A Study on Using Semantic Word Associations to Predict the Success of a Novel

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

[1]Department of Computer Science & Engineering, Shahjalal University of Science & Tech.
[2]Department of Computer & Information Science, Fordham University
[3]Google LLC
{syeda06,biddut12}@student.sust.edu,
hgorelick@fordham.edu, sabir.ismail01@gmail.com,
saiful-cse@sust.edu,mamin17@fordham.edu

Abstract

Many new books get published every year, and only a fraction of them become popular among the readers. So the prediction of a book success can be a very useful parameter for publishers to make a reliable decision. This article presents the study of semantic word associations using the word embedding of book content for a set of Roget's thesaurus concepts for book success prediction. In this work, we discuss the method to represent a book as a spectrum of concepts based on the association score between its content embedding and a global embedding (i.e. fastText) for a set of semantically linked word clusters. We show that the semantic word associations outperform the previous methods for book success prediction. In addition, we present that semantic word associations also provide better results than using features like the frequency of word groups in Roget's thesaurus, LIWC (a popular tool for linguistic inquiry and word count), NRC (word association emotion lexicon), and part of speech (PoS). Our study reports that concept associations based on Roget's Thesaurus using word embedding of individual novel resulted in the state-of-the-art performance of 0.89 average weighted F1-score for book success prediction. Finally, we present a set of dominant themes that contribute towards the popularity of a book for a specific genre.

1 Introduction

Every year a lot of literary fictions get published and only a few of them achieve the popularity. So it is very important to be able to predict the success of a book before the publisher commits a significant effort and resources for it. Many factors contribute to the success of a book. The story, plot, and character development, all have specific role in the popularity of a book. There are some other factors like the time when the book has been published, the author's reputation, the marketing strategy, etc that may also influence a book's popularity. In this paper, we only focus on understanding a set of concepts' associations extracted from the content of the book to predict its success.

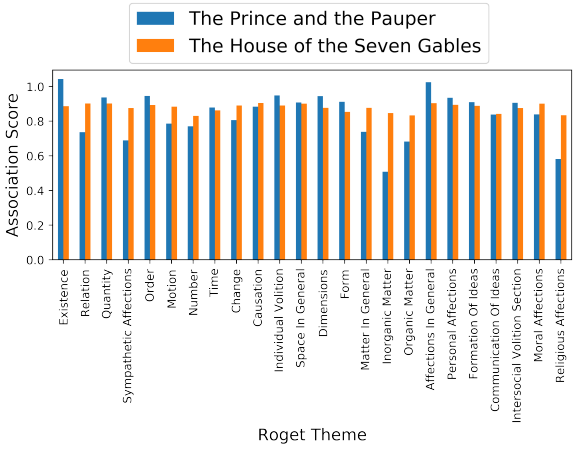

Figure 1: This figure represents average word embedding association scores for 24 themes as defined in the Roget's thesaurus. We observe that corresponding association scores for historical fiction books, such as the successful book *The Prince and the Pauper*, and the unsuccessful book *The House of the Seven Gables* are very different. The success of those books were defined using their corresponding *Goodreads-rating*.

According to the theory of word embedding, the vector representation of a word in the embedding space captures its semantic relationship with other words based on co-occurrence in the corpus. Kulkarni et al. (2015), and Hamilton et al. (2016a) developed methods for detecting the statistically significant linguistic change using word embedding. In the meantime, Caliskan et al. (2017) developed the concepts of word embedding association test (WEAT) to uncover the gender bias and ethnicity bias. Following these studies, Garg et al. (2018),

[*]Both authors contributed equally to this research.

Proceedings of the 10th Conference on Lexical and Computational Semantics, pages 38–51
August 5–6, 2021, Bangkok, Thailand (online)

and Jones et al. (2020) used 100 years of text data and demonstrated that word embedding can be used as a powerful tool to quantify historical trends and social change. For every time period, they warp the vector spaces into one unified coordinate system and construct a distance-based distributional time series for each word to track its linguistic displacement over time. Our idea is to use the associations of different semantically linked word groups or concepts in a book and investigate how its impact on book success prediction.

In this article, we study the efficacy of word associations to represent literature as a spectrum of individually organized concepts as a set of connoted words in the popular Roget's Thesaurus (Roget and Roget, 1886). We represent word association as the Euclidean distance between two words in the embedding space. To find the association of book content to a set of concepts, we compute the average Euclidean distance for each set of semantically linked word vectors of a book's normalized embedding space to the respective word representation in the global embedding space. The concept of word embedding normalization and the word association score has been used successfully in many recent research works for computing the gender associations (Jones et al., 2020).

In Figure 1, we show word associations of prominent themes for a successful book *The Prince and the Pauper* having *Goodreads-rating* > 3.5, and an unsuccessful book *The House of the Seven Gables* with a *Goodreads-rating* < 3.5. We observe that the average association score of each theme vary between these two books. We analyze the impact of these associations score for the success of each book, and obtain a set of dominant concepts that play an important role for a book success. In this paper, we include following research contributions:

- We developed necessary methods to represent a book as the spectrum of word associations for a set of semantically linked words.
- We present genre-wise book success prediction model using semantic word associations as features, and show that the model can achieve the best average weighted F1-score of 0.89.
- We derived a set of dominant features for each genre showing the impact of those features for interpreting the prediction of book success.

2 Related Work

In the earlier work, Ashok et al. (2013) used stylistic approaches, such as unigram, bigram, distribution of the part-of-speech, grammatical rules, constituents, sentiment, and connotation as features and used Liblinear SVM (Fan et al., 2008) for the classification task. They used books from total 8 genres, and they were able to achieve an average accuracy of 73.50% for all the genres.

van Cranenburgh and Koolen (2015) distinguished highly literary works from less literary works using textual features e.g. bigram. Vonnegut (1981); Reagan et al. (2016) worked on emotion along with the book for success prediction.

