Assignment 1

Nicolas Cridlig, Roberto Giordano, Federico Marchi, and **Alessio Pittiglio**
Master's Degree in Artificial Intelligence, University of Bologna
{ nicolasivan.cridlig, roberto.giordano2, federico.marchi9, alessio.pittiglio }@studio.unibo.it

Abstract

This study explores the application of Transformer and LSTM architectures for NLP, focusing on a binary classification problem for detecting sexism in text. Through comprehensive data cleaning and preprocessing techniques, we prepare the dataset to ensure robust model evaluation. The analysis is documented in a step-by-step manner to serve as a practical guide for researchers and practitioners. Our findings highlight the superior performance of Transformer-based models compared to LSTMs, albeit with increased computational costs during training and inference.

1 Introduction

Classification in NLP is often addressed using rule-based systems, in-context learning, or supervised machine learning. Rule-based systems require expert input and are costly to maintain, while in-context learning with LLMs is computationally expensive and lacks fine-tuning. Supervised methods like LSTMs and Transformers excel at contextual understanding but vary in performance and computational cost.

Our dataset is a subset of EXIST 2023 Task 1 on sexism detection (exi, 2023). It consists of tweets, noisy and informal text often laden with extraneous elements that interfere with text analysis. Therefore the natural language goes through regex-based preprocessing to remove noise and tokenized using GloVe embeddings.

> **Original:** *@RMatthewsPsyEdu The women will all be at home cooking for the family.*
> **Preprocessed:** *woman home cooking family*
> **Tokenized:** *2 145 2509 175 0 ... 0*

This eliminates noise while retaining semantic information in a dense format, as can be observed in the embedding space where similar concepts lie close to one another.

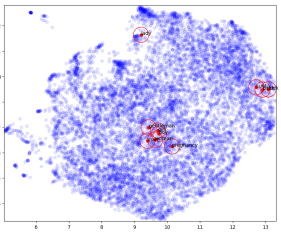

Figure 1: Embedding Space

Sequences are padded with zeros to ensure uniform length, and padding is excluded from the training loss calculation.

2 System description

Our contribution to the field is a pipeline called **Bird Bath** capable of processing 1.500 tweets per second. It is built on the excellent NLTK and Regex libraries.

The dataset is then fed into GloVe pretrained embeddings (Pennington et al., 2014) where we experimented with dimensions of [50, 100, 200, 300]. We utilize Keras to train two custom LSTM architectures with a frozen embedding layer. All out-of-vocabulary (OOV) words terms are randomly initialized to vectors with the same standard deviation as GloVe.

The Transformer model is based on the pre-trained variant of RoBERTa model, it's then fine-tuned for sexism speech detection, specifically designed for Twitter data. This model utilizes the tokenizer from the same pre-trained model to process input text. It is configured for binary classification. The model's architecture leverages the powerful RoBERTa transformer (Face, 2023), which is capable of handling long-range dependencies in text.

3 Experimental setup and results

The Baseline model is a shallow LSTM trained with binary cross entropy loss. It embeds input sequences using pre-trained embeddings of dimen-

sion 100, which are kept frozen. The model then passes the sequence through a bidirectional LSTM layer with 64 units, followed by a dense layer with a sigmoid activation for the classification. It has 1,112,109 parameters of which 84,609 are trainable.

The Model1 model is a deeper LSTM with same embedding as baseline which uses the same embedding layer. The model then passes the sequence through two bidirectional LSTM layers, again with 64 units. The first LSTM layer processes the entire sequence and outputs the full hidden states, while the second layer only outputs the final hidden state. This deeper architecture allows the model to capture more intricate patterns in the text. It concludes with a dense layer with a sigmoid activation for the binary classification task. It has 1,210,925 parameters of which 183,425 are trainable. Both the LSTM-based models used Adam as optimizer.

The Transformer model follows the architecture of RoBERTa model from Hugging Face (Face, 2023). We pad the textual data to 512 tokens to respect the prior length of the model. We use AdamW as optimizer. All models are trained with early stopping and adaptive learning rate. Dropout is used on the LSTM and weight decay on the Transforer to prevent overfitting. Batch size was set to 8 for all the models.

The table below presents the results achieved by the best-performing LSTM-based architecture, Model1, alongside those of the Transformer model(Cridlig, 2025).

	LSTM	**Transformer**
Accuracy	0.7517	0.8042
F1 score	0.7492	0.8028

Table 1: Best Performance on Test Set

4 Discussion

The inferior performance of the LSTM model with GloVe embeddings compared to the pretrained RoBERTa model for sexism classification in tweets can be attributed to several key factors. First, contextual embeddings provide a significant advantage; GloVe embeddings are static, and this means that each word has its own fixed meaning regardless the specific context in the tweets, whereas RoBERTa's contextual embeddings dynamically adjust based on the sentence, allowing it to capture nuanced meanings. Moreover the architecture differences play a crucial role. LSTMs process sequences step by step, which can limit their ability to capture long-range dependencies or global relationships. In contrast, RoBERTa's transformer architecture thanks to self-attention mechanisms, analyzes all parts of a sequence simultaneously, leading to richer sentiment analysis. RoBERTa also benefits from extensive pre-training, leveraging masked language modeling over large, diverse corpora to acquire a deep understanding of syntax, semantics, and domain-specific patterns in social media context. Moreover, we notice the LSTM performance is degraded on detecting sexist tweets, which is the minority class, while RoBERTa handles the data imbalance.

While Model1 achieved 5% less accuracy than RoBERTa, it does so at a reduced computational cost. To fine tune the transformer takes 1 hour for 5 epochs while the LSTM from scratch trains in under 5 minutes, on a Nvidia P100 GPU provided by Kaggle.

Both models outperformed the baseline, demonstrating that training and fine-tuning contributed to a better understanding of the meaning of the tweets. However, both models exhibit signs of overfitting, as indicated by the noticeable discrepancies in accuracy and macro F1-score across the training, validation, and test sets. The presence of OOV words in the test set may be one of the factors contributing to this overfitting. To improve model performance, data augmentation could be used increase data diversity, as well as hyperparameter tuning techniques, such as learning rate warm-up.

5 Conclusion

In this study, we compared LSTM and Transformer models for sexism detection in tweets. The Transformer model, a fine-tuned RoBERTa, demonstrated superior performance due to its ability to capture contextual relationships in text. Despite this, both models exhibit signs of overfitting, as evidenced by greater performance discrepancy across training and validation the longer training continues. To address this, future work could focus on data augmentation, hyperparameter tuning, and different architectures to enhance generalization. This study highlights the importance of model selection NLP tasks and suggests that Transformer-based models are particularly well-suited for capturing the complex, context-dependent nature of social media text, albeit at a higher computational cost

than LSTMs.

References

2023. Exist 2023 task: Explainable detection of sexism in social media. Accessed: 2025-01-07.

Marchi Pittiglio Cridlig, Giordano. 2025. Model weights. Accessed: 2025-01-13.

Hugging Face. 2023. Roberta model documentation. Accessed: 2025-01-13.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. https://nlp.stanford.edu/projects/glove/. Accessed: 2025-01-13.