Evaluation of E-bike Safety Management Policy Based on Text Mining

Enhua Zhang[1,a*], Weijie Wang[1,b], Tingli Zhao[1,c], Dong Chen[2,d]

[*a]zhangeh2022@163.com, [b]1813631150@qq.com,[c] 1583832650@qq.com,[d] 2632549341@qq.com

[1]College of Transportation Engineering, Nanjing Tech University, Jiangsu, China;
[2]College of Computer and Information Engineering, Nanjing Tech University, Jiangsu, China

Abstract: In view of the public comments on the safety management policy of electric bicycle on Weibo, this paper obtained the text data of related comments through crawler. snownlp and TF-IDF were used to conduct emotion mining and text feature extraction on the text data, and the LDA topic model of negative text was constructed to analyze the potential topic keywords of negative emotion. The results show that within one year after the implementation of Jiangsu E-bikes Management Regulations, positive and negative emotions account for almost the same proportion in Weibo comments, and the focus of both are “mandatory helmets using” and “prohibition of carrying adults”. Those with positive attitudes believe that mandatory helmet using and prohibition of carrying adults can improve safety of e-bike riding. The main reasons for the negative attitude were: prohibition of carrying adults, the rose price of helmets, inconvenient to carry helmets, wearing helmets when riding shared e-bikes and replacing with construction helmets. This study provides an effective tool for safety management policy evaluation and analysis of electric bicycles.

Keywords: Traffic management, E-bikes, Sentiment classification, Text feature extraction, LDA topic model

1 INTRODUCTION

E-bikes have become an important means of transportation for urban residents to travel short distances due to their convenient and economical transportation characteristics [1, 2]. At the same time, statistics show that traffic accidents caused by e-bikes that result in human casualties have continued to rise over the past 10 years [3]. E-bike traffic safety management has become an important task in urban traffic safety management. In May 2020, Jiangsu Province, as a province with large production and use of e-bikes, took the lead in promulgating China's first regulation on the management of e-bike traffic to provide a law enforcement basis for the safe management of e-bikes with a view to improving the current situation of e-bike traffic safety. The Jiangsu E-bike Regulations, which cover the management of e-bike traffic and penalties, have attracted strong public attention since their promulgation and implementation.

With the rapid development of mobile smart terminal technology, social media platforms have become the main place for people to express their insights and experiences [4], thus generating a large amount of social media data. In recent years scholars in the field of transportation have begun to pay attention to the application of social media data in transportation management. Joaquín Osorio-Arjona et al. analyzed the operation of metro stations and problems through the

ICPDI 2023, September 01-03, Chongqing, People's Republic of China

DOI 10.4108/eai.1-9-2023.2338797

social media data of Twitter users' discussions about the metro network (which mainly focuses on users who complained), which can better predict the change of users' moods due to changes in the transportation network [5].Collins et al. used social media data posted by Twitter users about public transportation to evaluate the satisfaction of Chicago rail passengers, which helped the transportation department to improve the public transportation system to provide passengers with a better travel experience [6].Qian Ye et al. analyzed the changes in people's attitudes toward the congestion pricing proposal in New York City by using Twitter data [7]. Haoliang Chang et al. [8] obtained information about traffic accidents and congestion posted by Sina Weibo users, located the areas where the accidents and congestion occurred, and analyzed the users' attitudes to determine the priority of mitigation measures. Shiliang Wang [9] obtained social media data posted by Sina Weibo users and demonstrated the value of social media in monitoring air quality trends and public responses, providing new ideas for air pollution monitoring.

Social media texts are mostly the public's immediate opinions or emotional expressions of events, which are rich in content, and can be analyzed through text mining to obtain the public's opinions and emotional expressions of events. Therefore, this paper takes the comments on Jiangsu Province Electric Bicycle Management Regulations in Sina Weibo as the research object, and analyzes the public's viewpoints and emotions on Jiangsu Province Electric Bicycle Management Policy through sentiment mining, feature extraction, and topic modeling, focusing on the reasons for the negative emotions to support the effective implementation of the E-bike Safety Management Policy.

