Analysis of gender sentiment expression in network based on TF-LDA algorithm

Ai Lyu [1, a], Chengyu Liu [2, b], Zhaoxing Ding[1, c], Junying Li[1, d] and Wenjun Zhang [1, e]

[1]College of Culture Communication, Shandong University, China;

[2] Business school, Shandong University, China.

[a] raven020912@163.com, [b] 861252746 @qq.com, [c] 3241538574 @qq.com, [d] 3089620183 @qq.com, [e] 2825702869 @mail.qq.com

Abstract. With the popularity of networks, social media has become an important way to spread gender awareness, in which sentiment expressions can especially reflect the development of feminism and public opinion atmosphere. The team took two of the hottest gender-related news events in 2022 and used crawler technology to collect network data of Weibo comments. Then, we used the TF-LDA model combined with the TF-IDF algorithm and LDA topic model, to screen the keywords respectively, then built a new text based on the filtered results, and made sentiment analysis based on the BosonNLP sentiment lexicons.

Keywords: Sentiment expression; Sentiment analysis; Gender; TF-LDA.

1. Introduction

Analysis of sentiment expression in network is an important way to understand social opinions from a macro level and is the basic premise of network public opinion governance. Our team will use sentiment analysis technology including Python to conduct quantitative analysis of online discourse. The research field mainly focuses on gender issues. On the one hand, gender issues have received a lot of public attention, with heated online discussions and abundant textual materials. On the other hand, gender-related news events are often highly controversial and opposites among expression are of typical analytical value.

Weibo is the social media platform the team chose to crawl data. Since 1995, Chinese feminism network media began to develop rapidly [1] among which Weibo is the most representative. According to sociologist Mark S. Granovetter [2], Weibo is one of the largest and most active "Weak Ties" network social platforms in China, with high research value. It provides rich materials for the study of gender sentiment expression.

2. Analysis process with TF-LDA algorithm

2.1 Data preprocessing

The research data used in this paper are all from Weibo. Our team chose "Attack on female diners in Tangshan China" and " Overturn abortion rights in USA" these gender-related news events as representatives, which once caused heated discussion among netizens on Weibo. We used web crawler technology to crawl the comment section of the news report, and convert the content of the comment section into plain text. In the early stage, our team conducted Jieba segmentation through existing corpus and self-built thesaurus, finally sorting out 3,600 pieces of data as the initial text for algorithm analysis.

Table 1. The number of comments scraped

Subject	Amount
Attack on female diners in Tangshan China	2100
Overturn abortion rights in USA	1000

2.2 Starting with TF-IDF algorithm

After getting the initial text data, we used the TF-IDF algorithm to judge the frequency of terms appearing in the text, in order to screen out the terms that played the most important role.

TF-IDF(Term Frequency–Inverse Document Frequency) aims to explore the importance of a term in a corpus or an article. The kernel of TF-IDF is that the importance of terms increases in direct proportion to the number of their occurrences in the file, but decreases in inverse proportion to the frequency of their occurrences in the corpus. TF-IDF is widely used in the information retrieval algorithm of search engines. The calculation formula is: TF-IDF = The number of occurrences of a term or term in a document/the total number of terms or terms in the document * log (the number of all documents/(the number of documents containing the term or term) +1).

Among them, TF(Term Frequency) represents the frequency of terms appearing in the text: $tf_{ij}=\frac{n_{i,j}}{\sum_k n_{k,j}}$; While IDF(Inverse Document Frequency) measures the universal importance of terms: $idf_i = \log\frac{|D|}{|\{j:t_i \in d_j\}|}$

After processing the text data with the TF-IDF algorithm, we finally sorted all the terms according to their importance, then selected the top 500 terms to form a new text database.

2.3 Using LDA topic model

In addition to the TF-IDF algorithm, LDA(Latent Dirichlet Allocation) is also one of the algorithms used by our team. LDA algorithm is mainly used to predict the topic distribution of texts. The core idea of the LDA is that a topic can be represented by a free distribution of words, while an article can be represented by a different distribution of topics. The topic of each text is given in the form of probability distribution, and then the text will be topic clustered or classified. This algorithm does not take the order of words into consideration, but uses the bag-of-words model to represent the text. We estimate the maximum likelihood of parameters α and β and establish a three-layer model of LDA: $I(\alpha, \beta) = \sum_{i=1}^{M} \log p(d_i|\alpha, \beta)$

In order to determine the number of optimal topics and ensure the rationality of LDA training results, we adopted the recognized index of Perplexity. Perplexity[3] is a deterministic judgment index carried out by the model when distinguishing topics, reflecting whether the model is applicable to new samples and can correctly distinguish topic division. The probability distribution model with low degree of perplexity can better predict the sample. We calculated perplexity on a number of different topics. The results show that the perplexity decreased with the increase in the number of topics, and the decline trend is becoming smaller. At the same time, considering that the number of experimental texts in this study is relatively small, the number of topics should not be set too high. We finally decided to set 10 as the optimal number of topics in the LDA experiment.

