An Empirical Study on the Clickbait of Data Science Articles in the WeChat Official Accounts

Shuyi Wang(✉) and Qi Wu

School of Management, Tianjin Normal University, Tianjin 300387, China
nkwshuyi@gmail.com, cathyqi055@gmail.com

Abstract. In the Internet age, clickbait is an effective method to attract people's attention, which usually uses some ways to achieve, such as exaggerating, omitting the details or using punctuation exceedingly. In order to attract the readers to click on the link, some news aggregator sites or social media will choose to use the clickbait. If the clickbait applied to scientific articles, not only will affects the article quality, but also will affects the development of relevant subjects. Thus, the purpose of this research is to explore whether there is a clickbait in the data science articles of WeChat official accounts. This paper collects the relevant data by using the shenjianshou platform, and then uses some steps to analyze data, including cleaning data, doing word segmentation, extracting keywords and building a regression model. According to the adjusted r-square value in the regression model, the model can only explains the change of 3.17% page views, which means that the clickbait phenomenon is not prominent in the data science articles of WeChat official accounts. Finally, the regression analysis results are discussed from subject perspective, writer perspective and reader perspective.

Keywords: Clickbait · WeChat official account · Data science
Keyword · Page view

1 Introduction

The clickbait refers to some very arresting titles on the Internet, and its purpose is to attract readers to click, in order to get a lot of benefit [1]. Writers will use some methods to achieve, including exaggerating, omitting key details or using punctuation exceedingly in the title [2]. For example, "Relying on this secret, he earned 10 billion in the stock market for 1 year", "She uncovered the sofa mat to see that …… heinous!" But these titles tend to hide a lot of advertisements or junk information, wasting people's valuable attention.

This article is one of the research results of the National Social Science Fund Youth Project "Research on the privacy protection of social media users based on information price dynamic disclosure" (Project Approval No.: 15CTQ017).

J. C. Hung et al. (Eds.): FC 2017, LNEE 464, pp. 131–140, 2018.
https://doi.org/10.1007/978-981-10-7398-4_14

In recent years, social media have gradually become the main channel for people to access information, and have become a platform for self-media to promote, such as Facebook, WeChat and so on. This trend has accelerated the transformation from traditional media to new media. Facing with the fierce competition in the online news market, many media regard the titles as a key factor in attracting people's attention [3]. At the same time, with the advent of fast-paced lifestyles and the spread of fast food culture, people are forced to develop a fast-food reading habit. Fewer readers have time and patience to read the entire article, most readers decide whether to read according to the article title. Plus, the article title is often limited by space in the Sina Weibo, WeChat official accounts and so on. This situation has accelerated the production of the clickbait.

At present, the degree of user activity in the WeChat is very high. It is reported that the number of WeChat official accounts exceeded 12 million in 2016, 52.3% of users use the WeChat official accounts to get the latest information via WeChat official accounts. However, the articles pushed in the WeChat official accounts only show the title, but do not display the content, so the title has become an important factor to make a difference in clicking the article link [4]. In order to obtain high page views, each WeChat official account will makes effort on the title. However, if the clickbait applied to scientific articles, not only will affect the article quality, but also will affect the development of relevant subjects. Thus, this paper focuses on the field of data science. Under the development rush of big data, data science has gradually become a popular research field [5]. In general, the purpose of this study is to explore whether there is a "clickbait" in the data science articles of WeChat official accounts.

The following is organized as follows: Sect. 2 is related work. Section 3 is data collection and processing. Section 4 is the part of building model. Section 5 is discussion. Section 6 is conclusion, which discusses the limitations and future prospects of this paper.

2 Related Work

In recent years, the emergence of clickbait on social media has been on the rise, this phenomenon is partly caused by people's curiosity. Readers click on the article link just to satisfy their curiosity and fill the knowledge gaps, which is called "information gap" theory of curiosity [6]. George Loewenstein explained the concept of "information gap", he thought that curiosity would be produced when people realized the knowledge gaps. The knowledge gaps will led to the feeling of pain and deprivation. At this time, people will be impulsive because they want to get back their lost knowledge at all hazards [6]. Clickbait just takes advantage of this psychological, create a "curiosity gap" to stimulate the reader's curiosity.

