technische universität
dortmund

Statistics meets Linguistics

Linguistic differences between AI-generated reviews and reviews written by German native speakers with English as their second language

Author: Aakash Goyal (229975)

Supervisors: Prof. Dr. Claus Weihs

Prof. Dr. Sarah Buschfeld

March 10, 2024

Contents

1. Introduction

In this digital era, e-commerce (or electronic commerce) has firmly established itself as a ubiquitous and popular choice for consumers buying goods and services. To decide on the right product, consumers are highly dependent on online reviews of the products or services. Online reviews play a pivotal role in guiding consumer decisions and act as a barometer for quality and trustworthiness. They also help the consumers gauge the worthiness of their investment in the product or service. However, with the advent of artificial intelligence (AI), relying on the authenticity of online reviews has become more challenging. It becomes a complex task to differentiate between a genuine customer review and an artificially generated review. These AI-generated reviews pose a significant threat as they mislead consumers and skew their perceptions of the products or services.

This report delves into the nuances of linguistics, which help in distinguishing between the AI-generated reviews and the reviews written by German native speakers with English as their second language. For this study, two products are initially selected, namely, the Apple Smartphone and Netflix, on which data is collected. A survey is designed and disseminated to German native speakers to gather their reviews about the selected products. A total of 21 people responded to the survey, and out of them, 18 provided valid responses for the study. In addition to this, 20 reviews for each product are generated using the chatbot technology ChatGPT (Chat Generative Pre-trained Transformer), developed by OpenAI, Inc. Secondly, a dataset is prepared by extracting different linguistic markers for each review submitted by respondents in the survey and generated by AI. Thirdly, using statistical or graphical methods, the demographic information of respondents and the prepared dataset of reviews with linguistic markers are scrutinized. Further, a conditional tree with all the linguistic markers is fitted, and the assumptions of the conditional tree are validated. Lastly, the evaluation metrics are recorded, and the results are interpreted.

In Section 2, a detailed overview of the data collection and data description is presented. In Section 3, statistical and graphical methods, including their properties and assumptions, are explained, which are used for the analysis. In Section 4, the introduced statistical and graphical methods are applied to the prepared dataset, and the results of the tasks are interpreted. In Section 5, all the findings of the analysis are summarized. Finally, in Section 6, the possibility of further research is addressed.

2. Data

In this section, the specifics of data collection method is presented and description of dataset is provided.

2.1. Data Collection and Preparation

The data is collected through a survey specifically designed and distributed randomly to German native speakers. A series of questions is created to gather detailed insights for this study. These questions cover various demographic aspects of the respondents, such as age, gender, highest academic level, nationality, second language, the age at which they started learning English, and their reviews of the selected products, such as Netflix and the Apple smartphone. A total of 21 responses to the survey are received. However, 3 responses are excluded from the study as they are not German nationals. This exclusion is carried out to maintain homogeneity in the data.

To collect data for AI-generated reviews, the chatbot technology ChatGPT (Chat Generative Pre-trained Transformer), developed by OpenAI, Inc., is employed. A prompt is inputted into the AI model, which then generates 20 reviews for each selected product. Further, the dataset is complete, and there are no missing values. From these reviews, a dedicated dataset is prepared where linguistic markers such as the number of personal pronouns, adjectives, compound sentences, word count, and spelling mistakes are extracted for each review. This extraction of linguistic markers is important for the subsequent analysis of linguistic nuances and patterns between the reviews written by German native speakers and those generated by AI.

2.2. Description of dataset

The dataset is prepared with different linguistic markers extracted for each review. It comprises 40 AI-generated reviews and 36 reviews written by German natives, making the total size of the dataset 76. It includes 6 variables: $compound\_sentences$ (number of compound sentences), $personal\_pronouns$ (number of personal pronouns), $adjectives$ (number of adjectives), $word\_count$ (count of words), $spelling\_mistakes$ (number of spelling mistakes), and $review\_type$ (indicating whether the review is written by a German native speaker or AI). The variables $compound\_sentences$, $personal\_pronouns$, $adjectives$, $word\_count$, and $spelling\_mistakes$ are of a discrete metrical type. The variable $review\_type$ is of a

discrete nominal type and serves as the response variable. It has two categories: Human and AI. The 'Human' category represents reviews written by German native speakers with English as their second language, while the 'AI' category represents reviews generated by chatbot technology. The dataset is of decent quality and contains no missing values.

2.3. Report Objectives

The main objective of the report is to analyze and classify whether a review is AI-generated or written by a German native speaker. More specifically, this study is targeted to determine whether the number of $adjectives$ (adjectives), $personal\_pronouns$ (personal pronouns), or $compound\_sentences$ (compound sentences) are reliable factors in ascertaining the authenticity of a review. To achieve this, firstly, a survey is conducted to collect review data for each product from German native speakers with English as their second language, and ChatGPT is prompted to generate 20 reviews for each product. Secondly, a dataset is prepared by extracting several linguistic markers such as $compound\_sentences$, $personal\_pronouns$, $adjectives$, $word\_count$, and $spelling\_mistakes$ for each review. Thirdly, bar plots are used to analyze demographic information such as age, gender, and the age at which respondents started learning English. Further, box plots are created to analyze the variability of linguistic markers within or between the types of reviews. Lastly, a conditional tree with all the covariates is fitted, and evaluation metrics, such as accuracy and balanced accuracy, are recorded, and the results are interpreted.