2 RESEARCH METHOD

2.1 Emotional analysis

Sentiment analysis is an important text information analysis and processing technology, whose goal is to automatically mine and analyze subjective information such as stance, point of view, opinion, emotion, etc. in the text, which contains sentiment basic unit extraction, sentiment classification, sentiment summary, sentiment retrieval and so on. Sentiment classification and text feature extraction are important research tasks in sentiment analysis, which need to be carried out on the basis of Chinese word segmentation. In recent years, with the improvement of corpus and the development of machine learning methods, statistical-based Chinese word segmentation methods have been widely used. Among them, jieba segmentation uses a probabilistic language model based on directed acyclic graphs, dynamic programming and Hidden Markov Models, which has high Chinese segmentation efficiency [10]. In addition, deactivated words can be added to it to eliminate texts with no real meaning, and customized dictionaries can be added to ensure that fixed collocations are not segmented.

2.1.1 Emotion classification

Sentiment classification is one of the multiple research tasks in sentiment analysis. Due to the short Chinese microblog text, strong emotions, language irregularities, etc., it is difficult for traditional text sentiment analysis techniques to obtain ideal processing results [11], while the snownlp natural language processing library based on the plain Bayesian algorithm has a higher accuracy rate in Chinese text analysis, which mainly has the functions of lexical annotation, sentiment value computation, sentiment categorization, extraction of keywords of the text, extraction of abstracts, etc.[12].

2.1.2 Text feature extraction

Currently TF-IDF (Term Frequency-Inverse Document Frequency) algorithm is the most widely used text feature extraction method, where TF refers to word frequency and IDF refers to inverse document frequency. TF-IDF can evaluate the importance of a word to a text set [13], if a word occurs more frequently in a text set and seldom occurs in other text sets, it is considered that the word has a good ability to differentiate between categories, that is, it is important to the text set is high.IDF is a measure of the general criticality of a word. TF-IDF is the product of TF and IDF, the algorithm tends to filter out the relatively common words and emphasize the words that are important to the document.

TF-IDF is calculated as follows:

$$TF_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

Where $n_{i,j}$ is the number of occurrences of the word in the file, $\sum_k n_{k,j}$ is the sum of occurrences of all words.

$$IDF_i = \log_{10} \frac{|D|}{1 + |d \in D: t \in d|} \tag{2}$$

Where $|D|$ denotes the total number of files in the corpus; $|d \in D: t \in d|$ denotes the number of files in which this word appears.

2.2 Text topic modeling

2.2.1 LDA（Latent Dirichlet Allocation）

LDA topic model, also called three-layer Bayesian probabilistic model, contains three-layer structure of words, topics and text sets, which is an unsupervised machine learning technique with good semantic dimensionality reduction [14, 15], and can be used to identify potential topic information in large-scale text sets or corpora. Chinese microblog text is a kind of short text with a large number of new vocabulary, and the probability of the same words appearing in different short texts is low, and the traditional word or phrase-based feature approach is difficult to accurately compute the similarity between Chinese texts [15], and the LDA topic model is suitable for topic mining in this environment.

The LDA topic model has a three-layer structure, which obeys a polynomial distribution from the text set to the topics and from the topics to the words.The LDA topic model is shown in Fig. 1.

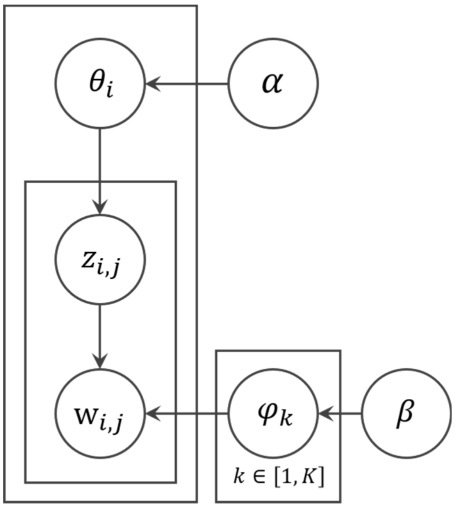

Fig. 1 Schematic diagram of LDA topic model

Where α and β denote the pre-specified parameters of Dirichlet distribution; θ denotes the polynomial distribution parameter of the topic in the document, obeying the Dirichlet distribution with parameter α; and φ denotes the polynomial distribution parameter of the word in the topic.

