CLTL at EXIST 2025: Identifying Sexist Memes Using an Ensemble of Shallow and Transformer Models

Notebook for the EXIST Lab at CLEF 2025

Ariana Britez[1,*], Ilia Markov[1]

[1]*Computational Linguistics & Text Mining Lab (CLTL), Vrije Universiteit Amsterdam, The Netherlands*

Abstract

We present the CLTL system developed for identifying and categorizing sexist memes (Task 2) in English at the EXIST 2025 Shared Task at CLEF 2025. The task consisted of three subtasks: (Task 2.1) sexism identification, where memes were classified as *sexist* or *not-sexist*; (Task 2.2) source intention classification, where sexist memes were further classified as either *direct* or *judgemental*; and (Task 2.3) sexism categorization, where memes were classified into one or more overlapping fine-grained classes: *ideological and inequality*, *stereotyping and dominance*, *objectification*, *sexual violence*, and *misogyny and non-sexual violence*. Our submissions were based on a hard majority voting ensemble strategy, where the component models included a multimodal model that combined the representations of Swin Transformer V2 and a pre-trained language model (RoBERTa or BERT), and the text-only models that used meme text and image captions as input. The text-only approaches included pre-trained transformer models (RoBERTa, BERT, and a BERTweet model fine-tuned for sexism detection) and a conventional machine learning approach, namely an SVM with stylometric and emotion-based features. Our experiments demonstrated that an ensemble that incorporates deep learning and conventional machine learning approaches is efficient for the sexist meme detection task. Our best runs with an ICM-Hard score of 0.2850 for Task 2.1, -0.0645 for Task 2.2, and -0.4214 for Task 2.3, were ranked 6th out of 18 runs, 4th out of 15 runs, and 5th out of 14 runs, on the English leaderboard, respectively.

Keywords

Multimodal Sexism Detection, Ensemble Learning, Transformer Models, Conventional Machine Learning Approaches

1. Introduction

Online platforms such as social media and discussion forums provide users with tools to create and share a wide range of information. With the rise of user-generated content, harmful content also increases. Focusing on gender, the Pew Research Center [1] reported that women were twice as likely as men to have experienced gender-based online harassment. This can manifest in the form of *sexism*, which involves expressions intended to spread, incite, promote or justify hatred on the basis of sex, not only in comments or posts but also in multimodal memes. While memes are often used for humorous or ironic effects, they are also employed to spread violence and aggression against women [2]. The automatic detection of sexism is necessary given that its spread could amplify social misbehaviour by supporting and even inciting hate crimes [3], as well as contribute to sexual stereotyping and gender inequalities offline [4]. Furthermore, identifying the different forms of sexism and the intention of the author could help to recognize patterns in how it is manifested online [5].

This paper details the participation of the CLTL team in the tasks related to English memes (Task 2) under the hard evaluation setup at the sEXism Identification in Social neTworks (EXIST) lab [6, 7] organized by CLEF 2025. EXIST addresses the identification of sexism in a broad sense, ranging from explicit misogyny to more subtle, implicit forms of sexist behaviour. The 2025 edition of the shared task targets the identification and categorization of sexism in tweets (Task 1), memes (Task 2) and TikTok videos (Task 3) in both English and Spanish. Each Task consisted of the following three subtasks: (1) sexism identification, where each instance had to be labelled as sexist or not; (2) source

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

*Corresponding author.

a.j.britez@student.vu.nl (A. Britez); i.markov@vu.nl (I. Markov)

intention detection, where sexist tweets had to be categorized as direct, reported or judgemental, while sexist memes and TikTok videos as either direct or judgemental; and (3) sexism categorization, where each sexist instance had to be further classified into overlapping fine-grained classes: *ideological and inequality*, *stereotyping and dominance*, *objectification*, *sexual violence*, and *misogyny and non-sexual violence*.

