Multilingual Sexism Detection in Memes, A CLIP - Enhanced Machine Learning Approach

Umera Wajeed Pasha

University of Galway, University Road, Galway, Ireland H91 TK33

Abstract

In this work, we use cutting-edge machine learning approaches to tackle the problem of sexism identification in memes. The study starts by importing and visualising a meme dataset, then pre-processing the images using techniques including cropping, scaling, and normalisation to get them ready for model training. A pre-trained model called CLIP is used to extract features, and the dataset is split into training and validation sets for memes in both Spanish and English. The collected features are used to train and assess a variety of machine learning models, such as Logistic Regression, SVM, XGBoost, Decision Trees, Random Forest, Neural Network, AdaBoost, and SGD. Accuracy scores, classification reports, and confusion matrices are used to evaluate performance. The Random Forest model performed the best out of all of them. After that, a JSON file containing the model's predictions about the occurrence of sexism in a test dataset is created. The results highlight how well-trained models and sophisticated machine learning approaches can identify hazardous content on social media, offering insightful information for future studies and useful applications that will help create safer online spaces.

Keywords

Sexism detection, Meme Analysis, Machine Learning (ML), Contrastive Learning, Learning with disagreement, Multilingual Natural Language Processing (NLP)

1. Introduction

Social networks have developed into an essential communication tool in the current digital era, enabling people to openly express their ideas and opinions but this transparency has also resulted in the spread of offensive material, such as sexism—a gender-based discrimination that mostly targets women. As sexism on social media is so widespread, automated solutions must be developed to identify and remove such offensive content. In order to solve this problem, the EXIST 2024 shared task challenges participants to develop models that can recognise sexist content in environments that are multilingual, specifically in Spanish and English [1].

The complex and context-dependent nature of the language used makes it difficult to automatically detect sexism. With differing degrees of effectiveness, conventional machine learning techniques like logistic regression and support vector machines (SVM) have been used. Transformer-based models have shown higher performance in natural language processing (NLP) tasks, such as sexism detection, more recently. Examples of these models are BERT [2], [3], RoBERTa [4], and their multilingual variations.

This research presents a way for identifying sexism in social networks by combining pre-trained embeddings with machine learning models. Key steps in the approach include loading and exploring datasets, pre-processing images, extracting features using the CLIP model [5], separating datasets, training and evaluating models. Memes were classified as sexist or non-sexist using a variety of machine learning methods, such as AdaBoost, SVM, XGBoost, Decision Trees, Random Forest, Logistic Regression, and SGD [6]. The highest-performing model, Random Forest, was then used to forecast whether sexism will be present in a test dataset. The outcomes were then stored in a JSON file for further analysis.

The dataset of memes used in this study has been annotated for sexism. To ensure high-quality input for

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

U.Pasha1@universityofgalway.ie (U. W. Pasha)

model training, the dataset is subjected to a thorough pre-processing protocol. The proposed method intends to support the ongoing efforts to develop more inclusive and safe online environments by fusing powerful machine learning models with sophisticated feature extraction techniques.

2. Background

Finding damaging and sexist content on social media has been a major field of study, with many different strategies and techniques put forth. In the beginning, sexism was frequently studied as a type of harassment or as a subcategory of hate speech. Character-level and word n-grams with logistic regression to classify tweets as racist, sexist, or neither along with other research have used conventional machine learning techniques like Random Forests, TF-IDF, and Support Vector Machines (SVM), utilising manually chosen features like emotion ratings and Bag of Words (BoW) [1].

Deep learning has greatly improved the performance of NLP tasks, including sexism detection. In particular, Transformer-based models like BERT [2], RoBERTa [4], and their multilingual variations (e.g., XLM-RoBERTa [3]) have made a substantial contribution to this improvement [7] [8]. The Transformer architecture enables these models to capture complicated language semantics and context, which is beneficial [9]. For example, trained on a large multilingual corpus, XLM-RoBERTa has shown greater performance in capturing multilingual context nuances, which makes it especially useful for jobs involving diverse languages.

