LLMs for hate speech classification [48], as well as assigning age and education level [49], age and gender [42], or various sociodemographic characteristics including age, ethnicity, and sexuality for the detection of sexism [40]. However, opinions differ as to whether assigning such characteristics to an LLM improves classification results [42] or have almost no effect [49] – depending, among other things, on the language

of the dataset, the model used and the specific prompting strategy.

In particular, Tian et al. [42] highlights the effectiveness of the persona-based prompting approach: a single LLaMA-3 model [50] assigned gender and age outperformed a more complex cascading strategy based on GPT models without persona assignment on the EXIST-2023 dataset [50]. Jiang et al. [41] demonstrated that instructing an LLM to classify texts such as tweets from the perspective of a person with specific sociodemographic characteristics can also be effective; however, overly complex prompts or too detailed descriptions of the persona can negatively impact model performance. For this reason, the present work focuses specifically on a single characteristic – gender – for the binary classification of sexism. Since, according to Aoyagui et al. [51], different model types are fundamentally capable of taking different perspectives when evaluating sexism, we do not limit ourselves to one model type, as Jiang et al. [41], for example, does, but compare the performance of four different models.

Despite their potential, it should not be forgotten that LLMs also have weaknesses, such as difficulties in recognising implicit sexism [39, 52] and a high sensitivity to the design of prompts [53]. Accordingly, it is interesting to investigate whether and to what extent LLMs outperform traditional approaches such as SVM.

3. Methodology

To address the problem of sexism detection in social media, SVM and LLM-based approaches were employed for Subtask 1.1, while the SVM system was utilised for Subtasks 1.2 and 1.3. We submitted a total of six runs. An overview of the approaches is provided in Table 1. Concerning Subtask 1.3, run *COMFOR_1* and *COMFOR_2* differ only in that hard labels were provided in the first run and soft labels in the second. The following paragraphs provide a detailed description of the individual methods.

Table 1
Description of the submitted system runs for all subtasks of Task 1.

Subtask	Run	Description
1.1	COMFOR_1	SVM classifier with GloVe embeddings and various lexical and statistical features trained on the pre-processed English training dataset, enriched with the translated Spanish training dataset
	COMFOR_2	Second-best combination of LLM model and persona using in-context learning with direct application to English non-preprocessed test data
	COMFOR_3	Second-best combination of LLM model and persona using in-context learning analogous to COMFOR_2
1.2	COMFOR_1	SVM classifiers with the same feature representation and the same training basis as Subtask 1.1 in a one-vs.-one approach
1.3	COMFOR_1	Classifier chain based on the SVM from Subtask 1.1 with hard class assignment
	COMFOR_2	Classifier chain based on the SVM from Subtask 1.1 with class probabilities

3.1. Data Description

The Twitter dataset for all subtasks of Task 1 addressed in this paper was initially created for the previous EXIST 2023 edition and described in detail by Plaza et al. [8]. In total, the dataset comprises 4,727 English and 5,307 Spanish tweets, which the organisers had already divided into training, validation and test data. Competition participants had the option of submitting results for only one language. We chose English because many of the features for SVM were language-dependent and required tools [54] and resources [23] that were only available in English at this scope. The organisers split the English tweets into a training dataset of 3,260 tweets, a validation dataset of 489 tweets and a test dataset of 978 tweets. To increase the training data set for the SVM, the Spanish training data set, comprising 3,660

tweets, was translated and combined with the English training data set. The translation was done with DeepL [55]. The combined training data set thus comprised 6,920 tweets.

A special characteristic of the annotations provided for the training data is that the "learning with disagreement" paradigm [56] was employed. That means that for each tweet, the individual annotations of all six annotators, along with their socio-demographic information, were provided instead of aggregated labels. An initial idea was therefore to account for the unreliability of individual annotators by merging the labels of the training dataset into a final label using a weighted majority vote, as suggested by Labudde and Spranger [57], for example. However, this approach was not feasible because the same six annotators did not annotate every tweet; instead, a total of 1,065 annotators were involved in the annotation process. Accordingly, it was not possible to make a statement about the reliability of annotators by evaluating their labelling behaviour.

Therefore, as is customary in the literature [58, 59], the labels for the three subtasks – both for the augmented training data and validation data – were summarised as follows based on the majority decisions of the annotators:

Subtask 1.1: In the first subtask, the binary decision of whether the tweet is sexist (YES) or not (NO), a simple majority decision was made, with YES being assigned in the event of a tie.

Subtask 1.2: If the majority label of the tweet in the first subtask was NO, the tweet was also assigned the label NO for the second subtask, which was to identify the author's intention. The reason for this was that only the intention of sexist tweets was to be determined [7]. Otherwise, a majority decision was made for the three intention categories. In the event of a tie, the corresponding tweets were removed from the dataset.

Subtask 1.3: In the third subtask, the multi-label classification of the targeted facets of a woman's life, only the sexist tweets (majority label YES) were given a label. Due to the design of this task as a multi-label classification task, a slightly modified procedure was necessary here: a label was assigned as soon as at least two annotators agreed on an assignment. If two annotators did not agree on at least one label for a tweet, this tweet was removed from the dataset.

The resulting class distribution of the training dataset, enriched with the translated tweets, is presented in Table 2, where NO indicates non-sexist tweets. For descriptions of the other classes, please refer to Plaza et al. [8].

Table 2
Number of tweets per category in the combined training data (English + automatically translated from Spanish) for the three subtasks of EXIST 2025, Task 1 [7]. The individual labels assigned by the annotators were aggregated into one label using the strategies described above.

Subtask	Category	# tweets
1.1	NO	4223
	YES	2697
1.2	NO	4223
	DIRECT	1617
	JUDGEMENTAL	579
	REPORTED	501
1.3	NO	4223
	STEREOTYPING AND DOMINANCE	1423
	IDEOLOGICAL AND INEQUALITY	1113
	OBJECTIFICATION	1103
	MISOGYNY AND NON-SEXUAL VIOLENCE	856
	SEXUAL VIOLENCE	675

As can be seen from Table 2, the training dataset used is highly unbalanced, particularly for Subtasks 1.2 and 1.3. For example, the DIRECT category contains approximately three times as many tweets as the JUDGEMENTAL and REPORTED categories.