The LDA topic model is generated as follows: (1) selecting a document d_i according to the prior probability $p(d_i)$; (2) generating the topic distribution θ_i of document d_i by sampling from the Dirichlet distribution with parameter α; (3) generating the topic $z_{i,j}$ of the jth word of document d_i by sampling from the polynomial distribution of topics θ_i; Dirichlet distribution with parameter β to generate the distribution φ_k of words corresponding to topic $z_{i,j}$; and sampling from the polynomial distribution φ_k of words to finally generate words $w_{i,j}$.Therefore, the probability that document d_i generates word w_j can be expressed as:

$$P(w_j \mid d_i) = \sum_{k=1}^{K} P(w_j \mid k) \times P(k \mid d_i) \tag{3}$$

2.2.2 Coherence test

The consistency index quantifies the semantic similarity between the top words in a topic, which is used to ensure the interpretability of the output topic of the LDA model. A higher consistency score represents better topic modeling [16]. In this paper, we use the UMass method as a consistency metric, which has been shown to match human judgments of topic quality [17]. The UMass metric defines a score based on document co-occurrence:

$$score(v_i, v_j, \varepsilon) = \log \frac{D(v_i, v_j) + \varepsilon}{D(v_j)} \tag{4}$$

Where $D(v_i, v_j)$ denotes the number of documents containing words v_i and v_j, $D(v_j)$ denotes the number of documents containing v_j. Then the coherence of UMass metrics is calculated by the formula:

$$C_{UMass} = \frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{P(v_i, v_j) + \varepsilon}{P(v_j)} \tag{5}$$

The computation of UMass metrics is based on the original corpus on which the topic model is trained, confirming that the LDA model learns the data present in the corpus. In this paper, the consistency test of the UMass metric is used to evaluate the effectiveness of LDA topic modeling and to determine the optimal number of topics.

3 DATA ANALYSIS AND RESULTS

3.1 Data Collection and Preprocessing

Based on Sina Weibo platform, this paper writes and implements a Python crawler script with selenium library [18] as the main tool. The pre-set keywords were used to retrieve the relevant microblogs and capture the comments on the Regulations on the Administration of Electric Bicycles in Jiangsu Province from May 2020 to April 2021, and a total of 7673 comment text data were obtained. The keywords were set to take into account the different expressions of the public for electric bicycles, such as battery car, small electric donkey, etc.; and the timeframe was limited to one year after the adoption of the May 2020 "Regulations on the Management of Electric Bicycles in Jiangsu Province".

The following principles were followed for data cleaning of the collected comment dataset:

(1) Uniformity of semantic expression. There may be different ways of expressing the same semantics, and there may be differences in expression between the official and the public, which will affect the effectiveness of the subsequent analysis, so it is necessary to unify the semantic expression;

(2) Text standardization. Initial microblog comments may contain html format tags, topic tags (#), links and @ symbols and other information without actual semantics, using Python regular expression re module to match and eliminate such symbols. The elimination will not affect the semantic expression.

3.2 Sentiment analysis

Sentiment classification is carried out on the basis of sentiment value calculation, using the python class library-snownlp for sentiment classification. In which the calculation of the sentiment value of the comment text is accomplished by the method of sentiments in snownlp, the sentiment value of the statement is taken as [0, 1], the larger the value indicates the more positive the sentiment tendency, and in order to represent the sentiment tendency of the statement in a more intuitive way, the interval of the sentiment value will be converted to [-0.5, 0.5]. The sentiment value will be divided into intervals of 0.01 length to categorize the sentiment of all the acquired comment text data. The statistics of the number of positive and negative sentiment texts showed that the percentage of comment texts with positive sentiment was 44.64%, and the percentage of comment texts with negative sentiment was 55.36%.

The IF-IDF algorithm is used to analyze the positive and negative sentiment classified texts respectively, and output the weights of the participles in the positive and negative texts, and Table 1 shows the keywords with the top ten weights.

