Success with Style: Using Writing Style to Predict the Success of Novels

Vikas Ganjigunte Ashok    Song Feng    Yejin Choi
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400
vganjiguntea, songfeng, ychoi@cs.stonybrook.edu

Abstract

Predicting the success of literary works is a curious question among publishers and aspiring writers alike. We examine the quantitative connection, if any, between writing style and successful literature. Based on novels over several different genres, we probe the predictive power of statistical stylometry in discriminating successful literary works, and identify characteristic stylistic elements that are more prominent in successful writings. Our study reports for the first time that statistical stylometry can be surprisingly effective in discriminating highly successful literature from less successful counterpart, achieving accuracy up to 84%. Closer analyses lead to several new insights into characteristics of the writing style in successful literature, including findings that are contrary to the conventional wisdom with respect to good writing style and readability.

1 Introduction

Predicting the success of novels is a curious question among publishers, professional book reviewers, aspiring and even expert writers alike. There are potentially many influencing factors, some of which concern the intrinsic content and quality of the book, such as interestingness, novelty, style of writing, and engaging storyline, but external factors such as social context and even luck can play a role. As a result, recognizing successful literary work is a hard task even for experts working in the publication industries. Indeed, even some of the best sellers and award winners can go through several rejections before they are picked up by a publisher.\(^1\)

Perhaps due to its obvious complexity of the problem, there has been little previous work that attempts to build statistical models that predict the success of literary works based on their intrinsic content and quality. Some previous studies do touch on the notion of stylistic aspects in successful literature, e.g., extensive studies in Literature discuss literary styles of significant authors (e.g., Ellegård (1962), McGann (1998)), while others consider content characteristics such as plots, characteristics of characters, action, emotion, genre, cast, of the best-selling novels and blockbuster movies (e.g., Harvey (1953), Hall (2012), Yun (2011)).

All these studies however, are qualitative in nature, as they rely on the knowledge and insights of human experts on literature. To our knowledge, no prior work has undertaken a systematic quantitative investigation on the overarching characterization of the writing style in successful literature. In consideration of widely different styles of authorship (e.g., Escalante et al. (2011), Peng et al. (2003), Argamon et al. (2003)), it is not even readily clear whether there might be common stylistic elements that help discriminating highly successful ones from less successful counterpart.

In this work, we present the first study that investigates this unstudied and unexpected connection between stylistic elements and the literary success. The key findings of our research reveal that there exists distinct linguistic patterns shared among suc-

\(^1\)E.g., Paul Harding’s “Tinkers” that won 2010 Pulitzer Prize for Fiction and J. K. Rowling’s “Harry Potter and the Philosopher’s Stone” that sold over 450 million copies.
cessful literature, at least within the same genre, making it possible to build a model with surprisingly high accuracy (up to 84%) in predicting the success of a novel. This result is surprising for two reasons. First, we tackle the hard task of predicting the success of novels written by previously unseen authors, avoiding incidental learning of authorship signature, since previous research demonstrated that one can achieve very high accuracy in authorship attribution (as high as 96% in some experimental setup) (e.g., Raghavan et al. (2010), Feng et al. (2012)). Second, we aim to discriminate highly successful novels from less successful, but nonetheless published books written by professional writers, which are undoubtedly of higher quality than average writings. It is important to note that the task we tackle here is much harder than discriminating highly successful works from those that have not even passed the scrutinizing eyes of publishers.

In order to quantify the success of literary works, and to obtain corresponding gold standard labels, one needs to first define “success”. For practical convenience, we largely rely on the download counts available at Project Gutenberg as a surrogate to quantify the success of novels. For a small number of novels however, we also consider award recipients (e.g., Pulitzer, Nobel), and Amazon’s sales records to define a novel’s success. We also extend our empirical study to movie scripts, where we quantify the success of movies based on the average review scores at imbd.com. We leave analysis based on other measures of literary success as future research.

