Syntax and Themes: How Context Free Grammar Rules and Semantic Word Association Influence Book Success

Henry Gorelick[1], **Biddut Sarker Bijoy**[*2], **Syeda Jannatus Saba**[*2], **Sudipta Kar**[3], **Md Saiful Islam**[2, 4], **Mohammad Ruhul Amin**[1]

[1]Fordham University
[2]Shahjalal University of Science and Technology
[3]Amazon Alexa AI
[4]University of Alberta
hgorelick@fordham.edu,
{biddut12,syeda06}@student.sust.edu, skar3@uh.edu,
mdsaiful1@ualberta.ca, mamin17@fordham.edu

Abstract

In this paper, we attempt to improve upon the state-of-the-art in predicting a novel's success by modeling the lexical semantic relationships of its contents. We created the largest dataset used in such a project containing lexical data from 17,962 books from Project Gutenberg. We utilized domain specific feature reduction techniques to implement the most accurate models to date for predicting book success, with our best model achieving an average accuracy of 94.0%. By analyzing the model parameters, we extracted the successful semantic relationships from books of 12 different genres. We finally mapped those semantic relations to a set of themes, as defined in *Roget's Thesaurus* and discovered the themes that successful books of a given genre prioritize. At the end of the paper, we further showed that our model demonstrate similar performance for book success prediction even when Goodreads rating was used instead of download count to measure success.

1 Introduction

Since its publication in 1868, approximately 1.78 million copies of Louisa May Alcott's *Little Women* have been sold, which equates to about 1,000 copies a month for 152 years. Every publisher in the industry hopes to find a manuscript that can sell even 10,000 copies in its lifetime. This begs the question: what makes Little Women a timeless success? Recently, researchers have attempted to use machine learning and natural language processing to answer this question, among others.

Predicting the success of a novel by analyzing its content is a challenging research problem. Thousands of new books are published every year, and only a fraction of them achieve wide popularity. Therefore, the ability to predict a book's success prior to publication would be exceptionally useful to the publishing industry and enable editors to make better decisions. Many factors contribute to a book's success including, but not limited to plot, setting, character development, etc. Additionally, there are some other factors that contribute to a book's popularity that an author and publisher cannot control like the time when a book is published, the author's reputation, and the marketing strategy. In this paper, we only focus on the content of the book to predict its popularity.

In this paper, we explore whether novels in a spe-

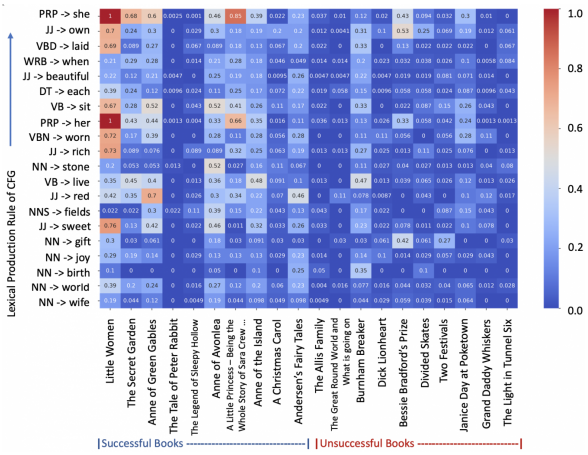

Figure 1: This figure represents the lexical production rules of context free grammar observed in both successful and unsuccessful books classified by Goodreads ratings. We present the count of 20 lexical rules, each normalized with respect to it's highest occurrence in any book, for 10 successful and 10 unsuccessful books from CHILDREN genre. We see that certain lexical rules occur more frequently in successful books than unsuccessful books.

[*]Both authors contributed equally to this research.

Proceedings of Recent Advances in Natural Language Processing, pages 463–474
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_053

cific genre have certain dominant themes in common based on stylometric features, and if so, what meaning we can attribute to those themes as it relates to a book's success. To attain this objective, we investigated ways to enable using the entire book's content for stylistic modeling using frequencies of lexical production rules of CFGs for each novel (see Figure 1) followed by semantic word association of those rules to *Roget's* Categories and Themes for better interpretation. In this work, we followed widely used feature reduction technique for SVM modeling to reduce the large lexical feature space for lengthy novels and experiment with different techniques using POS, Unigram, WordNet, and lexical production rules. In this article we present the following contributions:

- We built the largest dataset containing a total of 17,962 books. We included books from 4 additional genres and reclassified 2 of the genres included in (Ashok et al., 2013) as follows: Mystery→Detective; Love→Romance.
- We introduced our feature reduction methods to greatly improve prediction performance with our best model achieving 94% accuracy for success prediction.
- We mapped both WordNet's semantic word relations and context free grammar rules to a set of themes, as defined in *Roget's Thesaurus*. With these mappings, we discovered the themes that successful books of a given genre prioritize.

2 Related Works

Roget's Thesaurus is a widely used English-language thesaurus. A British lexicographer, Peter Mark Roget (1779–1869), created the thesaurus in 1805. The first version of the thesaurus comprised of nearly 15,000 words and was released to the public on 29 April 1852 (Roget and Roget, 1886). Since then each successive edition was improved with more words, with the most recent edition containing more than 100,000 words. In previous work, Jarmasz and Szpakowicz (2004) showed that *Roget's* is an excellent resource for measuring semantic similarity and *Roget's* word clusters have higher correlation than many other prominent word groups e.g., Wordnet (Miller, 1998; Jarmasz, 2012).