We performed a variety of experiments for each subtask related to memes (Task 2) in English in the hard evaluation setup, combining different approaches in a hard majority voting ensemble. We explored a multimodal approach with a pre-trained language model and a vision model that has demonstrated state-of-the-art performance in detecting harmful content in memes [8, 9, 10] as well as fine-tuned pre-trained language models and conventional machine learning approaches. Our best-performing run for sexism identification ranked 6th on the English leaderboard, achieving an ICM-Hard score of 0.2850 on the test set. This approach incorporated a multimodal model: Swin Transformer V2 [11] with RoBERTa [12], BERTweet fine-tuned for sexism detection [13], and an SVM model with stylometric and emotion-based features [14]. For the source intention detection task, our best run included Swin Transformer V2 with RoBERTa, RoBERTa instead of BERTweet as a pre-trained language model, and SVM, achieving an ICM-Hard score of -0.0645 on the test set and ranking 4th on the English leaderboard. In the sexism categorization task, our best-performing run included deep learning models in the ensemble, namely the multimodal Swin Transformer V2 with RoBERTa, RoBERTa and BERT [15] in a hierarchical approach where memes that are predicted as sexist in step one are further classified into fine-grained classes in step two. This ensemble achieved an ICM-Hard score of -0.4214 and ranked 5th on the English leaderboard. Our results demonstrate that an ensemble combining deep learning and shallow approaches is effective for detecting harmful content in memes, as has previously been shown for textual harmful content [16, 14, 17, 18].

2. Task and Data

2.1. Task Description

While this edition of the EXIST Shared Task focuses on identifying and categorizing sexism in tweets, memes and TikTok videos, we participated in the task related to memes (Task 2), which consists of the following three subtasks:

2.1.1. Task 2.1: Sexism Identification

This task involved a binary classification task, where each meme was classified as sexist or not. The following descriptions were taken from the EXIST 2025 overview paper [19], EXIST 2025 Annotation Guidelines and EXIST 2024 overview paper [20].

- **Sexist**: The Oxford English Dictionary defines sexism as "prejudice, stereotyping or discrimination, typically against women, on the basis of sex". Sexism encompasses any form of oppression or prejudice against women due to their sex. This discrimination can stem from different beliefs, such as stereotypes, the belief that men are superior to women, or an irrational hatred of women, commonly referred to as misogyny. The latter represents a more extreme, hate-driven form of sexism. The meme can be sexist itself, describe a sexist situation or criticize a sexist behaviour.
- **Non-sexist**: The meme does not prejudice, underestimate, or discriminate against women.

2.1.2. Task 2.2: Source Intention

This task also consisted in a binary classification, where identified sexist memes were further classified according to the intention of its author into direct or judgemental. The description of each class is provided below:

- **Direct**: The author intends to spread a message that is sexist by itself.
- **Judgemental**: The author intends to condemn a sexist situation or behaviour.

Table 1
Statistics of the training and evaluation sets used for Task 2.1.

Label	Total	Train		Eval	
		# Num	%	# Num	%
Sexist	**958**	863	56.41	95	55.56
Non-sexist	**743**	667	43.59	76	44.44
Total	**1,701**	**1,530**	**100**	**171**	**100**

2.1.3. Task 2.3: Sexism Categorization

This task involved a multi-label classification, where each meme belonged to one or more of the following fine-grained classes: ideological and inequality; stereotyping and dominance; objectification; sexual violence; and misogyny and non-sexual violence. The description of each type of sexist memes is provided below:

- **Ideological and inequality**: Memes that discredit the feminist movement to devalue, belittle and defame the struggle of women in all areas. Also included are memes that reject inequality between men and women, or present men as victims of gender-based oppression.
- **Stereotyping and dominance**: Memes that express false ideas about women that suggest they are more suitable or inappropriate for certain tasks. Also included are any memes that imply that men are somehow superior to women.
- **Objectification**: Memes where women are presented as objects apart from their dignity and personal aspects. Also included are memes that assume or describe certain physical qualities that women must have to fulfil traditional gender roles.
- **Sexual violence**: Memes where sexual suggestions, requests or harassment of a sexual nature are made.
- **Misogyny and non-sexual violence**: Memes where expressions of hatred and violence towards women are contained.

2.2. Dataset

The EXIST Shared Task included tasks in both English and Spanish. The dataset used for Task 2 contained over 5,000 memes collected from Google Images, distributed across the two languages and annotated by six annotators. For the training set, there were 2,000 memes per language, and for the test set, 500 memes per language. The hard labels for each subtask were determined using a probabilistic threshold: for Task 2.1, the class annotated by more than three annotators was selected as the hard gold label; for Task 2.2, the threshold to keep the hard gold label was more than two annotators; and for Task 2.3, more than one annotator. As a result, the size of the dataset was reduced whenever these thresholds were not met. The dataset provided by the organizers included the memes as well as the text contained in them.

