Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Research on AI-Generated Text Detection Based on Machine Learning Models

Moran Zeng

School of Advanced Manufacturing, Guangdong University of Technology, Guangzhou, Guangdong, 510000, China

3121009407@mail2.gdut.edu.cn

Abstract. The purpose of this research is to ensure the authenticity of information, guarantee the reliability and credibility of information sources, and prevent the spread of false information, fabricated data, and misleading content. In the academic field, detecting AI-generated papers, articles, and assignments helps maintain academic integrity, prevent academic fraud and plagiarism, and thus improve academic capabilities. This study summarize the characteristics of the three selected models, which are Logistic Regression, Support Vector Machine (SVM), and Naive Bayes (NB) Classifier. And provide recommendations and directions for improvement in the choice of detection models for AI-generated content. Through comparison of three models—logistic regression, SVM, and Naive Bayes—on the same dataset in terms of Accuracy, Precision, and F1-score, it is determined that logistic regression performs the best for this type of dataset. Logistic regression achieves superior performance with metrics exceeding 90%. SVM shows suitability for large datasets with metrics around 70% in this dataset. However, Naive Bayes, typically suitable for smaller datasets, performs poorly on this dataset, achieving only 50% accuracy.

Keywords: Support Vector Machine (SVM), Logistic Regression, Naive Bayes (NB) Classifier, Comparison, improvements, suggestions.

1. Introduction

Artificial intelligence-generated content (AIGC) technologies increasingly permeate into people's daily lives. By using these technologies, people can use natural language descriptions as input to generate various modalities of data, —such as text, code, images, speech, video, 3D models, scenes, and music [1].

The new combinations of words, phrases, and sentences generated by AIGC, including Chat GPT, are often indistinguishable from human-generated content. This convenient and efficient approach has been maliciously exploited. Some people use AI to generate fake documents. Students use AI tools to quickly complete creative assignments. And some researchers use AI tools to generate false research. These false data may lead to academic misdirection and catastrophic errors in certain fields. Furthermore, an increasing number of students are using AI tools to write papers or develop code. In this situation, these behaviors weaken students' academic abilities and affect the academic atmosphere, having a significant negative impact on the academic community. According to a survey conducted by Study.com on the use of Chat GPT by adult students (over 18 years old) in class and when completing assignments, the rate of using Chat GPT to complete assignments and perform in class was as high as 89% [2]. Therefore, detecting whether academic papers are generated by AI has become an urgent priority.

The current research in this field is in a developmental stage. From the related identification methods publicly available in this field, it can be seen that there is a lack of model training proficiency. The structures of different models vary, and they have different training times on the training set, as well as different numbers of iterations, leading to difficulties in selecting the recognition model [3].

This paper uses three different models, which are Logistic Regression, Support Vector Machine (SVM), and Naive Bayes (NB) Classifier, to predict whether the papers in the dataset are generated

by AI or completed by humans. The accuracy, macro F1, recall, and precision values of each model will be obtained, and a comprehensive comparison will be made to highlight the strengths, weaknesses, and deficiencies of each model. The results will be summarized and suggestions for improvement will be proposed. This study aims to provide effective recommendations for selecting models to detect AI-generated texts and to summarize the advantages of each model, ultimately leading to a conclusion on how to improve the determination of AI-generated texts using better models.

2. Method

2.1. Dataset

By constructing two types of papers as research data, one being academic papers written by scholars and the other being papers generated by AI. This study collects the scholar's materials from already published papers. The AI-generated papers will be created through natural language input specifying a certain paper topic and requirements. ChatGPT-4, as a leading representative of language models, is capable of top-tier content generation. Based on this, the study has decided to use ChatGPT-4 to generate the required paper materials. By providing ChatGPT-4 with natural language input for the topic, requirements, title, purpose, and author identity, it will generate Chinese paper abstracts [4].

2.2. Data Preprocess

This study selects 20 core papers from each different core professional journal as research samples. Core journals in the profession indexed by CSSCI include five different journals: Journal of Library Science in China, Journal of University Libraries, Library Forum, National Library Science Journal, and Library Science Research. These journals are highly influential in their respective fields and to some extent represent high-quality research papers in the discipline. Subsequently, specific natural language prompts and paper topics are input into ChatGPT4 to generate articles, which are then saved. It uses the following as natural language input for GPT-4."If you are a renowned scholar in the field of XX studies, I would like your assistance in drafting an abstract for a Chinese academic paper. I will provide a title for a Chinese academic paper, and I hope you can write the corresponding abstract based on this title. The first paper's title is XXX" [4].