Maharjan et al. (2017) used a set of hand-crafted features in combination with recurrent neural network and generated feature representation to predict the success, and obtained an average accuracy of 73.50% for the 8 genres. They also performed several experiments, including using all the features from Ashok et al. (2013), sentiment concept (Cambria et al., 2018), different readability metrics, Doc2Vec (Le and Mikolov, 2014) representation of a book, and unaligned Word2Vec (Mikolov et al., 2015) model of the book.

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

Jarmasz and Szpakowicz (2004); Jarmasz (2012) showed that Roget's has turned out to be an excellent resource for measuring semantic similarity and the words in Roget's word clusters have higher correlation than many other prominent word groups e.g., Wordnet Miller (1998). 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

In this study, we use the dataset introduced by Maharjan et al. (2017), a publicly available dataset

Genre	Unsuccessful	Successful	Total
Detective Mystery	60	46	106
Drama	29	70	99
Fiction	30	81	111
Historical Fiction	16	65	81
Love Stories	20	60	80
Poetry	23	158	181
Science Fiction	48	39	87
Short Stories	123	135	258
Total	349	654	1,003

Table 1: The book dataset originally introduced by Maharjan et al. (2017) is used in this research work for success prediction. Each book in this dataset belongs to one of the eight genres. Here we have the most number of books from the Short Stories genre(258) and the least number of books from the Love Stories genre(80).

comprising of total 1,003 books. All of these books are downloaded from the Project Gutenberg[1]. Details of the dataset are given in Table 1. Each of these books are labeled as either successful (1) or unsuccessful (0). The definition of the success of a book is based on Goodreads[2] ratings. A book is considered successful if it had been rated by at least 10 Goodreads users and has a Goodreads rating ≥ 3.5 out of 5. In this corpus, there are 349 unsuccessful books and 654 successful books. After downloading the books we used the NLTK API for data processing (Bird et al., 2009). For each book, we extracted the part-of-speech (PoS) tag frequencies using the Stanford CoreNLPParser, the Roget's Thesaurus category frequencies (Roget and Roget, 1886; Manning et al., 2014).

Linguistic Models

We utilized four linguistic models for our quantitative analysis. Two of the models - PoS and NRC are our own implementation of models used in Ashok et al. (2013) and Maharjan et al. (2018a). Our two additional models have not been used to make these types of qualitative conclusions until now. The linguistic models used in our frequency and association analysis are described below.

PoS: Part of Speech or PoS is a category to which a word is assigned in accordance with its syntactic functions. PoS provides context and classification to words that helps with better understanding of the purpose of word choice. We used NLTK PoS tagger to label our tokens.

LIWC: Linguistic Inquiry and Word Count (Pennebaker et al., 2015) is a text analysis program that counts words in psychologically meaningful categories. We used 72 LIWC categories for our experiments.

NRC: The distribution of sentiments is one way of looking at books. We used ten categories from NRC (trust, fear, negative, sadness, anger, surprise, positive, disgust, joy, anticipation) to quantify shifts in sentiment across the book.

Roget's Thesaurus: It is composed of 6 primary classes and each class is composed of multiple themes. There are total 24 themes that are further divided into multiple concepts. We used 1,019 word categories from the Roget's Thesaurus for the book success prediction.

4 Methodology

In order to predict the success of a book, one of our major research questions was how we can represent a book properly. We explored a wide range of feature sets and performed multiple experiments in order to find the most suitable feature set that can represent the concept, emotion and writing style of a book. In this section, we discuss the relevant methods that we used for the study of book success prediction.

4.1 Frequency Distribution

We explore 4 different word frequency distributions, such as (1) Roget's Thesaurus, (2) LIWC, (3) NRC and (4) PoS as the feature sets for the book success prediction. We first experimented with frequency distribution of Roget word categories to predict the success of a book. To perform this task, we compute the unit normalized word frequency distribution for each book. Here, frequency is computed for word groups rather than individual words. If a word falls under multiple word group its frequency contributes to all of them. The frequency count of a word group is the summation of frequencies of all the underlying words in that group. And finally, we apply the classifier as discussed in the subsection 4.4 for the book success prediction using Roget's word group frequency distributions as a feature of individual book. We repeat the above steps for creating three other feature sets based on the word frequency distributions of LIWC, NRC, and PoS for each book.

4.2 Association Score

To represent a book as a vector of concept association score, we first create the word embedding vectors from the respective book's content. We

