

An Effective Sentimental Analysis Model Based on spaCy

Yuanhao Shen *

Department of Computer Science, Xidian University, ShaanXi, China

* Corresponding author: 20009200039@stu.xidian.edu.cn

Abstract. As machine learning and the Internet continue to advance, the role of sentiment analysis in discerning positive and negative emotions within text has become increasingly pervasive. Its applications extend to enhancing human-computer interactions, monitoring mental health, and conducting business analyses. Numerous efforts have been dedicated to sentiment analysis and prediction. This article utilizes a dataset of 50,000 movie reviews, sourced from natural language processing (NLP) and included in the spaCy library, to contribute to this ongoing body of work. Two feature extractions, Count Vectorizer and TF-IDF are used, and three Machine Learning (ML) algorithms, Logistic Regression (LR), Decision Tree (DT) and the Multilayer Perceptron (MLP) are used to make predictions on the IMDB review data set with Python programming language. Comparing the experimental results, it can be found that different models have significant differences under different feature extractions, and using TF-IDF feature extraction combined with logistic regression model achieves the best accuracy (88.53%). It proves that LR, among three tested models, performs best in sentimental analysis.

Keywords: Machine learning, Natural language processing, sentimental analysis.

1. Introduction

Sentimental analysis is a method to identify, extract and categorize subjective information from text by mainly using Natural Language Processing (NLP) technology [1]. It has a diverse array of applications, such as monitoring people's mental health, improving interactions between humans and computers, and business analysis. Sentimental analysis can be applied to online chat supervision, where it can detect negative emotions and issue alert notifications to relevant personnel. This can effectively reduce the occurrence of depression or suicide cases.

The mainstream methods in sentimental analysis include Lexicon-based sentimental analysis, Machine Learning (ML) based sentimental analysis, and Deep Learning (DL) based sentiment analysis. Lexicon-based methods involve constructing sentiment lexicons to match sentiment words in the text for sentiment classification. It utilizes tokenization to split text into tokens, counts the occurrences of each word, and references an established lexicon to determine the subjectivity of each word [2]. ML based methods utilize training data to train classifiers, enabling automatic learning of sentiment features and patterns. Deep learning-based methods leverage deep neural networks to extract high-level features from text, leading to more accurate sentiment analysis.

In recent research, in order to automatically categorize the preprocessed IMDB movie reviews into different classes, a Long Short-Term Memory (LSTM) classifier including Adam optimizer is employed. A total of 10k reviews are taken into consideration. Among them, 5k reviews are positive sentiments and 5k reviews are negative sentiments. The results indicated that the highest accuracy can be achieved to 89.9% by the devised approach. In another one, the NLP approach, which uses ML algorithms and the DL methods, is utilized to classify the text-form reviews in the data set, which have been classified into separate categories of positive and negative. According to the observations, it is known that DL approaches generally performed higher accuracy than ML algorithms. Among the DL approaches, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), LSTM, RNN achieves a higher accuracy rate of 88%. That's because the feature extraction is supposed to be done manually when ML algorithms are used, while in the DL approach, no human intervention is required as the machine automatically performs feature extraction [3].

In this article, sentimental analysis is finished by making use of ML and DL approaches on the preprocessed IMDB review data set. IMDB data set has fifty thousand movie reviews for Natural

Language Processing (NLP) or text analysis. The text-form reviews are categorized as negative and positive reviews, which means designed for binary sentiment classification. It contains substantially more data than previous benchmark datasets.

In this study, incorporating NLP techniques, the study's focus lies in investigating the effectiveness of three commonly used ML models, namely Logistic Regression (LR), Decision Tree (DT), and Multi-layer Perceptron (MLP). The research objective is to evaluate the performance of these models on a 50k-size data set. By comparing the effects of different feature extraction techniques under the same preprocessing conditions, the aim is to evaluate the performance of these models.

