Enriching Hate-Tuned Transformer-Based Embeddings with Emotions for the Categorization of Sexism

Arianna Muti[1,*,†], Eleonora Mancini[2,†]

[1]*Department of Interpreting and Translation (DIT) - University of Bologna*

[2]*Department of Computer Science and Engineering (DISI) - University of Bologna*

Abstract

We present the results of the participation of our team **Unibo** in the shared task sEXism Identification in Social neTworks (EXIST). We target all three tasks: a) binary sexism identification, b) discerning the author's intention, and c) categorizing instances into fine-grained categories. For all the tasks, both English and Spanish data are to be considered. We compare two approaches to address this multilingual aspect: we employ machine translation to convert the Spanish data into English, allowing us to utilize a specially fine-tuned version of RoBERTa to detect hateful content, and we experiment with a multilingual version of RoBERTa to perform classification while preserving data in their original language. Furthermore, we predict emotions associated with each post and leverage them as additional features by concatenating them with the original text. This augmentation improves the performance of our models in Task 2 and 3. Our official submissions obtain F1=0.77 in Task 1 (13th position out of 69), macro-averaged F1=0.53 in Task 2 (4th position out of 35) and macro-averaged F1=0.59 in Task 3 (4th position out of 32).

Keywords

sexism, hate speech, hate-tuned transformers, emotion-based features

1. Introduction

The EXIST shared task aims at detecting sexism, ranging from explicit misogyny to other subtle forms of implicit sexist behaviours [1, 2]. This task distinguishes itself from other relevant tasks on sexism detection by encompassing not only posts explicitly recognized as sexist but also posts that document reported acts of sexism.

The EXIST shared task focuses on English and Spanish tweets and proposes three sub-tasks, where Tasks 2 and 3 are hierarchical with respect to Task 1.

Tasks are defined as follows:

- **Task 1 - Binary Sexism Detection**: systems are required to predict whether a post contains sexist expressions or behaviour.

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

[*] Corresponding author.

[†] These authors contributed equally.

arianna.muti2@unibo.it (A. Muti); e.mancini@unibo.it (E. Mancini)
https://www.unibo.it/sitoweb/arianna.muti2 (A. Muti); https://www.unibo.it/sitoweb/e.mancini/en (E. Mancini)
0000-0002-3387-6557 (A. Muti); 0000-0001-9205-3289 (E. Mancini)

CEUR Workshop Proceedings (CEUR-WS.org)

- **Task 2 - Source Intention**: if a post contains sexism, systems have to predict one of three mutually exclusive categories expressing the intention of the author. Intentions are classified as follows: *direct* (the intention was to write a message that is sexist by itself); *reported* (the intention is to report and share a sexist situation suffered by a woman or women in first or third person); *judgemental* (the intention is to condemn sexist behaviours).
- **Task 3 - Categorization of Sexism**: if a post contains sexism, systems have to predict one among 5 non-mutually exclusive subcategories, i.e., *ideological and inequality*; *stereotyping and dominance*; *objectification*; *sexual violence*; *misogyny and non-sexual violence.*

In this paper, we present our approach to address all three subtasks. To address the multilingual aspect of this task, we compare a multilingual model (*XLMR*[3]), where input data are mixed between English and Spanish, with a monolingual model (*Twitter-RoBERTa-base-hate*[1]), whose input data are translated from Spanish into English with Google Translate's API.

Furthermore, we explore the use of emotions as additional features. We predict the emotion for each tweet and we concatenate them to the original texts, forming an augmented representation that encapsulate both linguistic and emotional context. Our hypothesis is that the emotional dimension of sexist content may provide useful cues for its detection. The analysis of emotions might be beneficial for the detection of hate speech, as expressed hate should point to the author experiencing anger while the addressees are likely to experience fear [4]. We want to observe whether this applies also to the detection of sexism, especially in Task 2 and 3, where different emotions should be detected with respect to the intention of the author and the type of sexism. For instance, we expect reported sexism to co-occur with anger, sexual objectification with desire and misogyny with disgust, although people's emotional states can vary greatly based on their attitudes, and it is challenging to generalise the emotions of all individuals expressing and facing sexism. On the other hand, associating non-sexist tweets with either positive or neutral emotions can help the classification of such tweets, which might result in an overall increasing of the performance.

