Naive Bayes Classifier for Harry Potter Book Identification Using N-grams

Muhammed Muaaz Dawood (2425639)
Altaaf Ally (2424551)
Sayfullah Jumoorty (2430888)
Natural Language Processing
COMS4054A

I. Introduction

The objective of this project was to implement a Naive Bayes Classifier to identify which book a specific page belongs to within the Harry Potter series. The text was tokenized, and N-grams were extracted as features for the classifier. The impact of different N values (1, 2, 3, 4) on the model's performance was explored. Unigrams (N = 1) were used as a baseline to compare the effectiveness of higher-order N-grams (bigrams, trigrams, and four-grams). This allowed for an evaluation of how increasing the complexity of the feature representation affects the classifier's accuracy. The project was part of a broader exploration of the effects of tokenization and N-gram representations in text classification.

II. Contributions

1) **Altaaf Ally:** Data Preparation and Tokenization
2) **Sayfullah Jumoorty:** Model Development and Training
3) **Muhammed Muaaz Dawood:** Evaluation and Visualization

III. Methodology

A. Data Preparation

The dataset consisted of text from the seven Harry Potter books, split into pages. Each book was tokenized, with punctuation removed, and lowercased. The text was divided into N-grams, where N was varied between 1 and 4. These N-grams served as the features for the Naive Bayes Classifier.

The data was split into training, validation, and test sets with a ratio of 70:15:15. The classifier was trained on the training set, validated on the validation set, and evaluated on the test set.

B. Page Size and Limitations

For each book, we set the page size to 250 words, and limited each book to a maximum of 240 pages. This decision was made to ensure consistency in the dataset, as each book had a varying amount of pages. By standardizing the number of pages, we aimed to create a uniform training set, allowing for more reliable comparisons across different books and N-gram values.

The use of actual book pages in our classification approach presents several limitations. Firstly, the larger volume of text per page increases computational demands, requiring more processing time and memory. This approach also leads to a more extensive and potentially sparser feature space, which may impact the classifier's performance. Additionally, analyzing full pages might obscure subtle, line-level distinctions or unique phrases that could be more apparent in a line-by-line analysis, potentially reducing the granularity of the classification. These factors collectively affect the efficiency and potentially the accuracy of our book classification model.

C. Naive Bayes Classifier

A custom Naive Bayes Classifier was implemented without the use of external libraries like scikit-learn. The classifier calculates the probability of each class (book) given the input N-grams and predicts the class with the highest probability.

D. Evaluation Metrics

The classifier's performance was evaluated using accuracy on both validation and test sets. Additionally, a confusion matrix, seen in Figure 1, was generated to visualize the classification performance across different books.

IV. Results and Discussion

A. Impact of N-grams

The test accuracy for different values of N was recorded to evaluate the impact of varying N-gram sizes on the model's performance, and the results can be seen in Table I. The best performance was achieved with unigrams (N = 1) at 82.14% accuracy. As N increased to bigrams (N = 2), the accuracy decreased to 77.78%, and it continued to decline further with trigrams (N = 3) and four-grams (N = 4), achieving 59.92% and 40.87%, respectively. These results suggest that simpler models using unigrams were more effective for this text classification task, and that increasing N beyond unigrams led to a decline in performance, potentially due to overfitting or sparsity in higher-order N-gram representations.

TABLE I
TEST ACCURACY FOR DIFFERENT N-GRAM VALUES

N-gram Value	Test Accuracy (%)
1	82.14
2	77.78
3	59.92
4	40.87

B. Confusion Matrix

The confusion matrix (Figure 1) illustrates the classifier's performance across the seven books. The diagonal values indicate the number of correct predictions for each book. The classifier's predictions are uniform, meaning they do not favor one book over another. This ensures a balanced assessment across all categories.

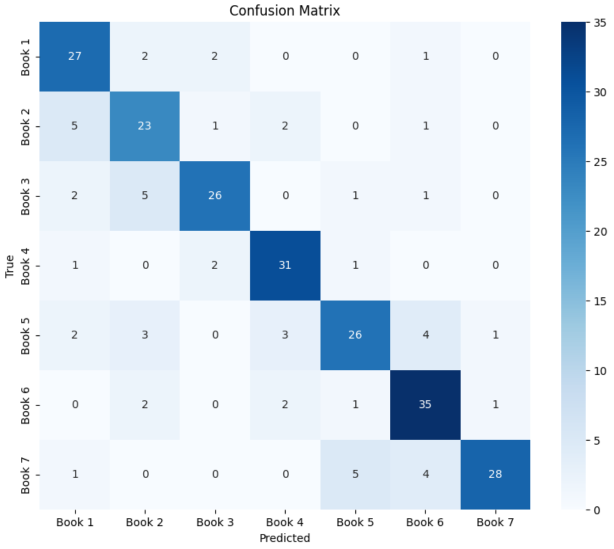

Fig. 1. Confusion Matrix for N=3

C. Insights

The exploration of N-gram sizes in this study reveals the delicate balance between model complexity and performance in text classification tasks. Contrary to the assumption that more complex N-grams would enhance the classifier's ability to capture contextual information, the results indicate that simplicity prevails. Unigrams (N = 1) emerged as the most effective feature, achieving the highest accuracy. This suggests that, for the Harry Potter series, individual word frequencies are sufficient for distinguishing between the books, without the need for more complex word combinations. The decline in accuracy with higher N-gram sizes may reflect the diminishing returns of adding more context, where the added complexity does not contribute to better classification but rather introduces potential noise. This outcome highlights a critical insight: in text classification, especially within a homogeneous set like the Harry Potter books, the simplest approach can sometimes be the most powerful.

V. CONCLUSION

The experiment demonstrated that a Naive Bayes Classifier can effectively classify pages from the Harry Potter series based on N-gram features, with unigrams providing the highest accuracy.