Syntactic features, such as CFG productions have been found to be very effective in different NLP tasks. Raghavan et al. (2010) used CFGs for authorship attribution achieving very high accuracy such as 96%. Rayson et al. (2002) presented systematic analyses based on lexical and syntactic features for genre detection of a literary works showing that novels involve more use of verbs and adverbs. On the other hand, Douglas and Broussard (2000) showed that informative writing tend to use nouns, prepositions, determiners and coordinating conjunctions more. CFGs were also used in several other works, such as gender attribution by tracing stylometric evidence by (Sarawgi et al., 2011), and native language detection by exploiting parse structures (Wang and Zong, 2011).

In the earlier work, Ashok et al. (2013) used stylistic approaches, such as unigram, bigram, part-of-speech distribution, grammatical rules, constituents, sentiment, and connotation as features and used Liblinear SVM (Fan et al., 2008) for the book success classification task. They used books from 8 genres, and they were able to achieve an average accuracy of 73.50% across all genres. Maharjan et al. (2017) used a set of hand-crafted features in combination with a recurrent neural network and generated feature representation to predict success. They obtained an average F1-score of 73.50% for 8 genres. In a more recent work by Maharjan et al. (2018a), they used the flow of emotion throughout a book for success prediction and obtained an F1-score of 69%.

In this paper, we used widely used feature reduction technique for SVM modeling. Guyon et al. (2002) used SVM weights for assigning ranks in the feature selection process. They verified that the top-ranked genes found by SVM have biological relevance to cancer and the SVM classifier with SVM selected features worked better than other classifiers in determining the relevant features along with the classification task.

3 Dataset Construction

3.1 Original Dataset

The original dataset from Ashok et al. (2013) is quite small as it only includes the first 1,000 sentences from 800 books split into 8 different genres, which are further split into successful and unsuccessful classes, each having 50 books. Additionally, many of the files included have less than 1,000 sentences, or contain automatically generated text from Project Gutenberg instead of the text from the proper novel. Finally, the books included are prelabeled with their successful/unsuccessful class

where download counts are absent, which limits further testing. Considering these issues, we decided to build upon (Ashok et al., 2013) by creating a cleaner and more complete dataset. Additionally, we present multiple models that are both more accurate and more general than the best performing model in (Ashok et al., 2013), unigram. From these models, we discovered more interesting and revealing qualities that separate successful from non-successful books.

3.2 New Dataset

We downloaded and used 17,962 English novels from Project Gutenberg: an online catalog of over 60,000 books, which are available to download for free in various formats (Gutenberg). We filtered the 60k books as follows: a) only English books, and b) only fiction books. We used a bash script[1] to harvest the novels from Project Gutenberg according to the webmaster's guidelines[2].

After downloading the books, we used the NLTK API for data processing (Bird et al., 2009). For each book, we extracted the unigram and bigram frequencies, the part-of-speech (POS) tag using the Stanford CoreNLPParser frequencies, the lexical and non-lexical context free grammar production rules also using the Stanford CoreNLPParser, the *Roget's Thesaurus* Category frequencies, and the WordNet Synset frequencies (Roget, 1852; Princeton University, 2010; Zhu et al., 2013). Like the authors of (Maharjan et al., 2018a), we also extracted the NRC Emotional Lexicon features and the Linguistic Inquiry and Word Count (LIWC) features from each book (Mohammad and Turney, 2013; Pennebaker et al., 2015). These emotional word mappings are highly valuable for some tasks, but the resulting models were not effective in our tests, and therefore not presented in this article.

Like in (Ashok et al., 2013), we also used the download count of each book to define success. In addition to predicting success classification for books in 12 unique genres, we also tested prediction performance independent of genre across the entire dataset. In both settings, we found an upper (υ^+) and lower (υ^-) download count threshold for classifying books of that genre as "successful" (with approx. more than 60% download count) or "not successful" (with approx. less than 40% download count) ensuring a balanced dataset (Table 1). We further collected Goodreads rating of 7,541 books out of 17,962 books that we discuss at the end of this paper.

GENRE	# BOOKS	υ^-	υ^+
Adventure	917	28	46
Children	3278	27	35
Detective	285	41	74
Drama	785	45	62
Fantasy	382	76	81
Fiction	5369	22	38
Historical Fiction	961	32	50
Humor	1024	14	24
Poetry	1664	34	50
Romance Fiction	634	34	48
Science Fiction	1748	44	58
Short Stories	915	35	49
All	17,962	35	37

Table 1: # of novels per genre and download count thresholds for unsuccessful ($\leq \upsilon^-$) and successful ($\geq \upsilon^+$) classes.

4 Methodology

4.1 Linguistic Models

We utilized 12 linguistic models for our quantitative analysis. 6 of the models are our own implementation of models used in (Ashok et al., 2013). Our 6 additional models have not been used to make these types of qualitative conclusions until now. These models include WordNet (Princeton University, 2010), *Roget's Thesaurus* (Roget, 1852), two models that map WordNet to different levels of *Roget's Thesaurus*, and two models that map context free grammar rules to *Roget's Thesaurus*. Mapping examples are given in Table 2 and explained below.

Unigram: The frequency of unique words in text.

Part-of-Speech Distribution: The authors of Ashok et al. (2013) demonstrated the value of PoS tag distribution in success prediction, and Koppel et al. (2006) presented the relationship between PoS tagging and genre detection and authorship attribution. Therefore, we reevaluated the application of PoS tag distribution for success prediction.