3.2. System 1: Support Vector Machine

For all three subtasks, runs were submitted that utilised a traditional classification approach employing an SVM. The implementation of this system is described in more detail in the following subsections.

3.2.1. Data Preprocessing

Several cleanup steps were performed. Initially, URLs and brackets were removed, as those were not relevant for classification. In addition, mentions beginning with an @ sign were replaced with the placeholder "person", as it was irrelevant who exactly was being addressed since no information about users was available. However, for the second sub-task, which aims to distinguish whether the tweet reports on sexism, describes it or is itself sexist [7] – it is crucial if someone is addressed at all. Emojis were detected using Unicode patterns and mapped to their corresponding word expressions based on a dictionary-based approach. They were then treated as ordinary tokens of the tweet. In this way, their symbolic meaning could be preserved.

Since hashtags can also contain sexist references [60], the original hashtags were used as the basis for two features (see subsubsection 3.2.2), and compound hashtags were broken down into individual words so that they could be considered part of the tweet, just like emojis. The Wordninja tool by Keredson [54] was used for this purpose, as it can reliably separate hashtags even if their components are not separated by capital letters. At the same time, this tool removes all special characters.

Finally, all tweets were converted to lowercase for normalisation and lemmatised using the English UDPipe language model by Straka and Straková [61]. Over- and undersampling strategies were not used for the final model, as preliminary experiments indicated that these strategies deteriorated the results.

3.2.2. Feature extraction

To generate the feature representation of the tweets for all three subtasks, various types of features were extracted and concatenated into a single feature vector for each instance. Those included a sparse term vector representation based on term frequency–inverse document frequency (TF-IDF) weighting, a dense embedding of the tweet, and additional numerical features such as the number of dictionary words and the frequency of specific token types, including emojis. Each feature type is described in more detail subsequently.

TF-IDF vectors Following the BoW approach, a TF-IDF weighted document-term matrix was generated, with each row containing the weighted term vector of a tweet.

GloVe-Embeddings In addition, the GloVe model by Pennington et al. [27] was used, which was trained on a large tweet corpus and is therefore particularly well suited for this data. Subsequently, to obtain a vector representation for a tweet, the arithmetic mean of the embeddings of the words occurring in that tweet was calculated.

Emojis The semantic information of the emojis was taken into account by including their word descriptions in the BoW representation and the embedding representation after preprocessing (see subsubsection 3.2.1). Their word descriptions were included in the bag-of-words representation and the embedding representation. To also take into account quantitative information about the use of emojis, the number of emojis per tweet was added as a numerical feature.

Hashtags The original hashtags served as the basis for two additional features: First, following the approach of Pamungkas et al. [22], a TF-IDF weighted bag-of-hashtags representation was created, analogous to the standard bag-of-words model. For each tweet, this approach yielded a sparse vector whose elements corresponded to the TF-IDF values of the hashtags appearing in that tweet. Second, the total number of hashtags per tweet served as an additional numerical feature.

Swear words Words with an offensive or hurtful character play a central role in the detection of sexist language. Since traditional dictionaries often do not include such slang expressions, a list of swear words from the website NoSwearing.com [23] was used. This source, which is continually expanded, was particularly well-suited for identifying these terms. The number of swear words per tweet was integrated as a numerical feature.

Sexist words With a special focus on sexist language, a separate word list was created. To this end, the 100 words with the highest Pearson correlation to the YES label (i.e. sexism) in the training data were extracted. The frequency of these presumably sexist words served as a numerical feature.

3.2.3. Feature selection

After feature extraction, feature selection was carried out using the approach proposed by Kuhn [62], based on correlation analyses, to eliminate redundancies and make the subsequent training of the model more efficient. Specifically, the Pearson correlation coefficient was computed for each pair of features. If the absolute correlation value between two features exceeded a threshold of 0.5, the feature with a higher mean absolute correlation to all other features was removed, as it was considered less impactful. As a result of feature selection, individual terms were removed from the TF-IDF vector, while all other feature types were retained without modification.

3.2.4. Classification approach

Using the extracted features, SVM models were trained for each subtask. A sigmoid kernel was chosen because Srivastava and Bhambhu [63] generally recommend a non-linear kernel for classification tasks, and the sigmoid kernel proved promising in initial experiments. The sigmoid kernel $tanh(\gamma u'v)$ was used, where γ was set to the inverse of the number of features.

The different types of classification tasks required the use of different strategies for using the SVM algorithm:

Subtask 1.1: Binary classification SVM could be applied directly to this task.

Subtask 1.2: Multi-class classification Following the one-vs-one approach [64], a binary classifier was trained for each pair of classes. The class of new tweets was determined using the voting strategy described by Chang and Lin [65].

Subtask 1.3: Multi-label classification To take into account possible correlations between the categories, an ensemble classifier chain was trained using the approach of [20]. For each category (i.e. each facet), a binary classifier was created and linked in such a way that the predictions of previous classifiers were incorporated into the following classifiers as additional features. Since there is no natural order of the categories, the order of the classifiers in the chain was chosen arbitrarily. The classifier chain provided the probability of tweets belonging to the categories. These soft labels represented the *COMFOR_2* run of the subtask for the soft-soft evaluation, as described by Plaza et al. [7]. To obtain hard assignments, a threshold value of 0.3 was set for the membership probabilities for a label. These hard labels represented the *COMFOR_1* run of the subtask for the hard-hard evaluation.

Subtasks 1.2 and 1.3 were treated as a hierarchical classification problem: The systems were trained only on sexist annotated training data and applied exclusively to tweets that were predicted as sexist in Subtask 1.1. This approach met the requirement to only classify sexist tweets further.

3.3. System 2: Large Language Models

For Subtask 1.1, the results of the traditional approach, based on SVM, were compared with those of LLMs using persona-based prompting in combination with in-context learning. The focus was on open-source LLMs rather than paid APIs such as GPT-4 [37]. Specifically, the performance of the following models was evaluated using validation data: Qwen2.5-32B [66], Gemma 3-27B [67], LLama-3-8B [50] and Mixtral 8x22B [68]. The analysis of open-source models for detecting sexism offers the particular advantage that this sensitive data can be better protected, as the models are run locally on the user's servers.

