5-15-2019

# UNFOLDING THE CLICKBAIT: A SIREN'S CALL IN THE ATTENTION ECONOMY

Wenping Zhang
*Renmin University of China*, wpzhang@ruc.edu.cn

Qiqi Jiang
*Copenhagen Business School*, qj.digi@cbs.dk

Chih-Hung Peng
*City University of Hong Kong*, chpeng@cityu.edu.hk

Recommended Citation

Zhang, Wenping; Jiang, Qiqi; and Peng, Chih-Hung, (2019). "UNFOLDING THE CLICKBAIT: A SIREN'S CALL IN THE ATTENTION ECONOMY". In Proceedings of the 27th European Conference on Information Systems (ECIS), Stockholm & Uppsala, Sweden, June 8-14, 2019. ISBN 978-1-7336325-0-8 Research Papers.
https://aisel.aisnet.org/ecis2019_rp/66

# UNFOLDING THE CLICKBAIT: A SIREN'S CALL IN THE ATTENTION ECONOMY

*Research paper*

Zhang, Wenping, Renmin University of China, Beijing, China, wpzhang@ruc.edu.cn

Jiang, Qiqi, Copenhagen Business School, Frederiksberg, Denmark, qj.digi@cbs.dk

Peng, Chih-Hung, City University of Hong Kong, Hong Kong SAR, chpeng@cityu.edu.hk

## Abstract

*The media consumption moving online was supposed to enable people to access information more effectively, instead, to some extent, people were yet overwhelmed. Clickbait, defined as "(On the Internet) content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page" (in Oxford English Dictionary), prevails and restrains people retrieve their desirable information effectively. In this exploratory study, we proposed two intriguing research conjectures, i.e. how rhetorical features in the clickbait influence the visiting traffic to the publisher at different levels and attempted to understand the antecedents and consequences of the prevalence of clickbait. In collaboration with a leading digital media company, we longitudinally collected a massive dataset in 2017. To test the research conjectures, we designed and developed a series of text mining methods and applied econometric analysis for empirical validation. The findings revealed 1) the rhetorical characteristics (hyperbole, insinuation, and visual rhetoric) could entice individual to click the baited headlines, 2) there was a quadratic (inverted U-shaped) relationship between number of clickbait posted by publisher and its visit traffic, and 3) such non-linear relationship was moderated by publisher's age.*

*Keywords: Clickbait, Text Mining, Rhetoric*

# 1 Introduction

Clickbait, defined as "(On the Internet) content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page" (in Oxford English Dictionary), prevails in the Internet, especially in the online media, recently. The prevalence of clickbait results from the unique revenue model in the Internet. The publishers employ the clickbait to increase the page views or clicks for monetization. Due to its deceptive nature, clickbait is widely criticized by both public institutions and companies. For instance, 28 EU member countries found that the online media tended to publish articles with "catchy, provocative and sensationalist front-page" headlines in lieu of delivering quality content and declared to improve such tendency (Orosa 2017); Facebook announced to take initiatives to significantly reduce the clickbait by intensifying the punishment[1]. Nevertheless, the clickbait is still prevalent even though various regulations for reducing the clickbait have been implemented (Chakraborty et al. 2016; Rony et al. 2017). We think the ineffectiveness on controlling the spread of clickbait results from the insufficient understanding of the "demand side" of clickbait. In particular, few studies have investigated the antecedents to the prevalence of clickbait from individual side and the consequence of employing clickbait to the publishers. This exploratory study attempts to answer these two questions, which will be instrumental in making more effective regulations in the future.

The information-gap theory of curiosity explained why individuals have intentions to click the baited headlines (Loewenstein 1994). Elaborately, it stimulates the curiosity when individual perceives a gap between her/his knowledge and her/his attention. Such gap generates "feeling of deprivation labelled curiosity" and motivates her/him to acquire the gapped information to alleviate the "feeling of deprivation". Prior literatures argued such baited stimuli could be developed in terms of using the rhetorical strategies. For instances, Flower and Hayes (1980) found rhetoric could construct the mystical stimuli to attract reader's attentions; similar conclusion was reached by McQuarrie et al. (1992), in which the rhetorical strategies were found to effectively attract consumer's attentions. In the marketing studies, rhetoric has been widely accepted as an effective strategy for designing appealing advertising messages (Phillips and McQuarrie 2002). Thereby, leading from such findings, we infer the individuals are more prone to click the rhetorically featured clickbait.

