Assignment 1

Samuele Centanni, Tomaz Cotic and Mattia Lodi
Master's Degree in Artificial Intelligence, University of Bologna
{samuele.centanni, tomaz.cotic, mattia.lodi3}@studio.unibo.it

Abstract

This report outlines our work on the first assignment for the Natural Language Processing course. The primary objective is to perform text classification using models of increasing complexity. We compared the performance of a Bidirectional LSTM, a pretrained RoBERTa model, and an SVM baseline. Our findings indicate that while the Transformer outperforms the Bidirectional LSTM, neither achieves outstanding results due to the unbalanced nature of the dataset and the inherent difficulty of the task.

1 Introduction

The task involves classifying tweets into four categories: *not-sexist*, *direct*, *judgmental*, or *reported*. The core comparison is between a single-layer *Bidirectional LSTM*, a two-layer *Bidirectional LSTM*, and a Transformer-based model, *RoBERTa*.

We also evaluated different embedding strategies, an alternative RNN architecture (GRU), and a comparison between two Transformer models: *RoBERTa* and *DeBERTa-v3-base*. Finally, all results were measured against a simple yet effective *SVM* baseline.

Following an assessment of the results, we analyzed the causes of classification errors, which we attributed to the **semantic similarity** between classes.

2 System description

We adapted an existing dataset to our requirements and defined three classification models. The pipeline differs between the RNN and Transformer architectures.

For the RNN models:

- We preprocessed the dataset by removing non-informative words and performing lemmatization.
- We generated text embeddings using a pretrained method.
- We defined and trained two BiLSTM models, one with a single layer and one with two layers.

For the Transformers:

- As RoBERTa and DeBERTa utilize specific preprocessing, we provided them with raw text.
- For the latter, following the RoBERTa documentation, we only performed minor preprocessing.
- We defined the tokenizer and analyzed how each transformer tokenizes input text.

3 Experimental setup and results

We performed classification using the three neural models, evaluating them via F1-score.
Moreover, thanks to visually informative plots such as *ROC curve*, *AUC curve* and *confusion matrix*, we were able to derive interesting conclusions in Section 4.

The LSTM models were trained for 10 epochs using a hidden dimension of 128 and a learning rate of 1×10^{-4}. For the Transformer-based models, we maintained the same training duration but reduced the learning rate to 2×10^{-5}.

The results are summarized in Table 1.

Model	F1-score (mean)	F1-score (best)
1-layer BiLSTM	0.3522	0.3573
2-layer BiLSTM	0.3618	0.3637
RoBERTa	**0.5013**	**0.5305**
DeBERTa	0.4736	0.5021

Table 1: Classification results for the neural models.

4 Discussion

4.1 Quantitative Analysis

Quantitative results indicate that RoBERTa-hate-speech outperforms all other models, including the more advanced DeBERTa-v3-base. This suggests that domain-specific fine-tuning on hate speech is more effective for this task than general architectural improvements.

Among RNNs, the 2-layer BiLSTM provided a marginal improvement over the 1-layer version, while the Bi-GRU performed worst ($F1 = 0.3420$), likely due to its simpler gating mechanism being less capable of capturing long-range dependencies in complex tweets.

The SVM baseline achieves a competitive overall accuracy but performs poorly in terms of F1-score due to its heavy bias toward the majority class (*non-sexist*). While it effectively identifies the most common label, it fails to generalize to minority categories, where even the BiLSTM shows a 7-8% improvement in recall, and the Transformer models demonstrate a far superior ability to distinguish subtle linguistic features.

4.2 Error Analysis

The confusion matrix reveals two primary failure modes: **minority absorption** and **semantic ambiguity**.

The *reported* class suffers from extreme imbalance; despite a high AUC, the low AP indicates that false positives dominate, while the *judgmental* class exhibits blurred linguistic boundaries that confuse even the Transformer.

In particular, the dataset exhibits a significant class imbalance, heavily skewed towards the *non-sexist* class, which constitutes approximately 70% of the training samples (over 2000 instances). Among the sexist categories, *direct sexist* is the most represented ($\sim 19\%$), while *judgmental* and *reported* are strictly minority classes, accounting for only $\sim 6.5\%$ and $\sim 4.5\%$ of the data respectively.

