SINAI at EXIST 2022: Exploring Data Augmentation and Machine Translation for Sexism Identification

Daniel García-Baena[1,*,†], Miguel Ángel García-Cumbreras[1,†], Salud María Jiménez-Zafra[1,†] and Manuel García-Vega[1,†]

[1]*Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Spain*

Abstract

This work presents the participation of the SINAI team in the EXIST 2022 shared task at IberLEF. Specifically, we have addressed *Task 1: Sexism identification* consisting of a binary classification of sexism. We have explored data augmentation and machine translation techniques for fine-tuning a Transformer model. In total, 45 systems have been submitted from all participating teams. The 3 runs sent by our team have been placed in positions 32, 33 and 35 with an accuracy of 73.16%, 72.78% and 72.02%, respectively, being 79.96% the best result obtained in the competition.

Keywords

Sexism Identification, Data Augmentation, Machine Translation, Transformers, Natural Language Processing

1. Introduction

The large amount of content posted daily on social networks and their rapid dissemination make it necessary to combat inappropriate behaviors, such as sexist conducts. Sexism is prejudice or discrimination based on sex. It can be expressed in different ways, such as directly, indirectly or descriptively [1], and can undermine women ideologically, sexually, with hatred, by objectifying them, etc. Most of the works and competitions organized so far focus only on one form of sexism, misogyny or hatred towards women [2, 3, 4, 5, 6], such as the AMI shared task [7, 8], on the automatic identification of misogyny in Twitter, and HatEval [9], on the detection of hate speech against immigrants and women. The shared task EXIST, sEXism Identification in Social neTworks, was created in 2021 [10] to address the main forms of sexism, including misogyny, inequality, objectification, sexual violence and dominance. Its goal is to promote the development of automatic systems to identify and classify sexism in English and Spanish languages.

IberLEF 2022, September 2022, A Coruña, Spain.

*Corresponding author.

†These authors contributed equally.

daniel.gbaena@gmail.com (D. García-Baena); magc@ujaen.es (M. García-Cumbreras); sjzafra@ujaen.es (S. M. Jiménez-Zafra); mgarcia@ujaen.es (M. García-Vega)

0000-0002-3334-8447 (D. García-Baena); 0000-0003-1867-9587 (M. García-Cumbreras); 0000-0003-3274-8825 (S. M. Jiménez-Zafra); 0000-0003-2850-4940 (M. García-Vega)

Table 1
Dataset statistics for *Task 1: Sexism identification*

Data	Class	Language		Total
		English	Spanish	
Training	**Sexism**	2,794	2,864	5,658
	Non-sexism	2,850	2,837	5,687
Test		526	532	1,058
Total		6,170	6,233	12,403

EXIST 2022 [11] is the second edition of the task and is organized at IberLEF workshop. The organizers proposed two tasks: *Task 1: Sexism identification* and *Task 2: Sexism categorization*. This work presents the system we developed for our participation in Task 1 that is based on exploring data augmentation and machine translation.

The rest of the paper is organized as follows. Firstly, Section 2 describes the task and the data provided. Secondly, Section 3 presents the proposed system for addressing Task 1. Following, Section 4 shows the results obtained and a discussion thereof. Finally, Section 5 completes the paper with some conclusions and the future work.

2. Task description

The shared task EXIST 2022 [11], organized at IberLEF workshop, aims to identify and categorize sexist content in social networks. Specifically, it proposes two tasks to detect sexism in English and Spanish. *Task 1: Sexism identification* is a binary classification task that consists of determining whether a text is sexist or not. *Task 2: Sexism categorization* is a multi-class classification task in which, for each text classified as sexist in the first task, it should be determined the facet of women that is undermined according to the following categories: (1) ideological and inequality, (2) stereotyping and dominance, (3) objectification, (4) sexual violence, and (5) misogyny and non-sexual violence. We participated in *Task 1: Sexism identification*.

During the systems development phase, the complete EXIST 2021 dataset [10] was provided as training set. This dataset contains 5,644 texts written in English and 5,701 in Spanish from Twitter and Gab. For testing the systems, 1,058 tweets written in Spanish and English and collected in January 2022 were supplied. The organizers chose not to release the gold labels of this dataset, so its distribution cannot be provided. The statistics of the EXIST 2022 dataset for Task 1 are presented in Table 1.