Besides, in the "attention economy" era, if the industry can catch the public's attention, it is easy to stand out from the competition. American economist Michael Goldhaber once said: "The information in the network is not only rich, but even flooded. With the development of information, the value is not information, but your attention." [7] Thus, these articles with ultra-high clicks have successfully attracted public's attention with the help of the article title.

At present, there are many methods to solve the clickbait problem. As the world's leading content recommendation platform, Taboola introduces a tool that allows people to remove articles they do not like, including clickbait articles [8]. Facebook has always attached great importance to the users' experience. As early as August 2014, Facebook proposed to boycott the clickbait, and built a system to detect the clickbait [9]. In August 2016, Facebook once again announced to ban the clickbait and made adjustments to the detection system, clickbait articles would be filtered out of the reader's news feed by using the sorting algorithm. In order to solve the clickbait phenomenon in Twitter, Potthast et al. [1] used the top 20 most prolific publishers on Twitter as the data set, and built a related detection model which can help readers reduce the clickbait articles in their news feed.

There are some scholars who have proposed other methods to detect the clickbait. Chakraborty et al. [9] aimed at detecting the clickbait from all online news media by using different characteristic between the "clickbait" articles and "unclickbait" articles. They created a browser extension that readers can filter out clickbait articles. Biyani et al. [2] extracted the various indicators from the article titles, body and web links of the news aggregators. According to these indicators, a machine learning model is created to detect the clickbait phenomenon automatically.

In addition, some scholars have studied the clickbait phenomenon from other angles. Blom et al. [10] have conducted an in-depth analysis of online news headlines from the pragmatics perspective. The results found that some commercial and tabloid media are more likely to use forward referring headlines to lure readers click. Pengnate [11] has studied from the psychology perspective. By using eye tracking devices to measure the change of pupil and evaluate the clickbait article on the degree of reader's emotional awakening. Chen et al. [12] have explored other potential methods from four aspects to detect clickbait automatically, including language, grammar, image and user's reading behavior.

By summarizing the literature, we found that few scholars have studied the clickbait of WeChat official accounts. In this paper, we choose the WeChat official accounts of data science field as the study object, and use tools to collect data about data science articles, including article title, article content, page view and upvote number. By building a regression model to verify whether there is a "clickbait" in the data science articles of WeChat official accounts.

3 Data Collection and Processing

3.1 Data Collection

We mainly use the shenjianshou platform to collect data in the process of data collection. By using the crawler interface of the WeChat articles collection (multi-keyword crawling) which provided by shenjianshou platform, to obtain data about data science articles. In the crawler keyword setting, we enter 12 keywords about data science, including data analysis, data mining, big data, machine learning, depth learning, artificial intelligence, data visualization, data collection, database, data warehouse, data retrieval and crawler. After starting the crawler, there are 8947 articles related to these keywords.

3.2 Data Cleansing

There may be some missing values in the data set. If we retain these missing values, they will affect the accuracy of research result, so data cleansing is required. First, we need to filter out all the articles that posted on the crawling day, because most WeChat official accounts are updated in a limited number in one day. Readers usually make a decision to read or not read the article within one day after updating. Therefore, the page views of articles will be changed significantly after the first day's uploading. Some articles we collected that may include just released less than one day or even less than an hour. These articles may be very popular ultimately, but the page view at the current crawling time is poor. If these articles calculate together, it will produce data perturbation. Next, we need to filter out the missing data of article title, article content, page view and upvote number. In addition, it is necessary to remove some problematic data, such as the upvote number is more than page view, duplicate articles, etc. After data cleaning, the number of data has been reduced to 6698.