Differing from individuals whose attraction to the clickbait were mainly driven by irrationally psychological motives, the publishers rationally created the clickbait for the monetization (Cook 2016; Rochlin 2017). In particular, from publisher's perspective, it monetizes the website in terms of increasing the site traffic (Rochlin 2017). For example, Farhi (2007) found that online publishers significantly relied on the web advertising revenues from site traffic to secure its survival. Thereby, opportunistically employing clickbait to attract large volume of visits is an effective and efficient approach for increasing publisher's revenues. Besides, with the increase of the number of visitors, the publisher can more easily acquire the potential subscribers. Schlosser et al. (2006) stated converting visitors to users were of great benefits and thus should be regarded as a task with highest priority in digital strategy management. Wang et al. (2005) found the subscription fees or subscriber payoffs significantly contributed to the income of online publishers. In this regard, clickbait, as a budget and effective approach for user acquisition, is widely popularized by online publishers (Lombardi 2017).

Although the clickbait prevailed in the Internet and online media, not many studies have investigated such phenomenon to date. Numbers of extant studies confounded the concept of clickbait and fake news though they were completely different in essence. In particular, clickbait targeted for attracting

[1] https://newsroom.fb.com/news/2017/05/news-feed-fyi-new-updates-to-reduce-clickbait-headlines/

users to click and even read whereas fake news was created to disseminate fictitious information for malicious purposes. Notably, differing from fake news which contained little factual basis, the content under the clickbait was dubiously valueless only. Our literature reviews suggested the current researches on clickbait could be classified into two streams, i.e. algorithm or application design for clickbait detection and discussing the consequence of clickbait. The former discussed the design and evaluation of different clickbait detection methods, which attempted to significantly reduce the clickbait by targeting and identification prior to its prevalence. However, with the evolvement of the clickbait, whether those proposed methods still effectively worked was in doubt. For the latter stream of research, these works mainly focused on the impact from the prevalence of clickbait. Such post-hoc evaluation did not provide tangible initiatives or suggestion for managing the deluge of clickbait in the online media. Collectively, although there was evident progress in the clickbait research, however, we thought the research of clickbait could be further broadened.

As stated previously, this study served as an exploratory work to articulate the “demand side” of the clickbait. Elaborately, we attempted to answer the questions at two different levels, namely, 1) why individuals clicked those baited headlines and 2) what happened to the publishers if they constantly created the clickbait. Answering these questions is not only vital to understand the antecedents to the prevalence of clickbait but conducive to enacting the regulations or policies to reduce the creation of clickbait in the future as well.

The reminder of this study was organized as follows. In the next Section, we introduced the background knowledge and concept of clickbait. A brief literature reviews in both two streams, i.e. clickbait detection method and consequence of the clickbait prevailing, were presented and discussed. Meanwhile, as an exploratory study, we proposed our research conjectures. Next, we presented our empirical validations, i.e. text mining and econometric analysis, to test the proposed conjectures. In particular, there were two parts in our empirics. These two empirics were conducted to test the conjecture at individual level and publisher level respectively. We discussed the findings and concluded this study in the last section.

# 2 Background, Concept, and Research Conjecture

Although clickbait has prevailed in the online media for decades, there were few studies discussing the clickbait in the information systems (IS) discipline. It was noted that the clickbait was different from the fake news (Chen et al. 2015). Thus, earlier studies which discussed fake news and its antecedents and consequences were not included in this section. After a comprehensive extent of literature reviews, we classified the research topics of clickbait into two main streams, i.e. the detection of clickbait and the consequence of clickbait prevailing. The former one was about design and evaluation of clickbait detection methods. For the latter, it investigated the various consequences, i.e. negative impact, resulting from the deluge of clickbait. More details were presented below.

