Dandys-de-BERTganim at EXIST 2025: a Multi-task Learning Architecture for Sexism Identification

Marc Hurtado[1], Aleixandre Tarrasó[1]

[1]Universitat Politècnica de València, Camí de Vera, sn 46022 València, Spain

Abstract

This paper presents Dandys, our system developed for the EXIST 2025 shared task on sexism identification in social media at CLEF 2025. The challenge comprises three interconnected tasks: binary sexism detection, source intention classification, and sexism type categorization. To address these tasks, we adopt a multi-task learning architecture with language-specific transformers for English and Spanish tweets, integrating demographic information from annotators as contextual signals. We enhance model generalization through data augmentation techniques such as back-translation and a punctuation-based augmentation method. Furthermore, we introduce a soft-labeling data reader to better reflect annotation disagreement, aligning with the Learning with Disagreement paradigm. Our results demonstrate the effectiveness of leveraging task interdependence, soft supervision, and multilingual modeling for addressing complex sociolinguistic classification problems. Given good performance across all tasks, our architecture is validated as both robust and effective.

Keywords

Sexism Detection, Social Media, Multi-task Learning, Data Augmentation, Annotator Demographics, Soft Labeling, Multilingual NLP, Learning with Disagreement, EXIST 2025

1. Introduction

Social media platforms have become key components of modern communication and information exchange. However, they also enable the spread of harmful and discriminatory content, including sexist and misogynistic expressions. Therefore, identifying and filtering such content is essential to promote a respectful and discrimination-free online environment. To address this challenge, we participate in EXIST (sEXism Identification in Social neTworks) [1] [2], a shared task initiative aimed at detecting different forms of sexism. Specifically, we participated in all three subtasks of Task 1, which focused on detecting sexism in social media texts.

- **Subtask 1.1: Sexism Detection.** This binary classification task determines whether a tweet contains sexist language or behavior.
- **Subtask 1.2: Source Intention Classification.** Once a tweet is labeled as sexist, this multiclass task categorizes the tweet based on the author's intent.
- **Subtask 1.3: Sexism Type Categorization.** This multilabel classification task assigns one or more specific types of sexist behavior to each tweet.

The EXIST dataset consists of annotated tweets in both English and Spanish. Each tweet is labelled by six different individuals, whose interpretations are inevitably shaped by their personal backgrounds. The dataset includes not only the six annotations but also the demography information (gender, age, ethnicity, education level, and nationality) of the six annotators. For this reason, we adopt the Learning With Disagreement paradigm, which embraces divergent annotations from multiple annotators. This approach enables models to learn from a wider range of perspectives and subjective interpretations, fostering fairer and more robust systems.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

mhurben@upv.es (M. Hurtado); atarsor@upv.es (A. Tarrasó)

https://github.com/Marxx01 (M. Hurtado); https://github.com/Sorni18 (A. Tarrasó)

2. Dandys-de-BERTganim System

In this section, we provide a comprehensive description of the components, design choices, and implementation details of the Dandys system that we developed for the EXIST 2025 shared task.

2.1. Data Augmentation

Data augmentation is crucial for improving model robustness and generalization, especially in low-resource or highly imbalanced settings like sexism detection on social media, where linguistic variability and noisy user-generated content can hinder performance. To increase the volume and diversity of our training data, we applied two complementary data augmentation strategies: back-translation and punctuation-based augmentation.

Back-translation is a technique that involves translating a piece of text from its original language into another language and then translating it back into the original language. This process helps generate new linguistic variants while preserving the semantic content of the text. For both English and Spanish, we used the state-of-the-art Helsinki-NLP opus-mt models[3] to translate the tweets. Figure 1 shows a schema of the Back-translation strategy and Table 1 presents an example of the back-translation process. Moreover, due to tweets being written in Spanish and English, the first translation is also used to augment the dataset of the other language.