Finally, it is worth mentioning that in order to evaluate the participants' systems, the organizers used the Evaluation Framework EvALL [12] and selected the Accuracy measure to rank the participants in Task 1.

3. System description

In this work, we explore the use of Transformers and two techniques:

1. **Data augmentation**. Given that Deep Learning systems improve their performance when they incorporate a larger volume of data [13, 14, 15], a search for task-related datasets has been performed (explained below).
2. **Machine translation**. Since there are languages that have hardly any dedicated resources for training Deep Learning systems, we are going to test how the use of datasets in other languages is affected by applying machine translation.

3.1. Data Augmentation and Machine Translation

The performance of most Machine Learning models, and Deep Learning models in particular, depends on the quality, quantity and relevance of the training data. Data augmentation is a set of techniques to artificially increase the amount of data. This includes making small changes to data or using Deep Learning models to generate the new one. Under this approach, our intention was to augment the training data with existing, available, task-related datasets. In addition, we used machine translation to translate all the Spanish data to English, as the system was developed to classify English texts. For this, we used the free and unlimited python library Googletrans [16], that implemented Google Translate API.

After performing a search for available and task-related datasets, we selected the following ones:

- AMI 2018 [17]. It is the dataset provided in the AMI 2018 task, whose objective was the automatic identification of misogynous content, both in Spanish and English languages, in Twitter. The AMI shared task was organized according to two main sub-tasks: *misogyny identification*, discrimination of misogynist contents from the non-misogynist ones, and *misogynistic behaviour and target classification*, recognition of the targets that can be either specific users or groups of women together with the identification of the type.
- Spanish MisoCorpus-2020 [18]. It is a balanced corpus regarding misogyny in Spanish. It is classified into three subsets: i) violence towards relevant women, ii) messages harassing women in Spanish from Spain and Spanish from Latin America, and iii) general traits related to misogyny.
- ISEP Sexist [19]. It is a dataset of sexist statements written in English. It presents more than 1,100 examples of statements of workplace sexism, roughly balanced between examples of certain sexism and ambiguous or neutral cases.

Table 2 shows some features and statistics of these three datasets, and the EXIST 2022 dataset provided by the organization.

3.2. Deep Learning model

We chose Hugging Face open source state-of-the-art machine learning library Transformers [20] and their free public collection of pre-trained models in order to review a currently popular choice for Artificial Intelligence developers and researchers. Specifically, we used: distilbert-base-uncased-finetuned-sst-2-english [21], one of the most downloaded text-classification/English

Table 2
Datasets features

Dataset	# docs	# language
EXIST 2022 (1)	11,345	English
AMI 2018 (2)	20,483	English/Spanish
Spanish MisoCorpus-2020 (3)	19,735	Spanish
ISEP Sexist (4)	12,482	English

models available on Hugging Face. This model is a fine-tune checkpoint of distilbert-base-uncased, fine-tuned on SST-2. distilbert-base-uncased-finetuned-sst-2-english reaches an accuracy of 91.3% on the dev set (for comparison, BERT bert-base-uncased version achieves an accuracy of 92.7%).

We selected this model not only due to its popularity but because it was fine-tuned on the widely used SST-2, the Stanford Sentiment Treebank dataset [22], that as a sentiment analysis corpus can be considered related to the EXIST 2022 dataset. We checked some of the other most downloaded text-classification Hugging Face available options as the already referenced bert-base-uncased [23] and nlptown/bert-base-multilingual-uncased-sentiment [24] before of selecting distilbert-base-uncased-finetuned-sst-2-english owing to its better accuracy results on the training stage.

3.3. Data preprocessing

In relation with data preprocessing, the texts were translated into English using the Python library Googletrans [16], and tokenized using the distilbert-base-uncased-finetuned-sst-2-english tokenizer provided by Hugging Face. Moreover, all URL links and mentions were removed.

3.4. Experiments

With the EXIST 2022 training data, we performed different experiments using the Pareto Principle or 80/20 rule (80% of the data for training and 20% for validation). Specifically, we used DistilBERT [21] with default parameters fine-tuned on different set of corpus: i) EXIST 2022, ii) EXIST 2022 and AMI 2018, iii) EXIST 2022 and Spanish MisoCorpus-2020, iv) EXIST 2022 and ISEP Sexist, and v) a combination of all the corpus (EXIST 2022, AMI 2018, Spanish MisoCorpus-2020, and ISEP Sexist). It should be noted that at no point did we perform hyperparameter fitting of the Transformer model.

