Original Paper

Data Analysis and Visualization of Newspaper Articles on Thirdhand Smoke: A Topic Modeling Approach

Qian Liu[1,2], BTech, MSc, PhD; Qiuyi Chen[1], BA; Jiayi Shen[3], MBBS; Huailiang Wu[3], MBBS; Yimeng Sun[1], BA; Wai-Kit Ming[4,5,6,7], MD, MPH, MMSCI, PhD

[1]School of Journalism and Communication, Jinan University, Guangzhou, China

[2]National Media Experimental Teaching Demonstration Center, Jinan University, Guangzhou, China

[3]School of Medicine, Jinan University, Guangzhou, China

[4]Department of Obstetrics & Gynaecology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

[5]Harvard Medical School, Harvard University, Boston, MA, United States

[6]Faculty of Medicine, The Chinese University of Hong Kong, Shatin, NT, China (Hong Kong)

[7]Li Ka Shing Faculty of Medicine, University of Hong Kong, Pokfulam, China (Hong Kong)

Corresponding Author:
Wai-Kit Ming, MD, MPH, MMSCI, PhD
Department of Obstetrics & Gynaecology
The First Affiliated Hospital, Sun Yat-sen University
58 Zhongshan 2nd Rd
Guangzhou,
China
Phone: 86 14715485116
Email: wkming@connect.hku.hk

Abstract

Background: Thirdhand smoke has been a growing topic for years in China. Thirdhand smoke (THS) consists of residual tobacco smoke pollutants that remain on surfaces and in dust. These pollutants are re-emitted as a gas or react with oxidants and other compounds in the environment to yield secondary pollutants.

Objective: Collecting media reports on THS from major media outlets and analyzing this subject using topic modeling can facilitate a better understanding of the role that the media plays in communicating this health issue to the public.

Methods: The data were retrieved from the Wiser and Factiva news databases. A preliminary investigation focused on articles dated between January 1, 2013, and December 31, 2017. Use of Latent Dirichlet Allocation yielded the top 10 topics about THS. The use of the modified LDAvis tool enabled an overall view of the topic model, which visualizes different topics as circles. Multidimensional scaling was used to represent the intertopic distances on a two-dimensional plane.

Results: We found 745 articles dated between January 1, 2013, and December 31, 2017. The United States ranked first in terms of publications (152 articles on THS from 2013-2017). We found 279 news reports about THS from the Chinese media over the same period and 363 news reports from the United States. Given our analysis of the percentage of news related to THS in China, Topic 1 (Cancer) was the most popular among the topics and was mentioned in 31.9% of all news stories. Topic 2 (Control of quitting smoking) was related to roughly 15% of news items on THS.

Conclusions: Data analysis and the visualization of news articles can generate useful information. Our study shows that topic modeling can offer insights into understanding news reports related to THS. This analysis of media trends indicated that related diseases, air and particulate matter ($PM_{2.5}$), and control and restrictions are the major concerns of the Chinese media reporting on THS. The Chinese press still needs to consider fuller reports on THS based on scientific evidence and with less focus on sensational headlines. We recommend that additional studies be conducted related to sentiment analysis of news data to verify and measure the influence of THS-related topics.

(JMIR Med Inform 2019;7(1):e12414) doi: 10.2196/12414

KEYWORDS

media concerns; topic modeling; third-hand smoke; tobacco; indoor air quality

Introduction

Thirdhand smoke (THS) is an important public health issue and has been an increasingly popular topic for decades in China since its first mention in 2009 [1]. Aggregating media reports from major media outlets and analyzing the coverage using topic modeling may help shed light on the role that the media plays in communicating this health concept. Thirdhand smoke consists of residual tobacco smoke pollutants that remain on surfaces and in dust after tobacco has been smoked. These pollutants are re-emitted into the gas phase (ie, off-gassing) or react with oxidants and other compounds in the environment to yield secondary pollutants [2]. Evidence supports the presence of THS in indoor environments. Thirdhand smoke is found in enclosed spaces where habitual smoking occurs, such as residences and automobiles [3]. This phenomenon is associated with health hazards: research has shown that residual nicotine from tobacco smoke absorbed onto indoor surfaces reacts with ambient nitrous acid to form carcinogenic tobacco-specific nitrosamines, which can cause significant levels of DNA damage in human cell lines [3]. The related acute and long-term risks of THS include disease and premature mortality. Children and infants are particularly susceptible to THS exposure [4].

Given the increasing interest in THS, the mass media has focused on delivering and communicating information on this topic to public audiences. Both in China and abroad, there have been media reports related to THS. According to search results of news articles mentioning THS, the United States ranks first in terms of topic mentions. However, previous studies have shown that fewer people are aware of the harms of THS than that of secondhand smoke [5]. Because the mass media is a key player in communicating health-related information, it could play a positive role in helping the public understand the risks of THS and the ways to protect themselves from it [6]. We therefore decided to compare this topic between China and the United States.

Multimodal data modeling combines information from different resources. Topic modeling refers to statistical models in which unstructured data are structured in accordance with latent themes to deal with multimodal data. Latent Dirichlet Allocation (LDA) is the most popular form of topic modeling and is a generative probabilistic modeling method for converting visual words into images and visual word documents [7-9]. Topic modeling has broad applications in various fields such as text mining [10], medicine [11-13], economics [14], and social network analysis [15]. To the best of our knowledge, however, there have been very few studies using LDA to evaluate the media's treatment of THS. As a result, we used LDA modeling method for our analysis despite its' being a common method. This paper aims to determine the current patterns and the role of mass communication related to THS.

Methods

Data Collection

We used the Wiser database for Chinese news content (from the Wise News website) and the Factiva database (from the Dow Jones website) to retrieve the international news articles. The Wiser database is an ever-growing, professional Chinese media content database that contains more than one million data entries. It is currently the best source for Chinese media content given its large volume of data and timely updates. Factiva is an international news and information database that includes nearly 33,000 premium sources such as licensed publications, influential websites, blogs, images, and videos. To get a general idea of the topic of THS, we conducted a preliminary investigation of the Factiva database and retrieved articles dated between January 1, 2013, and December 31, 2017. The Wiser database includes data published only within the last five years. However, many of the articles we flagged merely noted the topic of smoking without further elaborating on the issue of THS. We next narrowed down the entries to only those newspaper sources pertaining to THS; we believe that articles from newspapers are much more reliable than those from other sources. We used LDA to further analyze the Chinese news articles.

An LDA topic model is a model with a three-level hierarchical Bayesian model. The basic assumption of this model is a combination of words belonging to different topics [16]. LDA suggests that there may be multiple topics in an article and that the wording in that article or paper reflects the exact set of topics that the reporter wished to address. Using Gibbs Sampling techniques, a method that estimates the marginal distribution of interested variables, we can determine the topics among the data pool [17].

Processing

Data processing (Figure 1) was conducted before applying LDA modeling by using Python to do data cleaning and the Python package, Jieba, to do the segmentation [18,19]. First, the redundant and null data were chosen to be removed, followed by removal of irrelevant information such as advertisements. Next, word segmentation was conducted using the Jieba package. However, THS articles with a lot of terminology were calculated and added based on the calculation of the words' information entropy and term frequency to avoid the influence of unprofessional dictionaries. New terms such as "China Anticancer Association," "Family Doctor," "Shenzhen Municipal Government," "Air Circulation," "Air Purification," "Chinese Preventive Medicine Association," "Tobacco Monopoly," "Tobacco Market," and so on, were added to the dictionary for further analysis. Furthermore, common Chinese stop words were removed such as "a," "of", "it," etc. Next, a document-term matrix was built, and term frequency–inverse document frequency (TF-IDF) was used in the data processing.

Multiple studies have been conducted related to the choice of LDA topic number and the explanations of each topic. In previous research related to topic number, 10-30 topics were assigned nearly the same log-likelihood measure. Therefore, we adopted the number 10 for the topic parameter. By analyzing key words, topic content was generated accordingly. If we considered only the statistical measures, the results might not be interpretable by humans [20]. Therefore, by combining both statistical measures and manual interpretation, we selected 10 topics to analyze by using Python version 3.6.1 with the LDAvis tool [16]. We set λ=1 and recovered 10 topics and their keywords (Table 1). To elaborate on the topics, the topic name was generated based on the given keywords (Table 1) as well.

We sought topics that overlapped as in the visualization shown in Figures 2 and 3. In this two-dimensional plane, topics are represented as cycles whose centers are determined by the computed topics' distances [16]. We categorized the topics into three main primary groups: links to related diseases, air and particulate matter ($PM_{2.5}$), and control and restrictions (Table 1).

Figure 1. Data processing chart. LDA: Latent Dirichlet Allocation; TF-IDF: frequency–inverse document frequency.

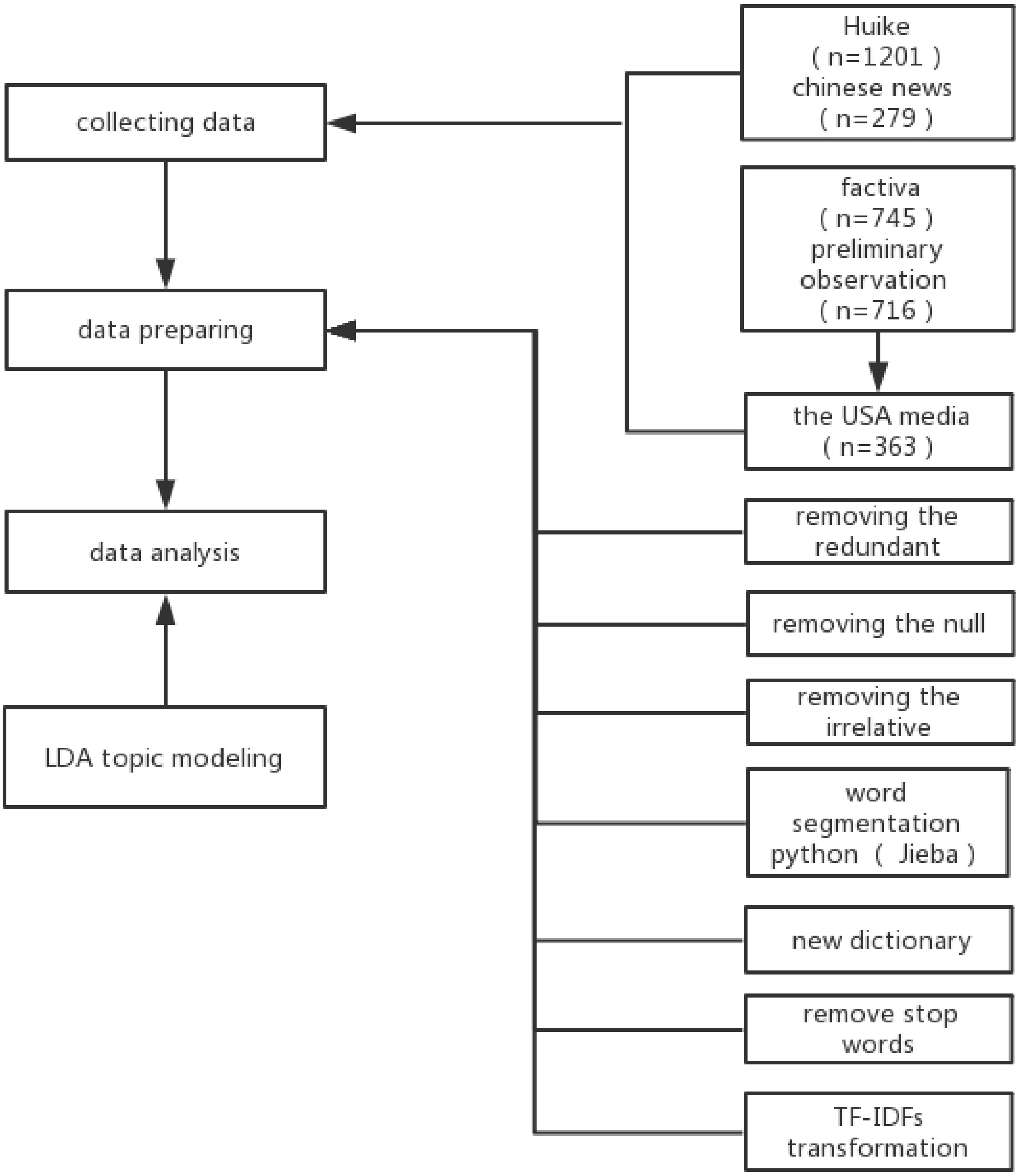


Table 1. Topic classification and keywords.

Classification group and topic name	Key words	Percentage (%)
Related diseases		
Topic 1: cancer	Lung cancer, cancer, tumor, treatment, patient, risk factor, pollution, professor, China, population, prevalence, diseases pattern, prevention of air pollution	31.9
Topic 5: risks of smoking	Smoker, movement, body, nicotine, content, quitting smoking, experts, symptom related to dead smokers	10.1
Topic 7: diseases induced by smoking	Asthma, citizen, hospital, doctor, patient, treatment, smoker, time, long-term, breath, chairman, cause	4.9
Topic 3: susceptible population	Children, research, food, contact, cause, influence, environment, increase, body, clothes, smog, content, reveal, professor, indoor, smoker, female	11.4
Topic 4: quitting smoking	Quit smoking, smoker, smoke, hospital, drug, breath, doctor, work, smoker, content, one kind, introduction, treatment, chairman	11.1
Topic 8: relevant research	Introduction, reveal, cigarette, children, smoke, officer, smoker, increase, factor, place, reason, environment, relevant	3.2
Air and PM$_{2.5}$		
Topic 6: air quality	PM$_{2.5}$, indoor, concentration, severe, microgram, air, pollution, smog, influence, kitchen, cooking fume	9.7
Control and restrictions		
Topic 9: classic smoking control case	Shenzhen, tobacco control, activity, citizen, place, rule, investigation, smoker, work, over, condition, patients, quit smoking, control, indoor, public place, increase	2.6
Topic 10: public control	Public place, quit smoking, ban tobacco, rule, place, professor children, indoor, China, tobacco control, body, influence, body, female worker, reveal	0.1
Topic 2: control of quitting smoking	Quit smoking, ban tobacco, third-hand smoke, public place, place, rule, ban, indoor, control, work, country, smoke, society, China, Beijing, relevant, smoker, outdoor	15

Figure 2. Intertopic distance map.

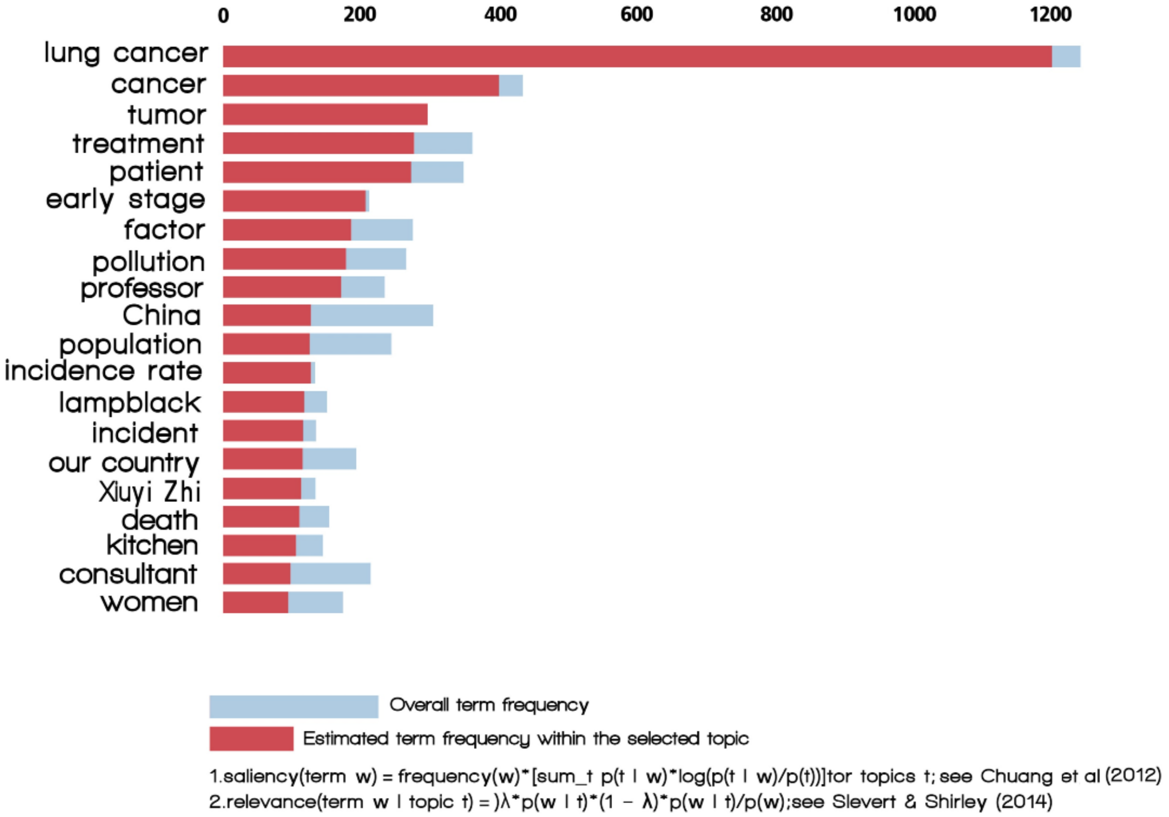
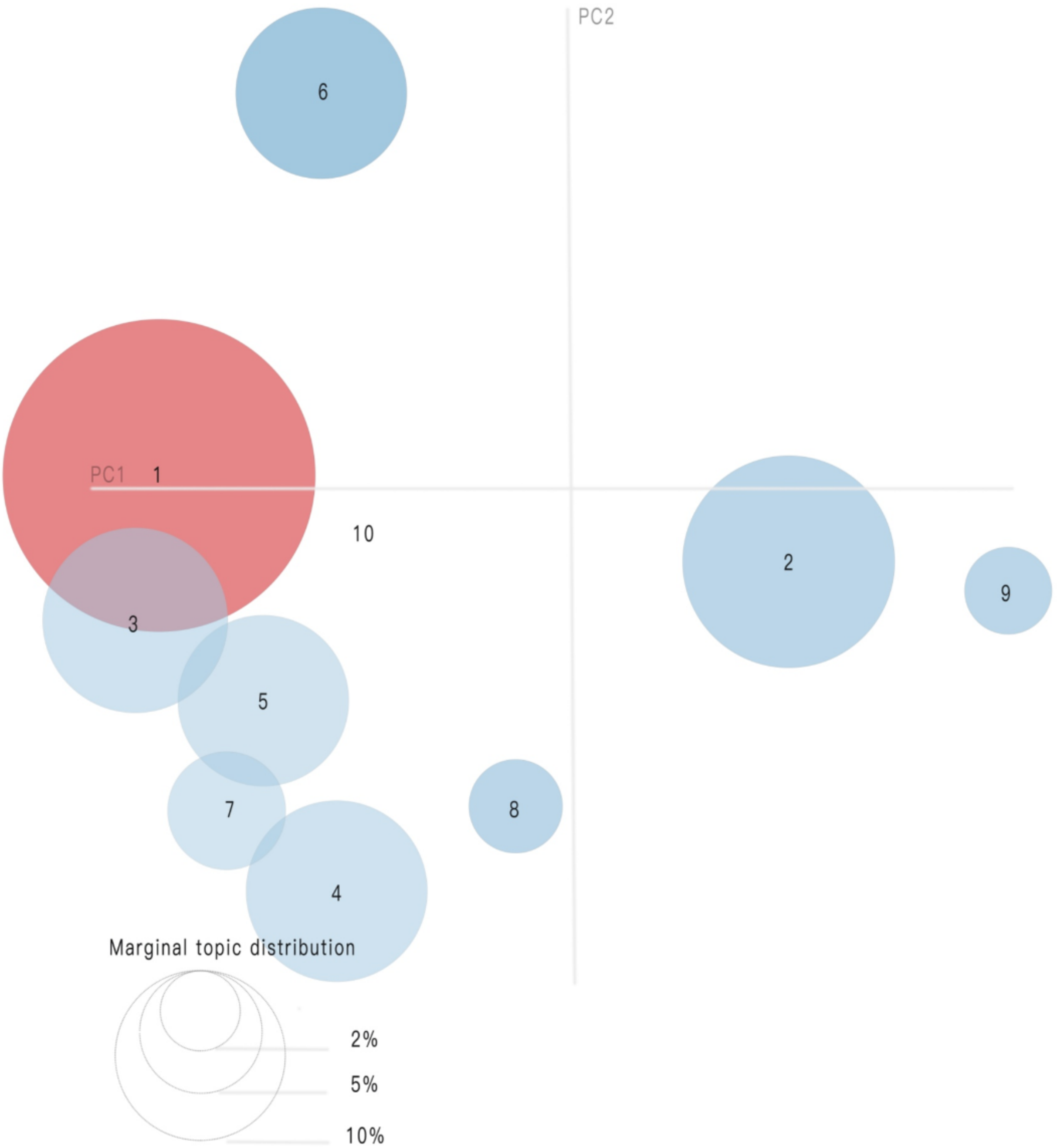


Figure 3. Top 20 most relevant terms for topic 1.



Results

We found 745 articles dated between January 1, 2013, and December 31, 2017. After excluding repeated articles, 716 news stories remained. This study collected 1201 news articles via the Wiser database platform. After culling repeated news and irrelevant entries, we recovered 288 news stories by selecting publications from only the largest media outlets in China representing the most influential mass media outlets. Finally, after the data-cleaning process was complete, we amassed 279 articles dated from January 1, 2013, through December 31, 2017. A comparison of the Chinese news and American news is listed in Table 2. Worldwide attention has been focused on THS [1,21], and the top five regions of origin of these articles and their corresponding prevalence are listed in Table 3. Not surprisingly, the United States ranks first, publishing roughly 152 articles on THS from 2013-2017. There has been an obvious decrease in interest in THS from the Chinese media within the last 5 years.

Figure 2 presents an overall view of the topic model. We plotted different topics as circles, where overall prevalence was calculated as the areas of the circles. The centers of each topic were determined by computing the distance between topics; we used multidimensional scaling to represent the intertopic distances on a two-dimensional plane [22]. PC1 indicates the transverse axis and PC2 indicates the longitudinal axis in Figure 2.

Figure 3 shows a bar chart in a descending order of the top 20 most useful terms, for interpreting a topic. The overlaid bars represent a given term’s corpus-wide frequency and the topic-specific frequency, as noted previously in the literature [23].

Figure 4 shows the percentage of news related to THS in China. Topic 1 (Cancer) was the most popular. Roughly 32% of all news stories noted this topic. Topic 2 (Control of quitting smoking) was included in approximately 15% news of THS-related news. Topic 3 (Susceptible population), Topic 4 (Quitting smoking), Topic 5 (Risks of smoking), and Topic 6 (Air quality) were each involved in roughly 10% of news stories. There were no news reports related to Topic 10.

Our results show that the Chinese press was less concerned with THS than the American press. Relative to 2013-2015, the number of reports in 2016 mentioning THS declined slightly. This trend indicates that the popularity of the topic might follow a worldwide trend (Table 3). We also observed a trend of increasing concern after a Chinese American scientist presented important findings about THS, linking THS to possible DNA damage that can cause diseases such as cancer [24,25]. This finding was widely reported both in China and the United States. It had an influence, particularly in government policy making (eg, legislation on smoke-free environments) [26]. Thus, there is a link between academic concerns and mass media concerns in China and the United States [5].

We also analyzed the three classifications, including links to diseases category, air and the $PM_{2.5}$ category, and the control and restrictions category (Table 1 and Figure 4). The first category, links to diseases, appeared in 72.6% of articles. This finding suggests that this category was frequently reported in China. Some health information can be misleading when delivered via the media, and new media platforms enable the rapid spread of news, some of it emotional or personal. This can include information that is false or misleading. As a result, the public still has a limited or even inaccurate understanding of THS. This situation was consistent with the findings of a prior study [27].

Roughly one tenth (9.7%) of the all-news stories were related to the air and $PM_{2.5}$ category. The media seemed to focus on sensational topics related to THS, even if some links were tenuous. Therefore, professionalism and credibility in reporting are of vital importance. Furthermore, the control and restrictions category was included in 17.1% of all stories. This category is a unique one. While the Chinese media emphasizes control and restrictions, the American media focuses on conveying to smokers that they should quit the habit.

In comparing the Chinese and American news, we found an enormous variation between reports on the same topic in terms of the following three aspects. The Chinese news focused on children and the elderly as victims, whereas the American news focused more on pregnant women. The Chinese media discussed a number of measures and restrictions the government or authorities took or should be taken related to smoking, while the American press focused more on how individuals could eliminate risks. Furthermore, Chinese news articles were, on average, shorter than American news articles.

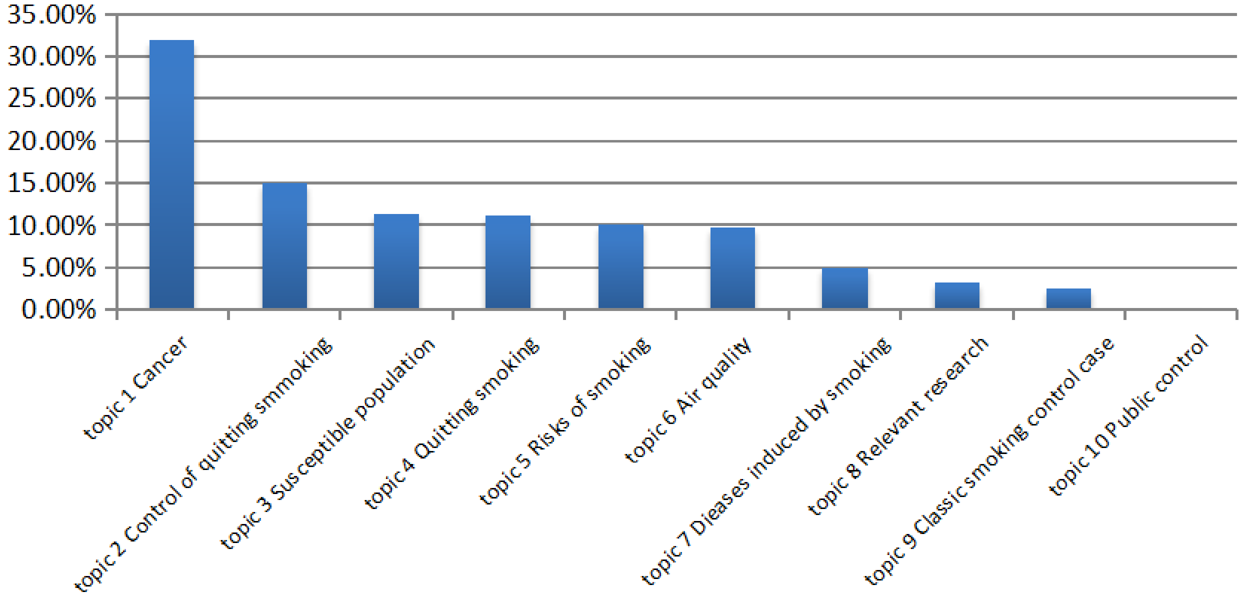
Table 2. Chinese and American news about thirdhand smoke.

Year	Chinese news articles (n)	American news articles (n)
2013	78	52
2014	57	114
2015	74	54
2016	29	38
2017	41	105
Total	279	363

Table 3. Countries publishing news about thirdhand smoke.

Country	Articles (n)
United States	152
United Kingdom	56
Canada	37
India	33
Australia	21

Figure 4. Topic percentage allocation.



Discussion

Principal Findings

Thirdhand smoke has been a growing topic for years in China. Topic modeling offers insights into understanding news reports related to THS. This study collected media reports from the United States and China and analyzed them using topic modeling. Specifically, we retrieved Chinese reports about THS from major newspaper and press outlets in China and compared this information with English-language news on the same subject from the United States. Our results revealed that the American press was more concerned with THS than the Chinese. In the Chinese media, three major concerns emerged: links of THS to disease, air and $PM_{2.5}$, and tobacco control and restrictions.

Our results indicate that the media has served a public health function by publishing reports that warn the general public about THS-related dangers such as lung cancer, asthma, tumors, and other diseases linked to tobacco. Furthermore, these articles emphasize the possible risks of THS and the susceptible populations (eg, children) in indoor environments. Smoking cessation and relevant research are also reported. Therefore, the media do communicate the risks associated with smoking, as well as information about prevention and smoking cessation.

For air quality and $PM_{2.5}$-related reports, the media attempts to connect sensational concepts with the topic, even if such concepts are not quite related. Therefore, professionalism and credibility are of vital importance. In addition, new media platforms enable the rapid spread of news that is more emotional or personal, or possibly false, which is a serious concern. Aided by new media platforms such as WeChat, Weibo, and Jinri Toutiao (the biggest new media platforms in China), information that may include misleading or exaggerated concepts can quickly be disseminated on personalized newsfeeds.

In terms of topics related to tobacco control and restrictions, the Chinese media emphasizes control and restrictions more than the United States. The American media focuses on helping smokers quit the habit.

Strengths and Limitations

Topic modeling is a new method that reveals the major topics in media reports and singles out several key concerns and findings related to the topics. Data analysis and the visualization of news articles can generate useful information. However, there is a limitation to be noted in our study. We included only major media databases, which might omit some news content from new media, such as WeChat posts. Therefore, we may have missed some news stories.

Conclusion

Thirdhand smoke is an important public health issue. Collecting media reports on THS from major media outlets and analyzing this subject using topic modeling can facilitate a better understanding of the role that the media plays in communicating this health issue to the public. We conclude that the Chinese press still needs to consider fuller reports on THS rather than simply reporting sensational headlines and needs to show more professionalism by not publishing articles that lack scientific evidence. We recommend that additional studies be conducted related to sentiment analysis of news data to verify and measure the influence of topics revealed from the reports. For example, scientists could measure the educational function of the media for public health or study the influence of misleading information about THS generated by news reports.

Acknowledgments

This study was funded by the National Social Science Foundation of China (18CXW021).

Conflicts of Interest

None declared.

References

1. Winickoff JP, Friebely J, Tanski SE, Sherrod C, Matt GE, Hovell MF, et al. Beliefs about the health effects of "thirdhand" smoke and home smoking bans. Pediatrics 2009 Jan;123(1):e74-e79 [FREE Full text] [doi: 10.1542/peds.2008-2184] [Medline: 19117850]
2. Burton A. Does the smoke ever really clear? Thirdhand smoke exposure raises new concerns. Environ Health Perspect 2011 Feb;119(2):A70-A74 [FREE Full text] [doi: 10.1289/ehp.119-a70] [Medline: 21285011]
3. Protano C, Vitali M. The new danger of thirdhand smoke: why passive smoking does not stop at secondhand smoke. Environ Health Perspect 2011 Oct;119(10):A422 [FREE Full text] [doi: 10.1289/ehp.1103956] [Medline: 21968336]
4. Quinlan K, McCaul K. Matched and mismatched interventions with young adult smokers: testing a stage theory. Health Psychol 2000 Mar;19(2):165-171. [Medline: 10762100]
5. Delgado-Rendon A, Cruz T, Soto D, Baezconde-Garbanati L, Unger J. Second and Thirdhand Smoke Exposure, Attitudes and Protective Practices: Results from a Survey of Hispanic Residents in Multi-unit Housing. J Immigr Minor Health 2017 Oct;19(5):1148-1155. [doi: 10.1007/s10903-016-0540-x] [Medline: 28074306]
6. Escoffery C, Bundy L, Carvalho M, Yembra D, Haardörfer R, Berg C, et al. Third-hand smoke as a potential intervention message for promoting smoke-free homes in low-income communities. Health Educ Res 2013 Oct;28(5):923-930 [FREE Full text] [doi: 10.1093/her/cyt056] [Medline: 23669213]
7. Zheng Y, Zhang Y, Larochelle H. A Deep and Autoregressive Approach for Topic Modeling of Multimodal Data. IEEE Trans Pattern Anal Mach Intell 2016 Dec;38(6):1056-1069. [doi: 10.1109/TPAMI.2015.2476802] [Medline: 26372202]
8. McLaurin E, McDonald A, Lee J, Aksan N, Dawson J, Tippin J, et al. Variations on a theme: Topic modeling of naturalistic driving data. Proc Hum Factors Ergon Soc Annu Meet 2014 Dec;58(1):2107-2111 [FREE Full text] [doi: 10.1177/1541931214581443] [Medline: 26190948]
9. Zhao W, Chen J, Perkins R, Liu Z, Ge W, Ding Y, et al. A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinformatics 2015;16 Suppl 13:S8 [FREE Full text] [doi: 10.1186/1471-2105-16-S13-S8] [Medline: 26424364]
10. Hassanpour S, Langlotz CP. Unsupervised Topic Modeling in a Large Free Text Radiology Report Repository. J Digit Imaging 2016 Feb;29(1):59-62 [FREE Full text] [doi: 10.1007/s10278-015-9823-3] [Medline: 26353748]
11. Kandula S, Curtis D, Hill B, Zeng-Treitler Q. Use of topic modeling for recommending relevant education material to diabetic patients. AMIA Annu Symp Proc 2011;2011:674-682 [FREE Full text] [Medline: 22195123]
12. Goyal N, Gomeni R. A latent variable approach in simultaneous modeling of longitudinal and dropout data in schizophrenia trials. Eur Neuropsychopharmacol 2013 Nov;23(11):1570-1576. [doi: 10.1016/j.euroneuro.2013.03.004] [Medline: 23602612]
13. Yu K, Zhang J, Chen M, Xu X, Suzuki A, Ilic K, et al. Mining hidden knowledge for drug safety assessment: topic modeling of LiverTox as a case study. BMC Bioinformatics 2014;15 Suppl 17:S6 [FREE Full text] [doi: 10.1186/1471-2105-15-S17-S6] [Medline: 25559675]
14. Bagust A, Beale S. Survival analysis and extrapolation modeling of time-to-event clinical trial data for economic evaluation: an alternative approach. Med Decis Making 2014 Dec;34(3):343-351. [doi: 10.1177/0272989X13497998] [Medline: 23901052]
15. Li A, Huang X, Hao B, O'Dea B, Christensen H, Zhu T. Attitudes towards suicide attempts broadcast on social media: an exploratory study of Chinese microblogs. PeerJ 2015;3:e1209 [FREE Full text] [doi: 10.7717/peerj.1209] [Medline: 26380801]
16. Blei D, Ng A, Jordan M. Latent Dirichllocation. J Mach Learn Res 2003 May 15;3:993-1022.
17. He B, De Sa C, Mitliagkas I, Ré C. Scan Order in Gibbs Sampling: Models in Which it Matters and Bounds on How Much. Adv Neural Inf Process Syst 2016;29 [FREE Full text] [Medline: 28344429]
18. Peng K, Liou L, Chang C, Lee D. Predicting Personality Traits of Chinese Users Based on Facebook Wall Posts. 2015 Oct 23 Presented at: 24th Wireless and Optical Communication Conference (WOCC); October 23-24, 2015; Taipei, Taiwan.
19. Adams B, McKenzie G. Crowdsourcing the character of a place: Character-level convolutional networks for multilingual geographic text classification. Transactions in GIS 2018 Jan 29;22(2):394-408. [doi: 10.1111/tgis.12317]
20. Grimmer J, Stewart B. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Polit. anal 2017 Jan 4;21(03):267-297. [doi: 10.1093/pan/mps028]
21. Holitzki H, Dowsett L, Spackman E, Noseworthy T, Clement F. Health effects of exposure to second- and third-hand marijuana smoke: a systematic review. CMAJ Open 2017 Nov 24;5(4):E814-E822 [FREE Full text] [doi: 10.9778/cmajo.20170112] [Medline: 29192095]
22. Chuang J, Ramage D, Manning C, Heer J. Interpretation and trust: designing model-driven visualizations for text analysis. 2012 Oct 05 Presented at: SIGCHI Conference on Human Factors in Computing Systems; May 05-10, 2012; Austin, Texas p. A-52.

23. Chuang J, Manning C, Heer J. Termite: Visualization Techniques for Assessing Textual Topic Models. 2012 May 21 Presented at: Proceedings of the International Working Conference on Advanced Visual Interfaces; 2012; Capri Island, Italy p. 74-77.
24. Hang B, Sarker A, Havel C, Saha S, Hazra TK, Schick S, et al. Thirdhand smoke causes DNA damage in human cells. Mutagenesis 2013 Jul;28(4):381-391 [FREE Full text] [doi: 10.1093/mutage/get013] [Medline: 23462851]
25. Chen Y, Adhami N, Martins-Green M. Biological markers of harm can be detected in mice exposed for two months to low doses of Third Hand Smoke under conditions that mimic human exposure. Food Chem Toxicol 2018 Dec;122:95-103. [doi: 10.1016/j.fct.2018.09.048] [Medline: 30253244]
26. Kennedy R, Ellens-Clark S, Nagge L, Douglas O, Madill C, Kaufman P. A Smoke-Free Community Housing Policy: Changes in Reported Smoking Behaviour-Findings from Waterloo Region, Canada. J Community Health 2015 Dec;40(6):1207-1215. [doi: 10.1007/s10900-015-0050-0] [Medline: 26070870]
27. Acuff L, Fristoe K, Hamblen J, Smith M, Chen J. Third-Hand Smoke: Old Smoke, New Concerns. J Community Health 2016 Dec;41(3):680-687. [doi: 10.1007/s10900-015-0114-1] [Medline: 26512014]

Abbreviations

LDA: Latent Dirichlet Allocation
TF-IDF: frequency–inverse document frequency
THS: thirdhand smoke
PM: particulate matter

Edited by G Eysenbach; submitted 06.10.18; peer-reviewed by C Zhang, A Majmundar; comments to author 25.10.18; revised version received 14.12.18; accepted 05.01.19; published 29.01.19

Please cite as:
Liu Q, Chen Q, Shen J, Wu H, Sun Y, Ming WK
Data Analysis and Visualization of Newspaper Articles on Thirdhand Smoke: A Topic Modeling Approach
JMIR Med Inform 2019;7(1):e12414
URL: http://medinform.jmir.org/2019/1/e12414/
doi: 10.2196/12414
PMID: 30694199