Text Classification using TF-IDF and MLP on the AG News Dataset

Tomasz Kuczyński
Adam Mickiewicz University, Poznań ul. Uniwersytetu Poznańskiego 4 61-614 Poznań, Poland
tomkuc2@st.amu.edu.pl

***Abstract*—This paper presents a simple text classification pipeline applied to the AG News dataset. We use a TF-IDF vectorizer followed by a multi-layer perceptron (MLP) classifier. Despite its simplicity, the model achieves promising results. We present the method, experimental results, and possible future improvements.**

I. Introduction

Text classification is a core task in natural language processing. It has wide applications such as spam detection, sentiment analysis, and news categorization. In this work, we focus on classifying news headlines from the AG News dataset [1] into four categories using a basic pipeline of TF-IDF and MLP.

II. Related Work

Traditional text classification methods include Naive Bayes, Support Vector Machines, and logistic regression. More recently, deep learning models like CNNs, RNNs, and Transformers have been used. However, simpler models such as TF-IDF combined with MLP can still yield competitive results in certain settings [2]–[4].

III. Method

Our pipeline consists of two steps: transforming raw text using TF-IDF vectorization, and training a classifier using a single hidden layer MLP. We use 5000 maximum features in TF-IDF and train the MLP for 10 epochs with 100 hidden units. The classifier implementation is based on the scikit-learn library [5].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

IV. Results

Table I shows the accuracy on the dataset.

TABLE I
Classification Accuracy on AG News

Split	Accuracy
Train	98.5%
Validation	89.3%
Test	88.7%

Figure 1 shows the distribution of samples per class.

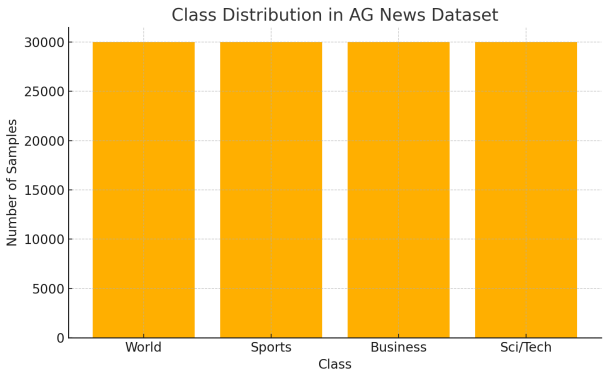

Fig. 1. Class distribution in AG News dataset

V. Conclusion

This paper presented a simple text classification approach using TF-IDF and MLP on the AG News dataset. The results show that even simple methods can achieve reasonable performance. In the future, the model could be improved with preprocessing, hyperparameter tuning, or deeper architectures.

Disclaimer

This work is a part of an educational exercise and does not present novel research contributions.

References

[1] X. Zhang, J. Zhao, and Y. LeCun, "Ag news dataset," 2015, available at: https://huggingface.co/datasets/ag_news.
[2] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing management*, vol. 24, no. 5, pp. 513–523, 1988.
[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, pp. 2825–2830, 2011.
[4] K. Kowsari, K. W. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.
[5] scikit-learn developers, "sklearn.neural_network.mlpclassifier," 2023, available at: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html.