Table: Justification of 11 features selection to predict multiple diseases (continue).

Justification for selection 11 features			
Feature Category	**Feature**	**Feature Score**	**Reasons**
Lexical	Noun Count	Correlation Coefficient: 0.157469 Chi-Square: 106.006848 Mutual Information: 0.156650	Noun count is strongly related to text meaning, as nouns often serve as key elements in sentence structure. The chi-square value of 106 and the mutual information score of 0.156650 confirm its significant role in distinguishing the target variable, making it a valuable feature for classification.
Semantic	Polarity	Correlation Coefficient: 0.150401 Chi-Square: 118.543829 Mutual Information: 0.763562	Polarity reflects sentiment, which is essential for understanding text's emotional tone. The strong chi-square result and high mutual information (0.763562) emphasize its importance in predicting the target variable, supporting its inclusion as a top feature.
Syntactic	Fragments Proportion Score	Correlation Coefficient: 0.129982 Chi-Square: 33.211240 Mutual Information: 0.040863	Justification: The proportion of fragments indicates the complexity and structure of text. While the correlation is modest, the chi-square value (33.211240) and mutual information (0.040863) show that fragments significantly differentiate the target variable, thus validating its importance.
Lexical	Long Words Count	Correlation Coefficient: 0.069164 Chi-Square: 72.522411 Mutual Information: 0.098928	Longer words are associated with more complex text, which can be indicative of certain target classes. The chi-square and mutual information results support its relevance, even though the correlation coefficient is relatively lower, suggesting its importance in the classification task.
Semantic	Irony Sarcasm Count	Correlation Coefficient: 0.060460 Chi-Square: 7.413642 Mutual Information: 0.025005	Irony and sarcasm often provide nuanced information about text meaning. Although the correlation is low, the chi-square result (7.413642) and mutual information (0.025005) demonstrate its utility in distinguishing the target variable, making it relevant for the feature set.
Lexical	Average Word Length	Correlation Coefficient: 0.055547 Chi-Square: 57.588631 Mutual Information: 0.998094	Justification: Average word length can indicate the complexity of the text. The high mutual information value (0.998094) and reasonable chi-square (57.588631) support its contribution to predicting the target variable, ensuring its place in the top features.
Readability	Automated Readability Index	Correlation Coefficient: 0.048720 Chi-Square: 1359.460186 Mutual Information: 1.238450	The Automated Readability Index measures text readability, which is a key determinant of text classification. The very high chi-square value (1359.460186) and mutual information (1.238450) confirm its strong predictive power, supporting its inclusion in the top features.

Table: Justification of 11 features selection to predict multiple diseases.

<table>
<tr><td colspan="4">Justification for selection 11 features (continue)</td></tr>
<tr><th>Feature Category</th><th>Feature</th><th>Feature Score</th><th>Reasons</th></tr>
<tr><td>Syntactic</td><td>Syntax Tree Depth Variability Score</td><td>Correlation Coefficient: 0.036676
Chi-Square: 32.145670
Mutual Information: 0.734970</td><td>The depth and variability of a syntax tree reflect text structure and complexity. While the correlation is modest, the chi-square (32.145670) and mutual information (0.734970) show that this feature significantly contributes to distinguishing the target variable.</td></tr>
<tr><td>Semantic</td><td>Coreference Resolution Density</td><td>Correlation Coefficient: 0.020171
Chi-Square: 10.912349
Mutual Information: 0.213544</td><td>Justification: Coreference resolution is crucial for understanding text coherence. Despite a low correlation coefficient, the chi-square (10.912349) and mutual information (0.213544) values highlight its importance in determining text meaning and differentiating target categories.</td></tr>
<tr><td>Semantic</td><td>Relationships Variation</td><td>Correlation Coefficient: 0.015332
Chi-Square: 1.834900
Mutual Information: 0.023761</td><td>Variation in relationships provides insight into text structure and meaning. While the correlation coefficient is weak, the chi-square and mutual information results confirm that this feature contributes to distinguishing the target variable.</td></tr>
<tr><td>Lexical</td><td>Lexical Diversity Score</td><td>Correlation Coefficient: 0.005532
Chi-Square: 9.123765
Mutual Information: 0.106017</td><td>Lexical diversity measures the variety of vocabulary in a text. Although the correlation is low, its chi-square value (9.123765) and mutual information (0.106017) suggest that it helps differentiate between text classes, making it relevant for the analysis.</td></tr>
<tr><td colspan="4">2 Features not considered (despite positive correlation coefficient score)</td></tr>
<tr><td>Syntactic</td><td>Clause Boundary Standard Deviation</td><td>Correlation Coefficient: 0.001656
Chi-Square: 0.822956
Mutual Information: 0.026603</td><td>While it has a positive correlation with the target variable, both the chi-square (0.822956) and mutual information (0.026603) are very low, indicating that this feature does not significantly contribute to distinguishing the target. As a result, it was removed from the top features.</td></tr>
<tr><td>Lexical</td><td>Lexical Sophistication Score</td><td>Correlation Coefficient: 0.000953
Chi-Square: 0.210168
Mutual Information: 0.12842</td><td>Justification: This feature also shows positive correlation with the target, but the low chi-square (0.210168) and mutual information (0.12842) values indicate that it does not have a substantial impact on the target variable classification. Thus, it was excluded from the selected features.</td></tr>
</table>