Context Free Grammar Rule Distribution: We also reevaluate the analysis of CFG rule distribution as presented in (Ashok et al., 2013), and use the same four categories:

- Γ: lexical production rules (productions where the right-hand symbol (RHS) is a terminal symbol (word)).

[1] https://www.exratione.com/2014/11/how-to-politely-download-all-english-language-text-format-files-from-project-gutenberg/

[2] https://www.gutenberg.org/wiki/Gutenberg:Information_About_Robot_Access_to_our_Pages

MODEL	ORIGINAL FEATURE	ROGET CATEGORY	ROGET THEME
WordNet	blaze, glitter, sunny	light	Organic Matter
	animal, heartbeat, revive	life	
Γ^G	Nom→Adj→bad Nom→Adj→illegal	Nom→Adj→wrong	Nom→Adj→Moral
	Nom→Adj→lawful Nom→Adj→unconstitutional	Nom→Adj→legality	

Table 2: Mapping to Roget examples for WordNet and Γ^G. For each model, the ORIGINAL FEATURES are combined in the ROGET CATEGORY column, which in turn is combined in the ROGET THEME column.

- Γ^G: lexical production rules prepended with the grandparent node.
- γ: nonlexical production rules (productions where the RHS is a non-terminal symbol).
- γ^G: nonlexical production rules prepended with the grandparent node.

WordNet: WordNet is large lexical database of English words. The WordNet database groups nouns, verbs, adjectives, and adverbs into sets of cognitive synonyms called Synsets. Each Sysnet expresses a distinct concept and is represented by a single word. Since Sysnets represent conceptual synonyms, they are able to be linked through conceptual and semantic relationships (Princeton University, 2010). WordNet has a total of 117,659 Synsets, each represented by a single, unique word, and our model uses the frequencies of these Synsets in each book. Not only does WordNet fit our semantic relation analysis methodology, but it has been used for the relevant task of metaphor identification in (Mao et al., 2018).

***Roget's Thesaurus*:** A tree structured thesaurus with six root nodes, which we will refer to as *Roget* Classes or Classes for short. Each Class is divided in sections, which results in 23 total sections. These sections represent 23 unique concepts that are both general enough to encompass a wide range of ideas, but also specific enough to retain clear meaning. Therefore, we refer to these sections as Themes, and they are the critical piece to interpreting the results of class prediction. Themes are further divided into subsections, levels, etc. before terminating in 1,039 groups of synonyms, which we will refer to as Categories. The Categories are comprised of 56,769 total words, with about half appearing in multiple Categories (Roget, 1852). Our *Roget* model uses the frequencies of these Categories in each book. Furthermore, the authors of (Aman and Szpakowicz, 2008) demonstrated the possible applications of *Roget's Thesaurus* for emotion detection with natural language processing, and (Kennedy and Szpakowicz, 2010) used the thesaurus for the related process of text summarizing.

Mapping WordNet to *Roget*: Since *Roget's Thesaurus* has fewer synonym groups than WordNet (1,039 vs. 117,659), and those groups are hierarchically abstracted with each of the 1,039 *Roget* Categories belonging to one of the 23 *Roget* Themes, we mapped WordNet's Synsets to *Roget's Thesaurus* to discover more meaningful insights into the distinct characteristics of successful novels. We mapped WordNet to *Roget* Categories (WNRC), and then subsequently to *Roget* Themes (WNRT).

Mapping Lexical Production Rules to *Roget*: Since the RHS of lexical production rules are words, they can also be mapped to *Roget's Thesaurus*. Using the RHS of the lexical production rules for each book we derived Γ^G to *Roget* Categories (Γ^GRC) and subsequently to *Roget* Themes (Γ^GRT).

4.2 Implementation

We used the sci-kit learn implementation of LibLinear SVM with 5-fold cross validation for class prediction (Pedregosa et al., 2011; Fan et al., 2008). To tune the weighted linear SVM parameter C, we used the tool gridsearchCV (Pedregosa et al., 2011) and performed a search over the values ranging $1e(-4to3)$. Part-of-speech tag features are scaled with unit normalization, while all other features are scaled using tf-idf. We used two strategies for the class prediction task: predicting class by genre and predicting class independent of genre. We chose this model over neural models as it gives us better scope to interpret book success with hand crafted features. After the initial training and testing of each model, we employed an exhaustive feature reduction method, similar to our success labeling process, to maximize performance (see Figure 2).

For a given model, we start with the mean feature weight learned during training. We remove all features from the dataset with $|weight|$ less than the $|mean|$ feature weight. Next, we train and test the model on this reduced fea-

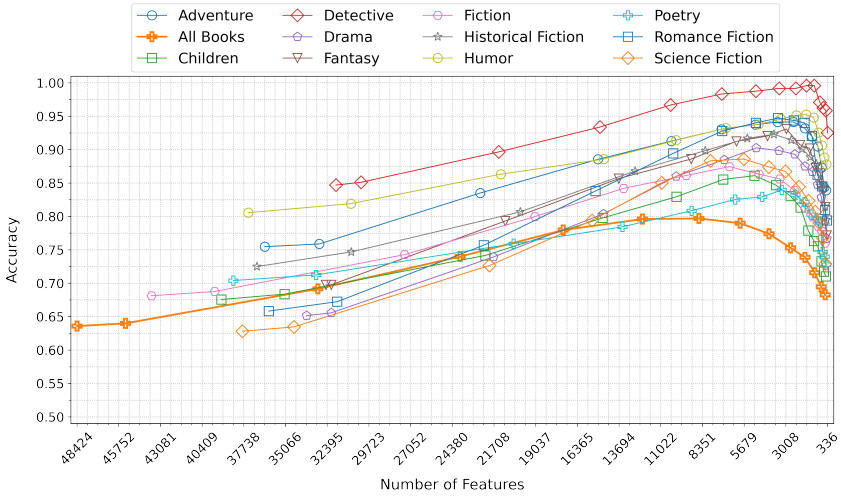

Figure 2: Feature reduction process: WordNet success prediction accuracy vs. number of features.

