The Wisdom of Weighing: Stacking Ensembles for a More Balanced Sexism Detector

Abhay Shanbhag[1,†], Suramya Jadhav[1,†], Atharva Date[1,†], Sumedh Joshi[1,†] and Sheetal Sonawane[1,†]

[1] *SCTR's Pune Institute of Computer Technology*

Abstract

The rise of sexism online has become increasingly prevalent as more and more people are found on social media with little regard for what they say behind a screen of anonymity. Sexism in the form of derogatory, biased, violent, and presumptuous remarks, apart from perpetuating gender inequality, also creates a hostile environment for women that calls for immediate attention. EXIST 2024, the fourth edition of the sEXism Identification in Social Networks task at CLEF 2024, aims to not only detect sexism but also capture its types, from explicit misogyny to other subtle expressions that involve implicit sexist behaviors. We provided solutions for three tasks, the first of which was the identification of sexism in both English and Spanish texts, whereas the second and third identified the more subtle categories and aspects of sexism. In this study, we introduce a robust classification system utilizing a stacking classifier composed of four LLMs, whose output probabilities feed into a LightGBM model to produce a consolidated prediction. Additionally, five supplementary models contribute to the final decision by providing weighted predictions based on their respective accuracies. This ensemble approach, leveraging both stacking and weighted averaging, ensures enhanced accuracy and reliability in classifying text as sexist or non-sexist.

Keywords

Sexism Identification, Source Intention, Sexism Categorization, BERT, RoBERTa, Ensemble Approach

1. Introduction

Sexism on social media is a widespread problem that reinforces prejudices and discrimination based on gender. Social media sites such as Facebook, Instagram, Twitter,etc. function as forums for the discussion and expression of sexist beliefs. This problem encompasses a variety of actions, ranging from blatantly sexist statements to subtly discriminatory small talk that reinforces stereotypical gender norms. Such content spreads easily thanks to social media's anonymity and wide audience, which frequently makes it challenging to monitor successfully. Women are disproportionately the victims of cyberbullying, where they face a variety of sexist acts such as body shaming, threats of violence, and disparaging remarks . Technologies like better content filtering algorithms and social interventions like victim support groups and public awareness campaigns are used in the fight against online misogyny. In order to create safer and more inclusive digital spaces, it is imperative to recognize and address sexism on social media.

In this Shared task, EXIST 2024 organised by Plaza et al. [1] and Plaza et al. [2], our research work aims to classify whether a given tweet contains sexist expressions or behaviours (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behaviour), and classify it according to two categories: YES and NO. The second subtask is a multi-class classification. For the tweets that have been predicted as sexist,the second task aims to classify each tweet according to the intention of the person who wrote it. One of the three following categories must be assigned to each sexist tweet: DIRECT, REPORTED, JUDGEMENTAL. The third subtask is a multi-label classification. For the tweets that have been predicted as sexist, the third task aims to categorize them according to the type of sexism. This is a multi-label task, so that more than one of the following labels may be assigned

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

[†] These authors contributed equally.

abhayshanbhag0110@gmail.com (. A. Shanbhag); 2018suramyajadhav@gmail.com (S. Jadhav); atharvad931@gmail.com (A. Date); sumedhjoshi463@gmail.com (S. Joshi); sssonawane@pict.edu (S. Sonawane)

to each tweet: IDEOLOGICAL-INEQUALITY, STEREOTYPING-DOMINANCE, OBJECTIFICATION, SEXUAL-VIOLENCE, MISOGYNY-NON-SEXUAL-VIOLENCE.