3. Statistical methods

In this section, several statistical methods are presented, which are later used for analyzing the data according to the project requirements.

3.1. Box Plot

The box plot is a graphical representation of the dataset. It uses a five-number summary to represent the dataset. There are five specific values in the five-number summary: minimum value, lower quartile, median, upper quartile, and maximum value.

In Figure 1, a box plot constructed with random data is shown. Quartiles divide the data into four equal subgroups. There are three quartiles, such as the lower quartile (Q_1), second

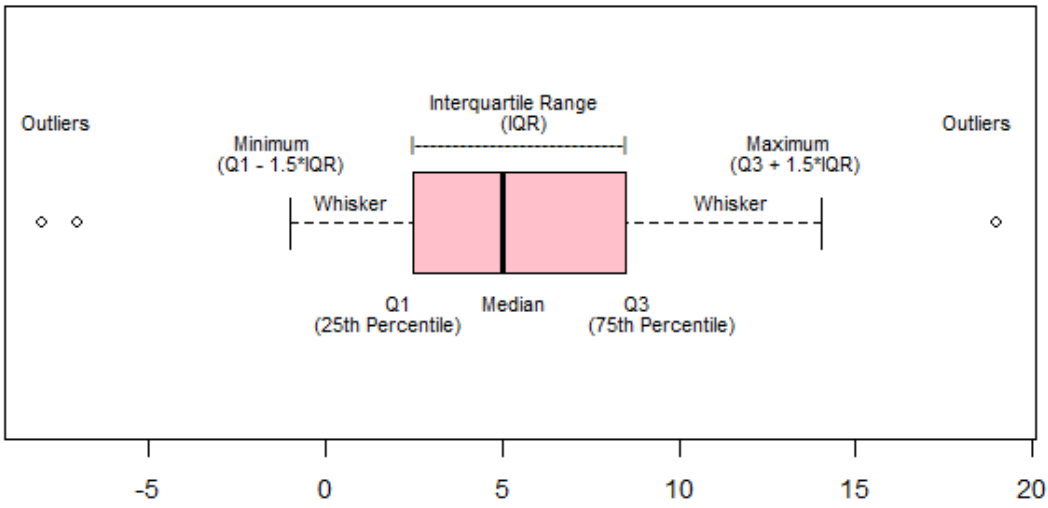


Figure 1: Boxplot

quartile (Q_2), and the upper quartile (Q_3). The left side of the box represents the lower quartile (Q_1). It separates the first one-fourth of the data. Similarly, the right side of the box represents the upper quartile (Q_3). This separates the third quarter of the data. The vertical line in the box represents the median, or the second quartile (Q_2). The length of the box, known as the inter-quartile range (IQR), is the difference between the upper quartile (Q_3) and lower quartile (Q_1) and is represented as $IQR = Q_3 - Q_1$. The inter-quartile range is the range of the middle 50% of the data. The lines that emanate from the box are called whiskers. The right whisker stretches out to the largest observation in the data, or $Q_3 + 1.5 * IQR$, whichever comes early. Similarly, the left whisker stretches out to the lowest observation in the data, or $Q_1 - 1.5 * IQR$, whichever comes early. The observations that fall beyond the left and right whiskers are known as outliers (Hay-Jahans, 2019, p. 137-139) and (Dodge, 2008, p. 55,56).

3.2. Bar Chart

A bar chart is a useful tool for representing the frequencies of data. Each bar, having equal width, represents the frequency of a data value by its height. There are two types of bar graphs: vertical and horizontal. In this report, vertical bar graphs are used. A bar graph consists of two axes: horizontal and vertical. In a vertical bar graph, data values are plotted on the horizontal axis, and their frequencies are depicted on the vertical axis. The bar graph is also used for comparing two or more data values. For comparison, bars that represent different data

values are placed side by side, and their heights, which indicate frequencies, can be compared (Hay-Jahans, 2019, pp. 111 - 113) and (Bluman, 2018, pp. 75 - 77).

3.3. Hypothesis Testing

The inferential approach allows to infer conclusions from a sample to the whole population. One of the common statistics used for such an approach is hypothesis testing. A hypothesis is the proposed solution to the problem. There are two hypotheses in hypothesis testing. One is the null hypothesis, which is denoted by H_0 and the other is the alternative hypothesis, which is represented by H_1. The null hypothesis is a null statement, which means that nothing new is happening and the old story is still true. While the alternative hypothesis means a new theory is true and/or something new is happening. Both hypotheses are mutually exclusive, which means the null hypothesis (H_0) is the complement of the alternative hypothesis (H_1). The null hypothesis is used for testing purposes. We either reject or fail to reject the null hypothesis. The null hypothesis is not rejected unless there is sufficient evidence that it is false.

