Assignment 1

Erfan Samieyan Sahneh
Master's Degree in Artificial Intelligence, University of Bologna
{ erfan.samieyansahneh}@studio.unibo.it

Abstract

This project explores the task of sexism detection using NLP techniques on English tweets. Initially, I implement RNN models with one or two layers of bidirectional LSTM to classify text as sexist or non-sexist. These models provide a baseline for performance, emphasizing simplicity and interpretability. In the second phase, I leverage a pre-trained RoBERTa model that was trained on hate speech detection, fine-tuned for this task, to evaluate the benefits of transfer learning. Interestingly, the baseline RNN models achieved performance comparable to the fine-tuned model, suggesting that simpler architectures can be competitive on this task. Our findings provide insights into the trade-offs between simple architectures and advanced pre-trained transformers for addressing societal challenges in online communication.

1 Introduction

Recent research has focused on developing automated methods to identify sexism in tweets. Traditional approaches have employed machine learning techniques, utilizing features such as n-grams and sentiment analysis to detect overtly sexist language. However, these methods often struggle with the nuanced and context-dependent nature of sexism, which can manifest in subtle or implicit forms.

To address these challenges, studies have explored the use of deep learning models, particularly RNNs and Transformer-based architectures. For instance, (Kalra and Zubiaga, 2021) investigated the application of LSTM networks and CNNs for sexism classification in tweets, highlighting the effectiveness of these models in capturing contextual information.

More recent advancements have leveraged pre-trained language models like RoBERTa, fine-tuning them for the specific task of sexism detection. (Aru et al., 2024) participated in the "EXIST 2024" challenge, focusing on sexism categorization in tweets using Transformer-based systems. Their approach demonstrated the potential of such models in achieving high accuracy in detecting and categorizing sexist content.

In this project, I used one- and two-layer bidirectional LSTMs with GloVe embeddings (Pennington et al., 2014) and a pre-trained RoBERTa model to detect sexism in tweets, evaluating their performance using the F1 score. The LSTM captures contextual dependencies with word embeddings, while RoBERTa (Barbieri et al., 2020) leverages advanced pre-trained representations. Comparing their performance highlights the trade-offs between simple architectures and Transformer-based models for sexism detection.

2 System description

In this project, I implemented two models for detecting sexism in tweets: a **Bidirectional LSTM-based model** and a **fine-tuned RoBERTa model**. Both models were evaluated using the **F1 score** for binary classification.

The second model is based on a pre-trained **RoBERTa** model, fine-tuned on our dataset. I used **AutoModelForSequenceClassification** from huggingface for the fine-tuning process.

The primary contribution is the design of the **Bidirectional LSTM model** and the fine-tuning of the **RoBERTa model** for sexism detection. and for the preprocessing part, I used my old projects. The F1 score was used to evaluate both models, highlighting their performance in detecting sexist content.

3 Experimental setup and results

3.1 Experimental Setup

I conducted experiments using two models: a Bidirectional LSTM with GloVe embeddings and a fine-tuned RoBERTa transformer. The Bidirectional LSTM model utilized pre-trained GloVe embeddings (300 dimensions) to represent input text, followed by a single(double) bidirectional LSTM layer with 128 hidden units and a dense output layer with a sigmoid activation for binary classification. For optimization, I used the Adam optimizer with a learning rate of 0.001.

Both models were evaluated using the F1 score, precision, recall, and accuracy.

I ran experiments with the LSTM model using three different random seeds for Tensorflow and Numpy, and reported the average performance and standard deviation. Then I chose the best model(seed) for further analysis.

3.2 Results

The results for both models are summarized in Table 1. The Bidirectional LSTM achieved an average test accuracy of 0.7646 ± 0.0216, with a weighted F1-score of 0.78. The fine-tuned RoBERTa model outperformed the LSTM model, achieving an accuracy of 0.83 and a weighted F1-score of 0.83.

Table 1: Performance Comparison of LSTM and RoBERTa Models

Model	**Accuracy**	**F1-Score**
LSTM (Average)	0.7646 ± 0.0216	0.78
RoBERTa	0.83	0.83

4 Discussion

4.1 Quantitative Results

The results show that the fine-tuned RoBERTa model outperforms the Bidirectional LSTM in terms of accuracy and F1-score. This is expected, as RoBERTa leverages contextual embeddings and a transformer architecture, making it better at capturing subtle patterns in text and a higher rate of precision.

Interestingly, the LSTM model performed closer to RoBERTa than anticipated, suggesting that the use of GloVe embeddings was effective. However, the RoBERTa model showed a better balance in handling both "Sexist" and "Not Sexist" classes, especially in terms of recall for the minority class.

4.2 Error Analysis

The first notable issue is the misclassification of tweets that contain language similar to sexist tweets but lack any reference to gender. For example, tweets that are offensive or toxic in tone but do not specify a gender were often incorrectly labeled as sexist. This suggests that the models rely heavily on surface-level patterns, such as the presence of offensive words, without fully understanding the context.

Another common error involves tweets that mention a specific gender but are not offensive or sexist. For instance, a tweet celebrating or neutrally describing a gender-related topic might be misclassified due to the model's sensitivity to gendered terms, even when no bias or sexism is present.

Also, both models struggled with subtle cases. For example:

- Tweets with positive tone but the same set of words (*"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"*) were often missed by both models.
- Ambiguous phrases like *"Going weak rn... PLAYING THE GUITAR?!?!?! MANSPREADING!?!??!?! I AM ON MY KNEES"* were hard to classify as sexist or non-sexist without deeper context.
- The LSTM model often misclassified tweets with figurative language due to its reliance on static embeddings.

These errors highlight challenges in detecting implicit bias and understanding context, especially for subtle forms of sexism.

5 Conclusion

I explored sexism detection in tweets using a Bidirectional LSTM with GloVe embeddings and a fine-tuned RoBERTa model. While the RoBERTa model outperformed the LSTM in terms of accuracy (0.83 vs. 0.78) and F1-score (0.83 vs. 0.78), the gap was smaller than expected, highlighting the effectiveness of the LSTM's design. Moreover, the difference between one and two-layer LSTM is negligible and in some cases, one layer got better results. However, both models faced challenges in detecting nuanced or implicit sexism, with frequent misclassifications of offensive but non-sexist language and benign gendered references.

The main limitations include the models' reliance on surface-level patterns and their difficulty in understanding subtle context.

6 Links to external resources

GitHub page.
Transformer best model weights

References

Giacomo Aru, Nicola Emmolo, Simone Marzeddu, Andrea Piras, Jacopo Raffi, and Lucia C. Passaro. 2024. Robexedda: Sexism detection in tweets. In *Conference and Labs of the Evaluation Forum*.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification.

Amikul Kalra and Arkaitz Zubiaga. 2021. Sexism identification in tweets and gabs using deep neural networks.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. Website.