2. Background and Dataset

We participated in tasks 1, 2, and 3. Total of 3,660 of tweets are in Spanish, and the rest of them in English. The annotated information helps in understanding how various demographic factors of annotators might influence the interpretation and classification of sexist content on social media, thus contributing to the development of more nuanced and effective detection algorithms.

```
"100001": {
"id_EXIST": "100001",
"lang": "es",
"tweet": "@TheChiflis Ignora al otro, es un capullo.El problema con este
 youtuber denuncia el acoso... cuando no afecta a la gente de izquierdas.
 Por ejemplo,en su video sobre el gamergate presenta como \"normal\" el
 acoso que reciben Fisher, Anita o Zöey cuando hubo hasta amenazas
 de bomba.",
"number_annotators": 6,
"annotators": ["Annotator_1", "Annotator_2", "Annotator_3",
 "Annotator_4", "Annotator_5", "Annotator_6"],
"gender_annotators": ["F", "F", "F", "M", "M", "M"],
"age_annotators": ["18-22", "23-45", "46+", "46+", "23-45", "18-22"],
"ethnicities_annotators": ["White or Caucasian", "Hispano or Latino",
 "White or Caucasian", "White or Caucasian", "White or Caucasian", "Hispano
 or Latino"],
"study_levels_annotators": ["Bachelor’s degree", "Bachelor’s degree",
 "High school degree or equivalent", "Master’s degree",
"Master’s degree", "High school degree or equivalent"],
"countries_annotators": ["Italy", "Mexico", "United States", "Spain",
 "Spain", "Chile"],
"labels_task1": ["YES", "YES", "NO", "YES", "YES", "YES"],
"labels_task2": ["REPORTED", "JUDGEMENTAL", "-", "REPORTED",
 "JUDGEMENTAL", "REPORTED"],
"labels_task3": [
   ["OBJECTIFICATION"],
   ["OBJECTIFICATION", "SEXUAL-VIOLENCE"],
   ["-"],
   ["STEREOTYPING-DOMINANCE"],
   ["SEXUAL-VIOLENCE"],
   ["IDEOLOGICAL-INEQUALITY", "MISOGYNY-NON-SEXUAL-VIOLENCE"]
]
}
```

Above is a glimpse of dataset provided by the organisers.

3. Related Work

Angel et al. [3] presents an innovative approach to effectively handling the Exist2023 dataset, a collection comprising both English and Spanish, with a focus on the informal language typical of social media platforms like Twitter, including emojis and hashtags. Their methodology introduces contrastive learning into the traditional fine-tuning language model pipeline, marking a departure from conventional approaches. Unlike standard fine-tuning methods aimed at learning an embedding space where similar samples cluster together, their contrastive learning technique adopts a regression setting. Here, labels are determined based on the fraction of annotators who agree with a specific classification, thereby accommodating the subjectivity and diversity of opinions inherent in the dataset.

Moreover, they leverage the diverse annotations from different annotators to enhance the prediction process, particularly in tasks such as sexism identification, which they frame as a regression problem predicting the fraction of annotators labeling a tweet as containing sexism. They then apply a threshold rule to convert these predictions into binary labels (i.e., "YES" or "NO"). Their study includes the submission of three variations: FT (Fine-Tuning), which entails standard fine-tuning of the language model; FreezeCL, where contrastive learning precedes fine-tuning, followed by freezing the model and training only the classifier head; and UnfreezeCL, similar to FreezeCL but allowing updates to all model parameters during fine-tuning. For contrastive learning, they conduct training for 10 epochs with a learning rate of 5e-5 and a batch size of 32, followed by fine-tuning for up to 20 epochs, matching the 30-epoch setting of FT, with early stopping, a learning rate of 1e-5, and a batch size of 128. They employ the AdamW optimizer and the transformers library for training on an NVIDIA V100 GPU with 32GB memory. The model with the lowest root mean square error (rmse) score on the validation set is saved for further evaluation.

In EXIST 2023, several teams participated in the competition, each employing various approaches to tackle the tasks. Some teams utilized multilingual models such as XLM-RoBERTa and BERT for classification tasks, demonstrating the versatility of these models across different languages. They incorporated both hard and soft labels, leveraging data augmentation techniques and fine-tuning on task-specific datasets to improve performance. In particular, Ersoy et al. [4] employed a cascade model using the output from Task 1 for Task 3, showcasing an effective transfer learning strategy. Vetagiri et al. [5] focused on training models exclusively on the provided dataset, generating both hard and soft labels for Task 1. Erbani et al. [6] took a fine-tuning approach on separate BERT models for each task, incorporating manual features and concatenating representations for improved classification. Additionally, Cordon et al. [7] experimented with ensemble models, including bert-large-uncased and distilbert, to enhance performance in Task 1. Chaudhary and Kumar [8] utilized a Bi-LSTM architecture for hard predictions in Task 1, while de Paula and da Silva [9] employed a variety of transformer models for both English and Spanish tasks. Hatekar et al. [10] optimized various models from HuggingFace and employed multilingual models with data augmentation techniques. Finally, Buzzell et al. [11] explored various approaches including SVM with TF-IDF and CNN models for Task 1, showcasing diversity in methodology among the participating teams.