While testing the hypothesis, it is possible to make errors about the null hypothesis. There are two types of errors, i.e., Type I error and Type II error. Type I error occurs when the true null hypothesis is rejected, also known as false positive. On the other hand, a Type II error occurs when the false null hypothesis is not rejected, also known as a false negative. Type I error represents the level of significance, and it is denoted by α. Type II error is denoted by β (Black, 2019, pp. 272-273).

	H_0 true	H_0 false
Reject H_0	Type I error(α)	Correct decision
Fail to reject H_0	Correct decision	Type II error(β)

Table 1: Type I and Type II errors

3.4. P-value and Level of significance

The p-value determines the strength of the evidence for the null hypothesis (H_0) to be rejected. The significance level is the threshold value at which the null hypothesis is tested. This significance level is determined before beginning an experiment. The null hypothesis (H_0) can be rejected if the significance level (α) is greater than the p-value. Conversely, the null hypothesis (H_0) is failed to be rejected if the significance level (α) is less than the p-value.

The significance level value (α) is well-defined for an experiment, and then the significance level value (α) and the p-value are compared to decide whether to reject or not reject the null hypothesis. For example, if the p-value is 0.03, then the null hypothesis (H_0) can be rejected for the significance level (α) of 0.05, while it cannot be rejected if the significance level (α) is 0.01 (R. Lyman Ott, 2015, pp. 257).

3.5. Conditional Inference Tree

A conditional inference tree is a type of decision tree that is based on recursive partitioning. Recursive partitioning is a statistical technique that divides the data into subsets that are homogeneous with respect to the response variable. In this technique, it is assumed that the conditional distribution of the response variable Y, given the predictors X, is determined by the function f that maps the predictors. It is represented by

$$D(Y|X) = D(Y|f(X_1, \ldots, X_m))$$

where X is an m-dimensional predictor vector $X = (X_1, \ldots, X_m)$ which is extracted from the sample space $\chi = \chi_1 \times \ldots \times \chi_m$ and Y is the response variable taken from sample space γ. The recursive binary partitioning algorithm for the random sample can be stated as :

1) With the help of statistical criteria, identify the covariate that distinguishes between various outcomes of the target variable most effectively.

2) Divide the dataset into multiple subsets based on the split on the selected variable.

3) Continue Steps 1 and 2 recursively until splitting is no longer possible, adhering to the predetermined criteria.

In addition to the above, a test statistic is calculated, which measures the change in the relationship between the response variable Y and the predictors X before and after the permutation. The higher value of the statistics represents a stronger association between the response variable and the predictor. The best split variable is selected by comparing the statistics from different predictors. The predictor with the highest statistics is selected for the next split. A p-value is used for such a comparison, as it remains unaffected by the measurement scales of the predictors.

Once the predictor is selected, the algorithm aims at dividing the chosen predictor into two separate groups. For predictors that offer several splits, this technique evaluates statistics for all the splits in the groups G or not G. The split with the highest statistics is selected as the

next best split. This process keeps on being repeated recursively until certain conditions are satisfied, which include:

1) a minimum criterion for performing the split is $1 - \alpha$, where α represents the level of significance. The algorithm stops splitting if it is not able to reject the global null hypothesis at the pre-specified α, i.e., 0.05. This parameter, α, serves a dual purpose by acting as a statistical benchmark and a hyperparameter which influences the size of the tree.

2) the minimum number of instances should be present in the node before splitting. If the node contains fewer instances than the threshold, no split occurs.

3) the minimum number of instances should be in a node following a split.

To fit a conditional inference tree, there are no conventional assumptions. But, it is important to pay attention to the independence of observations (Magali Paquot, 2022, pp. 611-616) and (Hothorn et al., 2015).

3.6. Evaluation Metrics

The capacity or ability of the model to classify two levels or categories of the response variable is assessed by the evaluation metrics of the model. Confusion matrix, accuracy, and balanced accuracy are some of the classification tools that are used in this report.

3.6.1. Confusion Matrix

It is an $N \times N$ matrix of correctly and wrongly predicted fitted values, where N is the number of classes being predicted. In this report, there are two classes in the target (AI or Human), so a 2×2 confusion matrix is obtained. Confusion matrix for binary classification is given as

		Actual		
		Positive	Negative	Total
Prediction	Positive	TP	FP	$TP + FP$
	Negative	FN	TN	$FN + TN$
	Total	$TP + FN$	$FP + TN$	

Table 2: Confusion Matrix

Elements of confusion matrix:

True Positive (TP): Values that are predicted positive by the model and were expected to be positive.

True Negative (TN): Values that are predicted negative by the model and were expected to be negative.

False Positive (FP): Values that are predicted positive by the model and were expected to be negative. This is also known as a Type-I error.

False Negative (FN): Values that are predicted negative by the model and were expected to be positive. This is also known as a Type-II error.

