I2C-UHU at EXIST2024: Learning from Divergence and Perspectivism for Sexism Identification and Source Intent Classification

Manuel Guerrero-García*, Manuel Cerrejón-Naranjo, Jacinto Mata-Vázquez and Victoria Pachón-Álvarez

I2C Research Group, University of Huelva, Spain

Abstract

In this paper, we present the contributions of the I2C-UHU team to the EXIST2024 Lab at CLEF 2024, focusing on the identification of sexism and the classification of source intent in social media texts. State-of-the-art transformer models are employed to address the complex and nuanced nature of sexist language. We adopt a two-fold approach: firstly, classifying tweets as sexist or non-sexist, and secondly, categorizing sexist tweets based on intent. Our innovative approach, employing Learning with Disagreement, incorporates diverse perspectives from multiple annotators, enhancing the robustness and accuracy of our models. We detail our data preprocessing, augmentation techniques, and hyperparameter optimization strategies. Our results in the competition demonstrated effectiveness, with our entries achieving positive rankings in the two tasks in which we participated. In Task 1, we secured the 10th position out of 70 participants on the hard labels leaderboard and the 13th position out of 40 for soft labels. In Task 2, we achieved the 11th position out of 46 participants for hard labels and the 17th position out of 35 in the best run for soft labels. Our findings provide a foundation for future research and practical applications in social media moderation and policy-making.

Keywords

Sexism identification, Learning with disagreement, Transformer models, Natural language processing

1. Introduction

In the EXIST2024 Lab at CLEF 2024[1], the I2C-UHU team addressed sexism on social media platforms through binary classification of tweets and classification based on author intent. The first task distinguishes between sexist and non-sexist content, crucial for filtering harmful language, while the second task classifies sexist tweets into direct, reported, and judgmental categories, providing deeper insights into manifestations of sexism. Utilizing transformer models and data augmentation, our approach aims for robustness and generalizability. By implementing "Learning with Disagreement" [2] we capture diverse perspectives from human annotators, enhancing model accuracy. The paper structure includes sections on related works, dataset description, methodology, results, and future research directions.

2. Related Works

In the realm of detecting sexist tweets, researchers use various methodologies to navigate the complexities of language and intent. Binary classification models serve as a foundational tool, offering a clear distinction between sexist and non-sexist content. However, the quest for a deeper understanding prompts the exploration of author intent, which requires delving into contextual cues and linguistic subtleties.

Task 1 of EXIST 2024 [3] is dedicated to binary categorization, where researchers have explored a spectrum of techniques. From traditional rule-based systems to cutting-edge deep learning architectures, the goal remains consistent: to accurately identify instances of sexism in tweets. Notable among these endeavors is the work of Burnap and Williams [4], who leveraged automatic classification techniques

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

manuel.guerrero790@alu.uhu.es (M. Guerrero-García); manuel.cerrejon886@alu.uhu.es (M. Cerrejón-Naranjo); mata@uhu.es (J. Mata-Vázquez); vpachon@dti.uhu.es (V. Pachón-Álvarez)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

to detect hate speech on Twitter. Their approach, which incorporated linguistic and contextual features, showcased significant accuracy in pinpointing problematic content.

Task 2, however, takes a deeper dive into the realm of author intent, recognizing that the mere presence of sexist language does not always imply malicious intent. To address this, researchers delve into the intricate interplay between language, context, and underlying motives. Waseem and Hovy [5] embarked on this journey by identifying predictive features for hate speech detection, underscoring the importance of contextual and demographic attributes in discerning the author's intent.

In sum, the exploration of related works underscores the multidimensional nature of detecting sexist tweets. While binary classification models provide a solid foundation, the pursuit of a more nuanced understanding necessitates the integration of author intent analysis and cutting-edge transformer models. These endeavors collectively advance our comprehension of sexism in online discourse and pave the way for more effective mitigation strategies.

3. Tasks and Dataset Description

In this section, the tasks in which participation was engaged and the datasets provided by the organizers are delineated.

3.1. Task 1: Sexism Identification in Tweets

Task 1 involves a binary classification problem where the objective is to determine whether a given tweet contains sexist expressions or behaviors. The classification is straightforward: each tweet is categorized as either sexist (*"YES"*) or not sexist (*"NO"*). Examples of sexist tweets include statements that directly express sexist sentiments, describe sexist situations, or criticize sexist behaviors. For instance, tweets that demean women's capabilities, perpetuate stereotypes, or contain derogatory comments fall into the *"YES"* category. Conversely, tweets that do not exhibit these characteristics are labeled as *"NO"*.

3.2. Task 2: Source Intention in Tweets

Task 2 is a multi-class classification task aimed at understanding the intention behind sexist tweets. This task only applies to tweets already identified as sexist in Task 1. The intention of the tweet's author is classified into one of three categories:

- **DIRECT:** The tweet itself is overtly sexist. For example, a tweet stating, *"A woman's place is in the home,"* directly conveys a sexist message.
- **REPORTED:** The tweet reports or describes a sexist incident or situation. An example is, *"Today, I saw a man harass a woman on the subway."*
- **JUDGEMENTAL:** The tweet condemns or criticizes sexist behaviors or situations. For instance, *"It's disgraceful how women are still paid less than men for the same work."*

Each of these categories provides insight into the various ways sexism can manifest and the different contexts in which it is discussed on social media.

3.3. Dataset Description

The dataset provided by the organizers contains over 8000 labeled tweets in English and Spanish, with balanced language distribution. The training dataset has 6920 tweets and the development dataset 1038 tweets. Provided in JSON format, each tweet includes attributes such as *"id_EXIST"*, *"lang"*, *"tweet"*, *"number_annotators"*, and detailed annotator information (*"annotators"*, *"gender_annotators"*, *"age_annotators"*, *"ethnicity_annotators"*, *"study_level_annotators"*, *"country_annotators"*). Labels are *"labels_task1"* for sexist content and *"labels_task2"* for author intent. The *"split"* attribute indicates the dataset subset and language. In Tables 1 and 2 examples of instances for Task 1 and Task 2 are described.