Rodríguez-Sánchez et al. [12] explored a variety of machine learning methods for the task of sexism detection. They compared the performance of logistic regression, SVM, random forest, bi-LSTMs, and mBERT on sexism detection in Spanish tweets. They found the neural models to be slightly better than the non-neural machine learning algorithms at detecting sexism in the dataset, although random forest achieved the highest precision. The bi-LSTM models were on par with mBERT in terms of F1, accuracy, precision, and recall. Rizvi and Jamatia [13] participated in the 2022 EXIST shared task Rodríguez-Sánchez et al. [14]. They experimented with logistic regression, Naive Bayes, and SVM systems and found that the logistic regression model worked best for both Spanish and English on both tasks. They used TF-IDF unigram and bigram representations as features for all three models. While their submission ultimately ranked 17th out of 19 submissions in the competition, with an official F1-score of 70.65% overall, their approach showed promise among the few submissions that did not implement pretrained transformer-based models.

Moldovan et al. [15] addressed the issue of sexism in Romanian. They used logistic regression, SVM, random forests, Ro-BERT, and mBERT to classify Romanian tweets as sexist or nonsexist. They used BOW-based representations, TF-IDF word representations, and sentence representations generated by mBERT and Ro-BERT as features for the non-neural models. The best performance was achieved with a fine-tuned Ro-BERT model; however, the best recall for non-sexist tweets was achieved by the random forest classifier using TF-IDF-based word representations, by a significant margin. Related to the topic of sexism detection is abusive language detection. Steimel et al. [16] investigated abusive language detection in English and German tweets using topic modeling and a number of neural and non-neural classifiers. They found that SGBoost performed best on the English data, while SVMs performed best on the German data. They also found that different sampling methods to address class imbalances led to drastically different outcomes regarding the two data sets. Their work provides evidence that the best classifier and techniques for one language cannot be assumed to perform well for other languages, even if the data sets share similarities. Thus, it is important to experiment with a variety of methods when handling multilingual data.

4. System Description

4.1. Text Preprocessing

A range of datasets, such as the Kirk et al. [17] EDOS dataset,Guest et al. [18] Misogyny, Rodríguez-Sánchez et al. [12] MeTwo and training datasets supplied by the organizers were used to fine-tune our models. While some of the models in our work are designed for Spanish text, some are limited to English. We separately translated all of the data into English and Spanish because our approach makes use of language-specific models. Siino et al. [19] mention the importance of preprocessing and showcase that Using a proper preprocessing strategy, simple models can outperform transformers in text classification. tasks. The preprocessing methods employed for our research work were lowercasing, eliminating mentions (like @username), eliminating hashtags, eliminating links, eliminating numerals, eliminating punctuation, and eliminating non-ASCII symbols. Hickman et al. [20] focus on providing empirically grounded recommendations for text preprocessing decision-making in text mining, considering the type of text mining, the research question, and dataset characteristics to enhance the validity and transparency of insights derived from natural language text data. Further preprocessing processes like stemming, Lemmatizing and stopword removal are redundant for the applicable contextualized transformer models. and doing so might negatively impact their performance. Following this preprocessing, two datasets were produced: one with all English text and the other with all Spanish text, both labeled as sexist and Non-Sexist in order to allow for fine tuning.

4.2. Task 1 : Identification of Sexism

4.2.1. Approach-1 : Finetuning English Model

Identifying subtle, context-specific instances of discriminatory language is the initial stage of sexism.detection. Sanh et al. [21] proposed DistilBERT, which is an excellent choice for sexism detection.task since it exhibits a high degree of proficiency in recognizing subtle verbal patterns. DistilBERT is a condensed version of BERT that retains 97% of BERT's language comprehension while being 40%smaller and 60% faster. Through knowledge distillation, a smaller model (DistilBERT) learns to replicate the behavior of a larger model (BERT) by focusing on its essential components. DistilBERT's compact and efficient architecture is especially helpful for tasks like sexism detection, where it's necessary to parse language subtly in order to identify discriminatory and biased remarks. DistilBERT's transformer design effectively captures context and dependencies in text, enabling it to distinguish between benign and harmful words. To create an English dataset for our study, we followed the instructions in Section4.1 to preprocess and integrate existing datasets. We then used this English dataset to fine-tune the DistilBERT model utilizing the hyperparameters shown in Table 1. As a result, the model performed better at interpreting the particular subtleties and variances associated with sexism than generic models.