Accuracy: It is a metric that measures the proportion of correct predictions made by the model out of the total number of predictions. It is calculated by dividing the number of correct predictions ($TP + TN$) by the total number of predictions ($TP + TN + FP + FN$). It can be observed as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Sensitivity: Out of all actual positive outcomes, it measures how effectively a model predicts positive outcomes. It is derived by dividing the number of true positives (TP) by the sum of true positives and false negatives ($TP + FN$).

Specificity: Out of all actual negative outcomes, it measures how effectively a model predicts negative outcomes. It is calculated by dividing the number of true negatives (TN) by the sum of true negatives and false positives ($TN + FP$) (Radek Silhavy, 2023, pp. 31, 44-45) and (Hilbe, 2015, pp. 86-88).

Balanced Accuracy: It is a method to measure the performance of the model for the unbalanced datasets. It is calculated as the average of the sensitivity and specificity of the model. It can be observed as

$$Balanced\ Accuracy = \frac{Specificity + Sensitivity}{2} \tag{2}$$

It considers the ability of the model to correctly identify positive outcomes as well as negative outcomes, especially when the dataset is imbalanced (Zhang et al., 2019).

4. Statistical analysis

In this section, the above-mentioned statistical methods are applied to the dataset for analysis, and the results of the tasks are interpreted. For all calculations and visualizations, the software R (R Core Team, 2022) and R packages rpart (Therneau and Atkinson, 2023), caret (Kuhn and Max, 2008), and party (Hothorn et al., 2006) are used. The value of level of significance is (α) is set at 0.05.

4.1. Descriptive Analysis

In this subsection, the focus lies on understanding the variables present in the dataset. The dataset consists of 5 variables, i.e., $compound\_sentences$, $personal\_pronouns$, $adjectives$, $word\_count$, $spelling\_mistakes$ and the response variable, i.e., $review\_type$.

As mentioned in Section 2, a survey is conducted to collect the dataset. Firstly, the demographic data of the respondents is explored. In Figure A1 in the Appendix, the bar plot of the genders of the respondents is depicted. The gender is divided into 3 categories, i.e., Diverse, Female, and Male. It is evident from the plot that most of the respondents to the survey are female.

In Figure 2, a bar plot for the review type depicts the distribution of the response variable. It is categorized into two types: 'AI' and 'Human'. The height of the bars indicates the frequency of each review type. The bar for review type 'AI' is taller than that of 'Human', depicting that the target variable $review\_type$ contains a higher count of AI reviews. The difference in the frequencies of the categories is quite small, and hence, it is inferred that the dataset is quite balanced.

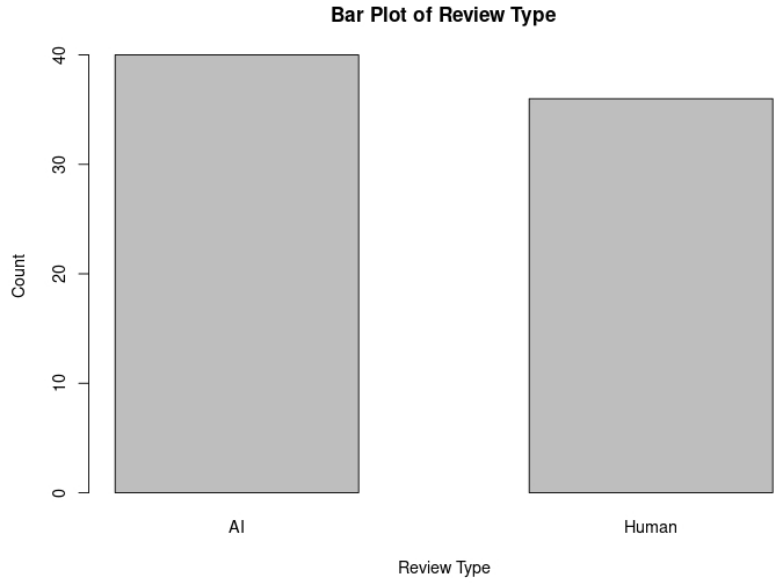
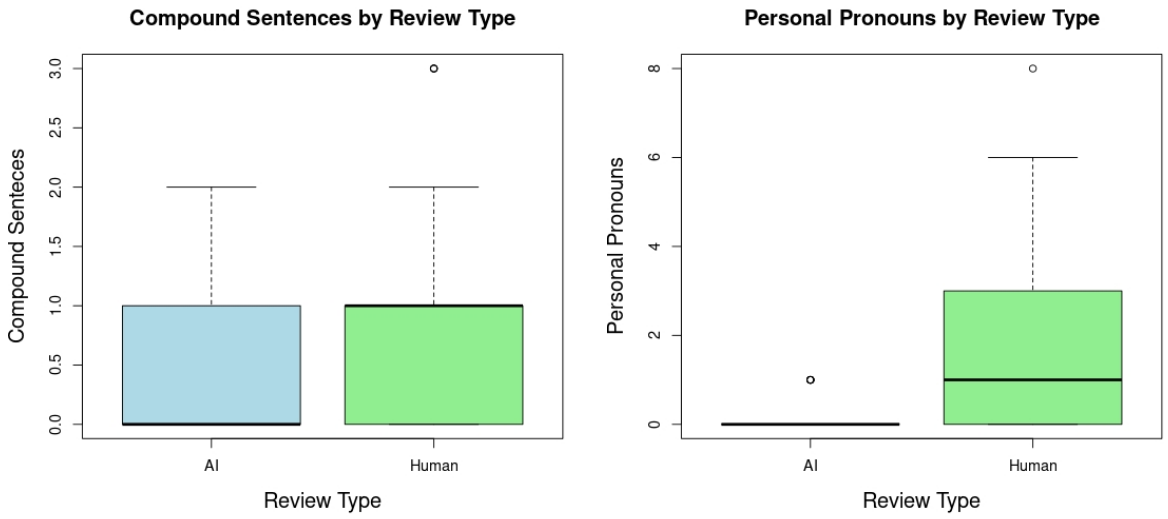


