Analyzing Financial Sentiment Regarding the Number of Neurons, Drop-out Value, & Max Sequence Length Value in LSTMs

ABSTRACT: Long Short-Term Memory Networks (LSTMs) are deep-learning architectures based on Recurrent Neural Networks. LSTMs are one of the most versatile deep learning architectures, with use cases ranging from language modeling to video recognition. In this study, three variables that influence the accuracy of LSTMs when predicting sentiment are investigated, the number of neurons, dropout value, and the max sequence length. It is hypothesized that a combination of 100 neurons, a max sequence length value of 256, and a dropout value of 0.2 will yield the most accurate results. A LSTM is implemented in the python programming language to predict sentiment and test this hypothesis. The results of the experiment showed that a combination of a drop-out value of 0, a neuron count of 100, and a max sequence length value of either 64 or 128 could most accurately analyze financial sentiment with an accuracy value of 90.75%. This research opens new fields of LSTMs for further research, such as, integrating LSTMs with existing architectures to develop innovative solutions, or exploring multi-layered LSTMs and unsupervised training.

KEYWORDS: LSTM, financial sentiment, sentiment, dropout value, max sequence length, neuron

1. INTRODUCTION: Computer science is the main driver of technological and worldwide development into a more advanced society. Due to the vast use cases of products of computer science such as, AI, VR, DL, LSTMs, etc., our society is fully dependent on computer science and its results. Consequently, there continues to be a constant venture to improve upon our

present developments in computer science by innovating new methods and educating the next generation.[6] One recent advancement in the computer science field is a Long Short-Term Memory Network (LSTM). LSTMs have a wide range of applicability, such as, speech recognition, language translation, and is even applicable in the gaming world.[7] The seemingly limitless potential of LSTMs raises the curiosity of many regarding the implementation of the algorithm in the financial sector. LSTMs can predict fluctuations in the financial field through sentiment analysis. Sentiment, or rather investors' sentiment, is one of the main factors that influences the frequent changes of the financial market.[4] Therefore the analysis of financial sentiment through machine learning is highly valuable as it has immense potential in the financial field. There are many factors that influence the accuracy of LSTMs when analyzing sentiment. In this study, we will attempt to optimize the accuracy of a one-layered LSTM when analyzing sentiment into three categories; positive, negative, and neutral. We will utilize three different variables to optimize the accuracy of the LSTM, the number of neurons, the dropout value, and the max sequence length of the LSTM. It is hypothesized that a combination of 100 neurons, a max sequence length value of 256, and a drop-out value of 0.2 will yield the most accurate results.

2. MATERIALS AND METHODS: To test our hypothesis, we tested a multitude of combinations of different values of the three variables investigated by compiling, training, and testing a single-layered LSTM. In this study, two datasets both consisting of financial statements, were used to train a single-layered LSTM. Firstly, we cleaned both datasets by using the text module built into Python. Using this module, we removed stop words and punctuation to ensure the data would be suitable for the LSTM to analyze. We then proceeded to use the Natural

Language Toolkit, scikit-learn, and the Keras modules in Python to tokenize and capture a numerical representation of the training dataset. The Keras module allowed us to compile the LSTM and change two of the three variables investigated – the number of neurons and the drop-out value of the LSTM. The scikit-learn module allowed us to set the max sequence length by creating a vector to store an iterative representation of financial statements. However, this results in the issue of shorter and longer sentences being different lengths due to the different number of words present in the statement. To solve this problem, we used the Keras module to pad the statements with empty spaces to such a way that all sentences are the same length when analyzed by the LSTM. For the LSTM to classify the financial statements into distinct categories, we created a label map that included three categories – positive, negative, and neutral. This allows the LSTM to categorize different financial statements in the testing dataset into their respective categories. To test many combinations of our three variables, we used four different values of each. This included the following values: Number of Neurons: 50, 100, 200, 500; Max Sequence Length: 64, 128, 256, 512; Drop-out Value: 0, 0.2, 0.3, 0.4. We used all possible combinations of these 12 values, leading to 64 different experiments.

3. RESULTS:

Max Sequence Length

	64	128	256	512
0.4	83.70%	83.26%	84.14%	86.34%
0.3	84.58%	85.02%	84.14%	85.02%
0.2	84.14%	82.82%	84.14%	85.90%
0	83.70%	83.70%	83.70%	85.90%

Figure 1 – 50 Neuron Testing

Max Sequence Length

	64	128	256	512
0.4	88.11%	90.31%	89.43%	84.14%

0.3	87.67%	88.11%	89.87%	84.58%
0.2	90.31%	89.43%	89.87%	82.38%
0	90.75%	90.75%	89.87%	83.70%

Figure 2 – 100 Neuron Testing

Max Sequence Length

	64	128	256	512
0.4	83.26%	84.58%	82.82%	84.58%
0.3	82.38%	84.14%	84.14%	85.02%
0.2	81.94%	82.38%	82.38%	83.70%
0	79.30%	85.02%	83.26%	81.94%