In this study, we do not attempt to separate out success based on literary quality (award winners) from success based on popularity (commercial hit, often in spite of bad literary quality), mainly because it is not practically easy to determine whether the high download counts are due to only one reason or the other. We expect that in many cases, the two different aspects of success are likely to coincide, however. In the case of the corpus obtained from Project Gutenberg, where most of our experiments are conducted, we expect that the download counts are more indicative of success based on the literary quality (which then may have resulted in popularity) rather than popularity without quality.

We examine several genres in fiction and movie

<table>
<thead>
<tr>
<th>Genre</th>
<th>#Books</th>
<th>$\tau^-$</th>
<th>$\tau^+$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adventure</td>
<td>409</td>
<td>17</td>
<td>100</td>
</tr>
<tr>
<td>Detective / Mystery</td>
<td>374</td>
<td>25</td>
<td>90</td>
</tr>
<tr>
<td>Fiction</td>
<td>1148</td>
<td>7</td>
<td>125</td>
</tr>
<tr>
<td>Historical Fiction</td>
<td>374</td>
<td>25</td>
<td>115</td>
</tr>
<tr>
<td>Love Stories</td>
<td>342</td>
<td>16</td>
<td>85</td>
</tr>
<tr>
<td>Poetry</td>
<td>580</td>
<td>9</td>
<td>70</td>
</tr>
<tr>
<td>Science Fiction</td>
<td>902</td>
<td>30</td>
<td>100</td>
</tr>
<tr>
<td>Short Stories</td>
<td>1117</td>
<td>9</td>
<td>224</td>
</tr>
</tbody>
</table>

Table 1: # of books available per genre at Gutenberg with download thresholds used to define more successful ($\geq \tau^+$) and less successful ($\leq \tau^-$) classes.

scripts, e.g., adventure stories, mystery, fiction, historical fiction, sci-fi, short stories, as well as poetry, and present systematic analyses based on lexical and syntactic features which have been known to be effective in a variety of NLP tasks ranging from authorship attribution (e.g., Raghavan et al. (2010)), genre detection (e.g., Rayson et al. (2001), Douglas and Broussard (2000)), gender identification (e.g., Sarawgi et al. (2011)) and native language detection (e.g., Wong and Dras (2011)).

Our empirical results demonstrate that (1) statistical stylometry can be surprisingly effective in discriminating successful literature, achieving accuracy up to 84%, (2) some elements of successful styles are genre-dependent while others are more universal. In addition, this research results in (3) findings that are somewhat contrary to the conventional wisdom with respect to the connection between successful writing styles and readability, (4) interesting correlations between sentiment / connotation and the literary success, and finally, (5) comparative insights between fiction and nonfiction with respect to the successful writing style.

2 Dataset Construction

For our experiments, we procure novels from project Gutenberg\footnote{http://www.gutenberg.org/}. Project Gutenberg houses over 40,000 books available for free download in electronic format and provides a catalog containing brief descriptions (title, author, genre, language, download count, etc.) of these books. We experiment with genres in Table 1, which have sufficient number of books allowing us to construct reasonably sized datasets.

We use the download counts in Gutenberg-catalog
as a surrogate to measure the degree of success of novels. For each genre, we determine a lower bound ($\tau_+$) and an upper bound ($\tau_-$) of download counts as shown in Table 1 to categorize the available books as more successful and less successful respectively. These thresholds are set to obtain at least 50 books for each class, and for each genre. To balance the data, for each genre, we construct a dataset of 100 novels (50 per class).

We make sure that no single author has more than 2 books in the resulting dataset, and in the majority of the cases, only one book has been taken from each author. Furthermore, we make sure that the books from the same author do not show up in both training and test data. These constraints make sure that we learn general linguistic patterns of successful novels, rather than a particular writing style of a few successful authors.

3 Methodology

In what follows, we describe five different aspects of linguistic styles we measure quantitatively. The first three correspond to the features that have been frequently utilized in previous studies in related tasks, e.g., genre detection (e.g., Kessler et al. (1997)) and authorship attribution (e.g., Stamatakis (2009)), while the last two are newly explored in this work.

I. Lexical Choices: unigrams and bigrams.

II. Distribution of Word Categories: Many previous studies have shown that the distribution of part-of-speech (POS) tags alone can reveal surprising insights on genre and authorship (e.g., Koppel and Schler (2003)), hence we examine their distributions with respect to the success of literary works.

