Text Mining Analysis on Users' Reviews for News Aggregator Toutiao

Text Mining Analysis on Users' Reviews for News Aggregator Toutiao

Jia Li[1,#], Bifeng Wang[2,#], Alexandra Jingsi Ni[3,#] and Qian Liu[4,*]

[1]International School, Jinan University, Guangzhou, Guangdong, China
[2]School of Art Education, Guangzhou Academy of Fine Arts, Guangzhou, Guangdong, China
[3]The Center for East Asian Studies, University of Turku, Turku, Finland
[4]School of Journalism and Communication, National Media Experimental Teaching Demonstration Center(Jinan University), Jinan University,Guangzhou, Guangdong, China

[*]Corresponding author's e-mail: tsusanliu@jnu.edu.cn

[#]Jia Li, Bifeng Wang and Alexandra Jingsi Ni share the co-first authorship

Abstract. This paper intent to improve consumer communication by text mining analysis with users' reviews. The news aggregator we focus on is: Toutiao, known as "today's headlines" in Chinese. It is the top news aggregator application run by Bytedance company in China. It utilizes AI algorithms to provide numerous news feed for its users. As new technologies are shaping the business strategy studies as well as online communication analysis, it requires innovative and effective analyses of unconventional data, such as the 12,290 online reviews on Toutiao we collected from Apple's App Store. Through the LDA topic modelling and sentiment analysis, our research has identified three major negative complains the consumers have regarding Toutiao application, namely: too many unsolicited advertisements, contents (vulgar content, time consuming video, privacy and copyrights infringement issue) and incompatibility with the latest Apple digital devices.

1. Introduction

Jinri Toutiao (today's headline) or Toutiao (headline) is a mobile News application offered by ByteDance. It is the most popular news application in Mainland China. However, on 10th April, 2017, the State Administration of Radio and Television (SAPPRFT) ordered Jinri Toutia to permanently shut down the user software and public accounts for "Neihan Duanzi" and comprehensively clean up the irregular and improper content which triggered strong resentment among internet users on their platform[1]. *Also,* on 29th of December 2017, regulatory authority had temporarily shut Jinri Toutiao down for 24 hours due to the alleged violations of Internet regulations, such as disseminating "*pornographic and vulgar content*"[2]. It is estimated that Toutiao has a daily active user over 120 million in mainland China alone.

2. Literature review

User-generated comments and reviews could reveal valuable information regarding the user experience of a particular application. Online ratings, reviews and comments of an application reflect actual perceptions and attitudes of the user. It could provide useful clues for application developer to

improve the quality of the application and maintain long-term focused customer loyalty. Therefore, it has been widely observed that digital interface, systemic analysis and customer supporting service are becoming more and more interdependent and integrated these days[3].

It's quite common that infringement (of copyrights) could be easily found in the user-generated contents. With the growing size of the so-called online communities and multiplying digital hubs, user do perform more than just one, and often predetermined, role[4].

3. Method

3.1. The LDA (Latent Dirichlet allocation) Topic Modeling

In the age of information overload, the LDA is commonly deployed to identify topics hidden in massive amount of textual data simply because the digital textual data generated day-by-day via the internet is way beyond the processing capacity of human beings. According to the literature consulted, the LDA topic model is a three-level Bayesian model arranged in hierarchical order. This model is grounded on the assumption that a document is a complex combination of words forming different topics[5]. Using Gibbs Sampling could readily identify topics[6]. Comprehensive studies have already been conducted to achieve the objective of parameter optimization through more efficient and effective solutions[7]. In one step further, some studies even present impressive visualizations of the topic modelling results[8] or empirical applications of the LDA modelling technique in different cases etc[6-7][9-10].

3.2. Sentiment Analysis

Sentiment analysis is a qualitative data-treating technique to excavate indications of attitudes from textual contents. Its top priority is to identify different streams of opinions according to many experienced researchers[11]. The Tencent Wenzhi system (nlp.qq.com) is a commercial platform to treat massive textual datasets through the assistance of AI technologies.

3.3. Data collection and refinement

We have collected our raw data (in the forms of online reviews and comments) from Tencent Fsight (http://fsight.qq.com/), in a preliminary search from October 10, 2017 to January 10, 2018, we gathered 12,290 reviews, after data cleaning, only 2,076 legitimate reviews and comments left. Applying sentiment analysis to the refined data, as shown in Table 1, we have calculated a scorea from 0 to 100. 0 represents the most negative, 100 represents the most positive and 50 represents a neutral position.