3. System Overview

A number of crucial processes are involved in the proposed method for detecting sexism in memes: importing and exploring datasets, pre-processing images, extracting features using the CLIP model, partitioning datasets, training and evaluating models.

3.1. Dataset Loading and Exploration

The collection includes memes as shown in Figure 1 with sexism annotations in both Spanish and English [10]. To comprehend the structure and substance of the dataset, the first stages are to load and visualise it. To provide a visual sense of the diversity and dispersion of the data, samples of memes with their related labels are displayed in this stage. More than 5,000 labelled memes in English and Spanish make up the EXIST 2024 Memes Dataset [11], [12]; 4,044 of the memes are categorised as training, and 1,053 as testing. The dataset makes sure that the two languages are distributed equally, which makes thorough multilingual analysis possible. Every meme is organised as a JSON object with comprehensive properties such as a distinct identifier ("id_EXIST"), the meme's language ("lang"), and the text that has been automatically retrieved from the meme ("text"). The filename ("meme") and the file's path ("path_memes") are also included in the dataset. The number of annotators, their unique identifiers, gender, age group, self-reported ethnicity, degree of education, and nation of residency are all carefully documented in the annotator data. Multiple annotators label each meme to indicate whether or not it contains sexist expressions or behaviours. "YES" or "NO" are examples of possible labels. This extensive annotation offers a strong basis for developing and testing machine learning models designed to identify sexism in memes. The organised method to guaranteeing fair and thorough data coverage for both training and testing phases is demonstrated in this detailed perspective of the dataset, which is depicted in the Table 1.

3.2. Pre-processing

Preparing images for feature extraction and model training requires a crucial step called pre-processing, which guarantees consistency and ideal input quality. In the pre-processing stage, images undergo several critical transformations to ensure consistency and optimal input quality for the model. First,

Figure 1: Dataset: EXIST: sEXism Identification in Social neTworks [11],[12]

Figure 2: Spanish Sexist Meme

Figure 3: Spanish Non-Sexist Meme

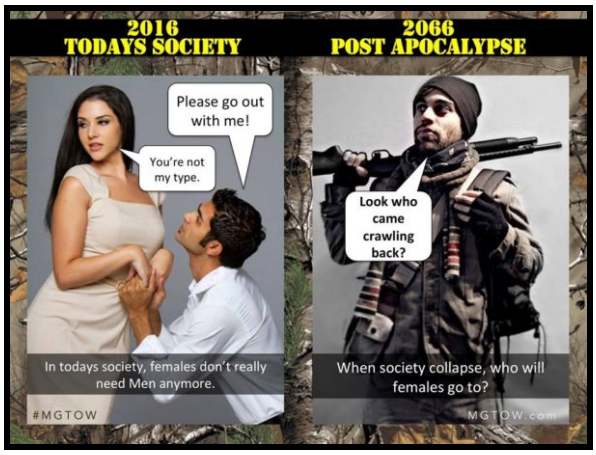

Figure 4: English Sexist Meme

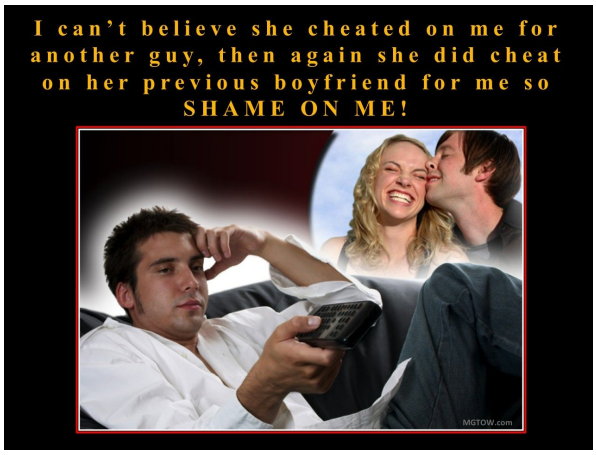

Figure 5: English Non-Sexist Meme

Table 1
Description of Datasets and Image Ranges

Dataset	Language	Image Count	Range
Training	Spanish	2034	110001-112034
	English	2010	210001-212010
Testing	Spanish	540	310001-310540
	English	513	410001-410513