The classification task was presented to all models using the same system and user prompt.

System prompt The system consists of an assignment to a persona or sociodemographic characteristic, a task description, and a definition of sexism. The following roles (personas) were each assigned to an LLM, which then annotated the entire English validation data set or test dataset:

- a male person
- a female person
- Alice Schwarzer, a well-known feminist

The selection of the first two roles – *a male person* and *a female person* – was based on the work of Tian et al. [42], Smith et al. [49] and on the assumption of the organisers of the shared task that the gender of human annotators can influence the assessment of sexism [8]. Since the classification task deals specifically with discrimination against women [8], it can be assumed that the "female" LLM tends to classify tweets as sexist more often than the "male" one. To further reinforce this effect, the persona of a prominent feminist, specifically Alice Schwarzer, was also introduced. The expectation here was that an LLM assigned such a role would be particularly sensitive to microaggressions and subtle forms of sexism. The idea of assigning a real personality to an LLM was taken up by Deshpande et al. [69], among others. They showed that this approach can significantly change both the attitude and the language of the model.

System Prompt

You are <persona>. Your task is to decide whether a tweet expresses ideas related to sexism:
Consider the following definition: Ideas about sexism can be expressed in the following three forms: the tweet is sexist itself, the tweet describes a sexist situation in which discrimination towards women occurs, or the tweet criticizes a sexist behaviour. Remember that you should annotate the tweet from the point of view of <persona>.

In addition, the LLM was provided with both the specific task and a definition of sexism via the system prompt—an approach also suggested, for example, by Reuver et al. [70]. The definition used is based on that provided by the organisers of EXIST 2025 [7] and has been slightly adapted to clarify which types of tweets are to be classified as sexist.

User prompt The user prompt consisted of the following parts: a repetition of the task and the assigned role, instructions on the desired output format, and examples of the classification task for in-context learning. Specifically, the one-class shot strategy recommended by Assis et al. [45] for hate speech detection was chosen, which involves providing only one annotated, preferably characteristic example for each of the two classes in this case. The motivation behind this is that a single, well-chosen

example can be more effective and helpful in guiding the LLM than several examples of lower quality. Therefore, the examples chosen were those mentioned by the organisers when presenting the tasks, rather than examples from the training data, where there were often contradictions between the human annotators [7].

User Prompt

Determine whether the input tweet contains sexism. You should only reply with "YES" or "NO". Do not provide explanations or notes. Respond with a single word. Respond "YES" if the tweet contains sexism and "NO" if it does not. Respond using JSON. Always remember that you should annotate the tweets from the perspective of <role>. Examples of classification are:

- People really try to convince women with little to no ass that they should go out and buy a body. Like bih, I don't need a fat ass to get a man. Never have. Label: YES
- @messyworldorder it's honestly so embarrassing to watch and they'll be like "not all white women are like that Label: NO

4. Results and Discussion

The following section compares the models using the validation data and discusses the final results of the competition based on the test data.

4.1. Comparison of Large Language Models

First, the performance of the LLMs was compared on the validation dataset, with the Information Contrast Measure (ICM) [71] score being the decisive factor. The organisers chose this score as the evaluation criterion for calculating the ranking [7]. The ICM score achieved and the normalised form, Macro ICM Norm, are presented in Table 3 for each of the four models, with the persona combination that yielded the highest ICM score for the corresponding model highlighted in bold. The results coloured in grey, without persona assignment, were obtained after the competition phase for further analysis.

Table 3
Results of the LLM-based approaches on the validation data. Combinations of four different LLMs and three different personas were compared, as well as a configuration without persona assignment. The results highlighted in grey without a persona were obtained after the competition phase.

Model	Persona (role)	ICM	Macro ICM Norm
Gemma 3-27B	male person	0.3925	0.7026
Gemma 3-27B	female person	0.3850	0.6987
Gemma 3-27B	Alice Schwarzer	0.3785	0.6953
Gemma 3-27B	-	**0.4119**	**0.7125**
Qwen2.5-32B	male person	0.3082	0.6591
Qwen2.5-32B	female person	0.3177	0.6639
Qwen2.5-32B	**Alice Schwarzer**	**0.3545**	**0.6830**
Qwen2.5-32B	-	0.2911	0.6502
Mixtral 8x22B	male person	0.2911	0.6502
Mixtral 8x22B	female person	0.3120	0.6610
Mixtral 8x22B	**Alice Schwarzer**	**0.3345**	**0.6726**
Mixtral 8x22B	-	0.2768	0.6428
LLama-3-8B	**male person**	**0.1383**	**0.5713**
LLama-3-8B	female person	0.1085	0.5559
LLama-3-8B	Alice Schwarzer	0.0596	0.5307
LLama-3-8B	-	0.0392	0.5202

The results presented in Table 3 reveal apparent performance differences between the models. Regardless of the assigned persona, Gemma 3-27B consistently achieves the best scores, while LLaMA-3-8B performs the worst. For example, Gemma 3-27B achieves approximately three times the ICM value of LLaMA-3-8B for male personas. This result supports the widely held assumption that model size is a key performance factor in LLMs, as explained by Bender et al. [72]. However, since Qwen2.5-32B lags behind Gemma 3-27B despite having a higher number of parameters, other influencing factors should not be ruled out, such as whether a model takes a perpetrator or victim perspective for classification, as described by Aoyagui et al. [51].

Furthermore, there was no tendency for the models to favour a particular gender for better classification results. While Gemma 3-27B and LLaMA-3-8B performed better when assigned the male persona, Qwen2.5-32B and Mixtral 8x22B achieved the highest scores in the role of Alice Schwarzer. Remarkably, none of the models achieved the best results with the unspecified female persona. One possible explanation for the different performance of the models depending on the assigned persona lies in the variation of the pre-training datasets [38, 66, 67, 68], which intensely influence the models' level of knowledge [73]. In particular, it must be investigated to what extent an LLM's knowledge of the persona assigned to it in the classification task affects its performance. The behaviour of the smallest model, LLaMA-3-8B, is particularly noteworthy: the ICM score for the generic female persona is twice as high as for the Alice Schwarzer persona. That could indicate that the model has difficulty adequately assuming the assigned role, possibly due to a lack of knowledge about the specific figure.