Table 1
Examples of instances for Task 1

id_EXIST	lang	tweet	annotators	labels_task1
101000	es	“No sean de esos que consiguen un peso y cambian con la gente. La plata no es culo.”	Annotator_91, Annotator_92, Annotator_93, Annotator_94, Annotator_95, Annotator_96	NO, NO, NO, NO, YES, NO
201573	en	“@Avigeek96 Well men kill women everyday”	Annotator_549, Annotator_550, Annotator_551, Annotator_552, Annotator_553, Annotator_554	NO, YES, YES, YES, YES, YES

Table 2
Examples of instances for Task 2

id_EXIST	lang	tweet	annotators	labels_task2
101000	es	“No sean de esos que consiguen un peso y cambian con la gente. La plata no es culo.”	Annotator_91, Annotator_92, Annotator_93, Annotator_94, Annotator_95, Annotator_96	-, -, -, -, JUDGEMENTAL, -
201573	en	“@Avigeek96 Well men kill women everyday”	Annotator_549, Annotator_550, Annotator_551, Annotator_552, Annotator_553, Annotator_554	-, REPORTED, JUDGEMENTAL, JUDGEMENTAL, JUDGEMENTAL, REPORTED

These instances were extracted from the file *"training.json"*, which contains 6920 instances, of which 3660 are in Spanish and 3260 are in English. In the case of the file *"dev.json"*, which contains 1038 total instances, the language distribution is 549 for Spanish and 489 for English.

In the training and development datasets, the distribution of ethnicities shows a predominant representation of the "White or Caucasian" group, followed by the "Hispanic or Latino" category. Additionally, regarding educational levels, the most common is "Bachelor’s degree," while the least represented are "Less than high school diploma" and "Doctorate." The class distributions for both the binary classification task (YES/NO) and the multiclass classification task (DIRECT, REPORTED, JUDGMENTAL) demonstrate substantial consistency between the training and validation datasets. The class distribution in the training and development datasets is depicted in Figure 1.

To effectively apply Learning with Disagreement techniques, it’s important to study how different annotator profiles are distributed across the labeled instances.

For training the dataset, there is an equal number of male and female annotators, with a total of 20760 annotators of each gender and a total of 41520. For the development dataset, the distribution is the same, 50% male (3114) and 50% female (3114), out of a total of 6228 annotators. The age distribution is also equitable across both datasets, with one third of annotators falling into each of the following age groups: 18-22 years, 23-45 years, and over 46 years.

4. Methodology

In this section, the methodology used to develop the model submitted to the competition is described.

As previously described, our approaches are based on the use of transformer-based language models. Given that the provided data is in both English and Spanish, four pre-trained models were chosen.

- **XLM-RoBERTa Base:** A pre-trained language model utilizing the RoBERTa architecture and trained on multiple languages. It excels in efficiently and accurately understanding and generating text in various languages[6].

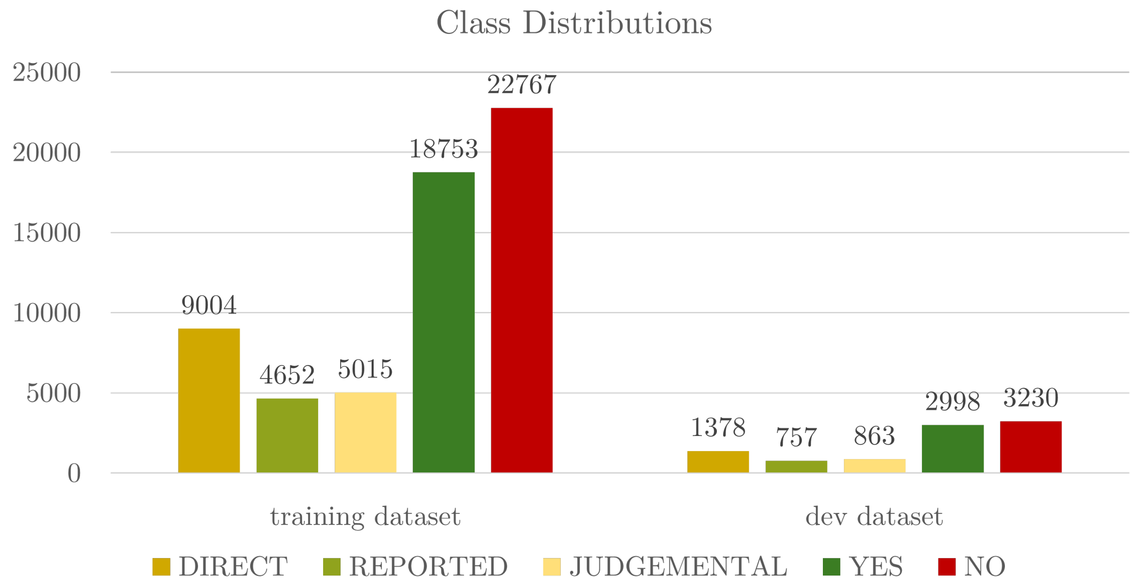

Figure 1: Class distribution in training and dev datasets

- **DeBERTa v3 Base:** A variant of BERT incorporating improvements in attention and word representation, resulting in better performance across a variety of NLP tasks such as text comprehension and language generation[7].
- **RoBERTa Base BNE:** A language-specific adaptation for Spanish of the RoBERTa Base model, trained on the Spanish Corpus from the Spanish Text Bank (BNE). It offers high performance in Spanish language processing tasks[8].
- **BERT Base Multi:** A version of BERT pre-trained in multiple languages and insensitive to case. It can comprehend and generate text in various languages without distinguishing between uppercase and lowercase[9].

4.1. Baseline

The first step in developing classification tasks was to establish an initial benchmark or baseline. This baseline establishes a fundamental methodology that serves as a reference point for comparing more advanced models. It sets a performance threshold that other models must exceed in text classification for our approaches. Two baselines, Version A and Version B, were developed for addressing both Task 1 and Task 2.

4.1.1. Baseline Version A

This approach focuses on training a multiclass classifier (NO, DIRECT, REPORTED, and JUDGEMENTAL) to address all labels for Task 1 and Task 2 simultaneously. The baseline model uses the competition's datasets without preprocessing and with arbitrary hyperparameter values. Both Spanish and English data are included. Models were trained and validated with the training dataset and tested with the development dataset unless otherwise specified for hyperparameter tuning[10].