the images are resized to a uniform dimension of 256 × 256 pixels, standardizing input sizes to reduce computational complexity and enhance processing performance. The images are then centrally cropped to 224 × 224 pixels, which helps eliminate extraneous background elements and focus on the main content of the memes. Following cropping, the pixel values are normalized to a range typically between 0 and 1, ensuring uniform feature scaling which accelerates the convergence process during training. The pre-processing pipeline also involves converting images to RGB format to maintain color consistency and handling any potential image loading errors. These meticulously crafted steps, performed using

Python libraries such as PIL for image handling and Torchvision for transformations, are essential for meeting the input specifications of the CLIP model [5], which relies on consistently processed images for precise feature extraction.

3.3. Feature Extraction

For feature extraction, we employ the state-of-the-art pre-trained model CLIP (Contrastive Language-Image Pre-training) [5], which is renowned for its efficaciousness in encoding text and images into a common feature space. CLIP is able to comprehend and categorise complicated multimodal input by using contrastive learning to align visual and textual representations. We extract high-dimensional feature vectors that capture the semantic content of the pre-processed images by feeding them into the CLIP model. The following machine learning models then use these attributes as inputs. The reason behind the selection of CLIP is its strong ability to capture the subtle correlations between textual and visual data, which makes it especially appropriate for applications like meme categorization where the quality of both text and image material is crucial [5].

3.4. Dataset Splitting

To guarantee an even distribution of memes in Spanish and English, the dataset is carefully divided into training and validation sets. In order to maintain the representativeness of the training data and ensure that the models trained on it can effectively generalise to new examples across other languages, stratified splitting is essential. The validation set acts as an impartial set to assess how well the machine learning models perform; the training set is utilised to fit the models. This process is necessary to determine how effectively the models will function in practical situations and to adjust hyperparameters to avoid overfitting. Additionally, by preventing the models from becoming biased in favour of any one language, the balanced distribution improves the models cross-linguistic applicability.

3.5. Model Training and Evaluation

Different machine learning models were trained to categorise the memes as sexist or non-sexist after feature extraction and dataset splitting. Logistic regression, Support Vector Machines (SVM), XGBoost, Decision Trees, Random Forest, AdaBoost, Neural Networks, and Stochastic Gradient Descent (SGD) are among the models that were assessed. After a thorough training process using the collected features, each model is assessed using confusion matrices, accuracy scores, and classification reports. These measurements offer a thorough evaluation of each model's effectiveness, pointing out both its advantages and disadvantages in terms of sexist content detection.

The Random Forest model outperformed the other models that were assessed, exhibiting the best classification accuracy and robustness as shown in Figure 6. This model is excellent at managing complicated datasets and reducing overfitting. It is well-known for its ensemble learning method, which integrates many decision trees. The Random Forest model is then used to predict if sexism will be present in the test dataset after it has been determined to be the top performer. The forecasts are then stored in a JSON file for additional examination, resulting in a structured output that is simple to understand and apply to reports and more study.

These intricate procedures, which include pre-processing, feature extraction, dataset splitting, model training, and evaluation, guarantee a strong and all-encompassing solution to the problem of sexism detection in memes. Every stage is meticulously crafted to optimise the efficacy and applicability of the models, hence augmenting the system's total efficiency in practical scenarios.