Table. 1 TF-IDF results of positive and negative text

Positive Keywords		Negative Keywords	
Keywords	Weight	Keywords	Weight
helmet	**0.72**	**helmet**	**0.63**
safety	**0.11**	**manned**	**0.10**
manned	**0.07**	**price increase**	**0.08**
ride	0.06	safety	0.07
public	0.05	facemask	0.06
put on	0.05	bicycle	0.06
price increase	0.04	ride	0.04
shared	0.04	penalty	0.04
travel	0.04	price	0.04
16 years old	0.04	16 years old	0.03

Overall, the keywords "helmet", "safety", "manned" and "price increase" have a higher weight in the positive and negative text sets. The keyword "helmet" corresponds to "driving and riding an electric bicycle should wear a safety helmet in accordance with the regulations" in the "Regulations on the Management of E-bikes in Jiangsu Province", and the keyword "manned" corresponds to "adults driving electric bicycles can only carry a minor under the age of sixteen" (prohibited to carry adults). The keyword "price increase" is due to the price increase of helmets after the promulgation of the "Jiangsu Province Electric Bicycle Management Regulations".

In the TF-IDF results of the textual data of positive comments, the top three keywords are "helmet" (0.72), "safety" (0.11), and "manned" (0.07), indicating that in the positively categorized comment text data, the public believes that "helmets" and "manned" (prohibiting the carriage of adults) are for safety reasons, and expresses their approval and support; the analysis results of the negatively categorized comment text data show that the top three keywords are "helmet" (0.63), "manned" (0.10) and "price increase" (0.08). It can be seen that "helmet" and "manned" are concentrated in both positive and negative texts, indicating that the public's concern is the same in both positive and negative texts.

3.3 LDA Topic Modeling for Negative Sentiment Text Sets

The negative comment text set was analyzed by LDA topic modeling. In building the model, the python LDA correlation module UMass coherence score function is used, and through the coherence test (Figure 2), the negative comment text data is clustered into 3 topics, and the 10 most relevant feature words to the topics are output, and then the topic distance map is output in Python using pyLDAvis [15].

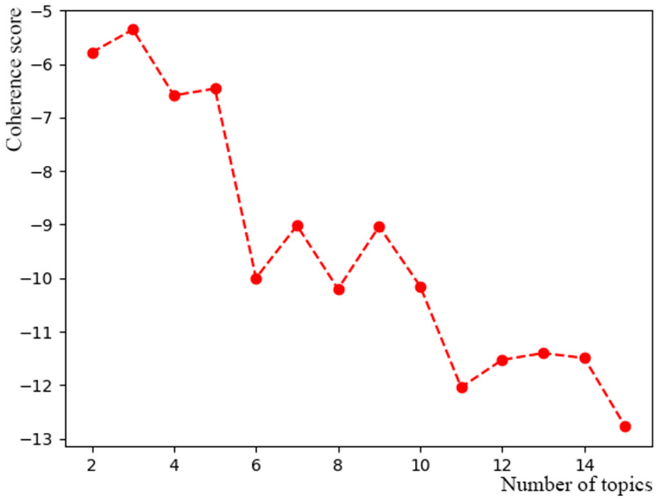

Fig. 2. Coherence scores of different topic numbers in the negative text

The size of the bubbles in the topic distance map indicates the probability of occurrence of the corresponding topic. The distance between the bubbles represents the degree of differentiation between the topics [19], the further the distance the higher the degree of differentiation, and if the bubbles overlap, it means that there is a cross section of feature words in these two topics. The results show (Figure 3) that the degree of difference between the three topics is high, indicating that the LDA topic modeling results are better, the distinction between topics is obvious, and their semantic associations are negligible.

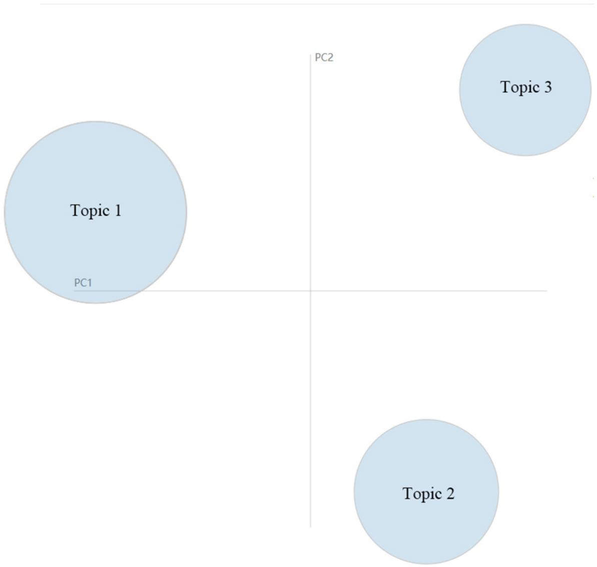

Fig. 3 Thematic distance mapping

The statistics of the top 10 most relevant feature words of the three topics are outputted separately (Table 2), and it is found that the frequency of the words “helmet” and “carrying adult passengers” is higher in the three topics, which is consistent with the TF-IDF results. The three topics focus on the public's negative sentiment towards “mandatory wearing of helmets” and “prohibiting the carriage of adult passengers” in the regulations.