[1]https://www.gutenberg.org/
[2]https://www.goodreads.com

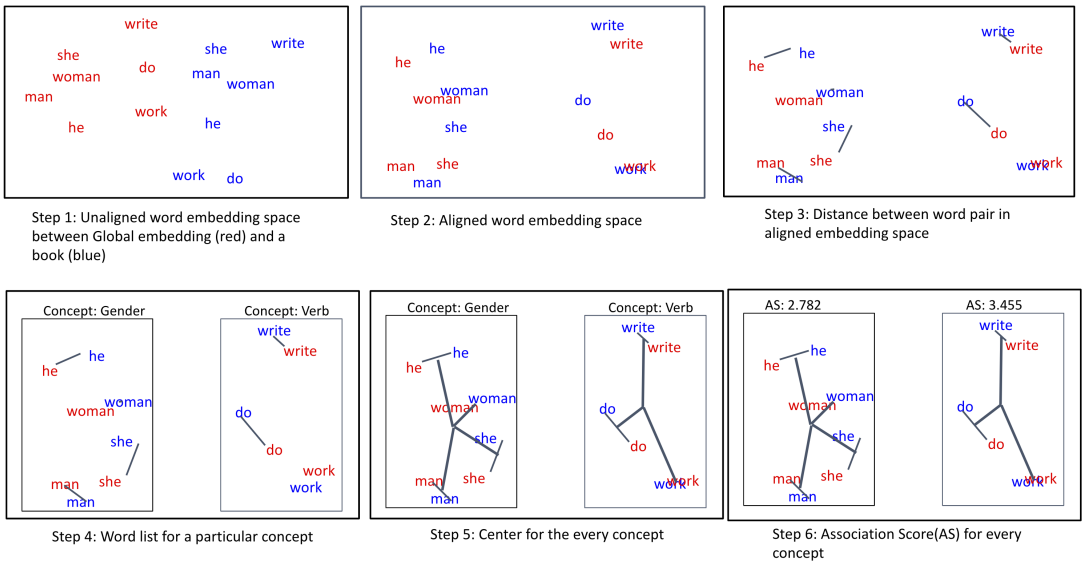

Figure 2: Steps in computing the concept association. At first, word vectors for the global embedding (fastText) and local embedding (individual book embedding) are aligned to a unified space (Steps 1 - 3). Then, for each word, we compute the Euclidean distance of its representative vector from the global and aligned local embedding. The Euclidean distance of all words of each concept are then averaged to calculate the association score (Steps 4 - 6).

then align each book embedding to a global embedding space so that each book can be analyzed with respect to a reference embedding space (Mikolov et al., 2018). To generate the word embedding of each book, we considered the fastText embedding generation methods (Bojanowski et al., 2017). On the contrary to Word2Vec and Glove, fastText treats each word in corpus like an atomic entity and generate a vector for each word. In fastText embedding, the vector representation for a word is created depending on its constituent character n-grams. This method generates better word embedding for rare words and out of vocabulary words.

To do the embedding space alignment, we use the methods described in the paper (Artetxe et al., 2018) including 4 other methods described in (Hamilton et al., 2016b; Kendall, 1989). Intuitively, we have two embedding space for each book, one is the original or local embedding of the book and the other is global fastText embedding. For every

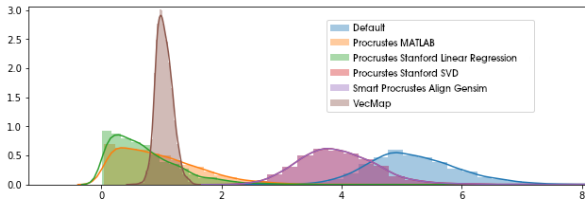

Figure 3: Distribution of the word association for Roget concept words using different alignments methods

word present in a book embedding, we calculate the Euclidean distance. The distribution of the distance using different alignment methods is shown in Figure 3 for the word embedding of 10 books. Ultimately, we use the method named VecMap (Artetxe et al., 2018) as it results in minimum distance after vector alignment.

To represent a book as a vector of concept association score, we first create the fastText word embedding vectors from the respective book's content. As a result, we obtain two individual embedding spaces, one for book and another for the global embedding space. We align the book embedding space to the global embedding space so that each book can be analyzed with respect to a reference embedding space (Refer to Figure 2: Steps 1 - 3). To find the concept association score, we compute the average Euclidean distance from the book's aligned embedding vectors to the global embedding vectors for each semantically linked word cluster. We depict the process in Figure 2 (Steps 4 - 6).

We use the wiki word embedding model (Bojanowski et al., 2017) as our global embedding space. It is trained on Wikipedia using fastText. For the compatibility of book embedding and global embedding, we use fastText to produce word embedding for each book individually. Each generated word vector is 300 dimensional. We use skip-gram as a training algorithm. We then tune the number of iterations over the book content (epochs) by

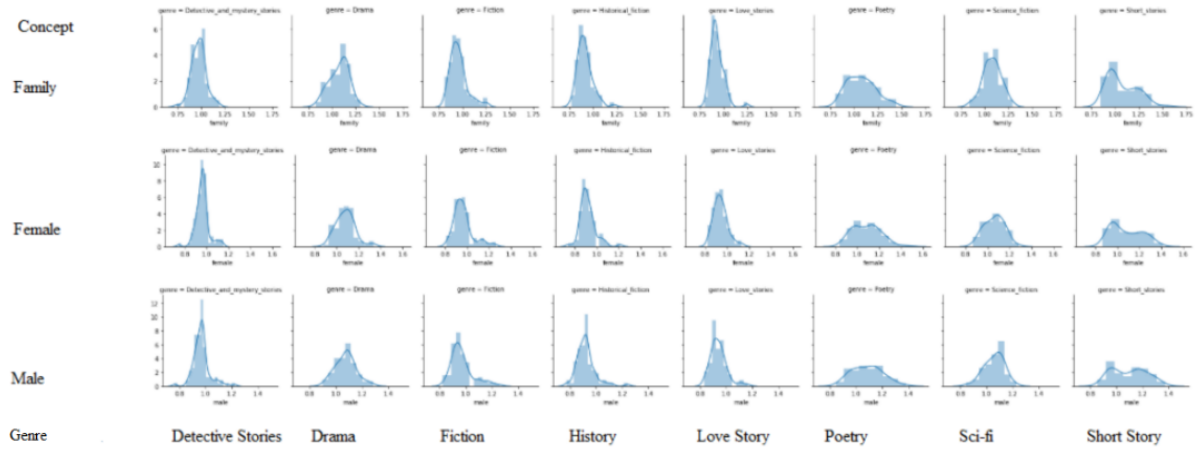

Figure 4: Association of different concepts with 8 genres. The x-axis is the mean association score of the words in a Roget concept, and the y-axis is the frequency observed for each book.

running 20 different experiments with a random selection of diverse values of epochs, and then select 50 as the epoch. To generate word embedding vectors for each book, we only consider those words that have a minimum word count 2.