Figure 2: Box plot for word count

In Figure 3(a), the box plots for the number of compound sentences for each review type are depicted. The median number of compound sentences for the review type 'AI' is 0, while for the review type 'Human', it is 1. The boxes represent the interquartile ranges (IQRs) and for both the review types, they are comparatively similar in size. One of the respondents to the

survey used 3 compound sentences, represented by a circle, in the review, and it is an outlier here.



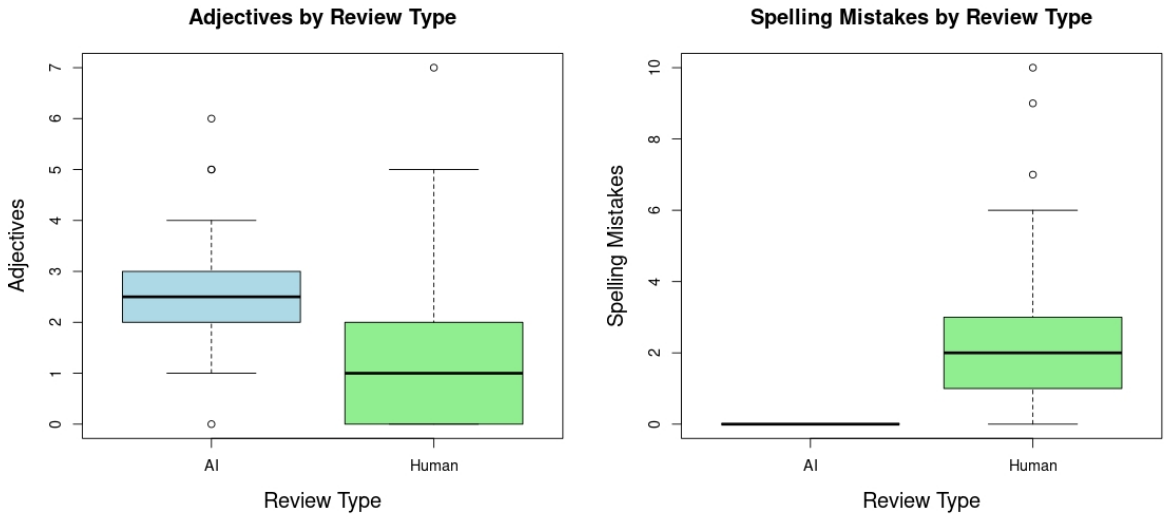
(a) (b)

Figure 3: Box plot for (a) Compound Sentences (b) Personal Pronouns

Figure 3(b) represents the box plots of the number of personal pronouns for each review type. From the plot, it is observed that the chatbot technology ChatGPT hardly uses personal pronouns, while German native speakers often use personal pronouns in their review writing.

In Figure 4(a), the box plot for the number of adjectives used in each review type is represented. The median for the AI review type is higher than the Human review type, indicating that AI uses more adjectives in reviews than Humans do. Moreover, there are outliers for both review types that are represented by the circles. In the case of AI-generated reviews, outliers are present above and below both the whiskers, which tells that there are reviews that contain a significantly higher or lower number of adjectives compared to the usual review. On the other hand, in the case of human-written reviews, an outlier is present above the upper whisker and suggests that there are human-written reviews that contain a significantly higher number of adjectives than the rest of the data.

In Figure 4(b), the box plot for the number of spelling mistakes in reviews written by AI and German native speakers (i.e. Human) is depicted. The inter-quartile range for the spelling mistakes done in the reviews written by human is relatively taller than that of AI. Moreover, the box for AI reviews is just a line which indicates that there is no spelling mistakes. The median value for reviews written by human is above zero, which indicates that human written contains more spelling mistakes than AI-generated ones.



(a) (b)

Figure 4: Box plot for (a) Adjectives (b) Spelling Mistakes

In Figure A2 in the Appendix, the box plot for the number of words in each review type is depicted. The IQR for the AI-generated review is quite narrower than that of human written reviews. It can be inferred that reviews generated by AI are more consistent in length than the reviews written by humans. Moreover, the comparison of the median values of word count indicates that AI tend to write longer reviews than humans at usual.

