The Secret to Popular Chinese Web Novels: A Corpus-Driven Study

Yi-Ju Lin
Graduate Institute of Linguistics, National Taiwan University, Taiwan
jl8394@gmail.com

Shu-Kai Hsieh
Graduate Institute of Linguistics, National Taiwan University, Taiwan
shukai@gmail.com

Abstract

What is the secret to writing popular novels? The issue is an intriguing one among researchers from various fields. The goal of this study is to identify the linguistic features of several popular web novels as well as how the textual features found within and the overall tone interact with the genre and themes of each novel. Apart from writing style, non-textual information may also reveal details behind the success of web novels. Since web fiction has become a major industry with top writers making millions of dollars and their stories adapted into published books, determining essential elements of "publishable" novels is of importance. The present study further examines how non-textual information, namely, the number of hits, shares, favorites, and comments, may contribute to several features of the most popular published and unpublished web novels. Findings reveal that keywords, function words, and lexical diversity of a novel are highly related to its genres and writing style while dialogue proportion shows the narration voice of the story. In addition, relatively shorter sentences are found in these novels. The data also reveal that the number of favorites and comments serve as significant predictors for the number of shares and hits of unpublished web novels, respectively; however, the number of hits and shares of published web novels is more unpredictable.

2012 ACM Subject Classification General and reference → Empirical studies; General and reference

Keywords and phrases Popular Chinese Web Novels, NLP techniques, Sentiment Analysis, Publication of Web novels

Digital Object Identifier 10.4230/OASIcs.LDK.2019.24

Category Short Paper

1 Introduction

Is there a common pattern for popular novels? This is a curious question among publishers, professional book reviewers, and even researchers from various fields. More recently, with an increase in employment of empirical methods in studies of linguistics, exploiting and combining computational tool into research of language and literature has increasingly been the object of study in recent years.

The goal of this research is to identify the textual properties along with the external factors that may contribute to popular web novels in Taiwan. As previous literature indicated [1, 6, 9, 11], stylistic features are essential in differentiating authorial style and text genre. The first part of this study examines the textual content of these popular online novels. In other words, several stylistic features and the overall sentiment tone of the top 3 hit novels were investigated by exploiting Natural Language Processing (NLP) techniques such as keyword extraction and sentiment analysis. The most prominent features that can discriminate different genres and styles the best, for example, high-frequency words, dialogue proportion, average sentence length, and lexical richness, were displayed to show the shared textual elements in popular web novels. Finally, as previous literature [3, 19] noted, examining

2nd Conference on Language, Data and Knowledge (LDK 2019).
Editors: Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski; Article No. 24; pp. 24:1–24:8
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

writing style alone does not define a novel's success. It has to do a lot with book promotion and reader's feedback. In this paper, the non-textual information such as the number of hits, shares, favorites, and comments was examined to predict the top hit web novels and published web novels. The study attempts to expand the understanding of the shared elements among popular Chinese web novels in Taiwan.

2 Related Work

The anatomy of successful literary works is an intriguing issue among publishers and aspiring writers alike, and even researchers from various fields. Related works on the stylistic features of literary works are abundant. A number of publications [2, 5, 6, 9] have focused on the stylistic aspects in characterizing different genres and styles of literature. It is showed that using linguistic cues in classifying genres of literary works is effective. These discriminating features include passive use, terms of address (e.g., Mr., Ms.), frequent words, punctuation cues, dialogue proportion. In addition, syntactic features are said to be helpful in distinguishing genres of novels. Juatze (2013) [8] reported that literature contains more complex (e.g., subordinating) sentences than chick literature (humorous novels on the challenges of being a modern-day urban female). However, it is worth noting that none of these previous works were done on language other than English. Therefore, whether these prominent stylistic markers shown in previous literature are also effective in differentiating genres and styles of Chinese novels is worth discussing. The work of Yu (2012) [18] pointed out the effectiveness of using function words for Chinese authorship attribution in different genres. It is also noted by Wu (2017) [17] that stylistic features such as average sentence length and vocabulary pattern are able to differentiate authorial styles.

A few studies were carried out on the quantitative connection between writing style and successful literature. Ganjigunte Ashok et al (2013) [3] revealed that there exist distinct linguistic patterns shared among successful literature. It indicated that popular novels use lots of conjunctions, while less popular books use more verbs, adverbs and foreign words. The groundbreaking study then built a model with surprisingly high accuracy (up to 84%) in predicting the success of a novel by using statistical methods.