ture set and record the accuracy. For each subsequent test, starting at a step value of 0.25, we take only the features with weights greater than or equal to $Mean(OriginalWeights) + (StdDev(OriginalWeights) * Step)$. This process continues, increasing the step value by 0.25 after each iteration, until one of the following conditions is met: 100% classification accuracy is achieved, maximum accuracy is found (determined if multiple consecutive subsequent feature sets produce decreasing performance), or the number of features is reduced to less than 1% of the original number of features. Additionally, as explained previously, the processes of mapping WordNet to *Roget's Thesaurus* is a feature reduction technique in its own right. Table 3 illustrates the degree of feature reduction when WordNet and Γ^G are mapped.

5 Experimental Results

Using the original small dataset (Ashok et al., 2013), the prediction accuracy for each model by genre is presented in Table 8 at Appendix A1, and highlights another primary reason for increasing the size of the dataset. As each of the models was found to be achieving 100% accuracy in success prediction, we were convinced that those models were overfitting the dataset. This observation further motivated us to build a much larger dataset. For the classification task on newly constructed dataset, we applied the 5-fold cross validation method on all the genre specific datasets for evaluating each machine learning model. While for the feature reduction task, features were reduced using training weights from the training set, then tested on the test set. We had continued reducing the training set until the resulting features did not improve performance on the remaining test set.

The prediction accuracy for each model across all books, and each model by genre, both before and after feature reduction are shown in Table 4 and Table 5, respectively. As illustrated in both settings, the performance of nearly every model improved after we reduced the features with γ^G showing the largest improvement of an average of 24.3% when reduced by genre and WordNet improving the most by 16.1% when reduced independent of genre.

The best performing models are indicated in bold

MODEL	# OF FEATURES	# OF FEATURES^R
WordNet	31,833	1,670
WNRC	840	272
WNRT	21	9
Γ^G	24,302	596
Γ^GRC	995	184
Γ^GRT	21	13

Table 3: Number of features of ADVENTURE books before/after reduction for WordNet and Γ^G models.

MODEL	ACCURACY	ACCURACY^R
Unigram	61.6	61.6
POS	61.1	61.1
Γ	64.3	77.8
Γ^G	64.2	**80.1**
γ	61.1	68.9
γ^G	59.5	71.5
Roget	65.3	66.2
WordNet	63.6	79.7
WNRC	67.6	68.8
WNRT	61.5	61.5
Γ^GRC	66.9	67.8
Γ^GRT	60.8	60.8

Table 4: Accuracy of prediction model for ALL BOOKS of new dataset, with original and optimal reduced feature set (R).

MODEL	GENRE												AVG
	Ad	Ch	De	Dr	Fa	Fi	Hi	Hu	Po	Ro	Sc	Sh	
Uni	72.3	63.8	80.6	60.8	67.5	62.5	63.5	73.2	64.8	62.9	63.2	60.1	66.3
Uni^{R}	76.9	71.6	84.7	69.4	73.3	62.5	69.1	81.9	73.3	67.1	78.1	67.2	73.0
POS	66.2	63.8	72.7	63.1	67.0	66.0	70.5	77.6	70.3	60.1	63.2	68.5	67.5
POS^{R}	66.2	63.9	73.5	63.8	69.4	66.2	70.7	77.6	70.6	63.6	63.8	69.4	68.2
Γ	74.6	66.5	85.5	66.0	68.5	68.9	72.4	78.7	70.0	65.4	62.0	69.5	70.7
Γ^{R}	95.0	86.5	99.6	89.7	92.5	87.8	92.8	97.4	88.1	95.7	90.7	91.6	92.3
Γ^{G}	74.9	67.0	85.9	67.0	69.3	67.8	71.2	80.6	70.5	66.7	62.5	69.1	71.0
Γ^{G^R}	**96.9**	**88.5**	99.6	**94.1**	**96.2**	**88.3**	**93.5**	**98.5**	**89.8**	**97.0**	**92.8**	**92.5**	**94.0**
γ	69.9	59.7	78.9	62.3	65.9	60.9	66.1	73.5	65.7	62.1	56.5	66.8	65.7
γ^{R}	93.6	79.1	99.2	87.3	93.3	80.0	86.8	94.4	83.8	91.4	83.4	87.2	88.3
γ^{G}	68.5	59.8	82.3	61.5	65.6	63.8	67.4	74.9	66.3	65.0	58.4	68.4	66.8
γ^{G^R}	95.1	83.7	**100.0**	89.2	92.8	82.5	87.9	96.6	87.0	96.7	90.1	91.8	91.1
Roget	68.9	66.9	79.1	66.0	68.3	69.4	72.9	80.2	70.5	65.3	64.4	70.6	70.2
Roget^{R}	81.5	71.2	91.1	75.8	82.1	72.7	78.2	84.0	74.6	77.3	70.5	79.0	78.2
WN	75.5	67.6	84.7	65.2	69.8	68.1	72.5	80.6	70.4	65.8	62.8	69.0	71.0
WN^{R}	94.1	86.1	99.6	90.3	93.1	87.5	92.3	95.3	84.0	94.7	88.6	89.5	91.3
WNRC	79.7	72.2	93.4	74.8	79.3	72.0	81.4	86.6	72.8	82.9	70.0	77.4	78.5
WNRC^{R}	90.2	76.1	97.5	86.3	93.1	75.2	90.3	93.2	78.4	92.4	76.3	85.0	86.2
WNRT	66.5	61.9	80.6	64.2	67.4	65.8	70.3	76.7	70.2	66.2	62.6	71.0	68.6
WNRT^{R}	68.0	62.7	82.7	64.9	68.4	65.9	71.2	77.4	70.5	68.7	63.3	72.3	69.7
Γ^{G}RC	86.1	84.0	92.4	81.2	87.4	74.5	82.8	88.0	76.2	87.3	74.4	78.9	81.9
Γ^{G}RC^{R}	92.9	77.6	98.8	89.4	95.4	77.9	90.7	93.6	82.2	95.4	79.8	87.3	88.4
Γ^{G}RT	75.5	63.2	76.7	63.6	62.3	66.2	73.5	79.3	70.3	6701	62.3	71.0	69.2
Γ^{G}RT^{R}	76.0	63.5	77.5	64.9	65.3	66.4	74.0	79.3	70.5	68.1	62.8	71.0	69.9