4.2. Conditional Tree

In this subsection, the conditional tree, explained in Section 3.5, is used to model the dataset that is collected and prepared in Section 2.1. The survey is distributed unconditionally to the participants. Moreover, there are no repetitive values in the dataset. Therefore, there is no relationship between the observations, and they are independent of each other. All the covariates, such as $compound\_sentences$, $personal\_pronouns$, $adjectives$, $word\_count$, and $spelling\_mistakes$ are considered at the time of training our conditional tree model. In Figure 5, a conditional tree fitted to all the covariates is depicted. Node 1 is the starting point of the tree, and it depicts that the first split in the data is based on $spelling\_mistakes$. The p-value (p-value<0.001) of the node 1 indicates that the number of spelling mistakes is the most statistically significant predictor. If there are spelling mistakes, the algorithm moves to Node 5, otherwise, it moves to Node 2. Following the left branch where spelling mistakes are smaller or equal to zero, the tree is further divided based on the usage of $personal\_pronouns$.

Similarly, the p-value (p-value = 0.002) indicates that this is a significant predictor. If personal pronouns are absent, the algorithm moves to Node 3, otherwise it moves to Node 4. In the end, three terminal nodes remain, as the tree does not find any other significant split at the set significance level of 0.05. These terminal nodes represent the model's final predictions based on the different decision rules.

Node 3 (n=37): This node represents the subgroup with no spelling mistakes and no personal pronouns, primarily containing AI-generated reviews.

Node 4 (n=8): This node represents the subgroup with no spelling mistakes and the presence of personal pronouns in the reviews. It is observed that the model prediction's are mixed but still AI- generated reviews is in majority.

Node 5 (n=31): This node represents the subgroup with spelling mistakes, predominantly contains human written reviews.

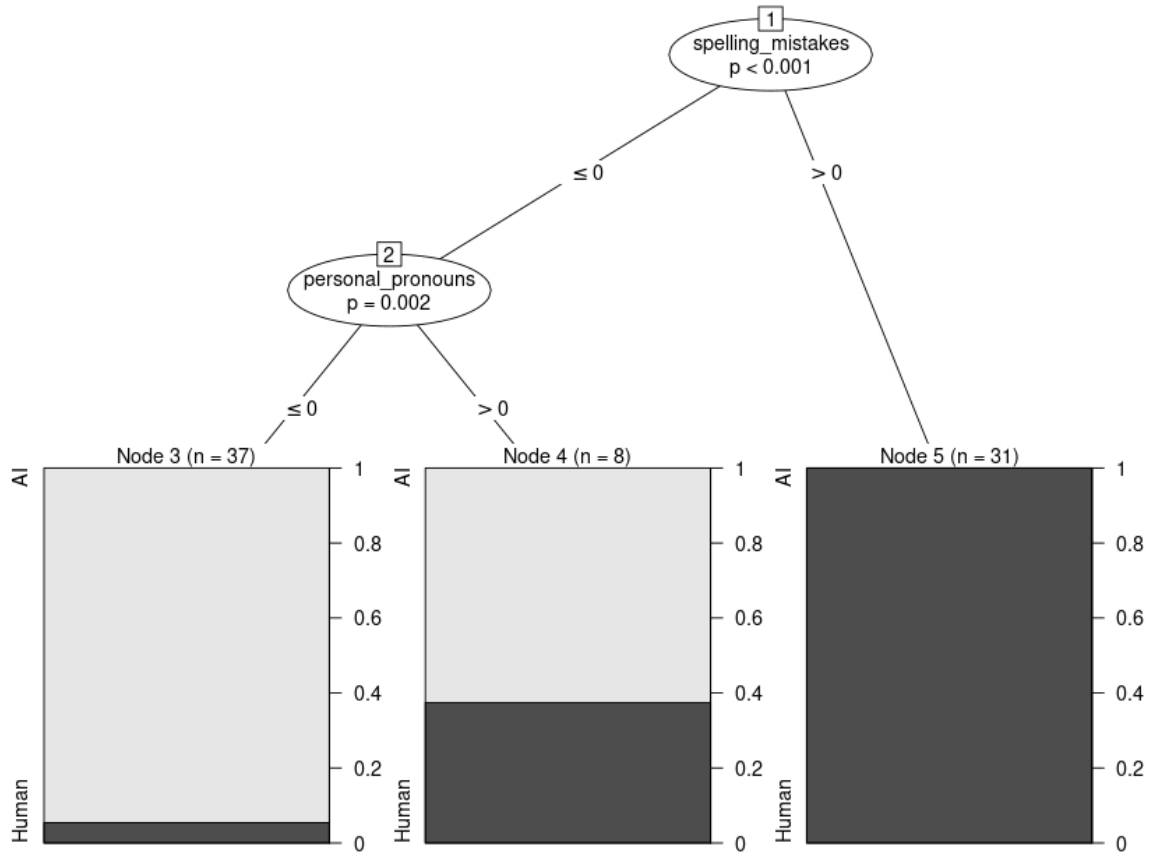


Figure 5: Conditional tree model

4.3. Evaluation Metrics

In this subsection, the performance of the model is evaluated using a confusion matrix and subsequent statistical metrics like accuracy and balanced accuracy. Table 3 represents the confusion matrix, and it indicates that out of 76 instances, the model predicts 40 instances of AI-generated reviews and 31 instances of human-written reviews correctly.