3.3 Preparation Work

It is necessary to analyze whether the article quality has an impact on the page view before the regression analysis. We sort the articles by page view, and pick out the top 20 articles, as shown in Table 1. Since the original data is in Chinese, so the tables and figures in this paper are translated from Chinese into English. Figure 1 is a histogram of twenty articles' page view. As can be seen from Fig. 1, the gap between these 20 articles' page view is big. The page view of ranked No.1 article is more than 100,000, while ranked No. 20 article's page view is only more than 60,000. The page views of these 20 articles are far more than the mean and median of the total page views. The mean is indicated by a solid line, and the median is represented by a dashed line. It is important to notice that the top six articles' page views are the same in Fig. 1. This is because that the WeChat in order to avoid the clickbait, plagiarism and other bad behaviors, decided to limit the page view into 100,000.

Table 1. The top 20 articles' titles

No.	Article title
1	Magical magic formula database
2	Artificial intelligence tearing human, just a matter of time? A little bit of fear…
3	Big data definition "city new owner", you find your place?
4	Artificial intelligence will be the next winning point that Baidu promoted?
5	Must read database maintenance announcement
6	Choosing voluntary by "Big Data" need offer 39,800 yuan, is "sky-high price" consultation reliable to help you fill in the voluntary college entrance examination?
7	Looking at unknown Beijing through big data: about salary, rich and poor, house, the Fifth Ring Road
8	Tonight dry goods: deep learning private placement
9	Sogou CTO Yang Hongtao: what kind of position to participate in artificial intelligence?

(*continued*)

Table 1. (*continued*)

No.	Article title
10	(The new regulations) The Supreme Court, the Supreme Procuratorate, the Ministry of Public Security "on the stipulation of electronic data collection and judgment"
11	Big data: get your insurance before 40 years old!
12	It is said that annual salary of more than 500,000 people, are hopelessly in love with the "data visualization"
13	Human lost! Artificial intelligence defeats the world championship, its strength is far more than that
14	The world's largest movie database IMDb selected the highest score movie of 25 years!
15	Huang Xiaoming PK artificial intelligence? Angelababy into a dilemma!
16	Database|08 card Chinese database is online, the player recommended recently launch
17	QQ space broadcast fighting of the iteration artificial intelligence, Alpha dog or has become a past
18	Encryption | Netease user database was leaked, Alipay, etc. may also be hidden
19	Li Yanhong: Do not panic for artificial intelligence, at best, it is " a wolf wearing skin sheep"
20	Artificial intelligence is fully rolled, is human thinking really superiority?

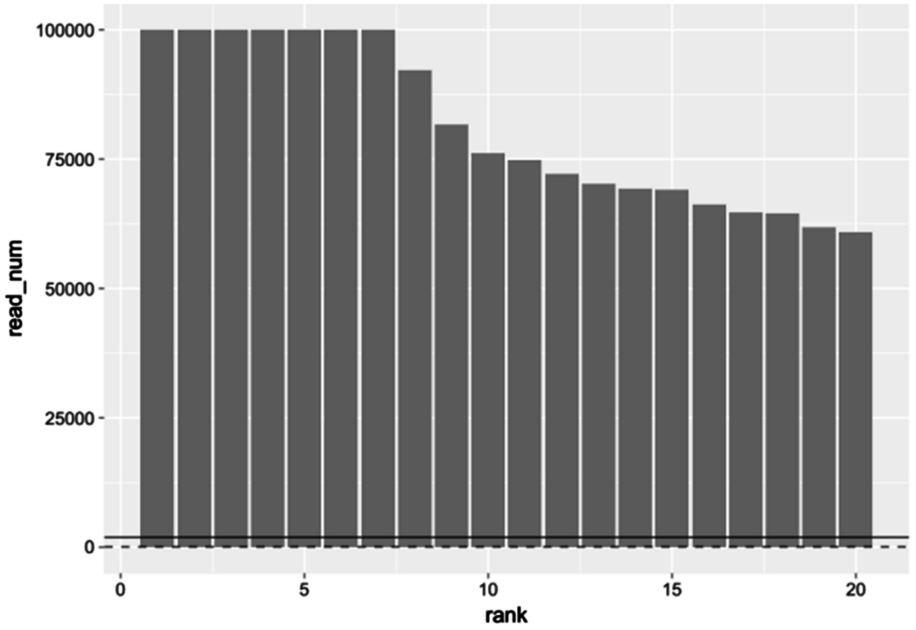

Fig. 1. Page views of the top 20 articles

The quality of article is expressed by the ratio of dividing the upvote number by the page view. Likewise, the mean is represented by a solid line, and the median is represented by a dashed line. There was no one articles whose quality values exceeded the mean, and there were seven articles that did not reach the median, as shown in Fig. 2. Thus, the article quality is difficult to guarantee, even the most popular article. Readers do not know what the quality of this article is, so the article title is the only information to based on.

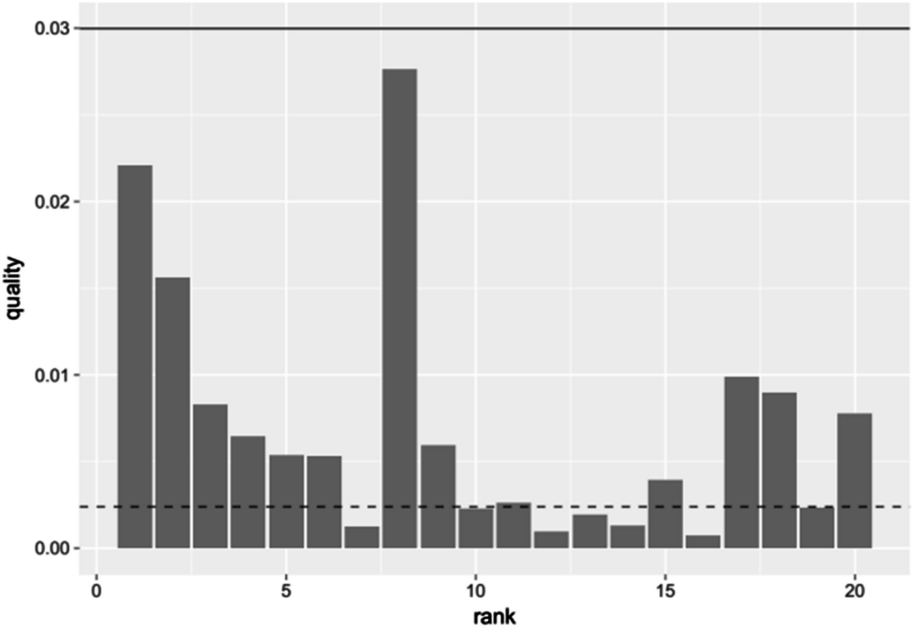

Fig. 2. The quality of the top 20 articles

4 Regression Analysis

4.1 Word Segmentation and Keyword Extraction

In order to remove some stop words and extract the keywords of titles, we need to do work segmentation. In this paper, we use the jieba tool to extract the keywords. Download the list of stop words from jieba tool, and add some new things into it, including common punctuations and the text sentence separated with space.

After the end of word segmentation, there will be tens of thousands of keywords. But not all keywords are useful, so the next task is to remove the useless keywords. Here we use the graphlab to get the keywords data into a dictionary list, each keyword will has a corresponding number. We pick up the three most representative keywords of each title according to the number, and count the frequency of each keyword. Then we sort out the top 200 keywords with higher frequency, and use these 200 keywords as feature variables to construct matrix. Table 2 lists the top 20 keywords with higher frequency.

Table 2. The top 20 keywords with higher frequency

Keyword	Frequency	Keyword	Frequency
Data	90	Learning	29
Database	64	Deep	27
Crawler	59	Application	26
Era	46	MySQL	22
Data analysis	43	Future	21
Data mining	39	Change	21
Visualization	37	Algorithm	21
What	37	Do	21
Artificial intelligence	35	Develop	21
Machine	34	Summary	19

4.2 Regression Model

In the previous section, we got the top 200 keywords with higher frequency. The independent variables are the top 200 keywords, the dependent variable is the page view. We build a regression model to explore the relationship between keywords and page views. To prevent the model from being over-fitting, we need to optimize the model. The more independent variables the model have, the more likely to be over-fitting. In this paper, we use the AIC (Akaike information criterion) rule to simplify the model. By constantly adjusting the AIC parameters to eliminate the lower correlation keywords, and improve the fitting quality of the model. Finally, after several adjustments, the number of independent variables is reduced to 30, including data mining, data analysis, learning, etc.

Since we mainly study Chinese keywords, so we translated the Chinese keywords into English, is to facilitate the understanding of the results of regression analysis, as shown in Table 3. In the regression model, the most critical value is adjusted r-squared. In general, the adjusted r-squared is one of the most important indicators to measure the quality of model. However, this conclusion does not apply to this model. Because this regression model is not used to predict the page view, but used to test whether using keywords is dominant in the contribution of page views. As can be seen from Table 3, adjusted r-square value is 0.0317, indicating that the model can explain the variance of 3.17% page views. Therefore, the keywords of data science articles' titles do not have a major effect on the page views, which means that the "clickbait" phenomenon in the data science articles of WeChat official account is not prominent.

Table 3. Regression analysis results

Keyword		Estimate	Std.Error	T value	Pr(>|t|)	Signifi-cance
Chinese	English					
(Intercept)		1955.7	267.3	7.317	2.82e-13	***
数据	Data	-704.1	281.2	-2.504	0.012297	*
数据库	Database	1416.2	287.5	4.927	8.56e-07	***
时代	Era	-753.2	364.9	-2.064	0.039032	*
数据分析	Data Analysis	637.2	254.0	2.509	0.012144	*
数据挖掘	Data Mining	1092.4	399.3	2.736	0.006243	**
可视化	Visualization	1392.6	454.7	3.063	0.002203	**
人工智能	AI	1371.6	334.6	4.099	4.19e-05	***
学习	Learning	576.4	302.3	1.907	0.056619	.
应用	Application	-860.2	402.4	-2.139	0.032477	*
通知	Notice	-1167.5	786.6	-1.484	0.137802	
期权	Option	-2658.3	1532.2	-1.735	0.082791	.
生活	Life	-2544.3	863.8	-2.945	0.003236	**
检索	Retrieve	-1475.7	456.5	-3.232	0.001234	**
领域	Field	-1109.0	712.1	-1.557	0.119426	
机器人	Robot	-1539.2	852.1	-1.806	0.070895	.
一篇	A piece of	6544.7	1436.1	4.557	5.27e-06	***
Python	Python	6102.5	2617.7	2.331	0.019773	*
到	To	979.0	471.2	2.078	0.037775	*
推荐	Recommend	1442.9	750.0	1.924	0.054429	.
关键词	Keyword	-3835.1	2448.1	-1.567	0.117264	
种	A kind of	1160.1	727.5	1.595	0.110832	
带来	Bring	-1975.5	1217.7	-1.622	0.104776	
三个	Three	4509.8	1857.2	2.428	0.015199	*
报告	Report	1187.5	644.1	1.844	0.065293	.
日	Day	-818.8	495.9	-1.651	0.098771	.
自己	Myself	5158.8	1401.8	3.680	0.000235	***
人类	Human	4030.9	816.7	4.936	8.18e-07	***
产业	Industry	-1135.2	614.8	-1.847	0.064859	.
干货	Dry Goods	950.9	665.8	1.428	0.153302	
施行	Implement	7352.8	2084.9	3.527	0.000424	***

Residual standard error: 6403 on 6667 degrees of freedom

Multiple R-squared: 0.03604 Adjusted R-squared: 0.0317

F-statistic: 8.308 on 30 and 6667 DF， p-value: < 2.2e-16

5 Discussion

In this section, we will discuss the research results from three aspects, including the perspective of subject, the writer's perspective and the reader's perspective.

From the perspective of subject, the core of data science is to extract potential and valuable information from massive data by mixing different elements, theories and techniques in different fields, and even transforming into products. Therefore, according to the characteristics of this subject, data science articles emphasize the authenticity and practicality, professional and technical aspects are also more prominent.

From the writer's perspective, the authors who will write data science articles, basically have a certain depth of professional knowledge, and be able to master professional skills about data science. Therefore, such articles will not blindly pursue the click rate and access traffic, but will pay more attention to the article content. The research process of data science articles is generally inseparable from the data, so author's attitude towards the article will be more rigorous.

From the reader's perspective, there are three kinds of readers who will choose to read data science articles. One is the beginner who is interested in data science field, one is engaged in the relevant professional staff, and the other one is scholar whose research direction is data science. Reading these articles is generally to fill the knowledge gap, strengthen the technological advantages, and understand the development trend of data science. These readers will not easily waste their attention, but pay more attention to the value of article content.

6 Conclusion

This paper mainly discusses whether there is a "clickbait" in the data science articles that published in the WeChat official accounts. By constructing the regression model, it is found that the clickbait phenomenon in the data science articles of WeChat official accounts is not prominent. Although this paper has made some research results, but there are still many shortcomings. Firstly, there are omissions of setting keywords in the data collection process, resulting in an incomplete data set, because the related keywords of data science are far more than twelve. Secondly, this article ignores the influence of the keywords on the clickbaits in the abstract. If the WeChat official accounts only publish one article within one day, the reader will also see a portion of the abstract in addition to the title. In the future study, we will improve the completeness of the data set, and try to use other methods to simplify the data analysis process.

References

1. Potthast, M., Köpsel, S., Stein, B., Hagen, M.: Clickbait detection. In: The 38th European Conference on Information Retrieval, pp. 810–817. Springer International Publishing (2016)
2. Biyani, P., Tsioutsiouliklis, K., Blackmer, J.: 8 amazing secrets for getting more clicks: detecting clickbaits in news streams using article informality. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, vol. 16, pp. 94–100 (2016)
3. Chen, Y., Conroy, N.J., Rubin, V.L.: News in an online world: the need for an 'automatic crap detector'. In: Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4 (2015)
4. Fang, J., Lu, W.: A study on influential factors of Wechat public accounts information transmission hotness. J. Intell. **35**(2), 157–162 (2016)
5. Wang, Y.F., Xie, Q.N., Song, X.K.: Review and prospect of overseas research on data science. Libr. Inf. Serv. **60**(14), 5–14 (2016)
6. Loewenstein, G.: The psychology of curiosity: a review and reinterpretation. Psychol. Bull. **116**(1), 75–98 (1994)
7. Jia, F.H.: Information awareness and attention economy. J. Intell. **1**, 89–90 (2002)
8. Bashir, M.A., Arshad, S., Wilson, C.: Recommended for you: a first look at content recommendation networks. In: Proceedings of the 2016 ACM Conference on Internet Measurement, pp. 17–24 (2016)
9. Chakraborty, A., Paranjape, B., Kakarla, S., Ganguly, N.: Stop clickbait: detecting and preventing clickbaits in online news media. In: The 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 9–16 (2016)
10. Blom, J.N., Hansen, K.R.: Click bait: forward-reference as lure in online news headlines. J. Pragmat. **76**, 87–100 (2015)
11. Pengnate, S.F.: Measuring emotional arousal in clickbait: eye-tracking approach. In: Twenty-Second Americas Conference on Information Systems (2016)
12. Chen, Y., Conroy, N.J., Rubin, V.L.: Misleading online content: recognizing clickbait as false news. In: Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, pp. 15–19 (2015)