Table 5: Accuracy (%) of classification results BY GENRE for new dataset, with/without feature reduction (R). Genre and model names are abbreviated; and best model performance is shown in bold.

in Table 4 and Table 5. When predicting novel success by genre and independent of genre, Γ^{G^R} shows the best results predicting a book's success class with an accuracy of 94.0% and 80.1%, respectively. Furthermore, when predicting success by genre, Γ^{G^R} achieves the highest accuracy for each genre except DETECTIVE. For DETECTIVE novels, γ^{G^R} outperforms all models with 100% accuracy.

Figure 2 illustrates the pattern of performance improvement that each model exhibits through the feature reduction process both by genre and independent of genre. As the number of features is reduced, the average accuracy for success prediction increases until the algorithm finds the best set of features and achieves peak performance. Then accuracy sharply drops as the feature set is reduced further. The fact that each model demonstrates such behavior validates the effectiveness of our feature reduction method.

6 Interpreting Book Success Prediction

While our reduced Γ^{G} and WordNet models display excellent performance in both test settings (by genre and independent of genre), the resulting feature sets are not self-explanatory. In other words, the respective lexical production rules and Synsets that the models deem most important do not necessarily highlight some interesting aspect of successful books. This is where *Roget's Thesaurus* proves most valuable.

We figured that if we looked up the *Roget* Theme of the RHS for each lexical production rule and the *Roget* Theme for each WordNet Synset we would find that the successful and unsuccessful books prioritize different Themes. With this hypothesis in mind, we mapped the reduced WordNet and reduced Γ^{G} models to new *Roget* models by first looking up the *Roget* Category of each Synset and RHS, respectively, from the reduced feature sets, and then summing the frequencies in each group of Sysnets/symbols. As we did with each previous model, we reduced the new WNRC and Γ^{G}RC models. From the WNRC^{R} and Γ^{G}RC^{R} models we mapped again, this time from *Roget* Categories to the 23 *Roget* Themes, which produced the WNRT and Γ^{G}RT models. Mapping examples are given in Table 2 and its outcome is detailed in the Figure 3.

We did not expect the performance of the WNRC and Γ^{G}RC models, since they were conceived strictly as intermediary maps between WordNet/Γ^{G} and *Roget* Themes. Γ^{G}RC produced the highest baseline results of all the models without any feature reduction used in our experiments with 81.9% average accuracy by genre. Furthermore, Γ^{G}RC^{R} accurately predicts success classification per genre

at an average rate of 88.4%. What's impressive about the accuracy of $\Gamma^{G}RC^{R}$, when compared to that of Γ^{G}, is the large difference in number of features used in each model as shown when predicting DETECTIVE novels in Table 3.

With these impressive results from $\Gamma^{G}RC^{R}$, we expected $\Gamma^{G}RT$ and $\Gamma^{G}RT^{R}$ to follow suit despite learning with a feature set of at most 13 features. However, this was not the case as $\Gamma^{G}RT^{R}$ predicts the success of a book by its genre with an average accuracy of only 69.9%. As previously stated, the motivation for the construction of WNRT and $\Gamma^{G}RT$ was strictly to find a common thread between successful novels in each genre. Therefore, the poor performance of the $WNRT^{R}$ and $\Gamma^{G}RT^{R}$ models does not undercut the reasoning behind its conception, and the high accuracy of $WordNet^{R}$, $WNRC^{R}$, $\Gamma^{G^{R}}$, and $\Gamma^{G}RC^{R}$ supports our claim that each is a general model that can reveal underlying characteristics of successful books.

Additionally, WNRT and $\Gamma^{G}RT$ do not improve performance after feature reduction when classifying independent of genre. This outcome also supports our original hypothesis as it shows that the models require each of the 23 *Roget* Themes in order to make the most accurate prediction. The lack of improvement in $WNRT^{R}$ and $\Gamma^{G}RT^{R}$ when predicting success class independent of genre also demonstrates the relationship between a novel's genre and its prioritization of certain Themes.