III. Distribution of Grammar Rules: Recent studies reported that features based on CFG rules are helpful in authorship attribution (e.g., Raghavan et al. (2010), Feng et al. (2012)). We experiment with four different encodings of production rules:

- $\Gamma$: lexicalized production rules (all production rules, including those with terminals)
- $\Gamma^G$: lexicalized production rules prepended with the grandparent node.
- $\gamma$: unlexicalized production rules (all production rules except those with terminals).
- $\gamma^G$: unlexicalized production rules prepended with the grandparent node.
IV. Distribution of Constituents: PCFG grammar rules are overly specific to draw a big picture on the distribution of large, recursive syntactic units. We hypothesize that the distribution of constituents can serve this purpose, and that it will reveal interesting and more interpretable insights into writing styles in highly successful literature. Despite its relative simplicity, we are not aware of previous work that looks at the distribution of constituents directly. In particular, we are interested in examining the distribution of phrasal and/or clausal tags as follows: (i) Phrasal tag percent (PHR) - percentage distribution of phrasal tags,⁴ (ii) Clausal tag percent (CLS) - percentage distribution of clausal tags.

V. Distribution of Sentiment and Connotation: Finally, we examine whether the distribution of sentiment and connotation words, and their polarity, has any correlation with respect to the success of literary works. We are not aware of any previous work that looks into this connection.

4 Prediction Performance

We use LibLinear SVM (Fan et al., 2008) with L2 tuned over training data, and all performance is based on 5-fold cross validation. We take 1000 sentences from the beginning of each book. POS features are encoded as unit normalized frequency and all other features are encoded as tf-idf.⁵

Prediction Results Table 2 shows the classification results. The best performance reaches as high as 84% in accuracy. In fact, in all genres except for history, the best performance is at least 74%, if not higher. Another notable observation is that even in the poetry genre, which is not prose, the accuracy gets as high as 74%. This level of performance is not entirely anticipated, given that (1) the test data consists of books written only by previously unseen authors, and (2) each author has widely different writing style, and (3) we do not have training data at scale, and (4) we aim to tackle the hard task of discriminating highly successful ones from less successful, but nonetheless successful ones, as all of them were, after all, good enough to be published.⁶

Prediction with Varying Thresholds of Download Counts Before we proceed to comprehensive analysis of writing style that are prominent in more successful literature (§5), in Table 3, we present how the prediction accuracy varies as we adjust the definition of more-successful and less-successful literature, by gradually increasing (decreasing) the threshold \( \tau^- \) (\( \tau^+ \)). As we reduce the gap between \( \tau^- \) and \( \tau^+ \), the performance decreases, which shows that indeed there are notable statistical differences in linguistic patterns between novels with high and low download counts, and the stylistic difference monotonically increases (thereby higher prediction accuracy) as we increase the gap between two classes.⁶

In our pilot study, we also experimented with the binary classification task of discriminating highly successful ones from those that are not even published (unpublished online novels), and it was a much easier task as expected.
This is particularly interesting as the size of training data set is actually monotonically decreasing (making it harder to achieve high accuracy) while we increase the separation between $\tau^-$ and $\tau^+$. 

5 Analysis of Successful Writing Styles

5.1 Insights Based on Lexical Choices

It is apparent from Table 2 that unigram features yield curiously high performance in many genres. We therefore examine discriminative unigrams for ADVENTURE, shown in Table 4. Interestingly, less successful books rely on verbs that are explicitly descriptive of actions and emotions (e.g., “wanted”, “took”, “promised”, “cried”, “cheered”, etc.), while more successful books favor verbs that describe thought-processing (e.g., “recognized”, “remembered”), and verbs that serve the purpose of quotes and reports (e.g., “say”). Also, more successful books use discourse connectives and prepositions more frequently, while less successful books rely more on topical words that could be almost cliché, e.g., “love”, typical locations, and involve more extreme (e.g., “breathless”) and negative words (e.g., “risk”).

