Leveraging N-Gram Part-of-Speech Tag Sequences for Automatic Genre Classification

Isabel Arvelo
Vanderbilt University
isabel.c.arvelo@vanderbilt.edu

ABSTRACT

This study investigates the distinctive linguistic structures between fiction and non-fiction texts using n-gram part-of-speech (POS) tag sequences. By analyzing a corpus of English texts from Project Gutenberg, we examine how these structures capture genre-specific language patterns and assess the effectiveness of incorporating longer n-gram sequences in a binary classification task. Statistical analysis reveals that verbs and pronouns are the most characteristic POS tags in fiction, while adpositions and conjunctions are more prevalent in non-fiction. Logistic regression models trained on unigram and 1–4-gram POS features achieve high accuracy in genre classification, with the 1-4-gram model performing slightly better. However, the marginal improvement suggests that additional information provided by longer n-grams may not significantly enhance the model's discriminative power. The findings offer insights into the linguistic differences between genres and highlight the potential of using POS tags for automated genre classification in various applications, such as book recommendation systems and library cataloging.

Keywords

N-grams, Part of Speech Tagging, Automatic Genre Classification

1. INTRODUCTION

Categorizing literary works into genres has had longstanding importance in literary theory and these categorizations have been evolving since the classical era [12]. Literary genres offer a taxonomy to classify texts, but the term itself is ambiguous as there is a rich body of work on competing approaches for defining genre and how it can be used in practice [7]. There is no widely accepted or standardized framework for what constitutes a genre and classification systems for texts vary greatly by context [4]. From a linguistic perspective, a genre is a collection of texts that possess a set of shared features, which differentiate them from texts in other categories. For the purposes of this paper, we will use Crown's text classification system [6] that groups text types into three broad categories: fiction, non-fiction, and poetry. Genre classification is not only important in literary studies, but with the increased digitization of books [26] and wealth of publicly available free-text data online, there are several potential applications in various tasks such as book discovery and organization and information retrieval [12]. Texts within a genre are characterized by unique syntactic and lexical features [18] and previous studies have explored the use of various linguistic features for automatic genre classification, including lexical n-grams [18], syntactic structures, and part-of-speech (POS) tags [2]. Part-of-speech tags have been used as features in several highly performant genre classifiers, but there has not been a lot of work considering n-gram sequences of POS tags. Lexical n-grams have been found to vary between different types of texts [21], indicating that n-gram sequences of POS tags may also have the potential to capture more varied and complex structures distinctive to each genre, but studies employing this approach [23] have not focused on the specific binary classification of fiction vs non-fiction.

2. PURPOSE STATEMENT AND RESEARCH QUESTION

The present study aims to address gaps in the literature by investigating two main research questions:

1. How do the linguistic structures captured by n-gram POS tag sequences differ between fiction and non-fiction texts, and what do these differences reveal about genre-specific language patterns?
2. To what extent can the incorporation of longer n-gram POS tag sequences, as opposed to individual POS tags, enhance the ability of a classification algorithm to differentiate between fiction and non-fiction texts?

3. LITERATURE REVIEW

Investigating the predictive value of part of speech as a discriminator has been a major area of focus in automatic genre classification research. Mendhakar and H S, developed an artificial neural network classifier to classify fictional and non-fiction texts using only 9 POS tags. Their results indicated an overall classification accuracy of 98% [16]. There have been several other studies that have investigated the use of POS tags in distinguishing between different text genres. Qureshi et al. built a logistic regression classifier to differentiate between fiction and non-fiction genres utilizing POS tags among nineteen other "low-level", "high-level" and derived features" [20] to classify texts from the Brown Corpus [9] and the British National Corpus [3]. The results demonstrated classification accuracies of 100% and 96.31% for each corpus, respectively. They also found that a classifier containing only two features - the ratio of adverbs to adjectives and adjectives to pronouns was just as performant as a classifier with a more exhaustive set of features.

In addition to genre classification, POS tags have also been used to study differences in language patterns between genres. Biber used POS histograms to capture genre-specific language differences [2], while Feldman presented statistics showing variations in the frequency of different parts of speech and syntactic structures between fiction and exposition [8].