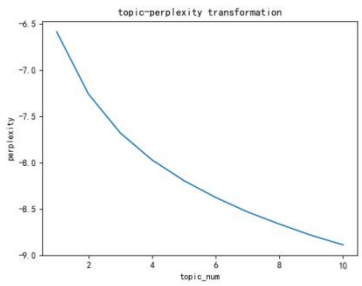

Fig. 1 Topic - perplexity transformation

After removing the stop words, we used LDA to separate 10 topics, select the top 50 words with the highest contribution under each topic, and store them in the new data text. This completed the preparation for the gender sentimental expression analysis phase.

0:00	1:00	2:00	3:00	4:00	5:00	6:00	7:00	8:00	9:00
"不能" (0.024)	"加油" (0.017)	"几个" (0.014)	"姐姐" (0.041)	"女生" (0.056)	"严惩" (0.059)	"搭讪" (0.097)	"没有" (0.023)	"没有" (0.039)	"搜狐" (0.014)
"好好" (0.018)	"团团" (0.017)	"性别" (0.012)	"真的" (0.023)	"不要" (0.024)	"男性" (0.018)	"性骚扰" (0.066)	"女孩" (0.020)	"不要" (0.034)	"男宝" (0.010)
"重判" (0.016)	"唐山" (0.013)	"阳刚" (0.009)	"没有" (0.021)	"真的" (0.024)	"女性" (0.015)	"这是" (0.041)	"希望" (0.019)	"坏人" (0.013)	"害怕" (0.009)
"没有" (0.013)	"中国" (0.013)	"不要" (0.009)	"死刑" (0.014)	"没有" (0.015)	"受害者" (0.013)	"不是" (0.037)	"社会" (0.010)	"不是" (0.013)	"媒体" (0.009)
"真的" (0.011)	"女人" (0.008)	"犯罪" (0.009)	"勇敢" (0.014)	"三个" (0.012)	"希望" (0.011)	"杀人" (0.031)	"保护" (0.010)	"事情" (0.011)	"时间" (0.008)
"身体" (0.008)	"不是" (0.008)	"视频" (0.008)	"不要" (0.010)	"透露" (0.011)	"保护" (0.010)	"未遂" (0.022)	"人民" (0.010)	"气死" (0.010)	"这件" (0.007)
"死刑" (0.007)	"没有" (0.008)	"医院" (0.008)	"希望" (0.008)	"视频" (0.008)	"正义" (0.010)	"骚扰" (0.015)	"国家" (0.008)	"不会" (0.009)	"几天" (0.006)
"女孩" (0.007)	"真的" (0.007)	"不能" (0.008)	"美好" (0.008)	"四名" (0.008)	"不要" (0.010)	"冲突" (0.014)	"真相" (0.008)	"惩罚" (0.009)	"搭讪" (0.006)
"加油" (0.006)	"男人" (0.007)	"对立" (0.008)	"难受" (0.007)	"标题" (0.008)	"事情" (0.008)	"耍流氓" (0.011)	"女孩子" (0.008)	"加油" (0.008)	"女生" (0.005)
"严惩" (0.006)	"希望" (0.007)	"评论" (0.008)	"抱抱" (0.007)	"完美" (0.007)	"不是" (0.008)	"不会" (0.010)	"四个" (0.007)	"热度" (0.007)	"不了了之" (0.005)
"女孩子" (0.005)	"社会" (0.006)	"发声" (0.006)	"愤怒" (0.007)	"出警" (0.006)	"施暴者" (0.008)	"唐山" (0.008)	"发声" (0.007)	"受害者" (0.007)	"老板" (0.005)
"报警" (0.005)	"晚上" (0.006)	"唐山" (0.006)	"发生" (0.006)	"勇敢" (0.006)	"可怕" (0.008)	"除恶" (0.008)	"真的" (0.006)	"有没有" (0.007)	"第一个" (0.005)
"很棒" (0.005)	"报道" (0.006)	"重伤" (0.006)	"垃圾" (0.005)	"官方" (0.006)	"法律" (0.007)	"扫黑" (0.008)	"拘留" (0.006)	"有人" (0.006)	"团团" (0.004)
"照顾" (0.005)	"打人" (0.005)	"事件" (0.006)	"证据" (0.005)	"两个" (0.006)	"人民" (0.006)	"女孩" (0.008)	"受伤" (0.006)	"强奸" (0.006)	"没有" (0.004)
"谁家" (0.005)	"绝望" (0.004)	"女生" (0.006)	"治安" (0.005)	"受害者" (0.005)	"家属" (0.005)	"打人" (0.007)	"凌晨" (0.006)	"家属" (0.006)	"女性" (0.004)
"女人" (0.005)	"孩子" (0.004)	"暴力" (0.006)	"支持" (0.005)	"不是" (0.005)	"犯罪者" (0.005)	"女性" (0.006)	"世界" (0.005)	"发声" (0.005)	"直播" (0.004)
"依旧" (0.005)	"伤害" (0.004)	"未遂" (0.005)	"心疼" (0.005)	"地方" (0.005)	"社会" (0.005)	"明明" (0.006)	"不要" (0.005)	"好好" (0.005)	"情况" (0.004)
"女生" (0.005)	"力量" (0.004)	"女性" (0.005)	"恶心" (0.005)	"没错" (0.005)	"小姐姐" (0.005)	"殴打" (0.006)	"法律" (0.005)	"后续" (0.005)	"发生" (0.004)
"做错" (0.005)	"行凶" (0.004)	"杀人" (0.005)	"国家" (0.005)	"事件" (0.005)	"消息" (0.005)	"女孩子" (0.005)	"好好" (0.005)	"女孩" (0.005)	"女孩" (0.004)
"不是" (0.005)	"判刑" (0.004)	"女孩" (0.005)	"不会" (0.004)	"公布" (0.005)	"女孩" (0.005)	"黑社会" (0.005)	"视频" (0.005)	"几年" (0.004)	"能量" (0.004)
"犯罪" (0.004)	"几个" (0.004)	"真的" (0.005)	"世界" (0.004)	"宝贝" (0.005)	"性别" (0.004)	"社会" (0.005)	"活着" (0.005)	"伤害" (0.004)	"姑娘" (0.004)
"社会" (0.004)	"视频" (0.004)	"彻查" (0.004)	"有没有" (0.004)	"女性" (0.005)	"罪犯" (0.004)	"上手" (0.004)	"姑娘" (0.005)	"法律" (0.004)	"矛盾" (0.004)
"干净" (0.004)	"法律" (0.004)	"发生" (0.004)	"转发" (0.004)	"保护" (0.004)	"恐怖" (0.004)	"单方面" (0.004)	"全国" (0.005)	"希望" (0.004)	"第一" (0.004)
"深夜" (0.004)	"说话" (0.004)	"世界" (0.004)	"十年" (0.004)	"害怕" (0.004)	"真的" (0.004)	"群殴" (0.004)	"女子" (0.004)	"施暴者" (0.004)	"五年" (0.004)
"活着" (0.004)	"女性" (0.003)	"打人" (0.004)	"小姐姐" (0.004)	"坏人" (0.004)	"不能" (0.004)	"谢谢" (0.004)	"不是" (0.004)	"勇敢" (0.003)	"群众" (0.004)
"出发" (0.004)	"不好" (0.003)	"拘留" (0.004)	"中国" (0.004)	"怀疑" (0.004)	"视频" (0.004)	"黑恶势力" (0.004)	"唐山" (0.004)	"出手" (0.003)	"报警" (0.004)
"拖进" (0.004)	"心理" (0.003)	"地方" (0.004)	"女生" (0.004)	"监控" (0.004)	"教育" (0.004)	"希望" (0.004)	"想象" (0.004)	"放弃" (0.003)	"人民" (0.004)