For the larger models, the assigned persona appears to have less influence on performance, especially in the Gemma 3-27B model, which raises the question of whether the assignment of personas makes a remarkable contribution to the classification result. For clarification, the classification task was subsequently repeated with all models, this time without persona assignment. The system prompt and user prompt remained identical to those in paragraph 3.3 and paragraph 3.3; only the text elements marked in red, which defined the persona, were omitted.

As Table 3 (results highlighted in grey) demonstrates, persona assignment led to a slight improvement in performance for all models except Gemma 3-27B. One possible explanation for the fact that the best-performing model did not show any improvement is that specifying a specific perspective is particularly helpful when a model is less confident in performing the task. However, it should also be noted that the difference in performance between prompting with and without persona was minor, especially for the larger models. This observation is also consistent with the findings of Civelli et al. [48], who suggest that persona assignments can increase the stylistic diversity of model responses, but do not necessarily lead to strongly different classification results.

4.2. Results on validation data

All systems used for submitting the runs were first evaluated on the English validation data (see Table 4). For Subtask 1.1, the SVM-based system (*COMFOR_1*) and the two most powerful LLM-based approaches were used. As shown in Table 3, the latter are Gemma 3-27B with an assigned male persona (*COMFOR_2*) and a female persona (*COMFOR_3*). The runs listed in Table 4 for Subtask 1.2 and Subtask 1.3 are based on the SVM system described in subsection 3.2.

It should also be noted that, for technical reasons, the ICM-Hard and ICM-Hard Norm scores could not be determined for the run *COMFOR_1* for Subtask 1.3 based on the validation data. The run *COMFOR_2* for Subtask 1.3 yielded so-called soft labels (see subsubsection 3.2.4). Accordingly, the metrics ICM-Soft and ICM-Soft Norm were calculated for this run, as described by Plaza et al. [9].

Subtask 1.1 When comparing the results of Subtask 1.1, it is noteworthy that the two LLM-based runs - *COMFOR_2* (Gemma 3-27B with male persona) and *COMFOR_3* (Gemma 3-27B with female persona) - outperformed the SVM baseline (*COMFOR_1*) in terms of the ICM hard score, the primary evaluation metric selected by the organizers, with a relative difference of 62.5%. All other calculated metrics also yielded lower scores for the SVM-based system compared to the LLM-based approaches. One possible reason for the lower performance of the SVM is the significant class imbalance in the training data

Table 4
Results of the systems submitted on the English validation data across all subtasks. *COMFOR_2* and *COMFOR_3* in Subtask 1.1 correspond to LLM-based runs; all other runs were generated using the SVM system.

Subtask	Run	Macro Precision	Macro Recall	Macro F1	ICM-Hard	ICM-Hard Norm	ICM-Soft	ICM-Soft Norm
1.1	COMFOR_1	0.77	0.74	0.75	0.24	0.62	-	-
	COMFOR_2	**0.81**	**0.81**	**0.81**	**0.39**	**0.70**	-	-
	COMFOR_3	**0.80**	**0.81**	**0.80**	**0.39**	**0.70**	-	-
1.2	COMFOR_1	0.47	0.48	0.47	-0.91	0.17	-	-
1.3	COMFOR_1	0.75	0.53	0.62	-	-		
	COMFOR_2	-	-	-	-	-	-10.89	0.00

(see subsection 3.1). This problem did not affect the LLM, as no further training or fine-tuning was performed on the training dataset.

Another limitation of the SVM approach arises from the large number of concatenated features. Although attempts were made to identify redundancy through correlation analyses, these experiments should be repeated with alternative correlation measures such as XICOR [74], which also capture non-linear relationships.

Furthermore, it is questionable whether the combination of different lexical feature types – such as GloVe embeddings and the BoW representation – offers added value. A systematic investigation of the respective contribution of individual feature types and their combinations to model performance is therefore necessary. More targeted feature engineering could further improve the results of SVM and may bring them closer to those of LLM-based approaches.

Concerning the results of the LLM approaches, it is remarkable that the macro precision, macro recall and macro F1 values hardly differ when assigning a male (*COMFOR_2*) or female persona (*COMFOR_3*). This result confirms the previous impression that the selected persona in Gemma 3-27B has little influence on the classification results.

Subtask 1.2 Regarding Subtask 1.2, it is noticeable that a significantly lower F_1 measure of only 0.47 and a negative ICM score were achieved. One possible reason for this result is that only tweets that were classified as sexist (YES) by the SVM in Subtask 1.1 were passed on to the SVM for fine-grained classification. Misclassifications by the SVM for Subtask 1.1, therefore, have a direct impact on the performance of the SVM for Subtask 1.2.

An alternative approach would be to address the task not as a hierarchical problem, but to directly classify the tweets into four categories, including the NO category, i.e., no sexism. However, this approach would exacerbate the problem of imbalance, as the non-sexist tweets strongly outnumber the tweets assigned to the various categories of sexist tweets. The fact that the imbalance is already problematic in the three categories of author intent is evident from the F_1 scores achieved for these individual categories:

- DIRECT: 0.76
- JUDGEMENTAL: 0.08
- REPORTED: 0.27

The F_1 measure for the overrepresented class DIRECT clearly exceeds that of the two underrepresented classes. In addition to class imbalance, the performance differences could also be due to the fact that feature types such as BoW or dictionary-based features are particularly well suited for detecting explicitly sexist tweets of the DIRECT category. Tweets in the JUDGEMENTAL category, however, describe (social) circumstances perceived as sexist [7] without necessarily containing terms with clear sexist connotations or swear words. Accordingly, it is not surprising that the SVM system with lexical features achieved the lowest F_1 score in this category.