Building on a large body of work using unigram POS tags, Tang and Cao examined whether n-grams of POS tags could be even more effective in discriminating between genres. Their study found

a strong relationship between the information captured by POS n-grams and the genre of a given text. This correlation tended to be stronger as the length of the n-grams increased, suggesting that longer POS sequences may capture more unique and informative syntactic patterns that can be leveraged for accurate genre classification [23].

These findings highlight the potential of using POS tags as features in machine learning techniques to accurately classify texts, with some studies achieving near-perfect accuracy rates. However, the variability in performance across different datasets and classification methods suggests that further research is needed to develop more robust and generalizable approaches.

The use of POS n-grams has also been explored in other NLP tasks, such as information retrieval and personality profiling. Lioma and van Rijsbergen proposed the use of POS n-grams, which they referred to as "POS contexts,"[14] to capture the grammatical and structural aspects of language. They suggested that these POS contexts could be used to identify the informative content of words for information retrieval. Wright and Chin demonstrated that the presence of certain POS n-grams improved the accuracy of personality prediction, suggesting that there is a relationship between syntactic structures represented by POS n-grams and writer personality [24].

While the use of POS n-grams has shown promise in various NLP tasks, there is still a need for further research to determine the optimal length of n-grams for specific purposes and to explore the potential of incorporating longer n-gram sequences in genre classification. As Gries and Mukherjee noted, although 4-grams have gained popularity in recent research, there is still uncertainty regarding the optimal value of n for various tasks [10]. The choice may depend on the specific application and research question at hand.

The findings from this research may provide insight into specific n-gram linguistic structures unique to fiction and non-fiction genres, as well as the potential benefits of incorporating longer n-gram POS tag sequences into genre classification algorithms.

4. METHODS

4.1 Data

The corpus for this project is made up of fiction and non-fiction English texts from Project Gutenberg using a dataset compiled by Nagyfi on HuggingFace [17] which contains more than 80% of all of the books written in English available on the site. Project Gutenberg is an online library offering over 70,000 free eBooks contributed by over 10,000 individuals, including a wide range of classic literature and historical texts. The project was founded in 1971 and is known as the oldest digital library. It is completely volunteer run and aims to digitize and archive cultural works and make them freely accessible to the public. After selecting 9999 texts from the HuggingFace dataset, the corpus was divided into three categories: fiction, non-fiction, and poetry. The texts were categorized based on the metadata available in the dataset which included the title, author, language, and subject. The subject field included terms like "*Fiction", "Detective and mystery stories",* and *"Doctrines".* Keywords related to fiction and non-fiction were used to categorize the texts as based on the subject tags. Poetry texts were explicitly filtered out, as this type of writing is materially different from the genres considered in this study.

A total of 2960 texts were randomly selected and texts with fewer than150 words were excluded. The final corpus was split almost evenly between fiction (n = 1477) and non-fiction texts (n = 1435). The text for each book was preprocessed to remove any metadata and non-textual content, such as table of contents, copyright information, and footnotes. The length of the texts across the corpus varies great due to its diverse composition including an entire collection of Aesop's Fables and the constitutional writings of Abraham Lincoln. The average word count is 55,165 (SD= 41,493) after data cleaning. On average, fiction texts to be longer, averaging 62,007(SD = 41,813) words per text as compared to nonfiction texts that had an average of 48,123 (SD = 39,971) words per text.

4.2 Part-of-Speech Tags

A part of speech tag is an annotation that identifies the syntactic function and properties of a word within a sentence or phrase. A tag set is a standardized set of labels that represent the different syntactic categories assigned to words in a text [10]. The part of speech tags for this analysis come from the Universal Dependency Tags [6] which are the coarse-grained POS tags in the python library spaCy. There are 17 core part-of-speech categories in the tag set, but 10 specific tags used were chosen based on previous work with POS n-grams [23] to reduce the computational complexity of the analysis: adjectives, adverbs, nouns, verbs, pronouns, adpositions, interjections, and conjunctions (subordinating and coordinating).