2. Methodology

2.1. Dataset

In terms of this study, a data set, created by Andrew Maas, that contains 50k movie reviews from IMDB is operated. 80% of the reviews in the data set are splitted for the purpose of training, while another 20% are intended to test three certain classifiers. The reviews in the data set can be distinguished into two parts, positive part and negative part. It refers to the IMDB rating system. According to dataset creator, the rating system allows viewers to score within the range from 1 to 10. If the score is ≤ 4 , the movie is labeled negative. If the score is ≥ 7 , the movie is labeled as positive sentiments. Besides, reviews with scores outside the above ranges are excluded. To each movie, there are not more than 30 reviews available for it. With a standard deviation of 172.91 words, the average length of each review is approximately 234.76 words. In summary, 88585 different words are contained in the data set [2].

2.2. Preprocess

2.2.1. Data Processing

In the study, a spaCy tool is used, including tokenization, removal of stop words and numbers, normalization and so on. In the data preprocessing phase, a series of steps to clean and normalize the text data are utilized. Firstly, the study mapped the sentiment labels to binary values (0 for 'negative' and 1 for 'positive') for easier classification. Then, the preprocess function removed non-alphabetic characters by using regular expressions to eliminate any special characters or symbols that may not contribute to the sentiment analysis. Next, the function eliminates any additional white spaces and trims leading and trailing white spaces to maintain text format consistency. In order to ready the text for analysis, it tokenizes the reviews and employs lemmatization to reduce each word to its base form. This process assists in minimizing noise and standardizing vocabulary. Finally, the lemmatized tokens are joined back into a sentence for further analysis.

2.2.2. Feature Extraction

Feature extraction is a crucial step in natural language processing and machine learning. In text data processing, the objective of feature extraction is to convert the text into a format that computers can comprehend and manipulate. Typically, text data is unstructured and consists of characters or words. The process of feature extraction converts text data into numerical feature vectors, so that machine learning algorithms can process and analyze them. Count Vectorizer (CV) is a commonly used method for feature extraction, which converts text into a feature vector representation based on word frequencies.

CV breaks down the text into words and calculates the occurrence count of each word in the text. It constructs a vocabulary where each word is associated with a unique index. Then, for each text sample, CV creates a vector where the value at each index position represents the number of occurrences of the corresponding word in the text.

CV can transform each review in the IMDB review dataset into a feature vector based on word frequencies. These feature vectors can be used as input for machine learning models to differentiate

between positive and negative reviews. CV provides a simple and intuitive way to capture important information in textual data and convert it into a numerical feature representation for processing by machine learning algorithms.

TF-IDF quantifies the significance of a term in a review document compared to the entire corpus. The TF-IDF value increases with the term's frequency in the document but is offset by its frequency across the corpus. Term Frequency (TF) means measuring the frequency of a term in a review document. Moreover, Inverse Document Frequency (IDF) measures the significance of a term in the entire corpus, emphasizing rare terms.

In the IMDB review dataset, each review can be represented as a TF-IDF vector, where the weight of each word corresponds to its TF-IDF score. These vectors serve as features for machine learning models, aiding in the discrimination of positive and negative reviews. A TF-IDF matrix can be constructed, with rows representing reviews and columns representing unique terms in the corpus. The cells of the matrix contain the TF-IDF scores for each term in the corresponding review. This transformation of reviews into TF-IDF vectors provides a feature-rich representation that is vital for machine learning.

2.3. Models

2.3.1. Logistic Regression

Logistic regression is a popular and versatile ML algorithm utilized for binary and multi-class classification tasks [4, 5]. It predicts the probability of a sample belonging to different classes by linearly combining the input features with weights and mapping the result through a non-linear function, typically the logistic function, to the probability range.

The fundamental concept of logistic regression is to establish a model based on the relationship between the known feature-label pairs in a given dataset and then employ this model to classify new data. With the goal of getting the predicted probabilities as near to the true labels as possible, the weights that parameterize the model are trained by either maximizing the likelihood function or minimizing the loss function.