4.2.2. Approach-2 : Finetuning Spanish Model

Sanh et al. [21] indicates RoBERTa is a powerful option for sexism detection because of its exceptional comprehension and analysis of delicate and complicated information. The transformer design is

Table 1
Parameters for Finetuning DistilBERT

Parameter	Value
learning_rate	2e-5
train_batch_size	16
eval_batch_size	16
seed	42
optimizer (Adam with betas)	0.9, 0.999
epsilon	1e-08
weight decay	0.01
num_train_epochs	5

Table 2
Parameters for Finetuning RoBERTuito

Parameter	Value
learning_rate	1e-5
train_batch_size	32
eval_batch_size	32
seed	123
optimizer (Adam with betas)	0.9, 0.999
epsilon	1e-08
weight decay	0.001
num_train_epochs	10

enhanced by a BERT version known as RoBERTa, which uses a strong encoder mechanism to effectively extract contextual information from text. This model's extensive pretraining on a large corpus of text makes it highly adept at identifying and evaluating minute biases and differentiating linguistic patterns. RoBERTa uses dynamic masking and larger mini-batches during training to improve the sensitivity and accuracy of its sexist content detection and classification. Because RoBERTa can represent complicated syntactic and semantic relationships and handle a wide range of linguistic terminology, it is a useful tool for detecting sexism. We opted to use the model developed by Liu et al. [22] as our version of RoBERTuito, which is based on RoBERTa. The model is accessible through HuggingFace and is fine-tuned on the EXIST2021 dataset. In our work we optimized the RoBERTuito model in our study using the hyperparameters indicated in Table 2 on all of the Spanish data that we obtained after processing and translating the external datasets specified in the 4.1 section into Spanish. As a result, the model outperformed generic models in its ability to comprehend the dataset's particular complexities and variations. This helped us navigate possible sexism that was presented in Spanish, which enhanced our approach and increased the detection accuracy.

4.2.3. Proposed Approach : Ensembling Multiple Models

In our proposed methodology, we have fine-tuned 3 different models: twitter-roberta-base-hate[5], distilbert-base-uncased [6], twitter-sexismo-finetuned-robertuito-exist2021[7] and have also directly used 2 pretrained models: xlm-roberta-base-misogyny-sexism-indomain-mix-bal[8], twitter-sexismo-finetuned-exist2021-metwo[9] from hugging face for sexism identification by ensembling the classifiers. The idea of creating a voting ensemble from neural classifiers has been explored by Siino et al. [23]. Let's see each one of them below. We fine-tuned the pretrained Roberta-twitter-hate model by Barbieri et al. [24] that had been trained for hate speech detection on our dataset using the parameters given in Table 3. Since this model was initially trained for hate speech detection, it had a better tendency to correctly classify sexist tweets that were aggressive in nature. A fine-tuned version of this model is thranduil2/results [10]. Next, we fine-tuned the distilbert-base-uncased [6] model on the existing dataset. Distilbert was used because of its high reliability and its lightweight architecture for text classification, and our new model now is thranduil2/sexismDistilbert [11]. Finally, for handling Spanish instances, we used the somosnlp-hackathon-2022/twitter-sexismo-finetuned-robertuito-exist2021, which was already finetuned on sexism-based tasks. Finally, we got NewSpanishFinetunedtrainepoch10 [12]. Croce et al. [25] employed a text vectorization layer to create Bag-of-Words sequences, which were then utilized by three distinct text classifiers (Decision Tree, Convolutional Neural Network, and Naive Bayes), culminating in the use of an SVM as the final classifier. Kang et al. [26] proposed an ensemble of text-based hidden Markov models using boosting and clusters of words produced by latent semantic analysis. We created an ensemble of four models that had the highest complementary error correction and trained a stacking classifier on the output scores of these models. For the stacking classifier, we used LightGBM, which is a lightweight and powerful gradient boosting model using the best parameters

