Proceedings of Machine Learning Research vol 245:1–11, 2024 2024 International Conference on Machine Learning and Intelligent Computing

ChatGPT Plagiarism in the Academic Field: Exploration and Analysis of Plagiarism Effects

Haiqiong Luo D22091100034@CITYU.EDU.MO
Faculty of Data Science, City University of Macau, Macau 999078, China
Hoiio Kong* HIKONG@CITYU.EDU.MO
Faculty of Data Science, City University of Macau, Macau 999078, China

Editors: Nianyin Zeng and Ram Bilas Pachori

Abstract

This article explores the plagiarism problem of ChatGPT in the education field in detail and proposes a study to evaluate its plagiarism effect. First, it introduces ChatGPT as an important innovation based on GPT technology, its application and potential impact in the field of natural language processing. The article then focuses on the plagiarism concerns caused by ChatGPT in academia and points out the countermeasures some schools have taken. Then, the article raises research questions, aiming to explore the actual effect of ChatGPT in paper plagiarism detection, and describes the research methods and premise assumptions in detail. By rewriting and plagiarism checking analysis of a series of papers in the education field, the study attempts to evaluate ChatGPT's performance in terms of plagiarism and explore factors that may affect the degree of plagiarism. Finally, the article conducts a detailed statistical analysis of the experimental results and puts forward suggestions for further research to emphasize the importance of academic integrity and the call for the correct use of ChatGPT.

Keywords: ChatGPT, Academic Plagiarism, Paper Rewriting, Education, Statistical Analysis

1. ChatGPT: The Rise of Artificial Intelligence in Education and Concerns about Plagiarism

ChatGPT demonstrated the ability of artificial intelligence models to generate content, and articles were quickly published about its possible uses and potential controversy. (S, 2022; F, 2022; A, 2022) Early adopters shared their experiences on social media, and the sentiment was mostly positive. (Haque et al., 2022) Articles lament the demise of traditional school essay assignments, (A, 2022; Stokel-Walker, 2022; E, 2022) as ChatGPT has been shown to produce high-scoring essays (Yeadon et al., 2023) and even express critical thinking. [8] The ethical and acceptable boundaries of ChatGPT use in scientific writing remain unclear [9].

2. Explore the Potential and Limitations of ChatGPT in Paper Plagiarism

2.1. Research Questions

This study aims to deeply explore the potential of ChatGPT in paper plagiarism to fully understand its plagiarism effect and conduct a corresponding detailed analysis. Through this research, it is expected to provide a more in-depth knowledge and understanding of ChatGPT's plagiarism problem in the academic field.A comprehensive and systematic assessment of the plagiarism effect of ChatGPT is not just a subjective judgment lacking scientific basis. Instead, rewrite the text and test it to see how it actually works. Draw objective conclusions by evaluating the rewritten results and combining them with scientific analysis.

2.2. Premise

In the study, the text rewriting method using ChatGPT followed specific rules and restrictions.

1. The rewriting process only focuses on the abstract and main text of the paper, and will not involve the rewriting of codes, titles, figures, formulas or references.

2. ChatGPT is used to rewrite every sentence of the entire article to ensure that the full text content is covered.

3. During the rewriting process, special attention needs to be paid to ChatGPT's rewriting of each sentence to ensure that each sentence has been adjusted accordingly. Although the number of words may increase or decrease after rewriting, this change will not affect the analysis and results of the study.

4. Importantly, this method is intended to rewrite the paper to verify ChatGPT's ability to rewrite text, rather than modify the data, code or references.

3. Research on the Effect of ChatGPT in Plagiarism of Educational Papers: Case Analysis Based on Peking University Core and CSSCI Articles

3.1. Research Process

1. When determining the scope of the research, we chose the field of education for the following reasons: First, we learned that papers in history, philosophy, education and other fields are mainly based on text descriptions and rarely involve a large amount of codes, charts, data or formulas. Therefore, during the rewriting process, deleting these contents will have little impact on the overall content of the article. Secondly, after reviewing the literature, we found that ChatGPT is widely used in education (Baidoo-Anu and Ansah, 2023; Tlili et al., 2023; Lo, 2023) and has a good rewriting effect on papers in the field of education, so we selected papers related to education as the research object.