Figure 6: Model Performance Summary Chart

Model	Accuracy	Precision (Not Sexist)	Recall (Not Sexist)	F1-Score (Not Sexist)	Precision (Sexist)	Recall (Sexist)	F1-Score (Sexist)	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-Score	Weighted Avg Precision	Weighted Avg Recall	Weighted Avg F1-Score
Logistic Regression	63.54%	0.62	0.6	0.61	0.65	0.67	0.65	0.63	0.63	0.63	0.64	0.64	0.64
SVM	67.99%	0.68	0.63	0.66	0.68	0.72	0.7	0.68	0.68	0.68	0.68	0.68	0.68
XGBoost Classifier	79.11%	0.79	0.77	0.78	0.79	0.81	0.8	0.79	0.79	0.79	0.79	0.79	0.79
Decision Tree	73.42%	0.72	0.73	0.73	0.75	0.73	0.74	0.73	0.73	0.73	0.73	0.73	0.73
Random Forest	80.10%	0.82	0.75	0.78	0.79	0.85	0.82	0.8	0.8	0.8	0.8	0.8	0.8
MLP Classifier (NN)	77.63%	0.78	0.74	0.76	0.77	0.81	0.79	0.78	0.78	0.78	0.78	0.78	0.78
AdaBoost Classifier	57.48%	0.56	0.55	0.55	0.59	0.6	0.59	0.57	0.57	0.57	0.57	0.57	0.57
SGD	58.47%	0.55	0.72	0.62	0.64	0.46	0.53	0.6	0.59	0.58	0.6	0.58	0.58

4. Results

Creating a system to recognise sexist material in memes was the focus of Task 4 of the EXIST 2024 shared task. The ICM-Hard metric, normalised ICM-Hard, and the F1 score for the positive class (F1_YES) are used to assess the performance of the model. With a focus on various evaluation situations, these metrics offer a thorough understanding of the model's capacity to detect sexist content. The outcomes of our contribution "Umera Wajeed Pasha_1.json" in three distinct evaluation contexts are shown below: all instances, Spanish instances, and English instances.

4.1. Overall Performance

In terms of the comprehensive assessment of every case, the system produced the following outcomes: Based on these findings, the model is ranked 36th out of all the participants. The system's capacity to

Table 2
Overall Performance Metrics for Sexist Content Detection

ICM-Hard	ICM-Hard Norm	F1_YES
-0.3083	0.3432	0.5956

manage intricate, hierarchical classification tasks is shown by the ICM-Hard score, a metric that takes into account the information content of both correct and incorrect classifications. A standardised view of this performance is given by the normalised ICM-Hard score (ICM-Hard Norm), while the F1_YES score emphasises the recall and precision for the positive class—in this example, the identification of sexist memes.

4.2. Spanish Instances

The system's performance increased when tested on Spanish instances, proving its capacity to manage multilingual data successfully: The model placed 30th in this category with these results. The pre-

Table 3
Performance Metrics for Spanish Instances in Sexist Content Detection

ICM-Hard	ICM-Hard Norm	F1_YES
-0.2216	0.3871	0.6032

processing and feature extraction strategies appear to be especially beneficial for Spanish language

memes, based on the enhanced scores in the Spanish context. When it comes to identifying sexist content in Spanish memes, a higher F1_YES score denotes improved memory balance and precision.

4.3. English Instances

In contrast, the evaluation on English instances highlighted areas for improvement in the system's performance: Here, the model ranked 37th. The model's difficulties in managing English language

Table 4
Performance Metrics for English Instances in Sexist Content Detection

ICM-Hard	ICM-Hard Norm	F1_YES
-0.3953	0.2993	0.5882

memes are indicated by the lower scores in the English context, which also point to possible areas for improvement in the pre-processing or feature extraction for English content. In this language area, there has to be a greater balance between recall and precision, as indicated by the relatively lower F1_YES score.

The performance measures in various scenarios highlight the advantages and disadvantages of this methodology. Although it showed diversity among languages, the Random Forest model, which was shown to be the best-performing model during training and validation, exhibited resilient performance overall. Future improvements, concentrating on customised feature extraction and pre-processing strategies to handle the unique qualities of English and Spanish memes, might be guided by the insights gained from these results.

5. Methodology Enhancement

It is clear from the insights from the existing results that improving the methods might greatly increase the resilience and efficacy of the sexism detection system. This section describes a number of possible improvements, with particular attention on sophisticated feature extraction methods, dynamic pre-processing pipelines, and hybrid approaches that combine numerous techniques for better outcomes.