Adjectives (ADJ) are descriptive words that modify nouns and add specificity to text by providing details such as size and color. Adverbs (ADV) modify verbs, adjectives, or other adverbs, offering information about the place, time, frequency or degree of an action or state of being. Nouns (NOUN) refer to people, places, things, or ideas. Verbs (VERB) indicate actions performed by a subject or the subject's state of being. Pronouns (PRON) replace nouns to avoid repetition and simplify sentences. Adpositions (ADP), account for both prepositions and postpositions, and typically express spatial or temporal relations. Interjections (INTJ) are expressive words like "Ah!" or "Heavens!" that convey strong emotion or sudden bursts of feeling. Coordinating conjunctions (CCONJ) connect words, phrases, or clauses of equal importance and subordinating conjunctions (SCONJ) introduce subordinate clauses and connect them to main clauses. Both types of conjunctions were grouped into one category for the purposes of this analysis. "X" is the tag used for words that do not fall into any of the defined part-of-speech categories, and is most often assigned to word fragments, foreign words and gibberish words [6].

An n-gram POS tag refers to a series of n adjacent part-of-speech labels that encode the grammatical context of a specific portion of text, such as a 2-gram (NOUN-VERB), 3-gram (ADJ-ADJ-NOUN), or 4-gram (PRON-VERB-ADJ-NOUN) of POS tags. Using a core set of 10 POS tags results in 100 possible bigrams, 1000 possible trigrams and 10000 possible 4-grams. To limit the complexity of the models, we only considered the topmost frequent 300 n-grams, adopting a similar approach to previous work with n-grams [23].

4.3 Statistical Analysis

4.3.1 Unigrams Across Fiction and Non-Fiction Texts

To compare the POS tag distributions between fiction and non-fiction texts, the non-parametric Mann-Whitney U test was used. This statistical test is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous. This non-parametric test was chosen because it does not assume normality and is robust to outliers. All the POS tag distributions are not normally distributed and according to Meissel and Yao parametric effect sizes can be used without any disadvantages,

regardless of whether the data meets the assumptions of normality [15]. The Mann-Whitney U test was used to test the null hypothesis that the distributions of each of the POS tags in fiction and non-fiction texts are the same.

Since the sample sizes are large, the Mann-Whitney U test is expected to be significant even for small differences [1]. Therefore, the effect size of the differences between the POS tag distributions in fiction and non-fiction texts was calculated using Cliff's Delta, a non-parametric measure of standardized mean difference (SMD) effect sizes [15].

Cliff's Delta quantifies the amount of difference between two groups of data. In this study, it was calculated by subtracting the probability that a randomly selected value (in this case, a specific POS tag) from the fiction genre is greater than a randomly selected value from the non-fiction genre, from the probability that a randomly selected value from the fiction group is less than a randomly selected value from the non-fiction group. The resulting value ranges from -1 to 1. A value of -1 indicates that all values in the fiction group are less than all values in the non-fiction group and value of 1 indicates that all values in the fiction group are greater than all values in the non-fiction group.

4.3.2 Classification Models

Two logistic regression models were developed and trained on the same dataset to predict the genre of a given text. Logistic Regression is a linear model that predicts the probability of a binary outcome. It is a good choice for text classification tasks because it is computationally efficient and interpretable.

The first model used the normalized counts of the 10 selected part-of-speech (POS) tags as input features. The second model included a combination of unigrams, 2-grams, 3-grams, and 4-grams derived from the same set of POS tags. By comparing the performance of these two models, we assessed the effectiveness of incorporating higher-order POS n-grams in capturing genre-specific linguistic patterns and improving the accuracy of genre classification.

When evaluating the performance of each classification model, three common metrics were used: precision, recall and F-1 score. Precision measures the percentage of correctly predicted positive instances out of all predicted positive instances. Recall measures the percentage of correctly predicted positive instances out of all actual positive instances and the F1-score is the harmonic mean of precision and recall, ranging between 0 and 1, providing a balanced measure of a model's performance.

Cross-validation was used to tune the hyperparameters of each logistic regression model to optimize its performance. Cross-validation is the process of splitting the data into training and testing sets multiple times to ensure that the model is robust to different splits of the data. This method provides a more accurate estimate of the model's performance on unseen data. This study used 5-fold cross-validation, which involves splitting the data into 5 equal parts and training the model on 4 parts while testing it on the remaining part. This process was repeated 5 times, with each part serving as the test set once. The final performance metric is the average of the performance metrics obtained in each fold. We used Grid Search Cross Validation to grid search over hyperparameters for each algorithm and then compare the performance of the best model (using optimal hyperparameters) for each of our feature sets.