2.3.2. Decision Tree

DT model is a popular and versatile ML algorithm used for various classification and regression tasks [6, 7]. It is known for its the capability to handle both categorical and numerical data, as well as its interpretability.

The Decision Tree algorithm is a model of decisions and predictions, which is designed like a tree. It recursively splits the data based on different features, creating a set of decision rules that lead to the prediction of the target variable.

In a classification context, DT algorithm splits the data based on feature values to maximize the purity of each resulting subset. It determines the best splitting criteria using measures like Gini impurity or information gain.

In a regression context, the Decision Tree algorithm splits the data to minimize the variance within each resulting subset. It uses measures like mean squared error or mean absolute error to determine the optimal splitting points.

2.3.3. Multilayer Perceptron

MLP is a versatile and powerful supervised learning model. It is widely used for various tasks [8, 9], including classification, regression, and pattern recognition. The MLP's capability to model complex non-linear relationships makes it a popular choice in modern machine learning.

MLP is a kind of feedforward neural network, which comprise multiple layers of interconnected nodes (neurons). One input layer, one or more hidden layers, and one output layer compose it. Except for the input layer Each layer, the network contains a set of neurons, and each neuron is connected to every neuron in the adjacent layers. Non-linearity is introduced by activation functions into the model, enabling it to approximate complicated relationships.

2.4. Parameters

The LR model uses the L2 penalty and a regularization parameter of 1.0. It fits the intercept and uses the lbfgs solver, with a maximum of 100 iterations.

The DT model uses the Gini criterion to measure impurity and does not specify a maximum depth. It splits nodes based on a minimum of 2 samples and requires a minimum of 1 sample in each leaf node. It considers all features for splitting and does not set a random state.

The MLP model has a hidden layer size of 100. It uses the ReLU activation function [10] and the Adam solver. The regularization parameter is set to 0.0001, and the batch size is automatically determined. Moreover, the learning rate is constant, and the model runs for a maximum of 200 iterations.

3. Results and Discussion

From Table 1, it can be observed that when using Count Vectorizer feature extraction, the logistic regression model showed the highest accuracy, reaching 87.61%. The accuracy of the decision tree model is 72.07%, and the accuracy of the multi-layer perceptron model is 84.57%. It can be seen that under this feature extraction method, the logistic regression model performs best on the IMDB sentiment analysis dataset.

Furthermore, observations revealed that, when employing TF-IDF feature extraction, the accuracy of the logistic regression model exhibited a slight improvement, achieving 88.53%. Conversely, the accuracy of the decision tree model showed a slight decrease in comparison to the CountVector feature extraction, reaching 71.92%. The accuracy of the multi-layer perceptron model has been improved, reaching 86.83%. Under this feature extraction method, the logistic regression model still performs best.

Table 1. Accuracy of all the results

	LR	DT	ML
CV	87.61%	72.07%	84.57%
TF-IDF	88.53%	71.92%	86.83%

Fig. 1, Fig. 2, Fig. 3, Fig. 4, Fig. 5 and Fig. 6 show the precision rate, recall rate and f1-score of the results from each model. Fig. 1, Fig. 2 and Fig. 3 use Count Vectorizer, and Fig. 4, Fig. 5 and Fig. 6 use TF-IDF.

The following conclusions can be drawn. TF-IDF feature extraction has slightly improved in logistic regression models, but its effect on decision tree models is not significant, while it has improved in multi-layer perceptron models.