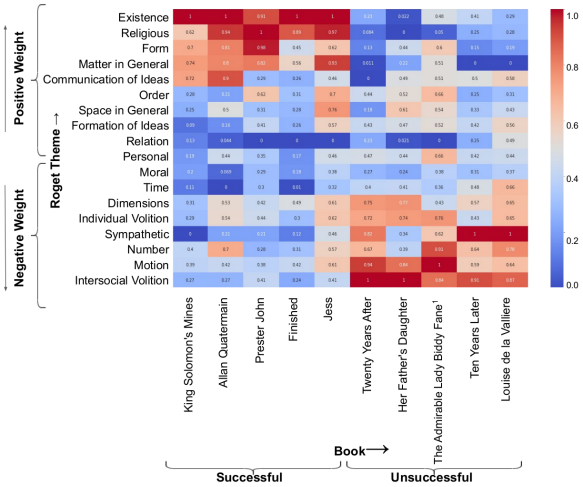

Figure 3: This heatmap presents how the mapping of Γ^{G} to RT helps to interpret success of ADVENTURE books. The plot presents both $+ve/-ve$ *Roget* Themes on the row, and successful/unsuccessful books on the column. Each cell represents the relative frequency of observing Γ^{G} in an RT. We observe that authors of successful books used certain CFGs that result in higher frequency in $+veRT$ cells, while the unsuccessful books have higher frequency in the $-veRt$ cells.

6.1 Successful Categories and Themes for a Genre

Figure 4 illustrates the top 30 discriminative positive and negative *Roget* Categories based on the model weights for CHILDREN's book success prediction. Greener Categories are positively weighted for success while redder Categories are negatively weighted. Specifically, we see positive weights for Themes of "Formation of Ideas" and "Related to Space." These Themes align with what readers should expect from CHILDREN's stories: developing new ideas (Formation of Ideas) as a character grows and has new experiences in the physical world around them (Related to Space). It's interesting however that "Communication of Ideas" shows negative weight for this genre. This suggests that CHILDREN's stories are more concerned with how a person grows and develops their own ideas, rather than how they communicate them. This pattern of the prioritization of expected Themes holds true across all genres, with few exceptions. Therefore, we can conclude that lexical choices focusing on Themes that conform to genre norms produce more successful novels.

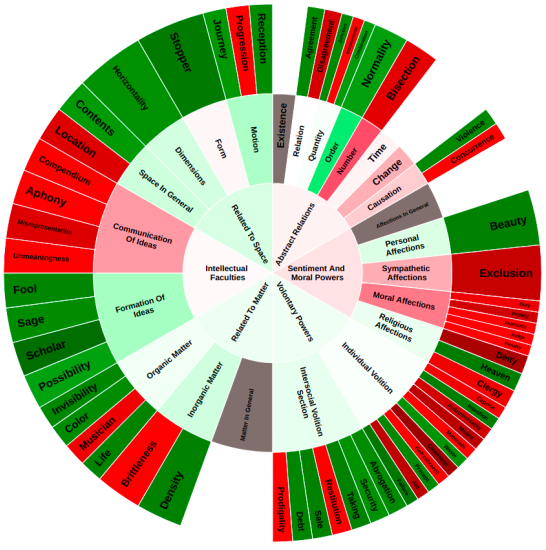

Figure 4: This sunburst presents a comprehensive review of the most discriminative *Roget* Category/Classes based on the classification model weight for a single genre, CHILDREN. We have considered top 30 discriminative features for both successful (Green) and unsuccessful books (Red).

THEME	WORDS	
	Successful	Unsuccessful
Affections	enthusiastic, lively, tenderness	inactive, sluggish, dull
Communication of Ideas	secret, untruth, language	school, grammar, taciturnity
Formation of Ideas	incredulity, impossibility, curiosity	dissent, sanity, memory
Moral	gluttony, impurity, selfishness	punishment, virtue, duty
Personal	expecting, blemish, hopelessness	aggravation, dejection, dullness

Table 6: Top 5 most important Themes for classifying CHILDREN novels and corresponding most predictive successful/unsuccessful thematic words

6.2 Thematic Analysis Based on Lexical Choices

After mapping the resulting feature weights of our WordNet^{R} and Γ^{G^R} models to *Roget* Themes, we were able to highlight the most important Themes when classifying the success of a novel given its genre. Table 6 gives the most important themes in predicting the success of CHILDREN'S novels and the successful and unsuccessful semantic word groups within those themes. These results clearly identify words associated with "school" and "grammar" as key contributors to unsuccessful CHILDREN'S novels, while words like "secret," "enthusiastic," and "selfishness" contribute to successful CHILDREN'S novels.

The indicated Themes align with intuitive expectations for CHILDREN'S books, especially the presence of FORMATION OF IDEAS and MORAL. To verify these results, we looked at the most downloaded CHILDREN'S book, *Little Women*. We ranked each book in the CHILDREN'S genre according to the frequency of each prioritized Theme listed in Table 7. Then, we looked to see where *Little Women* ranked for each of the Themes. *Little Women*'s use of the top Themes matches up as expected, as it ranks in the top three for four of the five most important Themes, and eighth for the fifth as shown in Table 7. The opposite is true for the least downloaded books, which all rank at the bottom for use of the most important Themes.