		Actual	
		AI	Human
Prediction	AI	40 (TP)	5 (FP)
	Human	0 (FN)	31 (TN)

Table 3: Confusion Matrix of the model

Moreover, 5 instances of human-written reviews are misclassified as AI-generated reviews by the model, but no instance of AI-generated reviews is misclassified. The model has achieved an accuracy of 93.42%, which indicates the high overall correctness of the model in predicting the type of reviews. Furthermore, the model achieves a sensitivity score of 100%, which reflects the perfect capability of the model to identify AI-generated reviews. The specificity of the model is 86.11%, which reflects the ability of the model to recognize human-written reviews. Although the dataset in Section 2.1 is quite balanced, with 40 instances of AI-generated reviews and 36 instances of German native speaker written reviews, the balanced accuracy of the model is still computed to get a detailed view of its performance. The model achieves a high balanced accuracy of 93.06%, which demonstrates its reliability in differentiating between both categories.

5. Summary

The main aim of the report was to analyze the nuances of linguistics, which would help in differentiating between the reviews generated by AI and the reviews written by German native speakers with English as their second language. The data was collected by conducting a survey based on consumer reviews of two products, i.e., the Apple smartphone and Netflix. The survey was sent to German native speakers with English as their second language to maintain homogeneity in the dataset. A total of 21 people responded to the survey, and out of them, only 18 provided valid responses. Further, the chatbot technology ChatGPT was used to generate 20 reviews for each product. After collecting the data, a dataset was prepared by extracting different linguistics markers from each review. These linguistics markers were $compound\_sentences$,

$personal\_pronouns$, $adjectives$, $word\_count$, and $spelling\_mistakes$. A target column $review_type$ was also added to the dataset, which marks whether the review is written by a German native speaker or generated by AI. After data preparation and processing, a descriptive analysis of the dataset was performed. The bar plot for the genders of the respondents, in Figure A1 in the Appendix, depicted that the most of the respondents were female. In Figure 2, the distribution of the response variable was observed and the it was concluded that the dataset was balanced. For the box plot of compound sentences in Figure 3(a), the median value was higher for the human category than the AI. Hence, it was inferred that German native speakers tend to use more compound sentences than AI. Similarly, for the box plot of personal pronouns in Figure 3(b), it was observed that humans tend to use more personal pronouns while writing a review than AI. Further, in Figure 4(a), it was inferred that AI uses more adjectives in its review writing than a German native speaker. In Figure 4(b), it was observed that AI makes no spelling mistakes while writing a review, but on the other hand, German native speakers tend to make spelling mistakes. In the end, in Figure A2 in the Appendix, it was concluded that AI writes longer reviews than German native speakers. Later, in order to model the dataset, a conditional tree model was fitted, and the results were evaluated. A conditional tree was fitted with all the covariates such as $compound\_sentences$, $personal\_pronouns$, $adjectives$, $word\_count$, and $spelling\_mistakes$. Out of all the covariates, $spelling\_mistakes$ and $personal\_pronouns$ were the influential variables at the significance level of 0.05. The model achieved an accuracy of 93.42% and the balanced accuracy of 93.06% which shows that the model can differentiate between the two categories effectively. In conclusion, the analysis indicates that variables $compound\_sentences$ and $adjectives$ are not reliable factors, as the conditional tree does not find them significant at the set significance level. Conversely, the covariate $personal\_pronouns$ stands out as a reliable determinant in ascertaining the authenticity of a user review.

6. Future Work

For further examination, several approaches can be explored to enhance the results of this study. Firstly, more data points could be collected, as there are only 76 data points in our study. It is important to expand the dataset to explore the subtleties within it. Secondly, the dataset exhibits a gender imbalance, as most of the respondents to the survey were females. So, equalizing gender representation in the dataset can help in examining the influence of gender on language use while writing product reviews. Finally, more linguistic markers can be

extracted from the data in order to further understand the nuances of the language, as this will help in differentiating AI-generated reviews from reviews written by German native speakers.

Bibliography

Ken Black. *Business Statistics: For Contemporary Decision-Making*. John Wiley & Sons, Inc., 2019. ISBN 9781119650584.

Allan G. Bluman. *Elementary statistics : A step by step approach*. 2 Penn Plaza New York 10121, 2018. ISBN 9781259755330.