2. When determining the time range for selecting papers, we selected China-related education papers, with the time range ending in 2020. These papers were published in Peking University core and CSSCI journals. We gave priority to papers with more citations and downloads to ensure the quality and influence of the research objects.

3. We collected fifty Chinese papers in the field of education, extracted their titles, abstracts, and text content, and deleted citations, references, charts, and other parts. The title, abstract, and text of each paper are placed in one paragraph to facilitate subsequent rewriting. We used ChatGPT to rewrite the abstracts and texts of these fifty papers twice, that is, to rewrite fifty paragraphs of text, and thus obtained the rewritten paper content.

4. In order to evaluate the rewriting effect, we used China National Knowledge Infrastructure (website of CNKI personal plagiarism check) to check the original text and the rewritten content for plagiarism.

5. Since the number of words in the original and rewritten 50 papers exceeds the word limit for plagiarism checking on CNKI (website of CNKI personal plagiarism check) , the titles, abstracts, introductions and conclusions of the original and rewritten papers were extracted as the content for plagiarism checking.

6. Put it into CNKI (website of CNKI personal plagiarism check) for plagiarism checking. Through the plagiarism checking report, we first made a simple comparison between the original text and the rewritten content of the fifty papers, and conducted statistical analysis.

7. Finally, we draw corresponding conclusions by analyzing the results.

3.2. Overview of CNKI Education Paper Information

The obtained paper was downloaded from China National Knowledge Infrastructure. Figure 1 is the detailed information of the paper. There are 50 papers in total. Only five of them are shown here.

Serial Number	Author	Article Name	Issuing Time	Number of citations	Download Times
1	Zhang Yan	On the Concept and Mode of "Internet Plus Education"	2016/2/20	1235	35104
2	TAN Chuanbao	Understanding the Concept of Labor Education: How to Understand the Basic Connotation and Characteristics of Labor Education	2019/2/10	1180	42144
3	Zhou Guping, Kan Yue	The Talents Supporting and Educational Solutions to the "Belt and Road Initiative"	2015/10/15	542	23878
4	Ye Haosheng	The Body and Learning:Embodied Cognition and Its Impact on Traditional View of Education	2015/4/15	760	20511
5	Xu Changfa	The Re-Development Logic of the Labor-Education in the New Era	2018/11/15	721	23794

Figure 1: Information about the papers used in this study. (These papers are public on CNKI).

4. Study the Experimental Process

First, use ChatGPT to rewrite each paper twice, sentence by sentence. The specific operation is shown in Figure 2.

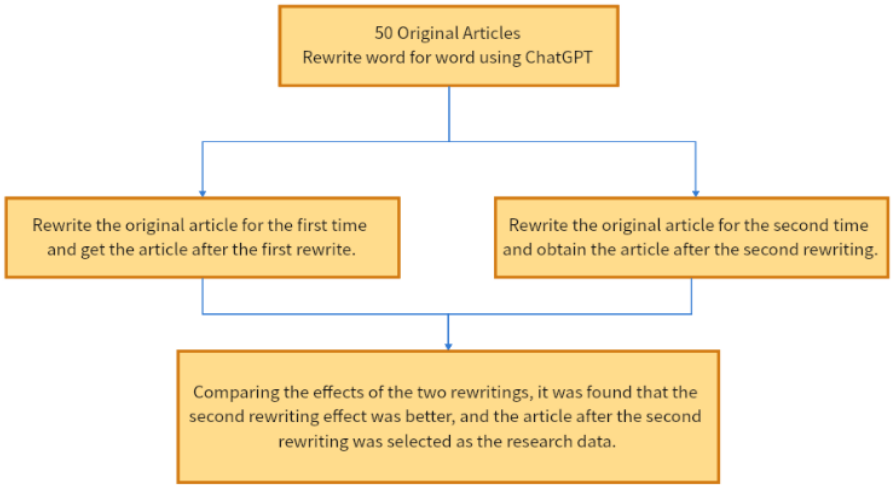

Figure 2: Rewrite the process with ChatGPT.

