Generalizable BERT-Based Cross-Media Sexism Classification

Notebook for the EXIST Lab at CLEF 2025

Tim Chopard[1,†], Darren Rawlings[2,†]

[1]University of Leeds, Woodhouse, Leeds, LS2 9JT, UK
[2]University of Groningen, Broerstraat 5, 9712 CP Groningen, The Netherlands

Abstract

Sexism has pervasive negative effects on both individuals and society. This paper presents a generalizable BERT-based approach to identifying and classifying the source intent of sexism across different social network channels. This approach focuses on individual models trained on the text of tweets and then applied to both Meme (image) and Video data using OCR and annotations respectively. The identification model performed well across all channels and the classification model performed well on both Tweets and Memes. This research suggests that a single model, fine-tuned on one media type can be effectively applied to multiple media types with minimal data preprocessing required.

Keywords

BERT, Sexism, Classification, Social Networks, Natural Language Processing

1. Introduction

Sexism has many far-reaching negative consequences, including reduced representation and decreased access and quality in healthcare [1]. Many of the stereotypes that underpin sexism are reinforced through media [2], one dominant media type being that of social networks [3].

Social networks are somewhat unique within media in that posts can, to an extent, be swiftly critiqued or countered[4]. This, alongside approaches to remove or reduce the reach of sexist content, is reliant on robust models to identify and classify sexist content within a variety of social media channels.

This paper describes a Bidirectional Encoder Representations from Transformers (BERT) [5] language model approach to identifying sexism within social media posts as part of EXIST (sEXism Identification in Social neTworks) [6]. EXIST 2025 provides clear guidelines on defining sexism which encompass content that includes either sexist expressions or behaviors. This definition is then further specified by splitting the sexist content into one of three categories, as described in Table 1. This provides a framework for classifying the content of social media that can be applied across multiple channels including text, image and video.

The approach taken was to train a model on a single media type and then test it both on unseen data from that type as well as the other media types. The goal was to create a generalizable model that could be applied across multiple media channels with minimal preprocessing or further fine-tuning required.

2. Data

For all three data sets there were accompanying metadata provided. These included data about the annotators, the labels, as well as the text or transcripts of speech present in the media.

The data concerning the annotators was not used in this research, and the conclusions reached by the annotators were assumed to be correct for labeling the data.

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

[†]These authors contributed equally.

timchopard@cloudberries.io (T. Chopard); d.rawlings@student.rug.nl (D. Rawlings)
https://cloudberries.io (T. Chopard); https://github.com/startung (D. Rawlings)

Table 1
The tiers of categories used in EXIST 2025 for classifying sexism and sexist behavior.

Is Sexist	Intention	Description
Yes	Direct	A goal to be sexist or exhibit sexist behavior
Yes	Judgemental	A description of sexism or sexist behavior with the intent to condemn it
Yes	Reported	Reporting direct or indirect experience of sexism or sexist behavior
No	-	The sample does not contain sexism

The labels provided for each individual Tweet, Meme or Video are shown in Table 2. All labels shown are present in the Tweets dataset, and all labels except "Reported" are present in the Memes and Videos datasets.

2.1. Tweets

As seen in Table 2, the provided data included 10034 tweets, split across training, dev (which, for this paper, was used as validation), and test datasets. Each of these datasets were split into approximately 53% Spanish language, and 47% English language tweets. The training and dev datasets were labeled with annotations from 6 annotators, with YES indicating the annotator's belief that the tweet was sexist and NO indicating the belief that it was not sexist. Demographic data for the annotators was also provided, but was not used in training of the models.

No preprocessing was applied to this data, all elements such as hashtags and emojis were retained. No attempts were made to correct spelling or grammar, and no data augmentation techniques were used.

Table 2
Summary of the provided data, showing the Dataset purpose, language and item counts.

Media	Dataset	Language	Count
Tweets	Training	Spanish	3660
	Training	English	3260
	Dev	Spanish	549
	Dev	English	489
	Test	Spanish	1098
	Test	English	978
Memes	Training	English	2010
	Training	Spanish	2034
	Test	English	513
	Test	Spanish	540
Videos	Training	English	1000
	Training	Spanish	1524
	Test	English	370
	Test	Spanish	304

2.2. Memes

The Memes dataset was comprised of 5097 images, along with accompanying metadata and labels. These data were split across training and test data. With a split of approximately 51% Spanish language, and 49% English language. These were composed of 4703 JPEGs and 394 PNGs of varying dimensions. These images were supplied with a provided copy of their text content generated by Optical Character Recognition (OCR).

For further analysis on the images, all were converted to PNG files using a standard sRGB color palette. A secondary description of the image was generated using Gemma 3 [7]. This provided additional text based data alongside the OCR.

