NLP Assignment 1

Francesco Baiocchi, Christian Di Buò and **Leonardo Petrilli**
Master's Degree in Artificial Intelligence, University of Bologna
{francesco.baiocchi2, christian.dibuo, leonardo.petrilli}@studio.unibo.it

Abstract

This work addresses the EXIST 2023 Task 1 on sexism detection, focusing on classifying tweets as sexist or not sexist. We employ two approaches: a custom Bidirectional LSTM model tailored for this task and a pretrained RoBERTa model. Our analysis identifies key error patterns, such as biases toward specific terms like 'woman' and the impact of UNK and OOV tokens on misclassification. These tokens are a key reason why the Transformer outperforms the BiLSTM, as its pretrained embeddings and contextual understanding enable it to better handle such cases.

1 Introduction

Online sexism can inflict harm on people, especially women, who are often targeted. Automatically detecting whether a sentence contains sexist references is an increasingly important task with applications in multiple fields, such as law and social media. Each year, the CLEF conference organizes the EXIST challenge, which focuses on sexism identification in social networks. The first task of this challenge involves a binary classification for sexism detection. Numerous approaches to this problem have been proposed, many of these rely on Transformer-based fine-tuned architectures such as RoBERTa, which has achieved impressive results (Vallecillo-Rodríguez et al., 2023).
We employed a standard NLP pipeline, starting with an initial phase where the data is loaded and analysed, followed by a preprocessing step, then training and finally an evaluation of three different models. We trained and tested each model using three seeds to identify the best performing architecture, whose test performance was then analysed and compared with a pre-trained model. The models were trained on a smaller version of the EXIST 2023 challenge dataset, consisting of tweets annotated by six different people as either sexist or not sexist.

2 System description

For the purpose of this task we implemented a pipeline involving data preprocessing, tokenization, model definition, training, evaluation, and error analysis. The preprocessing phase consisted of the following steps:

- Elimination of all Spanish tweets.
- Transformation of the task from a soft binary classification to a hard one by removing tweets which have an equal number of "yes" and "no" annotations.
- Elimination of hashtags, user mentions, URLs, emojis and special characters.
- Lemmatization of all the words to their base form.

After preprocessing, we created an embedding matrix initialized using pre-trained GloVe embeddings trained on Wikipedia and Gigaword 5 datasets (Pennington et al., 2014). OOV (Out-Of-Vocabulary) and UNK (unknown) tokens were mapped to random embeddings. We then implemented two models:

- **Baseline**: one bidirectional LSTM layer and a final dense layer.
- **Model1**: same as the baseline model, with one additional bidirectional LSTM layer.

Both models were trained and validated using the macro F1-score as the evaluation metric. Training was repeated with three different seeds and the model with the highest average score was selected. Finally, we imported and fine-tuned a pre-trained transformer model, **twitter-roberta-base-hate**, developed by Cardiff NLP and trained on ~58 million tweets (Group, 2021). The input tweets were cleaned using the same pipeline, excluding lemmatization and tokenization, as the model uses its own

tokenizer. This model was fine-tuned on our training data and compared with the best custom model on the test set.

3 Experimental setup and results

For the custom models the choice of hyperparameters had a minor effect on performance, with the best set being lstm_size = 64, batch_size = 32, initial_lr = $1e^{-3}$, weight_decay = $1e^{-2}$. Moreover, to prevent overfitting, we set early stopping with a patience of 10 epochs.

For the transformer fine-tuning, we used the following hyperparameter settings: learning_rate = $2e^{-5}$, batch_size = 8, weight_decay = $1e^{-2}$, and an early stopping with patience of 3 epochs. We report in Table 1 the macro F1 scores on the validation set for the custom models across three seeds. Additionally, Table 2 presents the metrics of the selected Custom Model and the Transformer on the test set.

Model	Seed 1	Seed 2	Seed 3
Baseline	0.804	0.842	0.793
Model 1	0.800	0.829	0.785

Table 1: Validation set macro F1 scores across three seeds.

Model	F1 score	Precision	Recall
Custom Model	0.748	0.755	0.745
RoBERTa Base	0.859	0.858	0.861

Table 2: Test set macro metrics.

4 Discussion

The results presented in Table 2 demonstrate a significant performance difference between the Custom Model and the Transformer. The RoBERTa Base model outperforms the Custom Model across all metrics, achieving a macro F1 score of 0.859 compared to 0.748 for the Custom Model.

Concerning the error analysis, we started at a high level by examining sentence length distributions for misclassified sentences and found no significant correlation with errors, indicating that sentence length does not strongly affect predictions. Since there appear to be no clear error patterns at the sentence level, we delved into a more fine-grained analysis: a word-level analysis. This approach aims to help us understand whether certain types of words systematically affect the model's predictions. For the Custom Model, we analyzed the presence of UNK and OOV tokens in sentences, distinguishing between all sentences and those incorrectly classified. By studying the correlation between these tokens and errors, we observed the following:

- **UNK tokens**: These appear frequently in both total and misclassified sentences, but their distribution does not show a significant correlation with misclassifications.
- **OOV tokens**: Although relatively rare overall, a significant number are found in misclassified sentences regarding sexism. These terms are closely related to sexism and play an important role in the task. (e.g *'gangbanged'*, *'phallocentrism'*, *'manspreading'*, and *'mansplaining'*)

For both models the distribution of key terms were analyzed, revealing a strong correlation between the word 'woman' and errors for non sexist sentences. This pattern reflects a dataset bias associating 'woman' with sexism even when the context is neutral or unrelated. By computing correlations, we validated the observed patterns: 'woman' is positively correlated with errors in label 0 (non sexist) and negatively correlated with label 1 (sexist), while OOV terms show a weak but notable correlation with errors in label 1.

Possible solutions to mitigate this bias include using sample weights to balance the importance of key terms during training, refining the dataset to provide more diverse and balanced contextual examples of terms like 'woman' and leveraging advanced embedding strategies, like contextual embeddings or subword tokenization, to better capture the semantics of OOV terms and reduce their impact on misclassifications.

5 Conclusion

In conclusion, the results align with our expectations, as RoBERTa outperformed the Custom Model due to its larger number of parameters, pretraining on a significantly larger corpus, and fine-tuning on the same task. This highlights the advantages of leveraging pre-trained transformer architectures for complex tasks like sexism detection. Additionally, our analysis underscored the importance of identifying and addressing biases, such as

the overassociation of 'woman' with sexism, which can lead to systematic misclassifications. Recognizing and mitigating such biases is crucial for improving model fairness and generalization in sensitive applications.

References

Cardiff NLP Group. 2021. twitter-roberta-base-hate. https://huggingface.co/cardiffnlp/twitter-roberta-base-hate. [Online; accessed 2-Jan-2025].

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

María Estrella Vallecillo-Rodríguez, Flor Miriam Plaza del Arco, Luis Alfonso Ureña-López, María Teresa Martín-Valdivia, and Arturo Montejo-Ráez. 2023. Integrating Annotator Information in Transformer Fine-tuning for Sexism Detection. [Online; accessed 8-Jan-2025].