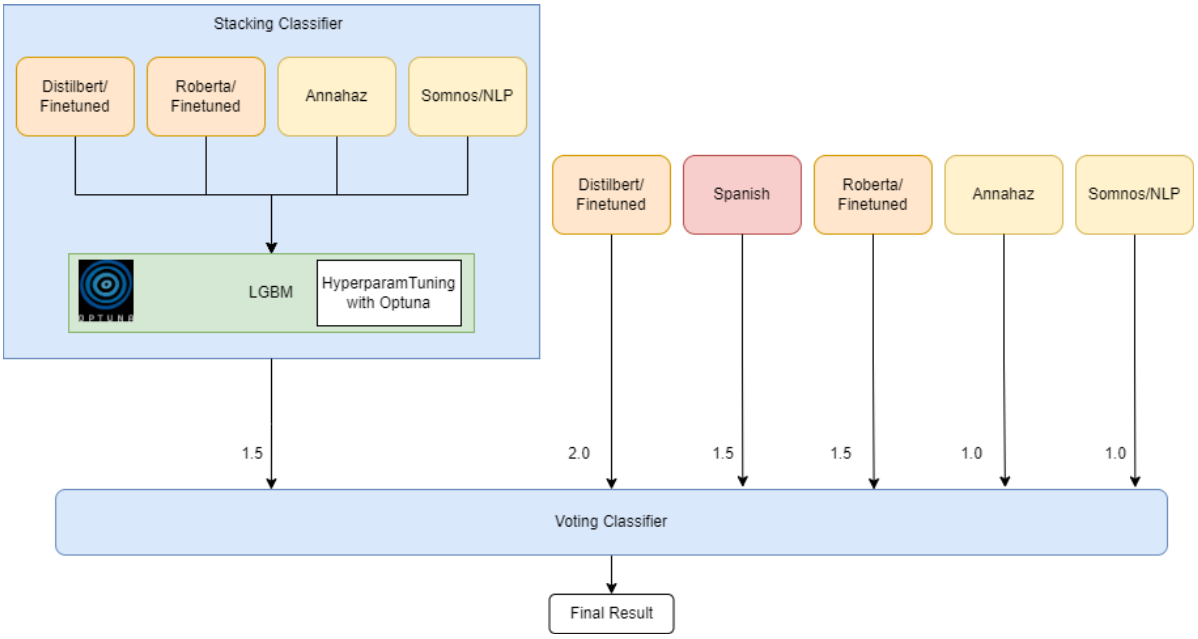

Figure 1: Architecture Diagram. Here models Distilbert finetuned and Roberta finetuned are finetuned models sexismDistilbert[11] and results[10] respectively. Annahaz[8] and SomnosNLP [9] are pretrained models from Hugging Face. The 'Spanish' model is also finedtuned model NewSpanishFinetunedtrainepoch10 [12]. For hyperparameter tuning of our LGBM, we used Optuna[4] framework for determining the correct set of hyperparameters. Numbers on arrow represent the weights for each of the votes given.

obtained by hyperparameter tuning as mentioned in Table 4.

Table 3
Training Parameters

Parameter	**Value**
Learning Rate	2×10^{-5}
Number of Training Epochs	10
Weight Decay	0.1
Per Device Train Batch Size	16

Table 4
Best Hyperparameters

Parameter	**Value**
λ_{L1}	$6.191822954187258 \times 10^{-8}$
λ_{L2}	0.3564434
Number of Leaves	253
Feature Fraction	0.9053014
Bagging Fraction	0.7948063
Bagging Frequency	1
Min Child Samples	5

To illustrate the functionality of our proposed architecture, we present the following operational workflow: Initially, a text sample is input into the system, where it undergoes analysis by a stacking classifier comprising four distinct models. Each model independently classifies the text as either sexist (1) or non-sexist (0). The output probabilities from these four models serve as input features for the LightGBM (LGBM) classification model, which subsequently synthesizes these inputs to produce a singular prediction of either 1 (sexist) or 0 (non-sexist). In addition to the stacking classifier, the proposed architecture integrates five supplementary models. These models are also tasked with classifying the same text, each outputting a prediction of either 1 (sexist) or 0 (non-sexist). To ensure a balanced and accurate final prediction, each of the six models—including the LGBM stacking classifier—is assigned a weight based on its accuracy performance. The predictions (0 or 1) from each model are then multiplied

[4] Optuna
[5] cardiffnlp/twitter-roberta-base-hate
[6] distilbert/distilbert-base-uncased
[7] somosnlp-hackathon-2022/twitter-sexismo-finetuned-robertuito-exist2021
[8] annahaz/xlm-roberta-base-misogyny-sexism-indomain-mix-bal
[9] somosnlp-hackathon-2022/twitter-sexismo-finetuned-exist2021-metwo

by their respective weights. The resulting weighted outputs are subsequently averaged. The decision rule applied to the averaged output is straightforward: if the mean value exceeds 0.5, the text is classified as sexist (1); if the mean value is 0.5 or less, the text is classified as non-sexist (0). This methodology ensures that the final classification leverages the strengths of multiple models, enhancing the overall accuracy and reliability of the system's predictions. Finally, we ensembled the Stacking Classifier itself with our best-performing fine-tuned Distilbert model as well as the fine-tuned Roberta-Twitter-Hate model, along with the fine-tuned Spanish RoBERETuito model, and passed the six models to a voting classifier. We performed a weighted average of the predictions of the six models to get our final prediction.

