Assignment 1

Matteo Fasulo, Maksim Omelchenko, Luca Babboni and **Luca Tedeschini**
Master's Degree in Artificial Intelligence, University of Bologna
{ matteo.fasulo, maksim.omelchenko, luca.babboni2, luca.tedeschini3 }@studio.unibo.it

Abstract

This report investigates methodologies for detecting sexist content in tweets, utilizing both English and multilingual datasets provided. Two primary approaches were evaluated: a Bidirectional LSTM model and the transformer-based model *Twitter-RoBERTa-base-hate*[1], fine-tuned on hate speech. Additionally, we assessed the performance of *XLM-RoBERTa-base*[2] for multilingual detection of English and Spanish tweets. The experimental results reveal the limitations of sequential models: bidirectional LSTMs with a single layer achieved an F1 score of 0.68, struggling to capture nuanced contextual information effectively. Adding a second layer to the LSTM architecture yielded similar performance. In contrast, transformer models achieved significantly better results. The fine-tuned *Twitter-RoBERTa-base-hate* reached an F1 score of 0.84 on English-only tweets, while *XLM-RoBERTa-base* demonstrated robust multilingual detection capabilities with an F1 score of 0.83.

1 Introduction

Deep learning models, including transformers like BERT, excel in tasks requiring rich contextual understanding by automatically learning features. However, they require significant computational resources and are prone to overfitting. Hybrid approaches that combine classical techniques with pre-trained embeddings can improve performance and adaptability but introduce additional complexity and are not universally applicable. Our approach for Long Short-Term Memory networks (LSTMs) utilizes the pre-trained embedding model *glove-twitter-50*[3], which provides over one million embedding vectors of dimension 50, specifically designed for tweets. To address the slight imbalance in our dataset, which is skewed towards non-sexist tweets, we incorporated class weights into the loss function during model training. Furthermore, we employed ensemble techniques, using majority voting across multiple seeds, to enhance the robustness of our results. Our experiments evaluated two primary models: a two-layer bidirectional LSTM and transformer-based models fine-tuned on our dataset. Inspired by previous research in the field (Muti and Mancini, 2023), we optimized hyperparameters through empirical testing and insights from related challenges (Plaza et al., 2023).

2 System description

The preprocessing pipeline consisted of several key steps. First, hard labels were generated based on annotator votes, categorizing tweets as sexist or non-sexist, and ties were discarded. Only English tweets were retained, unnecessary columns were removed, and labels were encoded as integers. Text cleaning followed the original GloVe preprocessing steps[4] by removing URLs, mentions, hashtags, punctuation, and elongations, replacing them with special tokens to minimize out-of-vocabulary (OOV) tokens. The text was then lemmatized and converted to lowercase. To address vocabulary gaps in GloVe, missing words from the training set were added, with their vectors computed as the average of the five most frequent co-occurring words. The LSTM architecture (Figure 1) was custom-designed with a pre-trained GloVe embedding layer and implemented with one and two layers with dropout layer to prevent overfitting. For the transformer model, a pre-trained *Twitter-RoBERTa-base-hate* was used with class weights in the loss function, yielding improved results. Experiments with Spanish tweets using the

[1] https://huggingface.co/cardiffnlp/twitter-roberta-base-hate

[2] https://huggingface.co/xlm-roberta-base

[3] https://nlp.stanford.edu/projects/glove/

[4] https://nlp.stanford.edu/projects/glove/preprocess-twitter.rb

XLM-RoBERTa-base model also showed favorable results without requiring pipeline modifications.

3 Experimental setup and results

For the bidirectional LSTM architecture, as described earlier, we used a batch size of 128 and a hidden dimension of 256. The model was trained for 60 epochs with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. After experimenting with various batch sizes ($\{64, 128, 256, 512, 1024\}$) and hidden dimensions ($\{128, 256, 512, 1024\}$), we determined this configuration to be optimal. The AdamW optimizer and a CrossEntropy loss function, weighted according to class distribution, were used to handle imbalance. A dynamic learning rate scheduler reduced the learning rate on a plateau based on validation loss, halving it with patience of 10 epochs. Early stopping was implemented to retain the model with the lowest validation loss (Figure 2). The primary evaluation metric was the macro F1 score, with accuracy also tracked. For the RoBERTa transformer model, we used a batch size of 16 and trained for 6 epochs with a learning rate of 5×10^{-6} and a weight decay of 1×10^{-5}. Class distribution disparities were handled using a weighted loss function. Both transformer and LSTM models were trained using three different seeds (1337, 42, 69), and predictions were aggregated using majority voting across seeds.

Model	F1 Macro	Accuracy	Tweet
LSTM (1 layer)	0.688	0.688	en
LSTM (2 layers)	0.688	0.688	en
RobertaHate	0.841	0.842	en
XLM-R	0.831	0.832	en, es

Table 1: Metrics for the ensembled models over the test set. Metrics for the validation set are available inside the Notebook.

