

Assignment 1

NLP Course Project

Jacopo Francesco Amoretti, Roberto Frabetti, Ivo Rambaldi

Master's Degree in Artificial Intelligence, University of Bologna

{ jacopo.amoretti, roberto.frabetti, ivo.rambaldi }@studio.unibo.it

Abstract

This work studies automatic sexism detection in social media posts as a multi-class classification task, distinguishing non-sexist content from three forms of sexism: DIRECT, JUDGEMENTAL, and REPORTED. We compare BiLSTM-based neural baselines using pre-trained GloVe embeddings with a transformer-based model fine-tuned from Twitter-RoBERTa. Experiments conducted across multiple random seeds show that the transformer model substantially outperforms recurrent architectures in terms of macro-F1. However, performance is limited on minority classes due to imbalance and subtle sexism.

1 Introduction

Note: as per instructions, this being an assignment report, the problem description and importance has been skipped

Standard approaches to this problem include neural models based on recurrent architectures and pre-trained word embeddings, as well as more recent transformer-based models relying on contextualized representations.

In this work, we compare these two paradigms by evaluating BiLSTM-based classifiers using pretrained GloVe Twitter embeddings against a transformer-based model fine-tuned from Twitter-RoBERTa.

Experiments are conducted across three random seeds and evaluated using macro-averaged precision, recall, and F1. Results show that the transformer-based model consistently outperforms recurrent baselines, though performance on minority classes remains limited due to class imbalance and subtle linguistic phenomena.

2 Background

Sexism detection is commonly studied as a subtask of abusive language and hate speech detection, but fine-grained categorization introduces additional complexity.

From a modeling perspective, social media language poses specific challenges such as informal syntax, and lexical variation, and frequent use of context-dependent expressions. Models based on static word embeddings typically struggle to represent such variability, whereas transformer-based models leverage contextualized representations and subword tokenization.

3 System description

We implement and evaluate three text classification systems: a BiLSTM baseline, a stacked BiLSTM variant, and a transformer-based classifier. All models share a common preprocessing, training, and evaluation pipeline to ensure fair comparison.

We implement BiLSTM baseline/stacked (GloVe embeddings, dropout, softmax) and fine-tune Twitter-RoBERTa-base-hate, replacing its head for 4 classes (using random initialization, keeping the other weights intact).

External libraries (TensorFlow/Keras and Hugging Face Transformers) are used for standard components, with adaptations limited to task-specific configuration and integration.

4 Data

Each post is labeled into one of four classes: non-sexist, DIRECT sexism, JUDGEMENTAL sexism, and REPORTED sexism. The original annotations include multiple annotators per instance and content in multiple languages.

In order to get a single label, majority-vote was applied. We then retain only English-language content. The resulting dataset contains 2,873 training posts, 150 validation posts, and 280 test posts. The class distribution is highly imbalanced: non-sexist content accounts for 2,014 instances, DIRECT sexism for 537 , REPORTED sexism for 184, and JUDGEMENTAL sexism for 138. This imbalance strongly influences both training dynamics and evaluation.

Text preprocessing includes normalization, removal of non-text elements (URLs, mentions, emojis), and lemmatization, followed by whitespace normalization. For the neural baselines, texts are further lemmatized using spaCy to reduce sparsity. A vocabulary of 9,073 unique tokens is built from the training split only. Pretrained GloVe Twitter embeddings are used for word representation, with out-of-vocabulary tokens initialized randomly.

5 Experimental setup and results

Models are evaluated on the validation and test dataset using macro-averaged precision, recall, and F1 to address class imbalance. LSTM models are trained with Adam and early stopping (patience = 3), Twitter-RoBERTa is fine-tuned for 3 epochs. All experiments are run with three random seeds; we report mean and standard deviation on validation.

Model	Precision
BiLSTM Baseline	0.444 ± 0.046
Stacked BiLSTM	0.437 ± 0.049
Twitter-RoBERTa	0.688 ± 0.067
Model	Recall
BiLSTM Baseline	0.391 ± 0.052
Stacked BiLSTM	0.420 ± 0.076
Twitter-RoBERTa	0.537 ± 0.004
Model	F1
BiLSTM Baseline	0.399 ± 0.056
Stacked BiLSTM	0.409 ± 0.054
Twitter-RoBERTa	0.482 ± 0.022

6 Discussion

The quantitative results confirm a clear advantage of the transformer-based model over recurrent baselines, but also highlight persistent weaknesses on minority classes (more in notebook). On validation, Twitter-RoBERTa reaches a macro-F1 of 0.482, corresponding to roughly a 20% relative improvement over the BiLSTM baseline (macro-F1 0.399) and a smaller but consistent gain over the stacked BiLSTM (macro-F1 0.409). The stacked model improves measures compared to the baseline but (at the cost of higher variance across seeds in recall), suggesting that deeper recurrence adds limited robustness given the dataset size. do not modify

The error analysis shows that most mistakes are

driven by severe class imbalance and subtle linguistic phenomena rather than by an inability to detect overt abuse. DIRECT sexism is usually captured when clear insults or slurs appear, whereas JUDGEMENTAL and REPORTED sexism are often confused with each other or collapsed into the non-sexist class. Additional errors come from sarcasm, implicit stereotypes, and noisy Twitter-style language. The BiLSTM is further penalized by the OOVs relative to GloVe, while RoBERTa’s subword tokenization mitigates but does not eliminate sparsity issues.

7 Conclusion

This work investigated fine-grained sexism detection in social media as a multi-class classification problem on the EXIST 2023 dataset. We compared recurrent neural baselines with a transformer-based model and observed a clear performance gap in favor of Twitter-RoBERTa, which achieved substantially higher macro-F1 scores. This outcome aligns with expectations, confirming that contextualized representations are better suited for noisy and implicit language than static word embeddings.

Despite these gains, performance on minority classes remains limited, with especially low recall for JUDGEMENTAL and REPORTED sexism. The results highlight class imbalance and subtle pragmatic phenomena as the main bottlenecks rather than architectural shortcomings. Future work should explore class-balancing strategies, discourse-aware modeling, or hierarchical formulations to better capture implicit and reported sexism.

8 Links to external resources

- Link to weights folder (My drive, public): https://drive.google.com/drive/folders/1-dnP99oPNfkoU1_Z6jDD4YaMKIk8HCO?usp=sharing

References