In our participation, we focused on the classification and categorization of memes in English. The test set consisted of 513 English memes. We split the training data using stratified sampling into 90% for training and 10% for evaluating our approaches. The statistics of the training and evaluation sets used in this work are presented in Tables 1, 2, and 3 for each subtask, respectively.

Our preprocessing steps included extracting the image captions from memes to implement our text-only models. This was done using the BLIP-2 vision-language model, specifically the version incorporating FlanT5$_{XL}$ fine-tuned on COCO [21]. The prompt used was "*ignore text on the image. a photo of*". However, since the resulting image caption still included text from the meme itself in 801 instances, either in addition to or instead of the description of the image, we removed the phrases that signalled the presence of meme text, such as "with the words", and "with the caption". After this process, 21 memes still required re-captioning, which was performed using the prompt *"a photo of"*.

Table 2
Statistics of the training and evaluation sets used for Task 2.2.

Label	Total	Train		Eval	
		# Num	%	# Num	%
Direct	**591**	525	69.44	65	75.58
Judgemental	**253**	231	30.56	21	24.42
Total	**844**	**756**	**100**	**86**	**100**

Table 3
Statistics of the training and evaluation sets used for Task 2.3.

Label	Total	Train # Num	Eval # Num
Ideological and inequality	**408**	369	39
Stereotyping and dominance	**480**	440	40
Objectification	**459**	416	43
Sexual violence	**213**	197	16
Misogyny and non-sexual violence	**180**	164	16
Total	**1,740**	**1,586**	**154**

Table 4
Meme, meme text and image caption from the BLIP-2 model.

Meme	Meme text	Image caption
	HATES IT WHEN YOU WIN HATES IT WHEN YOU LET HER WIN	a girl holding up two video game controllers

Table 5
Meme text and image caption combination for the text-only models.

Model	Text Representation
SVM	HATES IT WHEN YOU WIN HATES IT WHEN YOU LET HER WIN. a girl holding up two video game controllers
BERT	HATES IT WHEN YOU WIN HATES IT WHEN YOU LET HER WIN [SEP] a girl holding up two video game controllers
RoBERTa	HATES IT WHEN YOU WIN HATES IT WHEN YOU LET HER WIN </s> a girl holding up two video game controllers

The removal of phrases indicating the presence of meme text was repeated, resulting in the final dataset with image captions. An example of a meme taken from the training dataset (ID: 210568; pixelated for privacy reasons) and its image caption is provided in Table 4.

The meme text and image caption were combined, with the representation slightly changing for each model. Table 5 presents an example of how this was implemented for each model. The meme text and image caption were concatenated with a full stop for the SVM model, with the special token [SEP] for BERT and the special token </s> for RoBERTa. Before fine-tuning the RoBERTa model, the text was lowercased.

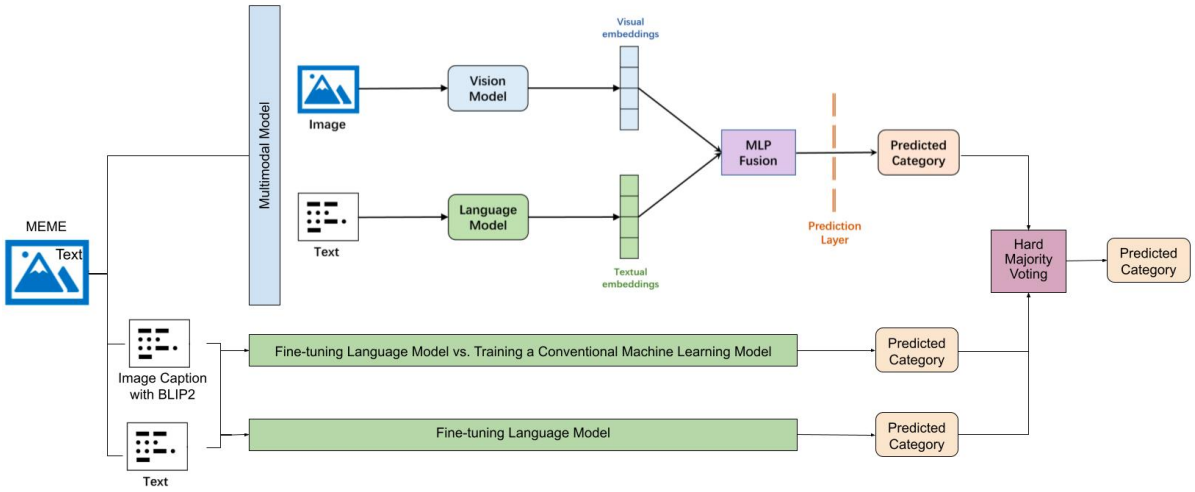

Figure 1: Overall architecture of our ensemble approach. The multimodal architecture (upper part of the figure) is based on Wang and Markov (2024) [10].