Therefore, each book of the dataset is represented using a feature vector of length 1,019 following the word category definition in Roget's Thesaurus. Figure 4 shows the distribution of different Roget concept associations for 8 different genres. From these distributions, it is clear that different concepts have different impact on each genre. We also perform the *Kolmogorov-Smirnov Test* (kol, 2008) to check whether these distributions are different or not. In most of the cases, we find that a pair of the the distributions are significantly different from each other as per the statistical test. Finally, we apply the classifier described in subsection 4.4 on the set of association scores of each book for book success prediction task.

4.3 Feature Selection

The feature selection process selects a subset of features that can efficiently describe the input samples. As a result, this step eliminates the interdependent and irrelevant variables, reduce effects from noise, and finally improve classification performance. Among various feature selection methods, we use the filter method (John et al., 1994) to identify relevant features. In this method, all the features are ranked based on a score or weight that is used to denote the feature relevance. This list of features is optimized or shortened depending on a defined threshold to improve the model prediction. We set the limit of shortened and selected feature length as 50 to prevent the loss of important information about a book.

In our experiments, we use the weighted linear SVM as a classifier. To predict the class of any testing sample x, the decision function for this classifier is given below.

$$f(x) = sgn(w^T \phi(x) + b) \quad (1)$$

If $f(x) < 0$, the book is predicted as unsuccessful and if $f(x) > 0$ the book is predicted as successful. Here, feature weight vector w in Equation 1 is determined by training the linear SVM classifier. This weight vector w can be used to find out the relevance of each feature (Guyon et al., 2002) . The feature values $\phi(x)$ in Equation 1 can only be positive for the book success prediction using both frequency and association analysis as feature. So the larger the value $|w_i|$ is, the more it contributes for deciding the sign of the decision function. It is worth mentioning that linear SVM classifier with optimized feature set is intuitively an efficient process as both the tasks use the same decision model. Thus selection of decision boundary for SVM and selection of relevant features are tightly connected (Bron et al., 2015).

4.4 Model Evaluation

For our prediction task, we used weighted linear SVM (Fan et al., 2008) as a classifier with L2 regularization over training data. We used grid-search in order to tune regularization hyperparameter C for weighted linear SVM. 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 $1e(-4to3)$. Then the best value of C was used as a regularization parameter for the weighted linear SVM. To mitigate the overfitting problem, we used 5-fold cross-validation to measure our performance. Thus, our dataset was randomly split into 5 equal segments, and results were averaged over 5 trials. In each trial, the model was trained on 4 segments and tested on the last segment.

We present the algorithms for *Association Score Calculation*, *Feature Ranking Based on Linear SVM Weights*, and *Training and Prediction* in the *Appendix Algorithms 1-3*.

5 Results

5.1 Baseline Model

Prior works have been done on book success prediction using the dataset introduced in Maharjan et al. (2017). Among them, some of the best weighted F1-scores for the book success prediction tasks are 0.69 for Book2Vec (DBoW+DMM) (Maharjan et al., 2017), 0.67 for the Emotion Flow (Maharjan et al., 2018a), 0.71 for Annotated char-3gram(AC3) (Maharjan et al., 2019), and 0.75 for the genre attention with RNN method (Maharjan et al., 2018b) which achieved the state-of-the-art performance. We set the weighted F1 score of 0.75 as our baseline result and proceed to our experiments.

5.2 Book Success Using Word Group Frequency

Our first set of experiments were devised using PoS, NRC and LIWC feature sets having 10, 44, 72 features respectively. As we decided 50 as the lowest number of selected features in subsection 4.3, we did not apply the feature selection method for PoS and NRC categories. Table 2 shows that feature set using PoS and NRC word frequencies could obtain average weighted F1 scores of 0.65 and 0.67 respectively. After employing the feature selection method for LIWC, we obtained an average weighted F1 score of 0.69 which is a slight improvement over the previous two methods but it still fails to outperform the baseline result.

5.3 Book Success Using Roget's Word Group Frequency

For this modeling task, we started with the semantic word association scores of 1,019 Roget's thesaurus concepts as features. As discussed in the methodology section, we performed feature selection for optimized model performance. As a result, this method yielded a performance gain of 0.88 average weighted F1 score beating the baseline results by a large margin (Table 2). In order to investigate the interpretability of the results we obtained from Roget frequency, we dived deeper into the analysis and explored the discriminative features for classifying successful and unsuccessful books for different genres. The visualization we produced for "Detective and Mystery Stories" is placed in the *Appendix Figure 9*. Although we obtained a result that outperformed the state-of-the-art performance using this analysis, it fails to discover more meaningful insights than association analysis that we discussed in the following subsections 5.4 and 5.5.

5.4 Book Success Using Word Association

As all our previous experiments are based on frequency distribution of lexical features, they failed to capture the essential semantic features that have an enormous impact on book success. To deal with this problem, we performed an association analysis using Roget's word groups that were cataloged based on semantic meaning as discussed in subsection 4.2. The feature selection result for each genre are presented in the Figure 5. It can be observed that as we keep filtering out irrelevant features, the performance for book success prediction for each genre increased. But after reaching a certain level, further feature reduction caused a monotonous decline in performance as it discarded important features. The best result obtained using Roget Association is an average weighted F1 score of 0.89 which outperforms not only the baseline results but also the state-of-art result we obtained using Roget's word group frequency (Table 2).

