Sentiment Analysis on Internet Movie Database (IMDb) Movie Review Dataset: Hyperparameters Tuning for Naïve Bayes Model

Haoran Li

Department of Material Science and Engineering, Tokyo Institute of Technology, Yokohama, 226-8503, Japan
1912231219@mail.sit.edu.cn

Abstract. Sentiment classification plays a crucial role in understanding and analyzing text data, particularly in domains like social media and online reviews. In this study, the influence of three key parameters on the accuracy of sentiment classification was investigated by applying Naive Bayes classifier to the Internet Movie Database (IMDb) movie review dataset. To explore the impact of the training set ratio, the proportion of data allocated to the training set was varied while keeping other parameters constant. Results indicate that increasing the training set ratio from 50% to 90% leads to a gradual improvement in classification accuracy. This finding suggests that a larger training set provides more representative samples for learning, enhancing the model's ability to generalize. Subsequently, the impact of the maximum features parameter—which establishes the feature space's dimensionality—was investigated. By changing the number of features taken into account, it is found that a larger value of max features, like 4096, produces better accuracy. Additionally, the impact of the smoothing parameter alpha on classification accuracy was investigated. The experiments showed that different alpha values, such as 0.1, 0.5, and 1, had minimal influence on the accuracy. This suggests that the Naive Bayes classifier is relatively robust to variations in the smoothing parameter in the context of sentiment classification. The findings emphasize the significance of a larger training set and an optimal number of features for improving accuracy, meanwhile the influence of the smoothing parameter appears to be limited in this context.

Keywords: Sentiment Classification, Machine Learning, Naive Bayes Classifier, Hyperparameter Tuning.

1 Introduction

Movie Review, whether conducted by an individual or through collective means, serve as a critical analysis of a movie aimed at expressing opinions about its various aspects. The goal of movie reviews is to make it easier for manufacturers to get client input so they may further develop their products. On the other side, users might analyze a product objectively by examining other people's reviews, which may affect

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_73

their judgments regarding whether or not to purchase the goods [1]. By comprehending and valuing film reviews, individuals can make informed choices regarding their selection of future movies to watch. Furthermore, these reviews often possess an emotive quality, commonly falling into either a negative or positive categorization, thereby allowing for a broad classification of their overall sentiment.

Nowadays, massive reviews on many websites can be easily accessed, such as Douban and Netflix. These platforms serve as avenues for expressing opinions in diverse formats. One illustration may be websites that allow users to leave personal evaluations on movies, like Amazon or Rotten Tomatoes [2, 3]. These evaluations often include more language and are lengthier. Other types of websites, such as posts on social networks like Twitter or article evaluations on Digg [4, 5], frequently have brief comments. Given the voluminous nature of textual content in reviews, it becomes pertinent to ascertain expeditiously whether a particular review conveys a positive or negative sentiment, as deciphering the intended sentiment from such texts can be arduous and obscure. It is important to figure out the precise emotion of a certain text since sentiment analysis has always been challenging due to obstacles such as slang terms, misspellings, short forms, repetitive characters, application of regional language, and new emerging emoticons [6]. With regard to extracting the important information from these review texts, it turns out that machine learning plays a key role in sentiment analysis.

In past decades, major improvements have been realized in the area of artificial intelligence (AI), leading to the emergence of various technologies. These technologies encompass a range of algorithms, such as support vector machines, Bayesian classifiers, random forests, to highly hardware-dependent technics such as deep neural network [7]. These algorithms have been used in a variety of industries, including healthcare and education. For natural language processing, machine learning has assumed a vital role in recent years. Hasan et al. substantially improved the accuracy of sentiment analysis based on Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer [8], and the classifier can be applied to any device with high performance. Some models such as naïve bayes seems to be simple and easy to understand, as bayes classifier only considers the frequency of the words, while such algorithm is proved to have good performance on predicting the polarity emotion of the texts. Dey claims that Naïve Bayes yield better results than K-Nearest Neighbor (K-NN) approach and can be used successfully in analyzing movie reviews [9].

As mentioned above, the selection of an appropriate algorithm in classification tasks significantly impacts the results, and the optimization process within machine learning also holds importance [10]. Based on the Internet Movie Database (IMDB) movie review data set, a Naïve Bayes classifier was used in this article to determine sentiment's polarity. (i.e. positive/negative). Finally, the classification accuracy is compared under different data split ratios and bag-of-words model parameter settings. As the training set ratio increases from 50% to 90%, the accuracy slightly improves. When the max features value increases to 4096, the best accuracy of 0.86 is achieved. It can be inferred that a larger max features value allows the model to capture more semantic information, which indicates that having sufficient data and an expressive feature representation is important for training high-performance machine learning models.

2 Method

2.1 Data Preparation

IMDb created by Stanford University is a widely recognized and extensively used online repository of film-related data, encompassing a vast array of movie information and user reviews [11].