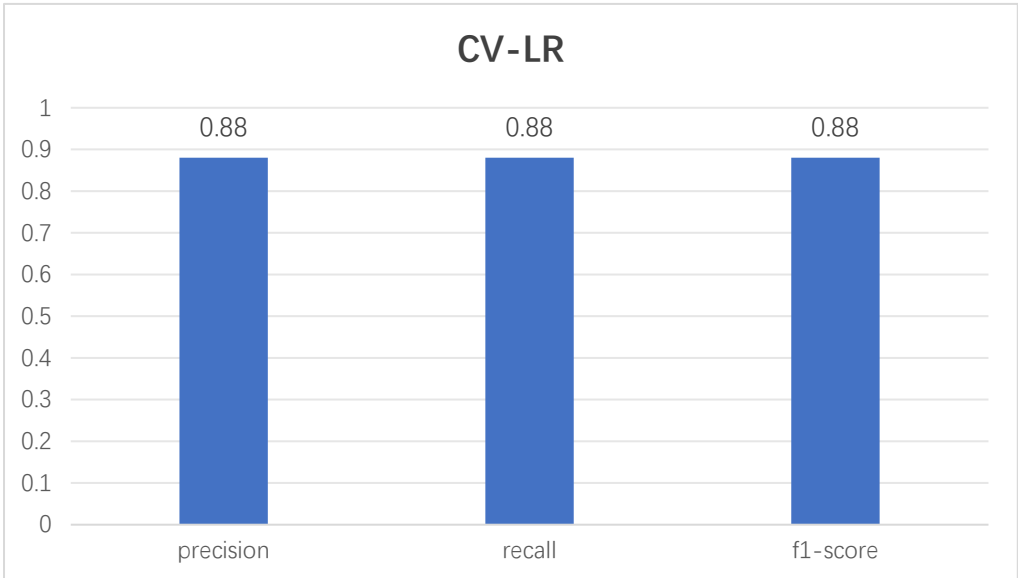


Figure 1. Results from CV and LR (Photo credit: Original).

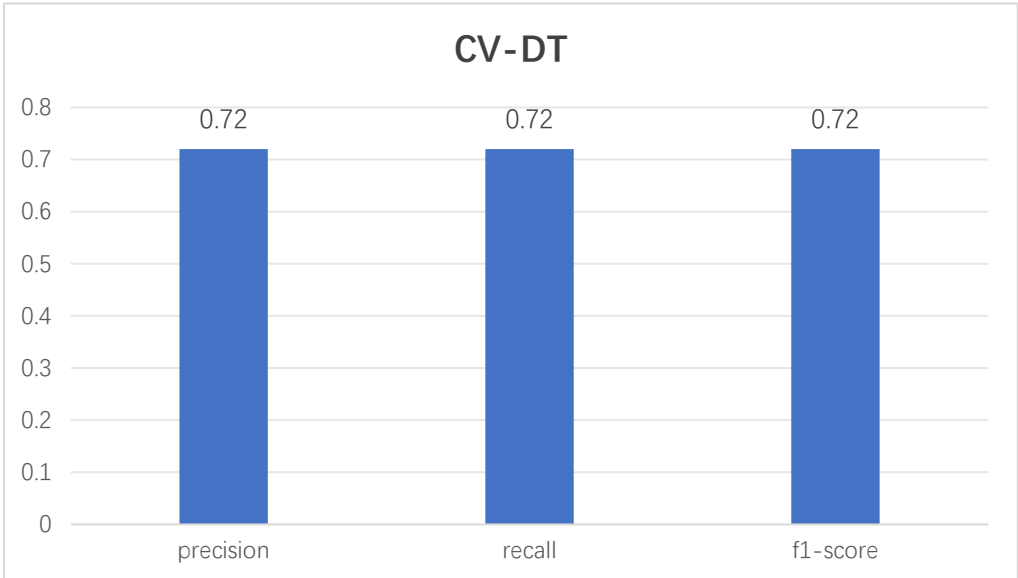


Figure 2. Results from CV and DT (Photo credit: Original).

