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

Anonymous EACL submission

Abstract

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. 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 18,000 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 then mapped context free grammar rules and WordNet's semantic word relations 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. In other words, if you want to write a bad children's book, write about keeping quiet in school.

1 Introduction

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.

Furthermore, the motivation for this work was to improve upon the state-of-the-art in book success prediction as presented by Ashok et al. We attempted to reproduce the results of (Ashok et al., 2013) using the same dataset[1]. In addition to implementing and testing the models used for success prediction in (Ashok et al., 2013), we also implemented and tested four new models on the dataset. Finally, we also incorporated the feature reduction techniques detailed in the **Methodology** section of this paper.

Previous Work

The authors of (Ashok et al., 2013) were the first to use statistical stylometry to predict the success of a novel based only on the contents of its first 1,000 sentences. They used stylistic approaches, such as uni-gram, bi-gram, distribution of the parts-of-speech, context free grammar rules, constituent tags, and sentiment and connotation values as features with a Linear SVM (Fan et al., 2008) for the classification task. The authors used books from 8 total genres, and their best performing model was able to achieve an average accuracy of 73.5% across all genres.

In (Maharjan et al., 2017), Maharajan et al. used a set of hand crafted features in combination with a recurrent neural network and generated feature representation to predict the likelihood of novel success. The best model presented by the au-

[1] Ashok et al. provided their dataset, however they were not able to provide their code.

thors of (Maharjan et al., 2017) was able to predict a novel's success with an average accuracy of 73.5% across 8 genres. They also performed several experiments, including using all the features used in (Ashok et al., 2013), sentiment concepts (Cambria et al., 2018), different readability metrics, doc2vec (Le and Mikolov, 2014) representation of a book, and unaligned word2vec (Mikolov et al., 2015) models of a book.

More recently, Maharajan et al. used the flow of the emotions across books for success prediction and obtained an F1-score of 69% (Maharjan et al., 2018). They divided the book into chunks, counted the frequency of emotional associations for each word using the NRC emotion lexicon (Mohammad and Turney, 2013), and then employed a recurrent neural network with an attention mechanism to predict both the genre and the success.

Contributions

In our attempt to reproduce the results presented in (Ashok et al., 2013), we discovered various issues with the dataset[2] including, but not limited to, its size, contents, and uniformity. Nonetheless, the results of our reproduction are summarized and compared to the original results in (Ashok et al., 2013) in Table 1. The full results of our attempted reproduction can be found in Supplementary Table 1. This original dataset 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 prelabled with their successful/unsuccessful class, 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), uni-gram. From these models, we discovered more interesting and revealing qualities that separate successful from non-successful books.

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.

Table 1: Average success prediction accuracy by genre per model presented in (Ashok et al., 2013) (α) vs. our attempted reproduction (β)

MODEL	α	β
Unigram	70.3	66.0
Bigram	67.2	68.5
POS	64.5	62.7
Γ	71.6	64.6
Γ^G	73.5	62.0
γ	66.3	54.8
γ^G	69.2	55.5

2 Dataset Construction

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 used a bash script[3] to harvest the novels from Project Gutenberg according to the webmaster's guidelines[4].

After downloading the books, we used the NLTK API for data processing (Bird et al., 2009). For each book, we extracted the uni-gram and bi-gram frequencies, the part-of-speech (POS) tag using the Stanford CoreNLPParser frequencies, the lexical and non-lexical context free grammar production

[2] https://www3.cs.stonybrook.edu/~songfeng/success/

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

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

Table 2: # of novels per genre and download count thresholds for unsuccessful ($\leq \upsilon^-$) and successful ($\geq \upsilon^+$) classes for the WordNet model

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

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., 2018), 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 a measurement of success. In addition to predicting success classification for books split into 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 threshold for classifying books of that genre as "successful" or "not successful."

We performed an exhaustive search to find these thresholds by setting the class labels according to an incrementally widening download margin, and then training and testing each model at each increment. Starting at the median number of downloads, we label all books with downloads above the median as successful and all books below the median as unsuccessful. We train and test the given model with these labels and record the accuracy. Next, we move the upper bound to the first value greater than the median downloads and the lower bound to the first value less than the median downloads, assign new class labels to the remaining books included in the wider margin, and then train and test the model again with these new labels. This continues until there are less than 100 books in either class, or until all margins that produce classes of equal length are found. This method ensures that the dataset is balanced between successful and unsuccessful books for training a robust model. Each model achieved its best performance with a different class label margin, which we then used for the remainder of classification testing as shown for the WordNet model in Table 2.