Misclassifications are most frequent between *neutral* and *direct* categories, suggesting a high degree of overlap in their textual patterns.

4.3 Future Work

Since dataset balancing via oversampling did not yield significant improvements, we conclude that class imbalance is not the sole cause of performance degradation. Future efforts should focus on enhancing data quality and refining class definitions, as minority class performance likely requires task-specific modeling beyond simple resampling.

5 Conclusion

In this report, we evaluated various neural architectures for tweet classification, ranging from RNNs to Transformer-based models. As expected, RoBERTa achieved the best results, significantly outperforming the SVM baseline and BiLSTM models. However, a surprising finding was that DeBERTa-v3-base underperformed compared to the domain-specific RoBERTa-hate-speech, highlighting that specialized pretraining is more impactful than architectural depth for this task. Furthermore, the 2-layer BiLSTM showed only marginal gains over the 1-layer version, while the Bi-GRU proved to be the least effective RNN architecture.

The main limitation of our approach remains the poor performance on minority classes, which did not improve significantly despite the application of oversampling techniques. This suggests that the difficulty lies not only in the class imbalance but also in the high semantic ambiguity and blurred boundaries between categories such as *judgmental* and *reported*.

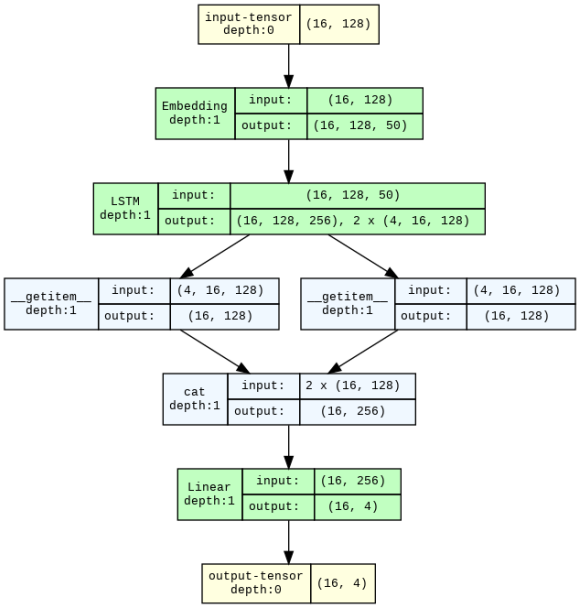

Figure 1: BiLSTM model architecture.

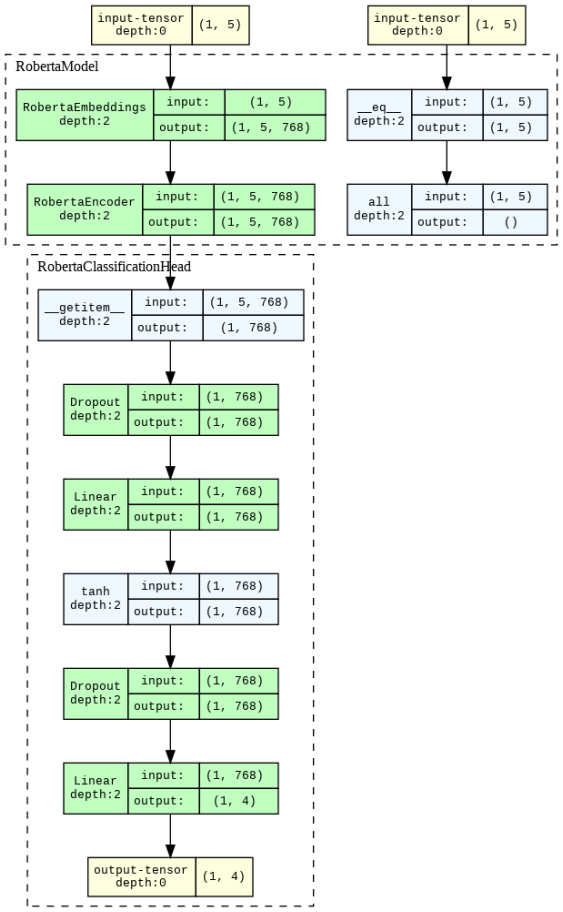

Figure 2: RoBERTa model architecture.