Yadolah Dodge. *The Concise Encyclopedia of Statistics*. Springer, New York, NY, 2008. ISBN 9781451115611. doi: https://doi.org/10.1007/978-0-387-32833-1.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. Taylor & Francis Group, LLC, Boca Raton, Florida, 2019. ISBN 9781138329164. doi: https://doi.org/10.1201/9780429448294.

Joseph M. Hilbe, editor. *Practical Guide to Logistic Regression*. CRC Press, Taylor and Francis Group, 2015. ISBN 978-1-4987-0958-3.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15 (3):651–674, 2006. doi: 10.1198/106186006X133933.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. ctree: Conditional inference trees. *The comprehensive R archive network*, 8, 2015.

Kuhn and Max. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008. doi: 10.18637/jss.v028.i05. URL `https://www.jstatsoft.org/index.php/jss/article/view/v028i05`.

Stefan Th. Gries Magali Paquot, editor. *A Practical Handbook of Corpus Linguistics*. Springer Cham, 2022. ISBN 978-3-030-46218-5. doi: https://doi.org/10.1007/978-3-030-46216-1.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

Michael Longnecker R. Lyman Ott. *An Introduction to Statistical Methods and Data Analysis*. Cengage Learning, Inc., Boston, USA, 06 2015. ISBN 9781305269477.

Zdenka Prokopova Radek Silhavy, Petr Silhavy, editor. *Data Science and Algorithms in Systems*. Springer Cham, 2023. ISBN 978-3-031-21438-7. doi: https://doi.org/10.1007/978-3-031-21438-7.

Terry Therneau and Beth Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2023. URL `https://CRAN.R-project.org/package=rpart`. R package version 4.1.23.

Hao Zhang, Zhuolin Li, Hossain Shahriar, Lixin Tao, Prabir Bhattacharya, and Ying Qian. Improving prediction accuracy for logistic regression on imbalanced datasets. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 918–919, 2019. doi: 10.1109/COMPSAC.2019.00140.

Appendix

A. Figures

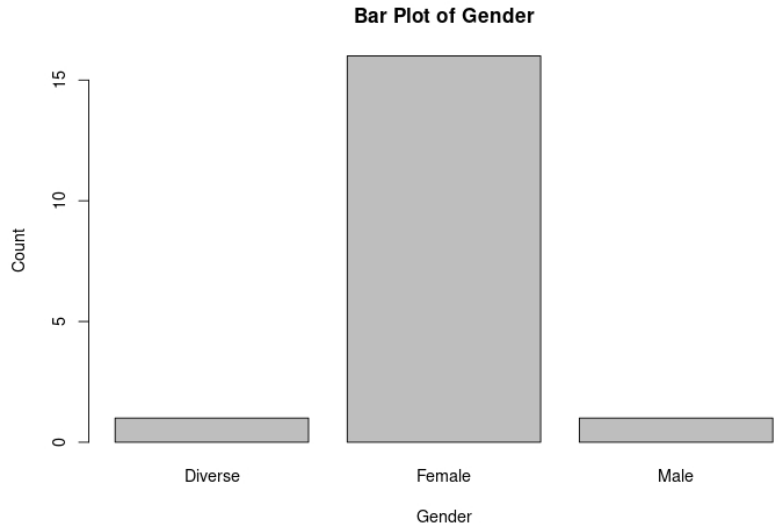


Figure A1: Box plot for word count

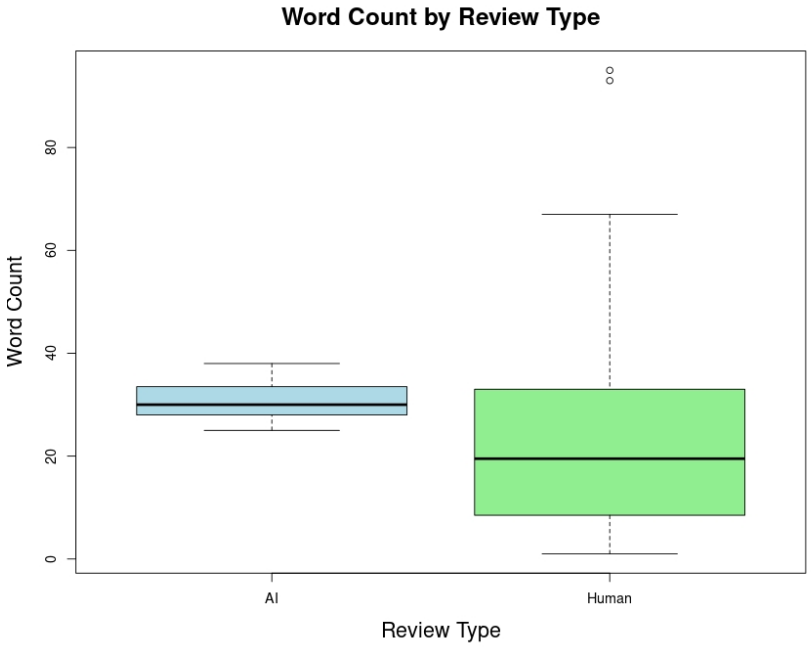


Figure A2: Box plot for word count