3. Methodology

Task 2.1 was framed as a binary classification problem. Memes labelled as *sexist* were further classified in Task 2.2 into *direct* or *judgemental*, as described in Section 2.1.2. For Task 2.3, we explored two different approaches: one using a hierarchical classification setup in which only memes identified as *sexist* in step one were further classified into fine-grained classes in step two, and another approach using a flat classification setup in which all instances were classified at once. Given that Task 2.3 is a multi-label classification task, in which each meme can be assigned to one or more categories, we implemented a binary relevance strategy. This approach decomposes the multi-label problem into independent binary classifiers for each class label [22].

For each subtask, we implemented a hard majority voting ensemble strategy that combined models trained and/or fine-tuned on both multimodal data and text-only data (meme text and image captions). Figure 1 provides an overview of our approach. While the component models were trained with either hierarchical or flat strategy for Task 2.3, the ensemble was implemented with a hierarchical approach.

Based on previous research [16, 14, 17, 18], which demonstrated promising results for detecting textual harmful content using an ensemble that combined deep learning and conventional machine learning approaches, we were interested in exploring whether this approach would also generalize to a multimodal setup. However, given that higher results were achieved during evaluation using a combination of multimodal and pre-trained models for some of the subtasks in Task 2, each run included an ensemble approach that combined different component models that showed the best results on the development set.

3.1. Component Models

3.1.1. Multimodal Approaches

Our multimodal approach incorporates the vision model **Swin Transformer V2** [11] with one of the following pre-trained language models: **BERT** [15] or **RoBERTa** [12]. Swin Transformer V2 builds upon the original Swin Transformer [23], a general-purpose vision model with a hierarchical architecture which representations are computed using shifted windows. It builds a hierarchical representation by starting with small image patches and gradually merging neighbouring patches in deeper layers. In this design, self-attention is computed locally within non-overlapping windows that partition the image. Swin Transformer V2 can process high-resolution images and employs a self-supervised pre-training method to reduce reliance on large amounts of labelled data [11].

In the multimodal approach, Swin Transformer V2 was used to extract visual features while the pre-trained language model was used for extracting contextualized textual embeddings. The resulting visual and textual representations were concatenated and passed through a Multilayer Perceptron (MLP) fusion module [24], followed by a prediction layer to classify each instance (Figure 1, top). This system was selected because it has shown state-of-the-art performance for the classification of harmful memes in English [8], as well as in other languages, such as Arabic [9] and Spanish [10].

3.1.2. Text-only Approaches

The text-only models implemented included: pre-trained **BERT** [15], **RoBERTa** [12], and a **BERTweet** fine-tuned for sexism identification [13], and a conventional **SVM** approach trained with stylometric and emotion-based features [14]. The encoder models were widely applied in the detection of sexism and misogyny [20, 25]. In our implementation, these models were fine-tuned for each subtask using the textual representation of memes (i.e., meme text and image captions).

BERTweet is a large-scale pre-trained language model for English Tweets. While it has the same architecture as BERT-base, this model was trained using the pre-training procedure from RoBERTa [26]. We specifically use *BERTweet-large-sexism-detector*[1] fine-tuned for sexism detection [13] using the Explainable Detection of Online Sexism (EDOS) dataset [27].

Among supervised machine learning algorithms for text classification, SVM have been widely used for the detection of hateful content. The SVM approach applied has proven to effectively reduce the false positive rate when combined with transformer models [17] and has also demonstrated strong performance for non-English languages [18]. We implemented an SVM with stylometric and emotion-based features [17], which incorporates part-of-speech tags, function words, and emotion-conveying words and their associations from the the NRC emotion lexicon [28]. These features were vectorized with a *tf-idf* weighting scheme. Unigrams were extracted for the POS, function word and emotion word features, while bigrams were used for the emotion association features. We also extracted character n-grams (with n=3–6) from the meme text and image captions using a *tf* weighting scheme and combined them with the aforementioned features. The model was built with the liblinear implementation of SVM from scikit-learn, with the regularization parameter (C) optimized through a grid search.

