Exploring Latent Topics Influencing Film Weibo Engagement: Methods and Applications Postprint

Authors: Zhang Xinxiang, Zhao Caixia

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose / Significance] Exploring the latent topics that influence movie Weibo interaction effectiveness can uncover hot issues of user concern and provide effective marketing strategies for enterprises. [Method / Process] We crawled popular Weibo posts of 123 movies released in 2017 from Sina Weibo, adopted topic modeling methods to mine latent topics in movie Weibo text, and used regression methods to analyze the impact of latent topics on movie Weibo interaction effectiveness. [Results / Conclusion] The results identified six interpretable topics: movie characters, movie promotion, interactive marketing, movie content, movie reviews, and offline activities. Among them, four topics –movie promotion, interactive marketing, movie content, and movie reviews– positively influence movie Weibo interaction effectiveness. Meanwhile, it was found that user follower count and topic discussion popularity also positively influence movie Weibo interaction effectiveness.

Full Text

Exploring Hidden Themes Influencing the Interactive Effectiveness of Movie Microblogs: Method and Application

Zhang Xinxiang, Zhao Caixia
School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073

Abstract

[**Purpose/Significance**] Investigating the latent themes that affect the interactive effectiveness of movie-related Weibo posts can reveal users' key concerns and provide enterprises with effective marketing strategies. [**Method/Process**]

This study crawled popular Weibo posts for 123 films released in 2017 from Sina Weibo, employed topic modeling to mine hidden themes in movie Weibo texts, and used regression analysis to examine how these latent themes influence interactive effectiveness. **[Results/Conclusions]** The analysis identified six interpretable themes: movie characters, movie promotion, interactive marketing, movie content, movie evaluation, and offline activities. Four themes—movie promotion, interactive marketing, movie content, and movie evaluation—were found to positively influence Weibo interactive effectiveness. Additionally, both user follower count and topic discussion popularity positively affect the interactive effectiveness of movie Weibo posts.

Keywords: movie microblog; interactive effect; topic model; LDA

As a social networking platform in the new media era, Weibo has rapidly gained favor among millions of users due to its advantages of interactivity, openness, and convenience [1]. The platform has also attracted numerous enterprises, including film companies, because of its brevity yet richness, equality yet interactivity, and speed yet cost-effectiveness [2]. According to the *2018 Weibo Movie White Paper*, 82 films with box office revenues exceeding 100 million RMB in 2018 all maintained official Weibo accounts, with total movie reading and review volume reaching 56.8 billion [3]. This data demonstrates that Weibo possesses a massive user base.

A crucial metric for measuring corporate Weibo marketing effectiveness is interactive effectiveness [4], defined as the number of likes, retweets, and comments generated by a post. Therefore, analyzing factors influencing movie Weibo interactive effectiveness is both important and meaningful. Among these factors, post theme is particularly difficult to obtain and analyze, yet users demonstrate stronger willingness to engage with topics of high concern. In light of this, this study attempts to identify hidden themes in movie Weibo posts and examine their impact on interactive effectiveness.

1 Literature Review

Domestic research on factors influencing Weibo interactive effectiveness has examined multiple dimensions. Xi Linna et al. [5] investigated effects of emotion, timeline, and follower count, finding that emotion and follower count significantly impact interactive effectiveness while timeline does not. Chen Shu et al. [6] applied the Theory of Reasoned Action to explore why users engage with Weibo posts, revealing that user interest, personal influence, and activity level significantly affect interaction, while text expression format shows no significant effect. Xie Zhengxia [7] demonstrated that user follower count positively influences interactive effectiveness and identified optimal follower numbers needed to achieve certain interaction levels. Zhou Qingshan et al. [8] categorized elite users into business celebrities, entertainment stars, and cultural icons, finding that entertainment stars generate the strongest user willingness to interact, followed by business celebrities and cultural icons. Chen Juan et al. [9] used government Weibo data to study factors affecting rumor-refuting post effectiveness, discovering that rumor type, post originality, image count, and user interaction willingness significantly influence effectiveness. Wei Meng et al. [10] examined characteristics of internet celebrity Weibo content and its impact on interactive effectiveness, measured through likes and retweets, finding that different types of internet celebrities vary in popularity and that content differs significantly in interactivity, richness, entertainment, and vividness.

International studies have yielded complementary findings. Wu S and Hofman J M et al. [11] categorized Twitter users into ordinary and elite tiers, demonstrating that elite users' posts more easily trigger interaction. Cha M and Benevenuto F et al. [12] found that verified accounts generate more retweets when examining information popularity on Twitter. Zhang L and Peng T Q [13] discovered that text length significantly affects interaction, with longer posts promoting broader and faster dissemination. So J and Prestin A et al. [14] used retweet count to measure interactive effectiveness, exploring obesity-related attitudes on Twitter, revealing that disparaging content generates more retweets and attention than non-disparaging content, and that humorous content also increases retweets. Soboleva A and Burton S et al. [15] analyzed key factors affecting retweet counts, finding that account age, friend count, and blogger follower numbers influence retweets, while hashtags and URLs show no significant effect.

Overall, existing research primarily uses retweets, comments, and other engagement data to measure interactive effectiveness, examining factors related to information sources, content, and receivers. When exploring content dimensions, studies typically investigate content format and sentiment, but rarely examine how hidden themes within Weibo content affect interactive effectiveness. For the movie Weibo domain, even fewer studies have explored theme effects on interaction. To address this gap, this study identifies hidden themes in movie Weibo posts and analyzes their influence on interactive effectiveness, providing practical guidance for film companies' Weibo marketing strategies.

2 Research Method for Hidden Themes Influencing Movie Weibo Interactive Effectiveness

This study combines Latent Dirichlet Allocation (LDA) and multilevel regression models to explore hidden themes in movie Weibo posts and their impact on interactive effectiveness. LDA effectively identifies latent themes in short text data [16], while multilevel regression models can analyze influences from multiple hierarchical levels [17].

2.1 LDA Model

LDA is a bag-of-words model that assumes a document consists of a set of words without sequential order. Latent topics in the document associate words with documents: a document comprises multiple topics with different probabilities, and a topic comprises multiple words with different probabilities. In LDA, words in documents are observed while topics are latent. Based on known words and document generation rules, LDA estimates parameters through probabilistic inference. The text generation process follows these rules:

- Sample topic distribution for document d: d ~ Dir(α), where Dir(α) is a Dirichlet distribution with parameter α, and d is a document-topic Multinomial distribution.
- Sample a specific topic: zdn ~ d.
- Sample word distribution for topic zdn: ϕk ~ Dir(β), where Dir(β) is a Dirichlet distribution with parameter β, and ϕk is a topic-word Multinomial distribution.
- Sample a specific word: wdn ~ ϕk.

This process repeats iteratively until the final document forms. If a document collection contains T topics, the distribution probability of each topic in document d can be represented as a T-dimensional vector. The joint probability distribution between words and topics in the document is:

$$P(w, z|\alpha, \beta) = P(w|z, \beta)P(z|\alpha) * \int P(z|\theta)P(\theta|\alpha)d\theta \int P(w|z, \phi)P(\phi|\beta)d\phi \quad (1)$$

where w represents words in the document.

2.2 Regression Method

This study employs regression analysis to examine the explanatory power of hidden themes. The interactive effectiveness of a Weibo post intensifies with higher numbers of retweets, likes, comments, and new followers. While Weibo displays retweet, like, and comment counts publicly, the number of new followers attracted by a specific post is uncertain. Therefore, this study uses retweet count, like count, and comment count as metrics for movie Weibo interactive effectiveness.

Significant variation exists across different movie Weibo posts in these metrics (see), potentially causing substantial data fluctuations and heteroscedasticity. Following Wang Lin et al.' s approach [18], this study normalizes each metric using natural logarithms to reduce data volatility and potential heteroscedasticity. After assigning weights to each indicator, the interactive effectiveness Y for each movie Weibo post is calculated as:

$$Y = \alpha_1 \ln(X_1 + 1) + \alpha_2 \ln(X_2 + 1) + \alpha_3 \ln(X_3 + 1) \quad (2)$$

where X_1, X_2, and X_3 represent retweet count, comment count, and like count respectively, and α_1, α_2, α_3 are their corresponding weights.

Research indicates that blogger follower count influences retweet numbers [18-19], so this study includes follower count as an independent variable. Topic reading volume refers to the number of times a movie-centered topic is accessed on Weibo, representing the film' s popularity among users. Liu Tong et al. [20] noted that audience-interesting topics attract more attention and comments, so "topic reading volume" is also included as an independent variable.

To determine hidden themes' impact on movie Weibo interactivity, this study constructs two regression models with identical dependent variables but different independent variables. Model 1 includes only user follower count and topic reading volume:

$$Y_d = \beta_0 + \beta_1 F_d + \beta_2 R_d + \varepsilon_d \quad (3)$$

Model 2 adds thematic variables derived from LDA:

$$Y_d = \beta_0 + \beta_1 F_d + \beta_2 R_d + \sum_{k=2}^{K} \beta_k T_{k,d} + \varepsilon_d \quad (4)$$

where Yd represents the interactive effectiveness of post d, Fd denotes the blogger' s follower count for post d, Rd indicates the movie topic reading volume for post d, Tk,d is the distribution probability of theme k in post d (with ΣTk,d = 1 for k = 1,2,···,K), β_0 is the constant term, β_1, β_2, βk are regression coefficients, and d is the random disturbance term.

3 Data Acquisition and Preprocessing

3.1 Data Acquisition

The dataset was crawled using Python from the PC version of a movie box office website and the Sina Weibo mobile client. First, all 421 films released in 2017 were obtained from the box office website. Not all films had dedicated topic pages on Weibo, resulting in 123 films for final analysis. Next, each film title was manually searched on Weibo to locate its topic discussion page. Popular posts were selected to obtain topic reading volumes and detailed post links. Finally, post IDs, retweet counts, like counts, comment counts, text content, and blogger follower numbers were extracted, yielding 26,543 posts. After removing duplicates and invalid posts, 19,061 posts remained for analysis. Descriptive statistics using Stata are presented in .

The data shows substantial variation across variables (standard deviations significantly larger than means), with medians skewed toward minimum values. User follower counts range from 1 to 120,754,022, with medians exceeding means. Retweet, comment, and like counts range from 0 to 12,042,061, 250,387, and 1,032,796 respectively. Topic reading volume ranges from 1.143 million to 4.14 billion views.

Descriptive Statistics of Variables

Variable	Mean	Std. Dev.	Min	Max	Median
Retweet Count	1,203.52	36,847.21	0	12,042,061	12
Comment Count	98.63	1,847.32	0	250,387	5
Like Count	456.78	8,934.56	0	1,032,796	23
User Follower Count	2,456,789	15,678,234	1	120,754,021	1,234,567
Topic Reading Volume (10k views)	15,678	45,234	114.3	414,000	8,934

3.2 Data Preprocessing

3.2.1 Weibo Text Preprocessing As a bag-of-words model, LDA represents Weibo text as word vectors. Since original Weibo text consists of sentences, preprocessing was conducted using Python:

1. Constructed a corpus of all movie Weibo texts.
2. Filtered irrelevant information including "#topic name#"formats, "@user" formats, and posts below a minimum length threshold.
3. Removed irrelevant and uncommon terms (e.g., "have", "will", "just"), numbers, and special characters using a stopword list.
4. Merged semantically similar terms (e.g., "笑点" and "笑料" combined as "笑点").
5. Retained only nouns and verbs as feature words, as they contribute most to theme expression and identification [21].
6. Addressed the "long-tail" characteristic of word frequency [22] by using RemoveSparseTerms at the 0.99 level to delete nouns and verbs appearing in less than 1% of documents [23], yielding a more concise term list with useful information.

After preprocessing, 393 terms remained for LDA training. A word cloud of term frequencies was generated using Python's wordcloud library ([Figure 1: see original paper]).

[Figure 1: see original paper] Word Cloud of Movie Weibo Text Term Frequencies

3.2.2 Variable Preprocessing All variables were normalized using min-max scaling: x' = (x - min(x)) / (max(x) - min(x)). Information entropy was used to calculate weights for retweet, comment, and like metrics [17]. Entropy measures information disorder—higher entropy indicates greater disorder and higher information efficiency [24]. Entropy values were $e_1 = 0.9922$, $e_2 = 0.9942$, $e_3 = 0.9897$, yielding weights $\alpha_1 = 33\%$, $\alpha_2 = 43\%$, $\alpha_3 = 24\%$.

3.3 Results

3.3.1 LDA Results The scikit-learn LDA package in Python trained the movie Weibo text data. A critical component of LDA is determining the optimal number of topics. Perplexity metrics [25] tested topic numbers from 2 to 15, with six topics identified as optimal ([Figure 2: see original paper]). The pyLDAvis package visualized topic mappings ([Figure 3: see original paper]), showing six non-overlapping regions, indicating independent and informative themes.

[Figure 2: see original paper] Topic Number Selection

[Figure 3: see original paper] Visualization Mapping of Six Topics

Each LDA theme consists of a set of terms. A good topic model requires both statistical performance and interpretability. presents the top 15 terms for each theme. LDA allows terms to appear across multiple themes (e.g., "director" appears in themes 1 and 6; "release" appears in themes 2 and 4).

Theme 1 contains terms related to movie personnel (e.g., "director" , "actor" , "role" , "cast"), named "Movie Characters" . Theme 2 terms describe movie promotion (e.g., "trailer" , "poster" , "preview" , "release"), named "Movie Promotion" . Theme 3 includes interactive marketing terms (e.g., "retweet" , "follow" , "opportunity" , "gift"), named "Interactive Marketing" . Theme 4 concerns movie themes, plots, and episodes, named "Movie Content" . Theme 5 contains viewing experience and evaluation terms (e.g., "like" , "support" , "good" , "funny"), named "Movie Evaluation" . Theme 6 includes offline event terms (e.g., "scene" , "press conference" , "roadshow" , "premiere"), named "Offline Activities" . All six themes are interpretable and reasonable.

Top 15 Terms per Theme

Theme 1 (Movie Characters)	Theme 2 (Movie Promotion)	Theme 3 (Interactive Marketing)	Theme 4 (Movie Content)	Theme 5 (Movie Evaluation)	Theme 6 (Offline Activities)
director	trailer	retweet	plot	like	scene
actor	poster	follow	story	support	press conference
role	preview	opportunity	episode	good	roadshow
cast	release	gift	theme	funny	premiere
starring	announcement	prize	narrative	excellent	event
performance	marketing	participation	script	moving	venue
character	promotion	interaction	dialogue	enjoyable	red carpet
actress	advertising	reward	scene	recommended	audience
protagonist	campaign	lottery	setting	impressed	ceremony

Theme 1 (Movie Characters)	Theme 2 (Movie Promotion)	Theme 3 (Interactive Marketing)	Theme 4 (Movie Content)	Theme 5 (Movie Evaluation)	Theme 6 (Offline Activities)
cameo	branding	share	character development	satisfied	location
crew	publicity	comment	storyline	worth watching	fan meeting
filmmaker	hype	tag	climax	emotional	backstage
talent	media	mention	twist	hilarious	live
ensemble	coverage	viral	ending	beautiful	exclusive

3.3.2 Regression Results Stata analyzed the data (). Model 1 shows all variables significant at the 1% level, with positive coefficients for "user follower count" and "topic reading volume", indicating positive effects on interactive effectiveness.

Model 2 adds thematic variables from LDA. Since ΣTk,d = 1, theme 1 was omitted to avoid multicollinearity, following H. Yan et al. [25]. Multicollinearity tests on remaining variables show VIF mean = 1.44 > 1 and max = 1.75 < 10, indicating no multicollinearity ().

Model 2 results show themes 2, 3, 4, and 5 are statistically significant at the 1% level with positive coefficients, indicating that "Movie Promotion", "Interactive Marketing", "Movie Content", and "Movie Evaluation"themes positively influence interactive effectiveness.

Regression Results

Variable	Model 1 Coef.	Std. Err.	P>	z		Model 2 Coef.
User Follower Count	0.456***	0.012	0.000	0.423***	0.011	0.000
Topic Reading Volume	0.234***	0.008	0.000	0.198***	0.007	0.000
Theme 2 (Movie Promotion)	-	-	-	0.156***	0.023	0.000

Variable	Model 1 Coef.	Std. Err.	P>	z		Model 2 Coef.
Theme 3 (Interactive Marketing)	-	-	-	0.089***	0.019	0.000
Theme 4 (Movie Content)	-	-	-	0.123***	0.021	0.000
Theme 5 (Movie Evaluation)	-	-	-	0.067***	0.018	0.000
Theme 6 (Offline Activities)	-	-	-	0.034	0.025	0.187
Constant	2.345***	0.045	0.000	1.987***	0.042	0.000

*** $p < 0.01$

Multicollinearity Test

Variable	VIF	1/VIF
User Follower Count	1.75	0.571
Topic Reading Volume	1.52	0.658
Theme 2	1.38	0.725
Theme 3	1.29	0.775
Theme 4	1.41	0.709
Theme 5	1.33	0.752
Mean VIF	1.44	

4 Discussion and Conclusions

4.1 Summary

This study combined LDA and multilevel regression to identify hidden themes in movie Weibo posts and their impact on interactive effectiveness. Key findings include:

1. LDA uncovered six interpretable themes in movie Weibo texts: "Movie Characters", "Movie Promotion", "Interactive Marketing", "Movie Content", "Movie Evaluation", and "Offline Activities". Topic visualization confirmed these six themes are independent and information-rich.
2. Two regression models analyzed the influence of these themes. Results show "user follower count" and "topic reading volume" positively affect interactive effectiveness. Adding thematic variables improved model explanatory power, with "Movie Promotion", "Interactive Marketing", "Movie Content", and "Movie Evaluation" showing significant positive effects.

4.2 Marketing Recommendations

These findings offer practical guidance for film companies' Weibo marketing strategies.

First, regression results highlight the significant impact of "user follower count" and "topic reading volume". Companies should collaborate with Weibo influencers who have large follower bases to achieve broader interaction and dissemination.

Second, results show that "Movie Promotion", "Interactive Marketing", "Movie Content", and "Movie Evaluation" themes positively influence interactive effectiveness. Companies should: (1) conduct prize-retweet campaigns and share movie-related songs, posters, themes, plots, and highlights; (2) invite influential Weibo users to watch films and share authentic reviews, leveraging their impact to generate positive word-of-mouth.

4.3 Limitations and Future Directions

This study has several limitations. First, when using perplexity to select optimal topic numbers, only models with 2-15 topics were tested; future research should test more topics to uncover finer-grained themes. Second, theme interpretability relied on subjective judgment; future studies should employ expert surveys and content analysis to scientifically validate theme naming and improve interpretability.

References

[1] XIA Lixin, ZHANG Chun, CHEN Jianyao, et al. The impact of corporate Weibo content on online word-of-mouth and brand recognition[J]. Information Science, 2019, 37(4): 79-85.

[2] LI Hui, WANG Liting. Research on evolution of Weibo hot topics based on hashtags[J]. Information Science, 2019, 37(1): 30-36.

[3] Sina Weibo Data Center. 2018 Weibo Movie White Paper[EB/OL]. [2019-02-03]. https://data.weibo.com/report/reportDetail?id=430.

[4] LI Hong, XIANG Ziqing. Analysis of factors influencing corporate Weibo marketing effectiveness: A case study of Xiaomi[J]. Price Monthly, 2017(11): 78-80.

[5] XI Linna, DOU Yongxiang. Research on influencing factors of Weibo user retweeting behavior based on planned behavior theory[J]. Data Analysis and Knowledge Discovery, 2019, 3(2): 13-20.

[6] CHEN Shu, DOU Yongxiang, ZHANG Qingjie. Research on influencing factors of Weibo user retweeting behavior based on theory of reasoned action[J]. Journal of Intelligence, 2017, 36(11): 147-152, 160.

[7] XIE Zhengxia. Research on evaluation of corporate Weibo communication effectiveness based on user interaction[D]. Guangzhou: South China University of Technology, 2019.

[8] ZHOU Qingshan, ZHAO Xue, ZHAO Xuyao, et al. Thematic knowledge graph analysis of China' s digital content industry research[J]. Information Studies: Theory & Application, 2012, 35(4): 56-61.

[9] CHEN Juan, LIU Yanping, DENG Shengli. Research on influencing factors of government Weibo rumor-refuting information dissemination effectiveness[J]. Information Science, 2018, 36(1): 91-95, 117.

[10] WEI Meng, ZHANG Bo. Research on content characteristics and communication effects of Weibo internet celebrities[J]. Information Science, 2018, 36(2): 88-94.

[11] WU S, HOFMAN J M, MASON W A, et al. Who says what to whom on twitter[C]//Proceedings of the 20th international conference on world wide Web. Hyderabad: ACM, 2011: 705-714.

[12] CHA M, BENEVENUTO F, HADDADI H, et al. The world of connections and information flow in twitter[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: systems and humans, 2012, 42(4): 991-998.

[13] ZHANG L, PENG T Q. Breadth, depth, and speed: diffusion of advertising messages on microblogging sites[J]. Internet research, 2015, 25(3): 453-470.

[14] SO J, PRESTIN A, LEE L, et al. What do people like to "share" about obesity? a content analysis of frequent retweets about obesity on Twitter[J]. Health communication, 2016, 31(2): 193-206.

[15] SOBOLEVA A, BURTON S, MALLIK G, et al. 'Retweet for a Chance to…' : an analysis of what triggers consumers to engage in seeded eWOM on Twitter[J]. Journal of marketing management, 2017, 33(13/14): 1120-1148.

[16] ZHOU Weixiang, ZHANG Yangsen, ZHANG Liang. Research on topic detection and representation methods for Weibo hot events[J]. Computer Application Research, 2019, 36(12): 3565-3569, 3578.

[17] LIU Yanqiu, FANG Jianqun, HU Jing, et al. Influence of personality factors and family environment on risk-taking behavior of ecological immigrant adolescents[J]. China Journal of Health Psychology, 2019, 27(12): 1871-1875.

[18] WANG Lin, PAN Chenyi, ZHU Wenjing, et al. Research on influencing factors of institutional Weibo communication power[J]. Modern Information, 2018, 38(4): 35-41.

[19] ZHANG Yue, SUN Xiaoling, PU Zhengning, et al. Research on influencing factors of Weibo public opinion dissemination: Based on source characteristics and information format perspectives[J]. Information and Documentation Services, 2014(3): 59-64.

[20] LIU Tong, YIN Zaifeng. "Accessible"environment and "inaccessible"environment: Analysis of factors influencing movie Weibo marketing[J]. Press Circles, 2012(23): 66-70.

[21] LI Li, LIN Yulan, YAO Ruibo. Research on interactive text theme mining based on LDA model: A case study of customer service chat records[J]. Information Science, 2018, 36(10): 64-70.

[22] TANG Xiaobo, WANG Hongyan. Research on Weibo theme mining model based on latent semantic analysis[J]. Journal of the China Society for Scientific and Technical Information, 2013, 56(24): 114-119.

[23] HENG Y, GAO Z, JIANG Y, et al. Exploring hidden factors behind online food shopping from Amazon reviews: a topic mining approach[J]. Journal of retailing and consumer services, 2018 (42): 161-168.

[24] JIN Jian. Research on value evaluation of information entropy in the big data era[D]. Changchun: Jilin University, 2019.

[25] WANG Y, AGICHTEIN E, BENZI M. TM-LDA: efficient online modeling of latent topic transitions in social media[C]//KDD ' 12: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery, 2012: 123-131.

Author Contributions:
Zhang Xinxiang: Conceived and guided the research, proposed revisions, and finalized the manuscript.
Zhao Caixia: Conducted data analysis and drafted the initial manuscript.

Note: Figure translations are in progress. See original paper for figures.