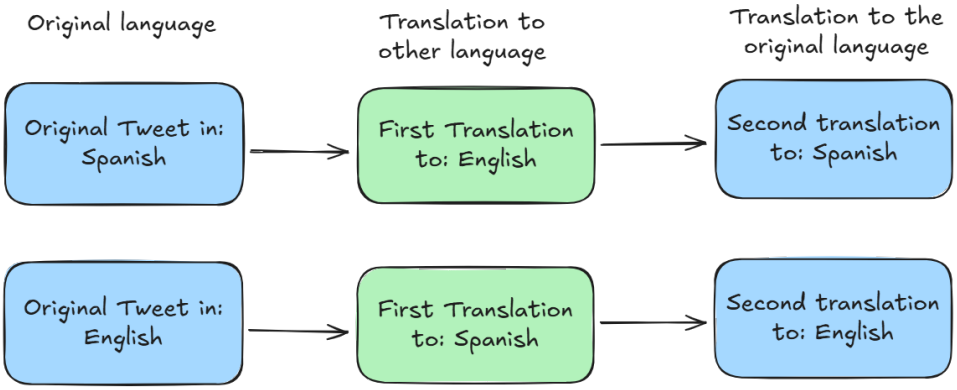

Figure 1: Schema of the Back-Translation strategy used for data augmentation

Table 1
Example of back-translation process. The original Spanish tweet is first translated to English and then translated back into Spanish, introducing subtle lexical and syntactic variation. We also clean the information, removing URLs, emojis, and @ symbols.

Lang	Original Sentence	First Translation (EN)	Back-Translation (ES)
ES	@Steven2897 Lee sobre Gamergate, y como eso ha cambiado la manera en la cual nos comunicamos en el internet. Los fanboys de Halo están tóxicos pero los fanboys de otras comunidades/juegos también han querido coger pauta con eso	Read about Gamergate and how that has changed the way we communicate on the internet. Halo fans are toxic but fans from other communities/games have also wanted to get along with that.	Lea sobre Gamergate y cómo eso ha cambiado la forma en que nos comunicamos en Internet. Los fans de Halo son tóxicos, pero los fans de otras comunidades/juegos también han querido llevarse bien con eso.

In addition, we implemented a punctuation-based augmentation method known as AEDA[4] (An Easier Data Augmentation). This technique involves inserting random punctuation marks into the sentence without altering its meaning. Figure 2 shows a schema of the combination of Back-translation and AEDA-Augmented strategies. We used a punctuation insertion ratio of approximately 0.3, meaning that punctuation was added before about 30% of the words. This approach helped enrich the syntactic variability of the dataset, which in turn supports better generalization in model training. Table 2 presents an example of the AEDA-Augmented process.

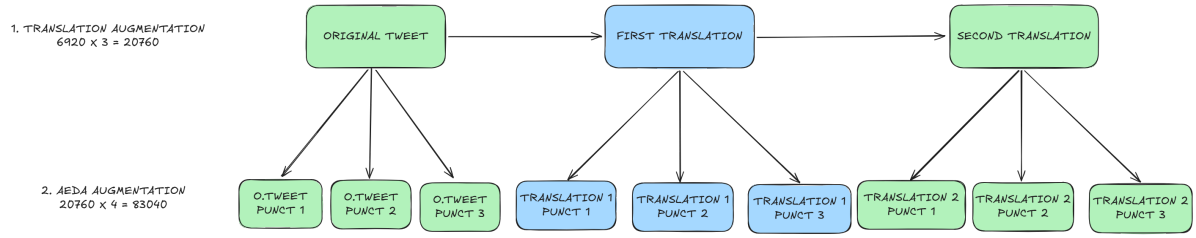

Figure 2: AEDA

Table 2
Example of AEDA augmentation applied to a Spanish tweet. The method inserts random punctuation into the original sentence to introduce variability while preserving the original meaning.