2.3. Videos

The Video dataset was also split into train and test datasets, with 2525 training videos and 674 test videos. All videos were mp4 files of varying duration, with accompanying transcriptions.

A Whisper speech-to-text model [8] was used to annotate the videos, generating an additional transcript to the one provided in the video metadata. Ultimately, this secondary transcript was not used in the final method due to time constraints.

3. Methods

For the competition submission two methods were used, they varied on the method used to create the outputs and how the data was loaded for training.

The first method (labeled bergro_1) used a pair of pretrained BERT models, fine-tuned on the tweets in the training data to output a single soft classification label. This label was used directly for the soft evaluation then mapped to YES/NO (with NO for all values less than 0.5, otherwise YES) for the hard evaluation. The data for this model was loaded with each tweet appearing once per epoch, with the label created by dividing the number of annotators indicating YES divided by the total number of annotators. Therefore, for a tweet which three annotators labeled the tweet YES (sexist) and three labeled it NO (not sexist) the label would be 0.5.

The second method used a pair of pretrained BERT models, fine-tuned on the tweets in the training data, but this time as a classifier with YES/NO outputs. The data for this model was loaded, with each tweet appearing once per annotator, with that annotator's label. As all tweets in the training and dev data provided were labeled by six annotators, the models were trained on the tweets six times per epoch, often with conflicting labels, reflecting the disagreement between annotators.

For the final competition, the models were trained using both the training and dev data sets to increase the amount of training data for the model.

3.1. Models

The approach to tackle this problem for both submission methods was to fine-tune a version of a BERT model. This approach was taken as these models work well for classification tasks and are fine-tuneable of consumer grade GPUs. Multilingual BERT variants exist [5], but initial testing revealed that single language models outperformed them on this task. Following initial trials one model was selected for each language.

3.1.1. English Model

For processing the English language tweets, DistilBERT [9], a smaller, more efficient version of the BERT (Bidirectional Encoder Representations from Transformers) language model [5] was used. Developed by researchers at Hugging Face, it is designed to be both smaller and faster than the original BERT model whilst maintaining comparable language understanding capabilities.

The key innovation behind DistilBERT is a technique called knowledge distillation. During its pre-training phase, a smaller *student* model (DistilBERT) is trained to mimic the full output distribution of a larger, pre-trained *teacher* model (BERT). This process transfers the rich *dark knowledge* from the teacher to the student.

For the bergro_1 approach the model was adapted to a single output label, to allow for the regression style approach, then fine-tuned on the training data using the dev data as validation data to optimise the hyperparameters.

3.1.2. Spanish Model

For the Spanish language model BETO: Spanish BERT [10] was selected, this model is larger than the DistilBERT model used for the classification of English language tweets. This model contains

approximately 110 million parameters, and is structured like the original BERT model, although the pre-training has some adaptations taken from later BERT derivatives such as RoBERTa [11]. The key difference, however, is the training corpora used which included 3 billion tokens of Spanish language content [12].

3.2. Image Descriptions

To augment the OCR generated content for the memes, the images were fed to an multimodal LLM to provide a description. The model used was Gemma3 [7] 12 billion parameters, quantized to Q4 using Quantization Aware Trained (QAT). The prompts were language specific, requesting a description of the image in 250 words.

The prompts used for this purpose were:

- **English**: "In 250 words or less describe the image. Do this without any preamble."
- **Spanish**: "En 250 palabras o menos, describa la imagen. Haga esto sin preámbulos." (this was machine translated from the English prompt).

For two Spanish memes the model ignored the requested word count and the text was cropped to 512 tokens to fit within the input size used in the BERT-based models.

3.3. Training metric

Training the bergro_1 model, which is a regressor model with a single output, used the Mean Squared Error (MSE) loss, which is the squared difference between the prediction probability distribution of the annotators and the actual distribution.

For training the bergro_2 model, which is a binary (YES/NO) classifier, accuracy was used as a metric. This measures the overall proportion of correct predictions.

3.4. Evaluation metrics

For the final results the metrics were provided by the EXIST 2025 organizers. As in previous iterations of EXIST, the organizers used both a *soft* evaluation and a *hard* evaluation. The soft evaluation was intended to measure the model's ability to capture disagreements. It achieves this by comparing the probability distribution of predicted labels with that of the labels provided by the annotators. In contrast the hard evaluation requires the system to predict a single label for every tweet and is compared against an absolute value calculated by a majority vote of the annotators' responses. For those instances for which there is no majority class (i.e. for subtask 1 there are 3 YES labels and 3 NO labels) were removed from this evaluation scheme. These are then combined into three metrics per evaluation type (soft/hard).

3.4.1. Hard evaluation

- **Information Contrast Model (ICM)**: How well the predicted outcomes align with group differences within the data [13].
- **Normalized Information Contrast Model (ICM Norm)**: A normalized version of ICM.
- **F_1:** A balanced accuracy measure taking into account false positives and false negatives [14].