The hyperparameters values used were: batch size of 32, learning rate of 2e-5, max length of 128, and weight decay of 0.01. The optimizer used was adamw_torch. The maximum number of training epochs was limited to 10 with an "early stopping" set at three epochs.

After training the chosen pre-trained models, the results for the Baseline Version A are presented in Table 3.

This classification strategy yields imprecise and low results. For example, the pre-trained XLM RoBERTa Base model achieved an F1 score of 0.8129 for the NO class, but only 0.3419 for the JUDGE-

Table 3
Results for Baseline Version A

Model	F1 for Baseline A
XLM RoBERTa Base	0.4983
Deberta v3 Base	0.4910
Roberta Base Bne	0.4599
Bert Base Multilingual	0.4388

MENTAL class. This pattern is consistent across other models, indicating difficulty in classifying all the labels together.

4.1.2. Baseline Version B

In Version B, the initial step involves classifying tweets into the two categories of Task 1 (YES and NO). Subsequently, tweets that are categorized as YES are further divided into the three distinct classes of Task 2 (DIRECT, REPORTED, and JUDGEMENTAL). The outcomes achieved with this "Baseline Version B" are detailed in Tables 4 and 5.

Table 4
Results for Baseline Version B, binary classification

Model	F1 for Baseline B
XLM RoBERTa Base	0.7807
Deberta v3 Base	0.7820
Roberta Base Bne	0.7584
Bert Base Multilingual	0.7618

Table 5
Results for Baseline Version B, multiclass classification

Trained Model	F1 for Baseline B
XLM RoBERTa Base	0.568331
Deberta v3 Base	0.555636
Roberta Base Bne	0.530543
Bert Base Multilingual	0.529283

As can be seen, the results improved significantly by breaking down the process into two classification phases.

4.2. Split Description for Training Framework

A schematic overview illustrating the distribution and creation of datasets employed for training the models is shown in Figure 2. These datasets are used in both Task 1 (annotated as v1.x-d) and Task 2 (annotated as v2.x-d). For example:

- The **v1.1 model** is trained and validated with the **v1.1-d dataset**:
  - **v1.1-d train** (training data)
  - **v1.1-d valid** (validation data)
- The **v1.3 model** is trained with the **v1.3-d dataset**:
  - **v1.3-d train** (training data)
  - **v1.3-d valid** (validation data)

- The **v2.1 model** is trained with the **v2.1-d dataset**:
  - **v2.1-d train** (training data)
  - **v2.1-d valid** (validation data)

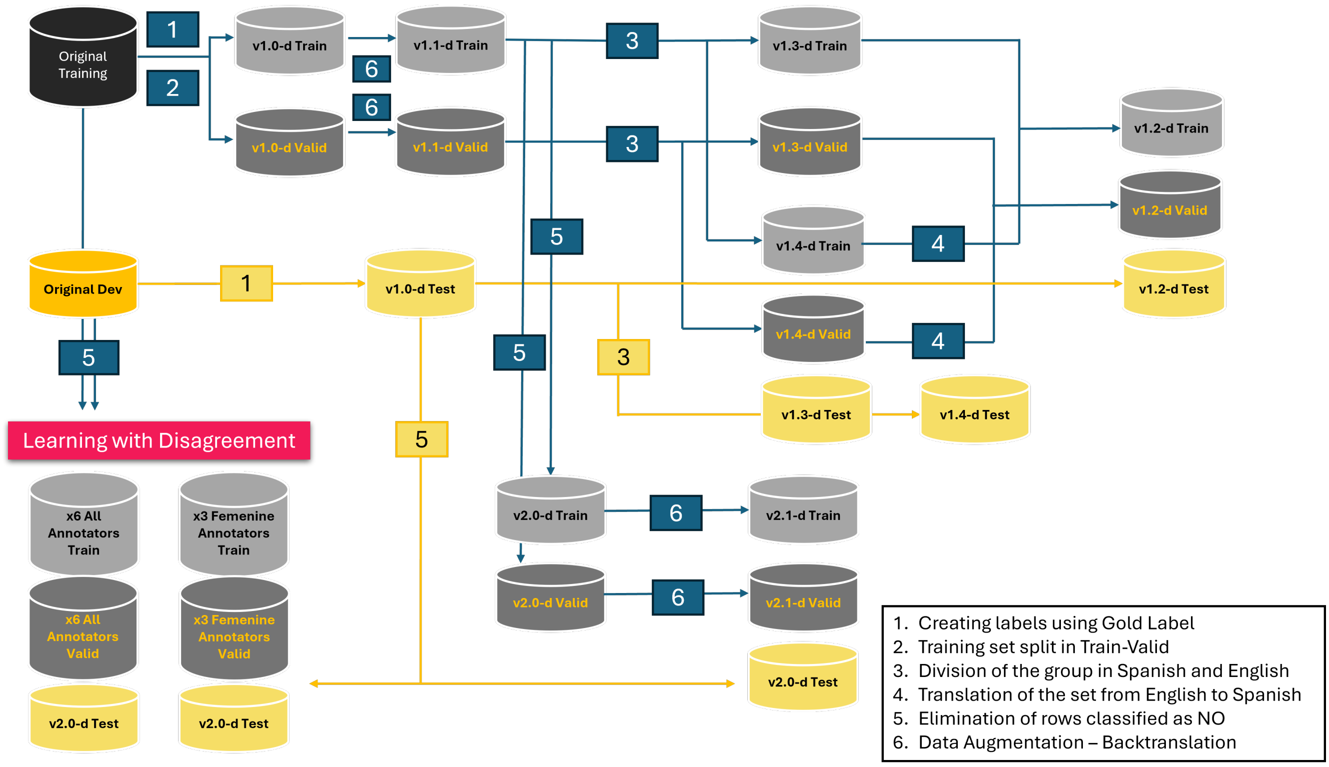

Figure 2: Datasets subdivisions for model training

4.3. Data Cleaning and Normalization

In the context of NLP, text preprocessing as data cleaning and normalization, are critical to ensuring texts are consistent and noise-free before use in machine learning models. Specifically, for cleaning tweets, the following techniques were employed:

- **Lowercase Conversion:** Ensures uniform treatment of words, eliminating the distinction between "Cat" and "cat", simplifying the dataset and reducing the number of unique features.
- **Removal of Links:** Eliminates web links present in tweets as they do not add semantic value and are often irrelevant to sentiment analysis or text meaning.
- **Removal of User Mentions:** Removes mentions of other users and retweets, which usually do not provide relevant information for semantic analysis and can introduce noise.
- **Removal of Hashtags:** Simplifies the text by removing hashtags, which may not be relevant for semantic analysis, focusing the analysis on words and phrases.
- **Removal of Emojis:** Although emojis convey emotions or contexts, their interpretation can be complex in textual analysis. Initial attempts to translate emojis into words did not improve results, thus they were removed to reduce noise and simplify analysis.