## 2.1 Design and Evaluation of Clickbait Detection Method

Research on clickbait detection could be earliest found in 2016. The work by Potthast et al. (2016) was regarded as one of the earliest works, in which they employed a dataset from Twitter to construct the first clickbait corpus containing 2,992 tweets and extracted 767 pieces of clickbait. In particular, they designed a detection approach by building upon 215 semantic features and incorporating the application of a series of simple random forest classifiers, which achieved 0.76 precision[2] and 0.76 recall[3]

[2] Precision (positive predictive value) is the fraction of relevant instances among the retrieved instances.

[3] Recall (sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.

respectively. Inspired by this attempt, Cao and Le (2017) increased to 331 semantic features and applied a more sophisticated algorithm, i.e. random forest regression (RFR), to target potential clickbait in the social media. This approach achieved outperformance, i.e. the 0.82 accuracy[4] and the 0.61 F1-score[5] respectively. Massive studies employing similar design thinking were proposed with the constantly improved performance. We tabulated the most representative studies in Table 1 below.

| Studies | Algorithm(s)Studies | Validation Dataset | Performance |
|---|---|---|---|
| Potthast et al. (2016) | Simple random forest classifiers | 2,992 tweets | 0.76 precision<br>0.76 recall |
| Cao and Le (2017) | Random forest regression (RFR) | 21,000 headlines/titles | 0.82 accuracy<br>0.61 F1-score |
| Anand et al. (2017) | A neural network (NN) architecture based on recurrent neural network (RNN) | 15,000 news headlines | 0.98 accuracy<br>0.98 F1-score |
| Gairola et al. (2017) | Recurrent neural network (RNN) | a collection of 19,538 posts | 0.65 F1-score |
| Zhou (2017) | Self-attentive neural network (SANN) | 102,045 tweets | 0.86 accuracy<br>0.68 F1-score |
| Zheng et al. (2017) | gradient boosting decision tree (GBDT) | Two datasets including 32037 and 11193 articles respectively. | 0.75 precision,<br>0.81 recall. |
| López-Sánchez et al. (2017) | A case-based reasoning methodology | One dataset contains a total of 32,000 headlines from digital newspapers | 0.99 accuracy |
| Biyani et al. (2016) | gradient boosting decision tree (GBDT) | 4073 webpages | 0.749 F-score |
| Chakraborty et al. (2016) | Support Vector Machines (SVM), Decision Trees, and Random Forests | 200 news | 0.93 accuracy |
| Rony et al. (2017) | SoftMax classifier | 32, 000 news headlines | 0.98 accuracy |

*Table 1. Brief Review of Clickbait Detection Approach (Algorithmic Design)*

Building upon the subtle algorithmic design of clickbait detection, certain studies developed the application of clickbait detection. For instance, Chakraborty et al. (2016) designed a browser extension for warning the potential clickbait by comparing the feature distributions between trained clickbait and non-clickbait dataset; Rony et al. (2017) designed and developed an automatic bot and integrated it into a web browser to help users avoid clicking the baited headline in the social media. However, either algorithm-based design or application-based design overly discussed the computational capability for detecting the clickbait and avoiding the potential clicking but rarely took account into dynamics

[4] Accuracy is the faction of the correctly retrieved instances.

[5] F1-score is the overall performance. It refers the balance (trade-off) between precision and recall. The higher score of F1-score indicates a better overall performance.

from the user side. As depicted previously, curiosity served as a prominent factor nourishing the prevalence of clickbait, which drove individuals to acquire the missing information to alleviate such feeling (Loewenstein 1994). Thus, understanding user's intention to click those baited headlines is imperative.

However, to the best of our knowledge, only limited studies considered users' characteristics in designing the clickbait detection method. Zheng et al. (2017) argued that behavioural traits were helpful to design clickbait detection and included them as semantically textual features in their application design. To validate it, they proposed method considering behavioural traits, i.e. gradient boosting decision tree (GBDT), and achieved better performance than those conventionally computational methods. López-Sánchez et al. (2017) argued the perception of clickbait was contingent upon individuals and applied deep learning and metric learning to improve the adaptability of their clickbait detection model. In addition, prior literatures in communication and media research unveiled the rhetorical language is more attracted to individuals (Benoit and Smythe 2003; Scaraboto et al. 2012). Practically, the rhetoric could be widely found to frame the headlines for gain attentions from readers in both online and traditional media. In this regard, we offered our first conjecture regarding to clickbait in the individual demand side:

**Research Conjecture 1:** *The extent of rhetoric features embedded in a clickbait is positively associated with the number of baited individuals.*

## 2.2 Consequence of Clickbait Prevailing

Besides the research on clickbait detection approaches, another stream of studies discussed the consequence or impact of clickbait prevailing. Intuitively, the primary objective of creating clickbait was for increasing the clicking likelihood. The publishers leveraged the clickbait to attract individual attentions, which enticed them to click such intriguing headline (Anand et al. 2017; Potthast et al. 2017). For example, publisher employed the social sharing functionality to promote the clickbait in the social media (Rubin 2017). Given that the headlines were prominently displayed in the shared newsfeeds in the social media, amounts of users there would be enticed to click the clickbait and be redirected to the publisher's site. Considering that the extent of visit served as primary antecedent to publisher's income, the clickbait, which helped to increase visit traffic to the publisher account, should not be sorely resisted by the publishers.

However, as depicted previously, from user's perspective, experiencing the clickbait engendered negative perception to the publishers. Accumulatively, the users were very likely to abandon the publisher if it constantly created the clickbait. For instance, Beleslin et al. (2017) asserted that using clickbait for attracting users was a risky strategy because the negative attitudes towards the clickbait could be progressively cultivated over time; Scacco and Muddiman (2016) comparatively studied the factual headlines and clickbait. They reached similar conclusion that individual presented negative attitudes and reactions when they were perennially exposed to the clickbait. Previous literatures argued individual behaviors could be viewed as a manifestation of their attitudes (Ajzen 1985). To this end, the publisher, which constantly created the clickbait, might receive the accumulated negative attitudes from its users. This would eventually make the users abandon the focal publisher. Hence, the user churn happened. To this end, we could find that although using the clickbait helped publishers accumulate the visit traffic firstly, such traffic would drop when they continued increasing the creation of clickbait over time. Thus, we postulated an inverted U-shaped relationship between the number of clickbait (created by a publisher) and the visit traffic (to the focal publisher).

Besides, given that the clickbait largely spread in social media through the Word-of-Mouth effect, e.g. repost or click "like" button (Fulgoni and Lipsman 2017), the aged publishers, which had more subscribers, could exert higher influence on increasing visit traffic through creating clickbait. More specifically, more people from their larger user base could be influenced and enticed to click those baited headlines. But at the same time, by continuing increasing the creation of clickbait, those aged publishers would also face the higher chances in losing more existing subscribers. The existing users might be annoyed with the clickbait and choose to discontinue subscribing to this publisher. This causes the decrease in the visit traffic. On the basis of the foregoing inference, we had the second conjecture regarding to clickbait in the publisher demand side:

**Research Conjecture 2:** *The number of clickbait posted by an aged publisher has a stronger inverted U-shaped relationship with its visit traffic than for younger publishers.*

# 3 Methods

## 3.1 Study 1: Individual-level Investigation

### 3.1.1 Data of Study 1

To test our first conjecture, we collected data from a leading digital media company (hereinafter referred to as *New-Media* to preserve the company's identity) in China. New-Media operates more than 500 official publisher accounts in a leading social media in China with more than 10 million subscribers. Each publisher account had its own designated editor and published articles in its own niche like "Fitness", "Lifestyle", or "Photographing" etc. All articles from these official accounts could be shared within the social media for free reading. In the bottom of each article, there was a series of sponsored advertisement. This contributed the key revenues to New-Media. Our sample contained 7,481 articles published from July 1st, 2017 to December 31th, 2017. After consulting some editors of publisher accounts and the managers from New-Media, we set a one-week lag for collecting our article-level data, i.e. number of visits and number of "Likes". The primary reason was the timeliness of each article was estimated as 5-7 days.

### 3.1.2 Measurement and Analysis of Study 1

The first research conjecture postulated the positive relationship between extent of rhetorical features in the clickbait and the number of baited individuals. Thus, each article was used as a unit of analysis. The dependent variable was the number of visits of each article. As mentioned previously, we counted this number of each article after 7 days since its initial publication. Remarkably, to remove the clicks created by bots, we set a threshold $\tau$, refers to the interval between the timestamp of hitting the headline and the timestamp of landing the article page, to filter out the fake clicks. In particular, we removed all clicks whose $\tau$ values were fewer than 5 seconds. The dependent variable was denoted as $RDS_i$.