Subtask 1.3 Both the problem of imbalance and the dependence on the classifier's performance from Subtask 1.1 also affect Subtask 1.3. Nevertheless, the macro F_1 measure for this task was higher than for Subtask 1.2. This result was achieved through comparatively high macro precision, but at the expense of lower macro recall. The decisive factor in this trade-off is primarily the threshold value at which a tweet is assigned a label, which is why systematic testing of different threshold values could lead to performance improvements.

The results of the soft evaluation were considerably weaker, with a negative ICM soft score of -10.89 and an ICM soft norm of 0.0. A possible reason is the training on an aggregated gold standard (see subsection 3.1), where agreement between two annotators was sufficient for label assignment. Thus, annotation uncertainties were not taken into account, which may have affected the prediction of the actual proportion of annotators who assigned a label [7].

4.3. Results on test data

The final evaluation by the competition organisers resulted in several rankings: on the one hand, the systems were compared only with the gold standard of one language, and on the other hand, with the gold standard for both languages [7]. The results of the hard-all evaluation, i.e. comparison of the predicted hard labels with the hard labels of the entire gold standard, can be found in Table 5. In addition to the evaluation values calculated by the organisers, this table also shows the ranking within the total number of submitted results.

Table 5
Results of the submitted runs on the English and Spanish instances of the test data for all subtasks in the hard-hard evaluation. The predicted class labels were compared with the aggregated labels of the annotators.

Subtask	Run	Rank	ICM-Hard	ICM-Hard Norm	**F1** (YES)
1.1	COMFOR_1	153 (of 160)	-0.4300	0.2839	0.3510
	COMFOR_2	149 (of 160)	-0.3278	0.3352	0.4379
	COMFOR_3	**147 (of 160)**	**-0.3061**	**0.3461**	**0.4609**
1.2	COMFOR_1	107 (of 140)	-0.9562	0.1891	0.2088
1.3	COMFOR_1	103 (of 132)	-1.3994	0.1751	0.3077

For all subtasks, even the first one, the submitted runs are listed at the bottom of the rankings. However, these results can mainly be explained by the fact that we only submitted results for English tweets. Therefore, the results and ranking of our runs, based solely on the English test data, which can be found in Table 6, are more interesting.

Table 6
Results of the submitted runs on the English instances of the test data for all subtasks in the hard-hard evaluation.

Subtask	Run	Rank	ICM-Hard	ICM-Hard Norm	**F1** (YES)
1.1	COMFOR_1	141 (of 158)	0.2033	0.6038	0.6067
	COMFOR_2	115 (of 158)	0.4384	0.7237	0.7185
	COMFOR_3	**77 (of 158)**	**0.4928**	**0.7515**	**0.7313**
1.2	COMFOR_1	93 (of 139)	-0.2203	0.4238	0.3106
1.3	COMFOR_1	78 (of 131)	-0.4635	0.3864	0.4683

Table 6 clearly shows that LLM-based models continue to outperform SVM on the test data for Subtask 1.1. In contrast to the results of the validation data, *COMFOR_3*, i.e. Gemma 3-27B with a female persona, achieved a slightly higher ICM hard score than *COMFOR_2*. Ranked 77th out of 158, this run is roughly in the middle of the submitted results. Accordingly, there is also room for improvement for the LLM-based system. In addition to the aspects already discussed in subsection 4.2, it should

also be investigated whether, as emphasised by Jiang et al. [41], less information-overloaded prompts could be more effective, for which persona assignment is combined with zero-shot prompting instead of in-context learning. The results of the SVM on the English test data were similar to those on the English validation data, with the low ranking once again highlighting the system's weaknesses.

Moreover, the SVM-based systems achieved poor results in Subtasks 1.2 and 1.3 with negative ICM hard scores. However, the fact that over 20 % of the teams in Subtask 1.2 and 10 % in Subtask 1.3 achieved an ICM hard score of 0.0 [75] underscores the challenge of reliably predicting aggregated labels in these classification tasks, considered subjective [7, 51].

Finally, the results for the soft-soft evaluation of Subtask 1.3 are shown in Table 7, both for all test data and for English data only.

Table 7
Result of the *COMFOR_2* ranking for Subtask 1.3 on the test data in the soft-soft evaluation. The first row refers to the evaluation on all instances, the second to the evaluation on the English instances.

Dataset	Rank	ICM-Soft	ICM-Soft Norm
All instances	29 (of 53)	-9.7944	0.00
English instances	27 (of 52)	-10.1922	0.00

What is particularly notable here is that the ICM Soft norm is zero in both cases. However, when comparing these results with those of the other participants, it becomes apparent that over 20 participants achieved a score of zero in this measure, which highlights that predicting the proportion of annotators who assigned a label also poses challenges. In this context, the organisers of the Shared Task themselves emphasised that label ambiguity and the lack of agreement between annotators make this task particularly difficult [75].

5. Conclusion

The aim of the first task of the EXIST-2025 task was to detect sexism in tweets and then to classify sexist content in more detail. In this paper, we presented our approach to these subtasks specifically for English-language tweets. For binary sexism detection (Subtask 1.1), two approaches were compared: On the one hand, LLMs were used in combination with persona-based prompting and in-context learning. On the other hand, a traditional machine learning approach was used, in which an SVM was trained on a dataset enriched with tweets translated from Spanish into English. Features based on a GloVe model pre-trained specifically for Twitter, as well as additional lexical and statistical features, were used for the SVM. The SVM was also used for the other subtasks: to classify the author's intention (Subtask 1.2) and to detect the facet of a woman's life affected in the tweet (Subtask 1.3). For multi-label classification in Subtask 1.3, the SVM served as the basis for the classifier chain method.

For Subtask 1.1, our best model, the LLM Gemma 3-27B with an assigned male persona, achieved an ICM score of 0.4928 on the English test data, outperforming the SVM approach, which achieved an ICM score of 0.2033. In the other two subtasks, however, negative ICM values were achieved in each case, a result primarily due to the strong class imbalance in the training data, which negatively impacted the performance of the SVM.

One possible approach to mitigate this problem is to follow the example of Lee et al. [76] and utilise LLMs to generate additional training instances for underrepresented classes artificially. Furthermore, since LLMs delivered better results than classic models in Subtask 1.1, it makes sense to extend their use to Subtasks 1.2 and 1.3. To date, it has been demonstrated that assigning a gender-specific persona has had only a minor impact on the classification results. Therefore, further potential can be tapped by assigning alternative perspectives, such as a specific cultural background.