5.2 Distribution of Sentiment & Connotation

We also determine the distribution of sentiment and connotation words separately for each class (Table 5) to check if there exists a connection with respect to successful writing styles.\footnote{We use MPQA subjectivity lexicon (Wilson et al., 2005) and connotation lexicon (Feng et al., 2013) for determining sentiment and connotation of words respectively.} We first compare distribution of sentiment and connotation for the entire words. As can be seen in Table 5 – Bottom, we find substantial differences in all genres. In particular, discriminative unigrams that characterize less successful novels involve significantly more sentiment-laden words.

5.3 Distribution of Word Categories

Summarized analysis of POS distribution across all genres is reported in Table 6. It can be seen that prepositions, nouns, pronouns, determiners and adjectives are predictive of highly successful books whereas less successful books are characterized by higher percentage of verbs, adverbs, and foreign words. Per genre distributions of POS tags are visualized in Figure 1. Interestingly, some POS tags show almost universal patterns (e.g., prepositions (IN), NNP, WP, VB), while others are more genre-specific.

In Relation to Journalism Style

The work of Douglas and Broussard (2000) reveals that informative writing (journalism) involves increased use of nouns, prepositions, determiners and coordinating conjunctions whereas imaginative writing (novels) involves more use of verbs and adverbs, as has been also confirmed by Rayson et al. (2001). Comparing their findings with Table 6, we find that highly
success of successful books. Production rules are overly specific to gain more generalized, interpretable, high-level insights however (Feng et al., 2012). Therefore, similarly as word categories (POS), we consider the categories of nonterminal nodes of the parse trees, in particular, phrasal and clausal tags, as they represent the gist of constituent structure that goes beyond shallow syntactic information represented by POS.

Table 8 shows how the distribution of phrasal and clausal tags differ for successful books when computed over all genres. Positive (negative) DIFF values indicate that the corresponding tags are favored in more successful (less successful) books when counted across all genres. We also report the number of genres (#Genres) in which the individual difference is positive / negative.

In terms of phrasal tags, we find that more successful books are composed of higher percentage of PP, NP and wh-noun phrases (WHNP), whereas less successful books are composed of higher percentage of VP, adverb phrases (ADVP), interjections (INTJ) and fragments (FRAG). Notice that this observation is inline with our earlier findings with respect to the distribution of POS.

In regard to clausal tags, more successful books involve more clausal tags that are necessary for complex sentence structure and inverted sentence structure (SBAR, SBARQ and SQ) whereas less successful books rely more on simple sentence structure (S). Figure 2 shows the visualization of the distribution of these phrasal and clausal tags.

It is also worth to mention that phrasal and clausal

Table 5: Top: Distribution of sentiment (connotation) among entire unigrams. Bottom: distribution of sentiment (connotation) among discriminative unigrams. 'S' and 'C' stand for sentiment and connotation respectively.

Table 6: Top discriminative POS tags.

5.4 Distribution of Constituents
It can be seen in Table 2 that deep syntactic features expressed in terms of different encodings of production rules consistently yield good perfor-
Figure 2: Difference between phrasal and clausal tag percentage distributions of more successful and less successful books across different genres. Specifically, we plot $D^- - D^+$, where $D^+$ is the phrasal tag distribution (in %) of more successful books and $D^-$ is the phrasal tag distribution (in %) of less successful books.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FOG index</td>
<td>9.88</td>
<td>9.80</td>
</tr>
<tr>
<td>Flesch index</td>
<td>87.48</td>
<td>87.64</td>
</tr>
</tbody>
</table>

Table 7: Readability: Lower FOG and higher Flesch indicate higher readability (numbers in Boldface).

tags alone can yield classification performance that are generally better than that of POS tags, in spite of the very small feature set (26 tags in total). In fact, constituent tags deliver the best performance in case of historical fiction genre (Table 2).

**Connection to Readability** Pitler and Nenkova (2008) provide comprehensive insights into assessment of readability. In their work, among the most discriminating features characterizing text with better readability is increased use of verb phrases (VP). Interestingly, contrary to the conventional wisdom – that readability is of desirable quality of good writings – our findings in Table 2 suggest that the increased use of VP correlates strongly with the writing style of the opposite spectrum of highly successful novels.