Our official submissions obtain macro-averaged F1=0.77 in Task 1 (13th position out of 69), F1=0.53 in Task 2 (4th position out of 35) and F1=0.59 in Task 3 (4th position out of 32). Our results show that enriching the embeddings with emotions can help the model to better distinguish the different nuances of sexism in Task 2 and 3, while it does not affect the performance of Task 1.

The paper is organised as follows: Section 2 provides a summary of related work. Section 3 describes the dataset for each task. Section 4 describes our methodology for the classification and the prediction of emotions. Section 5 and 6 present the models employed and the experimental setup, along with the results. Section 7 concludes with a summary of our findings.

2. Related Work

While many relevant shared tasks have been focusing on the detection of misogyny [5, 6, 7, 8, 9], some have tackled the detection of sexism as well; i.e. the past two editions of sEXism Identification in Social neTworks [10, 11] and the recent SemEval shared task on the Explainable

[1] https://huggingface.co/cardiffnlp/twitter-roberta-base-hate

Detection of Online Sexism (EDOS) [12]. For what concerns the past editions of EXIST, participants had to classify posts into the following categories: ideological and inequality; stereotype and dominance; objectification; sexual violence; and misogyny and non-sexual violence. The languages targeted were English and Spanish in the Gab and Twitter domains. In both editions, the majority of participants exploited transformer-based systems for both tasks. Some managed to improve the performance with data augmentation techniques, via back translation techniques [13] or task-related existing datasets [14]. Plaza-Del-Arco et al. [4] used the emotion detection task as the auxiliary one in a multi-task learning setup, by training a shared model with the Universal Joy dataset, achieving an improvement in the performance compared to their baseline.

Other researchers have explored the use of emotions as features for the identification of sexism. For instance, Markov et al. [15] investigates the use of emotion-based features in the context of multilingual hate speech detection. To encode emotions, they used the 14,182 emotion words and their associations with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive) from the NRC emotion lexicon [16]. Despite the existence of prior works that touch upon the role of emotions in detecting sexism, the exploration of emotions as a means for detecting sexism remains relatively limited. Our approach stands out from existing ones as we incorporate emotions as additional features by representing them in the embedding space along with the text.

3. Dataset

The EXIST dataset encompasses various forms of sexist expressions and related phenomena, including descriptive or reported assertions where the message involves sexist behaviors. The EXIST 2023 dataset boasts an extensive collection of over 10,000 labeled tweets, including both English and Spanish languages. The training, development and test sets comprise 6,920 tweets, 1,038 tweets, and 2,076 tweets respectively. A quite balanced distribution between the two languages is maintained, a little bit skewed towards the Spanish language, while the distribution among classes in Tasks 2 and 3 is quite unbalanced. For the class statistics, please refer to the overview paper of the shared task [17].

4. Methodology

We devise two different scenarios - with and without emotions - to assess the performance of the two models - monolingual and multilingual - for sexism identification across multiple tasks.

We treat Task 1, 2 and 3 as independent: the models for Task 2 and 3 are not influenced by the decision of the model for Task 1. We motivate this choice with a prior study that demonstrates that when it comes to hierarchical tasks, it is better to feed the model with all training data, even if half of the instances are judged as not sexist, since this helps the classification of non-sexist posts, and therefore increase the overall performance [18].

To assess the impact of considering emotions as an additional feature, we firstly employ the original tweet as input for the models. Then, we predict the emotion for each post and we concatenate the predicted emotion to the tweet as an additional string. Only one emotion is predicted for each post. As a result of this process, the embedding space will include the emotional information as part of the input sequence. This can potentially influence the positioning of the posts in the embedding space, as the emotions carry additional semantic content that can affect the overall representation of the text. The models have the opportunity to learn and distinguish patterns that align with both the text and the associated emotions, potentially improving their ability to classify sexist posts by considering the emotional cues alongside the textual information.