4 Discussion

The results in Table 1 align with our expectations: transformer-based models significantly outperform sequential models. Both single-layer and two-layer LSTM architectures achieved identical F1 scores of 0.688, demonstrating that increasing the depth of the LSTM did not lead to any improvement. In contrast, the transformer-based *Twitter-RoBERTa-base-hate* model provided a substantial performance boost, achieving an F1 score of 0.841—an improvement of 15.3% over the LSTMs. This improvement can be attributed to the attention mechanism in transformers, which effectively captures contextualized information in tweets. The *XLM-RoBERTa-base* model also performed well, achieving an F1 score of 0.831, demonstrating its robustness for multilingual detection without requiring significant pipeline modifications. Additionally, the nearly identical accuracy and F1 scores across models indicate balanced precision and recall, even with the class imbalance in the dataset, suggesting the effectiveness of using a weighted loss function. For instance, in the sentence "*My baby called me mommy Sha for the first time today twice!!! Y'all don't understand how hype that made me. Baby girl has autism and getting her to talk without being prompted has been a challenge. She's come so far*," the transformer model correctly identified it as non-sexist, while the LSTMs failed. Similarly, in the explicitly sexist sentence "*Aughhh I still got an exam tomorrow. I hate women*," only the transformer model correctly identified the sexism. However, certain limitations remain, as demonstrated by the following non-sexist sentence: "*Ladies, don't let anyone body shame you in any way...you are fat so what..you are beautiful my dear...be confident in yourself...peace*," which was incorrectly classified as sexist by all models. To make our work more accessible and interactive, we developed a dashboard where users can test our models and check whether a given tweet is classified as sexist or not.

5 Conclusion

In this work, we explored methods for detecting sexist content in tweets, comparing sequential and transformer-based models. While bidirectional LSTMs struggled to capture tweet context, transformer-based models, such as *Twitter-RoBERTa-base-hate*, excelled on English data, and *XLM-RoBERTa-base* demonstrated robustness on multilingual data. Despite meeting expectations, challenges remain in identifying nuanced or implicit sexism, highlighting the need for advanced architectures. Limitations include reliance on pretrained embeddings, difficulty with implicit cases, and overfitting to patterns in the training data. Future work should integrate attention mechanisms into LSTMs and explore sentiment and emotional context for improved performance.

6 Links to external resources

- Github repository
- Dashboard

7 Images

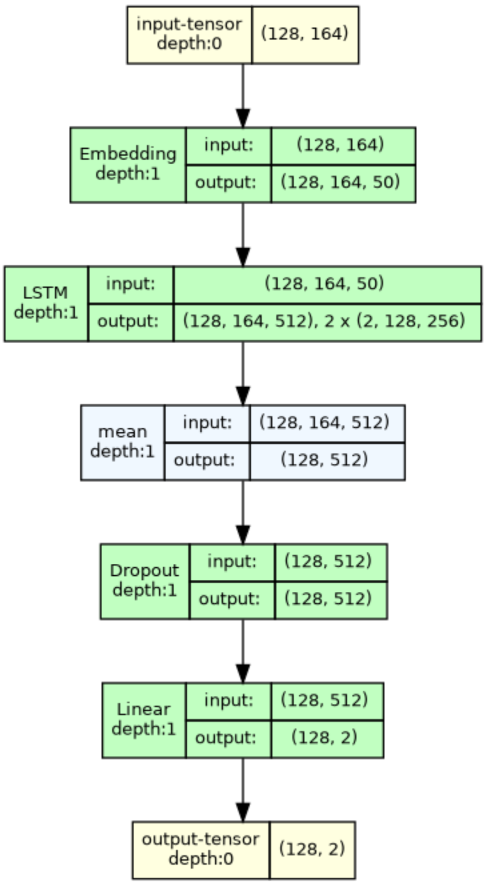

Figure 1: One layer LSTM architecture for the first batch of data

References

Arianna Muti and Eleonora Mancini. 2023. Enriching hate-tuned transformer-based embeddings with emotions for the categorization of sexism. In *CLEF 2023: Conference and Labs of the Evaluation Forum*, Thessaloniki, Greece. CEUR Workshop Proceedings. DOI avaiable on request.

Laura Plaza, Jorge Carrillo-de Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2023. Overview of exist 2023 – learning with disagreement for sexism identification and characterization. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18–21, 2023, Proceedings*, page 316–342, Berlin, Heidelberg. Springer-Verlag.

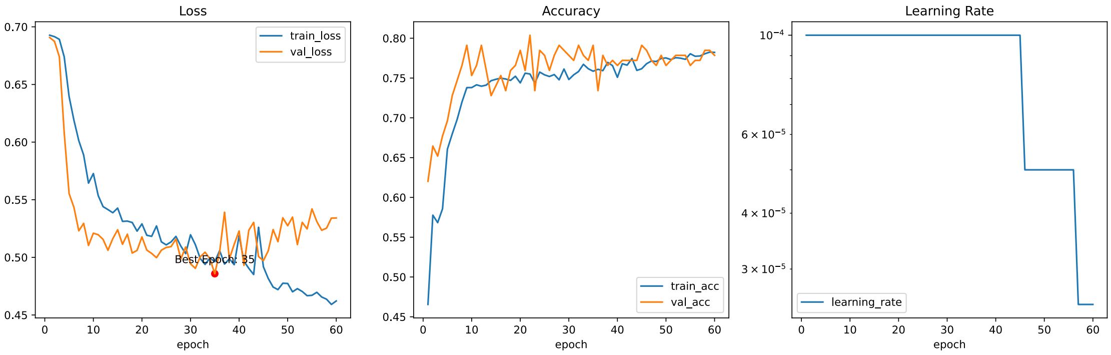

Figure 2: Two layer LSTM training history with loss, accuracy and learning rate update over epochs