Our Thematic observations hold true for each genre, but there is not one Theme shared by all 12 genres. This adheres to the observation we made about WNRT and Γ^{G}RT and each model's lack of improvement after feature reduction for predicting success across all books independent of genre.

7 Experiments with Goodreads Rating

The discoveries made in our research are just the beginning of what can be done with our dataset. In addition to the data utilized for this project, we also extracted Goodreads Rating[3] as proposed in

[3]https://www.goodreads.com/

THEME	RANK
Communication of Ideas	2
Formation of Ideas	2
Personal	2
Moral	3
Affections	8

Table 7: Ranking the use of the most important CHILDREN'S themes for #1 downloaded CHILDREN'S book, *Little Women* relative to other CHILDREN'S books in the dataset

(Maharjan et al., 2019). We could collect the rating for 7,541 books from a total of 17,962 books scraped from Project Gutenberg, where each book has been rated by at least 5 readers. We labeled all the books having average rating $>= 3.5$ as successful, and < 3.5 as unsuccessful (presented in Table 10 at Appendix A2). In Appendix A1, Table 9 demonstrates the performances of previous and our models, respectively. When predicting novel success by genre, Γ^{G} shows the best results predicting a book's success class with an average weighted F1-score of 92.2% outperforming previous state-of-the-art results(75% for the genre attention with RNN method (Maharjan et al., 2018b)) as well. This result validates the applicability of our proposed model for book success prediction.

8 Conclusion

We created the largest dataset for evaluating book success, and presented a novel study of how context free grammar rules and semantic word association of influence a book's success. Our empirical results demonstrate that our large dataset combined with our feature reduction technique can predict a book's success with better accuracy than the current state-of-the-art methods. The analysis performed in this project shows the relationship between thematic word groups and a book's popularity, with our best model that uses context free grammar lexical production rules (Γ^{G^R}) achieving a prediction accuracy of 94.0%. Finally, we illustrated that readers expect certain themes to be prioritized over others based on a book's genre, and the proper use of those themes directly contributes to a book's popularity.

References

Saima Aman and Stan Szpakowicz. 2008. Using roget's thesaurus for fine-grained emotion recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1764, Seattle, Washington, USA. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Dan Douglas and KATHLEEN M. Broussard. 2000. Longman grammar of spoken and written english. *TESOL Quarterly*, 34(4):787–788.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Project Gutenberg. Project gutenberg. (n.d.).

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422.

Mario Jarmasz. 2012. Roget's thesaurus as a lexical resource for natural language processing. *arXiv preprint arXiv:1204.0140*.

Mario Jarmasz and Stan Szpakowicz. 2004. Roget's thesaurus and semantic similarity. *Recent Advances in Natural Language Processing III: Selected Papers from RANLP*, 2003:111.

Alistair Kennedy and Stan Szpakowicz. 2010. Evaluation of a sentence ranker for text summarization based on roget's thesaurus. In *Text, Speech and Dialogue*, pages 101–108, Berlin, Heidelberg. Springer Berlin Heidelberg.

Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. 2006. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 659–660, New York, NY, USA. Association for Computing Machinery.

Suraj Maharjan, John Arevalo, Manuel Montes, Fabio A González, and Thamar Solorio. 2017. A multi-task approach to predict likability of books. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1217–1227.

Suraj Maharjan, Sudipta Kar, Manuel Montes-y Gómez, Fabio A Gonzalez, and Thamar Solorio. 2018a. Letting emotions flow: Success prediction by modeling the flow of emotions in books. *arXiv preprint arXiv:1805.09746*.

Suraj Maharjan, Deepthi Mave, Prasha Shrestha, Manuel Montes, Fabio A. González, and Thamar Solorio. 2019. Jointly learning author and annotated character n-gram embeddings: A case study in literary text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 684–692, Varna, Bulgaria. INCOMA Ltd.

Suraj Maharjan, Manuel Montes, Fabio A. González, and Thamar Solorio. 2018b. A genre-aware attention model to improve the likability prediction of books. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3381–3391, Brussels, Belgium. Association for Computational Linguistics.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and wordnet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James Pennebaker, Roger Booth, Ryan Boyd, and Martha Francis. 2015. Linguistic inquiry and word count: Liwc2015.

Princeton University. 2010. "about wordnet". Https://wordnet.princeton.edu/.

Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. 2010. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 conference short papers*, pages 38–42.

Paul Rayson, Andrew Wilson, and Geoffrey Leech. 2002. Grammatical word class variation within the british national corpus sampler. In *New frontiers of corpus research*, pages 295–306. Brill Rodopi.

Peter Mark Roget. 1852. *Roget's Thesaurus*. Project Gutenberg. Http://www.gutenberg.org/files/10681/10681-h/10681-h.htm.

P.M. Roget and J.L. Roget. 1886. *Thesaurus of English Words and Phrases: Classified and Arranged So as to Facilitate the Expression of Ideas and Assist in Literary Composition*. T. Y. Crowell.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proceedings of the fifteenth conference on computational natural language learning*, pages 78–86.

Zhiguo Wang and Chengqing Zong. 2011. Parse reranking based on higher-order lexical dependencies. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1251–1259.

Muhua Zhu, Yue Zhang, Wenliang Chen, Min Zhang, and Jingbo Zhu. 2013. Fast and accurate shift-reduce constituent parsing.