From the growing body of literature on applying sentiment analysis techniques to the text of fiction, it seems clear that using sentiment analysis for understanding fiction emotion provides another way of analyzing the genres and writing style of fiction. Sentiment Analysis is, at present, widely applied in the areas of product and movie reviews (Hu et al 2004 [7]; Sreejith et al 2017 [14]), whereas for this paper, we have tried to use it in longer texts like novels. Landt (2010) [10] examined English fan-translated version of the demo of the Japanese visual novel to discover the overall tone of the text, including how the tone changed as the story progressed as well as how sentiment was used to portray the various characters. Landt's (2010) study is useful since the overall tone of the top 3 hit novel was examined in the paper to analyze the relationship between the overall tone of fiction and its theme.

According to [16, 19], examining textual content alone is not enough in determining the popularity of a book. It is showed that non-textual information such as sales number, online reviews, and readers' interactive feedback on media platform also play a role in book success. However, previous studies focused mainly on printed books. Little effort has been devoted to combining online rating into determining the success of web novels.

3 Methodology

3.1 Data and Sample Selection

There is a growing base of web fiction offered for free. The novels analyzed in this study were all extracted from a free online novel website, Mirror fiction. The site is one of the best platforms in Taiwan currently offering original stories online. Some of the website top stories have received several millions of views. Established in 2017, Mirror fiction aims to create a platform which enables more creators' works to be officially published, authorized and adapted into published books, films and television works. Over hundreds of categories of fiction are listed on the site, including fan fiction, mystery, romance and thriller…etc. Readers are able to keep, share, and comment on the works.

Our corpus consists of 9 top hit novels of different genres. The rankings used here was based on the information released by Mirror fiction website. Using text mining techniques such as keyword extraction and sentiment analysis, the textual content of 9 web novels were analyzed. Furthermore, non-textual information (number of hits, share, and comment) of 59 web novels was investigated to predict top hit and top shared web novels. In addition, the discriminating features between printed web novels and web novels were also explored.

3.2 Text Analysis Tool and NLP techniques

A web-based text analysis tool, HTML5 Text Analyzer, was utilized in the study to identify stylistic features of web novels. The site provides detailed statistics of your text such as lexical richness, sentence length, function word proportion, dialogue proportion, and punctuation proportion. The prominent features that can most depict the genres, plot and themes of the web novels were chosen and presented in the paper.

NLP techniques are utilized to characterize the writing style of top hit novels as well as its interaction with different genres. Term frequency counts, as a way of keyword extraction, were computed for the top 3 hits and 3 different genres of novels(top 2 hits are extracted in each genre). The term frequency list displays the term frequency counts after removing stop words and unrelated terms (e.g. character names). Another way of extracting keywords is by using TF-IDF. It assumes words with high word frequency in the given documents and low document frequency in the whole collection of documents are of high importance. This way of extracting keywords enables higher discrimination power between documents. For example, common words like "the", "of" and "a", which appears in many novels, will be scaled down. Words that frequently appear in few novels, like "firm", "painted scroll", "dating" for indicating the plot of a particular novel, will be scaled up. Furthermore, the current study attempts to find the tone of these top hit novels by applying sentiment analysis. Python package such as Jieba and SnowNLP were used for processing Chinese text. Jieba help segmentation of the novels' text while SnowNLP analyzes the sentences of a novel and outputs sentiment score that indicates the probability of showing positive emotion.

In [19], the author highlighted the importance of combining readers' online ratings and reviews into determining popular fictions. The numerical statistics on Mirror fiction may reveal something about the success of novels. Therefore, 7 unpublished web novels and 11 published web novels were selected from the website. The number of hits, favorites, shares, and comments of these novels were selected as predictors to predict the factors that make web novels "publishable" by using regression model under statistical software R. Notice that these numerical variables (e.g., number of hits, shares, comments, and favorites) indicated the popularity before the novels have been published into hardcover.

4 Results and Discussions

4.1 Textual properties of top hit novels

What are the lexicon that is most frequently used in top hit novels? The most frequently used words of the top 3 hits were extracted and analyzed. In general, the frequency list is able to depict the genre and setting of the story by only looking at the top 10 frequent terms. For example, *Ghost Mansion*, one of the top hit novels extracted from Mirror fiction website, often uses terms like gong si ‘firm’, gong zuo ‘work’ , nu ren ‘women’ in the story. These frequent terms picture the setting of the novel, which focuses on the social lives and relationships of young professional women. Such kind of settings is often being categorized as “chick literature” which tells the story of the personal growth of a woman or deals with modern issues in women’s lives.

