12/04/2024
Soichiro Tanabe

# Enhancing Sentiment Analysis Performance with LSTM Models: A Comparative Study

## Background and context

Sentiment analysis serves as a cognitive tool for discerning the emotional sentiment conveyed within a text, focusing on categorizing and forecasting users' sentiments and opinions expressed in reviews (Mahadevaswamy and Swathi, 2023). Alaba et al. (2018) state that deep neural networks have shown good performance in sentiment analysis. In particular, LSTM (Long Short-Term Memory) networks have the ability to effectively learn from sequential data by considering the information from previous outputs (Gers et al., 2002), which is considered to be ideal for sentiment analysis. In fact, various studies of sentiment analysis have been conducted using LSTM. Alaba et al. (2018) used the combination of CNN and LSTM for sentiment analysis, demonstrating that a CNN-LSTM model performed better than previous sentiment analysis models. Despite achieving notable accuracy improvements by analysts using a CNN-LSTM model, there also remained room for further enhancing sentiment classification accuracy. For example, Mahadevaswamy and Swathi (2023) focused on Bidirectional-LSTM, which resulted in higher accuracy than LSTM.

## Aim and Objectives

For the previously mentioned studies, comparisons among models were conducted to evaluate how well LSTM models perform. Alaba et al. (2018) compared LSTM models with previous machine learning models, while Mahadevaswamy and Swathi (2023) compared Bi-LSTM, LSTM, and CNN models. However, it was expected that changing hyperparameters, such as the number of LSTM modules, would impact performance. Therefore, research was conducted on how performance changes according to varying the number of LSTM modules.

Additionally, it was also considered that performance depends on how the model is trained in LSTM. For example, Alaba et al. (2018), who conducted sentiment analysis by training different analysis levels (Character Level, Character N-Gram Level, and Word Level), demonstrated different results. Likewise, research was conducted on the amount of vocabulary used for training and the number of frequent words stored.

## Methods

To conduct sentiment analysis using an LSTM model, the Amazon review dataset from Kaggle (https://www.kaggle.com/datasets/kritanjalijain/amazon-reviews) was used. The dataset has already been adapted into ratings (0: negative, 1: positive). After importing training, validation, and test data, new training datasets of sizes 10%, 25%, and 50% were created to evaluate the impact of data volume. These new training datasets, along with a full training dataset, were applied to the LSTM model.

Additionally, encoders with 100-word, 250-word, and 500-word vocabularies were created to investigate whether the number of stored frequent words impacts LSTM models.

Firstly, the impact of dataset size on performance was investigated. LSTM models were trained using 10%, 25%, and 50% of the training data. For each model, the following training hyperparameters were used:

• Use 4 LSTM modules
• Use a 100-word encoder
• Train for 15 epochs

Secondly, the effect of the number of LSTM modules on model performance was examined. Models with 8 LSTM units and 12 LSTM units were created and trained. For each model, the following training parameters were utilised:

• Create and use a 250-word encoder
• Use 25% of the training data
• Train for 15 epochs

Thirdly, the investigation focused on how the size of the encoder's vocabulary impacts model performance. Models using encoders with 250-word and 500-word vocabularies were created. For each model, the following training parameters were used:

• Utilise a model with 4 LSTM modules
• Use 25% of the training data
• Train for 15 epochs

Finally, the best model was considered based on experiments regarding dataset size, the number of LSTM modules, and the size of the encoder's vocabulary. Hyperparameters were set based on previous experiments.

## Results

**The size of dataset investigation:**

| | ACCURACY | PRECISION | RECALL |
|---|---|---|---|
| 10% | 0.7533 | 0.7514 | 0.7574 |
| 25% | 0.7621 | 0.7738 | 0.7410 |
| 50% | 0.7705 | 0.7508 | 0.8101 |

Analysing the classification metrics, as the percentage of the training dataset usage increased, the accuracy also increased. Specifically, the models trained with 10%, 25%, and 50% of the training dataset achieved accuracy percentages of 0.7533, 0.7621, and 0.7705, respectively.

However, when considering precision, the model trained with 25% of the training dataset showed the highest percentage at 0.7738. On the other hand, the recall metric reached its highest value of 0.8101 for the model trained with 50% of the training dataset. Consequently, the model trained with 50% of the dataset yielded the best performance overall.

**The number of LSTM modules investigation:**

| | ACCURACY | PRECISION | RECALL |
|---|---|---|---|
| 8 LSTM | 0.8423 | 0.8342 | 0.8546 |
| 12 LSTM | 0.8454 | 0.8721 | 0.8096 |

Regarding accuracy and precision, training with 12 LSTM modules showed slightly better results compared to training with 8 LSTM modules. The accuracy and precision values were 0.8454 and 0.8721, respectively, with 12 LSTM, while they are 0.8423 and 0.8342 with 8 LSTM modules.