4.3. Task 2 : Source Intention of Sexism

BERT (Bidirectional Encoder Representations from Transformers), given by Devlin et al. [27], is a state-of-the-art natural language processing (NLP) model known for its exceptional performance across various language understanding tasks. Its key strength lies in its ability to capture bidirectional contextual information, enabling a nuanced understanding of word meaning by considering both preceding and succeeding contexts. Pre-trained on extensive text corpora using self-supervised learning tasks, such as masked language modeling and next sentence prediction, BERT learns rich semantic representations of language. This pre-training allows for fine-tuning on specific tasks, making BERT highly adaptable and effective for multi-class classification tasks. Its capability to comprehend complex relationships within text, coupled with its contextual understanding, makes it an ideal choice for tasks requiring nuanced classification of textual data. The BERT model was fine-tuned using the existing dataset for task 2, and the parameters used are given in Table 5.

Table 5
Task 2 Parameters

Parameter	Value
learning_rate	5e-05
train_batch_size	16
eval_batch_size	32
seed	42
optimizer (Adam with betas)	0.9, 0.999
epsilon	1e-08
lr_scheduler_type	linear
num_epochs	5

Table 6
Task 3 Parameters

Parameter	Value
learning_rate	5e-05
train_batch_size	16
eval_batch_size	16
seed	42
optimizer (Adam with betas)	0.9, 0.999
epsilon	1e-08
lr_scheduler_type	linear
num_epochs	7

4.4. Task 3 : Categorisation of Sexism in Tweets

For task3 the BERT model given by Devlin et al. [27] was finetuned using BERT Tokenizer on EXIST dataset for task 3 data with parameters as mentioned in Table 6. All the dataset was pre-processed and converted to English first using Google Cloud API.

5. Metrics Used

ICM is a similarity function that generalizes point-wise mutual information (PMI) and can be used to evaluate system outputs in classification problems by computing their similarity to the ground truth categories. The general definition of ICM is:

$$\mathrm{ICM}(A, B) = \alpha_1 \, \mathrm{IC}(A) + \alpha_2 \, \mathrm{IC}(B) - \beta \, \mathrm{IC}(A \cup B) \tag{1}$$

[10] thranduil2/results

[11] thranduil2/sexismDistilbert

[12] Suramya/NewSpanishFinetunedtrainepoch10

Where IC(A) is the information content of the item represented by the set of features A, etc. ICM maps into PMI when all parameters take a value of 1. In Amigó and Delgado, the general ICM definition is applied to cases where categories have a hierarchical structure and items may belong to more than one category. The resulting evaluation metric is proven to be analytically superior to the alternatives in the state of the art. The definition of ICM in this context is:

$$\mathrm{ICM}(s(d), g(d)) = 2I(s(d)) + 2I(g(d)) - 3I(s(d) \cup g(d)) \tag{2}$$

Where I() stands for Information Content, s(d) is the set of categories assigned to document d by system s, and g(d) the set of categories assigned to document d in the gold standard.

$$I(\{\langle c, v\rangle\}) = -\log_2 \left(P\left(\{d \in D : g_c(d) \geq v\}\right)\right) \tag{3}$$

6. Results

In our study, we employed a series of approaches aimed at enhancing the accuracy of our model. Different Approaches and their Rankings with corresponding score can be seen in Table 7. In this table 1a represents model explained in Section 4.2.1, 1b is the model presented in Section 4.2.2 and 1c is proposed model which uses ensembling concept as explained in Section 4.2.3 for Task 1. 2a and 3a are the models made by finetuning BERT whose details are mentioned in section 4.3 and 4.4 respectively. EN and ES denote the English and Spanish languages respectively, indicating our models performances on these specific texts present in the test dataset.