Our experiments with deep learning models were conducted on the Google Colaboratory platform with an NVIDIA A100 GPU. For the multimodal model, we used the PyTorch framework along with the AutoGluon library. The experiments were performed with consistent hyperparameters: a base learning rate of 1e-4, a decay rate of 0.9 using cosine decay scheduling, a batch size of 8, a maximum of 10 training epochs, and optimization via the AdamW optimizer. For the text-only pre-trained models, the Transformers library was used and hyperparameter optimization was performed with Optuna, specifically for the batch size, learning rate, weight decay, and number of epochs. After optimizing the parameters on the evaluation set, our models were trained on the entire original training dataset (training + development) before making the final predictions on the test set.

3.2. Ensembles per Subtask

We describe the runs submitted for each subtask below, all of which combine the predictions of the component models in a **hard majority voting ensemble** strategy.

3.2.1. Task 2.1: Sexism Identification

- **Run 1**: (1) Swin Transformer V2 with RoBERTa, (2) RoBERTa, (3) SVM.
- **Run 2**: (1) Swin Transformer V2 with RoBERTa, (2) BERT, (3) SVM.
- **Run 3**: (1) Swin Transformer V2 with RoBERTa, (2) BERTweet, (3) SVM.

[1] https://huggingface.co/NLP-LTU/bertweet-large-sexism-detector

Table 6
Results for Task 2.1 on the evaluation and test sets in English.

Set	Run	Ranking	ICM-Hard	ICM-Hard Norm	F1 YES
	Baseline (Swin Transformer V2+RoBERTa)		0.2198	0.6109	0.7416
Eval	1	n/a	0.1395	0.5704	0.75
	2		0.1787	0.5902	0.7553
	3		**0.2439**	**0.6230**	**0.7835**
	1	7	0.2282	0.6159	0.7363
Test	2	10	0.1714	0.5870	0.7377
	3	**6**	**0.2850**	**0.6447**	**0.7611**

3.2.2. Task 2.2: Source Intention

Memes were first classified into sexist or not using the model that showed the best results for binary classification on the evaluation set (run 3, Table 6). The memes identified as sexist were further classified into direct or judgemental in step two using the following models within the majority voting ensemble:

- **Run 1**: (1) Swin Transformer V2 with RoBERTa, (2) BERTweet, (3) SVM.
- **Run 2**: (1) Swin Transformer V2 with RoBERTa, (2) RoBERTa, (3) SVM.
- **Run 3**: (1) Swin Transformer V2 with RoBERTa, (2) Swin Transformer V2 with BERT, (3) BERTweet.

3.2.3. Task 2.3: Sexism Categorization

The component models in runs 1 and 3 were trained with a flat approach. In contrast, the component models in run 2 were trained using a hierarchical approach, where memes were first classified into sexist or not using our best-performing model during evaluation (run 3, Table 6). Then, the memes identified as sexist were further classified into the types of sexism with the models listed below combined in an ensemble:

- **Run 1:** (1) Swin Transformer V2 with RoBERTa, (2) Swin Transformer V2 with BERT, (3) RoBERTa.
- **Run 2**: (1) Swin Transformer V2 with RoBERTa, (2) RoBERTa, (3) BERT.
- **Run 3**: (1) Swin Transformer V2 with RoBERTa, (2) RoBERTa, (3) BERTweet.

4. Results

We submitted the ensemble models that showed the best results on the development set to be evaluated on the official test set employed in the shared task. The ICM (Information Contrast Measure) metric was used as the official evaluation metric for all subtasks. A normalized version of ICM (ICM Norm) was also reported, along with the F1 score of the positive class for Task 2.1 and the macro F1 score for the remaining tasks. We also implemented a multimodal baseline model for each subtask, Swin Transformer V2 with RoBERTa, and evaluated it on the evaluation set.

The results for Task 2.1 are presented in Table 6. The run that combined Swin Transformer V2 with RoBERTa, BERTweet and SVM (run 3) achieved the best performance on both the evaluation and test sets. It ranked 6th among all submitted runs for English, with an ICM-Hard score of 0.2850. We observe an improvement in performance from 0.2439 to 0.2850 in terms of ICM-Hard metric on the test set, but a drop in F1 of the positive class from 0.7835 to 0.7611. The difference among the three submitted runs was the pre-trained language model incorporated in the ensemble as a text-only approach. BERTweet fine-tuned for sexism detection performed best for this task, followed by RoBERTa, which achieved an ICM-Hard score of 0.2282 and was placed 7th in the ranking. Yet, performance dropped when BERT was used, yielding an ICM-Hard score of 0.1714 and ranking 10th among the

Table 7
Results for Task 2.2 on the evaluation and test sets in English.