An example of the data cleaning carried out is presented in Table 6.

Table 6
Data Cleaning and Normalization

Original Tweet	Cleaned and Normalized Tweet
Collab betweet WeAreEqual X @TaravaNFT ? YOU ALREADY KNOW IT. Join our Discord on how to join our exclusive Giveaway : https://t.co/x3stzfLLmh. #NFT #NFTGiveaway #art	collab betweet weareequal x ? you already know it. join our discord on how to join our exclusive giveaway : .

4.4. Data Augmentation and Hyperparameter Search

Data augmentation is a crucial technique in natural language processing (NLP) to enhance the performance of machine learning models by artificially expanding the dataset. Various strategies, including back-translation, have been employed to improve model robustness and generalization. Recent studies have demonstrated the effectiveness of data augmentation in text classification tasks, emphasizing its importance in handling diverse linguistic patterns and enhancing model accuracy [11, 12]. Back-translation, in particular, has been highlighted as a powerful augmentation technique, transforming text into a target language and then translating it back to the source language to generate varied paraphrases while preserving the original meaning [13, 14, 15].

4.4.1. Oversampling with Backtranslation

Oversampling addresses class imbalance [16] by generating syntactic and lexical variations through backtranslation, increasing dataset diversity without altering meaning [17]. Since the datasets are unbalanced, it is necessary to employ a balancing technique. In this case, the number of rows for the REPORTED and JUDGEMENTAL classes has been increased through backtranslation, while the original number of rows has been maintained for the DIRECT class. Using Helsinki-NLP/opus models from the OPUS project [18], tweets in Spanish are translated to English, then German, and back to Spanish. An example of data generation through backtranslation for a tweet in Spanish is shown in Table 7.

Table 7
Example of data generation through backtranslation for a tweet in Spanish

Original Tweet	New Tweet Generated with Backtranslation
Se supone q me tengo q avergonzar d ser mamá? Jajajajaajajaja naaaa	¿Debería avergonzarme de ser madre?

For tweets in English, they were translated from English to German, then from German to Spanish, and finally from Spanish back to English. An example of a newly generated instance is shown in Table 8.

Table 8
Example of data generation through backtranslation for a tweet in English

Original Tweet	New Tweet Generated with Backtranslation
Easy to throw rocks and hide behind your gender or sexual identity #onhere	Easy to throw stones and hide behind your sex or sexual identity #onhere

4.4.2. Hyperparameter Search

Hyperparameter tuning optimizes model performance by selecting optimal values for non-learned parameters. Optuna [19] helps define and iteratively optimize the hyperparameter search space. Exhaustive search (grid search) explores all possible combinations but is computationally expensive. To expedite experiments, the training and validation datasets were reduced to 80% of the original size. To implement exhaustive search using Optuna, a hyperparameter search space was defined, as shown in Table 9.

Table 9
Hyperparameter Search Space

Hyperparameter	Value Range
Batch Size	[8, 16, 32]
Learning Rate	[3e-5, 5e-5]
Weight Decay	[0.001, 0.01, 0.1]

In reference to the metrics obtained after hyperparameter optimization and the application of the previously explained techniques, the results are explained in Tables 10 and 11.

Table 10
F1 scores Task 1

Model	Baseline	Data augmentation + Hyperparameters
XLM RoBERTa Base	0.7807	0.7876
Deberta v3 Base	0.7820	0.7871
Roberta Base Bne	0.7584	0.7616
Bert Base Multilingual	0.7618	0.7640

Table 11
F1 scores Task 2

Model	Baseline	Data augmentation + Hyperparameters
XLM RoBERTa Base	0.5945	0.6095
Roberta Base Bne	0.4795	0.4905
Deberta v3 Base	0.5801	0.5968

4.5. General Training Configuration

Training was conducted using the Trainer class from Hugging Face, incorporating optimized hyperparameters. The adamw _torch[20] optimizer was employed for updating model weights, with evaluations conducted at the end of each epoch and models saved periodically. The best model, determined by the F1 metric, was loaded. Training was halted using the EarlyStoppingCallback if no improvements were observed. These strategies were then tested on the structured dev dataset. The RTX 4070 graphics card was utilized for its high performance and capability to manage intensive processing tasks, thereby ensuring efficient and speedy development and execution of complex models.

4.5.1. Identifying Sexism in Tweets - Version: v1.x

To train the final models that will generate predictions on the test data provided by the competition for Task 1, we selected the two best-performing models based on their metrics during the training process.

To train the v1.1 model, data from the v1.1-d dataset was used. Each tweet in this dataset is labeled by six annotators in both the training and validation sets. To obtain the majority label, following the competition guidelines to obtain the gold label, the votes were averaged, selecting the labels that received two or more votes from among the six possible annotators. In case of a tie, the instance in question was completely excluded. A multilingual model, XLM-RoBERTa-Base, was trained to handle both English and Spanish instances simultaneously. Figure 3 shows this process.