For logistic regression, C is a hyperparameter that controls the regularization, a technique used to prevent overfitting by penalizing large coefficients in the model. Smaller values of C specify stronger regularization. Since the default is 1, we used a range of values around that on the log scale in the grid search to specify the optimal value for our specific text classification task. The models were optimized to maximize their weighted F-1 score, which is a common metric for binary classification tasks that measures the trade-off between precision and recall.

4.3.3 N-Gram Feature Importance

Given the way the number of POS n-grams scale exponentially as n increases, making pairwise comparisons of each n-gram between fiction and non-fiction text is not as straightforward or meaningful when working with 2-, 3-, and 4- grams.

Instead, by comparing the coefficients of different n-gram POS tags in each model, we can identify the most important or "key" linguistic features that characterize and differentiate fiction and non-fiction texts. This concept of keyness helps us understand the distinctive n-gram POS tag patterns and markers associated with each genre, providing insights into the use of language and structure of each type of text.

Each coefficient in both fitted logistic regression models represents the degree to which the presence of a specific n-gram POS tag influences the classification of a text as either fiction or non-fiction. Specifically, the value represents the change in the log-odds of a text being classified as non-fiction versus fiction when the corresponding n-gram POS tag is present in the text, assuming all other variables remain constant, and the magnitude indicates the strength of the association between the tag and the text's classification. Larger absolute values (positive or negative) suggest a stronger influence of the n-gram POS tag on the classification, while values closer to zero indicate a weaker association.

Associated code and details about the data used in this analysis can be found at: https://github.com/isabelarvelo/n-gram-pos-genre-classification.

5. RESULTS

5.1 Unigrams

5.1.1 Unigrams Across Fiction and Non-Fiction Texts

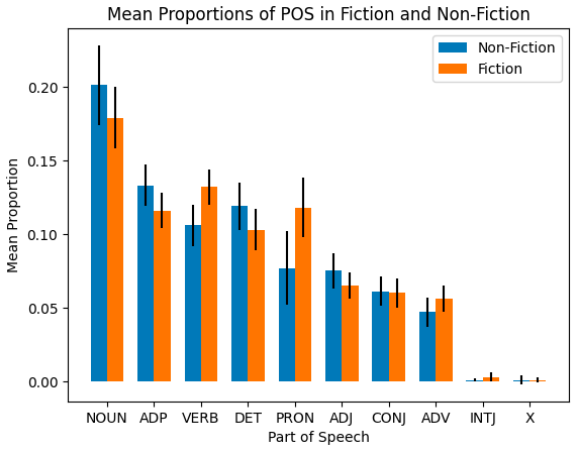

Figure 1. Distribution of Normalized Counts for Each POS Unigram in Fiction and Non-Fiction Texts

In non-fiction, there's a higher prevalence of nouns (Mean = 0.201, SD = 0.027) and adpositions (Mean = 0.133, SD = 0.014). Conversely, fiction is characterized by a greater use of verbs (Mean =

0.132, SD = 0.012) and pronouns (Mean = 0.118, SD = 0.02), Adjectives and adverbs show moderate usage across both genres, with slightly higher averages in non-fiction (MEAN = 0.075 SD = 0.012) and fiction (Mean =0.056, SD = 0.009) respectively. The standard deviations across these POS tags suggest a degree of variability within each genre, pointing to diverse authorial styles and narrative techniques. Since multiple comparisons were made (one for each POS tag), Bonferroni's correction was applied to adjust the significance level and reduce the likelihood of making a Type I error (concluding that there is a significant difference between the groups when there is no actual difference). Bonferroni's correction divides the original significance level by the number of comparisons made, resulting in a more conservative threshold for determining statistical significance.

After applying Bonferroni's correction, the Mann-Whitney U test results showed that the distributions of all unigram POS tags aresignificantly different between fiction and non-fiction texts. This finding suggests that the usage of each POS tag differs between the two genres, and these differences are unlikely to have occurred by chance.