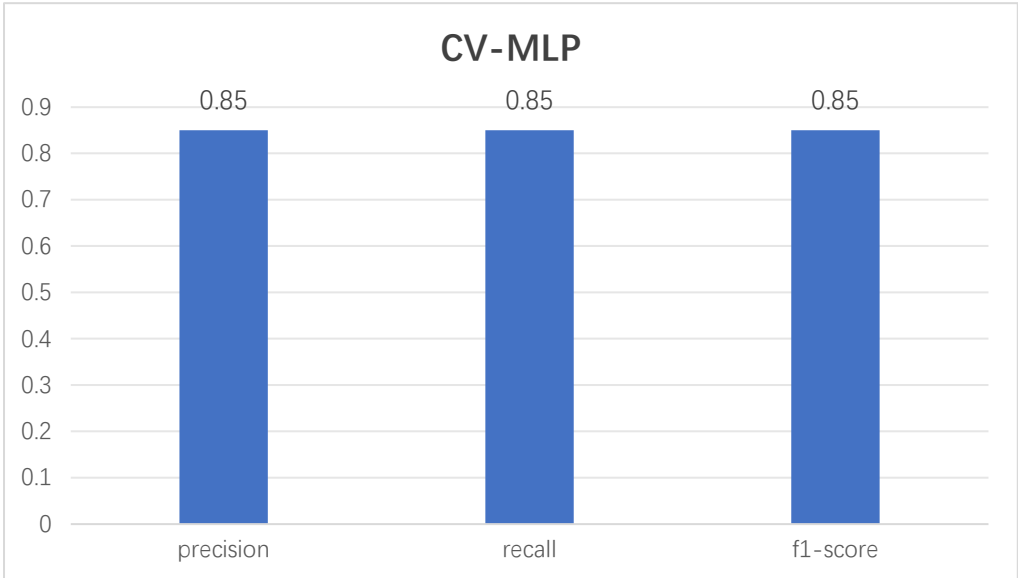


Figure 3. Results from CV and MLP (Photo credit: Original).

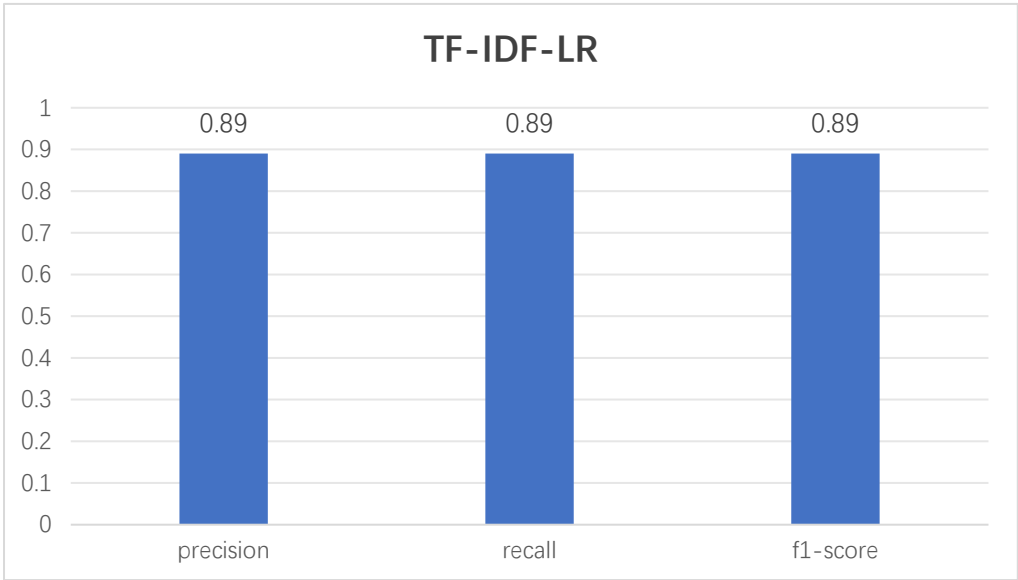


Figure 4. Results from TF-IDF and LR (Photo credit: Original).

