Assignment 1

Hassan Mujtaba Ahmed, Second Student, and **Third Student**
Master's Degree in Artificial Intelligence, University of Bologna
{ hassan.ahmed5, name2.surname22, name3.surname33 }@studio.unibo.it

Abstract

This report evaluates sexism detection in English tweets from the given assignment dataset. We compare custom Bidirectional LSTM architectures using GloVe embeddings against a fine-tuned Twitter-roBERTa-base-hate Transformer model. Our findings indicate that while the Transformer achieved a superior macro F1-score of 0.46, severe class imbalance prevented all models from successfully identifying the 'JUDGEMENTAL' category.

1 Introduction

The goal of given dataset and assignment is to classify the author's intention into four categories: non-sexist ('-'), DIRECT, REPORTED, and JUDGEMENTAL. Sexism detection is a complex NLP challenge due to the informal nature of social media and the need to understand nuanced human intentions. We employ a dual approach to compare the effectiveness of established RNN-based sequence modelling with state-of-the-art Transformer self-attention mechanisms. All experiments were conducted across three random seeds to ensure robust evaluation.

2 System description

- Preprocessing: Tweets were cleaned of emojis, hashtags, mentions, and URLs. For LSTMs, we applied lemmatisation and stemming, while the Transformer used the Twitter-roBERTa tokenizer.
- Architectures: The Baseline Bi-LSTM utilized 256 units, while the Stacked version added a second layer of 128 units. Both utilized non-trainable 50-dimensional GloVe embeddings. The Transformer model was fine-tuned for one epoch with a learning rate of $2e-5$.

3 Experimental setup and results

Models were evaluated using Macro F1-score, Precision, and Recall. Numerical results are summarized below:

Model	MacroF1	Accuracy	Precision	Recall
Best Bi-LSTM (Baseline)	0.38	0.75	0.46	0.37
Bi-LSTM Ensemble (Top 4)	0.36	0.74	0.46	0.35
Transformer (Twitter-roBERTa)	0.46	0.78	0.46	0.47

Table 1: Comparison of different models in the assignment

Note: Bi-LSTM scores represent the best seed performance (321) as recorded in the logs

4 Discussion

The primary obstacle was the heavy data skew toward non-sexist tweets, which comprise the vast majority of the 2,867 training samples.

1. The 'Judgemental' Gap: Neither the Bi-LSTMs nor the Transformer could identify the JUDGEMENTAL class, resulting in a 0.00 recall for that label.
2. Misclassification: Both models frequently mislabelled DIRECT sexist messages as non-sexist, such as the tweet "hot girl cant get nowher without a gps".
3. Comparative Performance: The Transformer proved significantly better at differentiating the two largest classes ("-" and "DIRECT") compared to the custom RNNs.

5 Conclusion

The Transformer (Twitter-roBERTa) model provided the best overall performance, yet the failure to detect JUDGEMENTAL content highlights the limitations of training on imbalanced datasets. We should implement over-sampling for minority classes or weighted loss functions to penalise majority-class bias.

References

Jason Brownlee. 2021. How to develop a bidirectional LSTM for sequence classification in Python with Keras. Machine Learning Mastery.

CardiffNLP. n.d. cardiffnlp/twitter-roberta-base-hate. Hugging Face Model Hub.

Anish Nama. 2023. Understanding bidirectional LSTM for sequential data processing. Medium.

Samarth Sarin. 2018. Simple guide for LSTM and GloVe embeddings. Kaggle Notebook.