The strategy adopted to add emotions as a feature involved several steps which are summarized in Figure 1. The following is a detailed description of the proposed approach:

1. **Emotion Models**: We employ two pre-trained models to infer emotions in tweets, namely EmoRoBERTa and Emotion English DistilRoBERTa-base. We will refer to this models as EmoDistilRoBERTa. Please, refer to Section 5 for a detailed description of these models.
2. **Translation Preprocessing**: Since the pre-trained models are trained on English data, a translation preprocessing phase was performed for Spanish tweets. The Google Translator, provided by the *deep_translator*[2] Python library, was utilized for this purpose. The tweets were translated from Spanish to English to ensure compatibility with the pre-trained emotion models.
3. **Pre-processing Steps**: Before inferring emotions on the translated tweets, several pre-processing steps were undertaken to align the data with the training data. This involved removing retweets, URLs, hashtags, and mentions from the tweets. By performing these pre-processing steps, the input data was prepared to match the format and content of the training data used for the emotion models.
4. **Emotion Prediction**: The pre-processed and translated tweets were then fed into the two emotion models to predict the corresponding emotions for each tweet. The models classified the tweets into different emotion labels based on their training.
5. **Augmenting the Dataset**: For each tweet, additional information was added to the dataset. This included:
    - *Translated Tweet*: The tweet translated from Spanish to English.
    - *Translated Tweet Original*: The tweet translated from Spanish to English in its original form (i.e. without performing pre-processing on it).
    - *Emotion*: The predicted emotion label for the tweet.
    - *Emotion ES*: The translation of the predicted emotion label into Spanish.
    - *Emotion Tweet Original*: The original tweet combined with the emotion label. In the case of Spanish tweets, this corresponds to the translated version of the emotion.
    - *Emotion Tweet Translated*: In the case of Spanish data, this corresponds to the translated version of the tweets, while in English tweets remain in their original language.

Two different versions of the dataset were created, one for each model used for predicting emotions. Each version contains the augmented data with the respective emotion predictions. Table 4 in Appendix A shows examples of the tweets before and after concatenating the emotions predicted by both models.

[2] https://github.com/nidhaloff/deep-translator/tree/master

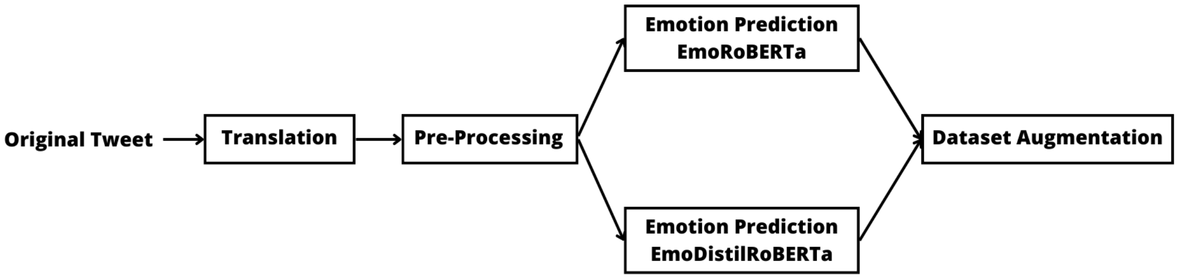
Original Tweet
Translation
Pre-Processing
Emotion Prediction EmoRoBERTa
Emotion Prediction EmoDistilRoBERTa
Dataset Augmentation

Figure 1: Steps involved in the dataset augmentation process.

5. Models Description

5.1. Emotion Classification Models

In our study, we employ two distinct models for emotion prediction, namely *EmoRoBERTa* and *Emotion English DistilRoBERTa-base*, both designed specifically for emotion detection tasks.

- **EmoRoBERTa**[3] is a pre-trained model that has undergone training on the GoEmotions[19] dataset. It possesses the capability to predict a wide range of 28 different emotions[4].
- **Emotion English DistilRoBERTa-base**[5] is a distilled version of the RoBERTa model, which has been trained on a diverse set of six datasets. These datasets comprise emotion-labeled texts sourced from various platforms such as Twitter, Reddit, student self-reports, and utterances from TV dialogues. This model focuses on predicting the six emotions defined in Ekman's emotion model[20] plus the neutral class[6].