Appendix

A1 Results

Model	Genre								Avg
	Adventure	Fiction	Historical	Love	Mystery	Poetry	Sci-Fi	Short	
Unigram	76.0	71.0	64.0	66.7	64.0	58.4	62.0	66.0	66.0
Unigram^{R}	92.7	86.0	73.0	84.9	89.0	79.4	88.0	83.0	84.5
Bigram	69.0	74.0	66.0	73.8	69.0	68.8	56.0	71.0	68.5
Bigram^{R}	93.8	92.0	88.0	86.9	92.0	88.6	85.0	93.0	89.9
POS	58.5	66.0	65.0	62.6	50.0	66.7	57.0	76.0	62.7
POS^{R}	71.9	73.0	69.0	69.6	61.0	71.9	61.0	76.0	69.2
Γ	64.6	65.0	58.0	74.7	67.0	65.7	52.0	70.0	64.6
Γ^{R}	**100.0**	99.0	**98.0**	**100.0**	**100.0**	95.8	94.0	98.0	98.1
Γ^{G}	65.7	60.0	63.0	62.7	64.0	67.7	47.0	66.0	62.0
$\Gamma^{G^{R}}$	**100.0**	**100.0**	**98.0**	**100.0**	**100.0**	**98.9**	99.0	**100.0**	**99.5**
γ	56.5	56.0	45.0	62.6	53.0	55.3	52.0	58.0	54.8
γ^{R}	**100.0**	**100.0**	97.0	**100.0**	**100.0**	97.9	98.0	96.0	98.6
γ^{G}	56.5	61.0	52.0	57.6	53.0	54.2	54.0	56.0	55.5
$\gamma^{G^{R}}$	**100.0**	**100.0**	**98.0**	**100.0**	**100.0**	97.9	**100.0**	96.0	99.0
WordNet	74.9	69.0	58.0	75.7	66.0	72.8	50.0	74.0	67.6
WordNet^{R}	**100.0**	**100.0**	97.0	**100.0**	**100.0**	96.8	96.0	98.0	98.5

Table 8: Accuracy (%) of classification results By Genre for (Ashok et al., 2013) dataset, with/without feature reduction (R) (*best performance in bold*)

Model (Reduced)	Genre												Avg
	Ad	Ch	De	Dr	Fa	Fi	Hi	Hu	Po	Ro	Sc	Sh	(W. F1)
Γ^{G} Ashok et al. (2013)	-	-	-	-	-	-	-	-	-	-	-	-	**73.5**
Genre Attention with RNN Maharjan et al. (2018b)	-	-	-	-	-	-	-	-	-	-	-	-	**75.1**
Γ	94.0	89.0	95.3	88.6	93.0	85.3	90.6	85.7	86.3	93.6	**89.3**	90.9	90.1
Γ^{G}	**94.6**	**90.5**	98.0	91.5	95.2	90.5	**91.1**	90.5	86.1	**96.1**	88.7	**93.9**	**92.2**
γ	90.1	82.6	95.2	89.8	91.1	79.1	85.0	87.7	88.2	92.3	84.1	86.4	87.6
γ^{G}	97.5	87.4	**99.4**	91.9	**96.9**	85.1	90.3	88.2	**86.4**	92.7	88.9	91.3	91.3
Roget	79.3	70.3	85.0	72.4	73.0	64.5	66.3	74.9	69.4	70.5	66.0	74.5	72.2
WN	94.0	85.9	95.3	88.9	93.0	**86.9**	**91.1**	80.9	81.2	95.3	84.1	87.1	88.6
WNRC	91.5	77.5	95.9	86.3	93.4	76.2	88.8	92.6	82.1	90.7	80.5	86.4	86.8
WNRT	68.7	60.1	80.3	68.9	75.4	56.5	65.6	70.1	62.6	68.3	58.2	62.4	66.4
ΓRC	91.6	76.6	96.7	89.2	93.8	75.7	87.5	89.4	82.4	92.0	83.8	89.0	87.3
ΓRT	64.1	59.0	81.3	64.6	66.7	55.9	67.3	70.2	67.9	58.6	55.6	70.8	65.2
Γ^{G}RC	96.2	80.4	99.3	**92.1**	96.3	79.2	90.2	**95.9**	84.9	94.6	81.5	87.8	89.9
Γ^{G}RT	74.7	57.1	82.3	67.0	63.8	57.9	66.8	67.3	56.9	66.9	59.9	62.4	65.3

Table 9: Average weighted F1-score for book success prediction using *Goodreads* rating. Part 1 of this table contains highest results of previous studies. Part 2 presents the results from the experiments with reduced feature set described in this article. Genre and model names abbreviated and the best performance is shown in bold font.

A2 Goodreads Dataset

GENRE	BOOKS	GR (SB)	GR (UB)	GRC
Adventure	917	285	97	383
Children	3,278	929	331	1,260
Detective	285	116	68	184
Drama	785	263	153	416
Fantasy	382	189	53	242
Fiction	5,369	1,461	722	2,183
Hist. Fiction	961	391	115	506
Humor	1,024	104	61	165
Poetry	1,664	441	140	581
Roma. Fiction	634	210	103	313
Sci. Fiction	1,748	388	581	969
Sho. Stories	915	214	125	339
Total	17,962	4,992	2,549	7,541

Table 10: This table presents the number of book ratings we collected from the Goodreads website for 12 genres. Here, GR stands for Goodreads, while SB, UB and GRC stands for successful books unsuccessful books and Goodreads count, respectively. We could collect a total of 7,541 book ratings from Goodreads, as opposed to the total 17,962 downloaded books from the Project Gutenberg website.