Figure 3 – 200 Neuron Testing

Max Sequence Length

	64	128	256	512
0.4	86.87%	81.94%	83.26%	81.06%
0.3	88.11%	82.38%	78.85%	83.26%
0.2	89.43%	87.50%	82.82%	82.38%
0	86.34%	84.14%	84.14%	82.82%

Figure 4 – 500 Neuron Testing

In these series of experiments various data points have been accumulated from various experiments across the three variables investigated, the number of neurons, max sequence length, and the drop-out value. There are four figures that describe the data collected from four single-layered LSTMs trained on the parameters specified. **Figure** 1 described a single-layered LSTM made of 50 neurons, **Figure 2** described a single-layered LSTM made of 100 neurons, **Figure 3** describes a single-layered LSTM made of 200 neurons, and **Figure 4** describes a single-layered LSTM made of 500 neurons. As shown in **Figure 1**, a max sequence length value of 512 coupled with a drop-out value of 0.2 or 0 yielded the highest accuracy of 85.90% when utilizing a single-layered LSTM made of 50 neurons. However, in **Figure 2**, the highest accuracy of 90.75% is found in two different max sequence length values, 64 and 128. This accuracy was only found when there was no drop-out value present (0) in both max sequence length values when a single-layered LSTM made of 100 neurons was tested. The data collected also shows that **Figure 3,** a single-layered LSTM made of 200 neurons, performed the worst when testing in terms of accuracy. It is also shown that the data collected from the LSTM in **Figure 2** has the highest accuracies overall when compared to the other LSTMs and their respective data. The highest accuracy of all the LSTMs was found in **Figure** 2, with a combination of max sequence length values of 64 and 128, and a drop-out value of 0. Therefore, different single-layered LSTMs with different amounts of neurons can have different levels of accuracies when compiled with different values of the max sequence length and drop-out value.

4. DISCUSSION: It was hypothesized that a combination of 100 neurons, a max sequence length value of 256, and a drop-out value of 0.2 will yield the most accurate results. This hypothesis was partially correct as **Figure 2**, a single-layered LSTM made of 100 neurons did yield the highest accuracy. However, this hypothesis was wrong as it hypothesized that a drop-out value of 0.2 and a max sequence length value of 256 would yield the highest accuracy when predicting financial sentiment. This was incorrect as when the drop-out value variable was eliminated (drop-out value = 0) and the max sequence length was set to 64 or 128, the single-layered LSTM trained on those parameters performed the best when analyzing the sentiment of financial statements.[7] Analyzing sentiment, especially investor's sentiment plays a significant role when predicting the financial market.[4]

5. CONCLUSION: In conclusion, with regards to the three variables investigated, a single layered LSTM made of 100 neurons trained with a max sequence length value of 64 or 128 and a drop-out value of 0 will yield the highest accuracy when testing the LSTM to categorize/analyze sentiment into three categories, positive, negative, and neutral from financial statements. The

continuation of the study of machine learning and LSTMs when analyzing sentiment is crucial to advance society and deep learning. On top of that, there are already new ways to advance this current technology arising, including unlabeled, unsupervised training data as a new method to train LSTMs when using them for their Natural Language Processing services (NLP).

6. ACKNOWLEDGEMENTS: I would like to thank my mentors Dr. Erik Sakk and Tim Adamson for helping me to write this article.

7. BIBLIOGRAPHY:

[1] J. Wang, L.-C. Yu, K. R. Lai, and X. Zhang, 'Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model', in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin, Germany: Association for Computational Linguistics, 2016, pp. 225–230. doi: 10.18653/v1/P16-2037.

[2] X. Huang et al., 'LSTM Based Sentiment Analysis for Cryptocurrency Prediction'. arXiv, Oct. 17, 2021. Accessed: Jul. 25, 2023. [Online]. Available: http://arxiv.org/abs/2103.14804

3]Y. Ma, H. Peng, T. Khan, E. Cambria, and A. Hussain, 'Sentic LSTM: a Hybrid Network for Targeted Aspect-Based Sentiment Analysis', Cogn Comput, vol. 10, no. 4, pp. 639–650, Aug. 2018, doi: 10.1007/s12559-018-9549-x.

Database

[4] J. Z. G. Hiew, X. Huang, H. Mou, D. Li, Q. Wu, and Y. Xu, ‘BERT-based Financial Sentiment Index and LSTM-based Stock Return Predictability’. arXiv, Jul. 07, 2022. Accessed: Jul. 25, 2023. [Online]. Available: http://arxiv.org/abs/1906.09024

[5] D. Araci, ‘FinBERT: Financial Sentiment Analysis with Pre-trained Language Models’. arXiv, Aug. 27, 2019. Accessed: Jul. 25, 2023. [Online]. Available: http://arxiv.org/abs/1908.10063

[6] C. Almelek and Ö. Güzeldereli, ‘Using Genetic Algorithms to Solve Computer Science Olympiad Optimization Problems’, IJHSR, vol. 4, no. 4, pp. 13–18, Aug. 2022, doi: 10.36838/v4i4.3.

[7] G. Van Houdt, C. Mosquera, and G. Nápoles, ‘A review on the long short-term memory model’, Artif Intell Rev, vol. 53, no. 8, pp. 5929–5955, Dec. 2020, doi: 10.1007/s10462-020-09838-1.