SINAI at SemEval-2023 Task 10: Leveraging Emotions, Sentiments, and Irony Knowledge for Explainable Detection of Online Sexism

M. Estrella Vallecillo-Rodríguez, Flor Miriam Plaza-del-Arco
L. Alfonso Ureña-López, M. Teresa Martín-Valdivia

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{mevallecillo, fmplaza, laurena, maite}@ujaen.es

Abstract

This paper describes the participation of SINAI research team in the Explainable Detection of Online Sexism (EDOS) Shared Task at SemEval 2023. Specifically, we participate in subtask A (binary sexism detection), subtask B (category of sexism), and subtask C (fine-grained vector of sexism). For the three subtasks, we propose a system that integrates information related to emotions, sentiments, and irony in order to check whether these features help detect sexism content. Our team ranked 46th in subtask A, 37th in subtask B, and 29th in subtask C, achieving 0.8245, 0.6043, and 0.4376 of macro f1-score, respectively, among the participants.

1 Introduction

Sexism is defined by The Oxford Dictionary as prejudice, stereotyping, or discrimination, typically against women, on the basis of sex[1]. Sexism is in our daily life when people underestimate the opinions expressed by women, in conversations both oral or written that contains sayings and fixed expressions. Nowadays, with social media platforms, sexist comments are very frequent and they are widespread quickly promoting more sexist reactions. Furthermore, sexism can be expressed in different forms, making detecting sexist comments a difficult task.

Due to this reason, several shared tasks and many academic events are organized by the scientific community. For example, HateEval (Basile et al., 2019) is based on hate speech detection against women and immigrants. Other tasks such as EVALITA (Fersini et al., 2018a) or the task on Automatic Misogyny Identification (AMI) (Fersini et al., 2018b) are focused on misogyny detection. However, sexism is not only misogyny (violence against women) which detection consists of a binary classification task, but also contains different types of sexism depending on how it is expressed, making the detection of sexism a multi-class classification task. For this reason, there are shared tasks such as EXIST (Rodríguez-Sánchez et al., 2022) that promote the development of tools to detect sexism and categorize it according to the facet of the women that are undermined or the EDOS shared task.

This paper describes the participation of SINAI in the Explainable Detection of Online Sexism (EDOS) shared task (Kirk et al., 2023) aimed at the identification of sexist language and its categories. The main purpose of this shared task is to promote the development of different English language models to detect sexism and the explainability of this phenomenon. For this purpose, the organizers proposed three different subtasks. Subtask A: Binary Sexism Detection is related to identifying if a comment is sexist or not. Once a post has been classified as sexist, the subtask B: Category of sexism consists of classifying the sexist posts into four different classes, and subtask C: Fine-grained Vector of Sexism aims to predict one class between eleven different classes that are more specific than the subtask B. Our team SINAI has participated in the three subtasks. Our proposal for addressing these subtasks is the integration of external knowledge into classifiers to more accurately predict the specific task. For example, in hate speech detection some previous works have incorporated knowledge from sentiment and emotion analysis which are related to the expression of offensiveness. For instance, (Plaza-del-Arco et al., 2021, 2022; Halat et al., 2022) proposed a novel approach that uses a multi-task learning paradigm to combine different phenomena that are inextricably related to the expression of offensive language such as sentiments, emotions, target, irony, sarcasm, and constructiveness, among others. Pérez et al. (2022) evaluated the impact of incorporating contextual information in hate speech related to the news posted on so-

[1] https://www.oed.com/

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 986–994

cial media. In misogyny detection, Frenda et al. (2018) introduce an approach based on aesthetic features captured by character n-grams, sentiment information, and a set of lexicons built by analyzing misogynistic tweets.

The rest of the paper is structured as follows: In Section 2 we describe the main strategy used to develop the system for the shared task. The data and the experimental methodology are described in Section 3. The evaluation results from the development and test phases are shown in Section 4. Finally, we conclude with a discussion in Section 5.

2 System Overview

In this section, we describe the system we develop for the Explainable Detection of Online Sexism shared task at SemEval 2023.

We propose a system that incorporates some extra-linguistic information related to emotions, sentiments, irony, or a combination of these features, such as emotions + sentiments or emotions + sentiments + irony, of the posts in the dataset, provided by the organizers of the task. The reason for incorporating the information of polarity, emotion, and irony to detect sexism is that sexist comments are often emotional and express a negative polarity and emotion towards the recipient. In addition, such comments are usually expressed using literary figures such as irony or mockery, to mask the sexist content.