3 Methodology

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 3 and further explained below.

- **Uni-grams:** The frequency of unique words in the 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)).
  - 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^{G}RC) and subsequently to Roget Themes (Γ^{G}RT).

Implementation

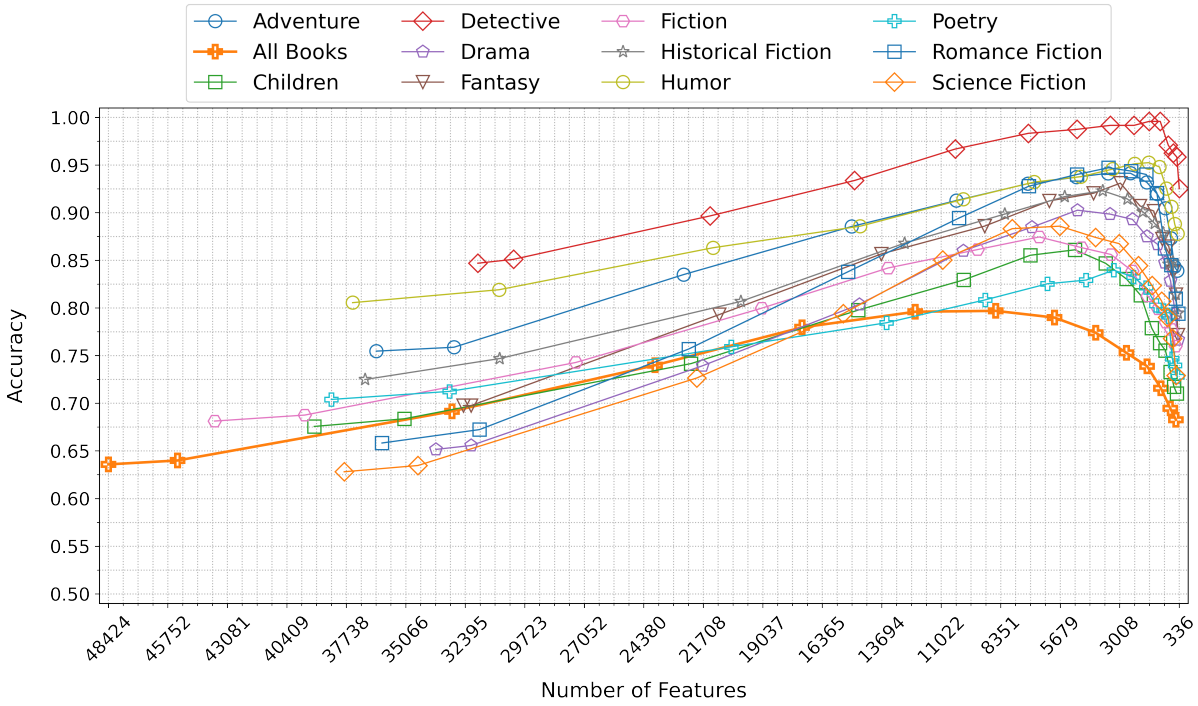
We used the sci-kit learn implementation of LibLinear SVM with 5-fold cross validation for class prediction (Pedregosa et al., 2011; Fan et al., 2008). Part-of-speech tag features are scaled with unit normalization, while all other features are scaled using tf-idf. We used two strategies for the class prediction task: predicting class by genre and predicting class independent of genre.

After the initial training and testing of each model, we employed an exhaustive feature reduc-

Table 3: 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.

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	

Figure 1: Feature reduction process: WordNet success prediction accuracy vs. number of features



tion method, similar to our success labeling process, to maximize performance. 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 feature 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.

Table 4: Number of features of HISTORICAL FICTION books before/after reduction for WordNet, WNRC, and WNRT

MODEL	# OF FEATURES	# OF FEATURES^{R}
WordNet	36,343	3,776
WNRC	818	354
WNRT	23	12

Additionally, as explained previously, the processes of mapping WordNet to *Roget's Thesaurus* is a feature reduction technique in its own right. Table 4 illustrates the extreme degree of feature reduction when WordNet and Γ^{G} are mapped.