When examining the effect sizes, several patterns emerge. For adjectives (ADJ), a Cliff's Delta value of -0.499704 indicates a medium to large negative effect size, suggesting that adjectives are more prevalent in non-fiction compared to fiction. In contrast, adverbs (ADV) exhibit a positive value of 0.511729, implying a higher frequency in fiction.

Nouns (NOUN) show a medium negative effect size (-0.495123), indicating they are more common in non-fiction texts, while verbs (VERB) demonstrate a large positive effect size (0.865608), strongly suggesting their greater frequency in fiction.

Adpositions (ADP) have a large negative effect size (-0.670082), signifying a more significant presence in non-fiction, whereas interjections (INTJ) display a large positive effect size (0.713295), indicating a notably higher occurrence in fiction. Similarly, pronouns (PRON) exhibit a large positive effect size (0.822085), suggesting their frequent use in fiction compared to non-fiction. The difference in the usage of conjunctions (CONJ) between the two genres is not very pronounced, as evidenced by the relatively smaller effect size.

These findings highlight distinct linguistic patterns across genres, with certain POS tags like verbs and pronouns being more dominant in fiction, while adpositions are more prevalent in non-fiction.

5.1.2 Unigram Genre Classification Model

The logistic regression model using unigram POS tags performs well in distinguishing between fiction and non-fiction texts, achieving high precision (87% for fiction, 91% for non-fiction), recall (91% for fiction, 86% for non-fiction), and F1-scores (0.89 for fiction, 0.88 for non-fiction). The overall accuracy of 89% and the weighted average of 0.89 indicate that the model effectively captures the linguistic differences between the two classes based on the unigram POS tag features and the model is slightly better at detecting fiction texts, but its non-fiction predictions are a bit more reliable.

For this model, POS tags with large negative coefficients strongly indicate that a text is more likely to be fiction, while POS tags with large positive coefficients strongly suggest that a text is more likely to be non-fiction. We will compare the coefficients that are the most predictive as a way of assessing keyness, or the degree to which each part-of-speech (POS) tag is characteristic of or strongly associated with fiction or non-fiction.

Table 1. Unigram Logistic Regression Classification Report

	precision	recall	f1-score	support
Fiction	0.87	0.91	0.89	295
Non-fiction	0.91	0.86	0.88	288
Accuracy			0.89	583
Weighted Average	0.89	0.89	0.89	583

Table 2. Unigram Logistic Regression Coefficients

Fiction	POS	Non-Fiction	POS
-32.87	VERB	14.74	ADP
-32.34	PRON	10.89	CONJ
-7.44	ADV	7.49	ADJ
-1.97	INTJ	6.14	DET
		3.38	NOUN
		0.32	X

5.2 N-gram POS Tag Sequences

5.2.1 1- 4 grams Classification Model

Comparing the two models, the 1–4-gram model appears to perform slightly better overall. It has a higher precision (+1%) and recall (+1%) for both genres. However, these differences are marginal, especially considering that this model has 700 more features than the unigram model.

Table 3. N-gram Logistic Regression Classification Report

	precision	recall	f1-score	support
Fiction	0.88	0.92	0.90	295
Non-fiction	0.92	0.87	0.89	288
Accuracy			0.90	583
Weighted Average	0.90	0.90	0.90	583

5.2.2 1-4 grams Across Fiction and Non-Fiction Texts

To gain insights from the 710 coefficients in the 1-4 gram logistic regression model, we focused on the top 10 and bottom 10 coefficients. These coefficients represent the n-grams that have the strongest associations with either fiction or non-fiction texts.

By examining the top 10 coefficients, we can identify the n-grams that are most indicative of fiction texts. These n-grams have the highest positive coefficients, meaning that their presence in a text strongly suggests that it belongs to the fiction genre. Similarly, the bottom 10 coefficients correspond to the n-grams that are most characteristic of non-fiction texts. These n-grams have the lowest

negative coefficients, indicating that their presence is a strong indicator of the non-fiction genre.

In addition to determining the most "key" n-grams for each genre, analyzing the top and bottom coefficients allows us to observe the distribution of 1-, 2-, 3-, and 4-grams among the most important features in the model. By counting the occurrences of each n-gram type (unigrams, bigrams, trigrams, and 4-grams) within the top 10 and bottom 10 coefficients, we can determine whether certain n-gram lengths tend to be more informative for genre classification.