Table 3 shows the results obtained in all the experiments, always using distilbert-base-uncased-finetuned-sst-2-english for pretraining. The best accuracy result was obtained by the EXIST 2022 and Spanish Misocorpus-2020 combination (0.7968), followed by the combination of all datasets (0.7917). We focused on accuracy, as it was the measure selected by the organizers to rank the systems for Task 1. In all cases, accuracy and F1 results from Table 3 refer to the training phase.

Table 3
Training results

Datasets	Accuracy	F1
EXIST 2022 (1)	0.7616	0.7615
(1) + AMI 2018 (2)	0.7606	0.7605
(1) + Spanish MisoCorpus-2020 (3)	**0.7968**	0.7967
(1) + ISEP Sexist (4)	0.7753	0.7749
(1) + (2) + (3) + (4) (5)	**0.7917**	0.7915

Table 4
Approach tested in each run

Run	Approach
Run 1	DistilBERT fine-tuned with EXIST 2022 (1)
Run 2	DistilBERT fine-tuned with EXIST 2022 + Spanish MisoCorpus-2020 (3)
Run 3	DistilBERT fine-tuned with EXIST 2022 + AMI 2018 + Spanish MisoCorpus-2020 ISEP Sexist (5)

4. Results and discussion

This section presents the results obtained in the test phase of *Task 1: Sexism identification*. In total, 45 systems were evaluated. The organizers selected accuracy for ranking the systems and each participating team could submit 3 runs. We selected our 3 runs based on the experiments carried out on the training phase. Table 4 shows the approach tested in each of the runs.

The performance of our 3 approaches by language (EN, ES and ALL) together with the results of the baseline system (TF-IDF+SVM) provided by the organisers are reported in Table 5. The best result, for each language and overall, has been obtained with the system trained on the corpus of the task itself (Run 1). In all cases, the baseline has been exceeded with similar values, improving by about 6% on F1.

In no case has a fine tuning of hyperparameters been performed, as we wanted to test the performance of including different datasets and even the use of datasets from other languages with the incorporation of automatic translation. For a more in-depth analysis we would need the test set with the gold labels, which at the time of writing this paper is not available.

If we compare the results of our system with the rest of the participants and the baseline, we can see that the difference between the best result and ours, in all cases, does not reach one tenth, being above the average of the systems presented. Table 6 shows the values obtained by our best system, for each language and in general, as well as the best result, the baseline and the average of all the systems presented.

Table 5
Results in the test data

Lang	Method	Accuracy	Precision	Recall	F1	Ranking
All	Baseline	0,6928	0,6919	0,685	0,6859	x
All	1	0,7316	0,7353	0,7362	**0,7315**	32
All	2	0,7278	0,7277	0,7295	0,7272	33
All	3	0,7202	0,7181	0,7187	0,7184	35
En	Baseline	0,7167	0,7092	0,7053	0,7068	x
En	1	0,7529	0,7570	0,7632	**0,7520**	32
En	2	0,7529	0,7548	0,7613	0,7517	33
En	3	0,7529	0,7476	0,7520	0,7489	34
Es	Baseline	0,6692	0,6747	0,6673	0,6649	x
Es	1	0,7105	0,7126	0,7113	**0,7103**	33
Es	2	0,7030	0,7029	0,7028	0,7029	34
Es	3	0,6880	0,6880	0,6875	0,6875	38

Table 6
Comparison of results

Lang	Method	Accuracy	Precision	Recall	F1
All	baseline	0,6928	0,6919	0,6850	0,6859
All	best result	0,7996	0,7982	0,7975	0,7978
All	our run	0,7316	0,7353	0,7362	0,7315
All	average	0,7247	0,7266	0,7264	0,7229
En	baseline	0,7167	0,7092	0,7053	0,7068
En	best result	0,8422	0,8388	0,8365	0,8376
En	our run	0,7529	0,7570	0,7632	0,7520
En	average	0,7708	0,7694	0,7716	0,7665
Es	baseline	0,6692	0,6747	0,6673	0,6649
Es	best result	0,7801	0,7808	0,7805	0,7801
Es	our run	0,7105	0,7126	0,7113	0,7103
Es	average	0,7338	0,7370	0,7344	0,7330

5. Conclusions and future work

In this paper we have presented the participation of the SINAI team in *Task 1: Sexism identification* of the EXIST@IberLEF 2022 shared task. The aim of our experiments was to test how Transformer models behaves incorporating different related datasets in the training stage. The main conclusion is that using the task dataset itself for training gives better results than incorporating other related datasets.

