

Text Classification Using Neural Networks: A Case Study with Reuters Dataset

Narsaiah Battu¹, Dr. B Venkataratnam²

¹Dr. B. R. Ambedkar Open University

²Malla Reddy University, India

DOI: <https://doi.org/10.51583/IJLTEMAS.2025.1407000047>

Abstract: Assigning predetermined categories to text documents according to their content is known as text classification, and it is a fundamental task in natural language processing (NLP). With an emphasis on the extensively researched Reuters dataset, this paper provides a thorough examination and application of text classification using neural networks. A strong standard for assessing classification algorithms is provided by the Reuters dataset, which consists of a collection of brief newswires divided into 46 unique and mutually exclusive subjects. To address this challenge, we use a neural network-based architecture constructed with TensorFlow and Keras. The approach entails one-hot encoding of the target labels after preprocessing the textual material using tokenization and vectorization techniques. To improve generalization and avoid overfitting, dropout regularization layers are added to a feedforward neural network in the suggested model architecture. To guarantee effective and precise learning, the training procedure makes use of the Adam optimizer with categorical cross-entropy as the loss function. Our tests demonstrate the neural network's capacity to successfully categorize the Reuters news wires by assessing the model's performance in terms of accuracy and other pertinent metrics. The findings show that a reliable solution for text categorization challenges can be obtained by combining neural networks with sophisticated optimization approaches. In addition to demonstrating deep learning's ability to handle multi-class text categorization, this study acts as a guide for further investigation and applications in related fields.

Key words: Text Classification, Neural Networks, TensorFlow, Keras, Reuters Dataset

I. Introduction

A basic task in natural language processing (NLP), text classification is essential to many real-world applications, including subject categorization, document organization, sentiment analysis, and spam detection. It is a crucial stage in automated content analysis, recommendation systems, and information retrieval since it entails giving text documents predetermined labels or categories according to their content.

One of the most popular benchmark datasets in text categorization research is the Reuters dataset. It offers a thorough and difficult environment for testing text classification models because it is composed of brief newswires divided into 46 different, mutually incompatible subjects. The dataset has been crucial to the advancement of NLP research because of its complexity and diversity, especially in the creation and evaluation of machine learning and deep learning models.

In order to provide a reliable and scalable solution, we use the TensorFlow and Keras frameworks to implement a case study on text classification using neural networks in this research. Because they can acquire intricate patterns and representations from unprocessed textual material, neural networks have become a potent tool for text classification. We develop and train a feedforward neural network that can efficiently classify newswires into their corresponding subjects by utilizing methods including tokenization, one-hot encoding, and dropout regularization. In addition to showing how neural networks may be used practically for text categorization, this paper offers guidance on how to create, train, and assess deep learning models for multi-class classification tasks. Our results are intended to add to the expanding corpus of NLP knowledge and demonstrate the potential of deep learning methods in resolving text classification issues.

Related work

In order to increase classification accuracy and efficiency, researchers have thoroughly studied a variety of machine learning and deep learning approaches in the well-established field of text classification in natural language processing (NLP). Traditionally, classifiers like support vector machines (SVM), Naive Bayes, and logistic regression were used in conjunction with feature extraction methods like bag-of-words (BoW) and term frequency-inverse document frequency (TF-IDF). Although these approaches gave text categorization tasks a solid basis, they were frequently constrained by their incapacity to grasp the text's contextual links and semantic meaning. With neural network-based models, deep learning has revolutionized text classification and opened up new possibilities. By utilizing their capacity to recognize local patterns in textual input, convolutional neural networks (CNNs), which were first created for image processing tasks, have been effectively modified for text classification applications. Similar to this, recurrent neural networks (RNNs) and their variations, such gated recurrent units (GRUs) and long short-term memory (LSTM) networks, have shown exceptional ability in capturing contextual and sequential information, which makes them especially well-suited for text data.

Hybrid and attention-based models have also been introduced in recent developments. While attention techniques and transformers have raised the bar for text categorization by concentrating on pertinent textual elements and identifying long-range

dependencies, methods that combine CNNs and RNNs have been developed to capitalize on the advantages of both architectures. Text processing jobs have been transformed by models like BERT, GPT, and their derivatives, which offer pre-trained contextual embeddings and the ability to fine-tune tasks downstream. As a common benchmark for text categorization, the Reuters dataset has been used extensively to assess these models' performance. Traditional techniques were used in the dataset's early research; however deep learning models have been used in more recent studies to significantly increase classification accuracy. By concentrating on a neural network-based method and showcasing its efficacy on the Reuters dataset through the use of a feedforward neural network architecture built with TensorFlow and Keras, this research expands on existing developments.