Table 4. N-gram Logistic Regression Coefficients

Fiction	POS	Non-Fic	POS
-26.70	VERB	11.94	ADP
-25.31	PRON	9.94	CONJ
-10.28	PRON_VERB	8.91	NOUN_ADP
-8.70	VERB_PRON	5.52	ADP_NOUN
-8.39	NOUN_VERB	4.86	ADJ
-7.50	PRON_NOUN	4.61	DET
-6.45	ADV	4.31	NOUN_CONJ
-5.39	ADP_PRON	3.32	CONJ_NOUN
-5.15	VERB_ADV	3.16	NOUN_ADP_NOUN
-4.29	DET_NOUN	3.13	NOUN

For fiction texts, the most important features are unigrams "VERB" and "PRON", and bigrams "PRON_VERB", "VERB_PRON", and "NOUN_VERB". For non-fiction texts, the most important features are unigrams "ADP", "CONJ", and "ADJ", and bigrams "NOUN_ADP" and "ADP_NOUN". The results show that unigrams and bigrams are the most influential n-grams for both genres, with only one trigram appearing in the top 10 for non-fiction texts and no 4-grams present in the top 10 for either genre.

6. DISCUSSION

Verbs and pronouns appear to be the most distinctive parts of speech (POS) tags for differentiating between fiction and non-fiction texts. This finding is evident in two aspects of the analysis:

- The comparison of unigram frequencies across genres shows that fiction texts have higher normalized counts of verbs and pronouns compared to non-fiction texts.
- In the first logistic regression model, which uses only unigram POS tags as features, the coefficients for verbs and pronouns have the highest absolute values. This indicates that the presence of these POS tags strongly contributes to the classification of a text as fiction.

This makes sense because fiction often involves narration, which heavily rely on action words (verbs) to describe events, characters' actions, and plot progression. Fiction stories often revolve around characters and their experiences, so pronouns are essential for describing their actions and relationships without the redundant use of proper nouns.

On the other hand, adpositions (ADP) and conjunctions (CONJ) are most strongly associated with non-fiction texts. This observation is consistent with the characteristics of non-fiction writing, which often deals with complex ideas and concepts requiring detailed explanations and clear connections between ideas. Adpositions help establish relationships between words, while conjunctions link ideas and create coherent arguments, both of which are crucial in non-fiction writing.

When comparing the two classification models, the very similar performance suggests that the additional information provided by the 2–4-gram POS tags may not significantly improve the model's ability to distinguish between fiction and non-fiction texts compared to using only unigram POS tags. Given the marginal improvement in performance achieved by the 1–4-gram model and the increased complexity and effort required for n-gram labeling, it may not be worthwhile to invest the extra resources in developing and implementing the more complex model.

Classification performance is not the only consideration when choosing the best model; factors such as computational resources, time constraints, and the interpretability of the results are also important factors. The unigram model, being simpler and more straightforward, offers a more practical solution for most applications. It requires less computational power, is easier to interpret, and can be trained and deployed more efficiently. The slight gain in accuracy offered by the 1–4-gram model may not justify the additional time and effort needed to create and maintain the more intricate model, especially when considering the scalability and resources required for real-world implementations. Therefore, unless there is a specific need for the highest possible accuracy, the unigram model presents a more cost-effective and efficient approach for distinguishing between fiction and non-fiction texts based on their POS tag distributions.

Further analysis reveals that longer n-grams do not appear to be more strongly associated with genre. Among the most "key" POS n-grams for each genre, unigrams are the most common, followed by bigrams, with no 4-grams present at all. Interestingly, seven of the ten most distinctive n-grams for fiction included one of the two most distinctive unigrams (verbs and pronouns), while eight out of the ten for non-fiction included one of the two most distinctive unigrams (adjectives and conjunctions). This suggests that there is significant overlapping information between the most "key" n-grams and that the association of these longer n-grams could be driven by the unigrams within them.