For the independent variable, i.e. extent of rhetorical features, we designed and developed a series of text mining methods to extract different rhetorical characteristics in the article headlines. By referring to prior studies in both clickbait and linguistic research (e.g. Anand et al. 2017; Biyani et al. 2016; Deighton 1985; Hart and Daughton 2015; Vatz 1973; Zhang et al. 2018), we have semantically and syntactically featured four prominent rhetorical characteristics, i.e. hyperbole, insinuation, puzzle, and visual rhetoric, and computationally measured such four rhetoric with our proposed framework.

***Hyperbole*** refers to use of exaggeration as a rhetoric. Overstatement or overexpression was widely found in this rhetorical characteristic. For instances, "*Make millions of dollars within 10 days!*" is a representative hyperbole. We manually constructed the hyperbole lexicon from clickbait headlines in our sample to detect the hyperbolic terms in the headlines. The hyperbolic strength of a headline could be measured by:

$$Hyperbole\_Strength = \frac{No.of\ hyperbole\ terms\ in\ headline}{No.of\ total\ terms\ in\ headline}$$

***Insinuation*** refers to a rhetoric strategy to appeal the readers. To do so, editors usually adopted tempting or metaphorical words or expressions. For example, editors may use the sentence like "*All successful people have these good living habits, check it now!*" to tempt readers to click. Similar as hyperbole, it can be measured by:

$$Insinuation\_Strength = \frac{No.of\ insinuation\ terms\ in\ headline}{No.of\ total\ terms\ in\ headline}$$

***Puzzle***: From linguistic perspective, puzzle could be created in two ways: keeping pronoun(s) in the headline and creating question-based headline (e.g. echo question or interrogative). For the former, it could be measured in similar ways as previous two measurements. An example of this rhetorical clickbait can be "*This kid opens a present. You won't believe what happens when they see what's inside!*". For the latter, a typical example is "*Can you solve this ancient riddle? Most people failed!*". To measure it, we firstly defined a set of figurative forms (e.g. interrogation, rhetorical questions, and elliptical sentences), denoted by $S$. Afterwards, we defined a set of figurative forms of each headline, denoted by $E$. In this regard, we used the following formula to depict the extent of puzzle in the question-based headlines:

$$Figurativeness\_Strength = \begin{cases} 1 & if\ S \cap E \neq \emptyset \\ 0 & else \end{cases}$$

Collectively, the overall measurement of the extent of puzzle in each headline could be written as:

$$Puzzle\_Strength \\ = \alpha \cdot Figurativeness\_strength + (1-\alpha) \cdot \frac{No.of\ pronomial\ terms\ in\ headline}{No.of\ total\ terms\ in\ headline}$$

Where $\alpha$ is a smoothing factor, e.g. we can set $\alpha = 0.5$ if both types of puzzles contribute equally in constructing the puzzle.

***Visual Rhetoric***: There are three types of visual rhetoric widely used in the clickbait, i.e. symbols, digits, and pictures.

We manually constructed a symbol set $Sym$ and another set $Sym'$ denoting the symbols in the headlines. Previous literatures indicated that the headlines with more symbols could attract more attentions (Cao and Zhang 2012; Cui et al. 2009), thereby the symbol strength can be measured as: $symbol\_strength = |Sym'|/|Sym|$. The extent of digits could be measured as similar as the measurement of hyperbole or insinuation. Given the fact that there was at most one picture, as a thumbnail under each headline, we employed a binary value, 1 denoted the inclusion of a picture, to depict whether there was a picture or not. Eventually, the visual rhetoric of each headline can be measured by:

$$visual\_rhetoric\_strength = \beta \cdot symbol\_strength + \gamma \cdot digits\_strength + \delta \cdot picture\_strenth$$

Where $\beta, \gamma$ and $\delta$ are smoothing factors and satisfy $\beta + \gamma + \delta = 1$ (e.g. $\beta = \gamma = \delta = 1/3$).