The architecture of the system we develop can be seen in Figure 1. The system incorporate information associated with different concepts such as emotions, sentiments, and irony. These concepts are associated with distinct tasks to extract this information, such as emotion classification which recognizes the emotions expressed within a text, sentiment classification which analyzes the polarity of the text as positive or negative, and irony detection which is based on the identification of whether a text is ironic or not. In the system we propose, to extract this information from the posts in the dataset, we use models pre-trained on the specific task related to the information that we are interested in. These models give us the scores of the texts in the different categories in which the models classify. For example, the emotion classifier gives us the probability that the text is related to the emotion of anger, disgust, and joy, among others. Then, the first step to developing our system is to obtain the different scores of the text in the different categories of emotions, polarity, and irony for each post in the dataset. Secondly, we provide the posts to the tokenizer and the model. The model returns the [CLS] token of the last hidden state and we concatenate this token with the emotion, polarity, and irony scores. The concatenation between the [CLS] token and the scores are then passed to the head classifier (feed-forward network), which will give us a prediction about the classification of the data.

3 Experimental Setup

3.1 Data

To run our experiments, we use the dataset provided by the organizers. The dataset is composed of posts from Gab and Reddit that can contain hateful and sexist language. A set of 20,000 posts are annotated as sexist or not sexist and the sexist posts are designated with more specific labels related to the type of sexism in order to provide explanations about how a post is considered sexist. Table 1 shows the dataset size. For subtask A we have two labels related to whether a post is sexist or not (1. Sexist, 2. Not Sexist). In subtask B a sexist post is classified into four categories (1. threats, plans to harm, and incitement, 2. derogation, 3. animosity, 4. prejudiced discussions). To address subtask C, each sexist post is annotated with eleven classes related to a fine-grained type of sexism (1. threats of harm, 2. incitement and encouragement of harm, 3. descriptive attacks, 4. aggressive and emotive attacks, 5. dehumanizing attacks and overt sexual objectification, 6. casual use of gendered slurs, profanities, and insults, 7. immutable gender differences and gender stereotypes, 8. backhanded gendered compliments, 9. condescending explanations or unwelcome advice, 10. supporting mistreatment of individual women, 11. supporting systemic discrimination against women as a group).

Dataset	#Instances
Train	1,4000
Development	2,000
Test	4,000

Table 1: EDOS dataset splits. Training, development, and test sizes.

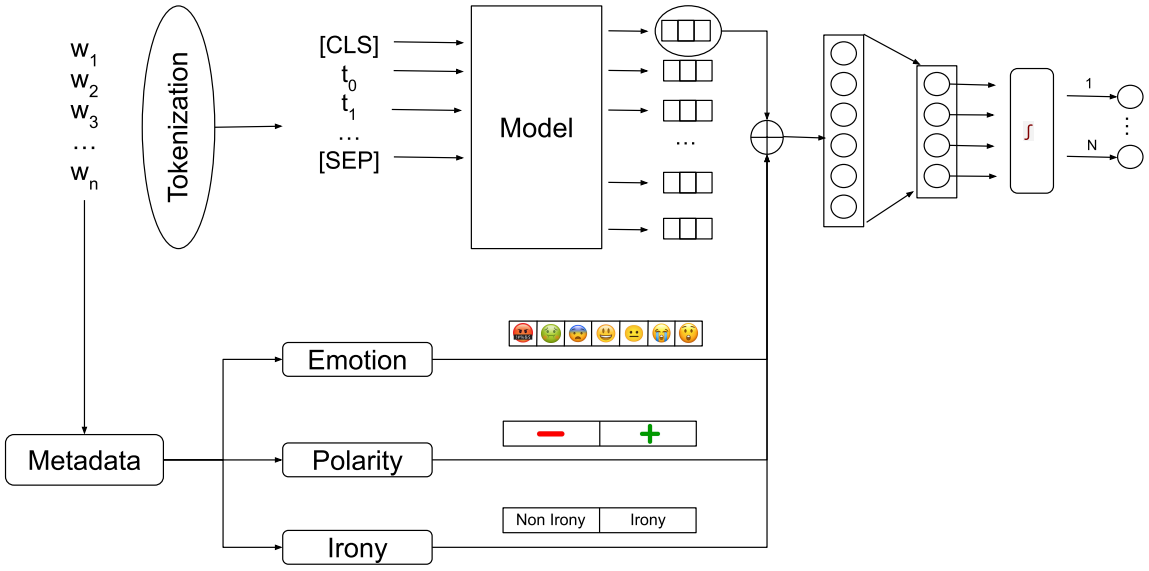

Figure 1: Proposed system for the EDOS task. The system incorporates information related to emotions, polarity, and irony in the texts. N represents the number of output nodes and depends on the task because it corresponds to the number of labels to classify.