However, in terms of recall, 8 LSTM modules achieved a better score than 12 LSTM modules (8 LSTM: 0.8546, 12 LSTM: 0.8096). In conclusion, from the perspective of accuracy, increasing the number of LSTM modules tends to improve performance.

**The size of the encoder's vocabulary investigation:**

| | ACCURACY | PRECISION | RECALL |
|---|---|---|---|
| 250 VOCAB | 0.8237 | 0.8047 | 0.8550 |
| 500 VOCAB | 0.8656 | 0.8436 | 0.8977 |

Comparing accuracy, precision, and recall, training with a vocabulary length of 500 performed better than training with a vocabulary length of 250. While the accuracy, precision, and recall with a vocabulary length of 250 were 0.8237, 0.8047, and 0.8550, respectively, those with a vocabulary length of 500 were 0.8237, 0.8047, and 0.8550, respectively. Therefore, increasing the vocabulary length resulted in better performance.

**Considered model:**

The experiments so far suggested that increasing the volume of the dataset, the number of LSTM modules, and the length of the vocabulary for an encoder leads to better performance. Therefore, it was expected that the best model is as follows:

- 12 LSTM modules
- an encoder with a 500-word vocabulary
- Full dataset

This was the result of the model using mentioned hyperparameters.

| | ACCURACY | PRECISION | RECALL |
|---|---|---|---|
| 12 LSTM /500 VOCAB /FULL DATA | 0.8935 | 0.9057 | 0.8785 |

As anticipated, this model exhibited the best performance in terms of accuracy, achieving a score of 0.8935. Additionally, precision also achieved high scores with 0.9057.

## Evaluation of results

Through the experiments, it was discovered that the number of LSTM modules, the volume of training data, and the size of the encoder's vocabulary impact the performance. Accuracy improved with a larger dataset volume, a higher number of LSTM modules, and a larger encoder size. The highest accuracy achieved was 89.35%, attained with the following settings: 12 LSTM modules, an encoder with a 500-word vocabulary, and the full dataset. This surpassed the LSTM result obtained by Mahadevaswamy and Swathi (2023), which was 85.75%.

## Discussion of wider implications

Even with consistent hyperparameters and dataset volume, LSTM models trained on different datasets may yield varying performances. Consequently, there are limitations to comparing models with the LSTM model proposed by Mahadevaswamy and Swathi (2023). Additionally, Mahadevaswamy and Swathi (2023) achieved 91.4% accuracy using Bi-LSTM. Therefore, there is room for improvement through further experiments using Bi-LSTM.

Furthermore, linguistic complexity and cultural nuances demand a nuanced approach to dataset selection for sentiment analysis, as demonstrated by Alaba et al. (2018) in their sentiment analysis of the Arabic language. It is crucial to capture the subtleties of sentiment expression accurately. In this study, an English dataset was trained and directly applied to LSTM, while three different analysis levels were employed for Arabic sentiment analysis. Language differences can significantly impact model performance.

## Conclusion

This study investigated the impact of various factors on the performance of LSTM models in sentiment analysis. It was revealed that the number of LSTM modules, the volume of training data, and the size of the encoder's vocabulary significantly influence accuracy. Notably, higher accuracy was achieved with a larger dataset volume, increased number of LSTM modules, and a larger encoder size, with the optimal configuration yielding an accuracy of 89.35%.

Comparison with prior research conducted by Mahadevaswamy and Swathi (2023) demonstrated that Bi-LSTM models outperformed traditional LSTM models, indicating the potential for further enhancements in model performance. Therefore, future experiments will explore the application of Bi-LSTM models using the Amazon dataset to validate these findings.

As Alaba et al. (2018) demonstrated the necessity of tailored approaches for sentiment analysis, it is important to consider linguistic complexity and cultural nuances in dataset selection for sentiment analysis. Therefore, not only should future investigations continue to explore advanced model architectures like Bi-LSTM, but they should also explore optimal approaches to languages in dataset to enhance the accuracy and utility of sentiment analysis systems.

## Reference


Alayba, A.M., Palade, V., England, M. and Iqbal, R. (2018) 'A Combined CNN and LSTM Model for Arabic Sentiment Analysis', *Machine Learning and Knowledge Extraction*, pp. 179–191. Available at: https://doi.org/10.1007/978-3-319-99740-7_12 (Accessed 10 April 2024)..

Gers, F.A., Eck, D and Schmidhuber, J. (2002) ' Applying LSTM to time series predictable through time-window approaches', *Perspectives in Neural Computing 2002*, vol. 9999, pp. 193–200. Available at: https://doi.org/10.1007/978-1-4471-0219-9_20 (Accessed 10 April 2024).

Mahadevaswamy, U.B. and Swathi, P. (2023) 'Sentiment Analysis using Bidirectional LSTM Network', *Procedia Computer Science*, 218, pp. 45–56. Available at: https://doi.org/10.1016/j.procs.2022.12.400 (Accessed 10 April 2024).