Table 1. User views Distribution and Sentiment Analysis Score

Data Source	Number of Comments	Average of Sentiment Score
360 Mobile Assistant	79	51
App Store	138	44
Vivo app	485	58
Wandoujia	683	57
Mi app store	538	53
Yingyongbao	150	53
Total	2076	55

We collected 1051 views regarding the Toutiao news application from App Store from 9th of January, 2017 to the 9th of January, 2018. It's noteworthy that Toutiao has some duplicated comments, after deleting 55 repeated, 1096 reviews and comments via the App Store over the time period from the 9th of January, 2017 to the 9th of January, 2018 were collected. Systematically analysis was done after the preparing process of the data in our research.

3.4. Processing the data

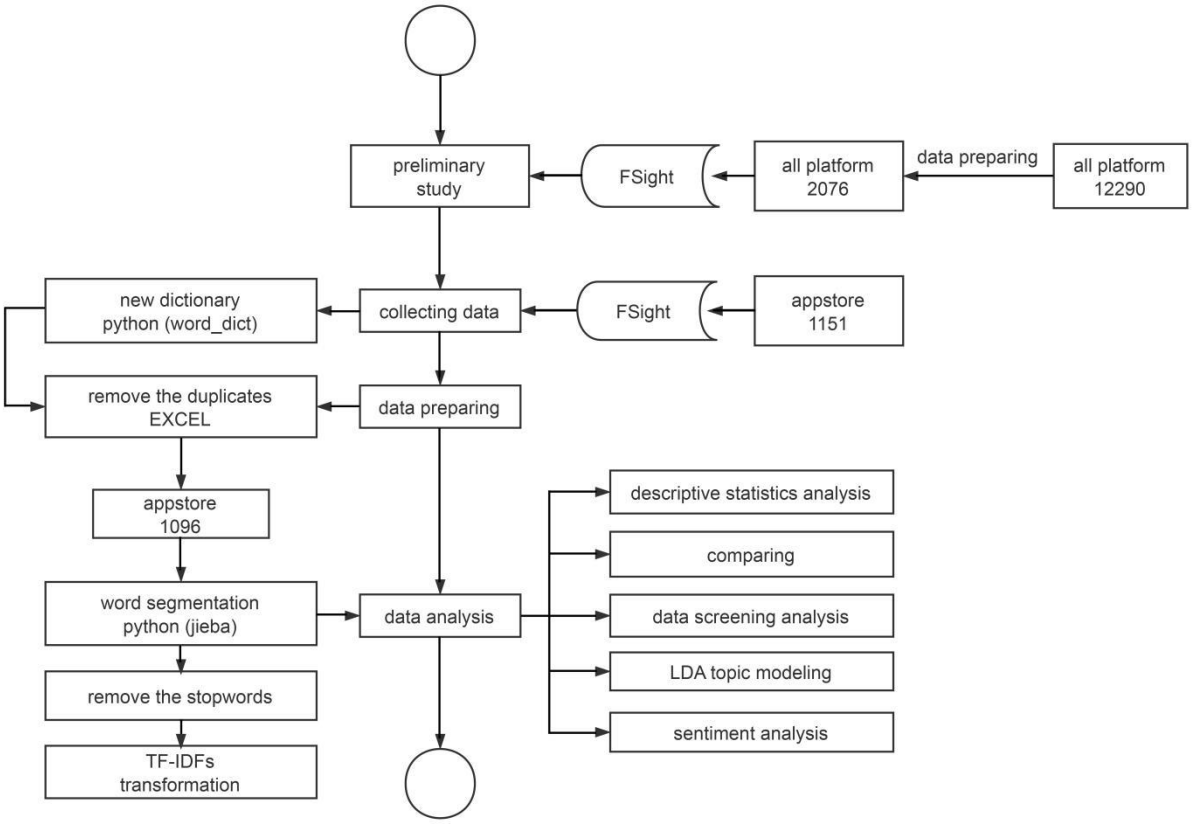

Figure 1. Collection, preparing and analysis of data.

As shown in Figure 1, before applying the LDA modelling and sentiment analysis, the collected data has gone through a process called segmentation since texts encoded in the Chinese language are formed by strings of equally spaced characters. The 1096 comments collected are preliminarily processed with the available open-source segmentation systems that are widely used by researchers, such as the python package Jieba[12-14].

The online reviews and comments contain rich and diverse linguistic traits and variations. Unconventional vocabularies are added to the dictionary for analytical purposes, such as “Toutiao”, and “Penzi” (literally means haters in English).

Having done the above, words bearing little information, such as “one”, “of”, “and”, “or” have also been filtered purposely. The LDA modelling requires a document-term matrix (DTM) to process. We have established a DTM to meet the requirement and test the refined datasets based on the Inverse Document Frequencies (the IDFs). This approach could help to assign weights to different terms of varied frequencies so as to demonstrate the unique profile of a given piece of textual content.