As mentioned earlier, our modeling experiments were performed using the genre-wise 5-fold cross validation. To further identify any overfitting characteristics in the modeling we computed the area under precision-recall curve (AUC of PR-curve). As our dataset is not balanced, we used PR-curve to validate or interpret our result. In the *Appendix Figure 1*, we show genre-wise precision-recall plots, where we draw a combined precision-recall curve of 5-fold cross-validation. Most of the combined results are above an AUC of 0.90 except Detective and Short Stories having that slightly less than 0.90 AUC. This proves that our model performed very well in this imbalanced dataset.

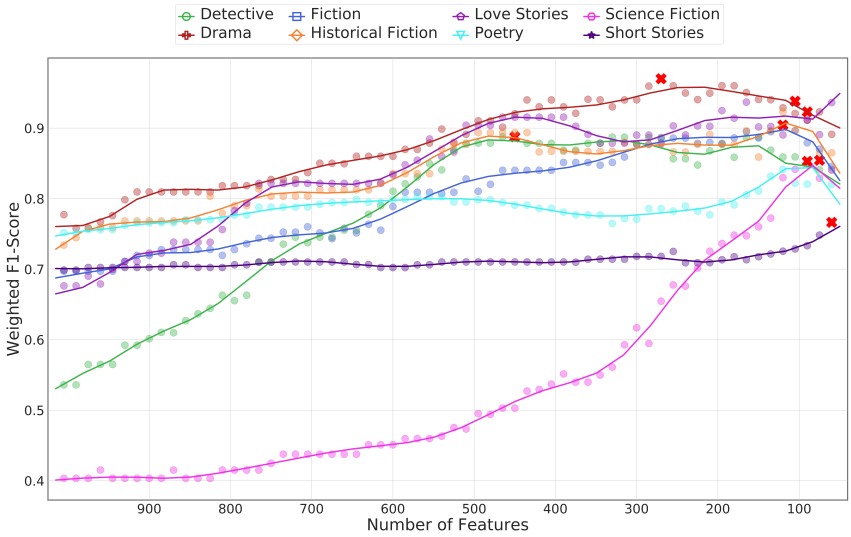

Figure 5: We performed feature selection process for each of the 8 genres. This figure represents the weighted F1-score achieved for different feature sets. Here the max length of the feature set is 1,019. Thus, at each iteration a single feature was filtered based on its weight. For each genre, we select the set of features that obtain the highest F1-score. The best performance for each genre for a particular feature set is marked with *X*. This plot shows an interesting insight that it is not necessary to use more than 500 concepts/features to represent a book.

Method	**Genre (Weighted F1)**								**Average**			
	Detect	**Drama**	**Fiction**	**Hist**	**Love**	**Poet**	**Scien**	**Short**	**Acc.**	**W. F1**	**Pre**	**Rec**
Roget Association	0.90	0.97	0.90	0.92	0.95	0.86	0.86	0.77	**0.89**	0.89	0.86	0.87
Roget Frequency	0.92	0.90	0.86	0.93	0.87	0.83	0.89	0.83	0.88	0.88	0.85	0.86
LIWC Frequency	0.68	0.68	0.69	0.74	0.74	0.82	0.49	0.70	0.69	0.69	0.63	0.64
PoS Frequency	0.70	0.68	0.66	0.65	0.61	0.77	0.47	0.68	0.64	0.65	0.6	0.61
NRC Frequency	0.58	0.69	0.62	0.72	0.65	0.75	0.59	0.72	0.66	0.67	0.62	0.62

Table 2: Genre-wise classification results

5.5 Result Interpretation

To explore the importance of semantic word associations in book success prediction, we present sunburst plot of reduced feature set. In figure 6, we observe that "Detective and Mystery" is the most interesting since it goes against expectations in a way that makes sense. Specifically, we would probably expect the *Intellectual Faculties*, *Related To Matter*, and *Abstract Relation* categories to be positively associated with stories about solving a crime/mystery using intellect, evidence, and abstract relationships. However, it appears that the most popular stories of this genre actually favor things that have less to do with evidence and more to do with characters and their choices/feelings. This is illustrated by the positive associations of *Voluntary Powers*, *Related To Space*, and *Sentiment and Moral Powers*. In other words, it seems readers like it best when a detective solves a mystery because he/she is "the good guy" who makes the right choices, rather than through real detective work.

Among all the 24 themes, *Intellectual Faculties* shows some interesting insights about the success prediction of a book. So we'll discuss about the impact of this theme in classifying books across different genres. The top features that the weighted linear SVM classifier determined for successful poetry books are *Analogy*, *Obscurity*, *Overestimation*, etc. This sheds light on the writing style of many of the greatest poems where the poets show a connection between materialistic and abstract entities while keeping some room for the readers to perceive the same poem with their own different flavor of apprehension. This finding is further validated by the presence of *Perspicuity* as one of the top features for unsuccessful poetry books. For example, take the following poem -

O my Luve is like a red, red rose
That's newly sprung in June;
O my Luve is like the melody
That's sweetly played in tune.

— Robert Burns

Figure 6: The large sunburst presents a comprehensive review of the most discriminative Roget classes, themes and concepts for a single genre, "Detective Mystery". While the small circles represent the disriminative feature distribution across multiple genres for a common Roget class, "Intellectual Faculties". We consider the top 30 discriminative features for both successful and unsuccessful books. Discriminative features for successful and unsuccessful books are colored with green and red respectively.