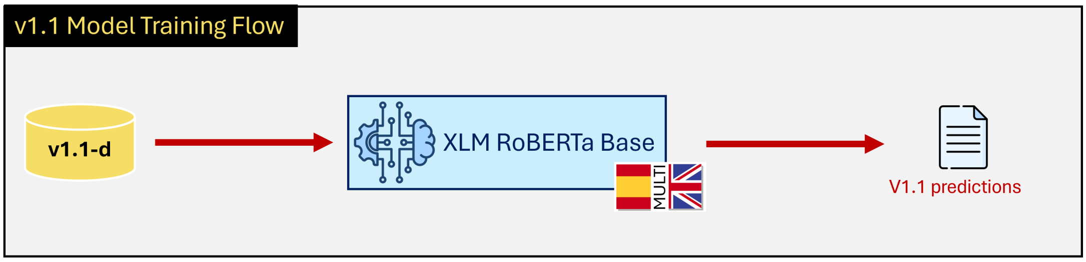

Figure 3: Training flow for model v1.1

Subsequently, this model was used to predict the labels of the data in the official competition test set. The results are presented indicating the majority predicted label for each instance of the test set, followed by the score_label, which represents the similarity score assigned by the classifier to the majority predicted label on a scale of 0 to 1.

To obtain the hard label, the majority predicted label was selected. Regarding the soft label, since it is a binary classifier (YES or NO), the score_label value was assigned to the majority class in each case, and the value of the minority class was calculated as 1 minus the score_label. It is important to note that the sum of the label values in the soft results should not exceed 1. The results model's evaluation is shown in the Table 12

Table 12
Version v1.1 Evaluation models' Results

Model	**F1 score**
XLM RoBERTa Base	0.850

The training process for model v1.2 is almost identical to the previously described process, but with some key differences regarding the models used and the workflow structure. For training, two datasets were used: v1.3-d for English instances and v1.4-d for Spanish instances. As for the models, DeBERTa v3 Base was used for English and RoBERTa Base Bne for Spanish. The figure 4 shows the process.

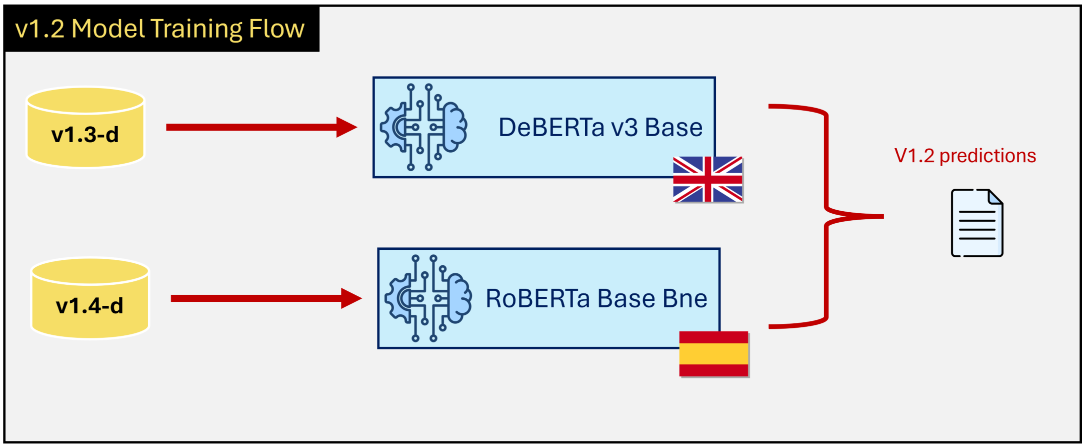

Figure 4: Version v1.2 Evaluation models' Results

The workflow began with the separate training of the two models: the DeBERTa v3 Base model was used for the English instances of dataset v1.3-d, and the RoBERTa Base Bne model was employed for the Spanish instances of dataset v1.4-d. The models' evaluation results are shown in Table 13

Table 13
Versions v1.3 and v1.4 Evaluation models' Results

Model	F1 score
XLM RoBERTa Base (v1.3)	0.854
DeBERTa v3 Base (v1.3)	0.859
XLM RoBERTa Base (v1.4)	0.826
RoBERTa Base Bne (v1.4)	0.863
BERT Base (v1.4)	0.818

Table 14
Version v1.2 - Predictions English + Spanish

Model	F1 score
DeBERTa v3 base	0.8589
RoBERTa Base Bne	0.8630
Final Average	0.8617

4.5.2. Intent Classification in Sexist Tweets - Model Versions

Model Version v2.1 was designed to address the second task of the competition, which focuses on classifying the intentionality of tweets previously categorized as sexist by model version v1.2 (Source Intention in Tweets). This task follows the initial classification of sexist messages and seeks to categorize such messages according to the author's intent, thus providing insights into the role of social media in issuing and spreading sexist messages. In this task, a classification between three classes DIRECT, REPORTED, and JUDGEMENTAL is proposed.

The training data comes from dataset version v2.1-d, containing only instances of the three classes, excluding instances categorized as NO, thus avoiding introducing noise in the training data and refining the model's accuracy. Only hard labels were generated for the final predictions, as the model does not return the score label of predicted classes as minority. The Figure 5 shows the process. Obtained results are shown in the Table 15

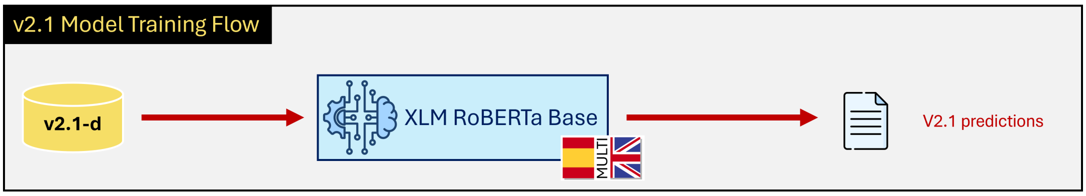

Figure 5: Training flow for Model Version 2.1

Table 15
Performance of Model Version 2.1

Model	F1 score
XLM RoBERTa Base	0.501

The next model applies Learning with Disagreement because it considers and leverages the differences in opinion among multiple human annotators when labeling the training data. This approach captures a greater diversity of perspectives, which is especially useful in subjective or complex tasks where there may be significant disagreement about the correct labels.

This method improves the model's predictions by integrating multiple viewpoints, creating a more robust and representative training dataset. Additionally, the soft labels resulting from this process enable the model to capture the uncertainty and variability inherent in human annotations, leading to better generalization and performance in real-world situations where data may not be clear or fully defined. The Figure 6 shows the process. Obtained results are shown in the Table 16.