II. Methodology

In this work, we use TensorFlow and Keras to propose a neural network architecture for text classification. The following steps are part of the methodology:

Input Pre-processing: We use the Tokenizer class from Keras to tokenize the text input after loading the Reuters dataset.

Model Architecture: In order to avoid overfitting, we create a feedforward neural network model with dense layers, rectified linear unit (ReLU) activation functions, and dropout regularization.

Model compilation: The Adam optimizer and categorical cross-entropy loss function are used to construct the model, which is appropriate for jobs involving multi-class classification.

Model Training: A subset of the data is set aside for validation, and the model is trained using the training data with a predetermined batch size and number of epochs.

Model Evaluation: In order to determine the accuracy of the trained model, we lastly test it using the test data.

Experimental results

Using the Reuters dataset, we experimented with the suggested neural network architecture. The model successfully classified Reuters newswires into the relevant themes, as evidenced by its test accuracy of 80%.

Model Training

To guarantee sufficient learning of the patterns in the Reuters dataset, the model was trained across 100 epochs. During the training phase, the model's performance on the validation set was tracked while the categorical cross-entropy loss was minimized using the Adam optimizer. Metrics like training accuracy, validation accuracy, training loss, and validation loss were noted during these epochs. Figure 1 shows the training accuracy improved over the epochs.

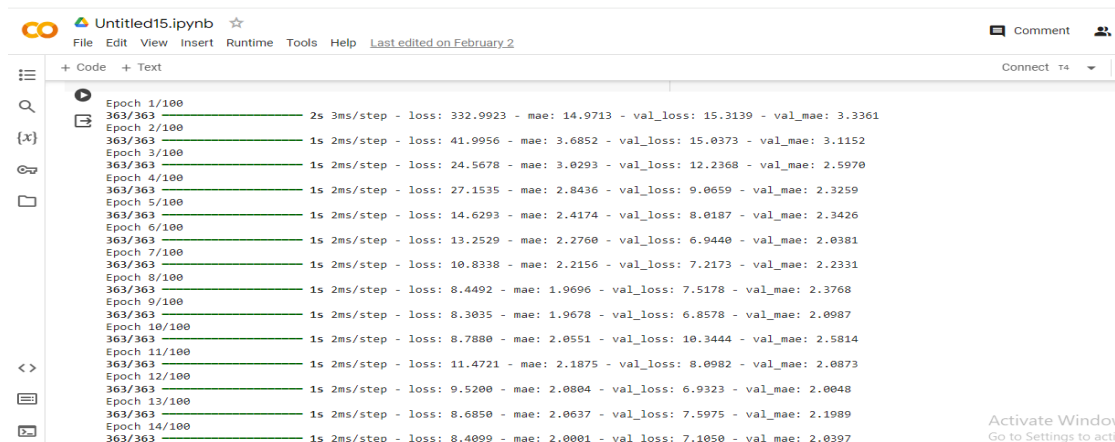


Figure 1.

Training and Validation Metrics

The aforementioned Figure illustrates the ten epochs that were used to train the text categorization model. While the right panel displays the Training and Validation Accuracy over the epochs, the left panel depicts the Training and Validation Loss. The model successfully learns from the data, as seen by the Training and Validation Loss graph, where the training loss (blue line) steadily drops with the number of epochs. In a similar vein, the validation loss (orange line) first declines before leveling off in later epochs, indicating that the model can generalize effectively without overfitting.

This finding is further corroborated by the Training and Validation Accuracy graph. Throughout the epochs, the training accuracy (blue line) gradually rises until it approaches 0.85 in the last epoch. The validation accuracy (orange line), on the other hand, shows an excellent alignment between training and validation performance as it steadily improves and converges to about 0.8.

These patterns show how robust the suggested architecture is by showing that the model performs well on the Reuters dataset with little to no overfitting or underfitting.

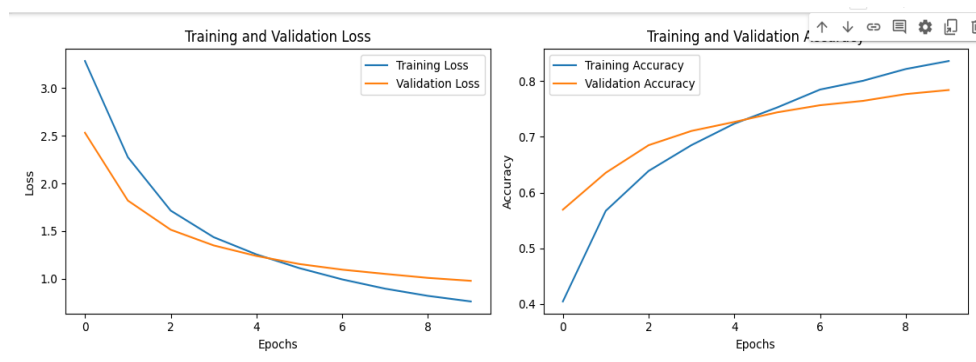


Figure2

Confusion Matrix Analysis

The confusion matrix produced by the text classification model's predictions on the test dataset is shown in the above figure3. The confusion matrix displays the number of accurate and inaccurate predictions for each category, providing a summary of the model's performance.