Additionally, tf-idf algorithm was used in this study. In is found that top 10 terms of top 3 hits ranked by tf-idf give a more accurate depiction on the story’s genre and setting. However, it must be noted that since the algorithm gives higher weights to terms that are common in one document but unique among all others documents, character’s name is given higher weights in the tf-idf list. The issue, however, has been resolved by removing character’s names. The algorithm showed a clearer picture of the setting of the novel. In chick literature, terms like xin wen ‘news’, zhu bo ‘anchor’, qi hua ‘marcom’, gong zuo ‘work’ that related to the modern issues in women’s lives appear on the list. On the other hand, expressions in classical Chinese like shi fu ‘master’, gong zhu ‘princess’, ming yue ‘bright moon’ picture the setting of historical novel.

Stylistic features provide another way of characterizing writing style of top hit novels. It is also believed that styles of the novel can be distinguished along certain textual features (Argamon 2006[2]; de Haan 1997[4]; Juatze 2013[8]). Table 1 shows the features that can most depict the themes and topic of top hit novels. Table 2 demonstrates the most discriminating features in different genres. *The Mantra* shows the lowest Simpson’s Index, which means it has the highest lexical richness. Simpson’s D [1](Simpson 1949, as presented in Tweedie & Baayen, 1998[15]) is calculated by:

$$D = \sum_{i=1}^{V} f_v(i,N) \frac{i}{N} \frac{i-1}{N-1}$$

where N refers to the total number of tokens, V to the number of types, and fv(i,N) to the numbers of types occurring i times in a sample of length N. I interpret this to mean that diverse vocabularies are used in *The Mantra* (historical fiction) to help the reader get immersed in the historical events and settings which are farther away from normal people’s real life.

Jodie Archer and Matthew L. Jockers’ *The Bestseller Code: Anatomy of the Blockbuster Novel*(2016)[1] showed that readers of bestsellers liked shorter sentences. It is shown in our data that popular web novels have shorter sentences than others, since the average length of Chinese novel is around 23 words per sentence [17]. Furthermore, interesting findings were revealed, showing that historical novels have the shortest sentence compared to the other two types of novel. This is reasonable since most of the historical novels are written in a mixture of modern vernacular Chinese and written classical Chinese, a traditional style of written Chinese that appears extremely concise and compact compared to modern spoken form of Chinese.

[1] A measure of lexical richness, calculate the frequency of different words in the writings. The smaller the value, the higher the lexical richness

As de Haan (1997)[4] noted, dialogue plays a part in differentiating genres of fiction texts. The proportion of dialogue and narrative will vary depending on the story's setting and genre. Our findings echoed with de Haan's (1997) study in some way. Furthermore, it is also found that dialogue is related to narrative voice(the format through which a story is communicated) of the story. As observed from Table 1, a relatively high proportion of dialogue is used in *My Heart Belongs to You*(Romance), while *Ghost Mansion*(chick literature) showed a low percentage of dialogue. The result could be interpreted in two ways. First, higher proportion of dialogue is used as a strategy in third person narration novel like *My Heart Belongs to You* to clarify the complicate relationship between characters while a first-person viewpoint story such as *Ghost Mansion* requires much less. Second, romance like *My Heart Belongs to You* requires a lot of dialogue because the relationship among the male and female characters is complex while a chick literature such as *Ghost Mansion* involves only characters in workplace.

Function words are said to be effective in distinguishing different writing styles and genres of novels [18]. However, as shown in Table 1, function words are not the distinguishing features in discriminating different writing style. Similarly in Table 2, function words are not able to discriminate historical novels and Romance.

The lexical choice and stylistic features, however, cannot depict the tone and emotion embedded in the story. As some researchers (Landt 2010 [10]; Sreejith et al 2017 [14]) argued, sentiment analysis of literary works is a useful tool in analyzing fiction. It is further highlighted in Landt's (2010) study that more research should be done on the interaction between the genre of text and its overall tone. It is revealed in this study that the overall tone of the text is highly related to its genres and themes. For example, *Ghost Mansion* has a lower overall sentiment score. This is due to the fact that the theme of the story is about the collusion between politicians and real estate tycoons. Although the story is categorized as a chick literature, the major part of the story is exposing the sordid underbelly of modern urban society. *My Heart Belongs to You*(romance) has the highest overall sentiment score among the three books. As a typical romance novel, there might be conflict that hinders the couple's relationship, but romance is still the overriding element in this kind of story. This explains the fact that the novel has the most positive overall sentiment.