3.4.2. Soft evaluation

- **Information Contrast Model-Soft (ICM-Soft)**: A modification of the original ICM metric in which the Information Contrast is estimated from both the soft ground truth values and the soft predictions [15].
- **Information Contrast Model-Soft Norm (ICM-Soft Norm)**: A normalized version of ICM-Soft.
- **Cross Entropy**: The difference between the predicted probability distribution and the ground truth distribution.

3.5. Source Intention

For training the Source Intention model, only the Tweets dataset was used. The model was trained on the proportion values for each class, as shown under intention in table 1. Two separate models were trained, as for the sexism identification. One model for the English content; DistilBERT, and one model for the Spanish content; BETO: Spanish BERT. Cross validation was performed in order to tune the hyperparameters, with the final hyperparameters shown in Table 3. The model was then used to predict tweets from a Soft likelihood estimates for each class from a previously unseen dataset.

Table 3
Hyperparameters used for training the BERT models in Source Intention classification

Hyperparameter	Value DistilBERT	Value BETO: Spanish BERT
Epochs	10	8
Learning Rate	3e-5	3e-5
Maximum Length	140	140
Batch Size	16	16

This model was also used to predict the Soft estimates for both Videos and Memes. As both of these datasets did not include the "Reported" label, this was removed from the predictions with the remaining values being normalized to once again sum to one.

Hard values were inferred from the soft values by selecting the class with the maximum predicted probability (argmax) for each instance.

4. Results

4.1. Sexism Identification

As seen in Table 4 the best results for task 1.1 were achieved with the bergro_1 model, using a regression approach to tweet classification. This outperformed bergro_2 in which a hard label was directly generated. Table 4 shows the runs submitted to the EXIST lab organizers.

Table 4
The results and rankings for the lab submission towards task 1.1 showing all measured metrics.

Submission	Hard Rank	ICM	ICM Norm	F_1	Soft Rank	ICM-Soft	ICM-Soft N.	CE
bergro_1	39/160	0.5194	0.7611	0.7654	18/67	0.6382	0.6023	0.7921
bergro_2	104/160	0.4289	0.7156	0.7350				

Table 5
Hard and Soft metrics for the combined BERT models trained on the Tweets test and dev datasets, and the scores for selecting majority classes. For the Tweets model the results are from the test dataset. For the Memes and Videos the results are from the training data from the respective datasets, note the model was not trained on these datasets.

Data	Model	ICM	ICM Norm	F_1	ICM-Soft	ICM-Soft Norm	Cross Entropy
Tweets	BERT (bergro_1)	0.5194	0.7611	0.7654	0.6382	0.6023	0.7921
	Majority	-0.4413	0.2782	0.0000	-2.3585	0.1218	4.6115
Memes	BERT (OCR)	0.0265	0.5136	0.6240	-1.1543	0.3185	2.4318
	BERT (Gemma3)	0.0907	0.5466	0.6441	-1.0470	0.3354	2.3909
	Majority	-0.4038	0.2947	0.6821	-2.3568	0.1212	4.4015
Videos	BERT	0.0741	0.5371	0.6887	-0.1499	0.4737	2.3045
	Majority	-0.4244	0.2858	0.0000	-1.2877	0.2740	4.4285

4.2. Source Intention

As seen in Table 6, the models trained on source intention from the Tweets training data performed well on both an unseen Tweets dataset as well as on the Memes and Video data. For each dataset the Majority score was also included, this shows the metrics for selecting the most common labels in each dataset.

Table 6
Hard and Soft metrics for the combined BERT models trained on the Tweets, and the scores for selecting majority classes.

Data	Model	ICM	ICM Norm	F_1	ICM-Soft	ICM-Soft Norm	Cross Entropy
Tweets	BERT	0.1542	0.5482	0.4974	-3.4845	0.2204	1.7011
	Majority	-0.9504	0.1910	0.1603	-5.4460	0.0612	4.6233
Memes	BERT	-0.4319	0.3511	0.4017	-1.8513	0.3063	1.4841
	Majority	-1.0445	0.1369	0.1839	-5.0745	0.0000	5.5565
Videos	BERT	-0.4988	0.3272	0.4337	-3.1683	0.1378	1.3987
	Majority	-0.7537	0.2155	0.2375	-3.1337	0.1663	4.4354

The BERT model performed well on both Tweets and Memes, notably outperforming the majority selection, suggesting much better than random inference. This applied to both the hard and soft labels. The performance was less good on Videos, slightly outperforming the majority selection on hard labeling, and almost matching it on soft labeling.