Table 7
Results Table of our team **maven**

		Teamwise Rank	Approach	ICM-Hard	ICM-Hard Norm	F1 SCORE
EN	task1	12	1a	0.3395	0.6732	0.6747
			1b	0.1926	0.5983	0.6512
			1c	0.5107	0.7606	0.7359
	task2	15	2a	-0.1333	0.4539	0.4123
	task3	8	3a	-0.3093	0.4242	0.4245
ES	task1	14	1a	0.4129	0.7065	0.7394
			1b	0.4129	0.7065	0.7394
			1c	0.4857	0.7429	0.7784
	task2	14	2a	-0.0033	0.499	0.4859
	task3	6	3a	-0.2401	0.4464	0.4662

7. Limitations and Future work

7.1. Limitations

Pre-trained transformer-based models often struggle with sarcasm detection, leading to misclassification of such instances. Specialized handling of sarcastic comments is necessary to improve model accuracy. In tasks 2 and 3, where data availability for discerning source intention is limited, the model tends to exhibit lower accuracy rates. Increasing the volume of data for fine-tuning enables the model to grasp finer nuances and glean hidden insights more effectively. Additionally, the model's performance is heavily contingent on the quality of the data it is trained on. Contextual ambiguity and the presence of code-mixed sentences, where multiple languages are used within the same sentence, diminish the model's efficiency. Therefore, ensuring clear context and providing diverse, high-quality data are crucial steps in enhancing the model's overall performance.

7.2. Future Work

In future work, employing Large Language Models (LLMs) for detection purposes could significantly enhance the performance of sarcasm detection systems. Furthermore, exploring varying ensembling methods such as Bagging Classifiers employed by Chen et al. [28]. These ensembling techniques can also be extended to downstream multi-class and multi-label classification tasks as demonstrated by Miri et al. [29]. Moreover, exploring advanced aggregation methods, such as incorporating probabilistic means from multiple models and employing dedicated algorithms for assigning weights, holds promise for improving overall accuracy. It's imperative to select models that complement each other's False Positive (FP) and False Negative (FN) rates, ensuring comprehensive coverage of cases without bias towards any particular outcome. Additionally, integrating data augmentation techniques, such as introducing noise to text or scraping data from diverse social media platforms and video transcripts, can substantially augment the dataset, thereby enriching the model's training corpus and potentially enhancing its robustness.

8. Conclusion

In this research work, we aimed at addressing the problem of sexism detection, categorization and finding intention in tweets with both English and Spanish content. As social media grows more and more ingrained in our daily lives, sexism continues to be a significant social issue that is drawing increasing attention. Today, it is imperative to address and minimize misogyny on these platforms. Given this, our study aimed to develop efficient multilingual sexism identification systems by fine-tuning models and utilizing ensembles of several models. Quality external datasets are beneficial as they improve performance of models. This outcome highlights how crucial it is to look at possible integration of ensemble-based methods into traditional pipelines in order to further produce accurate outputs. Oversampling techniques allowed for the prevention of model bias for each unique class group. Simple Aggregation of ensemble based and transformer model gave better results as compared to their individual performance.

References

[1] 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.

[2] 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.

[3] J. Angel, S. T. Aroyehun, A. F. Gelbukh, Multilingual sexism identification using contrastive learning, in: Conference and Labs of the Evaluation Forum, 2023. URL: https://api.semanticscholar.org/CorpusID:264441422.

[4] B. I. Ersoy, G. Radler, S. Carpentieri, Classifiers at exist 2023: Detecting sexism in spanish and english tweets with xlm-t, in: Conference and Labs of the Evaluation Forum, 2023. URL: https://api.semanticscholar.org/CorpusID:264441369.

[5] A. Vetagiri, P. K. Adhikary, D. P. Pakray, A. Das, Leveraging gpt-2 for automated classification of online sexist content, 2023.

[6] J. Erbani, E. Egyed-Zsigmond, D. Nurbakova, P.-E. Portier, When multiple perspectives and an optimization process lead to better performance, an automatic sexism identification on social media

with pretrained transformers in a soft label context, in: Conference and Labs of the Evaluation Forum, 2023. URL: https://api.semanticscholar.org/CorpusID:264441461.

[7] P. Cordon, J. Mata, V. Pachón, J. L. Domínguez, I2c-uhu at clef-2023 exist task: Leveraging ensembling language models to detect multilingual sexism in social media, in: Conference and Labs of the Evaluation Forum, 2023. URL: https://api.semanticscholar.org/CorpusID:264441322.

[8] A. Chaudhary, R. Kumar, Sexism identification in social networks, in: Conference and Labs of the Evaluation Forum, 2023. URL: https://api.semanticscholar.org/CorpusID:264441505.

[9] A. F. M. de Paula, R. F. da Silva, Detection and classification of sexism on social media using multiple languages, transformers, and ensemble models, in: IberLEF@SEPLN, 2022. URL: https://api.semanticscholar.org/CorpusID:252015736.