Fig. 2 LDA results (partial)

2.4 Reflection: advantages and disadvantages of the TF-LDA algorithm

As a new algorithm created and used in this experiment, TF-LDA has the common advantages of the two algorithms: 1. The content is more comprehensive, solving the shortcoming of TF-IDF's heavy reliance on corpus. When using LDA topic models, the corresponding theme and meaning are clearer. It not only solved the problems of TF-IDF, but also combined the advantages of TF-IDF's convenient operation and wide coverage, which can strengthen the accuracy of the analysis conclusion. 2. Although LDA topic models can extract corresponding topics and keywords covered by topics, the weak correlation between topics is easy to lead to the instability of content. So if we only use LDA topic models, the keywords under multiple topics are actually repeated, but they actually have different contributions, so it is difficult to choose. Combination with TF-IDF algorithm can help solve this problem, by which a wealth of words can be selected from, with more concentrated content and closer relevance.

However, there are still some deficiencies in the team's approach: 1. The sensitivity of TF-IDF algorithm to rare words, place names and personal names is very low. Additionally, its extraction of complex statements is somewhat shallow. 2. The scoring standard of BosonNLP emotional dictionary is closer to that of written language, while network language is more free and irregular.

In response to the above problems, the team will take measures to upgrade the BosonNLP sentiment lexicons, introduce more words, and separate modal particles, adjectives and adverbs for more accurate ratings.

2.5 Scoring with BosonNLP sentiment lexicons

In the analysis of sentiment expression tendency, scholars have made an endless stream of research achievements in recent years, and most of which form the sentiment lexicon for further analysis. If the construction of such sentiment lexicon is based on pure manual processing, it will cost a lot of manpower and material resources, and its specificity is poor, difficult to cover a variety of different fields. Therefore, many scholars have also studied the construction methods of sentiment lexicon. Our team constructed the sentiment lexicon by combining knowledge base (BosonNLP sentiment lexicons) and corpus (obtained comment data) together. BosonNLP sentiment lexicons, is a universal corpus produced by combining various classified data such as news, microblog and comments, and it contains new words such as irregular network terms and misspellings. According to the data on BosonNLP's official website, the accuracy rate of word segmentation and part-of-speech tagging was 94.62% [4].

Through the TF - IDF and LDA algorithm, we have already picked out the most important target words respectively. Then we took the same words from the two to build a new corpus, to ensure that the words in the corpus are comprehensive. Finally, this corpus data is imported into the BosonNLP sentiment lexicons and scored against keywords in the lexicons. We then got the total and average sentiment score of the text, as well as the proportion of positive or negative emotions.

2.6 Analysis result

First, we find that words such as "gender", "female" and "human rights" appeared very frequently in the comments below the two events. It can be seen that the female consciousness of Chinese netizens has been awakened to a large extent. They are eager to speak out in the event of violation of women's rights and provide their feedback to the parties or relevant departments.

Secondly, in the comments on " Overturn abortion rights in USA" and "Attack on female diners in Tangshan China", netizens both had more negative emotions than positive ones, which accounted for 58.4% and 41.6% in all. This shows that with the emergence of relevant issues, some netizens use relatively more negative words to express their inner anger. Particularly in the beginning of this case, the news reported the male behavior as "chatting up" rather than "sexual harassment", which caused some netizens' dissatisfaction. In addition, anger is the easiest emotion to interact with other emotions, thus spreading the fastest comparatively. Although this anger emotion can speed up the dissemination of information to some extent, making it widely discussed on the network and form a public topic, in the long run, it is not conducive to the rational solution of the event itself.

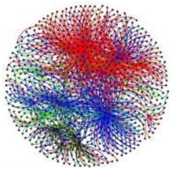

Fig. 3 The spread of various emotions among Weibo users [5]
The red line represents anger

Thirdly, among the topics modeled by the LDA algorithm, negation including "don't" and "can't" account for a large proportion (up to 16%). It can be seen that, besides the discussion of the news event itself, the public is also concerned about how they should behave. "Don't go out late at night as woman", " can't go to America", etc. The demand for individual behavior reflects the public's anxiety to avoid violations of women's right. Therefore, the systematic management of social problems is the real outlet of gender sentiment expression in network.

2.7 Prospect

As for the research framework of text sentiment analysis used in this paper, there are still the following deficiencies which need to be further improved and perfected.

1. The choice of text corpora data sets: Our team only crawled the text data from comment section on the Weibo platform. In other words, the data source is relatively simple. However, most of the users of the Weibo platform are youngsters [7], so they can not cover the sentiment expressions of people of all ages. In future studies, the representativeness of data set selection needs to be improved. Network text data from different media platforms can be used comprehensively as the data basis for sentiment expression analysis.

2. The text sentiment analysis: This paper made a simple dichotomy of sentiment expression, that is, set the sentiment score of positive emotion text as positive number, while score of negative emotion text as negative number. However, it lacks the processing of texts with relatively vague emotional tendency and neutral attitude. Therefore, in the future research, it is necessary to make a more fine-grained distinction on the emotion analysis of the text and conduct multidimensional sentiment tendency research.

3. Summary

Through analysis of gender sentiment expressions, we can get a sense of the general attitude towards gender issues and the current atmosphere of online discussion. Although network ecology has a negative emotional tendency, overall still looking for solutions to the problem. Sentiment is not the opposite of fact, but the constituent element of cognitive psychology and rational system, as well as a resource of social mobilization and social integration[6]. However, in the post-truth era, we must also strengthen the governance of sentiment expression, guard against the excessive use of sentiment, and avoid such gender issues becoming particular memes for chasing clout.

References

[1] Yi Xiaorong, The Research of Feminism Images in Network Media, 2018, Jinan University, MA thesis.

ISSN:2790-1688 Volume-5-(2023)

[2] Mark S. Granovetter. The Strength of Weak Ties. American Journal of Sociology，1973（6）. 1360-1380.

[3] ARUN R, SURESH V,MADHAVAN C E V, et al. On finding the natural number of topics with latent Dirichlet allocation: some observations[C] //Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2010： 391-402.

[4] bosonnlp. com/tag. Html

[5] technologyreview.com

[6] The Emotions, Prejudices and “the Miracle of Aggregation” in the Public Opinion: from the Concept of the “Post-truth”. Chinese Journal of Journalism & Communication. 2019, 41(1): 115-132s

[7] data.weibo.com/report