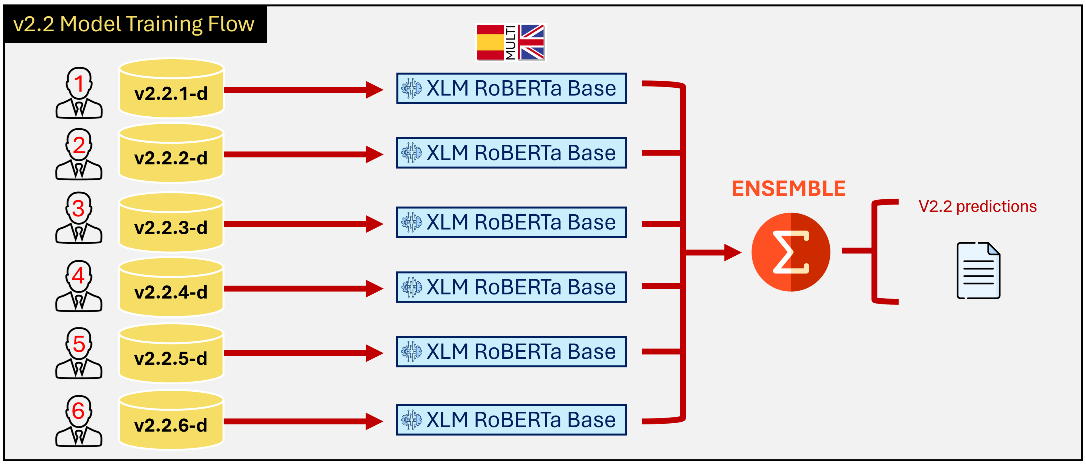

Figure 6: Training flow for Model Version 2.2

Table 16
Version 2.2 Evaluation models' Results

Model	F1 score
XLM RoBERTa [Ann_1]	0.576
XLM RoBERTa [Ann_2]	0.546
XLM RoBERTa [Ann_3]	0.509
XLM RoBERTa [Ann_4]	0.508
XLM RoBERTa [Ann_5]	0.517
XLM RoBERTa [Ann_6]	0.509
Ensembler	0.527

The training flow of the model shown in the image can be explained in detail, focusing on how the disagreement among annotators is handled and how soft labels are generated. Here's the step-by-step explanation:

1. Training data comes from six groups of annotators differentiated by gender and age: ["F 18-22", "F 23-45", "F 46+", "M 46+", "M 23-45", "M 18-22"]. Each group of annotators has provided labels for the training data.
2. Six datasets (v2.2.1-d, v2.2.2-d, v2.2.3-d, v2.2.4-d, v2.2.5-d, and v2.2.6-d) are used to train six instances of the XLM-RoBERTa Base model. Each dataset corresponds to the annotations of one of the six mentioned groups.
3. The six trained models are combined using an ensemble method. This process integrates the outputs of the different models to produce a more robust final prediction. The ensemble calculates a weighted average (sum) of the predictions of the six models.
4. To generate the soft labels, the proportion of annotators who voted for each label is taken into account. For example, if 2 out of 6 annotators labeled a data point as "DIRECT", the soft label for "DIRECT" would be 2/6 = 0.33333. This process is repeated for the other labels, "REPORTED" and "JUDGEMENTAL".

In the previous task (Task 1), the data was classified into the classes "YES" and "NO". If a data point was classified as "YES" with a probability of 0.80, this value is used to adjust the soft labels of Task

2. For example, if the soft label for "DIRECT" is 0.33333, the adjusted value would be 0.33333 * 0.80 = 0.26666. This adjustment is performed for all sub-classes of "YES" ("DIRECT", "REPORTED", and "JUDGEMENTAL").

This process must be done for the YES label when it is the majority class in Task 1, as well as to predict the percentage of this when it is the minority class in Task 1. In conclusion, the extremely low probability of the different YES classes in the instances that have been classified by the models of version 1 as NO is also being calculated.

Finally, Model Version 2.3 follows the same guidelines as Version 2.2, explained above, but the training data comes from three groups of annotators differentiated by gender and age: ["F 18-22", "F 23-45", "F 46"]. Only female groups have been selected to train the models that will compose the ensemble. The Figure 7 shows the process. Obtained results are shown in the Table 18.

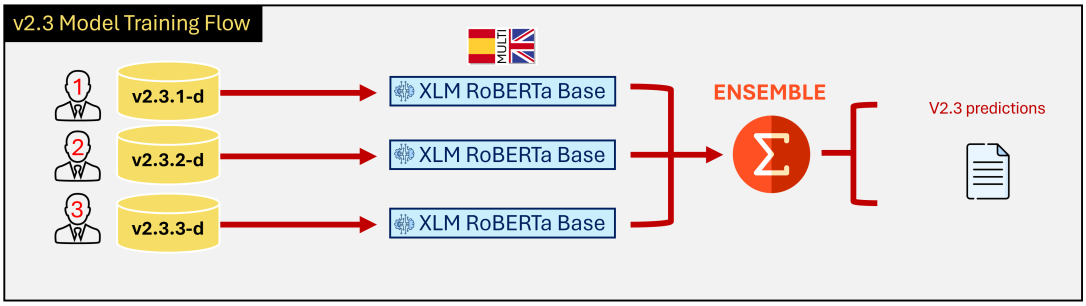

Figure 7: Training flow for Model Version 2.3

Table 17
Version 2.3 Evaluation models' Results

Model	F1 score
XLM RoBERTa [Ann_1]	0.5755
XLM RoBERTa [Ann_2]	0.5460
XLM RoBERTa [Ann_3]	0.5086
Ensembler	0.5434

4.6. Error Analysis

4.6.1. Task 1

This section provides a detailed analysis of errors made by the models in Task 1: Sexism Identification in Tweets, focusing on classification discrepancies between YES and NO classes. By scrutinizing misclassifications, patterns and insights into challenges faced by the models are aimed to be identified. Additionally, potential strategies to improve classification performance, especially for the minority class (YES), are explored. Examples are presented in Table 18.