Collectively, we used *HYP<sub>i</sub>*, *INS<sub>i</sub>*, *PUZ<sub>i</sub>*, and *VIS<sub>i</sub>*, to represent these four types of rhetorical characteristics respectively. Indeed, there should be more than four rhetorical features in the linguistic presentation. However, these four features were the most representative ones which were widely used in the clickbait after consulting the practitioners from New-Media. The methods for extracting more rhetorical characteristics were recommended in future study. We have indicated it in the limitation in the subsequent section.

After implementing these foregoing measurements to extract the rhetorical features from each headline, we recruited three student assistants to manually check the validity of the results. In particular, we randomly picked up 1000 headlines and related articles, and sequentially presented the headline and the related content to the assistants. They were requested to vote whether the headline was a clickbait or not. The results confirmed the validity and accuracy of our measurement. In particular, the headlines which were not labelled as a clickbait were assigned null values in any rhetorical dimension. In contrast, the headlines which had positive values in any rhetorical dimension, regardless of the numerical values, were labelled as clickbait.

We also included a rich set of control variables. These control covariates included headline length ($HDL_i$), number of likes ($LKN_i$), number of shares ($SRN_i$), whether this article was a lead article ($ILP_i$) in the issue. The headline length ($HDL_i$) was counted by the number of words. Remarkably, the publishers selected certain articles as lead articles and set such articles in the top at their social media page occasionally. We used a binary value ($ILP_i$) to indicate whether an article was a lead article (if yes, $ILP_i =1$) or not. It was noted that the topic of each article was not included. Although each publisher had its own niche, however, there were lots of overlaps among their published articles. For example, the topic of the article, "Diet Recipe", from the publisher "Fitness Girl" might have overlap with the articles from another publisher named "Healthy Life". In this regard, controlling the demographics of publisher might result into the contradiction instead. The details of variable definitions and descriptive statistics were presented in Table 2 below.

| Variable | Notation | Mean | Std. Dev | Min | Max |
|---|---|---|---|---|---|
| Number of Clicks of Article *i* | $RDS_i$ | 19090.421 | 18670.623 | 146 | 420697 |
| Extent of hyperbole in the headline of Article *i* | $HYP_i$ | 0.302 | 0.256 | 0 | 0.796 |
| Extent of insinuation in the headline of Article *i* | $INS_i$ | 0.296 | 0.373 | 0 | 0.823 |
| Extent of puzzle in the headline of Article *i* | $PUZ_i$ | 0.272 | 0.236 | 0 | 0.639 |
| Extent of visual rhetoric in the headline of Article *i* | $VIS_i$ | 0.219 | 0.282 | 0 | 0.701 |
| Length of headline of Article *i* | $HDL_i$ | 20.932 | 6.638 | 4 | 64 |
| Number of "Likes" of Article *i* | $LKN_i$ | 200.429 | 285.541 | 0 | 4931 |
| Number of "Sharing" of Article *i* | $SRN_i$ | 406.214 | 813.082 | 0 | 23852 |
| Whether Article *i* is a lead article or not Front Page | $ILP_i$ | 0.263 | 0.440 | 0 | 1 |

*Table 2. Variables Used in Article Level Study (7,481 observations)*

Correlations among the studied variables were reported in Table 3; the majority of the bivariate correlations were below the recommended 0.70 threshold level. Only two pairs of variables had slightly higher value in the correlation matrix. To rule out the concern of collinearity, we calculated the variance inflation factors (VIF) of each variable. The maximum estimated value of VIF is 6.53, which is lower than the recommended threshold, i.e. 10.0 (Cohen et al. 2003).

| Variables | $HYP_i$ | $INS_i$ | $PUZ_i$ | $VIS_i$ | $HDL_i$ | $LKN_i$ | $SRN_i$ | $ILP_i$ |
|---|---|---|---|---|---|---|---|---|
| $HYP_i$ | 1 | | | | | | | |
| $INS_i$ | 0.150 | 1 | | | | | | |
| $PUZ_i$ | 0.260 | 0.337 | 1 | | | | | |
| $VIS_i$ | -0.210 | 0.113 | 0.206 | 1 | | | | |
| $HDL_i$ | -0.186 | -0.256 | -0.338 | 0.426 | 1 | | | |
| $LKN_i$ | 0.416 | 0.628 | 0.661 | 0.503 | -0.200 | 1 | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *SRN<sub></sub>* $_i$ | 0.476 | 0.646 | 0.578 | 0.701 | 0.377 | 0.722 | 1 | |
| *ILP* $_i$ | 0.667 | 0.544 | 0.482 | 0.501 | -0.338 | 0.487 | 0.406 | 1 |