Additionally, it has not yet been taken into account that the intense subjectivity of the task can lead to disagreements among annotators. A possible solution here could be a combination of LLMs

with classic methods such as SVM. In this case, the LLM would not act as the sole classifier, but as a kind of 'artificial additional annotator' that takes over a decision-making function in the event of label inconsistencies and can thus contribute to the consistency of the training dataset.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword, Improve writing style. Further, the authors used DeepL in order to: Text Translation. After using these tool/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] I. Weber, H. Vandebosch, K. Poels, S. Pabian, The ecology of online hate speech: Mapping expert perspectives on the drivers for online hate perpetration with the Delphi method, Aggressive Behavior 50 (2024) e22136. doi:10.1002/ab.22136.

[2] A. Guterres, United Nations Strategy and Plan of Action on Hate Speech, 2019.

[3] European Union Agency for Fundamental Rights (Ed.), Online Content Moderation: Current Challenges in Detecting Hate Speech, Publications Office, Luxembourg, 2023. doi:10.2811/332335.

[4] C. Dolezalek, M. Fielitz, C. Heindl, S. Jaki, K. Schwarz, Tracing Online Misogyny - Eine Analyse Misogyner Ideologien Und Praktiken Aus Deutsch-Internationaler Perspektive, Technical Report, Bundesarbeitsgemeinschaft Gegen Hass im Netz, Berlin, 2024.

[5] J. L. Gil Bermejo, C. Martos Sánchez, O. Vázquez Aguado, E. B. García-Navarro, Adolescents, Ambivalent Sexism and Social Networks, a Conditioning Factor in the Healthcare of Women., Healthcare (Basel, Switzerland) 9 (2021). doi:10.3390/healthcare9060721.

[6] T. M. Hansen, L. Lindekilde, S. T. Karg, M. Bang Petersen, S. H. R. Rasmussen, Combatting online hate: Crowd moderation and the public goods problem, Communications 49 (2024) 444–467. doi:10.1515/commun-2023-0109.

[7] L. Plaza, J. Carrillo-de-Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Proceedings of the 47th European Conference on Information Retrieval, volume 15576, Springer Nature Switzerland, Lucca, Italy, 2025, pp. 442–449. doi:10.1007/978-3-031-88720-8_65.

[8] L. Plaza, J. Carrillo-de-Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization, Vienna, Austria, 2023, pp. 316–342. doi:10.1007/978-3-031-42448-9_23.

[9] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 - Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. Di Nunzio, L. Soulier, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Grenoble, France, 2024, pp. 93–117. doi:10.1007/978-3-031-71908-0_5.

[10] S. Fan, R. A. Frick, M. Steinebach, FraunhoferSIT@EXIST2024: Leveraging Stacking Ensemble Learning for Sexism Detection, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS, Grenoble, France, 2024, pp. 993–1002.

[11] M. Sreekumar, S. Karthik, D. Thenmozhi, S. Gopalakrishnan, K. Swaminathan, Sexism Identification In Tweets Using Machine Learning Approaches, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS, Grenoble, France, 2024, pp. 1253–1259.

[12] S. Gross, J. Petrak, L. Venhoff, B. Krenn, GermEval2024 Shared Task: GerMS-Detect – Sexism Detection in German Online News Fora, in: B. Krenn, J. Petrak, S. Gross (Eds.), Proceedings of GermEval 2024 Task 1 GerMS-detect Workshop on Sexism Detection in German Online News Fora (GerMS-detect 2024), Association for Computational Linguistics, Vienna, Austria, 2024, pp. 1–9.
[13] P. Donabauer, Pd2904 at GermEval2024 (Shared Task 1: GerMS-Detect): Exploring the Effectiveness of Multi-Task Transformers vs. Traditional Models for Sexism Detection (Closed Tracks of Subtasks 1 and 2), in: B. Krenn, J. Petrak, S. Gross (Eds.), Proceedings of GermEval 2024 Task 1 GerMS-detect Workshop on Sexism Detection in German Online News Fora (GerMS-detect 2024), Association for Computational Lingustics, Vienna, Austria, 2024, pp. 39–47.
[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.
[15] A. Abdollah Zadeh, J. Leevy, T. Khoshgoftaar, A Survey on the Choice Between Binary Classification and One-Class Classification, in: Proceedings of the 27th ISSAT International Conference on Reliability and Quality in Design, International Society of Science and Applied Technologies, Virtual Event, 2022.
[16] A. D. Asti, I. Budi, M. O. Ibrohim, Multi-label Classification for Hate Speech and Abusive Language in Indonesian-Local Languages, in: Proceedings of the 2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE, Depok, Indonesia, 2021, pp. 1–6. doi:10.1109/ICACSIS53237.2021.9631316.
[17] E. Montañes, R. Senge, J. Barranquero, J. Ramón Quevedo, J. José Del Coz, E. Hüllermeier, Dependent binary relevance models for multi-label classification, Pattern Recognition 47 (2014) 1494–1508. doi:10.1016/j.patcog.2013.09.029.
[18] E. Alvares-Cherman, J. Metz, M. C. Monard, Incorporating label dependency into the binary relevance framework for multi-label classification, Expert Systems with Applications 39 (2012) 1647–1655. doi:10.1016/j.eswa.2011.06.056.
[19] M. R. Boutell, J. Luo, X. Shen, C. M. Brown, Learning multi-label scene classification, Pattern Recognition 37 (2004) 1757–1771. doi:10.1016/j.patcog.2004.03.009.
[20] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier Chains for Multi-label Classification, in: W. Buntine, M. Grobelnik, D. Mladenić, J. Shawe-Taylor (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 254–269.
[21] F. Fasoli, A. Carnaghi, M. P. Paladino, Social acceptability of sexist derogatory and sexist objectifying slurs across contexts, Language Sciences 52 (2015) 98–107. doi:10.1016/j.langsci.2015.03.003.
[22] E. W. Pamungkas, V. Basile, V. Patti, Misogyny Detection in Twitter: A Multilingual and Cross-Domain Study, Information Processing & Management 57 (2020) 102360. doi:10.1016/j.ipm.2020.102360.
[23] All Slang Network, NoSwearing.com: Swear Word List, Dictionary, Filter, and API, https://www.noswearing.com/, 2025.
[24] E. Bassignana, V. Basile, V. Patti, Hurtlex: A Multilingual Lexicon of Words to Hurt, in: Proceedings of the 5th Italian Conference on Computational Linguistics (CLiC-it), volume 2253, CEUR-WS, Turin, Italy, 2018, pp. 51–56. doi:10.4000/books.aaccademia.3085.
[25] D. S. Asudani, N. K. Nagwani, P. Singh, Impact of word embedding models on text analytics in deep learning environment: A review, Artificial Intelligence Review 56 (2023) 10345–10425. doi:10.1007/s10462-023-10419-1.
[26] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed Representations of Words and Phrases and their Compositionality, Advances in Neural Information Processing Systems 26 (2013) 1–8.
[27] J. Pennington, R. Socher, C. D. Manning, GloVe: Global Vectors for Word Representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Process-