[10] Y. A. Hatekar, M. S. Abdo, S. Khanna, S. Kübler, Iuexist: Multilingual pre-trained language models for sexism detection on twitter in exist2023, in: Conference and Labs of the Evaluation Forum, 2023. URL: https://api.semanticscholar.org/CorpusID:264441455.

[11] M. Buzzell, J. Dickinson, N. Singh, S. Kübler, Iu-nlp-jedi: Investigating sexism detection in english and spanish, in: Conference and Labs of the Evaluation Forum, 2023. URL: https://api.semanticscholar.org/CorpusID:264441789.

[12] 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.

[13] A. Rizvi, A. Jamatia, Nit-agartala-nlp-team at exist 2022: Sexism identification in social networks, in: IberLEF@SEPLN, 2022. URL: https://api.semanticscholar.org/CorpusID:252015919.

[14] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2022: sexism identification in social networks, Proces. del Leng. Natural 69 (2022) 229–240. URL: https://api.semanticscholar.org/CorpusID:239690039.

[15] A. Moldovan, K. Csuros, A.-m. Bucur, L. Bercuci, Users hate blondes: Detecting sexism in user comments on online romanian news, 2022, pp. 230–230. doi:10.18653/v1/2022.woah-1.21.

[16] K. Steimel, D. Dakota, Y. E. Chen, S. Kübler, Investigating multilingual abusive language detection: A cautionary tale, in: Recent Advances in Natural Language Processing, 2019. URL: https://api.semanticscholar.org/CorpusID:210063047.

[17] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, Semeval-2023 task 10: Explainable detection of online sexism, ArXiv abs/2303.04222 (2023). URL: https://api.semanticscholar.org/CorpusID:257405434.

[18] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, H. Margetts, An expert annotated dataset for the detection of online misogyny, 2021, pp. 1336–1350. doi:10.18653/v1/2021.eacl-main.114.

[19] M. Siino, I. Tinnirello, M. La Cascia, Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers, Information Systems 121 (2023) 102342. doi:10.1016/j.is.2023.102342.

[20] L. Hickman, S. Thapa, L. Tay, M. Cao, P. Srinivasan, Text preprocessing for text mining in organizational research: Review and recommendations, Organizational Research Methods 25 (2020) 114 – 146. URL: https://api.semanticscholar.org/CorpusID:229501282.

[21] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019). URL: https://api.semanticscholar.org/CorpusID:203626972.

[22] 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 abs/1907.11692 (2019). URL: https://api.semanticscholar.org/CorpusID:198953378.

[23] M. Siino, I. Tinnirello, M. L. Cascia, T100: A modern classic ensemble to profile irony and stereotype spreaders, in: Conference and Labs of the Evaluation Forum, 2022. URL: https://api.semanticscholar.org/CorpusID:262324480.

[24] F. Barbieri, J. Camacho-Collados, L. Neves, L. Espinosa-Anke, Tweeteval: Unified benchmark and comparative evaluation for tweet classification, ArXiv abs/2010.12421 (2020). URL: https://api.semanticscholar.org/CorpusID:225062026.

[25] D. Croce, D. Garlisi, M. Siino, An svm ensemble approach to detect irony and stereotype spreaders

on twitter, in: Conference and Labs of the Evaluation Forum, 2022. URL: https://api.semanticscholar.org/CorpusID:251471680.

[26] M. Kang, J. Ahn, K. Lee, Opinion mining using ensemble text hidden markov models for text classification, Expert Systems with Applications 94 (2018) 218–227. URL: https://www.sciencedirect.com/science/article/pii/S0957417417304979. doi:https://doi.org/10.1016/j.eswa.2017.07.019.

[27] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: North American Chapter of the Association for Computational Linguistics, 2019. URL: https://api.semanticscholar.org/CorpusID:52967399.

[28] H. Chen, Z. Zhang, S. Huang, J. Hu, W. Ni, J. Liu, Textcnn-based ensemble learning model for japanese text multi-classification, Comput. Electr. Eng. 109 (2023) 108751. URL: https://api.semanticscholar.org/CorpusID:258900728.

[29] M. Miri, M. B. Dowlatshahi, A. Hashemi, M. K. Rafsanjani, B. B. Gupta, W. S. Alhalabi, Ensemble feature selection for multi-label text classification: An intelligent order statistics approach, International Journal of Intelligent Systems 37 (2022) 11319 – 11341. URL: https://api.semanticscholar.org/CorpusID:252056607.