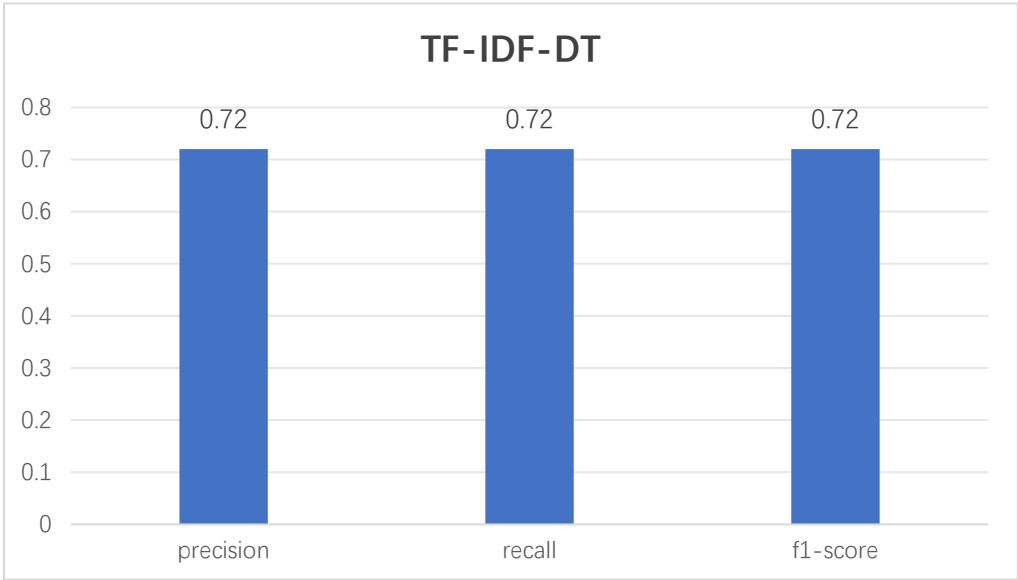


Figure 5. Results from TF-IDF and DT (Photo credit: Original).

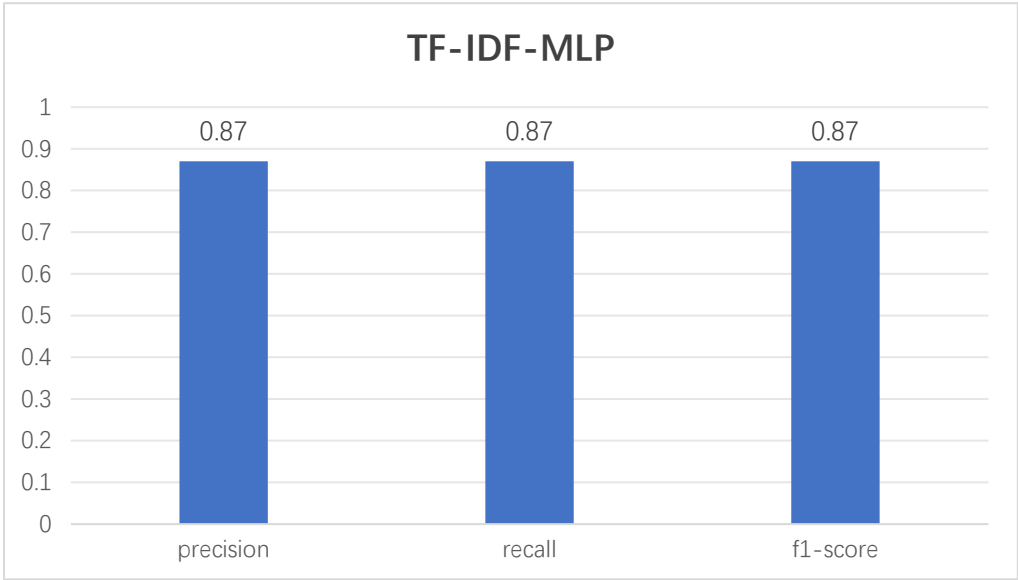


Figure 6. Results from TF-IDF and MLP (Photo credit: Original).