Table 18
Examples of instances for Task 1

Tweet	Labels	Predictions
Woman driving beside me a few minutes ago holding her phone to her ear with her shoulder, while holding a mug of coffee. Baby on Boardsticker on both rear windows.	NO	YES
Por qué todos los hombres cuando su novia o esposa está embarazada andan más de culeros que de costumbre	YES	NO

4.6.2. Task 2

This section analyzes errors encountered by models in Task 2: Source Intention in Tweets, focusing on classification accuracy across DIRECT, REPORTED, and JUDGEMENTAL categories. Through examination of misclassifications, factors influencing performance across these categories are aimed to be understood, and refinements to improve the model's ability to discern nuanced intentions in sexist tweets are discussed. Examples are provided in Table 19, and confusion matrices in Figure 8 depict prediction distributions for Task 2 models.

Table 19
Examples of instances for Task 2

Tweet	Labels	Predictions
Lo irónico es que en su mayoría sean hombres quienes apoyan la criminalización de las mujeres frente al aborto. Claro, a las mujeres hay que castigarlas, juzgarlas y señalarlas siempre, como si no fuera suficiente tener que cargar con el peso de una violación.	REPORTED	DIRECT
En total delirio esta tipa quiere legalizar el terrorismo. ¿Y esta escoria quiere definir los destinos de Chile? Permitirlo es de anti chilenos.	DIRECT	NO
If you don't vote, you ARE the problem. #VoteBlueIn2022 #WomensRights #GunControl #bookban #CivilRights #VotingRights	NO	REPORTED

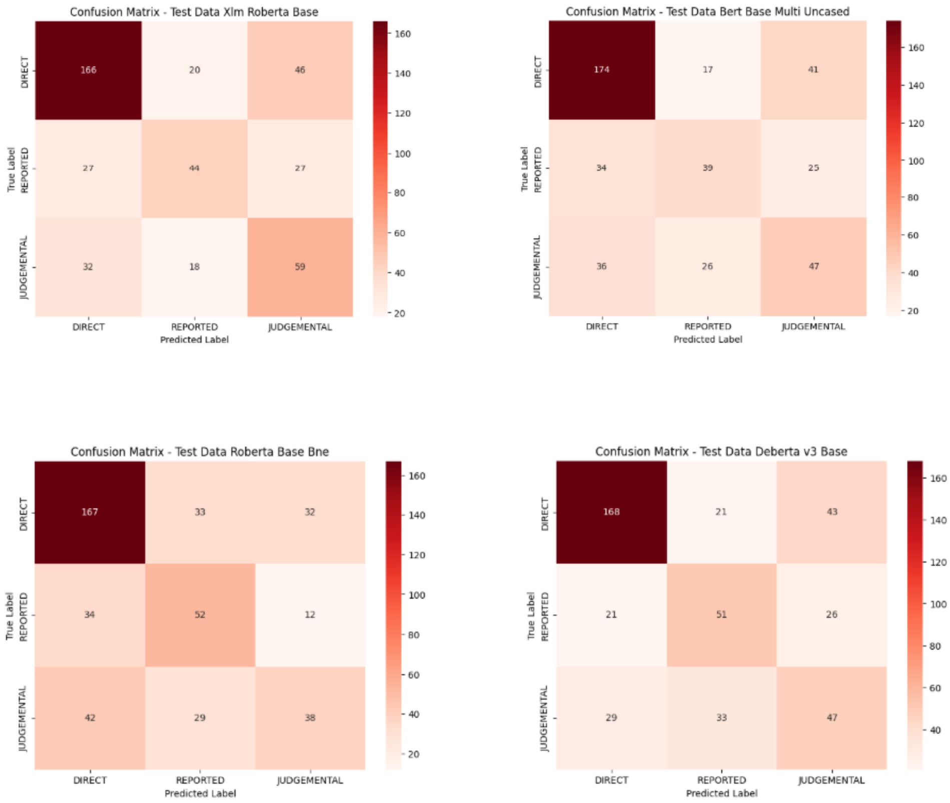

Figure 8: Confusion Matrices for Task 2 models test predictions

4.6.3. Error Analysis Conslusions

The analysis of errors in Task 1 and Task 2 uncovers various reasons for misclassifications. Many tweets feature nuanced language or context, challenging for models to interpret. For example, a tweet warning about sympathetic individuals may discuss predatory behavior broadly, misinterpreted by the model as sexist content. Tweets often employ sarcasm, idiomatic expressions, or ambiguous wording, leading to misclassification. A tweet about a woman multitasking while driving may be misconstrued as a gender stereotype critique rather than a comment on unsafe driving practices. Multilingual or culturally referential tweets add complexity. A Spanish tweet discussing men's behavior could be viewed contextually as commentary on male behavior patterns rather than explicit sexism.

5. Official Results

In Task 1, the best-performing strategy was a combination of models for different languages: RoBERTa Base BNE was used for classifying Spanish tweets, and DeBERTa v3 Base was employed for English tweets. This dual-model approach significantly outperformed other strategies, emphasizing the effectiveness of leveraging specialized models for each language. Following this, the multilingual model XLM RoBERTa Base also showed strong performance, though it was slightly behind the combined approach. In Task 1, Model v1.1 produced the run I2C-UHU_1, while v1.2 produced I2C-UHU_2. The official results for Task 1 are shown in the tables 20 and 21.

Table 20
HARD-HARD Evaluation EXIST 2024 Leaderboard Task1

Ranking	Run	ICM-Hard	ICM-Hard Norm	F1_YES
0	EXIST2024-test_gold.json	0.9948	1.0000	1.0000
-	-	-	-	-
10	**I2C-UHU_2.json**	0.5557	0.7793	0.7733
-	-	-	-	-
32	**I2C-UHU_1.json**	0.4651	0.7338	0.7513
-	-	-	-	-
68	EXIST2024-test_majority-class.json	-0.4413	0.2782	0.0000
-	-	-	-	-
70	EXIST2024-test_minority-class.json	-0.5742	0.2114	0.5698

Table 21
SOFT-SOFT Evaluation EXIST 2024 Leaderboard Task1

Ranking	Run	ICM-Soft	ICM-Soft Norm	Cross Entropy
0	EXIST2024-test_gold.json	3.1182	1.0000	0.5472
-	-	-	-	-
13	**I2C-UHU_2.json**	0.6871	0.6102	0.9184
-	-	-	-	-
18	**I2C-UHU_1.json**	0.5175	0.5830	1.0666
-	-	-	-	-
36	EXIST2024-test_majority-class.json	-2.3585	0.1218	4.6115
-	-	-	-	-
40	EXIST2024-test_minority-class.json	-3.0717	0.0075	5.3572

In Task 2, the best results were achieved using the Learning with Disagreement method with six groups of annotators (three male and three female). This approach outperformed the run that applied Learning with Disagreement with only three groups of female annotators. This finding suggests that having a more diverse set of annotators can enhance the model's performance by providing a broader range of perspectives, which likely leads to better generalization and robustness in the model's predictions.For Task 2, v2.1 generated the run I2C-UHU_1, v2.2 produced I2C-UHU_2, and v2.3 resulted in I2C-UHU_3. The official results for Task 2 are shown in the tables 22 and 23.