3.2 Selected Models

In order to extract the different features related to the phenomena of emotions, sentiments, and irony, we rely on pre-training language models that have been fine-tuned on the task associated. Regarding emotion classification, we use an emotion model based on a DistilRoBERTa model[2] (Hartmann, 2022), to perform the polarity classification, we use a polarity model based on a BERT model [3] and to detect the irony, we use a model based on a RoBERTa architecture[4] (Barbieri et al., 2020). All of the last models are selected due to the fact that they are the most downloaded models from Hugging Face[5]. On the other hand, we select two base models to experiment with the proposed architecture in order to validate the incorporation of knowledge from sentiments, emotions, and irony. These baselines are RoBERTa base[6] (Liu et al., 2019) and Twitter RoBERTa base hate[7] (Barbieri et al., 2020) which are based on a RoBERTa architecture, with the difference that the second model has been pre-trained with tweets and fine-tuned for the hate speech detection task. Twitter RoBERTa base hate is the RoBERTa base model fine-tuned for hate-speech detection using a dataset composed of Twitter comments. These comments were labeled as "hateful" or "not hateful". The fine-tuned data of this model do not include the sexism target but we consider that the knowledge obtained by this model in hate speech detection will help to perform the predictions in the EDOS task. More information about the pre-trained data and the model is shown on the model card of the Hugging Face repository[7].

3.3 Training Regimen and Hyper-parameters

During the development phase, we train the models on the training set and evaluate them on the validation set. In the test phase, we train the developed models on the train and validation sets. Later we evaluate the models using the test set.

Hyper-parameters: For hyper-parameter optimization, in the development phase we split the dataset into training (80%) and validation data (20%). Specifically, we optimize the learning rate, the weight decay, and the batch size. For the hyper-parameters optimization, we use Optuna (Akiba et al., 2019), an automatic hyper-parameter optimization software framework, with a grid search.

[2] https://huggingface.co/j-hartmann/emotion-english-distilroberta-base
[3] https://huggingface.co/fabriceyhc/bert-base-uncased-amazon_polarity
[4] https://huggingface.co/cardiffnlp/twitter-roberta-base-irony
[5] https://huggingface.co/
[6] https://huggingface.co/roberta-base
[7] https://huggingface.co/cardiffnlp/twitter-roberta-base-hate

The search space for each hyper-parameter appears in Table 2. and Table 3 shows the best hyper-parameters obtained for each model in the different subtasks.

Hyper-parameter	Options
Learning-rate	[2e-5, 3e-5, 4e-5, 5e-5]
Weight-decay	[0, 1e-2, 1e-3]
Batch-size	[8, 16, 32]

Table 2: Search space used to optimize hyper-parameters for the selected models in subtasks A, B, and C of the EDOS shared task.

To run all experiments, the models are implemented using PyTorch (Paszke et al., 2019), a machine learning framework based on the Torch library, used for applications such as computer vision and natural language processing. The experiments were run on a single Tesla A100 40GB GPU with 256 GB of RAM.

4 Results

In this section, we present the results obtained by the systems developed as part of our participation in SemEval 2023 Task 10. To evaluate our systems, we use the official competition metrics given by the organizers. Specifically, the macro F1-score for all of the subtasks. Furthermore, we add two additional metrics, the macro precision, and the macro recall scores. The experiments are conducted in two phases, the model selection phase and the evaluation phase, which are explained in the following subsections.

4.1 Model Selection

In order to select the best model for each subtask we trained the models selected in Section 3.2 with the train split provided by the organizers and evaluate them with the development subset. The results obtained in the development phase are shown in Tables 4, 5, and 6.