The Stanford IMDb Movie Review Dataset contains movie reviews sourced from the IMDb website, primarily employed for sentiment analysis tasks. The dataset includes reviews of 50,000 films, of which 25,000 were used to train the model and another 25,000 were used to test it. The distribution of reviews within both the training and test sets is representative, comprising a balanced mixture of positive and negative sentiments. Each movie review is labeled as positive or negative for a sentiment classification task.

However, the use of the Stanford IMDb movie review dataset is subject to some limitations and caveats. Since movie reviews are spontaneously written by users, they may contain a lot of subjective opinions and text noise. In addition, movie reviews vary in length and style, requiring proper preprocessing and feature extraction.

As for preprocessing steps in this article, first, the IMDb movie review dataset is loaded, then the training and testing sets are concatenated to facilitate subsequent preprocessing steps. Following that, the integer sequences within each sample are transformed into strings to enable tokenization, which converts the text sequences into sequences of word indices and constructs a vocabulary. The sequences are then padded and truncated to ensure a consistent length across all samples. After proportion of data allocated to the training set is determined, the integer sequences are converted back to text sequences and transformed into string arrays. Next, a bag-of-words model is created. By applying the bag-of-words model to text data, the text sequences are transformed into representations as bag-of-words vectors. In order to be utilized for further classification tasks, the function finally delivers the feature matrices for the training and testing sets, together with their respective labels.

2.2 Machine Learning Model

The Naive Bayes classifier, a prominent machine learning algorithm, has garnered significant attention and widespread adoption in the domain of classification tasks [12]. It relies on the Bayes theorem and presupposes that the classification characteristics are independent of one another [13]. Despite this "naive" assumption, Naive Bayes classifiers have demonstrated to be quite successful in various domains, including text classification, spam filtering, and sentiment analysis [14].

Based on the observed feature values, the Naive Bayes classifier determines the likelihood that a data instance fits to a given class. Bayes' theorem is employed which connects the prior likelihood of an event occurring to the conditional probability of an event [12]. In the context of classification, the algorithm calculates the conditional probability of a class given the observed features. All features are thought to be independent of one another when using naive Bayes classifiers. This presumption makes the calculation easier and the algorithm more effective [13]. Although the assumption rarely holds true in real-world scenarios, Naive Bayes classifiers often perform surprisingly well, especially when dealing with large feature spaces.

2.3 Hyperparameters

When using a Naive Bayes classifier for sentiment analysis, the following parameters can have an impact on the accuracy of the classifier:

Alpha controls the degree of smoothing applied to the feature probabilities, balancing between overfitting and underfitting [15]. Max features determine the maximum number of features considered during vectorization, allowing for control over the dimensionality of the feature representation [16]. The data set split ratio influences the model's ability to generalize, with a larger training set providing more representative samples while a smaller training set may lead to overfitting [17]. Careful consideration and selection of appropriate values for these parameters are crucial to achieve a balance between bias and variance in the model and optimize performance.

These three parameters will be examined in this article along with their combinations in order to understand how they affect the classifier's accuracy.

3 Results and Discussion

The augmentation of the train ratio, starting from 50% and progressing to 90%, exhibits a slight enhancement in accuracy shown in Fig. 1. This indicates that using a larger portion of the dataset for training allows the model to learn more effectively and improve its performance. However, the increase in accuracy is not significant, suggesting that the model is already capturing the important patterns and features from the data even with a smaller training set. The accuracy scores obtained for different train ratios are relatively close, ranging from 0.813 to 0.819. Overall, the model's performance is robust and not highly sensitive to variations in the train ratio. It indicates that the classifier demonstrates commendable generalization capabilities with respect to unseen data, thereby consistently attaining high accuracy across different train ratios.

As the max features value is increased from 256 to 4, 096, a clear trend of increasing accuracy can be observed. This indicates that including more features in the feature vector representation of the text data leads to improved performance of the classifier. With a larger vocabulary, the model can capture more detailed information and better discriminate between different classes of sentiment. However, it is crucial to acknowledge that the rate of improvement in accuracy diminishes as the max features value becomes larger. The result suggests that 2, 000 features might be a point of diminishing returns, where the inclusion of additional features beyond a certain threshold does not significantly enhance the model's performance.

In terms of the alpha parameter, the experimental results demonstrated that varying this value in the classifier does not have a significant impact on the accuracy. Default smoothing value (alpha = 1) already provides satisfactory performance in sentiment classification on the IMDb dataset.