ing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. doi:10.3115/v1/D14-1162.

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. doi:10.48550/ARXIV.1907.11692.

[29] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, L. Neves, TweetEval: Unified benchmark and comparative evaluation for tweet classification, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1644–1650. doi:10.18653/v1/2020.findings-emnlp.148.

[30] A. Azadi, B. Ansari, S. Zamani, S. Eetemadi, Bilingual Sexism Classification: Fine-Tuned XLM-RoBERTa and GPT-3.5 Few-Shot Learning, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, arXiv, Grenoble, France, 2025, pp. 958–965. doi:10.48550/arXiv.2406.07287. arXiv:2406.07287.

[31] H. Mohammadi, A. Giachanou, A. Bagheri, Towards Robust Online Sexism Detection: A Multi-Model Approach with BERT, XLM-RoBERTa, and DistilBERT for EXIST 2023 Tasks, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS, Thessaloniki, Greece, 2023, pp. 1000–1011.

[32] M. Usmani, R. Siddiqui, S. Rizwan, F. Khan, F. Alvi, A. Samad, Sexism Identification in Tweets using BERT and XLM Roberta, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS, Grenoble, France, 2024, pp. 1274–1279.

[33] E. Martinez, J. Cuadrado, J. C. M. Santos, E. Puertas, Notebook for the VerbaNex AI Lab at CLEF 2024, CEUR-WS, Grenoble, France, 2024, pp. 1107–1113.

[34] A. Das, N. Raychawdhary, T. Bhattacharya, G. Dozier, C. D. Seals, AU_NLP at SemEval-2023 task 10: Explainable detection of online sexism using fine-tuned RoBERTa, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 707–717. doi:10.18653/v1/2023.semeval-1.97.

[35] J. Panwar, R. Mamidi, DAP-LeR-DAug: Techniques for enhanced online sexism detection, in: M. Abbas, A. A. Freihat (Eds.), Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023), Association for Computational Linguistics, Online, 2023, pp. 51–58.

[36] Y. Kit, M. M. Mokji, Sentiment Analysis Using Pre-Trained Language Model With No Fine-Tuning and Less Resource, IEEE access 10 (2022) 107056–107065. doi:10.1109/ACCESS.2022.3212367.

[37] OpenAI, GPT-4 Technical Report, 2023. doi:10.48550/arXiv.2303.08774. arXiv:2303.08774.

[38] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and Efficient Foundation Language Models, 2023. doi:10.48550/ARXIV.2302.13971.

[39] A. Vetagiri, P. Pakray, A. Das, A deep dive into automated sexism detection using fine-tuned deep learning and large language models, Engineering Applications of Artificial Intelligence 145 (2025) 110167. doi:10.1016/j.engappai.2025.110167.

[40] A. R. Samani, T. Wang, K. Li, F. Chen, Large Language Models with Reinforcement Learning from Human Feedback Approach for Enhancing Explainable Sexism Detection, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics (COLING), Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 6230–6243.

[41] A. Jiang, N. Vitsakis, T. Dinkar, G. Abercrombie, I. Konstas, Re-examining Sexism and Misogyny Classification with Annotator Attitudes, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 15103–15125. doi:10.18653/v1/2024.findings-emnlp.887.

[42] L. Tian, N. Huang, X. Zhang, Large Language Model Cascades and Persona-Based In-Context

Learning for Multilingual Sexism Detection, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. Di Nunzio, L. Soulier, P. Galuščáková, A. García Seco De Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 14958 of *Lecture Notes in Computer Science*, Springer Nature Switzerland, Grenoble, France, 2024, pp. 254–265. doi:10.1007/978-3-031-71736-9_18.

[43] H. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 task 10: Explainable detection of online sexism, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2193–2210. doi:10.18653/v1/2023.semeval-1.305.

[44] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, ACM Computing Surveys 55 (2023) 1–35. doi:10.1145/3560815.

[45] G. Assis, A. Amorim, J. Carvalho, D. de Oliveira, D. Vianna, A. Paes, Exploring Portuguese Hate Speech Detection in Low-Resource Settings: Lightly Tuning Encoder Models or In-Context Learning of Large Models?, in: P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, R. Amaro (Eds.), Proceedings of the 16th International Conference on Computational Processing of Portuguese, volume 1, Association for Computational Lingustics, Santiago de Compostela, Galicia/Spain, 2024, pp. 301–311.

[46] U. Sahin, I. E. Kucukkaya, O. Ozcelik, C. Toraman, Zero and Few-Shot Hate Speech Detection in Social Media Messages Related to Earthquake Disaster, in: Proceedings of the 31st Signal Processing and Communications Applications Conference (SIU), IEEE, Istanbul, Turkiye, 2023, pp. 1–4. doi:10.1109/SIU59756.2023.10224056.