Here, the analogy between love and rose may arise a debate between the readers where one side will find the poem expressing that love is beautiful like a rose while the opposition might say this poem is indicating the delicacy and fragility of love. For the Love Stories genre, concepts like *Thought*, *Reasoning*, *Conversation*, *Perspicuity* work as important features for a successful book prediction. This goes against the normal way of thinking that a good love story book should only contain overflowing emotions, gestures that abandon earthly reasonings for the triumph of romance, etc. But it seems like the readers tend to prefer romantic books where lovers also consider their logical reasoning, worldly obligations while trying to win over their love. The 'Intellectual Faculties' section has an overall positive impact on detecting successful books of the Science Fiction genre. It is expected, as the main focus of successful science fiction books is towards many scientific revolutions or the main timeline of the story is set on futuristic utopian or dystopian civilization where new technology is introduced. We present the sunburst plot for all genres in the *Appendix Figures 2-8*.

6 Conclusion and Future Work

We present a novel study of word association of book content to predict the success of book and show that semantic word association features can be new vertical of the classification based task. Our empirical results demonstrate that word association and different types of concepts can be very useful to capture the book's literary content and can predict the book success with better accuracy. Rather than individual word frequency, the set of words with similar concepts has been proved to be more effective. We will continue our research work in this area and we intend to perform the experiments on a bigger data set in the future. We hypothesize that instead of preparing word embedding for individual book, we can retrain the global embedding using genre-wise data. This genre specialized embedding can help us to obtain a much better result for two reasons - as each embedding will be retrained on individual genre, the quality of generated embedding is expected to be better and it will represent the genre specific context for each word more explicitly.

References

2008. *Kolmogorov–Smirnov Test*. Springer New York, New York, NY.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Esther E Bron, Marion Smits, Wiro J Niessen, and Stefan Klein. 2015. Feature selection based on the svm weight vector for classification of dementia. *IEEE journal of biomedical and health informatics*, 19(5):1617–1626.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

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*.

Andreas van Cranenburgh and Corina Koolen. 2015. Identifying literary texts with bigrams. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 58–67, Denver, Colorado, USA. Association for Computational Linguistics.

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.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

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.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

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.

George H John, Ron Kohavi, and Karl Pfleger. 1994. Irrelevant features and the subset selection problem. In *Machine Learning Proceedings 1994*, pages 121–129. Elsevier.

Jason J Jones, Mohammad Ruhul Amin, Jessica Kim, and Steven Skiena. 2020. Stereotypical gender associations in language have decreased over time. *Sociological Science*, 7:1–35.

David G Kendall. 1989. A survey of the statistical theory of shape. *Statistical Science*, pages 87–99.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635.

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. 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.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

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.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1).

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.

K. Vonnegut. 1981. *Palm Sunday: An autobiographical collage*. New York: Delacorte Press.

Appendix

Algorithm 1: Association Score Calculate

Input: Books $b_i, i = 1, ..., n.$
Roget Concepts $r_i, i = 1, ..., c$
Global Word Embedding E_G
Output: The association vector
$a_i, i = 1, ..., n$

for $i \leftarrow 1$ **to** n **do**
  $P_{b_i} = Embedding(b_i)$
  $E_{b_i} = Align(P_{b_i}, E_G)$
  for $j \leftarrow 1$ **to** c **do**
    $W_G = Words(E_G)$
    $W_{b_i} = Words(E_{b_i})$
    $W_{r_j} = Words(r_j)$
    $W = W_G \cap W_{b_i} \cap W_{r_j}$
    for $k \leftarrow 1$ **to** $len(W)$ **do**
      $VG_{W_k} = E_G[W_k]$
      $VL_{W_k} = E_{b_i}[W_k]$
      $D_k = \sqrt{\sum_{l=1}^{L} (VG_{W_k,l} - VL_{W_k,l})^2}$
    end
    $a_{i,j} = AVG(D)$
  end
end

Algorithm 2: Feature Ranking Based on Linear SVM Weights

Input: Training sets, (x_i, y_i) , $i = 1, ..., l$.
Output: Sorted feature ranking list.

1. Use grid search to find the best parameter *C*.
2. Train a L2-loss linear SVM model using the best *C*.
3. Sort the features according to the absolute values of weights in the model.

Algorithm 3: Training and Prediction

Input: Training sets, testing sets.
Output: Predictions on nested subsets.

1. Use a feature ranking algorithm to compute the sorted feature list $f_j, j = 1, ..., n.$
2. For each feature size $m \in \{50, 51, ..., n\}$.
   (a) Generate the new training set that has only the first m features in the sorted feature list, $f_j, j = 1, ..., m.$
   (b) Use grid search to find the best parameter C.
   (c) Train the L2-loss linearSVM model on the new training set.
   (d) Predict the testing set using the model.