As a secondary way of probing the connection between readability and the writing style of successful literature, we also compute two different readability measures that have been used widely in prior literature (e.g., Sierra et al. (1992), Blumenstock (2008), Ali et al. (2010)): (i) Flesch reading ease score (Flesch, 1948), (ii) Gunning FOG index (Gunning, 1968). The overall weighted average readability scores are reported in Table 7. Again, we find that less successful novels have higher readability compared to more successful ones.

The work of Sawyer et al. (2008) provides yet another interesting contrasting point, where the authors found that award winning academic papers in marketing journals correlate strongly with increased readability, characterized by higher percentage of simple sentences. We conjecture that this opposite trend is likely to be due to difference between fiction and nonfiction, leaving further investigation as future research.

In sum, our analysis reveals an intriguing and unexpected observation on the connection between **readability and the literary success — that they correlate into the opposite directions**. Surely our findings only demonstrate correlation, not to be con-
fused as causation, between readability and literary success. We conjecture that the conceptual complexity of highly successful literary work might require syntactic complexity that goes against readability.

6 Literature beyond Project Gutenberg

One might wonder how the prediction algorithms trained on the dataset based on Project Gutenberg might perform on books not included at Gutenberg. This section attempts to address such a question. Due to the limited availability of electronically available books that are free of charge however, we could not procure more than a handful of books.  

6.1 Highly Successful Books

First, we apply the classifiers trained on the Project Gutenberg dataset (all genres merged) on a few extremely successful novels (Pulitzer prize, National Award recipients, etc). Table 9 shows the results of two classification options: (1) KL-divergence based, and (2) unigram-feature based. Although KL-divergence based prediction was not part of the classifiers that we explored in the previous sections, we include it here mainly to provide better insights as to which well-known books share closer structural similarity to either more or less successful writing style. As a probability model, we use the distributions of phrasal tags, as those can give us insights on deep syntactic structure while suppressing potential noises due to topical variances. Table 9 shows symmetrised KL-divergence between each of the previously unseen novels and the collection of books from Gutenberg corresponding to more successful (less successful) labels. For prediction, the label with smaller KL is chosen.

Table 9: Prediction on books beyond Gutenberg. Shaded entries indicate incorrect predictions.

---

Fused as causation, between readability and literary success. We conjecture that the conceptual complexity of highly successful literary work might require syntactic complexity that goes against readability.

6 Literature beyond Project Gutenberg

One might wonder how the prediction algorithms trained on the dataset based on Project Gutenberg might perform on books not included at Gutenberg. This section attempts to address such a question. Due to the limited availability of electronically available books that are free of charge however, we could not procure more than a handful of books.

6.1 Highly Successful Books

First, we apply the classifiers trained on the Project Gutenberg dataset (all genres merged) on a few extremely successful novels (Pulitzer prize, National Award recipients, etc). Table 9 shows the results of two classification options: (1) KL-divergence based, and (2) unigram-feature based. Although KL-divergence based prediction was not part of the classifiers that we explored in the previous sections, we include it here mainly to provide better insights as to which well-known books share closer structural similarity to either more or less successful writing style. As a probability model, we use the distributions of phrasal tags, as those can give us insights on deep syntactic structure while suppressing potential noises due to topical variances. Table 9 shows symmetrised KL-divergence between each of the previously unseen novels and the collection of books from Gutenberg corresponding to more successful (less successful) labels. For prediction, the label with smaller KL is chosen.

Based only on the distribution of 26 phrasal tags, the KL divergence classifier is able to make correct predictions.
predictions on 7 out of 10 books, a surprisingly high performance based on mere 26 features. Of course, considering only the distribution of phrasal tags is significantly less informed than considering numerous other features that have shown substantially better performance, e.g., unigrams and CFG rewrite rules. Therefore, we also present the SVM classifier trained on unigram features. It turns out unigram features are powerful enough to make correct predictions for all ten books in Table 9.

