Assignment 1

Tancredi Bosi and **Giovanni Grotto**
Master's Degree in Artificial Intelligence, University of Bologna
{ tancredi.bosi, giovanni.grotto }@studio.unibo.it

Abstract

This report explores detecting sexism in tweets using three models: one-layer and two-layer Bidirectional LSTMs, and the Twitter-RoBERTa-base Transformer. Evaluations showed the Transformer model significantly outperformed the others, with a 10-point accuracy improvement over the Baseline. The two-layer LSTM offered marginal gains. Analysis revealed challenges in handling sarcasm, ambiguous phrasing, and informal language, while experiments with embedding sizes highlighted potential trade-offs between representation richness and noise. These findings underscore the effectiveness of Transformer models and provide insights for refining sexism detection approaches.

1 Introduction

There are various methods to approach this task. The state-of-the-art methods rely on transformers like BERT, which leverage bidirectional attention to capture the contextual meaning of tweets. However, these methods come with the downsides of being computationally intensive to train and lacking interpretability.
In this report, we will try to solve this task by training two bi-directional LSTM models from scratch and fine-tuning a pre-trained version of RoBERTa specifically designed for hate speech detection.
We will compare these models on the same dataset and conduct analyses on the impact of out-of-vocabulary (OOV) tokens on the final results.
Additionally, we will investigate how the embedding dimension affects accuracy and F1 score.

2 System description

In the assignment three architectures are used:
Baseline: a Bidirectional LSTM with a Dense layer on top;
Model 1: a two layer Bidirectional LSTM with a Dense layer on top;
Transformer: Twitter-roBERTa-base for Hate Speech Detection.

In particular, Baseline and Model 1 are implemented through a class that we defined, which takes as input:
- ***input_size***: dimension of the input, set for both to 100, due to the GloVe embedding dimension.
- ***hidden_size***: dimension of the hidden state, set for both to 128.
- ***num_layers***: number of LSTM layers, set to 1 for "Baseline" and to 2 for "Model 1".
- ***output_size***: dimension of the output, set for both to 1 being the task a binary classification problem.

The "Transformer" model instead is loaded from HuggingFace. It is a transformer model trained on 58M tweets and finetuned for hate speech detection. In particular, the model is specialized in detecting hate speech against women and immigrants, thus perfect for our problem.

3 Experimental setup and results

To train the two architectures described above, we used the same hyperparameters: a batch size of 64, a learning rate of 0.001 with the Adam optimizer (all other parameters were set to their default values). These are standard values, and after experimenting with modifications to these hyperparameters, we did not observe any meaningful changes in performance.

To ensure a fair evaluation of each architecture, we also used different random seeds for training and evaluation, as specified in the assignment. This approach helps reduce variability in the results and provides a clearer indication of which model performs best.
This approach applies only to the first two models. Since training the transformer is highly

time-consuming, we were unable to train it using more than one random seed.

The evaluation metrics used to compare all the models include accuracy and F1 score. Additionally, we utilized the precision-recall curve to conduct a more detailed analysis of the differences in the results that can be seen in 1.

Metric	Baseline	Model 1	Transformer
Accuracy	0.73	0.75	0.84
Precision	0.73	0.76	0.84
Recall	0.73	0.74	0.84
F1 Score	0.73	0.74	0.84

Table 1: Performance metrics for Baseline, Model 1, and Transformer.

4 Discussion

Discussion of quantitative results

The Transformer model achieved the best performance across all metrics, significantly outperforming both LSTM-based models with a 10-point improvement in accuracy and F1 score over the Baseline. This result aligns with expectations, as Transformer models are known for their ability to capture contextual relationships in text, making them highly effective for tasks like sexism detection in tweets.

Model 1, with an additional LSTM layer, showed a marginal 0.02-point improvement in accuracy and F1 score over the Baseline, indicating that the added complexity captured only slightly more meaningful features within the dataset constraints.

Error analysis

In the error analysis, we first checked the distribution of the labels in the three given sets and we noted that there is a slight imbalance: non-sexist tweets appear more frequently than sexist tweets.

We then analyzed the confusion matrix and precision-recall curve of the three models' performances. Both the baseline and Model 1 demonstrate a tendency to produce slightly more false negatives than false positives. This behavior reflects the imbalance already observed in the dataset during both training and testing phases, where the models may struggle to correctly identify the minority class, leading to a bias toward underpredicting certain categories.

In contrast, the transformer model exhibits a different pattern: it generates more false positives than false negatives, indicating a more balanced recall between the two classes (0-predictions and 1-predictions).

Finally, we analyzed the misclassified samples. Here a representative example misclassified as sexist by all the three models:
"ladies dont have a miscarriage in louisiana if you do in addition to probably having your friends and family sued by some opportunistic yokel youll be charged with murder".
This example shows the behavior of the models: it is a long sentence, leading to create confusion to the model, uses aggressive words, such as "murder". In general, the majority of the errors is performed in cases similar to this one, specifically, the non-sexist tweets with aggressive words tend to be predicted wrongly as sexist.

An interesting observation concerns the presence of out-of-vocabulary (OOV) tokens, which appear to have no influence on the final performance. On average, the test set contains 2.48 OOV tokens per tweet. When we specifically examine tweets that were misclassified, the average number of OOV tokens per tweet is 2.03, showing that the presence of OOV tokens in a tweet does not significantly influence its classification.

Another observation is that changing the dimension of the GloVe embeddings can influence performance in unexpected ways. For instance, increasing the embedding size from 100 to 200 resulted in almost no performance gain. This may be due to the fact that increasing the embedding dimensions allows for capturing more meaning and details related to a token, but at the same time, it may introduce issues. Specifically, these additional details may not be necessary for the task, leading to increased complexity without any benefit. Moreover, larger embeddings mean more parameters to tune.

5 Conclusion

As expected, the transformer model is the best-performing one, achieving an accuracy improvement of 10 points over the baseline. In contrast, increasing the complexity of the LSTM model by adding an additional layer led to a slight, but not significant, improvement—only 0.02 points in accuracy.

References