# NLP Assignment 1: Sexism Detection in Tweet

**Nadia Farokhpay - 0001111417**

**Master's Degree in Artificial Intelligence, University of Bologna**

**nadia.farokhpay@studio.unibo.it**

## Abstract

This report details a comparative study of BiLSTM with GloVe embeddings and a fine-tuned RoBERTa transformer (cardiffnlp/twitter-roberta-base-hate) for sexism detection in tweets using the EXIST 2023 Task 1 dataset. After rigorous preprocessing (lowercasing, removal of URLs/mentions/hashtags/emojis, lemmatization with SpaCy), the RoBERTa model achieved a superior average macro F1-score of 0.8680 (±0.003) versus the BiLSTM's 0.830 (±0.006). Challenges in detecting sarcasm and implicit bias were noted, suggesting future work in data augmentation and hybrid models.

## 1 Introduction

Automated sexism detection is crucial for safer online social media. This project classifies tweets from the EXIST 2023 English dataset as sexist or non-sexist. Key contributions include:

1. A BiLSTM baseline
2. A fine-tuned RoBERTa model

## 2 System Description

### 2.1 Dataset & Preprocessing :

The English subset of EXIST 2023 (6,000 tweets) was used. Preprocessing involved: lowercasing, removing URLs, mentions, hashtags, and emojis, and lemmatization (SpaCy). GloVe embeddings (glove-wiki-gigaword-100) for BiLSTM covered 81.85% of the vocabulary; OOV words were randomly initialized. Class imbalance (60.4% non-sexist, 39.6% sexist) was handled by class weighting for RoBERTa.

### 2.2 Model Architectures :

- BiLSTM (Baseline): GloVe embeddings (100d, trainable) fed into a BiLSTM (64 units/direction), then a dense sigmoid layer. Approx. 530K parameters.
- RoBERTa (Transformer): cardiffnlp/twitter-roberta-base-hate (pre-trained on hate speech, 125M parameters) was fine-tuned using the Hugging Face transformers library.

### 2.3 Training Setup:

A stratified 80-20 train-validation split was used. Macro F1-score was the primary metric.

- BiLSTM: Adam optimizer, LR 0.001, batch size 32, 10 epochs. 3 runs (seeds 42, 43, 44).
- RoBERTa: AdamW optimizer, LR 1e-5, batch size 16, 5 epochs, class weighting. Early stopping (patience 2, delta 0.01 on validation macro F1). 3 runs (seeds 42, 43, 44).

## 3 Results & Evaluation

The performance of Phi-2 and Llama 3.1 across zero-shot, 2-shot, and 4-shot configurations is summarized below:

| Model | Avg. Macro | Std. Dev | Avg. Precisi | Avg. Recal |
|---|---|---|---|---|
| BiLSTM | 0.830 | ±0.006 | 0.842 | 0.819 |
| RoBERTa (fine-tuned) | 0.868 | ±0.003 | 0.881 | 0.855 |

RoBERTa outperformed BiLSTM by ~3.8 F1 points, showing better stability. Its pretraining on relevant data likely aided contextual understanding.

## 4 Error Analysis

A detailed examination of model misclassifications reveals several recurring patterns and model-specific challenges.

### BiLSTM Error Patterns

- **False Negatives ( Sexist misclassified as non-sexist ):**

Approximately 20% of BiLSTM's misclassifications were false negatives. These tweets typically involved implicit sexism or subtle bias that lacked overt indicators. For instance: "Another day, another comment about what I should wear..." While sexist in tone, such tweets lack explicit keywords, making them harder for BiLSTM to detect with static embeddings.

- **False Positives ( Non-sexist misclassified as sexist ):**

About 15% of errors stemmed from the model being overly sensitive to gendered language, regardless of context. For example: "Men and women are different, and that's okay." Here, the presence of gendered terms likely triggered the model despite the absence of hostile or sexist intent.

- **Ambiguity ( 5% ):**

Some tweets contained ambiguous or context-dependent expressions that even human annotators may find unclear. These included jokes, rhetorical questions, or culturally nuanced statements.

**RoBERTa Error Patterns**

While the fine-tuned RoBERTa model significantly reduced total error rates, its misclassifications also followed identifiable trends:

- **Sarcasm and Irony:**

Tweets involving sarcasm posed challenges, as understanding them often requires external knowledge or tone interpretation, e.g., "Oh sure, I just LOVE being told how emotional I am."

- **High Context Dependence:**

Tweets referencing previous interactions or societal norms without direct context often led to errors. Even with contextual embeddings, RoBERTa struggled to resolve these.

Overall, RoBERTa demonstrated better handling of implicit cues compared to BiLSTM, likely due to its pretraining on Twitter data and ability to model word dependencies in context. However, both models would benefit from targeted strategies for sarcasm detection, context resolution, and interpretability.

## 5 Conclusion

This study successfully demonstrated the effectiveness of deep learning models for sexism detection in tweets, using the EXIST 2023 Task 1 dataset. The comparative analysis established the superior performance of the fine-tuned RoBERTa transformer (cardiffnlp/twitter-roberta-base-hate) over a BiLSTM baseline with GloVe embeddings.

RoBERTa achieved an average macro F1-score of 0.868, surpassing the BiLSTM's 0.830. This improvement reflects RoBERTa's enhanced ability to capture nuanced contextual information in social media text, benefiting from its pretraining on large Twitter corpora that include hate speech and subtle biases.

While overall performance was strong, error analysis revealed persistent challenges in detecting sarcasm, irony, and context-dependent expressions. Nonetheless, the RoBERTa model demonstrated greater robustness and accuracy, highlighting the value of transformer-based approaches for moderating harmful content and supporting safer online environments.

## 6 References

1. Devlin, J. et al. (2019). BERT.arXiv:1810.04805
2. Pennington, J. et al. (2014). GloVe.EMNLP
3. Liu, Y. et al. (2019). RoBERTa.arXiv:1907.11692
4. SpaCy. https://spacy.io
5. Hugging Face Transformers. https://huggingface.co/transformers/