In Table 4, we can see the results obtained by the developed systems for subtask A. To compare the models we use the metric provided by the organizers (macro-F1). RoBERTa base model achieves the best result in macro-F1 score with the baseline strategy (0.8370). Adding extra-linguistic features to this model does not improve the baseline. Regarding the model that contains knowledge about hate speech (Twitter RoBERTa base hate), it does not outperform the result obtained by the RoBERTa base baseline (0.8370 to 0.8227). However, the strategies that incorporate knowledge from emotions and emotions + sentimients outperform the baseline in Twitter RoBERTa base hate. Therefore, as the difference in results between these two strategies is minimal, we selected the model that incorporates more external knowledge (emotions + sentiments) for the evaluation phase of subtask A. The reason for this selection is that we hypothesize that with more information the system will outperform the rest of the models in the evaluation phase.

In Table 5, results obtained in subtask B are presented. To compare the models we use the metric provided by the organizers (macro-F1). On the one hand, the RoBERTa base model that incorporates features related to emotions improves the baseline (0.6207 to 0.6542). All the strategies of this model except the one that includes emotions + sentiments yield better results than the baseline. On the other hand, Twitter RoBERTa base hate model on the baseline strategy outperforms the baseline of RoBERTa base model (0.6207 to 0.6437). However, adding extra information in Twitter RoBERTa base hate model does not help to detect the content of this subtask. Therefore, we selected the baseline of Twitter RoBERTa base hate as the candidate for the test phase due to the fact that we consider that the information about hate in this model will be useful for future predictions in sexism detection.

Table 6 shows results obtained in subtask C. To compare the models we use the metric provided by the organizers (macro-F1). Regarding the RoBERTa base model, as in the previous task, the baseline strategies incorporating external knowledge, except emotions + sentiments, improve the baseline scenario. Specifically, the model that uses irony information performs the best (from 0.4181 to 0.4470). In Twitter RoBERTa base hate, the baseline and the strategy that includes emotions, sentiments, and irony achieve the same result (0.4742). It can be noticed that the strategies used for Twitter RoBERTa base hate outperform the RoBERTa models that do not incorporate extra information by a large margin. Therefore, using Twitter RoBERTa base hate which is pre-trained on Twitter texts and fine-tuned on hate speech help to determine the fine-grained sexism. In this subtask, the best results are obtained by the irony strategy of Twitter RoBERTa base hate model, however, we select the

Hyper-parameter	Task A		Task B		Task C	
	RB	TRBH	RB	TRBH	RB	TRBH
Learning rate	4e-5	5e-5	3e-5	5e-5	4e-5	5e-5
Weight decay	1e-3	0	0	0	1e-3	1e-2
Batch size	32	8	32	32	8	32

Table 3: Result of the hyper-parameter optimization for each model in the different subtasks. RB: roberta-base, TRBH: twitter-roberta-base-hate.

	RB			TRBH		
Approach	P	R	F1	P	R	F1
baseline	0.8498	0.8263	0.8370	0.8372	0.8109	0.8227
emotions	0.8391	0.8265	0.8325	0.8454	0.8104	0.8256
sentiments	0.8362	0.8172	0.8260	0.8344	0.8058	0.8185
irony	0.8385	0.8147	0.8255	0.8413	0.8057	0.8211
emotions + sentiments	0.8323	0.8190	0.8253	**0.8372**	**0.8151**	**0.8252**
emotions+sentiments+irony	0.8361	0.8130	0.8235	0.8281	0.8010	0.8131

Table 4: RoBERTa Base (RB) and Twitter RoBERTa Base Hate (TRBH) results in subtask A for binary sexism Detection on EDOS 2023 development set. The selected model for the evaluation phase is shown in bold. P: Macro-averaged precision, R: Macro-averaged recall, F1: Macro-averaged F1- score.

	RB			TRBH		
Approach	P	R	F1	P	R	F1
baseline	0.6382	0.6113	0.6207	**0.6541**	**0.6377**	**0.6437**
emotions	0.6608	0.6506	0.6542	0.6488	0.6116	0.6251
sentiments	0.6400	0.6470	0.6412	0.6427	0.6222	0.6280
irony	0.6409	0.6447	0.6409	0.6422	0.6116	0.6211
emotions + sentiments	0.6167	0.6096	0.6109	0.6420	0.6135	0.6242
emotions + sentiments + irony	0.6364	0.6400	0.6366	0.6402	0.6179	0.6246

Table 5: RoBERTa Base (RB) and Twitter RoBERTa Base Hate (TRBH) results in subtask B for the category of sexism on EDOS 2023 development set. The selected model for the evaluation phase is shown in bold. P: Macro-averaged precision, R: Macro-averaged recall, F1: Macro-averaged F1- score.