Secondly, after accumulating experience and summarizing the rewriting process, we found that using ChatGPT for the second rewriting will be better than the first time. Therefore, we selected the results of the second rewrite and finally obtained the following paper rewrite results. In this result,

dark red is used to identify the title, blue is used to identify the abstract, and black is used to indicate the text content. Then, in the fifth point of explanation in Section 4, we need to extract the above-mentioned original text and the rewritten title, abstract, introduction and conclusion as the content for plagiarism checking. In this process, we will use different colors to identify the content of each part. Dark red represents the title, blue represents the abstract, green represents the introduction, and black represents the conclusion.

Subsequently, we will use China National Knowledge Infrastructure (website of CNKI personal plagiarism check) to check for duplication of the integrated original text and rewritten content. Among them, the even-numbered paragraphs are the original content, and the odd-numbered paragraphs are the rewritten content. Please check the corresponding duplication rate.

5. Perform Data Analysis on the Results

5.1. Summary of Plagiarism Check Results

Perform statistical analysis on the plagiarism check rate and compare before and after to understand the changes in the plagiarism check rate after increasing the number of papers.

We found that after rewriting using ChatGPT, the average duplication check rate dropped by 12.16%, with the maximum drop reaching 73.4%, and the minimum drop being -2.5%. It is worth noting that negative numbers indicate that the rewriting of ChatGPT may lead to an increase in the duplication check rate of articles. In order to gain a deeper understanding of the factors influencing the rewriting effect of ChatGPT, we conducted the following exploratory analysis.

5.2. Conduct Exploratory Analysis

5.2.1. EXPLORE THE FACTORS THAT AFFECT THE REDUCTION RATE OF DUPLICATION DETECTION

We will use Python to conduct a preliminary exploration of the above content, focusing on the impact of publication time, word count difference before and after rewriting, and number of citations on the duplication reduction rate to analyze whether there is a causal relationship between them.

We will draw scatter plots to observe the relationship between the duplication detection rate before rewriting, the duplication reduction rate, the number of citations, the word count difference before and after rewriting, and the time interval. The time interval here refers to the difference in days between the current time and the publication time. The specific effect is shown in Figure 3 with a total of 25 sub-pictures.

Among them, in the subgraph (a) (g) (m) (s) (y) in Figure 3, the histogram represents the distribution of each variable in the data set, showing the distribution of each variable itself. The x-axis of the histogram represents The value range of the variable, and the y-axis represents the frequency or density of the corresponding values. It can also be seen from Figure 3 that the duplication checking rate and reduction rate before rewriting, the number of citations, the difference in word count before and after rewriting, and the time interval do not show an obvious trend, which indicates that the relationship between the independent variables and the dependent variables may be nonlinear, rather than linearly.

We extract a specific scatter plot subgraph (u) and briefly explain it. It can be observed from the figure that as the duplication check rate before rewriting increases, the time interval does not show an obvious trend, which may mean that there is no linear relationship between the independent variable