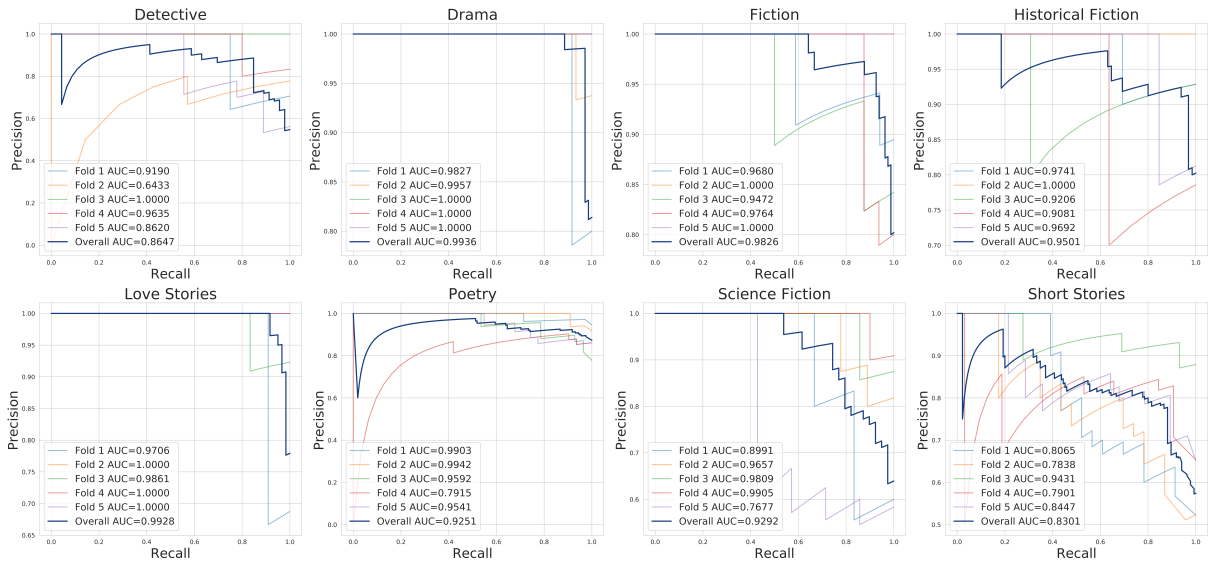

Figure 1: The precision-recall (PRC) plot shows precision values for corresponding sensitivity (recall) values for the association analysis. This PRC plot provides a model-wide evaluation.

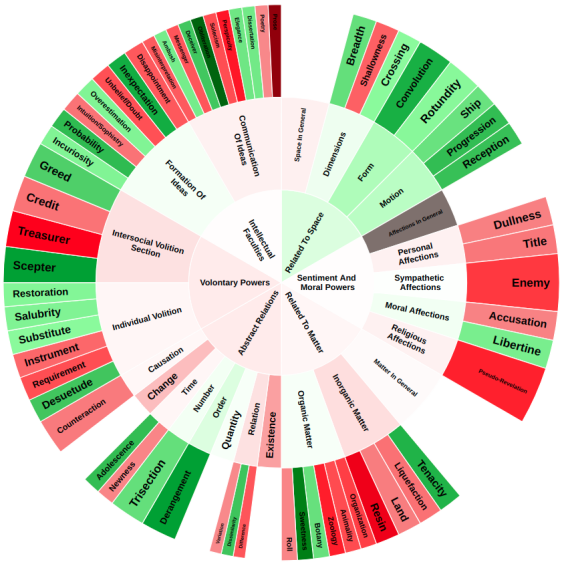

Figure 2: This figure presents a comprehensive review of the most discriminative Roget classes, themes and concepts for Roget Association Analysis of Drama genre. We consider the top 30 discriminative features for both successful and unsuccessful books. Discriminative features for successful and unsuccessful books are colored with green and red respectively.

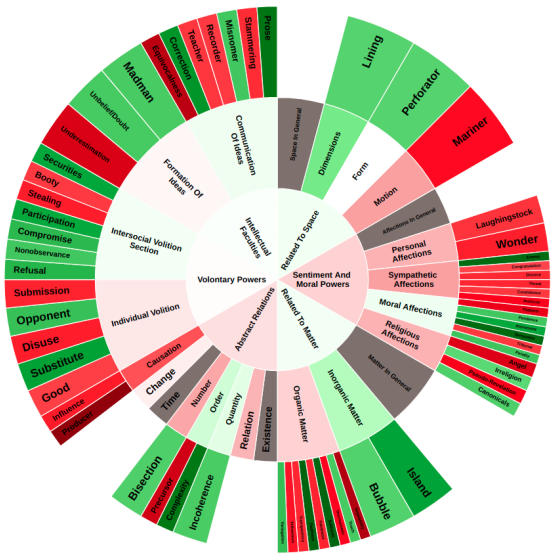

Figure 3: This figure presents a comprehensive review of the most discriminative Roget classes, themes and concepts for Roget Association Analysis of Fiction genre. We consider the top 30 discriminative features for both successful and unsuccessful books. Discriminative features for successful and unsuccessful books are colored with green and red respectively.

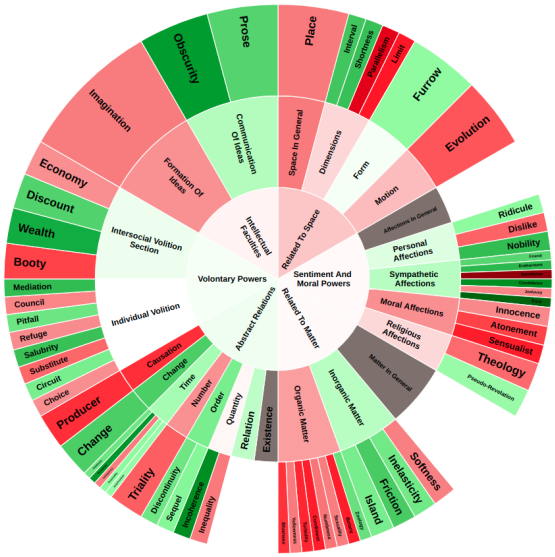

Figure 4: This figure presents a comprehensive review of the most discriminative Roget classes, themes and concepts for Roget Association Analysis of Historical Fiction genre. We consider the top 30 discriminative features for both successful and unsuccessful books. Discriminative features for successful and unsuccessful books are colored with green and red respectively.