Hemingway and Minimalism It is good to think about where and why KL-divergence-based approach fails. In fact, when we included Hemingway’s The Old Man and the Sea into the test set, we were expecting some level of confusions when relying only on high-level syntactic structure, as Hemingway’s signature style is minimalism, with 70% of his sentences corresponding to simple sentences. Not surprisingly, more adequately informed classifiers, e.g., SVM with unigram features, are still able to recognize Hemingway’s writings as those of highly successful ones.

6.2 Less Successful Books
In order to obtain less successful books, we consider the Amazon seller’s rank included in the product details of a book. The less successful books considered in Table 9 had an Amazon seller’s rank beyond 200k (higher rank indicating less commercial success) except Dan Brown’s The lost symbol, which we included mainly because of negative critiques it had attracted from media despite its commercial success. As shown in Table 9, all three classifiers make (arguably) correct predictions on Dan Brown’s book. This result also supports our earlier assumption on the nature of novels available at Project Gutenberg — that they would be more representative of literary success than general popularity (with or without literary quality).

7 Predicting Success of Movie Scripts
We have seen successful results in the novel domain, but can stylometry-based prediction work on very different domains, such as screenplays? Unlike novels, movie scripts are mostly in dialogues, which are likely to be more informal. Also, what to keep in mind is that much of the success of movies depends on factors beyond the quality of writing of the scripts, such as the quality of acting, the popularity of actors, budgets, artistic taste of directors and producers, editing and so forth.

We use the Movie Script Dataset introduced in Danescu-Niculescu-Mizil and Lee (2011). It includes the dialogue scripts of 617 movies. The average rating of all movies is 6.87. We consider movies with IMDb rating ≥ 8 as “more successful”, the ones with IMDb rating ≤ 5.5 as “less successful”. We combine all the dialogues of each movie and filter out the movies with less than 200 sentences. There are 11 genres (“ADVENTURE”, “FANTASY”, “ROMANCE”, “THRILLER”, “ACTION”, “COMEDY”, “CRIME”, “DRAMA”, “HORROR”, “MYSTERY”, “SCI-FI”) with 15 movies or more per class, we take 15 movies per class and perform classification tasks with the same experiment setting as Table 2.

Table 10, we show some of the example genres with relatively successful outcome, reaching as high as 89.3% accuracy in FANTASY genre. We would like to note however that in many other genres, the prediction did not work as well as it did for the novel domain. We suspect that there are at least two reasons for this: it must be partly due to very limited data size — only 15 instances per class with the rating threshold we selected for defining the success of

<table>
<thead>
<tr>
<th>Feature</th>
<th>Adven</th>
<th>Fanta</th>
<th>Roman</th>
<th>Thril</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>62.0</td>
<td>58.0</td>
<td>61.7</td>
<td>56.0</td>
</tr>
<tr>
<td>Unigram</td>
<td>62.0</td>
<td>81.3</td>
<td>70.0</td>
<td>80.0</td>
</tr>
<tr>
<td>Bigrams</td>
<td>73.3</td>
<td>84.7</td>
<td>80.8</td>
<td>76.0</td>
</tr>
<tr>
<td>Γ</td>
<td>66.0</td>
<td>81.3</td>
<td>70.0</td>
<td>76.0</td>
</tr>
<tr>
<td>Γ^G</td>
<td>62.0</td>
<td>69.3</td>
<td><strong>86.7</strong></td>
<td>60.0</td>
</tr>
<tr>
<td>γ</td>
<td>62.0</td>
<td>81.3</td>
<td>78.3</td>
<td>76.0</td>
</tr>
<tr>
<td>γ^G</td>
<td>69.3</td>
<td>77.3</td>
<td>77.5</td>
<td>68.0</td>
</tr>
<tr>
<td>Γ+Uni</td>
<td>62.0</td>
<td>85.3</td>
<td>70.0</td>
<td>76.0</td>
</tr>
<tr>
<td>Γ^G+Uni</td>
<td>54.7</td>
<td>81.3</td>
<td>70.0</td>
<td>76.0</td>
</tr>
<tr>
<td>γ+Uni</td>
<td>58.0</td>
<td><strong>89.3</strong></td>
<td>70.0</td>
<td>76.0</td>
</tr>
<tr>
<td>γ^G+Uni</td>
<td>58.0</td>
<td>84.7</td>
<td>70.0</td>
<td>76.0</td>
</tr>
<tr>
<td>PHR</td>
<td>46.0</td>
<td>42.7</td>
<td>65.8</td>
<td><strong>80.0</strong></td>
</tr>
<tr>
<td>PHR+CLR</td>
<td>76.7</td>
<td>31.3</td>
<td>65.8</td>
<td><strong>80.0</strong></td>
</tr>
<tr>
<td>PHR+Uni</td>
<td>62.0</td>
<td>81.3</td>
<td>70.0</td>
<td><strong>80.0</strong></td>
</tr>
<tr>
<td>PHR+CLR+Uni</td>
<td>62.0</td>
<td>81.3</td>
<td>70.0</td>
<td><strong>80.0</strong></td>
</tr>
</tbody>
</table>