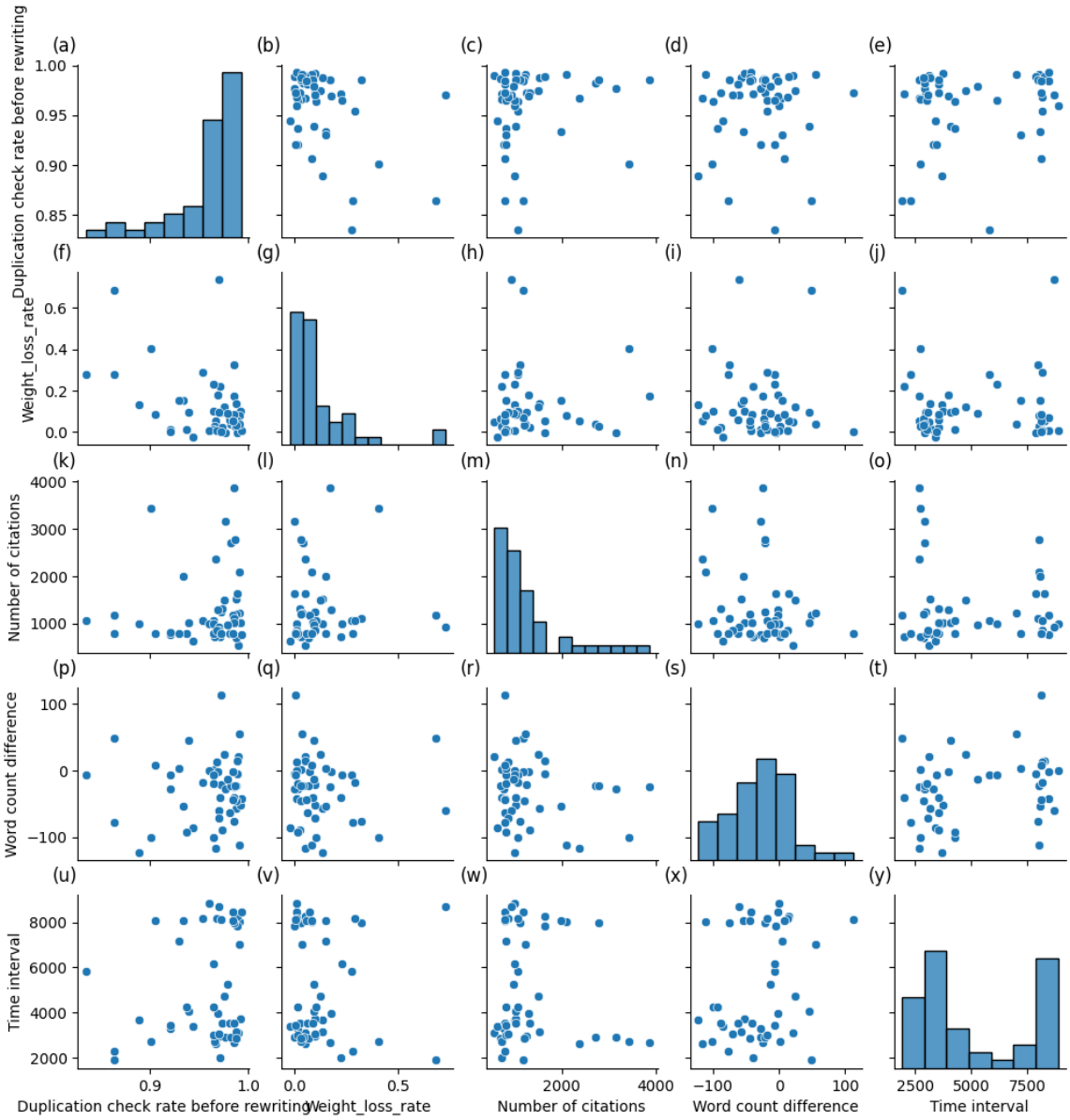

Figure 3: Histograms of duplication check rate before rewriting(a), weight loss rate(g), number of citations(m), word count difference(s), time interval(y). The bivariable relationship between this five variables(other subfigures).

and the dependent variable. Likewise, we can also observe correlations between several other sets of variables. In summary, we conclude that there is no obvious linear relationship between these variables.

5.2.2. VISUAL DISPLAY OF VARIABLE CORRELATION

We plotted a heat map of five variables, using heat maps to visually display the correlation between variables. In a heat map, variables with higher correlations are darker in color, while variables with lower correlations are lighter in color. The specific effect is shown in Figure 4. Observing the color distribution in the figure, we can see that the general colors are lighter, but the colors between the reduction rate and the duplicate check rate before rewriting are darker. This is because the reduced duplication rate is calculated from the pre-rewrite duplication rate.

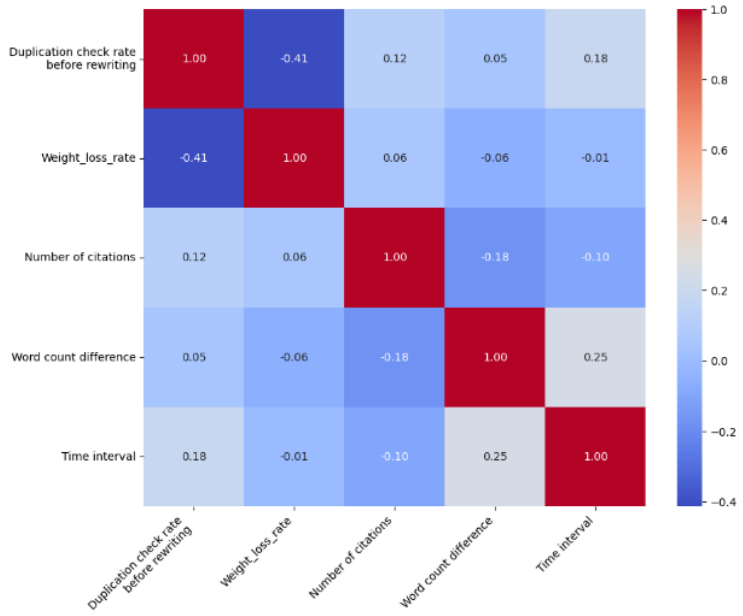

Figure 4: Heat map of the correlation matrix for the five variables.

5.2.3. SCATTER PLOT ANALYSIS AND NONLINEAR RELATIONSHIP VERIFICATION OF FITTED LINEAR REGRESSION MODEL

This plot is similar to a normal scatter plot, but with the addition of a fitted linear regression model. This straight line represents the best fit of a linear relationship between the independent and dependent variables. These fitted lines show the linear relationship between the independent and dependent variables. Specifically, for each subgraph, the model fits a straight line based on the given independent variable data, so that the predicted value of the dependent variable is as close as possible to the actual observed value. These straight lines are the fitting lines of the linear regression model, and they are used to represent the trend of the linear relationship between the independent variable and the dependent variable.

By observing this fitted line, you can gain a clearer understanding of the relationship between the independent and dependent variables, as well as the accuracy of the fitted model. If the fitted linear regression model roughly matches the distribution of points in the scatter plot, it probably means that the linear model fits the data well.

First, we drew scatter plots of the linear regression model for weight reduction rate, number of citations, difference in word count before and after rewriting, and time interval. The line in each subfigure represents the fitting line of the linear regression model. The line in each subfigure represents the fitting line of the linear regression model. The shaded area represents the confidence interval of the linear regression model. As shown in Figure 5a, b, c, each subfigure represents the linear relationship between an independent variable and the dependent variable. It can be observed that the scatter points of the three graphs do not completely match the fitting lines, and there are many outliers. This indicates that the relationship between the independent variable and the dependent variable may be non-linear rather than linear.

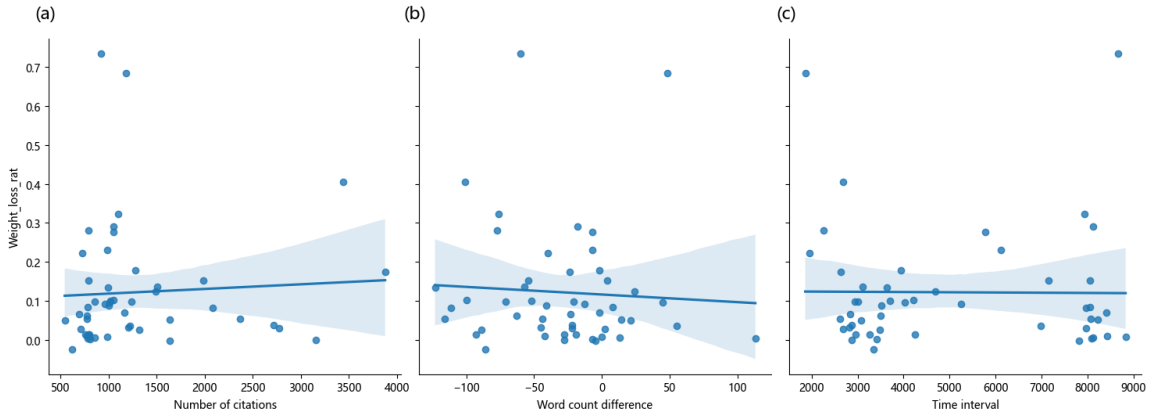

Figure 5: The linear relationship between number of citations(a), word count difference(b), time interval(c) and weight loss rate.

5.2.4. CONDUCT A SERIES OF MULTIPLE REGRESSION ANALYSES

(1) Multiple linear regression analysis

First, we use the trained model for fitting and get the following fitting multiple linear regression equation (1), where y represents the drop-down rate, x_1 represents the number of citations, x_2 represents the difference in word count before and after rewriting, and x_3 represents the time interval.

In the study, we explored the relationships between variables. First try to use the linear fitting method to obtain equation 1 to get a preliminary understanding of the interaction between them. However, subsequent validation analyzes showed that these variables did not follow a linear relationship. Therefore, the research focus shifted to an in-depth exploration of other types of relationships that may exist.

$$y = 0.10 + 1.02 * 10^{-5}x_1 - 0.0002x_2 + 6.1 * 10^{-7}x_3 \quad (1)$$

Next, we plotted a scatter plot of actual values versus predicted values, as shown in Figure 6a. Ideally, all points should be distributed along the diagonal ($y = x$), indicating that the predicted values are exactly the same as the actual values. However, in fact, the entire scatter point is not distributed along the diagonal ($y = x$), which indicates that the fitted straight line of the multiple linear regression model is not significant.

Next, we plotted the residual plot, as shown in Figure 6b, to check the fitting effect of the model. In a residual plot, the residuals should be evenly distributed above and below the horizontal reference line as the predicted value increases. However, in fact, the entire scatter points are not evenly distributed on the horizontal line, which indicates that the fitted straight line of the multiple linear regression model is not significant.

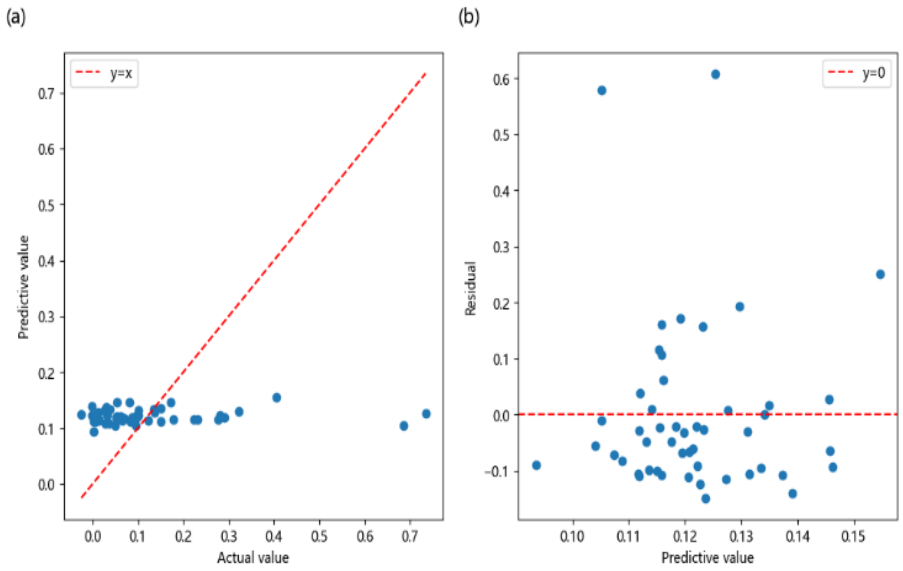

Figure 6: The relationship between actual values and predicted values(a), the relationship between predicted values and residuals from linear regression analysis.

(2) Multivariate Ridge Regression Analysis

When there is a multicollinear relationship between the independent variables of the regression equation, the ordinary least squares method (i.e. linear regression anal-ysis) can no longer be used to accurately analyze the regression equation. As early as 1962, Heer proposed an improved method. The least squares estimation method is called ridge regression. If there are multiple correlations between independent varia-bles, the ridge regression estimation method is a relatively stable method, and the standard deviation of the regression coefficient estimated by ridge regression is also small. (Hoerl and Kennard, 2000)

If you want to use ridge regression for analysis, you need to first determine whether there is a multicollinear relationship between the independent variables, and you can calculate the variance inflation factor (VIF). The variance inflation factor (VIF) method (Jiahui ZHU, 2021) is a method used to evaluate the linear relationship between model input variables. (Ma et al., 2024) The larger the VIF value, the more serious the collinearity between variables. Generally speaking, a VIF

greater than 10 may indicate the presence of multicollinearity. It can be calculated, as shown in Table 1. It can be seen that the VIFs are very small, indicating that the independent variables do not have multicollinearity, so the effect of using ridge regression analysis is the same as that of linear regression analysis.

Table 1: VIF value table

Feature	VIF
Number of citations	2.78752
Word count difference	1.389341
Time interval	2.360176

(3) Multiple Lasso regression analysis

Lasso regression was used to analyze the weight reduction rate, number of citations, word count difference before and after rewriting, and time interval, and the fitting equation 2 was obtained as follows. Among them, $y$represents the reduction rate, x_1 represents the number of citations, x_2 represents the difference in word count before and after rewriting, and x_3 represents the time interval. And it is not much different from the equation fitted by multiple linear regression. It is guessed that Lasso regression analysis is not suitable for this set of variables. Similarly, other fitting attempts were also made to Equation 2 to explore potential relationships between variables.

$$y = 0.108 + 1.02 * 10^{-5}x_1 - 0 * x_2 - 7.97 * 10^{-8}x_3 \tag{2}$$

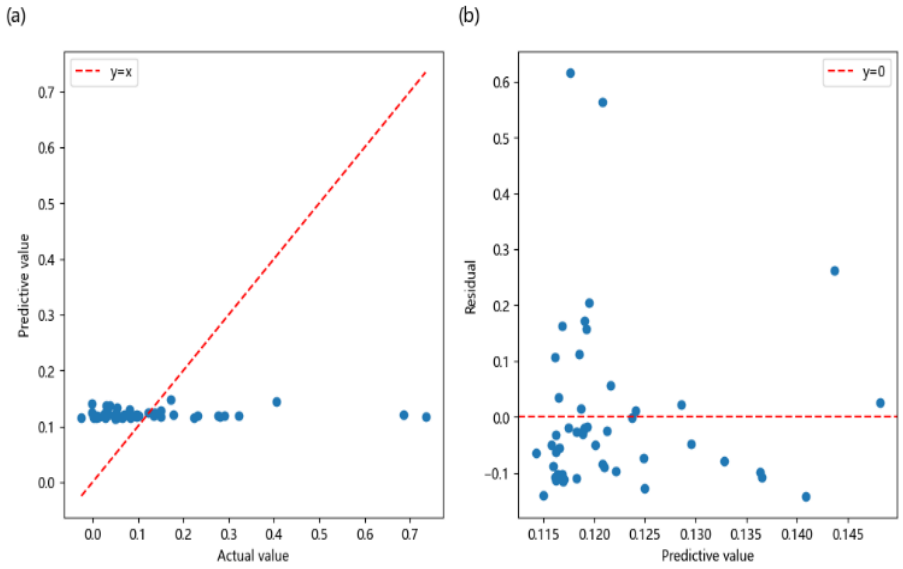

Figure 7: The relationship between actual values and predicted values(a), the relationship between predicted values and residuals(b) from Lasso regression analysis.

In order to verify the above conjecture, a scatter plot of actual values versus predicted values was drawn. As shown in Figure 7a, the entire scatter point is not distributed along the diagonal ($y = x$), which indicates that the fitted straight line of the multivariate Lasso regression model is not significant.Next, the residual plot is drawn, as shown in Figure 7b, to check the fitting effect of the model. In fact, the entire scatter points are not evenly distributed on the horizontal line, which indicates that the fitted straight line of the multivariate Lasso regression model is not significant.

5.3. Summary

Therefore, in summary, we can draw the conclusion of data analysis: no matter we use any multiple regression analysis method, the fitting effect is not good, and the fitting straight line is not significant. This shows that the duplication detection reduction rate has nothing to do with these factors. And it is calculated that after ChatGPT was rewritten, the duplication checking rate was reduced by 12.16% on average.