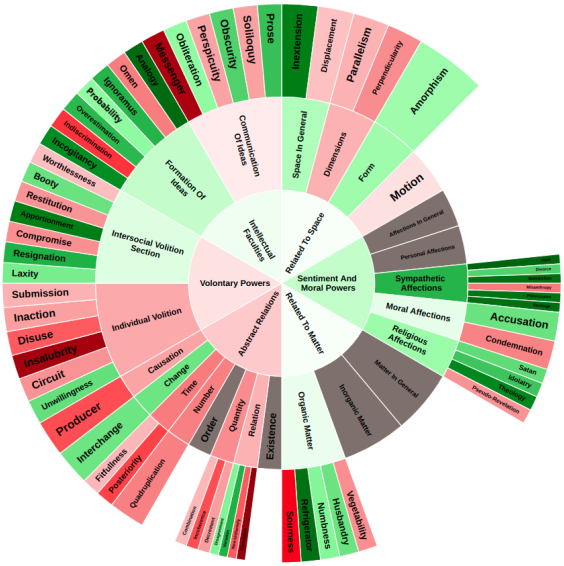

Figure 6: This figure presents a comprehensive review of the most discriminative Roget classes, themes and concepts for Roget Association Analysis of Poetry genre. We consider the top 30 discriminative features for both successful and unsuccessful books. Discriminative features for successful and unsuccessful books are colored with green and red respectively.

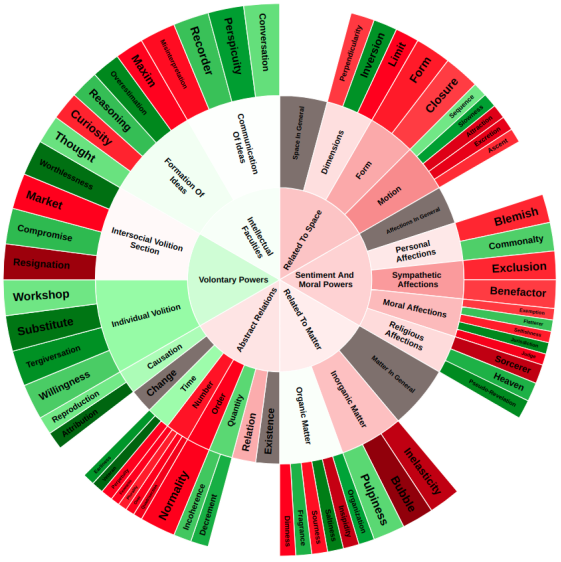

Figure 5: This figure presents a comprehensive review of the most discriminative Roget classes, themes and concepts for Roget Association Analysis of Love Stories genre. We consider the top 30 discriminative features for both successful and unsuccessful books. Discriminative features for successful and unsuccessful books are colored with green and red respectively.

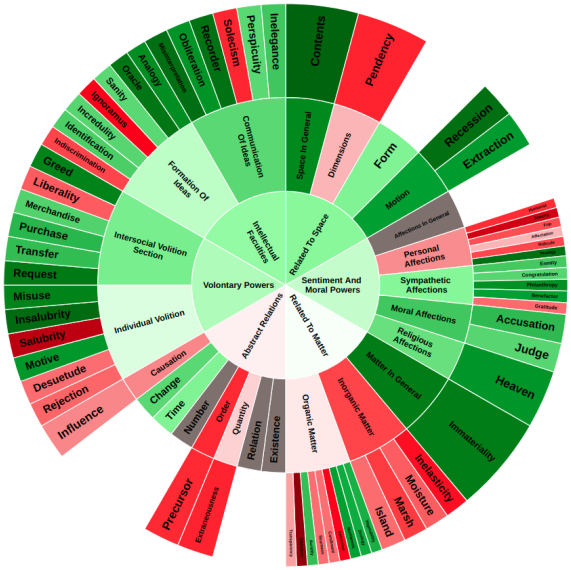

Figure 7: This figure presents a comprehensive review of the most discriminative Roget classes, themes and concepts for Roget Association Analysis of Science Fiction genre. We consider the top 34 and 26 discriminative features for successful and unsuccessful books respectively. Discriminative features for successful and unsuccessful books are colored with green and red respectively.

Figure 8: This figure presents a comprehensive review of the most discriminative Roget classes, themes and concepts for Roget Association Analysis of Short Stories genre. We consider the top 36 and 24 discriminative features for successful and unsuccessful books respectively. Discriminative features for successful and unsuccessful books are colored with green and red respectively.

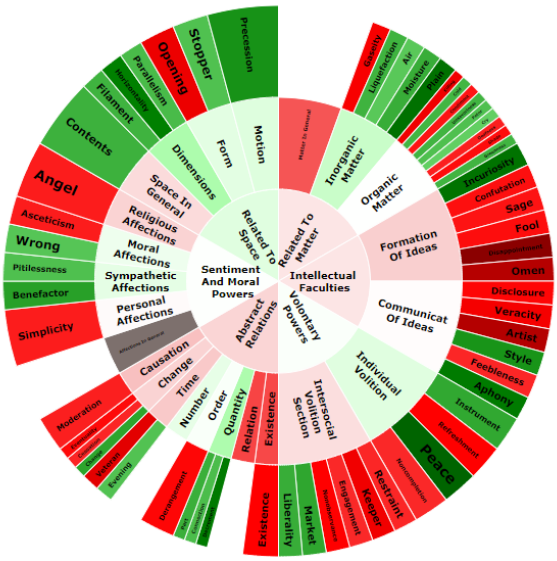

Figure 9: This figure presents a comprehensive review of the most discriminative Roget classes, themes and concepts for Roget Frequency Analysis of Detective and Mystery genre. We consider the top 30 discriminative features for both successful and unsuccessful books. Discriminative features for successful and unsuccessful books are colored with green and red respectively.