2.3. Model

There are three models—SVM, Logistic Regression, and Naive Bayes (NB)—that will be independently used to train models to detect whether the paper was generated by AI. Once the paper abstracts are generated successfully, these models will be utilized to determine if the papers were AI-generated or not.

Logistic Regression: a linear model used for binary classification problems. It calculates the probability of the input belonging to a certain class using the logistic function. In the logistic regression model, there is no strict requirement for a strong linear relationship between the model's independent variables and the dependent variable. When making predictions, continuous dependent variables are discretized and assigned different categorical values when input into the model. Additionally, each independent variable is enhanced for interpretability through weighted matrices (W) and vectors (b), and so on [5, 6].

Support Vector Machine (SVM): a supervised learning algorithm uses for classification and regression tasks. It finds the optimal hyperplane in the feature space to maxmise the margin between different classes. SVM is a small-sample machine learning method based on statistical learning theory. Unlike traditional machine learning methods, SVM does not only consider minimizing empirical risk but instead trains by evaluating structural risk minimization. In a limited dataset, the SVM model achieves an optimal trade-off between model complexity and learning capability to attain the best generalization performance, thereby completing the detection of AI-generated text [7].

Naive Bayes Classifier: Based on Bayes' theorem and the assumption of conditional independence between features. For AI generation detection, assuming independence among each feature word, this study applies the Naive Bayes model to detect AI-generated texts [8].

Each model has its unique advantages and application domains, so selecting the most suitable model depends on the specific problem characteristics and data features in practical applications.

Support Vector Machine (SVM) Performs well with high-dimensional and complex data, capable of effectively handling non-linear problems. SVM can adapt to different types of data features. It has strong generalization capabilities, suitable for small-sample and high-dimensional classification problems.

Logistic Regression is Simple and efficient, fast computation, easy to understand and implement. Suitable for handling linearly separable or approximately linearly separable data.

Naive Bayes Based on probability principles, performs well with small-scale data, and has good interpretability. Particularly effective in handling text classification and natural language processing problems, especially under the assumption of independence between features. Fast in training and prediction, insensitive to missing data.

2.4. Experiment

Due to the difficulty of machines in understanding natural language, this study utilizes the Term Frequency-Inverse Document Frequency (TF-IDF) method to transform text into vectors. The TF-IDF algorithm is a statistical method used to assess the importance of a word in a document set or a document in a corpus. It is typically employed to extract features from text, i.e., keywords. The importance of a word increases proportionally to its frequency in the document but is offset by the frequency of the word in the corpus. For instance, if a term or phrase has a high frequency (T) within an article but appears infrequently across other articles (thus having a low IDF), it is considered to have good discriminatory power and is suitable for text classification [9].

SVM, NB, and Logistic Regression methods are employed for prediction. The study divides the dataset into 30% for testing and 70% for training. Accuracy, Macro F1, Recall, and Precision values serve as metrics to evaluate the performance of different methods.

3. Result

The Accuracy, Precision, and F1-score values obtained after predicting papers written by scholars and those generated by ChatGPT-4 using the three classifiers are presented in Table 1 [4].

Table1. The comparison of performance evaluation results for three models.

Model	Accuracy	Precision	F1-Score
SVM	73.33%	70.97%	73.33%
NB	56.67%	52.73%	69.05%
LOGISTIC	93.33%	90.32%	93.33%

Based on the data, it is observed that Logistic Regression achieves all three metrics above 90%, demonstrating the best performance among the three classifiers. Specifically, it achieves an Accuracy and F1-score of 93.33%, effectively identifying AI-generated papers. SVM follows with all three metrics above 70%, showing relatively high accuracy in detecting AI-generated texts but with some gap compared to Logistic Regression. Lastly, Naive Bayes (NB) achieves an Accuracy of 56.67%, Precision of 52.73%, and F1-score of 69.05%, indicating limited capability in effectively identifying AI-generated texts in this dataset. Among the three models, Logistic Regression performs the best, followed by SVM, and lastly Naive Bayes [4].

4. Discussion

Based on the experimental data of this study, it was found that among the three models -SVM, NB, and Logistic Regression, Logistic Regression demonstrated the best performance and suitability for detecting AI-generated texts. The dataset used in this study is nonlinear, and Logistic Regression performs well in nonlinear data classification scenarios, showcasing its effectiveness in this context. It was observed that changes in the independent variables led to differences in performance evaluation metrics, emphasizing the importance of selecting optimal independent variables to achieve higher accuracy in future Logistic Regression studies focused on AI text detection.