4.2 What can numbers reveal about the success of web novels

Apart from the writing style, there are multiple factors that can determine the success of web novels. First, identifying popular tags in different genres of novel reveals readers' preference on the "topic" of the story. Our findings indicated that readers prefer topics on "urban", "workplace", and "modern". As shown in our analysis, the tags "modern", "urban", "workplace" are the most popular tags in chick literature. This result is highly in line with the main idea of chick literature, which often addresses issues of modern womanhood – from romantic relationships to female friendships to matters in the workplace. It is also showed that "time travel" is the most popular tags in historical novels. This is clearly related to the emergence of a novel genre called "alternative history" [13], which the protagonist (mostly women) travels from modern China to ancient China. "Urban" is showed to be the most popular tag in romance fiction writing. Adding "urban" flavor into romance novels make the story close to readers' real world since romance is mainly marketed to middle-class women who live in the urban area (Radway1991 [12]).

Web fiction has becoming a major industry with top writers making millions of dollars and their stories adapted into published books, TV, and movies. Therefore, determining essential elements of "publishable" novels is of importance. First of all, t-test analysis was conducted to evaluate the hypothesis that there is a significant difference in the number of hits, shares, comments, and favorites of unpublished and published web novels. The test was significant only on the number of favorites, t (16) = -2.7079, $p<0.05$. In other words, the number of favorites of published web novel is significantly greater than unpublished web novels. An explanation for this is that readers prefer to add the novel to "favorites" and download it before it is being published into paper book. This is reasonable since once the novel has been published, the reader is unable to view the book for free anymore.

Next, multiple regression analysis was then conducted to discover what factors result in a higher number of shares based on the number of favorites, hits, and comments (Table 3). In the group of unpublished web novels, our model shows that the number of favorites is the only significant predictor to explain the increase in the number of share. As for published web novel, the number of comments is the only predictor among the three variables that is able to explain the increase in the number of shares. We can infer from the result that for unpublished novels, whether it can be shared on social media (Facebook, Instagram, Twitter. . . etc.) depends heavily on the number of favorites from the website. On the other hand, for published web novels, the result is more apparent which shows that more comments lead to more shares on social media.

Finally, another multiple linear regression model was built to determine which variables contribute to the top hit of novels (Table 4). For unpublished web novels, the number of comments is prominent in predicting the number of hits. However, the situation becomes more complicated when it comes to predicting the number of hits in published web novel. There are no variables that can explain the number of hits in published web novels. One possible explanation for this is that once web novel has been published into paper book, multiple factors play in the role of the increase in the number of hits on the website. The factor that can mostly explain the number of hits is beyond the variables investigated in this study.

5 Conclusion

In this paper, the textual and non-textual features of popular web novels written with traditional Chinese have been analyzed. First, from examining the textual content of the most popular novels, it is evident that certain features such as keywords, function words and lexical diversity of the novel are highly related to the genres and writing style of the novel while dialogue proportion reveals something about the narrative mode of the story. Additionally, it is found that shorter sentences are favored by readers on Mirror fiction. The general sentiment in the novel is closely linked to the genre and themes of the story. This result is in line with Landt's (2010) [10] study, although no previous study had dealt with the issue in detail. Finally, the data reveal that the number of favorites and comments serve as significant predictors for the number of shares and hits of unpublished web novels, respectively. However, the number of hits and shares of published web novels is more unpredictable.

The current study makes an attempt to discover how the NLP techniques can help to explain popular web novels in Taiwan . However, since the study involved only "popular novels" but no "less popular novels", the discriminating features between highly popular ones from less popular ones cannot be determined. Another limitation concerns the sample size. Data collected in our research is too small to make a more accurate generalization on the writing styles among different genres. Given the exploratory nature of this study, it is hoped that it can serve as a basis for further study in exploring the secret to popular web novels in Taiwan.

References

1 Jodie Archer and Matthew L. Jockers. *The Bestseller Code: Anatomy of the Blockbuster Novel*. St. Martin's Press, Inc., New York, NY, USA, 2016.

2 Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *Text - Interdisciplinary Journal for the Study of Discourse*, 23:321–346, 2006.

3 V.G. Ashok, S Feng, and Y Choi. Success with style: Using writing style to predict the success of novels. *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1753–1764, 2013.

4 Pieter de Haan. More on the language of dialogue in fiction. *ICAME Journal*, 20, 1997.

5 Helena Montserrat Gomez Adorno, Germán Rios, Juan Pablo Posadas Durán, Grigori Sidorov, and Gerardo Sierra. Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts. *Computación y Sistemas*, 22(1), 2018.

6 David L. Hoover. Frequent Collocations and Authorial Style. *Literary and Linguistic Computing*, 18(3):261–286, 2003. doi:10.1093/llc/18.3.261.