	RB			TRBH		
Approach	P	R	F1	P	R	F1
baseline	0.4323	0.4116	0.4181	0.5173	0.4535	0.4742
emotions	0.4633	0.4128	0.4220	0.5198	0.4515	0.4679
sentiments	0.4425	0.4246	0.4305	0.4958	0.4386	0.4572
irony	0.4601	0.4478	0.4470	0.5071	0.4585	0.4747
emotions + sentiments	0.3998	0.3826	0.3866	0.5062	0.4469	0.4641
emotions + sentiments + irony	0.4415	0.4128	0.4229	**0.5173**	**0.4535**	**0.4742**

Table 6: RoBERTa Base (RB) and Twitter RoBERTa Base Hate (TRBH) results in subtask C for the fine-grained sexism detection on EDOS 2023 development set. The selected model for the evaluation phase is shown in bold. P: Macro-averaged precision, R: Macro-averaged recall, F1: Macro-averaged F1- score.

Approach	RB			TRBH		
	P	R	F1	P	R	F1
			Subtask A			
baseline	0.8228	0.8166	0.8197	0.8270	0.8214	0.8245
emotions + sentiments	0.8356	0.8270	0.8311	**0.8263**	**0.8228**	**0.8245**
			Subtask B			
baseline	0.6372	0.6411	0.6370	**0.6210**	**0.6003**	**0.6043**
			Subtask C			
baseline	0.4557	0.4390	0.4449	0.4545	0.4212	0.4261
emotions + sentiments + irony	0.4686	0.4705	0.4686	**0.4624**	**0.4257**	**0.4376**

Table 7: RoBERTa Base (RB) and Twitter RoBERTa Base Hate (TRBH) results in subtasks A, B, and C on EDOS 2023 test set. SINAI Team Submissions are shown in bold. P: Macro-averaged precision, R: Macro-averaged recall, F1: Macro-averaged F1- score.

model that incorporates features from sentiments, emotions, and irony because we consider that our strategy of incorporating different types of information could improve the results in the evaluation phase.

4.2 Model Evaluation

In the evaluation phase, we train our systems on the training and development sets and evaluate them on the test set. In Table 7 we present the results obtained in the different subtasks. For subtask A, the baseline that has been fine-tuned with hate speech outperforms the model that does not include this information (0.8197 to 0.8245). However, the incorporation of emotions and sentiments in this model does not help to detect sexist content more accurately. Finally, in Table 8 we can see that the final model selected in the pre-evaluation phase (Twitter RoBERTa base hate with emotions and sentiments) ranks 46th on subtask A with a macro-F1 score of 0.8245.

For subtask B, a comparison between the baseline models of RoBERTa base and the model fine-tuned with hate speech is shown. The RoBERTa base model achieves the best results for this subtask, however, the Twitter RoBERTa base hate model that contains hate speech information does not help to detect the different categories of sexism as we suppose in the pre-evaluation phase. In Table 9 we can see that this model ranked 37th in the leaderboard of subtask B.

Finally, in subtask C, we present the results of our baseline models and those strategies that incorporate emotions, sentiments, and irony knowledge. It can be observed that RoBERTa base outperforms the model that incorporates hate speech. In particular, in this subtask, the models that include extra information achieve the best performance. Therefore, the incorporation of emotions, sentiments, and irony help to detect fine-graned categories of sexism. In this subtask, we ranked 29th among the participants as can be seen in Table 10.

Ranking	Team	F1
1	PingAnLifeInsurance	0.8746
2	stce	0.8740
2	FiRC-NLP	0.8740
...	...	...
46	**SINAI**	**0.8245**
...	...	...
84	NLP_CHRISTINE	0.5029

Table 8: Ranking of participants system in subtask A of EDOS Shared Task.