The LDA model is determined by two parameters, namely the quantity of the topics and the Dirichlet hyper parameter $\alpha = \frac{50}{Topic\ number}, \beta = 0.1$[6]. With properly weighted parameters and systematic algorithm, sophisticated calculations could be performed to reveal even further the actual perceptions and attitudes of the user regarding Toutiao in a number of meaningful ways.

4. Result and discussion

With the aforementioned LDA topic modelling and sentiment analysis, our study intends to answer the following 3 research questions:

1. What are the quantitative results of sentiment analysis of the online comments made by the user of Toutiao?
2. What keywords could be generated, what topics and could be identified from these online reviews and comments and how they are related to the concerns of the user of Toutiao?
3. What categories could be classified after analysis of the topics?

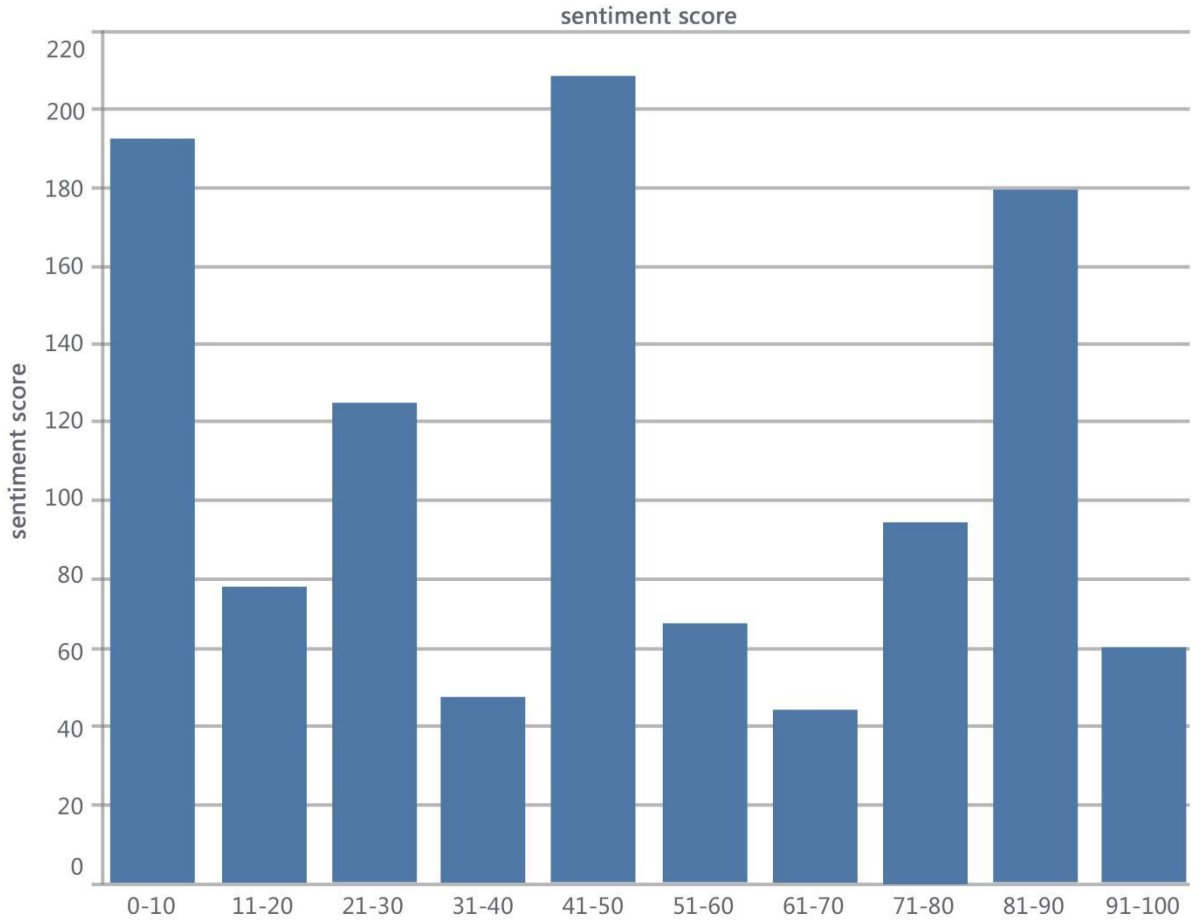

Figure 2. Sentiment analysis score.

For the first question, we have come up with an average sentiment analysis score of 44 and we could conclude from the results that user satisfaction is not very optimistic since the average score is actually below 50, i.e. the neutral position. One of the attributions that deserve to be mentioned is: two subdivisions of user tend to have an over-representation in the data collected. They are very likely to be A) extremely happy user and B) hyper- participatory user. They do appear at both ends in Figure 2, indicating that these users tend to provide significant amount of follow-up reviews and comments than the rest of other users.

To answer question 2 and 3, we have treated our datasets with Python in Version 3.6.1. The quantity of the topics is still one amongst the three parameters required. Many researchers have invested incredible effort to try to discover the optimal quantity of topics for LDA analysis and they have been trying a variety of innovative ways to interpret each and every of the topics they have identified [8]. According to what the statistical data suggests, the results do not automatically guarantee effective interpretations in the subsequent phase[15]. Therefore, we come up with 8 topics for further analyses, listed in Table 2. The 8 topics are specified and they are divided into 4 categories.

Table 2. Topic Classification and Keywords

Classification groups	Topic names	Key words	Proportion
1.Support voice	Topic 2: Pleasant User experience	Jinri Toutiao, readability, software, refresh, everyday, worthy	9.50%
	Topic 3: Good classification	Toutiao, channel, not bad, charge, smart phone	11.20%
	Topic 6: Praise	Support, good, hope, display, mark, upgrade, professional, five stars	8.00%
	Topic 4: Rich content	Like, download, useful news, good quality, content richness	4.90%
2. Too many unsolicited advertisements	Topic 1: Advertisement issue	Ads, rubbish, more, Toutiao, recommend, push, too much/many	28.10%
3. Contents (vulgar content, time consuming video, privacy and copyrights infringement issue)	Topic 5: Content problem	loading, comment, similar, duplicated content, too much/many Video, version, picture, vulgar, illegal,	15.80%
4.Incompatibility with the latest Apple digital devices.	Topic 7: Login and replay issue	Login, replay, solve, repair, Weibo, Update, download, Phone, Replay, disgusting	5.40%
	Topic 8: iphonex adaptation issue	iPhone X, compatibility, Refresh, Toutiao, not good, news, efficiency	5.40%

There are 4 positive topics revealing good user experience in the table. They account for up to 33.6% of all the reviews and comments collected by us. These reviews and comments reflect positive perceptions of Toutiao as a news application by its user, such as the content provided is well-organized and of high quality. In contrast, the 3 negative topics are illustrated as below:

(1) Many users state that there are too many unsolicited advertisements as shown in Table 3. They could seriously degrade user experience in many ways. Examples of typical comments that fall into this category are presented in Table 3:

Table 3. User' feedback on unsolicited advertisements with calculated sentiment analysis scores.

User' feedback on unsolicited advertisements	Time	Sentiment scores
Too much unsolicited advertisements, low quality news, irrelevant and redundant information.	2017-10-2303:29:04	23
There are just too many advertisements embedded in this application. It's no longer a news app but an ad app.	2017-12-3109:49:07	50
I am sick and tired of this news app and I have finally deleted it. I was a happy user once upon a time. Now I am seriously disgusted by its saturation with multifarious advertisements. I suppose other alternatives might work better for me.	2017-11-0710:06:39	38
It's a shameful news app! After upgrading, all I have been provided are endless advertisements!	2017-10-2119:14:41	15

(2) The shrinking quality of the content is also a major concern of the user of Toutiao. Quite a few users complained that the content is vulgar and meaningless. Privacy and copyrights infringement issues are also mentioned.

Recommended content, especially the short videos provided with the intention to prolong browsing time of the viewers, is a new function of Toutiao with the assistance of sophisticated AI technologies. However, some users are quite annoyed by this fancy new function, stating that: "too much highly similar contents!" while other claiming that "It's just boring and meaningless! No one likes it!"

(3) Incompatibility with the latest Apple digital devices, including low loading speed, frequent restarts after upgrading the application and login problems. In particular, iPhoneX compatibility-related issues have an average sentiment analysis score of 45, which is slightly below the neutral position. Many users of Toutiao are quite upset about the incompatibility of the application with their latest Apple digital gadgets. Many of them suggested that other content providers are doing (much) better than Toutiao in comparison. Examples of the negative comments regarding incompatibility are listed below:

Table 4. User' feedback on IPhone-X incompatibility issues.

User' feedback on IPhone-X incompatibility issues	Time	Sentiment analysis scores
iPhone X doesn't function properly with this app. It pauses each time after clicking the news tag. Really unpleasant user experience!	2017-11-12 13:41:10	25
Why the app has done nothing to try to accommodate iPhone X? Netease, Phoenix and others are doing much better. What's wrong with Toutiao?	2017-11-06 00:19:05	25
iPhone X is not well supported, the speed is too slow. Can't the app be a bit more efficient?	2017-11-0418:57:24	29

5. Conclusion

In conclusion, with the assistance of Latent Dirichlet allocation (LDA) and sentiment analysis, our research has deployed these two data-treating techniques to critically analyse the amount of 1,096 online comments we have carefully collected from the App Store regarding their user experience with Toutiao. The overall average sentiment score we have calculated is 53, just slightly above the neutral position, which signals potentially low customer satisfaction. Our research paper has also identified 8 topics from the online comments left by user of Toutiao. These 8 topics are classified into 4 main categories: **A)** positive feedback, **B)** malicious contents (*including vulgar content, time consuming video, privacy and potential violations of the intellectual property rights in many cases*), **C)** unsolicited advertisements and **D)** compatibility issues with the latest IPhone devices. Over the course of our research, we found that these active user of Toutiao are quite participatory rather than passive according to convincing empirical evidences. We believe that they are significantly empowered by digital media platforms to articulate their needs and wants as well as to defend their legitimate interests as user in the Age of the Internet, including even the, often underserved, niche user (with the newest end devices).

References

[1] Meng,J, "Chinese media regulator clamps down on Toutiao news site for posting vulgar content. *South China Morning Post*," May 1, 2018, [Online]. Available: http://www.scmp.com/tech/article/2141124/chinese-media-regulator-clamps-down-toutiao-news-site-posting-vulgar-content.

[2] Kristin, H, "China takes popular news app Toutiao offline for 24 hours over pornographic content," *South China Morning Post*. Feberary 1, 2018, [Online]. Available: http://www.scmp.com/news/china/policies-politics/article/2126197/china-takes-popular-news-app-toutiao-offline-24-hours.

[3] Turel, Ofir, and Yufei Yuan, "User acceptance of Web-based negotiation support systems: The role of perceived intention of the negotiating partner to negotiate online," *Group Decision and Negotiation* 16, no. 5, pp. 451-468, 2007.
[4] Ritzer, G., & Jurgenson, "Production, consumption, prosumption: The nature of capitalism in the age of the digital 'prosumer'," *Journal of Consumer Culture* 10(1), pp. 13–36, 2010.
[5] Blei, David M., Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research* 3, pp. 993-1022, Jan, 2003.
[6] Griffiths, T. L., & Steyvers, M, "Finding scientific topics," *Proceedings of the National Academy of Sciences 101*(suppl 1), pp. 5228–5235, 2004.
[7] Hoffman, M., Bach, F. R., & Blei, D. M, "Online learning for latent dirichlet allocation," *In advances in neural information processing systems*, pp. 856–864, 2010.
[8] Sievert, Carson, and Kenneth Shirley, "LDAvis: A method for visualizing and interpreting topics," In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pp. 63-70, 2014.
[9] Jiang, HanChen, MaoShan Qiang, and Peng Lin, "Finding academic concerns of the Three Gorges Project based on a topic modeling approach," *Ecological indicators* 60, pp. 693-701, 2016.
[10] Liu, Qian, Qiuyi Chen, Jiayi Shen, Huailiang Wu, Yimeng Sun, and Wai-Kit Ming. "Data analysis and visualization of newspaper articles on thirdhand smoke: a topic modeling approach." JMIR medical informatics 7, no. 1, e12414, 2019.
[11] Pang, Bo, and Lillian Lee, "Foundations and Trends® in Information Retrieval," *Foundations and Trends® in Information Retrieval* 2, no. 1-2, pp. 1-135, 2008.
[12] Peng, K.-H., Liou, L.-H., Chang, C.-S., & Lee, D.-S,"Predicting personality traits of Chinese users based on Facebook wall posts," *IEEE In Wireless and Optical Communication Conference (WOCC)* 24th, pp. 9–14, 2015.
[13] Zhang, Xiang, Junbo Zhao, and Yann LeCun. "Character-level convolutional networks for text classification." In *Advances in neural information processing systems*, pp. 649-657. 2015.
[14] Liu, Qian, Zequan Zheng, Jiabin Zheng, Qiuyi Chen, Guan Liu, Sihan Chen, Bojia Chu et al, "Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: Digital Topic Modeling Approach," *Journal of Medical Internet Research* 22, no. 4, e19118, 2020.
[15] Grimmer, Justin, and Brandon M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political analysis* 21, no. 3, pp. 267-297, 2013.

Acknowledgement
This paper was funded by National Social Science Foundation of China(no. 18CXW021).