Comparison of Machine Learning and Deep Learning for Text Classification on an Unbalanced Dataset

Seowoo Kim Aiffel / Seoul, South Korea
ygttsu@naver.com

Abstract

This study explores the performance of conventional machine learning models and a simple deep learning model in text classification tasks, particularly on unbalanced data. Naive Bayes and Gradient Boosting were selected as the machine learning models, while a one-dimensional Convolution Neural Network was chosen as the deep learning model. The results demonstrated that a simple deep-learning model might be effective and efficient for classifying unbalanced text datasets.

1 Introduction

In the era of big data, the need to classify text is accelerating. Text classification can be used for various purposes, including spam identification, emotion analysis, news text classification, etc. It has been suggested that both machine learning and deep learning model perform well in this task. For example, in Reuters dataset classification tasks, representative machine-learning models, such as Naive-Bayes model, Gradient Boosting model, showed decent results with 0.88 and 0.93 accuracy respectively (Zdrojewska et al., 2019). The Convolution Neural Network model (CNN), one of the popular deep-learning techniques for classification, also excels at classification, reaching 0.92 accuracy (Cai et al., 2018).

Even though previous research has proven the reliability of these models in terms of accuracy, we still don't clearly know which model would generally perform the best on unbalanced datasets. In reality, most data is skewed, and identifying each models' performance on raw dataset would be of help in deciding the most efficient model for applied areas.

Therefore, the main purpose of this research was to compare the performance of machine learning and deep learning models in classification tasks on unbalanced dataset.

2 Methods

The purpose of text classification is to determine the category of a given document, and the result of classification may involve multiple categories.

The study used the Reuters News dataset, which has a total of 46 classes. Exploratory Data Analysis (EDA) was conducted as a basic step to figure out if this dataset is skewed. For the deep learning model, the length of each news item was explored to decide the extent of truncation point for padding. As a result, the dataset was found to be skewed, mostly distributed in classes 3 and 4, accounting for 35.17

2.1 Preprocessing

To train machine-learning models, the news data was transformed into a matrix using TF-IDF. TF-IDF represents words' importance based on word frequency. The train and test dataset were fitted into this model before being adapted to a machine-learning model.

2.2 Machine Learning Model

Naive Bayes (NB) and Gradient Boosting (GB) were selected as machine-learning models. NB was selected as a baseline model. GB was selected as well, because the model showed the best precision and Recall score among machine-learning model in multi-class classification tasks (Zdrojewska et al., 2018).

2.3 Deep Learning Model

For deep learning model, 1D CNN was used. According to Kim (2014), simple CNN model showed its greatness on sentence-level classification tasks. 1D CNN is effective in classification tasks for relatively small dataset, thanks to its ef-

fective feature extraction methods, by capturing neighbors of each words in a sentence.

The model contains following part:

Part 1: Input layer. The preprocessed text data (padded data) is input into the model.

Part 2: Embedding layer. The layer converts the input text data into dense vectors of fixed size of 256. This layer helps in capturing the semantic meaning of the words.

Part 3: Dropout layer to prevent overfitting. The dropout rate was 0.2.

Part 4: Convolution Layer. The layer applies convolutional filter with 256 filters of size 3. It helps in extracting local features from the embedding vectors.

Part 5: Max-Pooling Layer. In this layer, Global Max Pooling was applied, so as to minimize the loss of feature information and reduces the computation required for the next layers as well.

Part 6: Dense Layer. In this layer, the activation function was ReLu function. This layer included 128 neurons as hidden units.

Part 7: Dropout Layer. This was the same with Part 3.

Part 8: SoftMax Layer. Regarding the purpose of the task, the last layer has softmax as its activation function. Softmax produced the probability distribution over the 46 classes indicating the likelihood of each class for a given input.

2.4 Procedure

To determine the extent to which word frequency is needed for optimal training in machine learning models, I tested the dataset with 1000, 3000, and 5000 words sorted according to their frequency. The number of words that showed the best result was then selected for training the deep-learning model.

3 Results

3.1 Machine-learning model

The GB model showed the better score including accuracy, precision, recall and F1-score in all trials. Regarding the number of words, in NB, including more words worsens the model's performance while in GB, the results was opposite.

3.2 Deep learning model

The best model was found at the third epochs. With the test dataset, this model's accuracy turned out to be 0.8054, and the loss was 0.8360.

Model	Accuracy	Precision	Recall	F1-Score
NB 1000	0.70	0.68	0.71	0.67
NB 3000	0.69	0.65	0.69	0.63
NB 5000	0.68	0.62	0.68	0.61
GB 1000	0.73	0.73	0.73	0.72
GB 3000	0.77	0.77	0.77	0.76
GB 5000	0.78	0.78	0.78	0.78

Table 1: Result of NB and GB. NB, Naive Bayes model, GB, Gradient Boosting model, 1000, 3000, 5000 is a number of words sorted based on its frequency.

[width=0.5]capture.png

Figure 1: Loss and Accuracy of 1D-CNN model.

4 Discussion

As result demonstrated, 1D CNN model outstripped all machine learning models in accuracy. The result implies that using simple deep learning model could be the efficient way in classifying unbalanced text data. Deep learning model could be the efficient option, in that the simple deep-learning model, such as 1D CNN requires shorter processing time for training than traditional machine learning models do, as well as higher accuracy.

5 References

Cai, J., Li, J., Li, W., & Wang, J. (2018). *Deeplearning Model Used in Text Classification. 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP).*

Yoon Kim, "Convolutional Neural Networks for Sentence Classification", EMNLP 2014, Part number. 1of1, pp. 1746-1751, Aug. 2014

Zdrojewska, A., Dutkiewicz, J., Jedrzejek, C., Olejnik, M. (2019). Comparison of the Novel Classification Methods on the Reuters-21578 Corpus. In: Choroś, K., Kopel, M., Kukla, E., Siemiński, A. (eds) Multimedia and Network Information Systems. MISSI 2018. Advances in Intelligent Systems and Computing, vol 833. Springer, Cham.