Assignment 1

Mattia Buzzoni, Mirko Mornelli and **Riccardo Romeo**
Master's Degree in Artificial Intelligence, University of Bologna
{ mattia.buzzoni, mirko.mornelli, riccardo.romeo }@studio.unibo.it

Abstract

Comparative analysis of neural network architectures for tweet classification: evaluating LSTM and Transformer models across English and Spanish datasets, with a focus on preprocessing techniques including spell correction.

1 Introduction

This report investigates the performance of two architectural approaches—recurrent networks (LSTMs) and attention-based models (Transformers)—in classifying tweets across English and Spanish datasets. We implemented two preprocessing methods, including spell correction, and analyzed the impact of these techniques on model performance. Our findings indicate that Transformer models consistently outperformed LSTMs, achieving an average 10% higher F1-Score and lower Binary Crossentropy values. Despite expectations, spell correction did not enhance performance. Notably, all models struggled with Spanish tweets, likely due to challenges in tokenization, slang, and cultural context. We propose modifications to the tokenization process to retain more information, such as using descriptive captions for emojis and maintaining hashtags within the text.

2 System description

We downloaded the dataset and encoded it as Pandas DataFrames, subsequently cleaning the tweets. A vocabulary was created based on the training dataset, and we utilized GloVe embeddings (Pennington et al., 2014) with a dimensionality of 100. All out-of-vocabulary (OOV) terms were added to the embedding matrix, where they were assigned random embeddings.

Next, we implemented the BiLSTM model and the BiLSTM+LSTM architecture using Keras and TensorFlow. Additionally, we obtained several Transformer-based models from Hugging Face: RoBERTa, trained for hate speech detection; a fine-tuned BERTweet (Dat Quoc Nguyen and Nguyen, 2020) specifically for sexism detection; and DeHateBert (Sai Saketh Aluru and Mukherjee, 2020), a fine-tuned model developed for hate detection in Spanish texts.

3 Experimental setup and results

We manually tuned the hyperparameters for all the models utilized in our study. Specifically, we selected a learning rate of 1×10^{-3} and the AdamW optimizer for the LSTM models. We opted not to include dropout layers, as they did not yield better results. The models were trained for a maximum of 30 epochs. Notably, we allowed the initial embedding layer of the LSTM models to be trainable, as this improved performance.

In contrast, the Transformer models we downloaded already included specific dropout values, so we only adjusted the learning rate to 2×10^{-5}. For all models, including the Transformer-based ones, we computed the macro-averaged F1 score and used Binary Crossentropy as the loss function.

We also implemented two types of preprocessing. The first followed the guidelines outlined in the assignment, while the second was tailored to the nature of the tweet-based dataset. In the second data-cleaning process, we performed spell correction on the tweets, based on the method described in Peter Norvig's blog here, using TextBlob for implementation.

Finally, we decided to incorporate the Spanish dataset, training the LSTM models on Spanish tweets and comparing their performance with that of DeHateBert.

4 Discussion

We observed a modest imbalance in the training dataset, with 60.4% of the tweets classified as non-sexist. Additionally, the dataset is relatively small, containing only 2,870 elements, factors that may

Architecture	Params	Activation	LR	WD	Epochs	Loss Val.	F1 Val.	F1 Test
Bi-LSTM + Dense	3.2M	sigmoid	1e-3	4e-3	30	1.18	0.77	0.75
Bi-LSTM + LSTM + Dense	4.8M	sigmoid	1e-3	4e-3	30	1.63	0.72	0.74
RoBERTa (hate detection)	125M	gelu	2e-5	0.01	1	0.34	0.88	0.85
BERTweet (sexism detection)	355M	gelu+linear	2e-5	0.01	2	0.37	0.90	0.85

Table 1: Results for every architecture on english dataset

Architecture	Params	Activation	LR	WD	Epochs	Loss Val.	F1 Val.	F1 Test
Bi-LSTM + Dense	3.2M	sigmoid	1e-3	4e-3	30	1.73	0.71	0.71
Bi-LSTM + LSTM + Dense	4.8M	sigmoid	1e-3	4e-3	30	2.41	0.71	0.70
DeHateBert	167M	tanh	2e-5	0.01	2	0.45	0.80	0.79

Table 2: Results for every architecture on spanish dataset

contribute to the low performance of the models. The implementation of spell correction did not yield better results, although it did reduce vocabulary size, leading to slightly faster training. However, the process was lengthy and ineffective for LSTM models.

All Transformer-based models outperformed the LSTMs in terms of Binary Crossentropy and F1-Score. We noted that the models struggled with classifying ironic tweets and those containing offensive language. Transformers demonstrated better performance in classifying some ironic tweets but struggled with extremely short and noisy tweets due to significant information loss in the tokenization process. For example, the tweet: *"BOUNCEEYYYY BOOBIEEEESSSSSS #hookup #BOUNCEY #boobs #boobie #tits"* is cleaned to: *"bounceeyyyy boobieeeesssssss"*, and tokenized to: *"[UNK] [UNK]"*.

Both LSTM models and the Transformer-based model DeHateBert did not perform well on the Spanish dataset, suggesting that the Spanish language presents intrinsic complexities compared to English. To enhance model performance, we recommend using an automated hyperparameter fine-tuner and addressing dataset imbalance through data augmentation. Finally, preprocessing tweets to retain more information about emojis and hashtags is suggested.

5 Conclusion

In this assignment, we evaluated the performance of recurrent networks (LSTMs) and attention-based models (Transformers) for tweet classification across English and Spanish datasets. Our results indicated that Transformer models consistently outperformed LSTMs, achieving an average of 10% higher F1-Score and lower Binary Crossentropy values. Interestingly, the implementation of spell correction did not enhance performance as anticipated, and all models struggled more with Spanish tweets due to factors like tokenization challenges and cultural context.

We identified the tokenization process as a primary limitation, suggesting that modifications could improve results. Potential solutions include retaining more information during data cleaning by using descriptive captions for emojis, preserving hashtags, and introducing special tokens for user mentions.

6 Links to external resources

- TextBlob
- RoBERTa
- BERTweet
- DeHateBert

References

Thanh Vu Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets.

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.

Punyajoy Saha Sai Saketh Aluru, Binny Mathew and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection.