Analyzing True Positive Rates of AI-Detectors for AI-Generated Text

Exploring the Impact of Article Types and Word Counts on the Accuracy of AI-Detectors in Identifying AI-Generated Text

Elijah Castro

June 10, 2023

Abstract

Our study aims to explore the differences in true positive rates between two AI-detectors designed to identify AI-generated text. Our investigation focuses on the relationship between the true positive rate and the choice of AI-detector, as well as its dependence on the article type and document length. Utilizing a logistic regression model, we identified significant predictors for the correct classification of documents by AI-detectors. We found that both the article type and the number of words in a document played crucial roles, while the specific AI-detector used did not demonstrate a significant association. Furthermore, our findings revealed that a specific threshold for the number of words and a particular article type greatly influenced the success rate of the AI-detector in correctly classifying documents. These insights offer valuable guidance to researchers aiming to improve the true positive rates of these particular AI-detectors.

Contents

Introduction

We aim to explore the factors that impact the AI-detector's ability to correctly classify documents as AI-generated. Our study utilizes data from two AI-detectors, three types of articles, and seven different levels of number of words, where each observation corresponds to one combination of AI-detector, topic area, and word count. Displayed below is a description of each variable included in the study.

Table 1: Variable Description and Notation

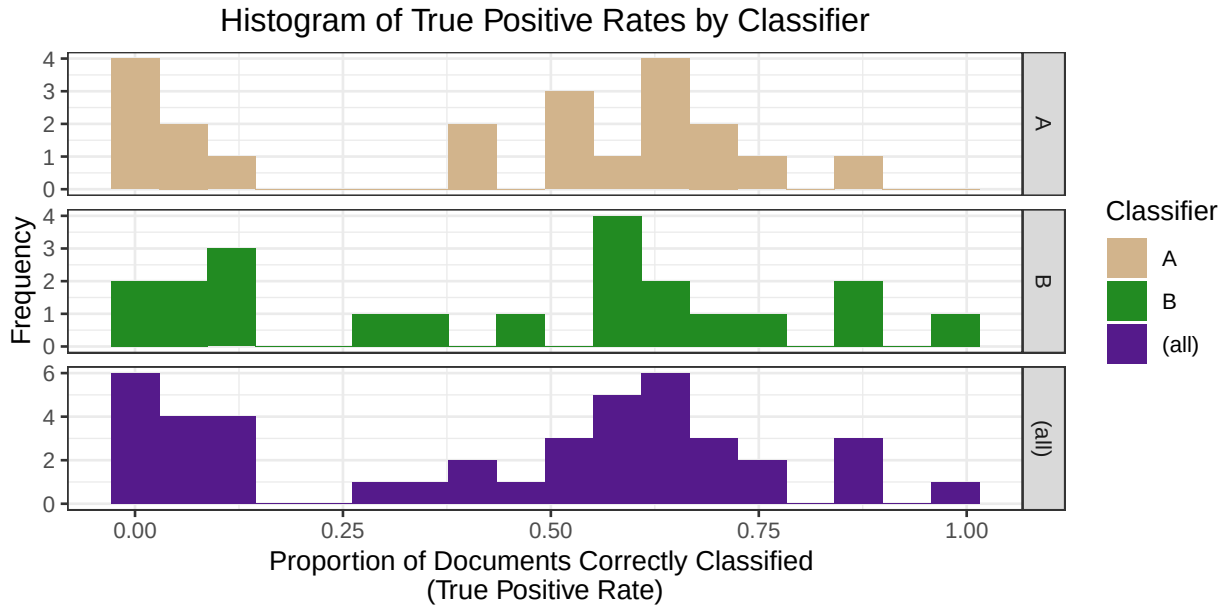
Variable	Description
mDoc	Total number of AI-generated documents studied for this row (18 in each row)
nPos	Number of documents that were correctly classified as AI-generated (True-positives) out of these 18 AI-generated documents
Classifier	Specific AI-detector used to classify these documents (either "A" or "B")
ArticleType	Type of Article ("FictionalNovel", "ArtHistory", and "Medical")
Words	Number of words in document divided by 1000 (0.1, 0.5, 1, 5, 10, 50, and 100, e.g, 0.1 corresponds to 100 words in document)

In the case of *mDoc*, all 18 documents within each row were AI-generated. Consequently, any document not classified as AI-generated would be considered a false negative. Additionally, it is important to note that distinct sets of 18 documents were evaluated for each observation, ensuring that there is no overlap between the documents represented in different observations within the dataset. Therefore, we can conclude that the data used in this study can be considered independent.

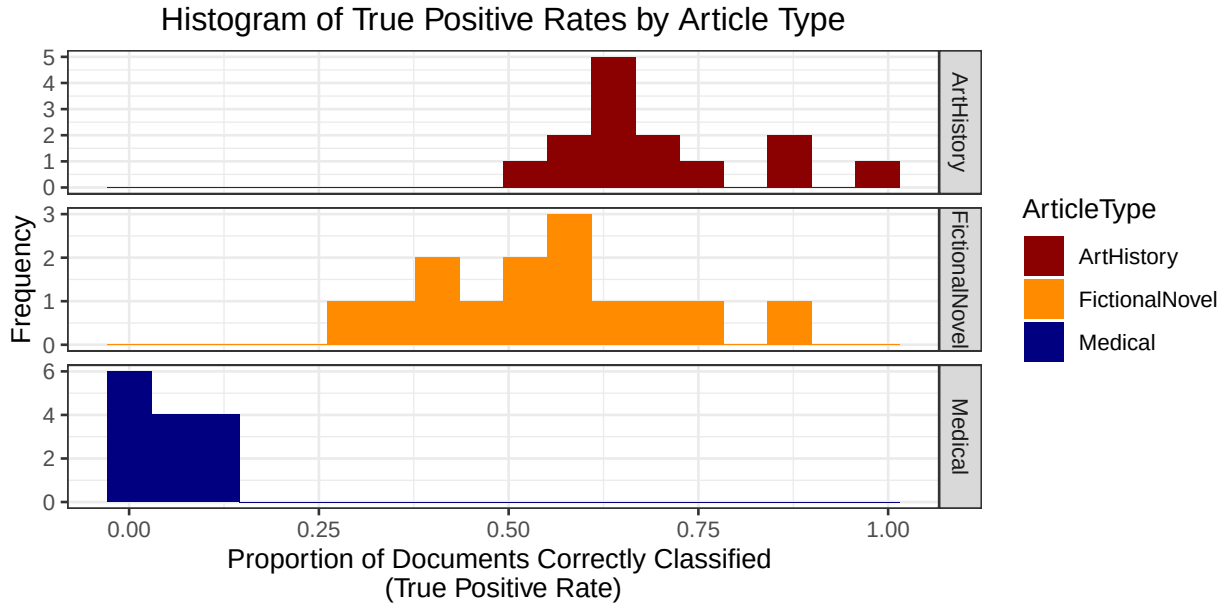
The main objective of this study is to develop a framework that accurately describes the relationship between the independent variables, namely *Classifier*, *ArticleType*, and *Words*, and the response variable (*nPos*, *mDoc* - *nPos*) for the AI-detectors investigated. This will be achieved through the application of various methods, including exploratory data analysis, model selection, and diagnostics. To conclude, we will present the results of the selected regression model and provide interpretations relevant to the research question at hand.

Exploratory Data Analysis

To explore the data and its relationships, we will first display histograms of the true positive rates by classifier and by article type.

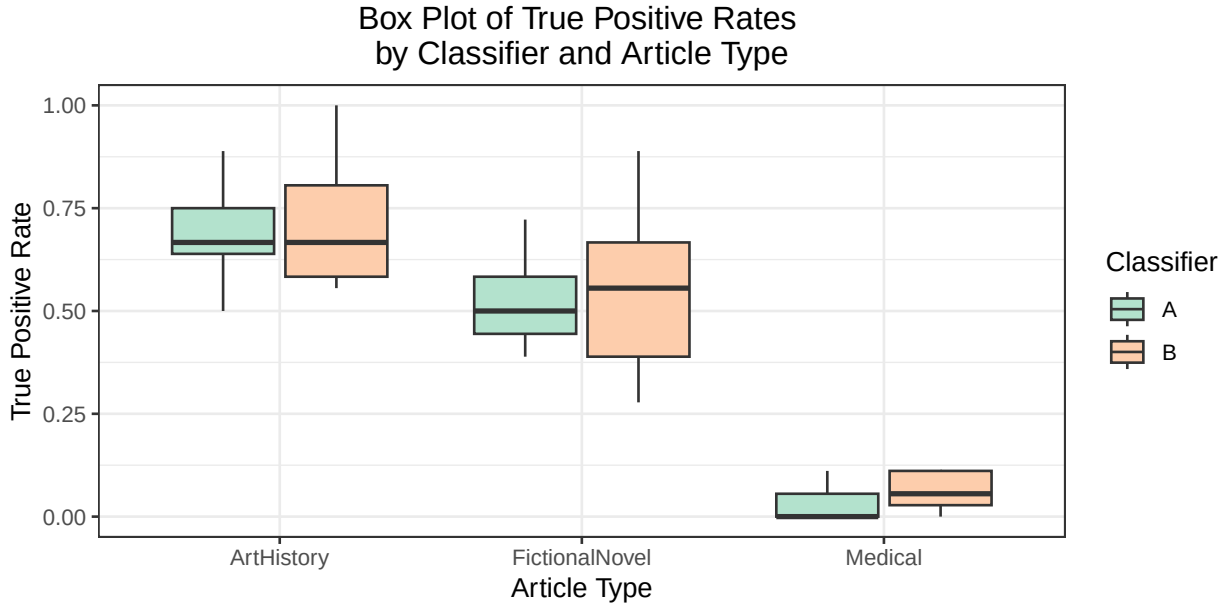


Above, we notice that the true positive rates for classifier A are slightly lower in comparison to classifier B. This suggests that, on average, classifier A has a somewhat lower tendency to correctly classify AI-generated documents compared to classifier B. This distinction, although extremely subtle, remains rather consistent across the range of true positive rates. Moreover, it is worth noting that among the classifiers in this dataset, only AI-detector B achieved the feat of correctly classifying all 18 documents, making it quite a unique observation.



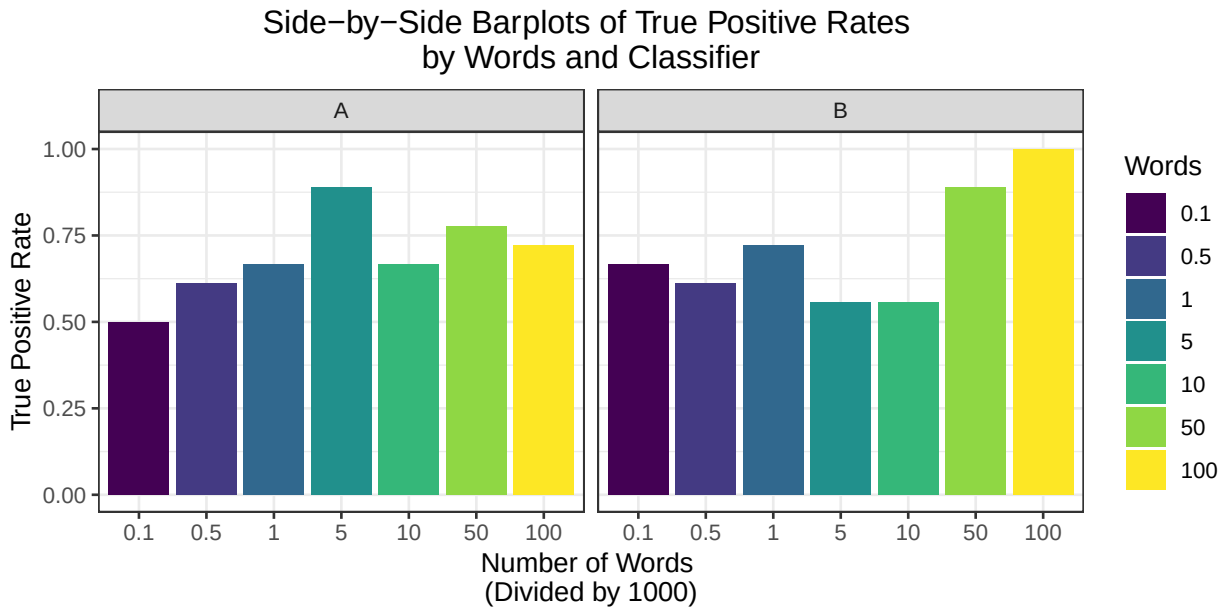
From this histogram, each article type exhibits distinct patterns in terms of their respective rates of correctly classifying documents. Art history articles consistently outperform the other types, showcasing a higher proportion of true positive rates. Fictional novel articles fall in the intermediate range, demonstrating a moderate level of success in correctly classifying documents as AI-generated. In contrast, medical articles face notable challenges, as they exhibit significantly lower rates of correctly classifying documents. This discrepancy highlights the considerable variation in the performance of different article types accurately identifying AI-generated text.

In order to gain a clearer understanding of how the true positive rates by article type are influenced by what classifier it corresponds to, we will present boxplots that depict the distribution of true positive rates for each combination of classifier and article type.



First and foremost, these boxplots provide further confirmation of the trends observed in the previous histogram. Specifically, art history articles exhibit the highest true positive rates, followed by fictional novel, and medical articles having the lowest rates. Within the art history category, the boxplots indicate that the average true positive rates are relatively consistent across both classifiers, suggesting that the choice between classifiers does not significantly impact the average true positive rate for these articles. Additionally, the boxplots shed light on the reason why classifier B possesses slightly higher true positive rates compared to classifier A. For both fictional novel and medical articles, classifier B demonstrates a higher mean true positive rate than classifier A. This disparity suggests that classifier B has a slightly better overall performance in correctly identifying AI-generated text for these particular article types. However, it is also worth noting that the boxes in the boxplots associated with classifier B are larger than those of classifier A across all article types. This larger box size implies that classifier B exhibits a greater variation in true positive rates compared to classifier A. So, while we do see a higher true positive rate for classifier B, on average, there is also potential for classifier B to have higher highs and lower lows in terms of the number documents it correctly classifies.

Lastly, in order to examine the relationship between true positive rates, classifiers, and the number of words in a document, we employ a side-by-side barplot to uncover any patterns or trends.



Upon comparing the true positive rates across different word counts, we again find that classifier B generally exhibits a slightly higher true positive rate compared to classifier A. Furthermore, there is a discernible trend indicating that as the number of words in a document increases, the true positive rate tends to follow a similar upward trajectory for both classifiers. In other words, a larger number of words in a document corresponds to a higher likelihood of accurate identification as AI-generated text.

Model Selection

Based on the preliminary findings of the data, we initially fit a logistic regression model to explore the relationship between the dependent variable and the independent variables. In this initial model, we included the independent variables *ArticleType* and *Words*, while excluding the predictor *Classifier* due to its insignificance (p-value much greater than 0.05).

To assess the adequacy of the model fit, we estimated the dispersion parameter, denoted as ϕ, which is a crucial consideration to ensure that the data does not exhibit overdispersion. The estimation from the initial logistic regression model yielded a dispersion parameter value of 1.4.

Unfortunately, the dispersion parameter value greater than 1 indicates the presence of overdispersion in our data. This poses a challenge as the overdispersion issue tends to underestimate standard errors, resulting in narrower confidence intervals. Consequently, this can lead to potential incorrect rejections of null hypotheses or acceptance of alternative hypotheses that may be false (Saefuddin and Setiabudi 2011).

The underlying cause of the overdispersion stems from the fact that the variance of true positives exceeds the mean. To properly account for the disparity between the variance and the mean of true positives, a more sophisticated model such as beta-binomial logistic regression may be necessary.

Beta-Binomial Logistic Regression Model

To address the issue of overdispersion, we employ a more sophisticated approach by constructing a Beta-Binomial logistic regression model. This model, compared to the simple logistic regression, accounts for the complexities inherent in the data. After fitting the Beta-Binomial model with the same primary predictors, we conduct a test to evaluate if overdispersion is present to such an extent that the Beta-Binomial model outperforms the simple binomial model.

One commonly used method to assess this is the likelihood ratio test (LRT). The null hypothesis assumes a Binomial distribution as the underlying distribution, while the alternative hypothesis suggests a Beta-Binomial distribution. The likelihood ratio test statistic, denoted as χ_1^2, is calculated as follows:

$$\chi_1^2 = -2(L_B - L_{BB})$$

Here, L_B represents the log likelihood value of the Binomial distribution at the maximum likelihood estimate (MLE), and L_{BB} represents the log-likelihood value of the Beta-Binomial model at the MLE. The resulting test statistic follows a χ^2 distribution with 1 degree of freedom, which represents the difference in the number of parameters between the two distributions (Kapourani 2018).

Upon performing the likelihood ratio test, we find that the p-value is significantly greater than 0.05. Consequently, we fail to reject the null hypothesis, suggesting that the data does not exhibit sufficient overdispersion to warrant the use of the Beta-Binomial model as a superior fit compared to the simple logistic regression model.

Hence, we determine the **final** model to be the logistic regression model discussed earlier. It is important to acknowledge that by selecting the logistic model, we sacrifice accuracy in favor of interpretability. The logistic regression model is less complex than the Beta-Binomial model, but its estimates still remain unbiased and accurate. Therefore, we can better comprehend the association between the true positive rate and the choice of AI-detector, while considering the influence of article type and the number of words in the document. Nevertheless, it is essential to exercise caution when interpreting the confidence intervals in this model, as they may exhibit a slightly lower precision.

Note: Code snippets illustrating the various steps undertaken during the model selection process will be provided in the Appendix at the end of this report.

Logistic Regression Model

The logistic regression model that represents this data will have the baseline category for *ArticleType* be `ArtHistory` and the baseline category for *Words* be `Words0.1`. Therefore, the intercept (β_0) in the equation represents 18 art history documents with 100 words each. A brief theoretical review of logistic regression and the model equation is as follows (Saefuddin and Setiabudi 2011):

Suppose there are k binomial observations written in the form of the proportion y_i/n_i, where y_i is the number of occurrences of the event and n_i is the total population number for each observation $i = 1, 2, \ldots, k$. In our case, y_i would be the number of correctly classified documents (true-positives) and n_i would be the total number of AI-generated documents studied for that observation (18 for each). Then, consider π_i, the probability of the event occurring. We can then express the logistic regression model for π_i with p predictors $(X_1, X_2, \ldots, X_p)$ as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$$

and for our data, as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \cdot \mathbb{I}(\textit{ArticleType} = \text{"FictionalNovel"}) + \beta_2 \cdot \mathbb{I}(\textit{ArticleType} = \text{"Medical"})$$
$$+\beta_3 \cdot \mathbb{I}(\textit{Words} = \text{"Words0.5"}) + \beta_4 \cdot \mathbb{I}(\textit{Words} = \text{"Words1"}) + \beta_5 \cdot \mathbb{I}(\textit{Words} = \text{"Words5"})$$
$$+\beta_6 \cdot \mathbb{I}(\textit{Words} = \text{"Words10"}) + \beta_7 \cdot \mathbb{I}(\textit{Words} = \text{"Words50"}) + \beta_8 \cdot \mathbb{I}(\textit{Words} = \text{"Words100"})$$

Note: Only one *Words* indicator and one *ArticleType* indicator will be active for each observation. This means that for each observation studied, the selection of the specific indicator variable is determined from the actual category of the *Words* variable or the *ArticleType* variable. Specifically, for every observation (18 documents at a time), only the relevant indicator variable will have a value of 1, while all other indicator variables within the same group will be assigned a value of 0. This ensures that the logistic regression model appropriately captures the effects of each specific category and avoids any overlap or confusion between different indicator variables. By employing this approach, the analysis accurately accounts for the unique contribution of each category within the *Words* variable and the *ArticleType* variable, allowing for a comprehensive understanding of their impact on the classification of AI-generated documents.

We now list the estimated coefficients of the model and their 95% confidence intervals in the table below:

Table 2: Logistic Regression Model: Estimated Coefficients, Associated 95% Confidence Intervals, and Respective p-values

Parameter	Estimate	95% Confidence Interval	p-value
β_0	0.436	(-0.061, 0.934)	0.086
β_1	-0.753	(-1.129, -0.376)	0
β_2	-4.034	(-4.695, -3.373)	0
β_3	0.274	(-0.376, 0.923)	0.409
β_4	0.055	(-0.593, 0.703)	0.869
β_5	0.219	(-0.43, 0.867)	0.509
β_6	0.274	(-0.376, 0.923)	0.409
β_7	1.294	(0.6, 1.987)	0.0003
β_8	1.226	(0.537, 1.915)	0.0005

Note: Despite the lack of statistical significance for certain levels of the *Words* variable, they are retained in the final model. This decision is based on our evaluation of each predictor's individual significance, which revealed that *Words* is statistically significant (with a p-value less than 0.05). Therefore, we include all levels of Words in the final model to ensure a comprehensive assessment of its impact on the outcome.

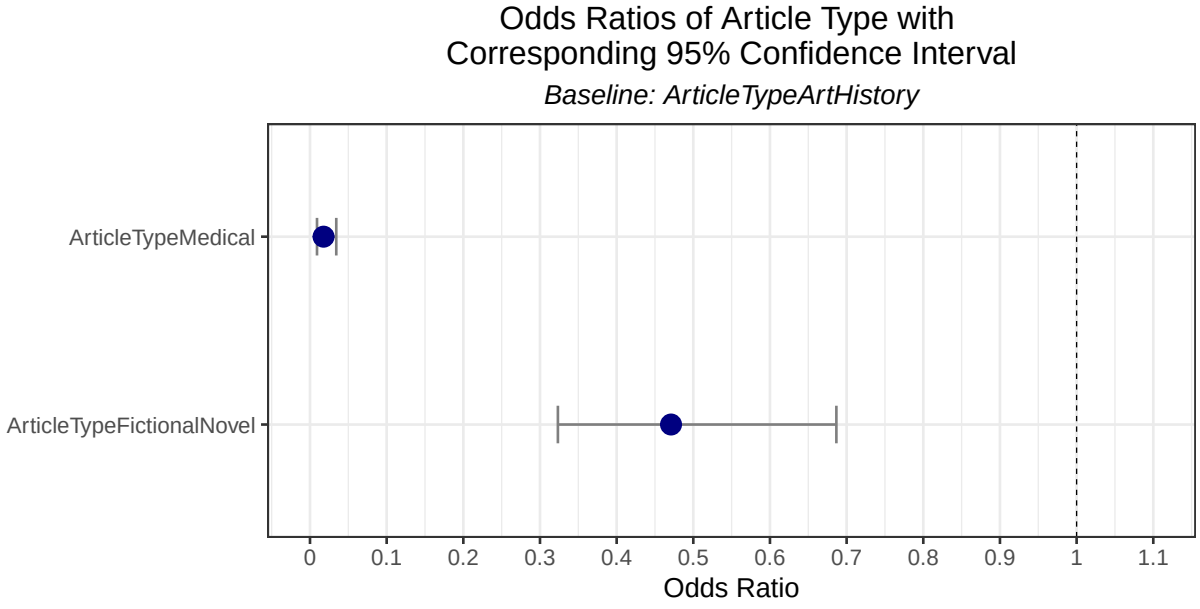
Results and Interpretations

Instead of trying to interpret the coefficient estimates on the logit scale (as listed in the table above), we can take the exponential of the coefficient estimates and their respective 95% confidence intervals to get the odds ratios.

Example: For the intercept, we can take the exponential of its coefficient estimate, and interpret: $\beta_0 = 0.436 \Rightarrow e^{0.436} = 1.547$.

Again, the intercept represents the odds of correctly classifying a document as AI-generated with the respective baseline categories. This means that when the document has a word count of 100 (baseline) and belongs to the `ArtHistory` category (baseline), the odds of the AI-detectors correctly classifying it as AI-generated are 1.547 times higher. In other words, the odds of successful classification increase by approximately 54.7% when the document falls under the baseline conditions. This implies that the AI-detectors are more likely to correctly identify AI-generated documents that have a word count of 100 and belong to the `ArtHistory` category, compared to other categories and word counts.

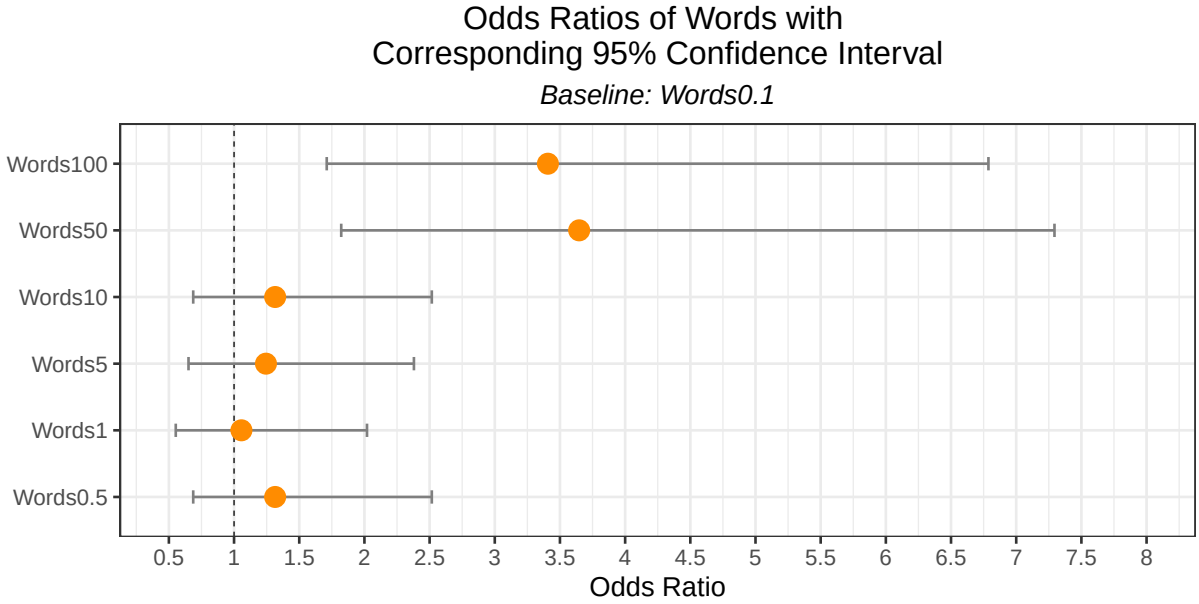
In order to provide a clearer perspective for the other variables, we will generate plots of the odds ratios along with their corresponding 95% confidence intervals. These plots will allow us to visualize the comparisons between different article types relative to their reference group (`ArtHistory`) and the comparisons between different numbers of words relative to their reference group (`Words0.1`). By examining these plots, we can derive meaningful interpretations of the odds ratios and their implications for the correct classification of AI-generated text by the AI-detectors.



Interpretations

- **ArticleTypeFictionalNovel**: The odds of correctly classifying a document as AI-generated for the `FictionalNovel` article type are approximately 0.47 times lower, on average. In other words, the likelihood of correctly classifying a document as AI-generated is reduced by approximately 53% when it belongs to the `FictionalNovel` category, compared to the `ArtHistory` category.

- **ArticleTypeMedical**: The odds of correctly classifying a document as AI-generated for the `Medical` article type are approximately 0.02 times lower, on average. In simpler terms, the likelihood of correctly classifying a document as AI-generated is significantly reduced by approximately 98% when it belongs to the `Medical` category, compared to the `ArtHistory` category.



Interpretations: Due to the large number of categories involved, we will focus our discussion on a select few. It is important to note that while we limit the discussion to these specific categories, the overall interpretation remains consistent. The only variation lies in the approximate values, rather than the fundamental meaning conveyed by the interpretation.

- **Words0.5**: The odds of correctly classifying a document as AI-generated for a document with 500 words (5000 divided by 1000 = 0.5) are approximately 1.3 times higher, on average. In simpler terms, the likelihood of correctly classifying a document with 500 words as AI-generated is approximately 30% higher compared to a document with 100 words (reference category).

⋮

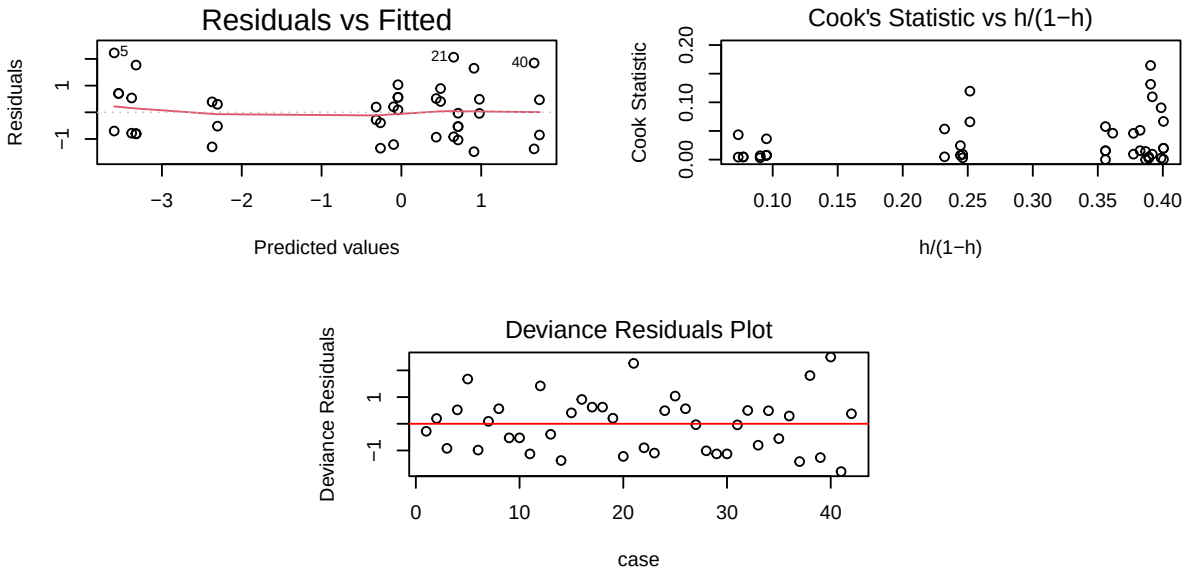
- **Words50**: The odds of correctly classifying a document as AI-generated for a document with 50000 words are approximately 3.6 times higher, on average. In simpler terms, the likelihood of correctly classifying a document with 50000 words as AI-generated is approximately 260% higher compared to a document with 100 words (reference category).
- **Words100**: The odds of correctly classifying a document as AI-generated for a document with 100000 words are approximately 3.4 times higher, on average. In simpler terms, the likelihood of correctly classifying a document with 100000 words as AI-generated is approximately 240% higher compared to a document with 100 words (reference category).

Based on the observed odds ratios, it is evident that the choice of article type plays a substantial role in the accurate classification of documents as AI-generated. The odds of correctly identifying a document in the `FictionalNovel` and `Medical` categories as AI-generated are significantly lower compared to the `ArtHistory` category.

Regarding the relationship between the number of words in a document and the odds of correctly classifying it as AI-generated, a clear trend emerges. As the number of words increases, the odds of correctly identifying a document as AI-generated tend to rise. Notably, the largest increase in odds is observed for documents containing 50,000 and 100,000 words.

Diagnostics

Here, we examine a few diagnostic plots to ensure our logistic regression model fits the data well (Newsom 2021).



