NLP Assignment 1

Flavio D'Amato , Luca Andaloro, Alessio La Torre, and **Lorenzo Venturi**
Master's Degree in Artificial Intelligence, University of Bologna
{ flavio.damato, luca.andaloro2, alessio.latorre, lorenzo.venturi14 }@studio.unibo.it

Abstract

This project tackles the EXIST 2023 Task 2 on Source Intention, which aims to categorize sexist messages according to the author's intention. We approached the task using two distinct architectures: Bidirectional LSTM (with one- and two-layer stacked variants) and a pretrained roBERTa transformer specialized in hate-speech detection. For both architectures we applied class weighting in order to mitigate the effects of the heavy class imbalance. Our results reveal a clear performance gap between the two approaches. The transformer model, even with a light fine-tuning on the task dataset, leverages its rich pretrained language representation, outperforming both LSTM variants. A qualitative analysis of common errors performing ensembling highlighted recurring issues, including limited training data of minority classes, text-cleaning steps that removed in particular cases semantic cues as quotation marks, missing contextual information in the tweets, specific unwanted OOV tokens due to syntax errors and difficulty in capturing irony.

1 Introduction

This report assesses the ability of both recurrent network and attention-based pretrained model to classify English tweets into 4 intention categories (non-sexist, direct, reported, judgemental). State-of-the-art approaches rely on pretrained, fine-tuned Transformers, as their broad linguistic knowledge helps mitigate the effect of imbalanced small datasets. We implemented a complete NLP pipeline, starting from the inspection of the dataset and a detailed data cleaning, as tweets often contain noisy or false informations. Each model was trained and evaluated across three random seeds. The experiments were conducted on a smaller version of the EXIST 2023 dataset, where tweets were annotated by six people as either non-sexist or as part of one of three sexist intention categories. The transformer architecture clearly outperformed the LSTM models, achieving over 0.1 higher macro F1-score on the test set, highlighting how an architecture with more parameters and pretrained on a large corpus can better capture semantics.

2 System description

We defined a complete pipeline for evaluating the different models. We started with a preprocessing phase in which we kept only English sentences and discarded items without a clear majority in the classification of the intention of the tweet. We then cleaned the text by removing hashtags, emojis, URLs, and special characters, followed by lemmatization in order to reduce all the terms to their base form. A vocabulary was created based on the training dataset and we also tried concatenating GloVe vocabulary reducing OOV terms, but no clear improvement was found. In order to build the embedding matrix, we used pretrained GloVe embeddings (Pennington et al., 2014). For handling OOV tokens, we experimented with both random initialization and neighbor-based method using top-k similar tokens, but saw no particular improvement in the second implementation. We fixed the embedding dimension to 50, as larger sizes did not provide meaningful gains while increasing training time. We implemented two LSTM-based models in PyTorch:

- Baseline model: single bidirectional LSTM with a dropout layer and a dense layer
- Stacked model: as the Baseline model with an additional LSTM layer

Finally, we imported and fine-tuned a transformer-based model, in particular twitter-roberta-base-hate (NLP, 2021), which was pretrained on a large corpus of $\sim$58M tweets. To preserve pretraining knowledge, we froze 8 out of its 12 layers, this was shown to reduce the overfitting quite considerably and led us to train for a few more epochs.

3 Experimental setup and results

For all the architectures, we fine-tuned the parameters in order to reduce overfitting and maximize the macro f1-score. For the LSTM models, we kept the same hyperparameters for both variants in order to better highlight their performance differences. In particular, we set: `lr=1e-4`, `batch_size=64`, `epochs=120`. For the transformer model, we fine-tuned it while freezing the first 6 layers, using `lr=2e-5`, `epochs= 5`, `batch_size=16`, `weight_decay= 0.01`. For the models, we applied a class weighting based on the square root of the inverse class frequency, a less accentuated method to highlight the loss of underrepresented classes, since balanced weighting was making LSTM overshoot on smaller classes.

Table 1: Mean validation set scores of the models across 3 random seeds

Model	Precision	Recall	F1-Score
Baseline	0.40 ± 0.03	0.44 ± 0.06	0.42 ± 0.04
Stacked	0.53 ± 0.07	0.44 ± 0.03	0.45 ± 0.03
Transformer	0.62 ± 0.03	0.61 ± 0.02	0.61 ± 0.01

Table 2: Mean test set scores of the models across 3 random seeds

Model	Precision	Recall	F1-Score
Baseline	0.32 ± 0.01	0.35 ± 0.02	0.33 ± 0.01
Stacked	0.39 ± 0.03	0.39 ± 0.02	0.38 ± 0.03
Transformer	0.52 ± 0.03	0.53 ± 0.02	0.52 ± 0.03

During error analysis, ensembling via highest confidence led to a slight improvement in performances.

4 Discussion

The main challenge of this task lies in the severe class imbalance: over 70% training samples are non-sexist, direct sexist reaches $\sim$18%, while reported and judgmental are both under 7%. All the models were trained with class weighting, yet, as shown in Table 2, transformer outperforms both LSTM variants, thanks to a big pretraining phase, giving it strong inductive biases and balanced semantic priors. Regarding LSTMs, we can see particularly in Table 1 some signs of overfitting passing from baseline to stacked, as between the two classes we see the major improvement on precision with respect to recall (the stacked gets more confident on examples it predicts correctly but fails on minority classes). Regarding the error analysis, we identified recurrent patterns behind misclassifications. Sentence length showed no correlation with the model errors. Looking at the number of OOV tokens, they did not exhibit linear correlation with mispredictions, although for many cases errors were linked to some unknown domain specific sexist terms as *phallocentrism* or to syntax errors that created UNK tokens from very useful terms, for instance the sentence *" 'gender harassment'.Female."* becomes *"gender <UNK>"* after cleaning because of missing spaces, leading LSTM models to fail on this input. Looking at common misclassified sentences between the models we found interesting patterns. The current text cleaning pipeline eliminates for some specific cases important semantics tokens, for example *quotation marks*, which are crucial for identifying reported speech, an example is *" "get changed, you look like a prostitute." "* which without the quotation marks is misclassified as direct sexism. Some ironic sentences also contribute to model errors, for example the sentence *"Calling A Man Bald Is Sexual Harassment"* is predicted as judgmental, but here and in many more cases even annotators' decision disagreed, we could define these cases as hard sentences. Emoji could also contain informative content, like symbol ♀ could be used to represent the target gender, and eliminating it removes important meaning. Finally, some tweets refers to external resources like videos or other posts, and without them classification is an ill-posed problem. Regarding future improvements, a tuned text-cleaning phase should be tested trying to reduce the errors cited and data augmentation could be introduced, trying to reduce the effects of the heavy class imbalance.

5 Conclusion

Identifying tweets is a challenging task, the transformer model outperform LSTM solutions as its larger capacity and the pretraining provide better internal biases. Class weighting is essential to overcome class imbalance, though it could be improved by adding data augmentation. A fine-grained data cleaning step should be developed, as tweets contain syntactic errors that introduce noise, finally, adding to the context cited external resources may help mitigate several of the remaining errors.

References

Cardiff NLP. 2021. cardiffnlp/twitter-roberta-base-hate model card. https://huggingface.co/cardiffnlp/twitter-roberta-base-hate.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.