5. Conclusion

5.1. Overview

The focus of the EXIST lab as a whole was to build models tuned to each data stream, Tweet, Meme and Video. This paper chose a slightly different path, namely to focus on a single generalized model that could be applied across channels. Using a generalized model reduced training time and resources required. Furthermore, the tweets set was the largest, and simplest to process.

5.1.1. Sexism Identification

For the trained BERT models, predicting a soft likelihood estimation and then mapping that estimation to a hard-label (bergro_1) outperformed the direct generation of a hard-label (bergro_2). This trained model performed strongly on tweets, and without further training also outperformed, the majority-class baseline for both memes (on both OCR generated text and LLM generated descriptions) and videos (on the provided transcripts). Using a multimodal LLM to provide a description outperformed using the OCR generated text context. This may be due to the spatial and non-textual information the multimodal LLM was able to provide.

5.1.2. Source Intention

The trained BERT model showed promise in predicting labels across different social media formats. Our results show that a BERT model trained only on Tweets data can notably outperform a majority baseline for both Tweets and Memes. This applied for Soft likelihood estimation as well as the Hard classification.

The model performed less well on the Videos dataset, outperforming the majority baseline in Hard classification, and roughly matching it in Soft likelihood estimation.

This suggests that a pre-trained BERT model can be tuned to identify both sexism and intent across multiple media channels using only a single channel for tuning. It also shows that this may not work universally, with certain similarities required between the channels.

5.2. Limitations

Amongst the challenges with detecting sexism, particularly direct sexism is that there are often efforts made to obfuscate the intent of a post, hashtags and slang terms used euphemistically to indicate ideological context [16, 17]. These methodologies change rapidly meaning that any system seeking to detect them also needs to dynamically update over time. This holds some importance within the context of this paper, as well as more broadly.

The data used in this paper spans from 2015 through to 2024. As such, models are also impacted by the evolution of language use through this time period. Language usage has been shown to change rapidly within social networks [18], with word and phrase meanings shifting over time.

5.3. Future Research

This paper has shown some promise in using BERT models trained on one media channel to classify sexism in another channel. This could be further explored to develop systems that could quickly be deployed to new and emerging media channels with minimal re-tuning required.

This research did not utilize all available information from the provided datasets. This includes the information about the annotators, and further information contained within the media, such as imagery in the videos. All this information is likely to hold value, and exploiting this in future research could be the key to unlocking improved accuracy and precision, whilst gaining deeper insights.

Declaration on Generative AI

The authors have not employed any Generative AI tools in writing this paper.

References

[1] P. Homan, Health consequences of structural sexism: Conceptual foundations, empirical evidence and priorities for future research, Social Science & Medicine 351 (2024) 116379. URL: https://www.sciencedirect.com/science/article/pii/S0277953623007360. doi:https://doi.org/10.1016/j.socscimed.2023.116379, gender, power, and health: Modifiable factors and opportunities for intervention.

[2] R. Stewart, B. Wright, L. Smith, S. Roberts, N. Russell, Gendered stereotypes and norms: A systematic review of interventions designed to shift attitudes and behaviour, Heliyon 7 (2021).

[3] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, IEEE Access 8 (2020) 219563–219576. doi:10.1109/ACCESS.2020.3042604.

[4] P. Carvalho, D. Caled, C. Silva, F. Batista, R. Ribeiro, The expression of hate speech against afro-descendant, roma, and lgbtq+ communities in youtube comments, Journal of Language Aggression and Conflict 12 (2024) 171–206.

[5] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[6] 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, J. Gonzalo, L. Plaza, A. García Seco de Herrera, 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), 2025.

[7] Gemma Team, Gemma 3, 2025. URL: https://goo.gle/Gemma3Report.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: International conference on machine learning, PMLR, 2023, pp. 28492–28518.
[9] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).
[10] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.
[11] 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, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.
[12] J. Cañete, Compilation of large spanish unannotated corpora, 2019. URL: https://doi.org/10.5281/zenodo.3247731. doi:10.5281/zenodo.3247731.
[13] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. URL: https://aclanthology.org/2022.acl-long.399/. doi:10.18653/v1/2022.acl-long.399.
[14] N. Chinchor, MUC-4 evaluation metrics, in: Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992, 1992. URL: https://aclanthology.org/M92-1002/.
[15] 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 (extended overview), in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[16] A. Perliger, C. Stevens, E. Leidig, Mapping the ideological landscape of extreme misogyny, International Centre for Counter-Terrorism, 2022.
[17] W. Zhu, H. Gong, R. Bansal, Z. Weinberg, N. Christin, G. Fanti, S. Bhat, Self-supervised euphemism detection and identification for content moderation, in: 2021 IEEE Symposium on Security and Privacy (SP), 2021, pp. 229–246. doi:10.1109/SP40001.2021.00075.
[18] T. Dembe, The impact of social media on language evolution, European Journal of Linguistics 3 (2024) 1–14.