4 Experimental Results

Classification Results - Original Dataset

The prediction accuracy for each model by genre is presented in Supplementary Table 1, and highlights another primary reason for increasing the

Table 5: Accuracy (%) of classification results for ALL BOOKS for the new dataset, before/after feature reduction

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^{G}RC	66.9	67.8
Γ^{G}RT	60.8	60.8

Table 6: Accuracy (%) of classification results BY GENRE for new dataset, with/without feature reduction (R) (genre and model names abbreviated; *best performance in bold*)

MODEL	GENRE												AVG
	A	C	De	Dr	Fa	Fi	Hi	Hu	P	R	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

size of the dataset. With the following models all achieving 100% accuracy in success prediction, we were convinced that those models were overfitting the dataset: Γ^{R} for ADVENTURE, LOVE, and MYSTERY books; Γ^{G^R} for ADVENTURE, FICTION, LOVE, MYSTERY, and SHORT STORIES books; γ^{R} for ADVENTURE, FICTION, LOVE, and MYSTERY books; γ^{G^R} for ADVENTURE, FICTION, LOVE, MYSTERY, and SCIENCE FICTION books; and WordNet^{R} for ADVENTURE, FICTION, LOVE, and MYSTERY books. This observation further motivated us to build a much larger dataset.

Classification Results - New Dataset

The prediction accuracy for each model across all books, and each model by genre, both before and after feature reduction are shown in Table 5 and Table 6, respectively[5]. 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 in Table 5 and Table 6. 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% prediction accuracy.

Figure 1 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

[5] All mapped to Roget models are mapped from their reduced base model (i.e. WordNet^{R})

Table 7: Top 5 most important Themes for classifying CHILDREN novels and corresponding most predictive successful/unsuccessful thematic words

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

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.

Interpreting Book Success

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, expected or otherwise. This is where *Roget's Thesaurus* proves most valuable.

Table 8: Number of features of DETECTIVE books before/after reduction for Γ^{G}, Γ^{G}RC, and Γ^{G}RT

MODEL	# OF FEATURES	# OF FEATURES^{R}
Γ^{G}	24,302	596
Γ^{G}RC	995	184
Γ^{G}RT	21	13

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. Then, 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 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 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 Γ^{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 8.

With these impressive results from Γ^{G}RC^{R}, we expected Γ^{G}RT and Γ^{G}RT^{R} to follow suit despite learning with a feature set of at most 23 features. However, this was not the case as Γ^{G}RT^{R} predicts the success of a book by its genre with an average accuracy of only 69.9% while learning from an average of only 14 features. As previously state, the motivation for the construction of WNRT and Γ^{G}RT was strictly to find a common thread between successful novels in each genre. Therefore, the poor performance of the WNRT^{R} and Γ^{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 Γ^{G}RC^{R} supports our claim that each is a general model that can reveal underlying characteristics of successful books.

Additionally, WNRT and Γ^{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 Γ^{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.

Successful Lexical Choices

After mapping the resulting feature weights of our WordNet^{R} and $\Gamma^{G^{R}}$ models to Roget Themes, we were able to highlight the most important Themes when classifying the success of a novel given its genre. Table 7 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.

Table 9: 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

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

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 9. 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 9. 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.

5 Future Work

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 bi-gram features from each book, but were not able to use these features in this work. In future work, we will continue to explore the impact of these features in addition to semantic word associations on book success.

We believe that we could achieve better results through the use of a different success surrogate metric. The scale of Project Gutenberg's catalog does not correspond to the website's popularity. As we continue this work we will therefore include ratings from popular sites such as Goodreads.com, Amazon, etc. to improve the our success labeling method. Another shortcoming of Project Gutenberg's catalog is that nearly all of the included books were published before 1924. So in order to further demonstrate the value of our results, we will incorporate contemporary novels into the dataset.

Finally, in (Ashok et al., 2013), Ashok et al. applied their models to movie scripts using ratings from IMDb.com to measure success. As we continue to explore the impact of semantic word association on book success, we will also apply our models to film/TV scripts.

6 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 ($\Gamma^{G^{R}}$) achieving a prediction accuracy of 94%. 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.

Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok. 2018. Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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.).

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.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

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. 2018. Letting emotions flow: Success prediction by modeling the flow of emotions in books. *arXiv preprint arXiv:1805.09746*.

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.

Tomas Mikolov, Kai Chen, Gregory S Corrado, and Jeffrey A Dean. 2015. Computing numeric representations of words in a high-dimensional space. US Patent 9,037,464.

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/.

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

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