For both residual plots (residuals vs fitted and deviance residuals), we find symmetry around zero, indicating that, on average, the model is accurately predicting the observed outcomes. We also find no obvious patterns or trends, i.e, the residuals are randomly scattered around zero. This suggests that there are no major issues with the fit of our model. Lastly, we also see that the residuals have consistent spread, indicating the data is dispersed relatively well.

In the plot of Cook's statistic, typically, two dotted lines are displayed: one vertical and one horizontal. Points above the horizontal line or to the right of the vertical line are indicative of potentially high influential points in the model. Conversely, if all points are below the horizontal line or to the left of the vertical line, the lines are not shown. Upon observing the plot in this analysis, we notice the absence of any lines, indicating the absence of influential points in our model.

Conclusions

In conclusion, our analysis reveals that there is no significant difference in the true positive rates between the two AI-detectors ("A" and "B") developed for detecting AI-generated text. However, we identified a notable dependence on the type of article and the number of words in the document. Specifically, we found that regardless of the AI-detector used, the highest success rates in correctly classifying documents were consistently observed when the article type was art history. Furthermore, documents with a larger number of words (50000 or 100000) also exhibited higher rates of accurate classification across AI-detectors.

These findings highlight the importance of considering both article type and document length when aiming to optimize the performance of AI-detectors for detecting AI-generated text. Art history articles consistently yielded more favorable results, suggesting a potential inherent characteristic that enhances detection accuracy. Additionally, the influence of document length indicates that longer texts may contain more discernible patterns or features that aid in distinguishing AI-generated content.

If we are to extend this research out further, there is potential to explore additional factors that could influence the performance of AI-detectors. By considering variables beyond article type and document length, we can gain a more comprehensive understanding of the factors that affect their efficacy. Furthermore, conducting investigations across various AI-detectors and incorporating new observations (still consisting of 18 documents each) would contribute to the applicability of our findings. This broader analysis would provide valuable insights into the wider applicability of our conclusions and deepen our understanding of the factors that contribute to the effectiveness of AI-detectors in detecting AI-generated text.

In closing, our study emphasizes the significance of considering contextual factors, such as article type and document length, in optimizing the performance of two AI-detectors for identifying AI-generated text. By leveraging these insights, researchers can advance the development of more accurate and robust AI detection systems in other various domains.

Appendix

Here, we display the relevant code snippets from the necessary sections of the report.

Introduction

```
# load in data set
AI_data <- read.csv("C:/Users/elija/OneDrive/UCSB Work/Statistics MA Work/PSTAT 230/MA Qual/AI_data.csv
AI_data[,3] <- factor(AI_data[,3])
AI_data[,4] <- factor(AI_data[,4])
AI_data[,5] <- factor(AI_data[,5])
```

Model Selection

Logistic Regression

```
#logistic regression
mod.l <- glm(cbind(nPos, mDoc-nPos) ~ Classifier + ArticleType + Words,
             data = AI_data, family = "binomial")
summary(mod.l)
step(mod.l)
```

```
#check variables individual significance to the model
Anova(mod.l, test.statistic = "LR", type = "III")
```

```
#compare models to see which one is better
mod.l2 <- update(mod.l, . ~ . - Classifier)
anova(mod.l2, mod.l, test = "Chisq")
```

```
#dispersion parameter
phi.hat <- summary(mod.l2)$deviance/summary(mod.l2)$df[2]
phi.hat
```

Beta-Binomial Logistic Regression

```
#beta-binomial model
mod.b <- betabin(cbind(nPos, mDoc-nPos) ~  Classifier + ArticleType + Words,
                 random = ~ 1, data = AI_data)

summary(mod.b)
```

```
#compare models to see which one is better
mod.b2 <- betabin(cbind(nPos, mDoc-nPos) ~  ArticleType + Words,
                  random = ~ 1, data = AI_data)
anova(mod.b2, mod.b)
```

```
#check model assumption - likelihood ratio test
#determined beta-binomial was not necessary
pchisq(2 * (logLik(mod.b2) - logLik(mod.l2)), df = 1, lower.tail = FALSE)
```

References

Kapourani, C. A. 2018. "Beta Binomial for Overdispersion." 2018. https://rpubs.com/cakapourani/Beta-Binomial.

Newsom. 2021. "Diagnostics for Logistic Regression." 2021. https://web.pdx.edu/~newsomj/cdaclass/ho_diagnostics.pdf.

Saefuddin, Asep, and Nur Andi Setiabudi. 2011. "The Effect of Overdispersion on Logistic Regression Analysis of Poverty in Indonesia." *International Journal for Statisticians*.