Lang	Original Sentence	AEDA-Augmented Sentence
ES	@Steven2897 Lee sobre Gamergate, y como eso ha cambiado la manera en la cual nos comunicamos en el internet. Los fanboys de Halo están tóxicos pero los fanboys de otras comunidades/juegos también han querido coger pauta con eso.	Lee : sobre Gamergate y ; como : eso ha : cambiado la manera en la cual nos comunicamos en ! el . internet Los fanboys de Halo están tóxicos pero los fanboys de , otras comunidades/juegos también han querido coger pauta con eso.

2.2. Dual dataset processing

During our experiments, we found it necessary to implement two distinct data processing components to address both the hard (single-label) and soft (distributional-label) subtasks:

- **Hard-Labeling Data Reader** Designed for the hard subtask, this reader applies a majority-vote strategy: each tweet is assigned exactly one label corresponding to the class chosen by the majority of annotators.
  - *Advantages*: Simplicity and speed in label generation.
  - *Limitations*: Discards annotation disagreement, potentially introducing label noise for ambiguous or context-dependent content.
- **Soft-Labeling Data Reader** Developed for the soft subtask, this reader preserves the full distribution of annotator responses. Instead of collapsing all annotations into a single class, it returns a probability vector reflecting class agreement. For example:

$$\left[0.40_{\text{Direct}},\ 0.30_{\text{Reported}},\ 0.30_{\text{Judgmental}}\right].$$

  - *Advantages*: Captures the subjectivity and uncertainty inherent in social media discourse. Also, enables the model to learn from disagreement patterns, improving probability calibration and robustness.

Both readers share the same preprocessing pipeline (normalization, tokenization, emoji handling), but differ in label construction:

- When the *hard* reader is used, the model is trained using standard cross-entropy loss on one-hot label vectors, which include an additional "ambiguous" class to represent uncertain cases.
- When the *soft* reader is used, the model is trained using Kullback–Leibler divergence loss (KLDivLoss) on distributional label vectors, following the *Learning with Disagreement* paradigm.

This dual-reader design allows our multi-task architecture to be trained in two complementary modes:

1. With hard labels (including the extra ambiguous class) for maximizing precision on clear-cut examples.
2. With soft labels and KL divergence supervision for capturing nuances and improving robustness on ambiguous cases.

3. Annotators Information

Another important component of our system involves leveraging the demographic metadata of the annotators[5]. Each tweet in the dataset is labeled by six different individuals, whose interpretations are inevitably shaped by their personal background[6]. To better capture this annotator-specific context, we designed a module that encodes demographic information—specifically, gender, age, ethnicity, education level, and nationality—into fixed-size vector representations. These embeddings aim to model the diversity of perspectives that arise from different social and cultural positions.

To achieve this, we precomputed a set of 78,900 sentence-level embeddings, each representing a unique combination of annotator demographic attributes, as shown in Table 3. Each sentence was then encoded into a dense vector using the LaBSE sentence transformer model and stored in a lookup table for efficient retrieval during training and inference.

Table 3
Annotators embeddings creation example.

Full Sentence	The {**index**} annotator is a {**gender**}, {**age**} years old, of {**ethnicity**} ethnicity, with {**level**} education level, and nationality {**nationality**}.
Example	The first annotator is a woman, 34 years old, of Asian ethnicity, with university education level, and nationality Japanese.

During training and inference, for each tweet, we retrieve the demographic descriptions of its six annotators and extract the corresponding embeddings from the table. These six vectors are then averaged to obtain a single representation that summarizes the annotator context for that tweet. This aggregated embedding is used as an additional input to the classifier, providing rich contextual signals that help the model to interpret annotation patterns more accurately—particularly in subjective or ambiguous cases. This process is shown in Figure 3

However, in this edition of the task, we observed that incorporating annotator embeddings did not lead to significant performance improvements. Unlike previous years, where annotator demographics were sometimes unevenly distributed—resulting in patterns or biases that the model could exploit—this year's dataset presents a more balanced and diverse allocation of annotators across tweets. As a result, the demographic information of annotators does not offer strong predictive cues about their labeling tendencies. This limits the added value of explicitly modeling their profiles, although we believe such

contextual representations remain a promising direction for future work, especially in scenarios where annotator bias is more prominent or unevenly distributed.