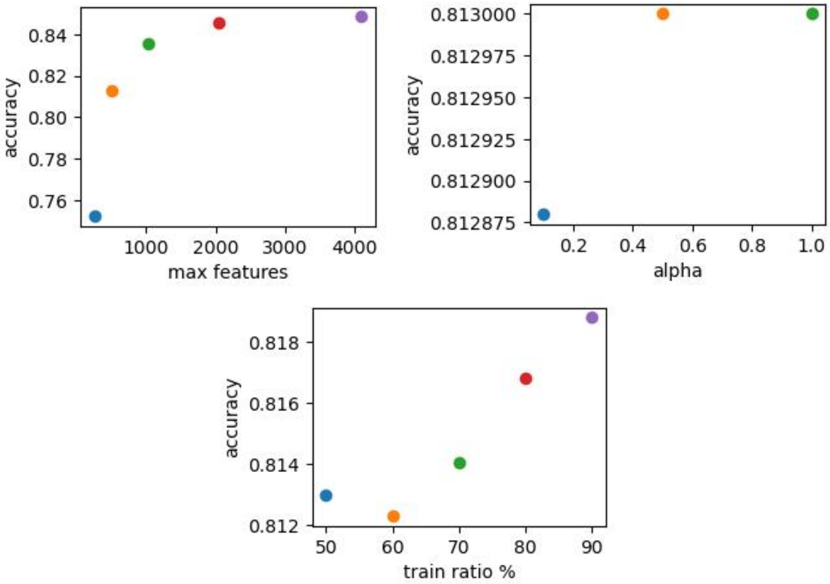

Fig. 1. Relationships between accuracy and max features, alpha, and training set ratio (Photo/Picture credit: Original).

When performing the grid search, the optimal combination that resulted in the highest accuracy was found to be alpha = 0.1 and max features = 4096, with an accuracy of 0.84852. This suggests that a larger number of features and increased training set ratio contribute to improved sentiment classification performance on the IMDb dataset.

4 Conclusion

This article focused on sentiment classification for the IMDb movie review dataset using a Naive Bayes classifier. The influence of three key parameters is investigated, namely train ratio, max features, and alpha. Additionally, a grid search was performed to identify the optimal combination of alpha and max features for maximizing accuracy. Experiments' findings demonstrated that accuracy was enhanced by raising the training set ratio. This indicates that a larger training set facilitates better model generalization. Moreover, increasing the max features also resulted in higher accuracy, suggesting that capturing more detailed information from the text contributes to better sentiment classification performance. Interestingly, varying the alpha value had minimal impact on accuracy, indicating the robustness of the Naive

Bayes classifier to changes in alpha. Through the grid search, the optimal values were found to be alpha = 0.1 and max features = 4096, resulting in an accuracy of 0.84852.

However, this study still has some limitations. It focused only on the Naive Bayes classifier without comparing it to other algorithms. The analysis was limited to IMDb movie reviews, and generalizability to other domains is uncertain. Future research should explore alternative models and investigate additional factors for improved sentiment classification. Ensemble methods and deep learning models could be explored to enhance performance.

References

1. Zhuang, L., Jing, F., Zhu, X. Y.: Movie review mining and summarization. In Proceedings of the 15th ACM international conference on Information and knowledge management, pp. 43-50. (2016).
2. Amazon online retailer web site, http://www.amazon.com (2023).
3. Rottentomatoes movie review site, http://www.rottentomatoes.com (2023).
4. Twitter social networking site, http://www.twitter.com (2023).
5. Digg social networking site, http://www.digg.com (2023).
6. Baid, P., Apoorva, G., Neelam, C.: Sentiment analysis of movie reviews using machine learning techniques. In: International Journal of Computer Applications 179, no. 7, pp. 45-49 (2017).
7. Mahesh, B.: Machine learning algorithms-a review. In International Journal of Science and Research (IJSR). [Internet] 9, pp. 381-386. (2020)
8. Hasan, Md, R., Maisha, M., Arifuzzaman, M.: Sentiment analysis with NLP on Twitter data. In 2019 international conference on computer, communication, chemical, materials and electronic engineering (IC4ME2), pp. 1-4. IEEE, (2019).
9. Dey, L., Sanjay C., Anuraag, B., Beepa, B., Sweta, T.: Sentiment analysis of review datasets using naive bayes and k-nn classifier. In arXiv preprint arXiv:1610.09982 (2016).
10. Probst, P., Anne-Laure B., Bernd B.: Tunability: Importance of hyperparameters of machine learning algorithms. In The Journal of Machine Learning Research 20, no. 1, pp. 1934-1965. (2019)
11. Harish, B. S., Keerthi, K., Darshan, H.: Sentiment analysis on IMDb movie reviews using hybrid feature extraction method. In International Journal of Interactive Multimedia and Artificial Intelligence, vol. 5, no. 5, pp. 109-114. (2019).
12. John, G. H., Langley, P.: Estimating continuous distributions in Bayesian classifiers. In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 338-345. Morgan Kaufmann. (1995).
13. Rish, I.: An empirical study of the naive Bayes classifier. In IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol. 3, pp. 41-46. (2001).
14. McCallum, A., Nigam, K.: A comparison of event models for naive Bayes text classification. In AAAI-98 Workshop on Learning for Text Categorization, vol. 752, pp. 41-48. (1998).
15. Zhang, H.: The Optimality of Naive Bayes. In Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004), pp. 562-567. (2004).
16. Pedregosa, F., Varoquaux, G., et al.: Scikit-learn: Machine learning in Python. In Journal of Machine Learning Research, 12(Oct), pp. 2825-2830. (2011).

17. Lewis, D.: Naive Bayes at Forty: The Independence Assumption in Information Retrieval. In European Conference on Machine Learning (ECML), pp. 4-15. (1998).