Table 22
HARD-HARD Evaluation EXIST 2024 Leaderboard Task2

Ranking	Run	ICM-Hard	ICM-Hard Norm	F1_YES
0	EXIST2024-test_gold.json	1.5378	1.0000	1.0000
-	-	-	-	-
11	**I2C-UHU_2.json**	0.1815	0.5590	0.4980
-	-	-	-	-
21	**I2C-UHU_1.json**	0.0418	0.5136	0.4708
-	-	-	-	-
24	**I2C-UHU_3.json**	0.0210	0.5068	0.4663
-	-	-	-	-
39	EXIST2024-test_majority-class.json	-0.9504	0.1910	0.1603
-	-	-	-	-
46	EXIST2024-test_minority-class.json	-3.1545	0.0000	0.0280

Table 23
SOFT-SOFT Evaluation EXIST 2024 Leaderboard Task2

Ranking	Run	ICM-Soft	ICM-Soft Norm	Cross Entropy
0	EXIST2024-test_gold.json	3.1182	1.0000	0.5472
-	-	-	-	-
17	**I2C-UHU_2.json**	-2.6952	0.2828	2.1440
-	-	-	-	-
22	**I2C-UHU_1.json**	-4.2278	0.1594	2.5245
-	-	-	-	-
27	EXIST2024-test_majority-class.json	-5.4460	0.0612	4.6233
-	-	-	-	-
35	EXIST2024-test_minority-class.json	-32.9552	0.0000	8.8517

6. Conclusions and Future Work

In this paper, the effectiveness of advanced transformer models in addressing the identification of sexism and the classification of source intent in social media texts has been demonstrated. The approach employed, which integrates Learning with Disagreement, facilitates the incorporation of diverse annotator perspectives, thereby enhancing the robustness and accuracy of the models. The methodology, consisting of classifying tweets as sexist or non-sexist and subsequently categorizing the intent of sexist tweets, has shown significant improvements in understanding and detecting nuanced sexist content. The results of the EXIST 2024 Leaderboard for Task 1 and Task 2 provide valuable insights into effective strategies for multilingual tweet classification and the impact of annotator diversity. For Task 1, superior performance was observed with the combination of language-specific models (RoBERTa Base BNE for Spanish and DeBERTa v3 Base for English), indicating the benefit of using specialized models tailored to individual languages. Meanwhile, Task 2 results indicated that Learning with Disagreement, utilizing a diverse set of annotators (both male and female), led to better outcomes compared to using only female annotators. This underscores the importance of diversity in annotation to capture a wider array of linguistic nuances and biases, thus improving the overall performance of the model. Future work will focus on refining the models by incorporating additional data sources and exploring more sophisticated ensemble methods. Additionally, efforts will be made to extend the research to other forms of harmful online content, applying the insights gained from this study to broader applications in social media moderation and policy-making. The insights derived from this research provide a valuable foundation for the development of more effective strategies to combat online sexism and other forms of digital harm.

Acknowledgments

This paper is part of the I+D+i Project titled *"Conspiracy Theories and hate speech online: Comparison of patterns in narratives and social networks about COVID-19, immigrants, refugees and LGBTI people [NON-CONSPIRA-HATE!]"*, PID2021-123983OB-I00, funded by *MCIN/AEI/10.13039/501100011033/* and by *"ERDF/EU"*.

References

[1] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[2] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, J. Artif. Int. Res. 72 (2022) 1385–1470. URL: https://doi.org/10.1613/jair.1.12752. doi:10.1613/jair.1.12752.

[3] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes (Extended Overview), in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[4] P. Burnap, M. L. Williams, Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making, Policy & Internet 7 (2015) 223–242. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.85. doi:https://doi.org/10.1002/poi3.85. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/poi3.85.

[5] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: J. Andreas, E. Choi, A. Lazaridou (Eds.), Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 88–93. URL: https://aclanthology.org/N16-2013. doi:10.18653/v1/N16-2013.

[6] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, arXiv preprint arXiv:1906.08237 (2019).

[7] J. He, Z. Gan, X. Liu, J. Li, J. Gao, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2021).

[8] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, M. Villegas, Spanish language models, 2021.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[10] T. Yu, H. Zhu, Hyper-parameter optimization: A review of algorithms and applications, 2020. arXiv:2003.05689.

[11] Author(s), A survey on data augmentation for text classification, Journal Name (2022).

[12] Author(s), Xlnet with data augmentation to profile cryptocurrency influencers, Journal Name (2023).

[13] Author(s), Backtranslate what you are saying and i will tell who you are, Journal Name (2024).

[14] Author(s), Data augmentation using back-translation for context-aware neural machine translation, Journal Name (2019).

[15] Author(s), Back-translation-style data augmentation for end-to-end asr, Journal Name (2018).

[16] S. Kobayashi, Contextual augmentation: Data augmentation by words with paradigmatic relations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, pp. 452–457.

[17] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A survey of data augmentation approaches for nlp, 2021. arXiv:2105.03075.

[18] M. Aulamo, J. Tiedemann, The OPUS resource repository: An open package for creating parallel corpora and machine translation services, in: M. Hartmann, B. Plank (Eds.), Proceedings of the 22nd Nordic Conference on Computational Linguistics, Linköping University Electronic Press, Turku, Finland, 2019, pp. 389–394. URL: https://aclanthology.org/W19-6146.

[19] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Aji, N. Bogoychev, A. Martins, A. Birch, Marian: Fast neural machine translation in c++, 2018, pp. 116–121. doi:10.18653/v1/P18-4020.

[20] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019. arXiv:1711.05101.