5.2. Sexism Classification Models

In our study, we employ two models for the classification of sexism: a monolingual model originally designed for hate speech detection and fine-tuned specifically for the task of sexism detection and classification, namely *Twitter-RoBERTa-base-hate* and *XLM-RoBERTa*, a general-purpose multilingual model to work with Spanish data, without the need of translating them into English.

- **Twitter-RoBERTa-base-hate** is a RoBERTa-base model trained on 58M tweets and fine-tuned for hate speech detection with the TweetEval benchmark [9]. This model predicts whether a tweet is hateful or not against immigrants and women. Since RoBERTa does not

[3] https://huggingface.co/arpanghoshal/EmoRoBERTa

[4] admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise + neutral

[5] https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

[6] anger, disgust, fear, joy, neutral, sadness, surprise

employ the Next Sentence Prediction loss, this model is more suitable for Twitter where most tweets are composed of a single sentence [21]. We chose Twitter-RoBERTa-base-hate over other hate-tuned Transformer-based models after initial experiments showed the former to outperform the others on Task 1.

- **XLM-RoBERTa**[3] (XLM-R) is a pre-trained language model architecture that extends upon the original RoBERTa model by incorporating multilingual capabilities. XLM-R takes text inputs in the form of word sequences. The input text is tokenized into subword units using the SentencePiece tokenizer, which splits words into smaller subword units to handle different languages' morphological variations. XLM-R employs translation language modeling during pre-training. It learns to translate sentences from one language to another by conditioning on the source language embeddings, allowing the model to develop cross-lingual understanding and to transfer knowledge across different languages. We chose XLM-R over mBERT [22] after initial experiments showed the former to outperform the latter on Task 1.

6. Experiments and Results

In this section we present the experiments performed for each task, along with the results on the development and test set.

6.1. Monolingual vs Multilingual Setting

Firstly, we want to assess whether a hate-tuned Transformer-based model, whose Spanish input data are translated into English, performs better than a general-purpose multilingual model which preserves the input data in its original language. To carry out this experiment, we compare the two baselines: XLM-R (multilingual setting) and RobertaHate (monolingual setting). We perform such experiment in Task 1.

Both RobertaHate and XLM-R are fine-tuned on our downstream task. We perform a minimum parameter selection tuning on the validation set (10% of the training set). We selected the highest performing learning rate $\in [1\text{e}{-}5, 2\text{e}{-}5, 1\text{e}{-}2]$; batch size $\in [4, 8, 16, 32]$; epochs in range $[1 - 10]$. The best configuration for both models is: lr = $2\text{e}{-}5$, batch size = 16, epochs = 4.

As shown in Table 1, translating data into English to employ a hate-tuned model performs slightly better than using a general-purpose multilingual model. Therefore, we do not consider XLM-R for the subsequent tasks.

6.2. Emotions as Additional Feature

We want to assess whether considering the emotions extracted from the tweets as features can help the classification of sexism in all three tasks. First, we need to assess whether it is better to rely on a smaller (EmoDistilRoBERTa) or wider (EmoRoBERTa) range of emotions. As shown in Table 1, in either ways, the performance does not shift in Task 1. However, in Task 2 the inclusion of emotions let us gain 0.04 points and 0.02 points with EmoRoBERTa and EmoDistilRoBERTa respectively, compared to our baseline (RobertaHate); in Task 3 we gain 0.02 points with EmoRoBERTa, while the score remains unaltered with EmoDistilRoBERTa.

Table 1
Experiments and macro-averaged F1 scores on the development set.

Setting	Task 1	Task 2	Task 3
XLMR	0.84	-	-
XLMR + EmoRoBERTa	0.84	-	-
XLMR + EmoDistilRoBERTa	0.85	-	-
RobertaHate	0.85	0.57	0.60
RobertaHate + EmoRoBERTa	0.85	0.61	0.62
RobertaHate + EmoDistilRoBERTa	0.85	0.59	0.60

Table 2
Experiments and F1 scores on the test set. For Task 1 the F1 score is computed on the positive class, while for Task 2 and 3 the macro-average is considered.