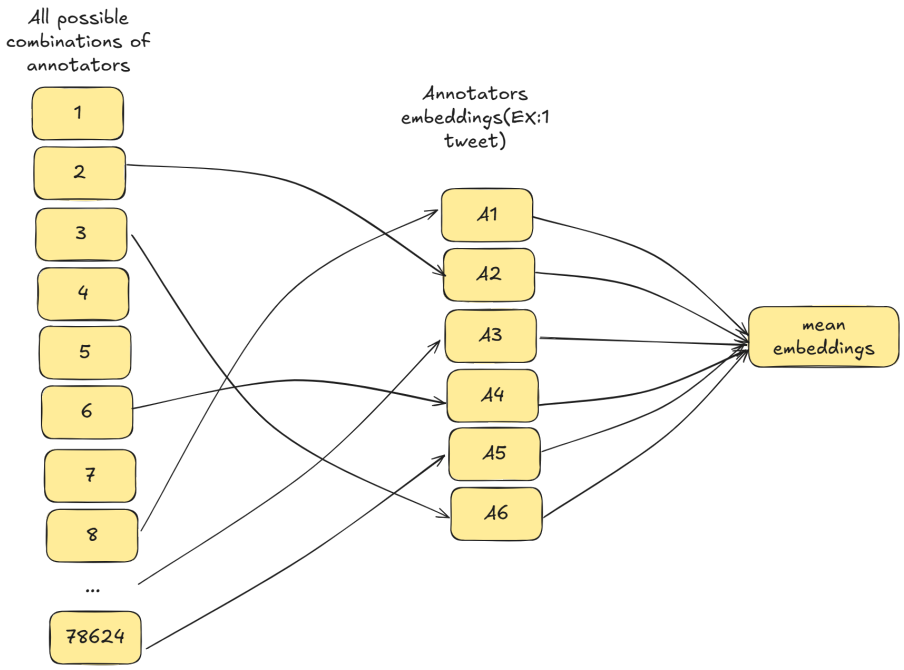

Figure 3: Annotators embeddings creation process.

4. Model Architecture and Fine-tuning Process.

In this section, we explain the architecture of the model. First, we will focus on the core architecture, and the different steps that are carried out within it. Later, we will explain the multitasking architecture implemented and how we have applied it to the different subtasks.

The backbone encoder shown in Figure 4 is a stack of Transformer layers that processes the input text to generate contextual embeddings. It begins with an input embedding layer that combines token embeddings with positional information. This is followed by multiple identical layers, each containing a multi-head self-attention mechanism and a feed-forward neural network. Residual connections and layer normalization are applied after each sub-layer to stabilize and enhance learning. As the input passes through these layers, the model captures increasingly complex patterns and relationships between words, resulting in a rich contextual representation of the input sequence.

Our model architecture is based on a multi-task[7][8] design that simultaneously handles all three classification subtasks. We developed two separate but identical model instances: one for English tweets and another for Spanish tweets. The English model is based on the xlm-roberta-base-sentiment transformer [9], while the Spanish model uses robertuito-sentiment-analysis [10] [11] [12].