6. Conclusion

This article aims to analyze the effect of using ChatGPT to plagiarize educational articles, and explore the specific numerical expression of the degree of plagiarism, as well as the factors that may affect the degree of plagiarism.First of all, we can confirm that using ChatGPT to reduce the duplication of articles is effective. On average, the duplication rate of each article can be reduced by about 12.16%. Secondly, when testing ChatGPT, we observed that as the number of rewrites increases, the rewrite effect will further improve. However, through the above multiple linear regression analysis, multiple ridge regression analysis, and multiple Lasso regression analysis, we found that the fitting effects are poor, and the fitting straight line is not significant, indicating that the reduced duplicate checking rate does not seem to be significantly related to other factors. Mainly Depends on the automatic rewriting effect of ChatGPT, this effect cannot be objectively analyzed through data.

Regarding the rewriting effect of ChatGPT, I think there are still some areas that can be improved in the experiment. First, we can conduct comparative experiments on the number of rewrites, such as comparing the effects of two rewrites and three rewrites, to verify whether the rewrite effect will further improve as the number of rewrites increases. Secondly, due to the large number of words in the fifty articles put together, CNKI cannot check all the contents for plagiarism, so we only selected the title, abstract, introduction and conclusion of each article for plagiarism checking. This selection method may affect the overall duplication checking effect, so we need to consider other methods to conduct duplication checking more comprehensively. Finally, we must remember the importance of academic integrity and plagiarism. When using ChatGPT to rewrite an article, ChatGPT is actually plagiarizing. We should maintain a sense of awe and follow academic ethics. At present, there are no clear regulations and instructions for the use of ChatGPT, but many people May misuse its functionality. Therefore, I hope that every academic practitioner can use ChatGPT with the correct attitude and method and draw reasonable conclusions from it.

References

Hern A. Ai bot chatgpt stuns academics with essay-writing skills and usability. 2022.

David Baidoo-Anu and Leticia Owusu Ansah. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *SSRN Electronic Journal*, 2023.

Whitford E. Here's how forbes got the chatgpt ai to write 2 college essays in 20 minutes. 2022.

Agomuoh F. Chatgpt: how to use the viral ai chatbot that took the world by storm. *Digital Trends*, 2022.

Mubin Ul Haque, Isuru Dharmadasa, Zarrin Tasnim Sworna, Roshan Namal Rajapakse, and Hussain Ahmad. "i think this is the most disruptive technology": Exploring sentiments of chatgpt early adopters using twitter data, 2022.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42:80 – 86, 2000.

Liying YU Jiahui ZHU. Coupling synergy measure of sci-tech innovation and financial development in china: screening based on vif-variation coefficient. *Journal of Shanghai University (Natural Science Edition)*, 27(4):785, 2021. doi: 10.12066/j.issn.1007-2861.2169.

Chung Kwan Lo. What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, 13(4), 2023. doi: 10.3390/educsci13040410.

[Liangyu Ma, Dongyan Cheng, Shuyuan Liang, Yanzhu Geng, and Xinhui Duan. Input feature selection method for wind turbine fault diagnosis based on lightgbm-vif-mic-sfs. *Thermal Power Generation*, 53:154–164, 2024.

Shankland S. Chatgpt: Why everyone is obsessed this mind-blowing ai chatbot. *CNET*, 2022.

C. Stokel-Walker. Ai bot chatgpt writes smart essays - should professors worry? *Nature*, 2022. doi: 10.1038/d41586-022-04397-7.

Ahmed Tlili, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T. Hickey, Ronghuai Huang, and Brighter Agyemang. What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education. *Smart Learning Environments*, 10(1):15, 2023. doi: 10.1186/s40561-023-00237-x.

Official website of CNKI personal plagiarism check. URL https://cx.cnki.net.

Will Yeadon, Oto-Obong Inyang, Arin Mizouri, Alex Peach, and Craig P Testrow. The death of the short-form physics essay in the coming ai revolution. *Physics Education*, 58(3):035027, apr 2023. doi: 10.1088/1361-6552/acc5cf.