NLP: Assignment 1

Alberto Genovese, Simone Reale, and **Alessandro Tutone**
Master's Degree in Artificial Intelligence, University of Bologna
{ alberto.genovese5, simone.reale3, alessandro.tutone }@studio.unibo.it

Abstract

This assignment aims to address the **EXIST 2023 Task 1** on sexism detection. Several models have been implemented to decide whether or not a given tweet contains or describes sexist expressions or behaviors. In the first part, two models were employed: a *Bidirectional LSTM* with just one LSTM layer and a second *Bidirectional LSTM* with twice the number of LSTM layers. The evaluation of the two models allowed us to compare their performances and their ability to solve the given task. In the second part, we adapted the ***Twitter-RoBERTa-base for Hate Speech Detection*** model, a *Transformer-based model*. The evaluation obtained demonstrates the potential of these models to address sexism detection. However, the results also highlight what are challenges and difficulties in understanding and accurately managing natural language.

1 Introduction

Sexism is a widespread issue in online social networks. While human supervision has traditionally been the primary approach to deal with this problem, thanks to the rapid advancement of artificial intelligence, the detection of such harmful contents can now be automatized. Several models such as *SVMs*, *RNNs*, *LSTMs*, and much more, are able to perform classification in various contexts. Although the task can be achieved in a very rapid way, these models have some explainability problems. Since we're addressing a binary classification task, the model is just required to tell if a given sentence is sexist or not without adding any explaination to its prediction. The lables are not always correct, indeed the nuanced nature of sexism, often expressed subtly or implicitly, arises further challenges for automated detection. To address these limitations, a combination of AI and human expertise is essential. Human moderators can provide valuable feedback to refine models and ensure an accurate detection of sexist content.

The experiments have been conducted by evaluating both the architectures introduced before with different setups. The *LSTM models (Baseline, Model 1)* were run with three different random seeds *(42, 347, 1337)* and the evaluation was done both on the average performances of each model and on their comparison, by keeping also the best seed. On the other hand, the *RoBERTa Transformer model* is run with the best hyperparameters setup found and its performances have been evaluated with respect to the LSTM models. What can be noticed from the experiments is that randomness is something that should not be overlooked, since changing the seed results in variations of the outcomes that cannot be neglected.

2 System description

During the first part of the assignment, we followed the architecture definition given by the instructions. Once the models were able to perform some predictions, we started to look for some papers that could help improving the performances of our two models. We came across an interesting paper: "*A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition*" (Zeyer et al., 2017). Considering this research, we achieved better performances for both our LSTM models by applying an hyper-parameter selection close to the one employed in the study. An additional small improvement, still suggested by the paper, was achieved when implementing the NAdam (Nesterov-accelerated Adaptive Moment Estimation) optimizer, namely an optimization algorithm that combines two techniques: the adaptive learning rates of Adam and the momentum-based acceleration of Nesterov Accelerated Gradient (NAG).

3 Experimental setup and results

For both the *Bidirectional LSTM* models, the following hyper-parameters have been used (all the values that are not mentioned are kept as default):

- *batch size:* **16**
- *training epochs:* **10**
- *optimizer:* **NAdam**
- *learning rate:* **1e-3**
- *metrics:* **accuracy** & **F1-score**
- *loss:* **binary cross-entropy**

On the other hand, slightly different hyper-parameters were used for the *Transformer-based* model (all the values that are not mentioned are kept as default):

- *batch size:* **16**
- *training epochs:* **10**
- *learning rate:* **5e-6**
- *epsilon:* **1e-7**
- *metrics:* **accuracy** & **F1-score**
- *loss:* **binary cross-entropy**

Model	Avg-F1 val	Seed	F1 test
Baseline LSTM	0.781	1337	0.735
Model 1 LSTM	0.777	347	0.746
Transformer	0.843	42	0.81

Table 1: Models results

4 Discussion

Considering the best outcomes of each model with the related seed, as expected, the worst results have been achieved by the "*Baseline LSTM*" model. Further improvements have been exploited by the "*Model 1 LSTM*", showing better generalization capabilities, even if its average F1-score on the validation set is lower. However, the best-performing model remained the *Transformer*, whose outcome is over the 10% better than LSTMs in general.

Looking at the errors, what can be evinced from these experiments is that the main causes of any lack of performance are related to the data distribution. For sure, class imbalance towards non-sexist tweets influences how the models learn, even if a deeper analysis of the tweets showed that there are some recurrent errors, due to the presence or absence of certain words. In particular, this behavior can be observed with gender-related tokens (in particular, the word "woman"), as shown in the following false-positive examples:

> *400237: "Yup I hate when men rape and kill women."*
>
> *400409: "Today women eye makeup. And those nails too."*

For any future development, it is recommended to work first on the data, trying to gather more balanced information or to apply some methods to re-balance the distribution. Regarding the model, more complex architectures can be implemented to optimize the results, as also described in the notebook. Indeed, any further deduction on the topic can be found in the error analysis part of the notebook.

5 Conclusion

Our work leads to the conclusion that implementing models, like *LSTM* and *Transformers* could be a reasonable choice to improve in tasks like *sexism detection*. The results achieved were not bad, but if considering a real use inside a more complex system, several considerations have to be done. A cost analysis would be crucial to understand which type of error to reduce (*False Positives* or *False Negatives*), depending on the application. In addition, the accuracy is still too low for a fully automated process, meaning that a human supervision is still required. As final considerations, the Transformer represents a big improvement with respect to LSTMs, not surprisingly, but still shows some room for refinements. Indeed, one point to still work on is removing the dependency of the model from single words and focusing on gathering more information from the semantic of the sentence.

References

Albert Zeyer, Patrick Doetsch, Paul Voigtlaender, Ralf Schlüter, and Hermann Ney. 2017. A comprehensive study of deep bidirectional lstm rnns for acoustic modeling in speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2462–2466.