4.3 Error Analysis

In order to identify the challenges faced by the systems in the detection of sexism we conducted an error analysis on the test set. To perform this analysis, we focused on subtask C as it is the most difficult subtask. In order to analyze the error of our model, we randomly selected 119 posts. On the one hand, we analyzed 53 posts that are wrongly labeled by the baseline of Twitter RoBERTa base hate model but correctly labeled by the model that

Ranking	Team	F1
1	JUAGE	0.7326
2	PASSTeam	0.7212
3	stce	0.7203
...	...	...
37	**SINAI**	**0.6043**
...	...	...
69	NLP_CHRISTINE	0.2293

Table 9: Ranking of participants system in subtask B of EDOS Shared Task.

Ranking	Team	F1
1	PALI	0.5606
2	stce	0.5487
3	PASSTeam	0.5412
...	...	...
29	**SINAI**	**0.4376**
...	...	...
63	shm2023	0.0632

Table 10: Ranking of participants system in subtask C of EDOS Shared Task.

incorporates emotions, sentiments, and irony. With this analysis, we aim to know the challenges faced by the baseline. On the other hand, we study 66 posts mislabeled by the best strategy that includes information related to emotions, sentiments, and irony of Twitter RoBERTa base hate model, in order to observe the difficulties presented in this task by the best model. Then, we selected the 4 most representative posts of the most common mistakes of each model.

Table 11 shows the posts mislabeled by the baseline strategy but correctly labeled by the best strategy that includes emotions, sentiments, and irony features. In examples 1 to 4 it can be observed that the emotion and sentiment predicted by the model correspond to negative emotions (disgust, anger) and negative polarity. We believe that these features are related to sexist expression and therefore this information might help the model to detect the different categories more accurately. Furthermore, in examples 1 to 3, the irony model identifies the post as irony which is also often used for the expression of sexist language to mask this content. Therefore, we conclude that some of the challenges faced by the baseline system could be addressed by

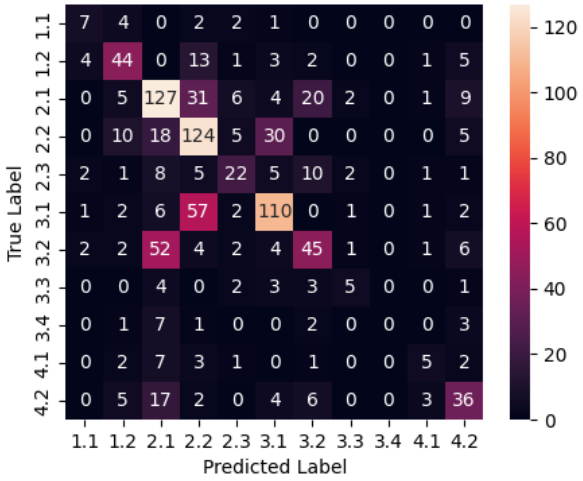

Figure 2: The confusion matrix of Twitter RoBERTa base hate (TRBH) model with emotions + sentiment + irony on the test set.

incorporating linguistic features that occur in the expression of sexism.

But we were also interested in identifying which challenges are faced by the best system, i.e. the one that incorporates feelings, emotions, and irony information. Therefore, we conducted an error analysis and identified the most common ones in Table 12. In examples with ID sexism2022_english-12298 and sexism2022_english-476 we can see that the emotion has been predicted with less probability, and in addition, the irony has not been detected. In the following example, with ID sexism2022_english-2110, we can see that the emotion model predicts a negative emotion (anger) but the sentiment predicted is positive polarity, therefore, the disagreement between the two models might cause difficulty in deciding the correct prediction. Finally, in the post with ID sexism2022_english-10146, the emotion prediction is neutral and the polarity is positive, two features that are not usually related to the expression of sexism.

In addition, we calculate the confusion matrix in order to perform deep analysis. Figure 2 shows the confusion matrix of Twitter RoBERTa base hate model with emotions + sentiment + irony. In this figure, we can observe that our model never uses the label 3.4 ("condescending explanations or unwelcome advice"). Perhaps this is due to the unbalanced data used during model training. The dataset used during this phase has fewer comments labeled with "condescending explanations or unwelcome advice". Furthermore, we can see that the