Table 10: Classification results on movie dialogue data (rating ≥ 8 vs rating ≤ 5.5).
movies. The second reason is due to many other external factors that can also influence the success of movies, as discussed earlier.

8 Related Work

Predicting success of novels and movies: To the best of our knowledge, our work is the first that provides quantitative insights into the unstudied connection between the writing style and the success of literary works. There have been some previous work that aims to gain insights into the secret recipe of successful books, but most were qualitative, based only on a dozen of books, focusing mainly on the high-level content of the books, such as the personalities of protagonists, antagonists, the nature of plots (e.g., Harvey (1953), Yun (2011)). In contrast, our work examines a considerably larger collection of books (800 in total) over eight different sub-genres, providing insights into lexical, syntactic, and discourse patterns that characterize the writing styles commonly shared among the successful literature. Another relevant work has been on a different domain of movies (Yun, 2011), however, the prediction is based only on external, non-textual information such as the reputation of actors and directors, and the power of distribution systems etc, without analyzing the actual content of the movie scripts.

Text quality and readability: Louis (2012) explored various features that measure the quality of text, which has some high-level connections to our work. Combining the insights from Louis (2012) with our results, we find that the characteristics of text quality explored in Louis (2012), readability of text in particular, do not correspond to the prominent writing style of highly successful literature. There have been a number of other work that focused on predicting and measuring readability (e.g., Kate et al. (2010), Pitler and Nenkova (2008), Schwarm and Ostendorf (2005), Heilman and Eskenazi (2006) and Collins-Thompson et al. (2004)) employing various linguistic features.

There is an important difference however, in regard to the nature of the selected text for analysis: most studies in readability focus on differentiating good writings from noticeably bad writings, often involving machine generated text or those written by ESL students. In contrast, our work essentially deals with differentiating good writings from even better writings. After all, all the books that we analyzed are written by expert writers who passed the scrutinizing eyes of publishers, hence it is reasonable to expect that the writing quality of even less successful books is respectful.

Predicting success among academic papers: In the domain of academic papers, which belongs to the broad genre of non-fiction, the work of Sawyer et al. (2008) investigated the stylistic characteristics of award winning papers in marketing journals, and found that the readability plays an important role. Combined with our study which focuses on fiction and creative writing, it suggests that the recipe for successful publications can be very different depending on whether it belongs to fiction or nonfiction. The work of Bergsma et al. (2012) is also somewhat relevant to ours in that their work included differentiating the writing styles of workshop papers from major conference papers, where the latter would be generally considered to be more successful.

9 Conclusion

We presented the first quantitative study that learns to predict the success of literary works based on their writing styles. Our empirical results demonstrated that statistical stylometry can be surprisingly effective in discriminating successful literature, achieving accuracy up to 84% in the novel domain and 89% in the movie domain. Furthermore, our study resulted in several insights including: lexical and syntactic elements of successful styles, the connection between successful writing style and readability, the connection between sentiment / connotation and the literary success, and last but not least, comparative insights between successful writing styles of fiction and nonfiction.

Acknowledgments This research was supported in part by the Stony Brook University Office of the Vice President for Research, and in part by gift from Google. We thank anonymous reviewers, Steve Greenspan, and Mike Collins for helpful comments and suggestions, Alex Berg for the title, and Arun Nampally for helping with the preliminary work.
References