Table 3. Correlation Table (Article Level Study)

Based on these variables, the model of our report level study could be described as follows:

$$f(\mathrm{RDS}_i) = \beta_0 + \beta_1\mathrm{HYP}_i + \beta_2\mathrm{INS}_i + \beta_3\mathrm{PUZ}_i + \beta_4\mathrm{VIS}_i + \beta_5\mathrm{HDL}_i + \beta_6\mathrm{LKN}_i + \beta_7\mathrm{SRN}_i + \beta_8\mathrm{ILP}_i + \varepsilon_i$$

Given that the dependent variable (*RDS$_i$*) was a typical counting variable, we employed both Poisson regression and Negative binomial regression to estimate the coefficients. The results were reported in Table 4 below.

| Variable | Coefficient | | | |
|---|---|---|---|---|
| | Poisson Regression | | Negative Binomial Regression | |
| | Model 1 | Model 2 | Model 3 | Model 4 |
| *HYP* $_i$ | 0.018**<br>(0.021) | 0.018***<br>(0.025) | 0.017*<br>(0.023) | 0.015**<br>(0.020) |
| *INS* $_i$ | 0.023*<br>(0.056) | 0.020***<br>(0.054) | 0.021**<br>(0.035) | 0.028***<br>(0.022) |
| *PUZ* $_i$ | 0.012*<br>(0.017) | 0.009<br>(0.013) | 0.007<br>(0.037) | 0.007<br>(0.011) |
| *VIS* $_i$ | 0.009*<br>(0.002) | 0.012*<br>(0.022) | 0.005*<br>(0.007) | 0.008**<br>(0.008) |
| *HDL* $_i$ | -- | -0.006*<br>(0.001) | -- | -0.004*<br>(0.001) |
| *LKN* $_i$ | -- | 0.088***<br>(0.081) | -- | 0.074***<br>(0.086) |
| *SRN* $_i$ | -- | 0.054***<br>(0.071) | -- | 0.044***<br>(0.087) |
| *ILP* $_i$ | -- | 0.112*<br>(0.351) | -- | 0.089**<br>(0.100) |
| cons | 12.936*<br>(0.420) | 13.674**<br>(0.489) | 13.080**<br>(0.523) | 13.013***<br>(0.689) |
| R2 | 0.436 | 0.631 | 0.502 | 0.7330 |
| *p-value <=0.1; **p-value <=0.05; ***p-value <=0.01 | | | | |

Table 4. Results for Report Level Model Estimation

The results indicated the negative binomial regression model fit the data better than the Poisson regression model. One of the plausible reasons is that Poisson regression makes a strong assumption that the variance is equal to the mean (Schilling and Phelps 2007). Our dependent variable was over-dispersed, which resulted into its variance exceeding its mean value. The dispersed data followed a gamma distribution, thereby the negative binomial distribution could achieve better estimated results (Hilbe 2011). The estimated results were interesting. It was found that the extent of rhetorical features, especially the hyperbole, insinuation, and visual rhetoric, enacted positive influences on the number of clicks of each article. Creating puzzles in the headlines had no significant contribution to increase the clicks of each article. To this end, our first research conjecture had been well tested. The exploratory results informed the clickbait characterized by hyperbolic, insinuating, and visual rhetoric could contribute to the increase in the number of the clicks of each article. Thus, setting rules or regulations to restrain using such rhetoric in the headlines could significantly contribute to reduce the prevalence of clickbait. But from another perspective, for those editors creating the headlines based on the factual basis, it encouraged them to use hyperbole, insinuation, and visual rhetoric, to increase their readership.