1. Improved Feature Extraction: Sophisticated feature extraction methods can be quite helpful in extracting the semantic and contextual details from memes, which frequently contain nuanced and intricate sexism indications. Using more complex models and methods can improve the quality of the features that are retrieved from memes textual and visual components. By aligning visual and textual elements into a shared embedding space, models that are built to handle both visual and textual data, such as VisualBERT or ViLBERT [13], can be integrated to provide a more thorough knowledge of memes and enable more accurate classification. Furthermore, by catching minute details and patterns that more basic models can overlook, using cutting-edge convolutional neural networks (CNNs) like EfficientNet, which offers a scalable and effective architecture, can enhance feature extraction from images. Graph Neural Networks (GNNs) can also be used to represent the relationships between various components inside a meme, capturing the dependencies and contextual relationships that are essential for comprehending the content, for memes with rich text-image interactions.

2. Dynamic Pre-processing: Preserving the consistency and quality of input features requires building dynamic pre-processing pipelines that can adjust to various data formats and language combinations. More efficiently, the heterogeneity in meme formats and content can be handled by a pre-processing architecture that is adaptable and versatile. By employing techniques like object identification to recognise and preserve the relevant portions of an image, adaptive scaling and cropping algorithms can guarantee that important portions of the images are not destroyed. Additionally, by handling various alphabets, special characters, and idiomatic expressions with customised approaches, normalisation techniques that take into account the unique characteristics of different languages

can improve text processing performance. Increasing the diversity of the training data through the application of data augmentation techniques like random cropping, rotation, and colour modifications can also help to create more resilient models that perform better when applied to previously unseen data.

3. Hybrid Approaches: Rule-based systems and machine learning models together can handle edge cases more skillfully and increase the system's overall accuracy. By combining the best features of probabilistic and deterministic techniques, hybrid approaches can offer a more complete solution. By using specified terms, phrases, or patterns that are suggestive of sexist content, rule-based filters might assist in identifying explicit and evident occurrences of sexism that machine learning models would overlook. By combining the predictions from various models, ensemble methods can increase overall performance by combining the strengths of various models, hence increasing the system's robustness and accuracy. Contextual data, such as user interaction patterns and social network metadata, can also offer extra insights that improve the detection of sexist content by illuminating the environment in which memes are shared and their possible effects.

6. Future Work

To significantly enhance the efficacy and robustness of the sexism detection system, several advanced strategies and techniques can be explored. These improvements focus on various aspects of the model development lifecycle, from data augmentation and pre-processing to model architecture and evaluation metrics.

6.1. Enhanced Data Augmentation

Using Generative Adversarial Networks (GANs) is one interesting way to increase the model's robustness. The current dataset can be enhanced by using GANs to produce artificially realistic yet synthetic meme images. By resolving class imbalance and broadening the pool of training samples, this strategy can improve the model's capacity to generalise across various forms of sexist material. Furthermore, the textual material within memes can be made more diverse by utilising textual data augmentation techniques like synonym replacement, paraphrasing, and back-translation. By ensuring that the model picks up strong features from a variety of linguistic expressions, these techniques raise the accuracy of the model even further.

6.2. Advanced Model Architectures

Text analysis performance in the system can be greatly improved by using transformer-based models, such as BERT [2], RoBERTa [4], and XLM-R. These models are particularly good at capturing contextual subtleties and intricate language semantics, which are essential for identifying nuanced instances of sexism. Additionally, investigating multimodal transformers that incorporate textual and visual inputs, such as VisualBERT or ViLBERT [13], can offer a comprehensive meme analysis. Predictions from several models can also be combined by using ensemble techniques like stacking and blending.[7] By combining the advantages of several models, this method lowers the chance of overfitting while enhancing prediction accuracy. Pre-trained models such as VGG [14], ResNet [15], or EfficientNet can be utilized for image feature extraction, or custom CNN architectures suited to the unique features of meme images can be created.