Table 2 Topic Keywords of Negative Text

Topic 1		Topic 2		Topic 3	
Words	Probability	Words	Probability	Words	Probability
cannot	0.009	helmet	0.010	bicycle	0.014
bicycle	0.007	electric	0.007	manned	0.010
electric	0.007	bicycle	0.007	electric	0.009
stipulate	0.007	price increase	0.006	helmet	0.007
helmet	0.007	inconvenient	0.005	outrageous	0.007
manned	0.007	safety	0.005	leader	0.005
why	0.005	manned	0.004	okay	0.005
now	0.005	prohibit	0.004	cars	0.004
safety hat	0.004	policy	0.004	cannot	0.004
shared	0.004	yet	0.003	comment	0.004

Topic 1 is the most common topic in the negative text dataset, in which "helmet" and "shared" are the feature words of topic 1. Manually cross-checking the original dataset of negative text, "shared" mainly reflects the problem of wearing helmets when riding shared e-bikes. At present, almost no shared e-bikes on the market are equipped with helmets, and users need to carry helmets with them if they want to use shared e-bikes, and the "mandatory wearing of helmets" regulation puts users of shared e-bikes in an awkward situation. "Safety hat" mainly represents "construction site helmet" in the negative sentiment text set, which is expressed as "Is it okay to use a construction site helmet?" "Helmets are too expensive, not as good as safety hat ", etc., reflecting the lack of awareness of some members of the public about the "correct wearing of safety helmets".

Keywords of topic 2 reflect the public's negative feelings about the phenomenon of helmet price increases. At the beginning of the promulgation and implementation of the "Jiangsu Province Electric Bicycle Regulations", the demand for electric bicycle safety helmets surged, the market supply was insufficient, and some online and offline merchants took advantage of the opportunity to raise the price, which triggered public dissatisfaction. Secondly, the public has negative feelings towards helmet carrying. At the early stage of the implementation of the regulations, the public was worried about helmet theft and thought that "helmet carrying and storing are very inconvenient"; at the same time, they showed negative feelings towards the prohibition of carrying adults.

The simultaneous occurrence of "outrageous" and "okay" in Topic 3 indicates that there is a high probability of both words appearing in the text of a single comment on Topic 3, and that the public, while recognizing one aspect of the Ordinance, also expresses serious doubts about another aspect, resulting in the overall negative sentiment of the text. serious skepticism, resulting in the overall expression of negative sentiment in this text. By manually confirming the texts representing topic 3 in the negative text dataset, it is found that "outrageous" is mainly directed at the regulation on "prohibiting the carriage of adults", while "okay" is mainly directed at the regulation on "helmet wearing". "Helmet Wearing" regulation. Topic 3 reflects people's recognition of "mandatory helmet wearing" while at the same time demonstrating a lack of understanding of the "restrictive passenger" policy regulations.

4 CONCLUSIONS

In this paper, natural language processing technology is applied to traffic safety management policy analysis using Python, and the public's emotional response is mined and analyzed by collecting public comments on microblogs about the promulgated and implemented Jiangsu Electric Bicycle Management Regulations, which effectively realizes the public reflective evaluation of the electric bicycle management policy. Compared with the traditional questionnaire survey, the microblog comment information mining method adopted in this paper is more economical, can reflect the real public sentiment, and has stronger scalability in the spatial and temporal dimensions, which can effectively improve the degree of public participation in the evaluation of policies and assist in the formulation and implementation of policies, provide a strong support for the evaluation of feedback on e-bike traffic management policies, and provide a reference for the evaluation of the management of e-bikes in other regions. It also provides a reference for the management of e-bikes in other regions.