Set	Run	Ranking	ICM-Hard	ICM-Hard Norm	Macro F1
	Baseline (Swin Transformer V2+RoBERTa)		-0.1153	0.4595	**0.5681**
Eval	**1**	n/a	**-0.0601**	**0.4789**	0.5320
	2		-0.0777	0.4727	0.5138
	3		-0.0978	0.4656	0.5371
	1	6	-0.0975	0.4662	0.4854
Test	**2**	**4**	**-0.0645**	**0.4776**	**0.4870**
	3	5	-0.0771	0.4732	0.4991

Table 8
Results for Task 2.3 on the evaluation and test sets in English.

Set	Run	Ranking	ICM-Hard	ICM-Hard Norm	Macro F1
	Baseline (Swin Transformer V2+RoBERTa)		**-0.2517**	**0.4463**	**0.5688**
Eval	1	n/a	-0.4012	0.4144	0.5099
	2		-0.4492	0.4042	0.5266
	3		-0.3310	0.4294	0.5112
	1	8	-0.7884	0.3325	0.3973
Test	**2**	**5**	**-0.4214**	**0.4105**	**0.4724**
	3	7	-0.6379	0.3645	0.4404

participating submissions. Given that our multimodal baseline model scored an ICM-Hard of 0.2198, the results obtained by the ensembles showed an improvement in terms of the ICM-Hard metric. This indicates that combining a multimodal model with a pre-trained language model and a conventional machine learning approach that captures stylometric and emotion-based features in an ensemble is beneficial for identifying sexism in multimodal content. This finding aligns with what has already been demonstrated for textual content. Nonetheless, the choice of the pre-trained language model incorporated into the ensemble can substantially affect its performance.

Table 7 shows the results for Task 2.2. In contrast to the results for Task 2.1, the ensemble model that achieved the best performance during evaluation (run 1), with an ICM-Hard score of -0.0601, was not the best-performing run on the test set. Run 1, which combined the same models that had performed best in Task 2.1–namely Swin Transformer V2 with RoBERTa, BERTweet and SVM–showed lower results for source intention identification on the test set. Instead, it was run 2, which used RoBERTa instead of BERTweet as pre-trained language model, that achieved the best performance among our submitted runs in English, ranking 4th with an ICM-Hard score of -0.0645. Run 3 followed with an ICM-Hard score of 0.0771, while run 1 ranked 6th with an ICM-Hard score of -0.0975. While our runs did not lead the ranking for the English task, all of them outperformed the winning approach in English in the source identification task at EXIST 2024, where the best-performing system achieved an ICM-Hard score of -0.1691 [20], compared to -0.0645 from our best run. Furthermore, all ensembles outperformed the baseline on the evaluation set in terms of the ICM metrics. However, the multimodal baseline achieved a higher macro F1 score than the ensembles on the evaluation set, with a macro F1 of 0.5681. The results in terms of ICM metrics highlight that combining an ensemble of deep learning and shallow approaches is effective for source intention identification.

The results for Task 2.3 are showed in Table 8. While our best performing ensemble during evaluation (run 3) included models trained with a flat approach, incorporating the multimodal Swin Transformer V2 with RoBERTa, RoBERTa and BERTweet, achieving an ICM-Hard score of -0.3310, this approach showed a drop in performance on the test set. Our best performing run (run 2) for sexism categorization was the ensemble trained with a hierarchical approach, obtaining an ICM-Hard score of -0.4214 on the test

set and ranking 5th among the English submissions. The ensembles with a flat approach in runs 1 and 3 resulted in an ICM-Hard score of -0.7884 and -0.6379 on the test set and ranked 8th and 7th in the English learderboard, respectively. These results show that a hierarchical approach improves the performance over the flat approach. However, including a second multimodal model (Swin Transformer V2 with BERT) in the ensemble yielded the worst performance among our runs, highlighting the importance of incorporating the text-only models into the ensemble in our remaining runs, which are likely to capture different information than the multimodal models. Given that ensembles improved the performance in Tasks 2.1 and 2.3, we expected that they would also generalize to this task. However, the ensembles in this task did not outperform the multimodal baseline, which scored an ICM-Hard score of -0.2517 on the evaluation set. Moreover, the results in this subtask reflect the complexity of multi-label classification considering the imbalanced distribution of the fine-grained classes in the training data. We note that while our best-performing run ranked 5th in the English leaderboard, it did outperform the winning approach in English in the sexism categorization task at EXIST 2024: -0.4214 vs. -0.5626 [20] in terms of ICM-Hard score.