There are a variety of potential applications for a genre classification model like the ones developed in this study. Online bookstores, libraries, and reading apps can integrate genre classification model to provide personalized book suggestions to their users. By analyzing the text of books and identifying their genres, the model can help readers discover new titles that align with their reading preferences. This not only enhances the user experience but also promotes book discoverability, exposing readers to a wider range of books within their favorite genres. Moreover, the model could assist librarians in cataloging and organizing large book collections. Manually categorizing books into genres can be a time-consuming and labor-intensive task. By automating the genre classification process, the model can save librarians significant time and effort. It would enable them to assign genre labels quickly and accurately to books based on their text, streamlining the cataloging process. Furthermore, the methodology used to develop this model can be

applied to other domains, such as movie and TV show recommendation systems, where scripts play a crucial role in determining the genre and target audience. By adapting the model to these contexts, it can provide valuable insights and recommendations to viewers, enhancing their entertainment experience. These preliminary results indicate that even a simple, low-resource, and easy-to-implement model that uses only POS tags as features can effectively classify texts into different categories.

It is important to note that while more specific genres or the inclusion of sub-genres could be more useful, it complicates the task because there is no objective hierarchy or straightforward way to choose one categorization over the other. For example, a novel set in a dystopian future with a strong female protagonist who challenges societal norms could be classified as Dystopian Fiction, Feminist Literature, or Social Commentary, depending on the perspective and context. With a model trying to predict a single subgenre, this ambiguity could lead to inconsistent labels and potentially reduce the model's accuracy. The subjectivity involved in assigning these types of labels could result in disagreements among annotators, making it challenging to create a reliable and consistent training dataset. By operating at a higher level of abstraction, the model developed in this study avoids the ambiguity and subjectivity associated with more granular genre classifications. This allows for a more consistent and reliable classification system that can still provide valuable insights and applications in various domains, such as book recommendation systems and content organization.

In conclusion, these preliminary results indicate that even a simple, low-resource, and easy-to-implement model that uses only unigram POS tags as features can effectively classify texts as fiction and non-fiction texts.

7. LIMITATIONS

The analysis relies on a relatively small subset of texts from the Project Gutenberg corpus that may not encompass the full diversity of fiction and non-fiction texts, potentially limiting the generalizability of the findings to other bodies of literature or contemporary works. Additionally, the model is only making predictions on fiction or nonfiction books, so it is not able to predict other genres and this performance may not generalize to a corpus that includes the text of books in other genres.

While efforts were made to preprocess the texts by removing extraneous textual content, this process was not perfect or validated with further testing. Errors in the cleaning process might have left some unrefined text, leading to variations in text quality and length that could impact the results.

Due to hardware constraints, this analysis used spaCy's small model which, while efficient and less resource-intensive, may not capture the linguistic nuances as effectively as the larger, more comprehensive models. This limitation impacts the depth and accuracy of POS tagging and, consequently, the overall findings of the study. Therefore, the results presented in this paper should be viewed in the context of these methodological constraints.

8. FUTURE WORK

This study focused on the binary classification of texts as either fiction or non-fiction. However, it may be more practical and informative to describe texts using a more nuanced approach, such as the generic facets framework proposed by Kessler et al [12]. This framework allows for a text to be characterized by multiple genres or attributes, rather than being limited to a single category. By describing texts in terms of specific structural or linguistic cues, such as "Narrative" or "Opinion," we can capture the key characteristics of a text that may be more relevant to certain applications [12]. This approach can also help in understanding and classifying texts that do not neatly fit into the original categories on which an automatic genre classification system was trained and would likely a require a multi-label classification model to allow for non-exclusive labels.

Another potential avenue that could be explored in future studies would be to expand the range of part-of-speech tags analyzed, with a particular focus on investigating patterns within specific types of more fine-grained parts of speech. This could provide deeper insights into the narrative styles and linguistic features characteristic of different genres.

9. ACKNOWLEDGMENTS

ChatGPT was used to proofread this manuscript.

10. REFERENCES

[1] Andrade, C. (2020). Sample Size and its Importance in Research. **Indian Journal of Psychological Medicine, 42**(1), 102–103. https://doi.org/10.4103/IJPSYM.IJPSYM_504_19

[2] Biber, D., & Armstrong, S. (1994). Using register-diversified corpora for general language studies. *Using large corpora*, 180-201.

[3] BNC Consortium. (2007). British national corpus. *Oxford Text Archive Core Collection*

[4] Chandler, D. An introduction to genre theory, 1997. URL http://www.aber.ac. uk/media/Documents/intgenre/intgenre.html.