The main findings of the study are as follows:

(1) Within one year after the promulgation of the regulation, the proportion of positive and negative emotions in the public microblog comments was basically the same, with slightly more negative emotional texts than positive ones, and there was no phenomenon of one-sidedness in the concentration of comments. Both positive and negative sentiments focused on "mandatory wearing of helmets" and "prohibiting the carriage of adults". People with positive sentiments believe that mandatory wearing of helmets and banning the transportation of adults can improve the safety of e-bike riding.

(2) In the set of negative sentiments, except for "banning the carriage of adults", the public mainly focuses on the issue of "wearing helmets", which is manifested in the following ways: increasing the price of helmets, inconvenience of carrying helmets, wearing helmets when riding e-bikes, and replacing helmets with helmets in construction. Helmets instead of helmets.

REFERENCES

[1] LIAO Cong, WU Lun, CAI Heng, et al. Spatial distribution and influencing factors of unsafe charging for electric bicycles in urban areas[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2021, 4: 671-678.

[2] LI Yan, NAN Si-rui, HU Wen-bin, et al. Lane transgressing risk model of electric bicycle on marking separation road section[J]. Journal of Chongqing Jiaotong University(Natural Science), 2021, 2: 13-20.

[3] ZHANG Jian-wei, SONG Fei, DANG Wen-xiu. Current situation and countermeasures of electric bicycle management—Taking S province as an example[J]. Journal of Shandong Police College, 2020, 1: 123-129.

[4] NE C S, STEINERT T, MKER H K. Gender- and diversity-oriented design of social media for participation in public transport[J]. HCI International 2020-Late Breaking Papers: Interaction, Knowledge and Social Media, 2020: 425-443.

[5] JOAQUÍN O A, JIRI H, RADEK S, et al. Social media semantic perceptions on Madrid Metro system: using twitter data to link complaints to space[J]. Sustainable Cities and Society, 2021, 64.

[6] COLLINS C, HASAN S, UKKUSURI S V. A novel transit rider satisfaction metric: Rider sentiments measured from online social media data [J]. Journal of Public Transportation, 2013, 2: 21-45.

[7] YE Q, CHEN X H, KALAN O, et al. Using LDA and LSTM models to study public opinions and critical groups towards congestion pricing in New York city through 2007 to 2019[J]. arXiv preprint arXiv:2008, 2020: 07366.

[8] CHANG H L, LI L S, HUANG J X, et al. Tracking traffic congestion and accidents using social media data: A case study of Shanghai[J]. Accid Anal Prev, 2022, 26(169):106618.

[9] WANG S, PAUL MJ, DREDZE M. Social media as a sensor of air quality and public response in china[J]. Journal of Medical Internet Research, 2015, 17(1).

[10] TANG Lin, GUO Chong-hui, CHEN Jing-feng. Review of chinese word segmentation studies[J]. Data Analysis and Knowledge Discovery, 2020, 4(2/3): 1-17.)

[11] LI Yong-gan, ZHOU Xue-guang, SUN Yan, et al. Research and implementation of chinese microblog sentiment classification[J]. Journal of Software, 2017,28(12):3183−3205

[12] PRAPHULA K J, RAJENDRA P, GAUTAM S. A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews[J]. Computer Science Review, 2021, 41.

[13] QIN P, XU W, GUO J. A novel negative sampling based on TFIDF for learning word representation[J]. Neurocomputing, 2016, 177.

[14] LUO S L, SYLVIA Y H. Understanding gender difference in perceptions toward transit services across space and time: A social media mining approach[J]. Transport Policy, 2021, 111.

[15] LI D, ZHANG Y, LI C. Mining public opinion on transportation systems based on social media data[J]. Sustainability, 2019, 11(15).

[16] ZHOU Y S, WANG X Q, KUM F Y. Sustainability disclosure for container shipping: A text-mining approach[J]. Transport Policy, 2021, 110.

[17] MIMNO D, WALLACH H M, TALLEY E. Optimizing semantic coherence in topic models[C]. Scotland:Edinburgh, 2011.

[18] LI Y F, PARAMJIT K D, DAVID L D. Two decades of Web application testing—A survey of recent advances[J]. Information Systems, 2014, 43: 20-54.

[19] SIEVERT C, SHIRLEY K E. LDAvis: A method for visualizing and interpreting topics[Z]. 2014.