In addition, for subtask 2 (source intention classification), we used a single multilingual model twitter-xlm-roberta-large-2022 for both Spanish and English. This choice was motivated by the benefits of training on a combined dataset, which provided more annotated examples per class and improved generalization. Two of our submissions relied exclusively on this multilingual model, trained on a broader set of tweets spanning both languages. This approach allowed us to leverage a larger and more diverse training set, resulting in slight performance gains, particularly on

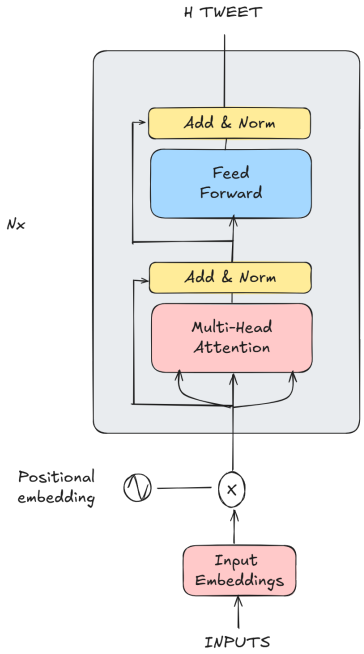

Figure 4: Backbone encoder.

underrepresented categories and borderline cases. The multilingual setup proved especially effective when Spanish and English data were merged, offering richer supervision and more robust generalization.

Once a tweet is processed by its respective transformer, the resulting tweet the embedding of the [CLS] token, is concatenated with the average annotator embedding. This joint representation is then passed through a shared encoder, which feeds into three separate output heads—each responsible for one of the classification tasks. Each output head consists of a dense layer with ReLU activation followed by another dense layer with a softmax or sigmoid activation function to produce the final class probabilities.

This architecture offers several advantages: by using a shared encoder, we reduce the total number of parameters, making the model more efficient. The multi-task learning setup promotes a shared representation learned across tasks, which can lead to improved performance by leveraging complementary information. Finally, the dual-model and multilingual strategies allow for both language-specific fine-tuning and cross-lingual knowledge transfer. A schema of the model architecture is shown in Figure 5.

A critical component in obtaining strong performance is the selection of key hyperparameters, such as learning rate, batch size, weight decay, dropout rate, and class imbalance handling strategies. To this end, we employed Optuna[13], an automatic hyperparameter optimization framework. Optuna enabled efficient exploration of the search space through Bayesian optimization and pruning, allowing us to identify optimal configurations for each model variant with reduced computational cost.

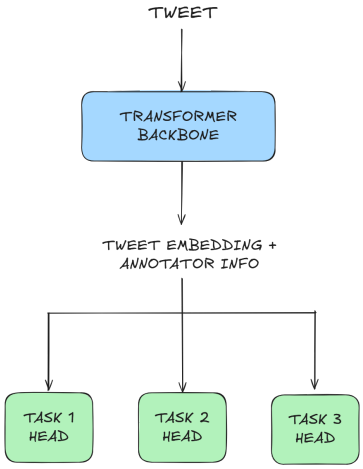

Figure 5: Full Model Architecture.

5. Results

The following tables present the official competition results for the EXIST2025 shared task on sexism identification. We compare our Dandys system against the gold-standards and the subtask winner across all three subtasks in both the soft (distributional labels) and hard (single-label) settings. We report Information Content Metrics (ICM), normalized ICM (ICM Norm), and task-specific performance measures: cross-entropy loss for soft subtasks, F1-score for binary detection (Task 1), and macro-F1 for the multiclass subtasks (Tasks 2 and 3). Notably, our performance on Task 2 (Source Intention) is good, demonstrating that our multi-task architecture and augmentation strategies are particularly effective for modeling intent.

Table 4
SUBTASK1 (Sexism Detection) – Soft setting

System	Team-Ranking	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2025-test_gold	-	3.1182	1.0000	0.5472
GrootWatch_1	1	1.0600	0.6700	0.8893
Dandys-de-BERTganim_1	8	0.6575	0.6054	0.7964

Table 5
SUBTASK1 (Sexism Detection) – Hard setting

System	Team-Ranking	ICM-Hard	ICM-Hard Norm	F1 YES
EXIST2025-test_gold	-	0.9948	1.0000	1.0000
Mario_1	1	0.6774	0.8405	0.8167
Dandys-de-BERTganim_2	33	0.5026	0.7526	0.7548