[47] K. Guo, A. Hu, J. Mu, Z. Shi, Z. Zhao, N. Vishwamitra, H. Hu, An Investigation of Large Language Models for Real-World Hate Speech Detection, in: 2023 International Conference on Machine Learning and Applications (ICMLA), IEEE, Jacksonville, FL, USA, 2023, pp. 1568–1573. doi:10.1109/ICMLA58977.2023.00237.

[48] S. Civelli, P. Bernardelle, G. Demartini, The Impact of Persona-based Political Perspectives on Hateful Content Detection, in: Companion Proceedings of the ACM on Web Conference 2025, ACM, Sydney NSW Australia, 2025, pp. 1963–1968. doi:10.1145/3701716.3718383.

[49] T. K. Smith, H. R. Nie, J. R. Trippas, D. Spina, RMIT-IR at EXIST Lab at CLEF 2024, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS, Grenoble, France, 2024, pp. 1237–1252.

[50] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The Llama 3 Herd of Models, arXiv e-prints (2024) arXiv–2407. doi:10.48550/arXiv.2407.21783.

[51] P. A. Aoyagui, K. Stemmler, S. A. Ferguson, Y.-H. Kim, A. Kuzminykh, A Matter of Perspective(s): Contrasting Human and LLM Argumentation in Subjective Decision-Making on Subtle Sexism, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI'25), volume 529 of *Chi '25*, Association for Computing Machinery, New York, NY, USA, 2025, pp. 1–16. doi:10.1145/3706598.3713248.

[52] C. P. Almendros, J. Camacho-Collados, Do Large Language Models Understand Mansplaining? Well, Actually..., LREC-COLING 2024 (2024) 5235–5246.

[53] L. Han, H. Tang, Designing of Prompts for Hate Speech Recognition with In-Context Learning, in: 2022 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, Las Vegas, NV, USA, 2022, pp. 319–320. doi:10.1109/CSCI58124.2022.00063.

[54] Keredson, Word Ninja, https://github.com/keredson/wordninja, 2025.

[55] DeepL SE, DeepL Translate: The world's most accurate translator, https://www.deepl.com/en/translator, 2025.

[56] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, SemEval-2021 Task 12: Learning with Disagreements, in: A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, X. Zhu (Eds.), Proceedings of the 15th International Workshop on

Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 338–347. doi:10.18653/v1/2021.semeval-1.41.

[57] D. Labudde, M. Spranger, Towards Inter-Rater-Agreement-Learning, in: Proceedings of the Tenth International Conference on Advances in Information Mining and Management (IMMM), IARIA Press, Lisabon, Portugal, 2020, pp. 10–14.

[58] C. Demus, J. Pitz, M. Schütz, N. Probol, M. Siegel, D. Labudde, DeTox: A Comprehensive Dataset for German Offensive Language and Conversation Analysis, in: K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen, Z. Talat (Eds.), Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Seattle, Washington (Hybrid), 2022, pp. 143–153. doi:10.18653/v1/2022.woah-1.14.

[59] E. Fersini, P. Rosso, M. Anzovino, Overview of the Task on Automatic Misogyny Identification at IberEval 2018, in: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR-WS, Sevilla, Spain, 2018, pp. 214–277.

[60] P. Chiril, F. Benamara, V. Moriceau, "Be nice to your wife! The restaurants are closed": Can gender stereotype detection improve sexism classification?, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2833–2844. doi:10.18653/v1/2021.findings-emnlp.242.

[61] M. Straka, J. Straková, Universal Dependencies 2.5 Models for UDPipe (2019-12-06), http://hdl.handle.net/11234/1-3131, 2019. Accessed on 2025-07-30.

[62] M. Kuhn, Data Sets and Miscellaneous Functions in the caret Package, Journal of Statistical Software (2011) 1–23.

[63] D. Srivastava, L. Bhambhu, Data classification using support vector machine, Journal of Theoretical and Applied Information Technology 12 (2010) 1–7.

[64] U. Kreßel, Pairwise Classification and Support Vector Machines, in: C. J. Burges, B. Schölkopf, A. J. Smola (Eds.), Advances in Kernel Methods, The MIT Press, 1998. doi:10.7551/mitpress/1130.003.0020.

[65] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 1–27. doi:10.1145/1961189.1961199.

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

[67] Gemma Team, Gemma: Open Models Based on Gemini Research and Technology, 2024. doi:10.48550/arXiv.2403.08295.

[68] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mixtral of Experts, 2024. doi:10.48550/ARXIV.2401.04088.

[69] A. Deshpande, V. Murahari, T. Rajpurohit, A. Kalyan, K. Narasimhan, Toxicity in chatgpt: Analyzing persona-assigned language models, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 1236–1270. doi:10.18653/v1/2023.findings-emnlp.88.

[70] M. Reuver, I. Sen, M. Melis, G. Lapesa, Tell me what you know about sexism: Expert-LLM interaction strategies and co-created definitions for zero-shot sexism detection, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Findings of the Association for Computational Linguistics: NAACL 2025, Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 8438–8467. doi:10.18653/v1/2025.findings-naacl.470.

[71] E. Amigó, F. Giner, J. Gonzalo, F. Verdejo, On the foundations of similarity in information access, Information Retrieval Journal 23 (2020) 216–254. doi:10.1007/s10791-020-09375-z.

[72] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the Dangers of Stochastic Parrots:

Can Language Models Be Too Big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, ACM, Virtual Event Canada, 2021, pp. 610–623. doi:10.1145/3442188.3445922.

[73] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, O. Levy, LIMA: Less Is More for Alignment, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, Nips '23, Curran Associates Inc., Red Hook, NY, USA, 2024, pp. 55006–55021. doi:10.48550/arXiv.2305.11206.

[74] S. Chatterjee, A New Coefficient of Correlation, Journal of the American Statistical Association 116 (2021) 2009–2022. doi:10.1080/01621459.2020.1758115.

[75] L. Plaza, J. Carrillo-de-Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos, in: J. Carrillo-de-Albornoz, A. G. S. de Herrera, J. Gonzalo, L. Plaza, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), CEUR-WS, Madrid, Spain, 2025.

[76] D.-H. Lee, J. Pujara, M. Sewak, R. White, S. Jauhar, Making large language models better data creators, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 15349–15360. doi:10.18653/v1/2023.emnlp-main.948.