ID	Post	Gold label	TRBH label	TRBH emotion + sentiment + irony			
				label	emotion	polarity	irony
sexism2022_english-10350	To female Bishops stop making fun of religion.	4.2	2.1	4.2	Disgust (60%)	N	I
sexism2022_english-1938	The whole sex shaming and cover up culture is a feminine invention - and it's a shit test. And so is [sex-positive feminism] ([URL] also a shit test.	4.2	2.1	4.2	Disgust (73%)	N	I
sexism2022_english-2862	Amy Schumer complaining about two black male comics being paid more than her from Netflix???? Check your white privilege bitch! [URL].	3.1	2.2	3.1	Anger (82%)	N	I
sexism2022_english-17935	Maybe if one of their women were molested... But no, they're all too ugly.	2.1	2.2	2.1	Anger (71%)	N	NI

Table 11: Twitter RoBERTa Base Hate (TRBH) error analysis between the baseline and emotions + sentiment + irony model. Examples of predictions from the EDOS 2023 test suite are shown, depicting improvement in the emotions+sentiments+irony strategy.

id	Post	Gold label	TRBH Emotion+Sentiment+Irony			
			label	Emotion	Polarity	Irony
sexism2022_english-12298	If only my ex was a nice woman like her and just put a bullet in my head instead of torturing me on a daily basis... Most women prefer the death by 1000000 cuts method.	2.1	1.2	Anger (53%)	N	NI
sexism2022_english-476	The crooked, drunken old hag deserves at least 3 life sentences!!	2.2	1.2	Anger (52%)	N	NI
sexism2022_english-2110	I picked 50% even though it's usually individualized, to balance out. I think sometimes the women is a naggy bitch, but we also all know that sometimes the guy is just a violent a-hole. But then again, she picked him.	2.2	3.1	Anger (74%)	P	NI
sexism2022_english-10146	Feminists want the advantages of being male and the advantages of being female and with none of the disadvantages of either.	3.2	2.1	Neutral (90%)	P	I

Table 12: Twitter RoBERTa base hate (TRBH) model with emotions + sentiment + irony error analysis. Examples of predictions from the EDOS 2023 test set are shown.

most predominant classes are the classes where the model performs the correct predictions.

5 Conclusion

This paper presents the participation of SINAI research group in the Explainable Detection of Online Sexism shared task at SemEval 2023. In all of the subtasks, we explore how different fine-tuned strategies of the RoBERTa model perform by adding extra-linguistic features (emotions, sentiment, irony). For the first subtask, we have seen that the incorporation of emotions and sentiments tends to help detect sexist content. In the second subtask, on the contrary, we observed that the baseline model RoBERTa helps to detect the different categories of sexism, however, the other strategies have not outperformed this model, as we expected in the pre-evaluation phase. Finally, in subtask C, we noticed that the incorporation of the different features (emotions, sentiments, irony) has contributed significantly to the detection of the different fine-grained categories of sexism. We conclude that, in general, the incorporation of extra-linguistic information helps the models to conduct the sexist tasks, especially when the number of classes increases, as we have observed in subtask C. In future work, we plan to explore other ways to add extra-linguistic information into the pre-trained language systems. In addition, we would like to analyze the impact of the model architecture while incorporating features.

Acknowledgements

This work has been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project PRECOM (SUBV-00016) funded by Ministerio de Consumo and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA@CLiC-it*.

Elisabetta Fersini, Paolo Rosso, and Mary E. Anzovino. 2018b. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *IberEval@SEPLN*.

Simona Frenda, Ghanem Bilal, et al. 2018. Exploration of Misogyny in Spanish and English tweets. In *Third workshop on evaluation of human language technologies for iberian languages (ibereval 2018)*, volume 2150, pages 260–267. Ceur Workshop Proceedings.

Sercan Halat, Flor Miriam Plaza-Del-Arco, Sebastian Padó, and Roman Klinger. 2022. Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language.

Jochen Hartmann. 2022. Emotion english distilroberta-base. `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/`.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. Semeval-2023 task 10: Explainable detection of online sexism.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Flor Miriam Plaza-del-Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and María-Teresa Martín-Valdivia. 2022. Integrating implicit and explicit linguistic phenomena via multi-task learning for offensive language detection. *Knowledge-Based Systems*, 258:109965.

Flor Miriam Plaza-del-Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and María Teresa Martín-Valdivia. 2021. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489.

Juan Manuel Pérez, Franco Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, and Viviana Cotik. 2022. Assessing the impact of contextual information in hate speech detection.

Francisco Rodríguez-Sánchez, Jorge Carrillo de Albornoz, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco-Remón, Maryna Makeienko, María Plaza, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2022. Overview of exist 2022: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 69(0):229–240.