Table 6
SUBTASK2 (Source Intention) – Soft setting

System	Team-Ranking	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2025-test_gold	-	6.2057	1.0000	0.9128
GrootWatch_1	1	-0.4385	0.4647	1.7711
Dandys-de-BERTganim_2	2	-0.7261	0.4415	1.3820

Table 7
SUBTASK2 (Source Intention) – Hard setting

System	Team-Ranking	ICM-Hard	ICM-Hard Norm	Macro F1
EXIST2025-test_gold	-	1.5378	1.0000	1.0000
Mario_1	1	0.4991	0.6623	0.5692
Dandys-de-BERTganim_2	4	0.3752	0.6220	0.5522

Table 8
SUBTASK3 (Sexism Type) – Soft setting

System	Team-Ranking	ICM-Soft	ICM-Soft Norm
EXIST2025-test_gold	-	9.4686	1.0000
GrootWatch_1	1	-1.1034	0.4417
Dandys-de-BERTganim_1	18	-8.7671	0.0370

Table 9
SUBTASK3 (Sexism Type) – Hard setting

System	Team-Ranking	ICM-Hard	ICM-Hard Norm	Macro F1
EXIST2025-test_gold	-	2.1533	1.0000	1.0000
Mario_1	1	0.6519	0.6514	0.6533
Dandys-de-BERTganim_1	10	0.2244	0.5521	0.5827

5.1. Analysis of the results

When comparing our system, Dandys-de-BERTganim, to the top two teams in the EXIST2025 challenge GrootWatch_1 and Mario_1, we observed both strengths and areas for improvement.

In soft settings, our model generally had lower ICM-Soft and normalized scores than GrootWatch_1. For example, in Subtask 1, we scored 0.6575 vs. their 1.0600, and in Subtask 3, our ICM-Soft dropped to –8.7671. However, our Cross Entropy was consistently better, indicating well-calibrated probability outputs. In Subtask 1 Soft, we achieved 0.7964 vs. GrootWatch_1's 0.8893, and in Subtask 2 Soft, we outperformed them again (1.3820 vs. 1.7711) while ranking 2nd overall.

In hard settings, our best result was in Subtask 2, where we ranked 4th, with a Macro F1 of 0.5522, close to Mario_1's 0.5692, the top performer. This suggests strong classification performance in source intention. However, in Subtask 1 Hard, we lagged behind, with an F1 YES of 0.7548 vs. Mario_1's 0.8167, and in Subtask 3 Hard, our F1 of 0.5827 was lower than GrootWatch_1's 0.6305.

Overall, while the top two teams led in label agreement and F1 scores, our system showed competitive results in probabilistic modeling and source intention classification. Future work should focus on improving alignment with annotations and boosting accuracy in stricter evaluation settings.

5.2. Conclusions and Future Work

Across all subtasks, the Dandys system demonstrates competitive performance. While there remains room for improvement in Task 1 and Task 3, our approach shows robustness in handling ambiguous and context-dependent examples. Most notably, in Task 2 (Source Intention), Dandys achieves high normalized ICM scores in both soft and hard settings, underscoring the benefits of our dual-reader design, data augmentation strategies, and multi-task learning framework for modeling nuanced intent in social media posts.

As future work, we plan to explore alternative strategies for integrating annotator demography information into the model, including position-agnostic embeddings and attention mechanisms for demographic information.

Declaration on Generative AI

During the development of this research work, the authors employed multiple generative artificial intelligence (AI) systems to enhance various aspects of the manuscript preparation process. Specifically, large language models (LLMs) including ChatGPT (OpenAI), DeepSeek (DeepSeek AI), Claude (Anthropic), and Gemini (Google DeepMind) were utilized for several critical purposes:

- **Writing Enhancement:** These tools assisted in improving grammatical accuracy, stylistic coherence, and overall readability of the manuscript.
- **Comprehensive Literature Review:** AI systems were used to conduct extensive searches of relevant literature, identify key research gaps, and summarize complex academic sources.
- **Code Optimization:** Where applicable, AI-assisted suggestions were implemented to refine computational methods, debug algorithms, and improve efficiency in software implementations.