6.3. Cross-lingual and Multimodal Models

Effective management of multilingual text data requires the use of cross-lingual embeddings, such as Multilingual BERT (mBERT). These embeddings improve the systems worldwide applicability by guaranteeing consistent performance across many languages. Creating shared embedding spaces for text and images through multimodal learning can greatly enhance the models comprehension of memes,

in addition to its cross-lingual capabilities. By capturing complex interactions between textual and visual aspects, pre-training models on big multimodal datasets strengthens the systems ability to interpret meme content.

6.4. Fine-tuning Pre-trained Models

General-purpose models can be tailored to the specifics of the target domain by fine-tuning pre-trained models on domain-specific datasets pertaining to sexism detection and social media analytics. This task-specific fine-tuning increases the relevance and accuracy of the models. Furthermore, the models performance on sexism identification can be improved by utilising layer-wise transfer from models that have already been pre-trained on comparable tasks, such hate speech detection, to leverage transfer learning. This method, which makes use of shared features across related domains, cuts down on training time and costs while offering a strong basis for the new work.

6.5. Multimodal Data Integration

Experimenting with fusion strategies, such as early, late, and hybrid fusion, is essential to capturing complementing information from memes textual and visual aspects. By fusing textual and visual elements, these methods offer a thorough comprehension of memes. Furthermore, the models comprehension of memes in the context of social media can be improved by employing contextual embedding that take into account the memes larger context, including user metadata and engagement metrics. By using this method, the model is guaranteed to capture the entire range of information included in memes, increasing the accuracy of detection [9].

6.6. Improved Evaluation Metrics

It is crucial to keep assessing hierarchical and multilabel classification problems using sophisticated metrics like ICM and ICM-Soft. These measures capture the complexities of sexism detection and offer a detailed assessment of model performance. Furthermore, user studies that assess the systems functionality in real-world situations and collect input for future improvement might yield insightful information. Through a user-centric evaluation, it is ensured that the model meets user expectations and works well in real-world applications.

7. Conclusion

This system for detecting sexist content in memes demonstrated moderate performance in the EXIST 2024 Task 4 shared task. The results indicate that while the model is competitive, there is considerable room for improvement, particularly in the English instances where it ranked lower. The system performed better on Spanish instances, which suggests that the pre-processing and feature extraction steps might be more effective for Spanish language content.
The outcomes highlight the intricacy of the task and the subtlety of sexist content, which presents serious difficulties for automated detection systems. Although this method, which fused sophisticated feature extraction with the CLIP model with conventional machine learning models, provided a strong basis, further improvements will be needed to increase its accuracy and durability.

Acknowledgments

We appreciate the platform that EXIST 2024 shared task organisers provided to further study on the identification of sexist content in social networks. I also like to express my appreciation to the annotators for their work in labelling the dataset, which made it possible to perform this study.

References

[1] A. Chaudhary, R. Kumar, Sexism identification in social networks., in: CLEF (Working Notes), 2023, pp. 891–900.

[2] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of naacL-HLT, volume 1, 2019, p. 2.

[3] A. DeLucia, S. Wu, A. Mueller, C. Aguirre, P. Resnik, M. Dredze, Bernice: A multilingual pre-trained encoder for twitter, in: Proceedings of the 2022 conference on empirical methods in natural language processing, 2022, pp. 6191–6205.

[4] 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, arXiv preprint arXiv:1907.11692 (2019).

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[6] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).

[7] S. Butt, N. Ashraf, G. Sidorov, A. F. Gelbukh, Sexism identification using bert and data augmentation-exist2021., in: IberLEF@ SEPLN, 2021, pp. 381–389.

[8] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).

[10] A. F. M. de Paula, R. F. da Silva, I. B. Schlicht, Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models, 2021. URL: https://arxiv.org/abs/2111.04551. arXiv:2111.04551.

[11] 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 Social Networks and Memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[12] 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 Social Networks and Memes (Extended Overview), in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[13] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, arXiv preprint arXiv:1908.03557 (2019).

[14] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.