5. Conclusion

In this work, we presented the CLTL system developed for sexism identification, source intention detection, and sexism categorization in Task 2, focusing on English memes, at the EXIST 2025 Shared Task. We evaluated different combinations of component models in a hard majority voting ensemble, including a multimodal model (Swin Transformer V2 paired with RoBERTa or BERT), pre-trained language models such as BERT, RoBERTa and a BERTweet fine-tuned for sexism detection, as well as an SVM incorporating stylometric and emotion-based features. Our results showed that an ensemble combining deep learning and shallow approaches is beneficial for both sexism identification and source intention classification in memes. Our best-performing run for sexism identification included an ensemble that grouped the predictions from Swin Transformer V2 with RoBERTa, BERTweet, and SVM, achieving an ICM-Hard score of 0.2850 and ranking 6th on the English leaderboard. The best-performing run for source intention detection combined Swin Transformer V2 with RoBERTa, RoBERTa and SVM, resulting in an ICM-Hard score of -0.0645 and ranking 4th on the English leaderboard. The runs for the sexism categorization task incorporated deep learning models in the ensemble, with the best-performing run including BERT, RoBERTa and Swin Transformer V2 with RoBERTa in a hierarchical setup. This approach achieved an ICM-Hard score of -0.4214 on the test set and ranked 5th on the English leaderboard. Our best-performing runs for source intention detection and sexism categorization outperformed the winning approaches to the same subtasks in English at EXIST 2024. One limitation of our approach is that the text-only models in the ensembles depend on the output of the image captioning tool used during preprocessing, which quality can significantly impact the overall performance.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 for grammar and spelling checks. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Duggan, Online Harassment 2017, 2017. URL: https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2017/07/PI_2017.07.11_Online-Harassment_FINAL.pdf.

[2] M. Paciello, F. D'Errico, G. Saleri, E. Lamponi, Online sexist meme and its effects on moral and emotional processes in social media, Computers in Human Behavior 116 (2021) 106655. doi:https://doi.org/10.1016/j.chb.2020.106655.

[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. doi:10.3233/JIFS-179023.

[4] E. Fersini, G. Rizzi, A. Saibene, F. Gasparini, Misogynous MEME Recognition: A Preliminary Study, in: AIxIA 2021 – Advances in Artificial Intelligence: 20th International Conference of the Italian Association for Artificial Intelligence, Virtual Event, December 1–3, 2021, Revised Selected Papers, Springer-Verlag, Berlin, Heidelberg, 2021, p. 279–293. doi:10.1007/978-3-031-08421-8_19.

[5] L. Plaza, J. Carrillo-de Albornoz, E. Amigó, J. Gonzalo, R. Morante, P. Rosso, D. Spina, B. Chulvi, A. Maeso, V. Ruiz, EXIST 2024: sEXism Identification in Social neTworks and Memes, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, 2024, pp. 498–504. doi:10.1007/978-3-031-56069-9_68.

[6] L. Plaza, J. Carrillo-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. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco 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.

[7] L. Plaza, J. Carrillo-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.

[8] Y. Wang, I. Markov, CLTL@Multimodal Hate Speech Event Detection 2024: The Winning Approach to Detecting Multimodal Hate Speech and Its Targets, in: A. Hürriyetoğlu, H. Tanev, S. Thapa, G. Uludoğan (Eds.), Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 73–78. URL: https://aclanthology.org/2024.case-1.9/.

[9] Y. Wang, I. Markov, CLTL at ArAIEval Shared Task: Multimodal Propagandistic Memes Classification Using Transformer Models, in: N. Habash, H. Bouamor, R. Eskander, N. Tomeh, I. Abu Farha, A. Abdelali, S. Touileb, I. Hamed, Y. Onaizan, B. Alhafni, W. Antoun, S. Khalifa, H. Haddad, I. Zitouni, B. AlKhamissi, R. Almatham, K. Mrini (Eds.), Proceedings of the Second Arabic Natural Language Processing Conference, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 501–506. URL: https://aclanthology.org/2024.arabicnlp-1.51/. doi:10.18653/v1/2024.arabicnlp-1.51.