The authors rigorously reviewed, fact-checked, and edited all AI-generated content to ensure accuracy, originality, and alignment with scholarly standards. Additionally, all critical insights, theoretical contributions, and final conclusions remain the original work of the authors. The use of these tools was strictly supplementary, and full responsibility for the research content, ethical considerations, and final presentation lies with the authors.

References

[1] L. Plaza, J. C. de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos (Extended Overview), in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[2] L. Plaza, J. C. de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[3] J. Tiedemann, M. Aulamo, D. Bakshandaeva, M. Boggia, S.-A. Grönroos, T. Nieminen, A. Raganato, Y. Scherrer, R. Vázquez, S. Virpioja, Democratizing neural machine translation with opus-mt, arXiv preprint (2022). URL: https://arxiv.org/pdf/2212.01936. arXiv:2212.01936, preprint.

[4] A. Karimi, L. Rossi, A. Prati, AEDA: An easier data augmentation technique for text classification, arXiv preprint (2021). URL: https://arxiv.org/pdf/2108.13230. arXiv:2108.13230, preprint.

[5] Y.-Z. Fang, L.-H. Lee, J.-D. Huang, NYCU-NLP at EXIST 2024: Leveraging Transformers with Diverse Annotations for Sexism Identification in Social Networks, in: Notebook for the EXIST Lab at CLEF 2024, 2024. URL: https://ceur-ws.org/Vol-3740/paper-93.pdf.

[6] Z. L. N. Z. J. G. L. Shengnan An, Zexiong Ma, Make your llm fully utilize the context, arXiv preprint (2024). URL: https://arxiv.org/pdf/2404.16811. arXiv:2404.16811.

[7] J. Zhang, K. Yan, Y. Mo, Multi-task learning for sentiment analysis with hard-sharing and task recognition mechanisms, Information 12 (2021) 207. URL: https://www.mdpi.com/2078-2489/12/5/207. doi:10.3390/info12050207.

[8] O. Sener, V. Koltun, Multi-task learning as multi-objective optimization, in: Advances in Neural Information Processing Systems, volume 31, 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/432aca3a1e345e339f35a30c8f65edce-Paper.pdf.

[9] J. Camacho-collados, K. Rezaee, T. Riahi, A. Ushio, D. Loureiro, D. Antypas, J. Boisson, L. Espinosa Anke, F. Liu, E. Martínez Cámara, et al., Tweetnlp: Cutting-edge natural language processing for social media, Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (2022) 38–49. URL: https://arxiv.org/abs/2206.14774.

[10] J. M. Pérez, M. Rajngewerc, J. C. Giudici, D. A. Furman, F. Luque, L. A. Alemany, M. V. Martínez, pysentimiento: a python toolkit for opinion mining and social nlp tasks, arXiv preprint arXiv:2106.09462 (2021).

[11] J. M. Pérez, D. A. Furman, L. Alonso Alemany, F. M. Luque, RoBERTuito: a pre-trained language model for social media text in Spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7235–7243. URL: https://aclanthology.org/2022.lrec-1.785.

[12] M. García-Vega, M. Díaz-Galiano, M. García-Cumbreras, F. Del Arco, A. Montejo-Ráez, S. Jiménez-Zafra, E. Martínez Cámara, C. Aguilar, M. Cabezudo, L. Chiruzzo, et al., Overview of tass 2020: Introducing emotion detection, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) Co-Located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, 2020, pp. 163–170.

[13] Optuna Development Team, Optuna: A hyperparameter optimization framework, Documentation, 2025. URL: https://optuna.readthedocs.io/en/stable/index.html, accessed: June 1, 2025.