7 Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, New York, NY, USA, 2004. ACM. doi:10.1145/1014052.1014073.

8 Kim Jautze, Corina Koolen, Andreas van Cranenburgh, and Hayco de Jong. From high heels to weed attics: a syntactic investigation of chick lit and literature. In *Proceedings of the Workshop on Computational Linguistics for Literature*, pages 72–81. Association for Computational Linguistics, 2013. URL: http://aclweb.org/anthology/W13-1410.

9 Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. Automatic Detection of Text Genre. In *Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics*, pages 32–38, 1997.

10 Matthias Landt. Sentiment Analysis as a Tool for Understanding Fiction. In *ACM 2010 Annual Meeting*, 2010.

11 Ying Liu and TianJiu Xiao. A Stylistic Analysis for Gu Long's Kung Fu Novels. *Journal of Quantitative Linguistics*, pages 1–30, 2018. doi:10.1080/09296174.2018.1504411.

12 Jeanice A. Radway. *Reading the Romance: Women, Patriarchy, and Popular Literature*. University of North Carolina Press, 1991. URL: http://www.jstor.org/stable/10.5149/9780807898857_radway.

13 Biwu Shang. Unnatural narratives in contemporary Chinese time travel fiction: patterns, values, and interpretive options. *Neohelicon*, 43:7–25, July 2016. doi:10.1007/s11059-016-0327-z.

14 D. Sreejith, M. P. Devika, Naga Santosh Tadikamalla, and Sanju Varghese Mathew. Sentiment Analysis of English Literature using Rasa-Oriented Semantic Ontology. *Indian Journal of Science and Technology*, 10(24), 2017. URL: http://www.indjst.org/index.php/indjst/article/view/96498.

15 Fiona J. Tweedie and R. Harald Baayen. How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, 32(5):323–352, September 1998. doi:10.1023/A:1001749303137.

16 Marc Verboord. Cultural products go online: Comparing the internet and print media on distributions of gender, genre and commercial success. *Communications*, 36(4):441–462, 2011.

17 Chin-Wei Wu. A Linguistic Stylistic Analysis of the Sentences in Wang Wen-hsing's Novel–Backed Against the Sea. *Journal of Chinese Literature of National Cheng Kung University*, 59:181–215, 2017. doi:10.1016/j.dcm.2018.03.003.

18 Bei Yu. Function Words for Chinese Authorship Attribution. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 45–53, Montréal, Canada, June 2012. Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/W12-2506.

19 Burcu Yucesoy, Xindi Wang, Junming Huang, and Albert-László Barabási. Success in books: a big data approach to bestsellers. *EPJ Data Science*, 7(1), April 2018. doi:10.1140/epjds/s13688-018-0135-y.

A Tables

Table 1 Statistics on the top 3 most popular novels.

	Text Length	Simpson's Index	Average Sentence Length	Dialogue Proportion	Function Word Proportion
Ghost Mansion	91084	0.0003	10.9383	0.05841	0.4836
The Mantra	94446	0.00022	7.31742	0.11776	0.4268
My Heart Belongs to You	45586	0.00078	10.4101	0.18222	0.4775

Table 2 Statistics on novels of different genres.

	Text Length	Simpson's Index	Average Sentence Length	Dialogue Proportion	Function Word Proportion
Chick Lit	167875	0.004882	10.96362	0.0700819	0.1007
Historical	160155	0.000044	7.32204	0.1407698	0.4303
Romance	58903	0.000245	9.916329	0.1488718	0.4904

Table 3 Regression results for predicting the number of shares with number the of favorites, hits, and comments.

Coefficients				
	Unpublished Web novel		Published web novel	
Independent variable	t value	p	t value	p
favorites	6.996	**.00599****	0.848	0.199
hits	-1.162	0.32911	1.197	0.27
comments	0.851	0.45742	-1.309	**.026***
significant values: *p<0.05, **p $<$ 0.01, ***p< 0.001	Adjusted R^2 : 0.9257, $p < 0.05*$		Adjusted R^2 : 0.7313, $p < 0.01 * *$	

Table 4 Regression results for predicting the number of hits with the number of favorites, shares, and comments.

Coefficients				
	Unpublished Web novel		Published web novel	
Independent variable	t value	p	t value	p
favorites	1.374	0.26315	0.848	0.424
shares	-1.162	0.32911	1.197	0.27
comments	6.205	**.00844****	-1.309	0.232
significant values: *p<0.05, **p <0.01, ***p<0.001	Adjusted R^2 : 0.9506, p<0.01**		Adjusted R^2: 0.289, p=.1578	