Setting	Task 1	Task 2	Task 3
RobertaHate	0.771	-	-
RobertaHate + EmoRoBERTa	0.771	0.528	0.590
top_1	0.810	0.571	0.629
top_2	0.805	0.548	0.629
top_3	0.802	0.548	0.584

To predict on the test set, we only considered the models that performed the best in the development set. Since for Task 1 we reached the same scores with and without adding emotions as features, we decided to test both models. For Task 2 and 3 we only tested our hate-tuned Transformer-based model with the inclusion of emotions. Since the organisers did not release the gold labels, we cannot perform additional experiments on the test set at the moment. Table 2 shows our official submission plus the comparison with the top three teams.

Additionally, we performed an analysis of the emotions associated to each tweet. In particular, for each tweet, we observed the correlation between certain categories of sexism and the emotions predicted. For this assessment, we consider EmoRoBERTa, which proposes a more fine-grained categorization of emotions and obtained better performance compared to EmoDistilRoBERTa in the development set (cf. Table 1). In Task 2, we observe a correlation between `non-sexist` tweets and *neutral* and *approval* emotions, while most of `reported sexism` is associated with *disgust*. `Direct sexism` is associated with *disgust*, *anger* and *amusement*. The latter does not occur in the other categories. When it comes to the type of sexism expressed (Task 3), as expected, the class `objectification` is also associated with positive emotions such as *admiration* and *amusement*, which are much less present in other classes, along with *disgust*, *neutral* and *anger*. The reason behind is that often sexual objectification can be overlooked as benevolent sexism [23], as in the following example: *Nice print, your honor, attractive even without a neckline, very rare in these times.* The other classes tend to occur mostly with *disgust*, *neutral*, *anger* and *fear*. `Ideological-Inequality` sees the presence of the emotion *surprise*, which ironically signals the presence of a cliché, like in the following example: *@BestKabest This gringa keeps crying over gamergate, what a "coincidence" that she* *has pronouns on her profile*. Finally, we observed a correlation between the emotion *neutral* and most implicit forms of sexism.

6.3. Error Analysis

In order to gain deeper insights into the influence of incorporating emotions in textual embeddings, we conduct an error analysis by comparing the predictions generated by two distinct models: the baseline model and the emotion-enhanced model. By contrasting the outputs of these models, our objective is to assess the extent to which including emotions affects prediction outcomes. We perform such analysis on the development set of Task 3, which we find the most suitable for its granularity of the classes, as they explain the kind of sexism expressed.

Firstly, we retrieve all instances misclassified by the baseline that are correctly identified by the enhanced-model. Table 3 shows the number of false positive and false negative cases for each class. The classes most influenced by the enhanced model are *Ideological and Inequality* and *Misogyny and Non-sexual Violence*, which are the classes that contain the most instances of implicit sexist language. For instance, in the sentence "It is impossible for a man to become a woman or a for a woman to become a man. For this reason, it is irresponsible and downright wrong (actually evil), to teach children that some cosmetic changes to their appearance can change their biological sex", there is no explicit hate against transexual people, although the implication that the author does not support the LGBT+ community is well captured in the predicted emotion 'disgust', which might have helped the enhanced-model to correctly identify it. Therefore, we speculate that including emotions in the textual embeddings might be a promising way to spot the most implicit cases of sexism and misogyny, which are the most difficult to detect.

When it comes to false positive instances, on the other hand, we noticed the presence of slurs or aggressive words in some tweets, which might have been misleading for the baseline. However, associating a positive emotion to such tweets might have helped the emotion-enhanced language model to understand that such slur is not used in a sexist way. However, to confirm that, an analysis on the embedding space would be necessary.

Another scenario in which emotions helped the classification is when instances contain irony or sarcasm. The following sentences "Look, look, how funny, women can't drive, hahaha, how funny."; "Some man moving my suitcase in the overhead luggage storage on a train to what he thinks is a better position (why?!) and now completely out of my reach for when I have to rush off the train in a couple of stops. Women can't arrange their own luggage, apparently." have been predicted as containing 'amusement' and 'surprise', which can signal the presence of irony and sarcasm, in this case towards discriminatory or belittling statements against women.

