Economics & Business Management
ISSN Print: 3079-5214
ISSN Online: 3079-5222
https://doi.org/10.63313/EBM.9245

A Study on the Impact of Social Media Data Mining from the Douban Platform on Book Topic Planning

Bo Zhang*. Xue Bi
School of Publishing, University of Science and Technology, Shanghai 20082, China
Email: immayday0329@163.com

How to cite this paper: Zhang, B., & Bi, X. (2026). A study on the impact of social media data mining from the Douban platform on book topic planning. Economics & Business Management, 7(3), 38–45. ISSN Print: 3079-5214, ISSN Online: 3079-5222.
https://doi.org/10.63313/EBM.9245
Published: 2026-11-16

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/



Abstract

Against the backdrop of "Internet Plus," this study analyzes reader needs using data from Douban's Top 250 books. Through Python-based data collection, text mining, and statistical analysis, it identifies two themes: individual life narratives and emotional resonance, and social structure and reality reflection. Findings reveal differences in emotional responses across book types, but no significant correlation between immediate emotions and ratings. The study demonstrates the value of user-generated content in informing publishing decisions and book selection.

Keywords

Douban Platform; Reader Demand; Book Topic Selection

1. Introduction

In the "Internet+" era, social media provides real-time insights into reader preferences and market trends. Traditional book selection relies heavily on editorial experience and bestseller lists. As a major Chinese book-review platform, Douban offers extensive user-generated data. However, research on applying such data to publishing remains limited. This study uses Python to collect data from Douban's Top 250 books, analyzing ratings and reviews to identify reader preferences and emotional feedback, providing data-driven insights for book selection.

2. Literature review

2.1. Research related to social media data mining

Data mining identifies potential patterns in large-scale social media data, including user behaviour, interactions, and content expression. Existing studies employ topic modelling, sentiment analysis, and behavioural analysis to explore user interests

and preferences, providing methodological support for reader research. In publishing, Wang Bin et al. highlight the applications of social media data in topic planning, targeted marketing, and user experience evaluation. Zhang Chunlei, Qin Lin et al. demonstrate that reader feedback helps identify content interests and potential topics. Lu Yan, Wang Qian et al. emphasise that browsing, purchasing, reviewing, and sharing data reveal reader preferences and publication performance. Overall, existing research explores the publishing applications of social media data, but systematic analysis of reader reviews requires further investigation.

2.2. Research on Book Topic Planning

Book topic planning involves determining topic directions by integrating reader needs, market trends, and content innovation. Peng Minglan highlights its importance in meeting reader needs, while Jiang Weiwei and Guo Mo emphasise research analysis, competitor studies, and topic innovation, suggesting that topics should balance reliability, innovation, and branding.With AI development, topic planning is increasingly data-driven. Chen Jinhua and Wang Yujia explore the application of "Internet+" and AI in topic planning and content production, while Song Wei and Dong Bing focus on data-driven market positioning and reader demand identification. However, Zhang Yahui highlights ethical risks and technical barriers in AI applications, noting that excessive rationalisation may weaken emotional elements and technical limitations may restrict practical applications. Therefore, data tools should complement rather than replace editorial judgement.

2.3. A Study on How Douban is Used in the Publishing Industry

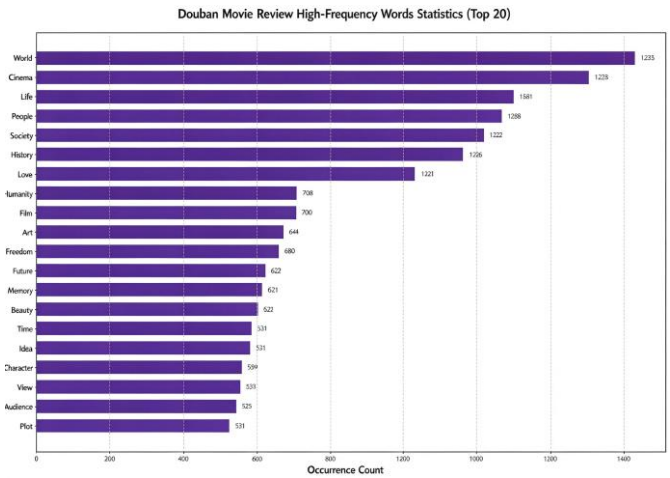
Existing research examines Douban's publishing value in content creation, reading services, and publishing collaboration. Li Xinyu, Gu Xueqiang et al. argue that community and user preference data help identify niche topics and reader needs. Yang Hua, Ge Liang et al. explore Douban Reading's role in submissions, book promotion, and reader interaction. Chen Lei and Shi Yongqin highlight how online self-publishing platforms connect authors with traditional publishers, facilitating quality works' publication. Overall, research recognises Douban's value in reader insights and topic planning, but largely focuses on qualitative applications. Quantitative analysis of large-scale reviews, particularly themes and sentiment, remains limited. Therefore, this study quantitatively analyses short reviews of Douban's Top 250 books. and grammar:

3. Data collection and research methods

This study uses Python to collect data from Douban's Top 250 books, covering 250 books and 23,528 short reviews. After data cleaning, 23,369 valid reviews remained. Methods include word frequency analysis, LDA topic modelling, SnowNLP sentiment analysis, Pearson correlation, and linear regression to identify reader interests,

emotional tendencies, differences across book categories, and relationships between sentiment and book ratings.Overall, readers' concerns center on "individual life narratives and emotional resonance" and "social structure analysis," with sentiment varying significantly across book categories, yet sentiment indicators show no significant correlation with book ratings—so Douban data insights should complement rather than replace editors' professional judgment.

Figure 1. Top 20 Frequent Words in Douban Short Reviews



LDA topic modelling extracts latent themes from word co-occurrence patterns, making it suitable for analysing Douban's colloquial short reviews. Using Python's Gensim library, this study builds an LDA model and evaluates topic coherence across K=2–8, selecting K=2 for better interpretability.The results reveal two main themes: "Individual Life Narratives and Emotional Resonance," involving love, life, fate, and freedom; and "Social Structure Analysis and Reality Reflection," involving society, politics, power, and economy. These findings indicate readers' dual interests in personal experiences and social issues. Accordingly, literary topics should emphasise emotional relationships and life experiences, while social science, history, and nonfiction topics should address real-world issues through valuable analysis.

Table 1. Topic categories and keywords

Topic	Keywords (sorted in descending order of weight)		
Topic 1: Narrative of Individual Life and Emotional Resonance	1. love (0.0029)	2. life (0.0025)	3. women (0.0023)
	4. female (0.0021)	5.suffering (0.0019)	6. existence (0.0018)
	7. fate (0.0018)	8. child (0.0017)	9.men(0.0017)
	10. choice (0.0015)	11.death (0.0015)	12. freedom (0.0014)
	13. nature (0.0013)	14. hardship (0.0012)	15. soul (0.0011)
Topic 2: Analysis of Social Structure and Reflection on Reality	China (0.0090)	2. society (0.0081)	3. politics (0.0041)
	Japan (0.0029)	5. United States (0.0023)	6. theory (0.0020)
	institution (0.0019)	8. male (0.0018)	9. analysis (0.0016)
	logic (0.0016)	11. power (0.0016)	12. economy (0.0015)
	13. research (0.0015)	14. civilization (0.0014)	15. knowledge (0.0013)

4. Sentiment analysis of Douban reviews

Sentiment analysis is used to identify the emotional tendencies in reviews,

helping to judge readers' acceptance of the book content and their emotional feedback. Positive reviews can reveal the content elements that resonate with readers, while negative reviews help identify shortcomings in the reading experience, providing a reference for topic selection, content refinement and expression optimisation.

4.1. Methods of sentiment analysis and sentiment distribution

This article uses SnowNLP to score short reviews, with scores ranging from 0 to 1, where higher values indicate more positive sentiment, and then statistics are compiled on the sentiment distribution of the reviews. After analysing 23,369 valid short reviews, it was found that 75.1% were positive, 23.0% were negative, and 1.9% were neutral, with overall feedback being mainly positive.

Sentiment analysis is used to identify the emotional tendencies in reviews, helping to judge readers' acceptance of the book content and their emotional feedback. Positive reviews can reveal the content elements that resonate with readers, while negative reviews help identify shortcomings in the reading experience, providing a reference for topic selection, content refinement and expression optimisation. This article uses SnowNLP to score short reviews, with scores ranging from 0 to 1, where higher values indicate more positive sentiment, and then statistics are compiled on the sentiment distribution of the reviews. After analysing 23,369 valid short reviews, it was found that 75.1% were positive, 23.0% were negative, and 1.9% were neutral, with overall feedback being mainly positive.

4.2. Analysis of Emotional Differences and Rating Correlations in Different Categories of Books

Simplified content: After completing the holistic feature extraction of Douban Top 250 book reviews, this study will further explore the differential structures reflected in reader feedback. The overall description of Douban short reviews helps us grasp reader consensus but may obscure audience reception and evaluation results for different categories of books. Therefore, the core task here is to examine, through statistical methods, whether there is a significant correlation between readers' emotional tendencies in reviews and the final book ratings for different types of books. This analysis aims to answer two key questions: first, do different types of books trigger varying degrees of emotional resonance or controversy among readers? Second, do emotional differences correspond differently to the book market? To better clarify the types of Douban Top 250 books, this paper combines Douban book tags and manual classification methods, categorising books into eight main types: literature and novels, science fiction and fantasy, social history, psychology and personal growth, mystery and suspense, essays, arts and culture, and natural science. Each category is ensured

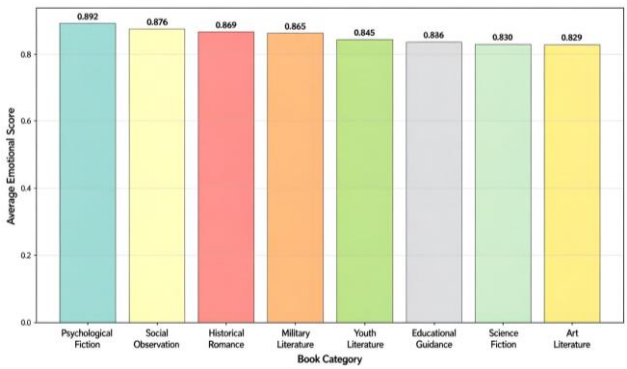
to have a sufficient sample size, with the classification roughly as follows.

Table 2. Topic categories and keywords

Book Title	Category
The Aristocratic Families of the Eastern Jin Dynasty	Social History
Letter from an Unknown Woman	Literary Fiction
The Wisdom of Life	Psychology & Personal Growth
Vegetation of the Human World	Prose & Essays
The Complete Sherlock Holmes (Vols. 1–3)	Mystery & Suspense

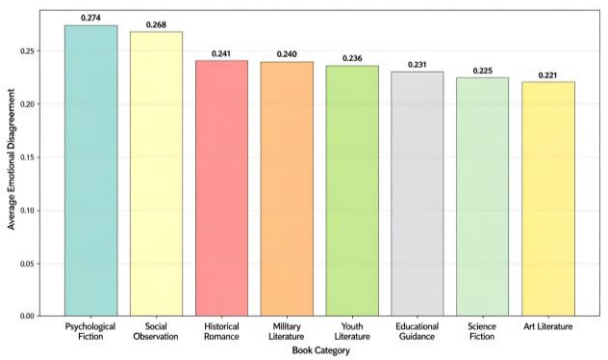
Using SnowNLP, we calculated and compared the average sentiment scores across book categories (Figure 12). Personal growth (0.892) and literary fiction (0.876) ranked highest, suggesting stronger positive emotional resonance. Although all categories scored above 0.83, the 0.06-point gap indicates differences in their emotional appeal.

Figure 2. Comparison of average sentiment scores for each book category



As shown in Figure 3, mystery and sci-fi/fantasy books have the highest divergence, whereas psychological growth books manage to get high emotional scores while keeping low divergence, suggesting their emotional comfort is quite universally effective.

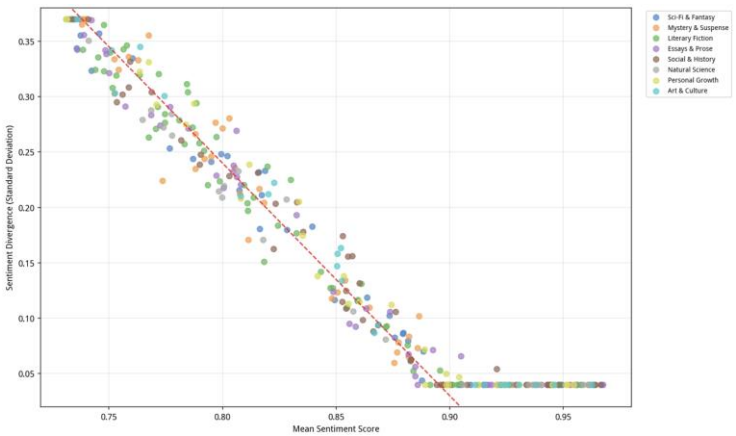
Figure 3. Comparison of emotional differences across book categories



To get a clearer picture of the relationship between the average sentiment scores and sentiment divergence of different book categories, a scatter plot was drawn. As shown in Figure 4, this chart displays the distribution of eight types of books

across two dimensions: 'reader emotional resonance intensity' (average sentiment score) and 'reader review controversiality' (sentiment divergence).

Figure 4. Comparison of emotional differences across book categories



The data reveals a generally negative correlation between average sentiment and divergence: books with stronger positive resonance tend to receive more consistent evaluations. Psychological growth books show high resonance and low controversy, while natural science books receive more varied assessments due to differences in readers' knowledge and values. Literary fiction displays greater divergence, reflecting differences in themes and narrative styles. Therefore, topic planning should adopt a category-based approach: emotional genres should prioritize resonance and broad recognition, while knowledge-based genres should emphasize intellectual depth and diverse perspectives. Successful topic selection requires differentiated positioning and marketing strategies based on genre characteristics.

To examine whether emotional features predict book performance, correlation and regression analyses were conducted using SPSS 27.0.

Pearson correlation showed no significant relationship between Douban ratings and average emotional score ($r=-0.031$, $p=0.629$) or emotional divergence ($r=-0.007$, $p=0.911$). However, these two emotional variables were strongly negatively correlated ($r=-0.830$, $p<0.001$), suggesting that positive reviews are more consistent, while negative reviews generate greater disagreement. Regression results also showed no significant predictive effect of emotional features on ratings (adjusted $R^2=-0.004$, $p=0.584$).

These findings suggest that short-review emotions do not determine established book ratings, which are more closely associated with lasting literary value. Therefore, publishers should combine social media insights with professional editorial judgment rather than rely solely on emotional responses when evaluating book topics.

5. Conclusion and Discussion

This study, through a multi-dimensional analysis of Douban's Top 250 book data, draws the following key conclusions: First, Douban reviews contain a wealth of information about readers' preferences. Word frequency and thematic analysis show that readers focus on two main areas: 'individual life narrative and emotional resonance' and 'social structure analysis and real-world reflection', providing clear guidance for differentiated selection of literary and social science books. Second, sentiment analysis reveals that the emotional reactions triggered by different types of books vary significantly. Books on psychological growth and literary novels are more likely to receive high and consistent positive feelings, while books on natural sciences and social history are more likely to spark controversial discussions. Finally, statistical analysis indicates no significant correlation between the emotional characteristics of readers' short reviews and the books' Douban ratings. This suggests that for classic books, their authority and reputation surpass individual readers' immediate emotional feedback, and publishers should use social media data as a tool for dynamic market insight, rather than the sole criterion for judging a book's classic status.

Reader reviews can reflect market trends and reading preferences, providing important guidance for publishers in understanding readers' needs. However, the immediacy and fragmented nature of short reviews make it difficult to fully assess a book's long-term value and deeper cultural significance, indicating that data insights should be combined with editors' professional judgment. For publishers, a data monitoring system can be established to incorporate reader feedback into the topic selection process, but it is important to avoid over-reliance on data that leads to homogeneous topics, balancing attention to market trends with the independence and creativity of content creation.

References

[1] Yin Zhangchi, Wang Li. Using Social Media User Data to Support Publishing Decisions [J]. Publishing and Distribution Research, 2018, (04): 25-29. DOI:10.19393/j.cnki.cn11-1537/g2.2018.04.006.

[2] Wang Bin, Yang Shulin, Ren Xuelei, et al. Application of Data Mining Technology in Digital Publishing [J]. Journal of Beijing Institute of Graphic Communication, 2016, 24 (06): 61-63. DOI:10.19461/j.cnki.1004-8626.2016.06.015.

[3] Zhang Chunlei. How Data Mining Enhances Children's Book Publishing Planning [J]. Editing Journal, 2020, (01): 117-120. DOI:CNKI:SUN:BJXZ.0.2020-01-031.

[4] Qin Lin. Research on the Application of Data Mining Technology in Book Topic Planning [J]. Qin Zhi, 2022, (02): 78-80. DOI:CNKI:SUN:QINZ.0.2022-02-026.

[5] Lu Yan. Book Marketing Strategies and Trends in the Digital Era [J]. Modern Business Research, 2024, (16): 35-37. DOI:CNKI:SUN:BDSP.0.2024-16-011.

[6] Wang Qian. Research on Optimising Traditional Book Publishing under the Media Integration Background [J]. Journalism Research Guide, 2023, 14 (24): 235-238. DOI:CNKI:SUN:XWDK.0.2023-24-072.

[7] Peng Minglan. Analysis of Innovative Strategies for Book Topic Planning in the "Internet+" Era [J]. China Media Science and Technology, 2022, (03): 71-73. DOI:10.19483/j.cnki.11-4653/n.2022.03.022.

[8] Jiang Weiwei. Exploration of Book Editing Topic Planning and Innovation Methods [J]. Collection, Writing, Editing, 2025, (04): 118-120. DOI:CNKI:SUN:CXBI.0.2025-04-023.

[9] Guo Mo. Common Problems and Countermeasures in Book Topic Planning [J]. Communication and Copyright, 2021, (04): 21-23. DOI:10.16852/j.cnki.45-1390/g2.2021.04.008.

[10] Chen Jinhua. Application of "Internet+" Thinking in Book Topic Planning [J]. China Press, 2024, (06): 134-135. DOI:10.13854/j.cnki.cni.2024.06.029.