Sexism Detection in Social Media
A Comparative Study of BiLSTM and Transformer-based Models

Ivo Rambaldi
University of Bologna
ivo.rambaldi@studio.unibo.it

Abstract

This report presents a text classification study on the detection of sexist content in social media posts. We compare a recurrent neural baseline based on a BiLSTM architecture with a transformer-based model fine-tuned from a pretrained RoBERTa checkpoint. The task is framed as a multi-class classification problem distinguishing non-sexist content from three types of sexism: DIRECT, JUDGEMENTAL, and REPORTED. We evaluate the models using standard classification metrics and provide an error analysis highlighting the main challenges of the task, particularly class imbalance and subtle linguistic phenomena.

1 Introduction

Online social platforms are a major channel for public discourse but also a frequent source of abusive and discriminatory language. Automatically identifying sexist content is therefore an important problem, both from a societal and a technical perspective. Compared to coarse binary hate-speech detection, fine-grained sexism classification introduces additional difficulty due to subtle distinctions between explicit, judgemental, and reported forms of sexism.

In this project, we address sexism detection as a supervised multi-class classification task. We investigate how different modeling choices affect performance, comparing a traditional neural architecture based on static word embeddings with a modern transformer-based model leveraging contextualized representations.

2 Dataset and Task Definition

The dataset consists of short social media posts from the EXIST 2023 shared task, annotated with four labels: non-sexist (‘-’), DIRECT sexism, JUDGEMENTAL sexism, and REPORTED sexism. The task is to assign exactly one label to each post.

We apply majority vote aggregation across annotators and filter English-only content, yielding 2,873 training samples, 150 validation samples, and 280 test samples. A key characteristic of the dataset is strong class imbalance: the non-sexist class represents 70.1% of samples (2,014), while DIRECT accounts for 18.7% (537), REPORTED for 6.4% (184), and JUDGEMENTAL for only 4.8% (138). This imbalance has a direct impact on both training dynamics and evaluation, particularly for macro-averaged metrics.

Text preprocessing includes removal of URLs, mentions, hashtags, and emojis, followed by spaCy lemmatization. This results in a training vocabulary of 9,073 unique tokens.

3 Models

3.1 BiLSTM Baseline

As a baseline, we use a bidirectional LSTM model operating on pretrained GloVe Twitter embeddings (100-dimensional, 27B token corpus). The architecture consists of an embedding layer, a BiLSTM encoder with 128 hidden units per direction, dropout (0.2), and a linear classification head. Out-of-vocabulary tokens (11.8% of training vocabulary) receive random Gaussian initialization. The model contains 1.14M trainable parameters and is trained with Adam optimizer, sparse categorical cross-entropy loss, and early stopping (patience=3) on validation loss.

While simple, this model provides a useful reference point and highlights the limitations of static word representations in handling noisy and context-dependent social media language.

3.2 Stacked BiLSTM

We also test a stacked variant with two BiLSTM layers (128 and 64 hidden units per direction) and increased dropout (0.3) to investigate whether additional depth improves performance.

3.3 Transformer-based Model

The main model is based on Twitter-RoBERTa-base-hate, a RoBERTa model pretrained on Twitter data and adapted for hate speech detection. A classification head is added on top of the [CLS] token representation, and the entire model is fine-tuned end-to-end for 3 epochs with learning rate 2e-5, batch size 16, and weight decay 0.01. Thanks to subword tokenization and contextualized representations, this model is better suited to handle lexical variation, short texts, and implicit meanings.

4 Experimental Setup

Models are trained using cross-entropy loss and evaluated on held-out validation and test sets. We report precision, recall, and F1-score, with particular emphasis on macro-F1 to account for class imbalance. All models are trained across three random seeds (42, 1337, 2025), and we report mean and standard deviation. The best-performing model is selected based on validation macro-F1.

5 Results

Table 1 shows validation performance across all models.

Model	Precision	Recall	F1
BiLSTM Baseline	0.496 ± 0.016	0.411 ± 0.028	0.424 ± 0.027
Stacked BiLSTM	0.495 ± 0.085	0.447 ± 0.056	0.440 ± 0.025
Twitter-RoBERTa	$\mathbf{0.688 \pm 0.067}$	$\mathbf{0.537 \pm 0.013}$	$\mathbf{0.549 \pm 0.004}$

Table 1: Macro-averaged validation metrics across three seeds. Bold indicates best performance.

The BiLSTM baseline achieves limited performance, with macro-F1 around 0.42, showing strong bias toward the majority class. The stacked BiLSTM shows only marginal improvement (0.440 F1) with higher variance, suggesting that additional recurrent depth provides limited benefit without sufficient training data.

The RoBERTa model consistently outperforms both baselines, reaching macro-F1 of 0.549 on validation, a 29.5% relative improvement over the BiLSTM baseline. Nevertheless, performance remains uneven across labels, with especially low recall for minority classes.

Table 2 presents per-class test results for the best RoBERTa model (seed 1337).

Class	Support	Precision	Recall	F1
Non-sexist (-)	196	0.78	0.91	0.84
DIRECT	52	0.67	0.60	0.63
JUDGEMENTAL	14	0.33	0.14	0.20
REPORTED	18	0.64	0.39	0.48
Macro avg	280	0.61	0.49	0.49
Weighted avg	280	0.74	0.77	0.74

Table 2: Per-class test set performance for Twitter-RoBERTa (seed 1337).

The gap between weighted F1 (0.74) and macro F1 (0.49) quantifies the impact of class imbalance. Non-sexist content achieves strong performance (F1=0.84), while JUDGEMENTAL performs poorly (F1=0.20) due to extreme underrepresentation and subtle linguistic characteristics.

6 Error Analysis

The error patterns are heavily influenced by class imbalance, with the non-sexist class acting as a strong attractor. Confusion matrix analysis reveals that 40% of DIRECT samples, 61% of REPORTED samples, and 86% of JUDGEMENTAL samples are misclassified as non-sexist.

DIRECT sexism is generally detected when explicit insults are present, while JUDGEMENTAL and REPORTED sexism are frequently confused with other or misclassified as non-sexist. Reported speech requires discourse-level understanding to distinguish reporting from endorsing sexism, which remains challenging even for transformer-based models. For example, posts like "the man sat next to me has just told his friend that he doesn't employ women because they're not committed enough" require understanding that the speaker reports rather than endorses the sexist view.

Additional errors arise from sarcasm, implicit stereotypes, and noisy language typical of social media. JUDGEMENTAL samples often express sexist attitudes through implications rather than explicit statements, making them difficult to separate from opinionated but non-sexist content.

The BiLSTM baseline's 11.8% out-of-vocabulary rate further limits performance, while the transformer's subword tokenization handles novel terms through character-level composition.

Overall, errors mostly stem from subtle linguistic phenomena and severe class imbalance rather

than from a failure to recognize overtly abusive language.

7 Conclusion

This project shows that transformer-based models substantially outperform recurrent baselines for fine-grained sexism detection, achieving 29.5% relative improvement in macro-F1. The benefits stem from contextualized representations, subword tokenization, and domain-adapted pretraining.

However, the task remains challenging due to dataset imbalance and the inherently subtle nature of some sexist categories. Future improvements are likely to come from better class balancing strategies (oversampling, class-weighted loss, focal loss) and targeted modeling of implicit and reported sexism, rather than purely increasing model complexity. Hierarchical classification or multi-task learning with related objectives may also help improve minority class performance.