7. Conclusion

In this paper we have presented the results of our participation in the EXIST shared task for the detection and categorization of sexism. Our approach showcased the efficacy of translating data into English and employing a hate-tuned Transformer-based model compared to a general-purpose multilingual model, when dealing with multilingual data. Furthermore, the inclusion of emotion-based features proved to be a valuable enhancement, boosting the performance in Task

Table 3
The number of false positive and false negative instances misclassified by the baseline, that the model enhanced with emotions classified correctly.

Class	FP	FN
Objectification	5	10
Stereotyping and Dominance	15	10
Ideological and Inequality	37	8
Misogyny and Non-sexual Violence	18	7
Sexual Violence	2	9

2 and Task 3, where emotional context can play a major role when associated with fine-grained categories of sexism. However, considering emotions does not affect, neither positively nor negatively, the binary classification of sexism in Task 1. Our error analysis showed that emotions can be a valuable feature to detect sexist implicit language or to reduce false positives in case of slurs or sarcasm, which often result in spurious correlations with the positive class. Our approach allowed us to pave the way for a study on the correlation between certain types of sexism and the emotion predicted. As a future research direction, further investigation should be conducted to explore the correlation between emotions and sexism. One way would be through the creation of gold labels with respect to emotions in relation to sexist tweets. Moreover, it would be helpful to analyse the embedding spaces of the two models, the one that includes emotions and the baseline, to visually observe how the inclusion of the emotions affects the understanding of sexist tweets by language models.

8. Limitations

The limitation of our study is that the specific impact of including emotions on the embedding space depend on the quality and relevance of the predicted emotions, and we do not have any control on that. Moreover, translating tweets might produce a bias in the prediction of emotions, as emotions are influenced by cultural and linguistic background, and the perception of emotions can vary across different languages. When translating tweets discussing sexism, the specific emotional context conveyed in the original language may not fully be carried over to the translated version. This could be due to differences in language structure, idiomatic expressions, or cultural connotations associated with certain emotions. As a result, the predicted emotions may not accurately capture the intended emotional nuances related to sexism.

References

[1] L. Plaza, J. Carrillo-de Albornoz, R. Morante, J. Gonzalo, E. Amigó, D. Spina, P. Rosso, Overview of exist 2023: sexism identification in social networks, in: Proceedings of ECIR'23, 2023, pp. 593–599. doi:10.1007/978-3-031-28241-6_68.

[2] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023 – learning with disagreement for sexism identification and characterization(extended overview)., in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum. Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro and Michalis Vlachos, Eds., Thessaloniki, Greece, 2023.

[3] 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, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://doi.org/10.18653/v1/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[4] F. M. Plaza-Del-Arco, S. Halat, S. Padó, R. Klinger, Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language, 2109.10255 (2022).

[5] E. Fersini, D. Nozza, P. Rosso, Overview of the Evalita 2018 task on automatic misogyny identification (AMI), in: EVALITA Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop 12-13 December 2018, Naples, Torino: Accademia University Press, 2018, pp. 59–66. doi:doi:10.4000/books.aaccademia.4497.

[6] E. Fersini, D. Nozza, P. Rosso, AMI@EVALITA2020: Automatic misogyny identification, in: Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), CEUR, 2020.

[7] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549. URL: https://aclanthology.org/2022.semeval-1.74. doi:10.18653/v1/2022.semeval-1.74.

[8] M. Anzovino, E. Fersini, P. Rosso, Automatic identification and classification of misogynistic language on twitter, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2018, pp. 57–64.

[9] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, 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, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://aclanthology.org/S19-2007. doi:10.18653/v1/S19-2007.

[10] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6389.

[11] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón,

M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6443.

[12] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 Task 10: Explainable Detection of Online Sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Toronto, Canada, 2023. URL: http://arxiv.org/abs/2303.04222. doi:10.48550/arXiv.2303.04222.

