Assignment 1

Battistoni Leandro, Ponzuoli Giovanni and **Rimondi Simone**
Master's Degree in Artificial Intelligence, University of Bologna
{ leandro.battistoni, giovanni.ponzuoli, simone.rimondi5 }@studio.unibo.it

Abstract

This work addresses the EXIST 2023 Task 2 on sexism detection through multi-class tweet classification. We compare a bidirectional LSTM, a stacked LSTM, and a pre-trained transformer (Twitter-RoBERTa-base), evaluating all models with macro F1-score across three random seeds. Results show a clear performance improvement from recurrent models to transformers, highlighting the effectiveness of attention mechanisms. We further analyze dataset statistics, embedding quality, and ensemble strategies, and provide an error analysis identifying common misclassification patterns, including class overlap and out-of-vocabulary (OOV) effects. Additional experiments explore advanced preprocessing and training techniques, offering insights into robust sexism detection.

1 Introduction

This work addresses the EXIST 2023 Task 2 on sexism detection, focusing on the multi-class classification of tweets according to the sexist intention expressed by the author. The task involves assigning each tweet to one of four classes and is characterized by challenges such as informal language, class imbalance, and semantic ambiguity.

Traditional text classification approaches rely on recurrent neural networks with static word embeddings, which are effective for modeling sequences but limited in capturing long-range dependencies and contextual meaning. Transformer-based architectures have recently become the state of the art, leveraging self-attention to model contextual relationships more effectively in short and noisy texts such as tweets.

Following a progressive modeling strategy, we compare a bidirectional LSTM baseline with a stacked LSTM variant and a pre-trained transformer model, Twitter-RoBERTa-base for hate speech detection. Models are trained and evaluated on the English subset of the EXIST dataset using three different random seeds, and performance is assessed through macro F1-score, precision, and recall. Additional experiments investigate preprocessing strategies, embedding initialization, and training techniques such as label smoothing and class reweighting.

Results show a clear performance improvement from recurrent models to the transformer-based approach, highlighting the importance of contextualized embeddings and attention mechanisms. The analysis further emphasizes the role of embedding quality and robustness across seeds, and provides insights into common error patterns through a detailed error analysis.

2 System description

The experimental pipeline follows a unified workflow consisting of dataset acquisition and cleaning, text preprocessing, vocabulary construction and embedding matrix initialization, model training and evaluation, ensemble creation, and error analysis. Three neural architectures are considered. A bidirectional LSTM classifier is implemented as a baseline, using pre-trained GloVe embeddings and following the Long Short-Term Memory architecture (Hochreiter and Schmidhuber, 1997). A stacked bidirectional LSTM extends the baseline by increasing model depth, while preserving the same training and evaluation protocol. Both recurrent models are implemented in Keras and trained from scratch, with original contributions including embedding handling, hyper-parameter tuning, and robustness analysis across three random seeds. Finally, a transformer-based model, Twitter-RoBERTa-base, is fine-tuned using PyTorch and the HuggingFace Transformers library. This model is based on the RoBERTa architecture (Liu et al., 2019), which builds upon the Transformer and self-attention mechanism introduced by Vaswani et al. (Vaswani et al., 2017). While the transformer architecture itself is not original, this work contributes through dataset adaptation, preprocessing strategies, training techniques such as label smoothing and class reweighting, and a consistent evaluation framework. An ensemble approach is explored by selecting the most confident predictions across seed-specific models, and a detailed error analysis is conducted to identify systematic failure patterns.

3 Experimental setup and results

This section reports the experimental setup and evaluation results. We refer to the bidirectional LSTM as *Baseline*, the stacked LSTM as *Stacked*, and the Twitter-RoBERTa-base model as *Transformer*. All models are optimized with Adam and evaluated using accuracy and macro-averaged precision, recall, and F1-score. Robustness is ensured by training each configuration with three random seeds, and results are reported using an ensemble that selects the most confident prediction per sample.

Model	**Epochs**	**Batch**	**LR**	**#Params**
Baseline	40	128	1e-3	1.2M
Stacked	40	128	1e-3	1.6M
Transformer	15	128	2e-5	124M

Table 1: Hyper-parameter configuration of the models.

Model	**Val**				**Test**			
	Acc	P_m	R_m	$F1_m$	Acc	P_m	R_m	$F1_m$
Baseline	0.70	0.42	0.33	0.34	0.66	0.47	0.40	0.41
Stacked	0.61	0.38	0.34	0.35	0.69	0.49	0.42	0.44
Transformer	**0.79**	0.51	0.48	0.49	0.76	**0.57**	**0.50**	**0.52**

Table 2: Validation and test performance (macro-averaged, ensemble over three seeds).

4 Discussion

Quantitative Results. The results show a clear improvement in performance as model complexity and representational power increase. The *Baseline* bidirectional LSTM achieves the lowest Macro F1-score, reflecting the limitations of shallow recurrent architectures with static embeddings. The *Stacked* LSTM benefits from increased depth, yielding moderate gains. The *Transformer* model significantly outperforms both recurrent approaches, confirming the effectiveness of self-attention mechanisms and contextualized representations for tweet-level classification. This advantage is further supported by subword tokenization and domain-specific pretraining on large-scale Twitter data.

The aggressive preprocessing pipeline reduces vocabulary sparsity and benefits LSTM-based models but may discard semantically rich signals such as emojis and hashtags. Despite this potential information loss, the transformer model remains robust and achieves the best overall performance.

Error Analysis and Limitations. Error analysis reveals that tweet length slightly affects LSTM-based models, while it has negligible impact on the transformer. Slang and informal language increase error rates across all systems, though less markedly for the transformer. No clear correlation is observed between the number of out-of-vocabulary tokens and misclassification rates. All models exhibit high-confidence errors, highlighting challenges posed by semantic ambiguity. Finally, given the subjective nature of sexism detection and moderate human agreement reported in prior work, the obtained F1-scores represent strong results within the intrinsic limits of the task.

5 Conclusion

This work investigated EXIST 2023 Task 2 on sexism detection by comparing recurrent and transformer-based architectures for multi-class tweet classification. A progressive modeling strategy was adopted, moving from a bidirectional LSTM baseline to stacked LSTM and finally to pre-trained transformers. Experimental results showed consistent performance improvement with increasing model complexity and the introduction of contextualized representations. The transformer-based approach achieved the best overall performance, highlighting the effectiveness of self-attention mechanisms and domain-specific pre-training for noisy social media text.

Despite these positive results, several limitations remain. The task is inherently subjective, with ambiguous language and moderate human agreement limiting achievable performance. Additionally, aggressive preprocessing may remove semantically relevant signals, and the ensemble strategy, while improving robustness, increases computational cost. Future work could explore softer preprocessing strategies, more advanced ensembling methods, and the use of multilingual or instruction-tuned models to better capture nuanced sexist expressions. Further analysis of confidence calibration and uncertainty estimation may also help mitigate high-confidence misclassifications.

6 Links to external resources

The source code is available at this repository.

References

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *Proceedings of EMNLP-IJCNLP*, pages 3558–3568.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.