Sentiment Analysis for FIFA 2022 Tweets: Comparative Study of Dense, RNN and LSTM Models

Oscar Muñoz Ramirez - Sebastian Erazo - Juan Diego Lora

1. Introduction

This study aims to classify tweets from the first day of the FIFA World Cup 2022 into three sentiment categories: positive, neutral, and negative. We explored three neural network architectures—Dense Neural Networks (DenseNN), Vanilla RNN (SimpleRNN), and LSTM—evaluating their performance with and without hyperparameter tuning via GridSearchCV.

2. Models Without GridSearch

DenseNN (Baseline)

- Accuracy: ~0.74
  F1-score (weighted): ~0.73

The Dense Neural Network served as the baseline model and delivered decent performance. However, it showed high variance across classes. The confusion matrix revealed that the model frequently misclassified neutral tweets as positive, indicating a bias. Additionally, signs of overfitting were noticeable as the training progressed, suggesting that while it learned well initially, it struggled to generalize during validation.

SimpleRNN (Vanilla RNN)

- Accuracy: 0.70
  F1-score (weighted): 0.70
  Kappa: 0.66

The SimpleRNN model performed slightly below the DenseNN in terms of accuracy, but it offered a notable improvement in temporal sequence modeling. It trained quickly and was able to represent word order better, but its validation loss began fluctuating after epoch 5, which could point to training instability or the need for better regularization. It still demonstrated good capacity to generalize, especially with class balance taken into account.

LSTM

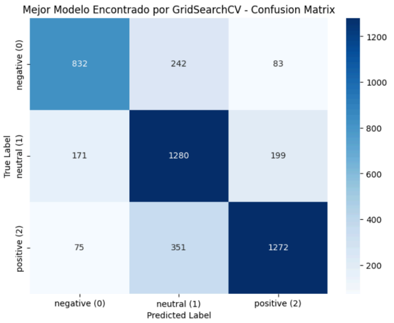
- Accuracy: 0.7376
  F1-score: 0.7376
  Kappa: 0.6015

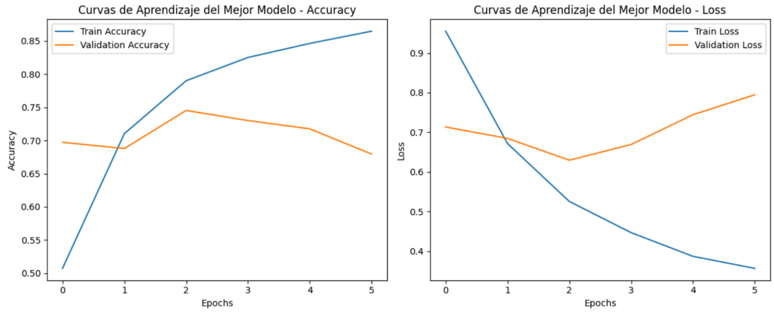
The LSTM model provided the most balanced results among all non-tuned models. Its architecture enabled it to capture long-term dependencies in the tweet sequences more effectively. The learning curves were stable across epochs, and it managed to retain solid class-wise performance without favoring any specific sentiment. As such, LSTM stood out as the most promising base model even before hyperparameter tuning.

3. Models With GridSearchCV

DenseNN + GridSearch

- Accuracy: ~0.75
  F1-score (weighted): ~0.74
  Kappa: ~0.62

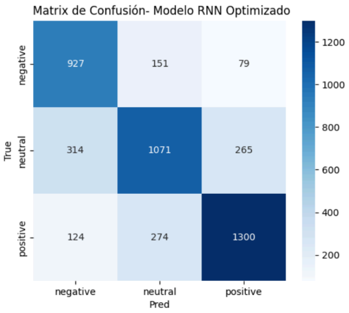


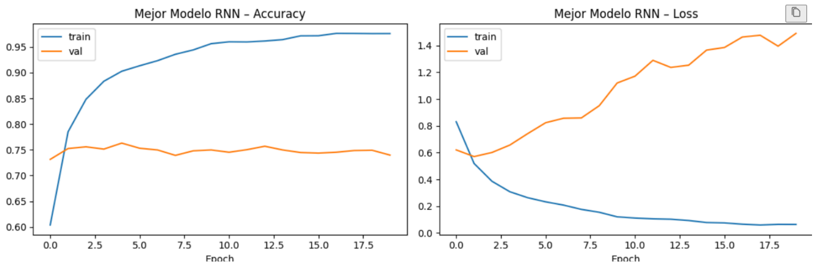


With hyperparameter tuning, the Dense Neural Network showed a measurable improvement in overall performance. Using 128 neurons in the first dense layer, a dropout rate of 0.4, and the RMSprop optimizer, the model achieved better balance between training and validation performance. The confusion matrix demonstrated a clearer separation between classes, especially with improved prediction accuracy for neutral tweets—a weakness in the untuned version. This highlights the value of tuning even in relatively shallow models.