[13] S. Butt, N. Ashraf, G. Sidorov, A. Gelbukh, Sexism identification using bert and data augmentation - exist2021, CEUR Workshop Proceedings 2943 (2021) 381–389. Publisher Copyright: © 2021 CEUR-WS. All rights reserved.; null ; Conference date: 21-09-2021.

[14] D. García-Baena, M. Á. G. Cumbreras, S. M. J. Zafra, M. García-Vega, Sinai at exist 2022: Exploring data augmentation and machine translation for sexism identification, in: IberLEF@SEPLN, 2022.

[15] I. Markov, N. Ljubešić, D. Fišer, W. Daelemans, Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection, in: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Online, 2021, pp. 149–159. URL: https://aclanthology.org/2021.wassa-1.16.

[16] S. M. Mohammad, P. D. Turney, Crowdsourcing a word–emotion association lexicon, Computational Intelligence 29 (2013) 436–465. doi:https://doi.org/10.1111/j.1467-8640.2012.00460.x.

[17] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023 – learning with disagreement for sexism identification and characterization. experimental ir meets multilinguality, multimodality, and interaction., in: Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023). Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro and Michalis Vlachos, Eds., Thessaloniki, Greece, 2023.

[18] A. Muti, F. Fernicola, A. Barrón-Cedeño, Misogyny and aggressiveness tend to come together and together we address them, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 4142–4148. URL: https://aclanthology.org/2022.lrec-1.440.

[19] D. Demszky, D. Movshovitz-Attias, J. Ko, A. S. Cowen, G. Nemade, S. Ravi, Goemotions: A dataset of fine-grained emotions, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 4040–4054. URL: https://doi.org/10.18653/v1/2020.acl-main.372. doi:10.18653/v1/2020.acl-main.372.

[20] P. Ekman, Basic emotions, in: T. Dalgleish, M. J. Power (Eds.), Handbook of cognition and emotion, 1999, pp. 45–60.

[21] F. Barbieri, J. Camacho-Collados, L. E. Anke, L. Neves, Tweeteval: Unified benchmark and comparative evaluation for tweet classification, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020, volume EMNLP 2020 of *Findings of ACL*, Association for Computational Linguistics, 2020, pp. 1644–1650. URL: https://doi.org/10.18653/v1/2020.findings-emnlp.148.

doi:10.18653/v1/2020.findings-emnlp.148.

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

[23] P. Glick, S. Fiske, Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women, Psychology of Women Quarterly 21 (1997) 119–135. doi:10.1111/j.1471-6402.1997.tb00104.x.

A. Appendix

Table 4 shows examples of the tweets before and after concatenating the emotions predicted by both models.

Table 4

Examples of the original tweets, the predicted emotions using both EmoRoBERTa and EmoDistilRoBERTa together with the translated emotions into Spanish, and the original tweets concatenated with the corresponding predicted emotions. When we employ the hate-tuned model for the classification of sexism, instead of Emotion Tweet Original we have the column Emotion Tweet Translated, which contains the tweet translated into English and the corresponding predicted emotion.

Original Tweet	Emotion EN (EmoRoBERTa)	Emotion ES (EmoRoBERTa)	Emotion EN (EmoDistilRoBERTa)	Emotion ES (EmoDistilRoBERTa)	Emotion Tweet Original (EmoRoBERTa)	Emotion Tweet Original (EmoDistilRoBERTa)
@gerardotc En estos casos, como en tantos otros, el castigo siempre va en la misma direcciòn. #Androcentrismo #Misoginia	DISGUST	ASCO	DISGUST	ASCO	@gerardotc En estos casos, como en tantos otros, el castigo siempre va en la misma direcciòn. #Androcentrismo #Misoginia ASCO	@gerardotc En estos casos, como en tantos otros, el castigo siempre va en la misma direcciòn. #Androcentrismo #Misoginia ASCO
@mmpadellan How is Ginni Thomas still living her best life as a free woman?	CURIOSITY	CURIOSIDAD	NEUTRAL	NEUTRAL	@mmpadellan How is Ginni Thomas still living her best life as a free woman? CURIOSITY	@mmpadellan How is Ginni Thomas still living her best life as a free woman? NEUTRAL