Regarding SVM, its performance indicates applicability in AI text detection under supervised environments. However, due to the relatively small size of the training set in this study, SVM did not achieve its optimal state. It is expected that SVM's performance metrics will improve with larger datasets. Therefore, improving SVM's training under weakly supervised scenarios should be a future research focus, given that real-world datasets often fall into such categories.

On the other hand, NB exhibited lower performance metrics in accurately detecting AI-generated articles in this study. The dataset used was moderate in size, and NB's performance benefits from the assumption of attribute independence, making it more suitable for smaller datasets. However, its performance tends to decrease with larger datasets. Future research could explore enhancing NB's performance by mitigating the impact of the independence assumption, such as through techniques like the Naive-Bayes Tree (NBTree) hybrid model, as suggested by Kohavi, which integrates NB within decision trees to reduce the constraints of its independence assumption [10].

5. Conclusion

The study concludes that the use of SVM, NB, and Logistic Regression models for detecting AI-generated text should be tailored to different datasets to achieve optimal performance. Blindly selecting a model may lead to issues such as low accuracy and inability to predict effectively. Among them, the Logistic Regression model performed the best, achieving accuracy, precision, and F1 scores all above 90%. This indicates that under this method, machine learning models can effectively distinguish between scholarly and GPT-4 generated abstracts. For non-linear datasets, this model provides the best results.

When dealing with large datasets, the SVM model should be chosen. Extensive training on large training sets allows SVM to perform optimally, whereas in smaller datasets, its performance may drop, as seen in this study with scores just above 70%. Conversely, the Naive Bayes model did not perform well in this study, achieving metrics slightly above 50% and failing to achieve high precision predictions. However, for smaller datasets, Naive Bayes can perform better. For future improvements to this model, developing hybrid models could enhance performance on small datasets.

Currently, there are numerous applications in various fields for detecting AI-generated content, such as OpenAI's GPT-3/GPT-4 Detector, which analyzes text features to assess whether they could be generated by GPT models. These tools are widely used in social media platforms, news agencies, and education to identify automatically generated content. Additionally, tools like Originality.AI are used in content creation platforms, academic publishing, and online education to detect AI-generated content and plagiarism.

In sectors like government regulation, financial services, and brand protection, Sensity AI is employed for deepfake detection, covering not only videos and images but also extending to textual content. Sensity AI offers a range of tools for identifying and labeling AI-generated content.

It must be acknowledged that while generative AIs like ChatGPT-4 can mimic scholarly writing to some extent, their limitations in professional knowledge, data validity, and theoretical understanding make them unable to fully replace scholars. AIGC significantly enhances productivity and information retrieval capabilities but misuse can lead to plagiarism and questioning the authenticity of generated materials. Developing AI generation detection can mitigate such risks to some extent.

References

[1] Xibin S, Lilei W. Research on the detection of ai-generated academic journal texts. Science and Publication, 2023, (08): 56-62.

[2] Tangermann V. 89 Percent of college students admit to using ChatGPT for homework, study claims wait, what!? [2023-04-27]. Available at: https://futurism.com/the-byte/students-admit-chatgpt-homework.

[3] Zhou M. Technical Defects of AIGC Paper Detection System and Responses of Academic Journals. Publishing and Printing, 2024, 1-10.

[4] Yibo W, Xin G, Zhifeng L, et al. Detection and comparative study of ai-generated and scholar-written chinese paper abstracts: a case study in library science. Journal of Information, 2024, 1-8.

[5] Hanxia Z. Scenario analysis suitable for linear regression and logistic regression. Automation & Instrumentation, 2022 (10): 1-4+8.

[6] Nibbering D, Hastie T J. Multiclass-penalized logistic regression. Computational Statistics & Data Analysis, 2021.

[7] Shifei D, Yuting S, Zhizhen L, et al. Review of support vector machine algorithm under weak supervision scenarios. Journal of Computer Science, 2024, 1-25.

[8] Bowen Z. Research on text classification algorithm based on naive bayes method. Xiangtan University, 2021.

[9] Sang X, He J, Chen M. Real-time monitoring and modeling of online public opinion based on tf-idf and lsi models. Mathematics in Practice and Theory, 2022, 52(11): 56-66.

[10] Hall M. A decision tree-based attribute weighting filter for naive bayes. Knowledge-Based Systems, 2007, 20: 120–126.