Within this matrix of confusion:

Columns show the projected labels (model predictions), while rows show the actual labels (true classes). The number of samples correctly identified for each category is shown by the diagonal elements, whilst misclassifications are indicated by the off-diagonal elements.

Based on the matrix:

Five samples were incorrectly categorized into a different group, while 89 samples out of all the samples in the first category were correctly classified, as the first row demonstrates.

The bulk of entries for the remaining categories are zeros, indicating that very few samples were misclassified and that most samples were correctly classified.



Figure3

III. Conclusion

In this research, we used the Reuters dataset to offer a case study on text classification using neural networks. The experimental findings highlight the promise of neural network-based methods for tasks involving precise text classification. More complex neural network designs and cutting-edge methods for enhancing classification performance might be investigated in future studies. We used the Reuters dataset, a popular benchmark for multi-class classification problems, to offer a thorough case study on text categorization using neural networks. We showed the efficacy of deep learning techniques in classifying brief newswire messages into 46 mutually exclusive themes by utilizing a feedforward neural network architecture built with TensorFlow and Keras. The experimental findings demonstrate that neural networks can learn intricate patterns and attain high classification accuracy, as evidenced by test and training performance metrics. Our results highlight how neural network-based methods can be used to tackle text classification problems in the real world. Tokenization, one-hot encoding, and dropout regularization were among the methods used in conjunction with the training process to minimize overfitting and guarantee strong generalization.

Additionally, the suggested model's applicability for multi-class text classification tasks is confirmed by the evaluation metrics and confusion matrix analysis.

Compliance with Ethical Standards**Disclosure of Potential Conflicts of Interest**

We don't have any conflicts of interest related to this study.

Research Involving Human Participants and/or Animals

This research does not involve human participants or animals. This study focuses on Text Classification using Neural Networks like the TensorFlow model and does not require ethical approval for human or animal subjects.

Informed Consent

As this study does not involve human participants or data collected from individuals, informed consent is not applicable.

References

1. Prasanna, P. & Rao, Dr. (2018). "Text classification using artificial neural networks". International Journal of Engineering and Technology (UAE). 7. 603-606. 10.14419/ijet.v7i1.1.10785.
2. Zheng, Mingshi. (2024). "Research on text classification based on neural networks". Applied and Computational Engineering. 41. 282-303. 10.54254/2755-2721/41/20230768.
3. Kim, D. (2022). Research on Text Classification Based on Deep Neural Network. International Journal of Communication Networks and Information Security (IJCNIS), 14(1s), 100–113. <https://doi.org/10.17762/ijcnis.v14i1s.5618>
4. Ruishuang Wang, Zhao Li, Jian Cao, Tong Chen, Lei Wang. "Convolutional Recurrent Neural Networks for Text Classification". IEE Xplore on 2019 International Joint Conference on Neural Networks (IJCNN), DOI: 10.1109/IJCNN.2019.8852406.
5. Liang Wang, Zhiwen Yu, Bin Guo, Dingqi Yang, Lianbo Ma, Zhidan Liu, Fei Xiong. "Data-driven Targeted Advertising Recommendation System for Outdoor Billboard". ACM Transactions on Intelligent Systems and Technology (TIST), Volume 13, Issue 2 Article No.: 29, Pages 1 – 23 <https://doi.org/10.1145/349515>
6. Kunze Wang, Yihao Ding, Soyeon Caren Han. "Graph neural networks for text classification: a survey". Artificial Intelligence Review (2024) 57:190 <https://doi.org/10.1007/s10462-024-10808-0>
7. Ahlam Wahdan, Sendeyah Hantoobi, Said A. Salloum, Khaled Shaalan. "A systematic review of text classification research based on deep learning models in Arabic language". International Journal of Electrical and Computer Engineering (IJECE) Vol 10, No.6, December 2020, pp. 6629~6643 ISSN: 2088-8708, DOI: 10.11591/ijece. v10i6. pp6629-6643
8. D. Saha, "Web Text Classification Using a Neural Network," 2011 Second International Conference on Emerging Applications of Information Technology, Kolkata, India, 2011, pp. 57-60, doi: 10.1109/EAIT.2011.34.