In the future, we plan to continue testing external resources to improve the training phase of the system by analyzing the contribution of each dataset, using different machine translation systems, and testing transfer learning systems as well as general topic datasets.

Acknowledgments

This work has been partially supported by Big Hug project (P20_00956, PAIDI 2020) and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government, and LIVING-LANG project (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe. Salud María Jiménez-Zafra has been partially supported by a grant from Fondo Social Europeo and the Administration of the Junta de Andalucía (DOC_01073).

References

[1] P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, M. Coulomb-Gully, He said "who's gonna take care of your children when you are at acl?": Reported sexist acts are not sexist, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4055–4066.

[2] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.

[3] S. Frenda, B. Ghanem, M. Montes-y Gómez, P. Rosso, Online hate speech against women: Automatic identification of misogyny and sexism on twitter, Journal of Intelligent & Fuzzy Systems 36 (2019) 4743–4752.

[4] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, R. Valencia-García, Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings, Future Generation Computer Systems 114 (2021) 506 – 518. URL: http://www.sciencedirect.com/science/article/pii/S0167739X20301928. doi:10.1016/j.future.2020.08.032.

[5] F.-M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López, M. T. Martín-Valdivia, Detecting misogyny and xenophobia in spanish tweets using language technologies, ACM Transactions on Internet Technology (TOIT) 20 (2020) 1–19.

[6] J. A. García-Díaz, S. M. Jiménez-Zafra, M. A. García-Cumbreras, R. Valencia-García, Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers, Complex & Intelligent Systems (2022) 1–22.

[7] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami), EVALITA Evaluation of NLP and Speech Tools for Italian 12 (2018) 59.

[8] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at ibereval 2018., IberEval@ SEPLN 2150 (2018) 214–228.

[9] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al., Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 54–63.

[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.

[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).
[12] E. Amigó, J. Carrillo-de Albornoz, M. Almagro-Cádiz, J. Gonzalo, J. Rodríguez-Vidal, F. Verdejo, Evall: Open access evaluation for information access systems, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 1301–1304.
[13] G. Marcus, Deep learning: A critical appraisal, arXiv preprint arXiv:1801.00631 (2018).
[14] T. Panch, P. Szolovits, R. Atun, Artificial intelligence, machine learning and health systems, Journal of global health 8 (2018).
[15] K. Y. Ngiam, W. Khor, Big data and machine learning algorithms for health-care delivery, The Lancet Oncology 20 (2019) e262–e273.
[16] Googletrans, Googletrans, 2022. URL: https://github.com/ssut/py-googletrans.
[17] M. Anzovino, E. Fersini, P. Rosso, Automatic misogyny identification, 2018. URL: https://drive.google.com/drive/folders/13UfLXcPTvT9bEAPP8tLj2quXGGa2gsTq.
[18] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, R. Valencia-García, Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings, Future Generation Computer Systems 114 (2021) 506–518. URL: https://www.sciencedirect.com/science/article/pii/S0167739X20301928. doi:https://doi.org/10.1016/j.future.2020.08.032.
[19] D. Grosz, P. C. Céspedes, Automatic detection of sexist statements commonly used at the workplace, CoRR abs/2007.04181 (2020). URL: https://arxiv.org/abs/2007.04181. arXiv:2007.04181.
[20] Hugging Face Transformers, Hugging face transformers, 2022. URL: https://github.com/huggingface/transformers.
[21] distilbert-base-uncased-finetuned-sst-2-english, distilbert-base-uncased-finetuned-sst-2-english, 2022. URL: https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english.
[22] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA, 2013, pp. 1631–1642. URL: https://aclanthology.org/D13-1170.
[23] bert-base-uncased, bert-base-uncased, 2022. URL: https://huggingface.co/bert-base-uncased.
[24] nlptown/bert-base-multilingual-uncased-sentiment, nlptown/bert-base-multilingual-uncased-sentiment, 2022. URL: https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment.