International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 9 Issue 5 Series II || May 2020 || PP 17-20

Research on the Focus of Chinese Netizens on Garbage Classification Based on LDA TopicModel ——Taking "Zhihu" as an Example

Hongde Liu
(School of Economics and Management, China University of Petroleum (Beijing), Beijing 102249, China)

ABSTRACT:*Studying the content of netizens' discussion on garbage classification is helpful to discover the potential focus of netizens. With the help of python crawler technology, this paper obtains a total of 1923 pieces of relevant text data of garbage classification on the "Zhihu" platform, and uses confusion calculation to determine the optimal number of topics. Then, subject the text by LDA topic model. Based on the time factor, a Date-Topic intensity evolution graph is drawn. The discussion topics of netizens show diverse characteristics. It mainly includes direct focus and indirect focus. The promulgation of each policy will cause a strong discussion among netizens within a period of time after its introduction, and medical waste is highly concerned by netizens since February 2020. Paying attention to the trend of increasing first and then decreasing, it is very important to do long-term publicity work.*

KEYWORDS: *garbage classification; LDA; Date-Topic intensity evolution; Zhihu; netizens*

Date of Submission: 06-05-2020 Date of Acceptance: 20-05-2020

I. INTRODUCTION

Due to the large population of our country, the huge amount of garbage put great pressure on the environment. Garbage classification is a reform of the traditional methods of garbage collection and disposal[1]. Starting from July 1, 2019, Shanghai is the first experimental unit to implement garbage classification [2]. Ensuring the people's front-end classification awareness and back-end treatment facilities is the important priority of the measures to promote garbage classification.Affected by COVID-19, offline publicity and education work are in a bottleneck period. It is imperative to promote online publicity and raise the awareness of netizens in garbage classification. With the advent of the web 2.0 era, the convenience of the social question and answer (Q&A) platform can be reflected[3]. "Zhihu" is the most popular social Q&A platform in China. Researchers often use the "Zhihu" platform to discover potential concerns of netizens.

We use Python to obtain the netizens' discussion about the garbage classification on the "Zhihu" platform. We use LDA topic model for topic extraction. On April 26, 2019, the Ministry of Housing and Urban-Rural Development issued a notice instructing prefecture-level and higher cities nationwide to fully initiate domestic garbage classification. We consider the time factor and make a Date-Topic intensity evolution graph from April 2019 to April 2020. By analyzing keywords under different topics to understand the netizens' focus on garbage classification, and provide suggestions for targeted online work. Through Date-Topic intensity evolution graph analysis, understand the public opinion-oriented trend of garbage classification.

II. LITERATURE REVIEW

The current research focuses on three aspects: policy benefit evaluation, technical method research and mass behavior research. In terms of policy benefit evaluation: Sisi Chen, Jialiang Huang, Tingting Xiao, et al (2020) [4]use LCA and LCI methods to study the impact of waste classification on carbon emissions from different domestic waste treatment methods. In terms of technical methods: Fengqi Zhou, Wenbo Zhang (2020) [5]point out that there are technical problems in the current garbage classification in the stages of collection, treatment, removal and transportation. In terms of mass behavior research: Bo Fan, Wenting Yang, Xingchen Shen. (2018) [6] take China and Singapore as examples to build motivation based on planned behavior theory-intention-behavior model, using structural equation model to empirical investigation results.

In the field of text research, LDA topic model is widely used.Guowen Li, Xiaoqian Zhu, Jun Wang, et al(2017) [7] use the LDA topic model to analyze China ’s financial stability report; Changyu Zhao, Yaping Wu, Jimin Wang(2019) [8] use LDA to analyze Twitter ’s text on the “Belt and Road” topic, and obtain the main focus of attention; Yan Lou, JialinYang, Lucheng Huang, et al(2020) [9] use LDA to extract topics, analyze the topics related to senior technology on"Zhihu", and obtain 10 topics that the public paid attention to in senior technology.In summary, there are few researches on the web text of garbage classification. Analyzing the

relevant texts on the social Q & A platform is helpful for understanding the netizens' focus, and the LDA topic model is widely used in text analysis.

III. METHOD INTRODUCTION

3.1 LDA Topic Model

LDA (latent Dirichlet allocation) is a probabilistic model of traditional clustering discrete data sets proposed by Blei et al. in 2003 [10]. It has been widely used in document clustering analysis in recent years [11-14]. The LDA model constructs a three-layer Bayesian structure of document-topic-word, and uses the form of probability distribution to intuitively display the topic of each document in the data set [13]. User can classify topics based on the results of topic probability distribution [14]. This calculation method, which considers probability based on the space vector model, is helpful for extracting popular points of interest and related words from a bunch of data sets[12].The probability of generating wordsw_j under the comment condition of garbage classification events $d_i$is expressed as[11]:

$$p(w_j|d_i) = \sum_{s=1}^{K} P(w_j|z = s) \times P(z = s|d_i) \quad (1)$$

$P(w_j|z = s)$indicates the probability that the word belongs to topic z, and $P(z = s|d_i)$ indicates the probability of the topic z in the garbage classification event comment d_i.

3.2 Perplexity analysis

The LDA topic model needs to determine the value of the topic number K. K in this paper is selected through the calculation of perplexity (equation 2). The smaller the value, the higher the degree of topic fit [9].

$$perplexity(D) = e^{-\frac{\sum_{d=1}^{l} \log(P(w_d))}{\sum_{d=1}^{l} N_d}} \quad (2)$$

l is the number of texts, and $N_d$is the total word frequency in the documentd.

IV. EXAMPLE ANALYSIS

We use Python to crawl 2,837 discussions on garbage classification under the topic of garbage classification on the "Zhihu" platform. After excluding duplicate data, incomplete data, data beyond the time span, and data not related to garbage classification, 1923 valid samples remained.

4.1 Extraction of Perplexity

With the help of python, calculate the perplexity (Fig.1). The smaller the perplexity, the higher the degree of topic fit. But too many topics will increase the difficulty of analysis. Therefore, considering the degree of topic fitting and analysis difficulty,we set the number of topics to 10.

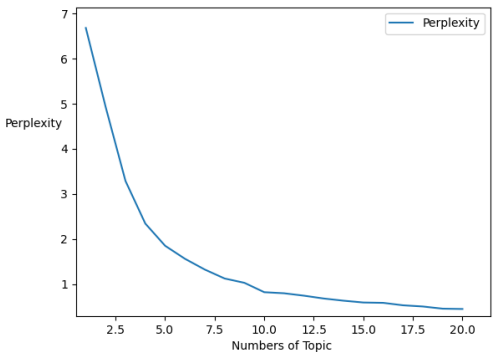

Figure 1 Graph of calculation result of perplexity

4.2 Topic Extraction

Figure2 Topic extraction results

Based on the Topic-Keyword document, we can find that the topic content is diverse. Mainly involves the following aspects: Commercial Market (Topic 1); the Equipment and Technology of Garbage Classification (Topic 2); Policy Plan (Topic 3); Garbage Disposal Process (Topic 4); Campus Education (Topic 5); Medical Waste (Topic 6); Cognitive Learning (Topic 7); Pilot City (Topic 8); Social Research (Topic 9); Experience Reference (Topic 10). We divide these topics into the direct focus and the indirect focus.

Topic 2, 3, 4, 6 and 8 are the direct focus. Netizens pay attention to Garbage Disposal Process and the Equipment and Technology of Garbage Classification. Popularizing common knowledge about garbage classification is helpful to meet the needs of netizens, and the correct technical description will also reduce the problems in the management process. Pilot cities such as Shanghai and Beijing have become the focus of attention of netizens. Doing a good job description and summarizing the results of the pilot cities will benefit the nationwide promotion of policies. The collection of Policy Plan could help to understand the sentiment of netizens, and plays a guiding role in garbage classification policies. The emergence of medical waste is largely related to the epidemic situation, and the disposal of medical waste should be done.

Topic 1, 5, 7, 9 and 10 are the indirect focus. With the emergence of garbage classification, relevant industrial chains have gradually taken shape. Many people have discovered the market opportunities. At the same time, garbage classification has awakened the social research work that has been sleeping for a long time, and a new round of research is being carried out in an orderly manner. The public's perspective of discussion has expanded from domestic to Japan, which is relatively advanced in implementing waste classification. Japan has been at the forefront of the world in waste classification, especially kitchen waste, and domestic scholars have also studied Japanese waste classification models in large numbers. The implementation of a policy needs to arouse public awareness, and people's cognitive learning and related education work are also received great attention.

4.3 Date-Topic Intensity Evolution Analysis

We use document-topic distribution to divide the modeling results into different time slices at a certain time granularity.Finally, wemake the Date-Topic intensity evolution graph from April 2019 to April 2020 according to the number of texts contained in each topic (Fig.3).

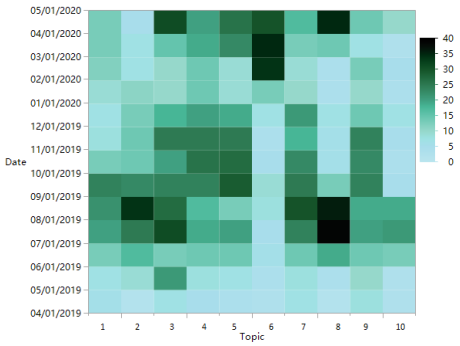

Figure 3 Date-Topic intensity evolution graph

Analyzing the Date-Topic intensity evolution graph, we find that in addition to Topic 6, the remaining Topic netizens pay more attention to July-October 2019, which may be due to the official implementation of the Shanghai garbage classification policy in July 2019. A downward trend began in November, and netizens are concerned about the landslide. However, the focus of attention begins to pick up in March 2020, because Beijing began implementing the garbage classification policy on May 1, 2020. We can draw a preliminary conclusion: the introduction of each policy will cause a strong discussion among netizens within a certain period of time after the introduction of the policy; education on garbage classification focuses on the school start time in September and the online class period since February 2020; Medical waste is received less attention before, but it has been highly concerned by netizens since the outbreak of the new coronavirus pneumonia; people's cognitive learning behaviors have shown a trend of increasing first and then decreasing, and it is vital to do

long-term publicity work; before starting the pilot Within a month, we must fully do our best to guide public opinion online.

V. CONCLUSION

In the short term, we must pay attention to the popularization of waste sorting equipment and technology and guide netizens to understand the correct waste disposal process; because the pilot cities are facing high social attention, we must promptly disclose the implementation process and results of transparent measures; There is a great need for medical waste related publicity work. In the long term, we need pay attention to school education and actively promote the purpose and significance of educational activities; encourage relevant manufacturers and research institutions to increase R & D efforts to develop new business service models and products, and publish typical examples of successful entrepreneurship on the platform; for foreign developed countries,we should learn from experience; the netizens' cognitive learning on garbage classification will reach its peak before and after the pilot cities begin to carry out garbage classification, but it will not be sustainable in the long run. Therefore, it is important to do long-term online learning.

REFERENCES

[1]. HuaGuo, Jiamei Li, Mingjie Qiu,et al.Discussion on the problem of waste classification management in China [J].EnvironmentandDevelopment, 2018,30 (02): 42-43.
[2]. Shijiang Xiao, Huijuan Dong, Yong Geng, et al. Policy impacts on Municipal Solid Waste management in Shanghai: A system dynamics model analysis[J]. Journal of Cleaner Production, 262.
[3]. Li M, Lu X, Chen L, et al. Knowledge map construction for question and answer archives[J]. Expert Systems with Applications, 2019, 141:112923.
[4]. Sisi Chen,Jialiang Huang,Tingting Xiao,et al. Carbon emissions under different domestic waste treatment modes induced by garbage classification: Case study in pilot communities in Shanghai, China[J]. Elsevier B.V.,2020,717.
[5]. Fengqi Zhou, Wenbo Zhang. Research on the characteristics of artificial intelligence applications and optimization paths in the field of garbage classification [J / OL]. Journal of Xinjiang Normal University (Philosophy and Social Sciences Edition), 2020 (04): 1-10 [2020-04-25].
[6]. Bo Fan,Wenting Yang,Xingchen Shen. A comparison study of 'motivation–intention–behavior' model on household solid waste sorting in China and Singapore[J]. Journal of Cleaner Production,2019,211.
[7]. Guowen Li,Xiaoqian Zhu,Jun Wang, et al. Using LDA Model to Quantify and Visualize Textual Financial Stability Report[J]. Procedia Computer Science,2017,122.
[8]. ChangyuZhao, Yaping Wu, Jimin Wang. Twitter text topic mining and sentiment analysis under the "Belt and Road Initiative" [J / OL]. Library and Information Work: 1-9. 2019.19.012.
[9]. Yan Lou, JialinYang, Lucheng Huang, et al. Hot spots and sentiment analysis of the elderly science and technology public based on the online question-and-answer community—Taking"Zhihu"as an example [J]. Journal of Information, 2020,39 (03): 115 -122.
[10]. Saiping Guan, Xiaolong Jin, Xueke Xu, et al. Clustering of news reviews based on WMD distance and neighbor communication [J]. Journal of Chinese Information Processing, 2017, 31 (05): 203-214.
[11]. Yang Yuan, Xiao Li, Yating Yang. Unsupervised word sense disambiguation in Uyghur language based on LDA topic model [J]. Journal of Xiamen University (Natural Science Edition), 2020, 59 (02): 198-205.
[12]. Tao Wang, Ming Li. Research on comment text mining based on LDA model and semantic network [J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2019, 36 (04): 9-16.
[13]. Bastani K, Namavari H, Shaffer J. Latent Dirichlet Allocation (LDA) for Topic Modeling of the CFPB Consumer Complaints[J]. 2018.
[14]. Miha Pavlinek, Vili Podgorelec. Text classification method based on self-training and LDA topic models[M]. Pergamon Press, Inc. 2017.

Hongde Liu. "Research on the Focus of Chinese Netizens on Garbage Classification Based on LDA TopicModel——Taking "Zhihu" as an Example." *International Journal of Engineering Science Invention (IJESI),* Vol. 09(05), 2020, PP 17-20.