[10] Y. Wang, I. Markov, CLTL at DIMEMEX Shared Task: Fine-Grained Detection of Hate Speech in Memes, in: S. M. Jiménez-Zafra, L. Chiruzzo, F. Rangel, F. Balouchzahi, U. B. Corrêa, A. Bonet Jover, H. Gómez-Adorno, J. A. González Barba, D. I. Hernández Farías, A. Montejo Ráez, P. Moral, C. Rodríguez Abellán, M. E. Vallecillo Rodríguez, M. Taulé, R. Valencia-García (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), 2024.

[11] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al., Swin Transformer v2: Scaling Up Capacity and Resolution, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12009–12019. doi:10.1109/CVPR52688.2022.01170.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A Robustly Optimized BERT Pretraining Approach, arXiv preprint arXiv:1907.11692 (2019).

[13] S. Al-Azzawi, G. Kovács, F. Nilsson, T. Adewumi, M. Liwicki, NLP-LTU at SemEval-2023 task 10: The Impact of Data Augmentation and Semi-Supervised Learning Techniques on Text Classification Performance on an Imbalanced Dataset, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto,

Canada, 2023, pp. 1421–1427. URL: https://aclanthology.org/2023.semeval-1.196/. doi:10.18653/v1/2023.semeval-1.196.

[14] I. Markov, N. Ljubešić, D. Fišer, W. Daelemans, Exploring Stylometric and Emotion-Based Features for Multilingual Cross-Domain Hate Speech Detection, in: O. De Clercq, A. Balahur, J. Sedoc, V. Barriere, S. Tafreshi, S. Buechel, V. Hoste (Eds.), 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/.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[16] B. van Aken, J. Risch, R. Krestel, A. Löser, Challenges for Toxic Comment Classification: An In-Depth Error Analysis, in: D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, J. Wernimont (Eds.), Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 33–42. URL: http://aclweb.org/anthology/W18-5105. doi:10.18653/v1/W18-5105.

[17] I. Markov, W. Daelemans, Improving Cross-Domain Hate Speech Detection by Reducing the False Positive Rate, in: A. Feldman, G. Da San Martino, C. Leberknight, P. Nakov (Eds.), Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, Association for Computational Linguistics, Online, 2021, pp. 17–22. URL: https://aclanthology.org/2021.nlp4if-1.3/. doi:10.18653/v1/2021.nlp4if-1.3.

[18] I. Markov, I. Gevers, W. Daelemans, An Ensemble Approach for Dutch Cross-Domain Hate Speech Detection, in: Natural Language Processing and Information Systems: 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15–17, 2022, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2022, pp. 3–15. doi:10.1007/978-3-031-08473-7_1.

[19] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, 2025, pp. 442–449. doi:10.1007/978-3-031-88720-8_65.

[20] 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 Tweets and Memes (Extended Overview), in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. Di Nunzio, L. Soulier, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction: 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, Proceedings, Part II, 2024, p. 93–117. doi:10.1007/978-3-031-71908-0_5.

[21] J. Li, D. Li, S. Savarese, S. Hoi, BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 19730–19742. URL: https://proceedings.mlr.press/v202/li23q.html.

[22] M.-L. Zhang, Y.-K. Li, X.-Y. Liu, X. Geng, Binary relevance for multi-label learning: an overview, Frontiers of Computer Science 12 (2018) 191–202. doi:10.1007/s11704-017-7031-7.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9992–10002. doi:10.1109/ICCV48922.2021.00986.

[24] X. Shi, J. Mueller, N. Erickson, M. Li, A. Smola, Multimodal AutoML on Structured Tables with Text Fields, in: 8th ICML Workshop on Automated Machine Learning (AutoML), 2021. URL: https://openreview.net/forum?id=OHAIVOOl7Vl.

[25] 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: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), 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.

[26] D. Q. Nguyen, T. Vu, A. Tuan Nguyen, BERTweet: A pre-trained language model for English tweets, in: Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 9–14. URL: https://aclanthology.org/2020.emnlp-demos.2/. doi:10.18653/v1/2020.emnlp-demos.2.

[27] H. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 Task 10: Explainable Detection of Online Sexism, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2193–2210. URL: https://aclanthology.org/2023.semeval-1.305/. doi:10.18653/v1/2023.semeval-1.305.

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