SimpleRNN + GridSearch

- Accuracy: 0.7258
  Precision: 0.72
  Recall: 0.71
  F1-score: 0.715
  Kappa: 0.68



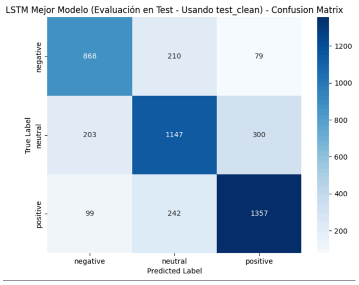


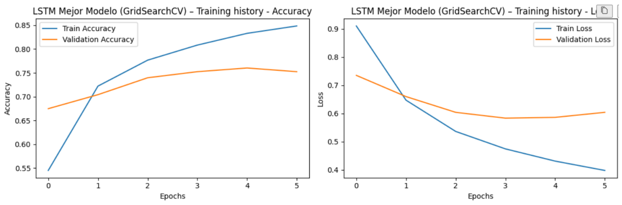
GridSearchCV significantly stabilized the SimpleRNN's performance. With 128 hidden units, a dropout of 0.5, and the RMSprop optimizer, the model achieved greater consistency and generalization. The precision and recall were balanced (0.72 and 0.71 respectively), and the model exhibited better class separation—particularly reducing false positives in the "positive" category. While it did not surpass LSTM in absolute terms, it closed the performance gap and became a viable lightweight alternative.

LSTM + GridSearch

- Accuracy: 0.7485
  F1-score (weighted): 0.7481
  Kappa: 0.6178

- Best Params: dropout=0.5, embed_dim=64, lstm_units=64, optimizer='rmsprop', batch_size=64, epochs=6





Confusion Matrix:

Negative: F1 = 0.75
Neutral: F1 = 0.71
Positive: F1 = 0.79

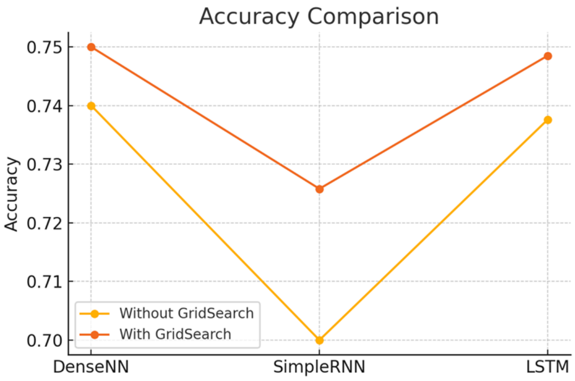
When optimized using GridSearchCV, achieved the best results overall. The selected configuration—64 LSTM units, 64-dimensional embeddings, 0.5 dropout, RMSprop optimizer, and a batch size of 64—led to the most balanced and accurate model across all metrics. The class-wise F1-scores were impressive: 0.75 for negative, 0.71 for neutral, and 0.79 for positive sentiment. Not only did it outperform in raw metrics, but it also showed excellent generalization and robustness, making it the most effective model for this classification task.

4. Comparative Analysis

Model	GridSearch	Accuracy	F1-Score	Kappa	Best Feature
Dummy	No	~0.33	~0.00	~0.00	Give us a baseline
DenseNN	No	~0.74	~0.73	-0.59	Fast and easy to train

DenseNN	Yes	~0.75	~0.74	~0.62	Better balance after tuning
SimpleRNN	No	0.70	0.70	0.66	Good recall, simpler model
SimpleRNN	Yes	0.73	0.715	0.68	Improved generalization
LSTM	No	0.7376	0.7376	0.6015	Strong performance, robust
LSTM	Yes	**0.7485**	**0.7481**	**0.6178**	**Best overall performance**

Visualizations of confusion matrices and training/validation curves confirm the superiority of LSTM with tuned parameters, especially in reducing imbalance errors and stabilizing loss.



5. Conclusion

LSTM with hyperparameter optimization achieved the best performance in sentiment classification of FIFA 2022 tweets, with high accuracy, balanced F1-scores, and the strongest Kappa index. GridSearchCV significantly improved model robustness across all architectures, confirming its importance for real-world NLP tasks.

Future Work: Integrate pretrained embeddings (e.g., GloVe), test on multilingual data, and extend the analysis with attention-based transformers like BERT.

6. References

Chollet, F. (2017). *Deep Learning with Python*. Manning Publications.

Goldberg, Y. (2016). *A Primer on Neural Network Models for Natural Language Processing*.

Journal of Artificial Intelligence Research, 57, 345–420.

Brownlee, J. (2017). *Sequence Prediction with Deep Learning*. Machine Learning Mastery.

Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine

Learning Research, 12, 2825–2830.