Fig. 7 shows the confusion matrix of the result of the LR model under the TF-IDF feature extraction, which has the best accuracy in the study. It’s obvious that the precision of LR model for positive reviews is a bit higher than that for negative reviews. In addition, the recall rate for negative reviews is higher than that for positive reviews.

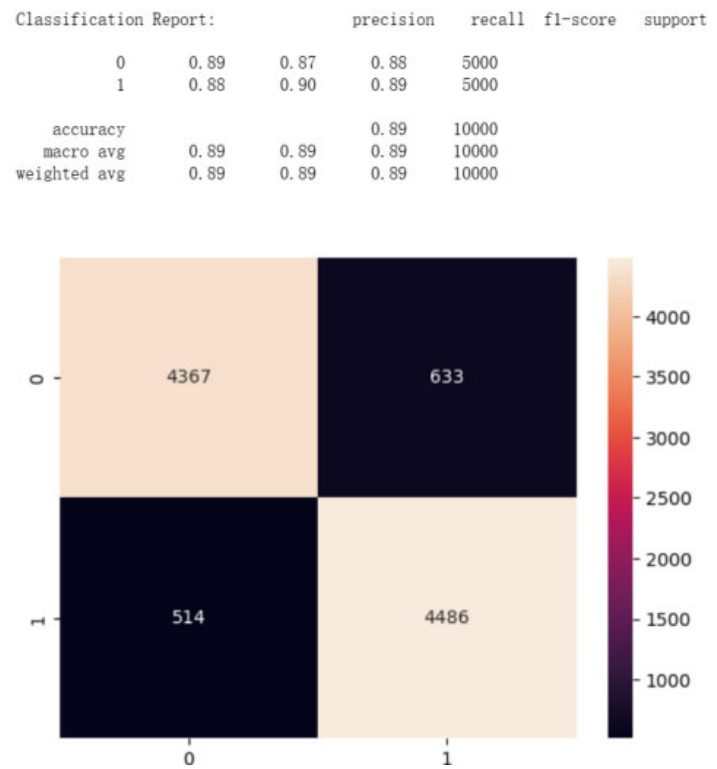


Figure 7. Confusion Matrix (Photo credit: Original).

4. Conclusion

Machine learning techniques have found successful applications in sentiment analysis. The future of sentiment analysis is anticipated to evolve toward greater intelligence, enabling more efficient analyses. This research highlights the superiority of logistic regression models over decision trees and multi-layer perceptrons in sentiment analysis. Additionally, it demonstrates that the use of TF-IDF brings about notable improvements. However, the study has certain limitations, including the use of a random seed (5) for dataset partitioning, which may not ensure strict balance between training and testing datasets, potentially leading to category imbalance.

References

[1] Cheng Q, et al. Short-Term load forecasting model and method for power system based on complementation of neural network and fuzzy logic. Transactions of China Electrotechnical Society, 2004, 19 (10): 53 - 58.

[2] Sailunaz K, Dhaliwal M, Rokne J. et al. Emotion detection from text and speech: a survey. Soc. Netw. Anal. Min. 8, 28, 2018.

[3] Qaisar S M. Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory, 2020 2nd International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2020, pp. 1 - 4.

[4] LaValley M P. Logistic regression. Circulation, 2008, 117 (18): 2395 - 2399.

[5] Menard S. Applied logistic regression analysis. Sage, 2002.

[6] Myles A J, Feudale R N, Liu Y, et al. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 2004, 18 (6): 275 - 285.

- [7] Song Y Y, & Ying L U. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27 (2), 130, 2015.
- [8] Qiu Y, Yang Y, Lin Z, et al. Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV. China Communications, 2020, 17 (3): 46 - 57.
- [9] Valanarasu J M J, Patel V M. Unext: Mlp-based rapid medical image segmentation network. International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2022: 23 - 33.
- [10] He J, Li L, Xu J, et al. ReLU deep neural networks and linear finite elements. arXiv preprint arXiv: 1807.03973, 2018.