[5] Crown, A. (2013). *Guide to Text Types: Narrative, Non-Fiction and Poetry*. National Literacy Trust. https://www.thomastallisschool.com/uploads/2/2/8/7/2287089/guide_to_text_types_final-1.pdf

[6] Marneffe, M. C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014, May). Universal Stanford dependencies: A cross-linguistic typology. In *LREC* (Vol. 14, pp. 4585-4592).

[7] Fakhruddin, W. F. W. W., & Hassan, H. (2015). A review of genre approaches within linguistic traditions. *LSP International Journal*, *2*(2).

[8] Feldman, S., Marin, M. A., Ostendorf, M., & Gupta, M. R. (2009, April). Part-of-speech histograms for genre classification of text. In *2009 IEEE international conference on acoustics, speech and signal processing* (pp. 4781-4784). IEEE. DOI= https://doi.org/10.1109/ICASSP.2009.4960700.

[9] Francis, W. N., & Kucera, H. (1979). Brown corpus manual. *Letters to the Editor*, *5*(2), 7.

[10] Gries, S. T., & Mukherjee, J. (2010). Lexical gravity across varieties of English: An ICE-based study of n-grams in Asian Englishes. *International Journal of Corpus Linguistics*, *15*(4), 520-548.

[11] Jurafsky, D., & Martin, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

[12] Kessler, B., Nunberg, G., & Schütze, H. (1997). Automatic detection of text genre. *arXiv preprint cmp-lg/9707002*.

[13] Klarer, M. (2013). *An introduction to literary studies*. Routledge.

[14] Lioma, C., & van Rijsbergen, C. K. (2008). Part of speech n-grams and information retrieval. Revue française de linguistique appliquée, (1), 009-022.

[15] Meissel, K., & Yao, E. S. (2024). Using Cliff's Delta as a Non-Parametric Effect Size Measure: An Accessible Web App and R Tutorial. *Practical Assessment, Research, and Evaluation*, *29*(1).

[16] Mendhakar, A. & H S, D. (2023). Parts-of-Speech (PoS) Analysis and Classification of Various Text Genres. Corpus-based Studies across Humanities, 1(1), 99-131. DOI= https://doi.org/10.1515/csh-2023-0002.

[17] Nagyfi, R. (2023). Sedthh/gutenberg_english · datasets at hugging face. *Hugging Face*. https://huggingface.co/datasets/sedthh/gutenberg_english

[18] Philipp Petrenz, Bonnie Webber; Stable Classification of Text Genres. Computational Linguistics 2011; 37 (2): 385–393. DOI= https://doi.org/10.1162/COLI_a_00052.

[19] Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, *5*(2), 221.

[20] Qureshi, M. R., Ranjan, S., Rajkumar, R., & Shah, K. (2019, August). A simple approach to classify fictional and non-fictional genres. In *Proceedings of the Second Workshop on Storytelling* (pp. 81-89). DOI= https://doi.org/10.18653/v1/W19-3409

[21] Rahmoun, A., & Elberrichi, Z. (2007). Experimenting N-Grams in Text Categorization. *Int. Arab J. Inf. Technol.*, *4*(4), 377-385.

[22] Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Text genre detection using common word frequencies. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

[23] Tang, X., & Cao, J. (2015). Automatic genre classification via n-grams of part-of-speech tags. Procedia-Social and Behavioral Sciences, 198, 474-478.. DOI= https://doi.org/10.1016/j.sbspro.2015.07.468.

[24] Wright, W. R., & Chin, D. N. (2014). Personality profiling from text: introducing part-of-speech N-grams. In *User Modeling, Adaptation, and Personalization: 22nd International Conference, UMAP 2014, Aalborg, Denmark, July 7-11, 2014. Proceedings 22* (pp. 243-253). Springer International Publishing.

[25] Santini, M. (2004, January). A shallow approach to syntactic feature extraction for genre classification. In *proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics* (pp. 6-7). Birmingham, UK.

[26] *33% growth for digital books from public libraries and schools in 2020 sets records*. OverDrive. (2021, February 8). https://company.overdrive.com/2021/01/07/33-growth-for-digital-books-from-public-libraries-and-schools-in-2020-sets-records/