## 3.2 Study 2: Publisher-level Investigation

### 3.2.1 Data of Study 2

The second research conjecture was to investigate 1) a quadratic relationship between number of creating clickbait and the visit traffic to the publisher, and 2) a moderating role of the publisher's age. Differing from previous article-level study, we selected 202 publisher accounts managed by New-Media for the publisher-level investigation. As depicted in previous sections, the change in the visit traffic resulted from the peer influence and the existing user churn. To precisely observe such fluctuation, fundamental amounts of user base in imperative. In other words, the clickbait posted by those publisher accounts only having few subscribers might not enact significant influence on its visit traffic. After consulting with the practitioners from New-Media, we focused on the publisher accounts having at least 50,000 subscribers. This resulted into 202 publisher accounts in our sample. These accounts were created for targeting 25 distinctive niches. Given that these accounts did not post an article every day, we created our panel as a month-level observation. In this regard, 6-month observation might not sufficient. Thus, we collected the data for the whole year of 2017. Thus, in this strictly balanced panel, the publisher account was the unit of analysis. In the end, we had 2,424 observations.

### 3.2.2 Measurement and Analysis of Study 2

The dependent variable ($AUN_{jt}$) in this study was the aggregated number of monthly unique visitors to the article published by each publisher account in the last month. This measurement was different from the number of visits or number of unique visits widely employed in the web analytics (Vellingiri et al. 2015). In other words, we only counted the number of unduplicated visitors by each month to the different articles published by the same publisher account in the previous month. For example, if a publisher account only published two articles in Jan 2017 and both articles were only read by one person in Feb 2017, this dependent variable would be equal to 2 for this publisher account in Feb 2017.

The Key predictors in study 2 include (1) monthly number of clickbait published by each account ($CFQ_{jt-1}$), and (2) the age (counted by number of months) of each publisher account ($AGE_{jt-1}$). Leading from the findings in Study 1, we applied our measurement for extracting rhetorical characteristics to identify the baited headlines. Notably, in this exploratory investigation, the weight of the rhetoric in each clickbait was not considered.

We also included a set of control variables, such as average number of visits to each article of each publisher account ($ARS_{jt}$), number of articles published by each publisher account ($ATN_{jt}$), number of newly added subscribers of each publisher account ($NRU_{jt}$), the total number of subscribers of each publisher account ($PTU_{jt}$), and the categorical niche of each publisher account ($PDM_j$). The details of variable definitions and descriptive statistics were presented in Table 5 below.

| Variable | Notation | Mean | Std. Dev | Min | Max |
|---|---|---|---|---|---|
| Visit traffic to publisher *j* at time *t* | $AUN_{jt}$ | 376,391.333 | 254,152.620 | 142,098 | 1,008,022 |
| Number of clickbait published by publisher *j* at time *t-1* | $CFQ_{jt-1}$ | 70.580 | 69.216 | 9 | 420 |
| Age of publisher *j* at time *t-1* | $AGE_{jt-1}$ | 31.320 | 16.891 | 7 | 106 |
| Average number of visits to | $ARS_{jt}$ | 14,560.588 | 13,025.447 | 94 | 420,697 |

| | | | | | |
|---|---|---|---|---|---|
| each article published by publisher *j* at time *t* | | | | | |
| Total number of articles published by publisher *j* at time *t* | $ATN_{jt}$ | 124.030 | 159.331 | 57 | 686 |
| Number of newly added subscribers of publisher *j* at time *t* | $NRU_{jt}$ | 2,196.000 | 3,609.802 | 559 | 36,295 |
| Total number of subscribers of publisher *j* at time *t* | $PTU_{jt}$ | 692,033.600 | 334,535.921 | 235,239 | 1,273,382 |
| Categorical niche of publisher *j* | $PDM_j$ | Categorical variables with 25 categories. | | | |

*Table 5. Variables Used in Publisher Level Study (2,424 observations)*

We presented the correlations among these variables in Table 6 below, where it was found the majority of the bivariate correlations were below 0.70. Besides, none of the variation inflation factor was higher than 10.0.

| Variables | $CFQ_{jt-1}$ | $AGE_{jt-1}$ | $ARS_{jt}$ | $ATN_{jt}$ | $NRU_{jt}$ | $PTU_{jt}$ |
|---|---|---|---|---|---|---|
| $CFQ_{jt-1}$ | 1 | | | | | |
| $AGE_{jt-1}$ | 0.474 | 1 | | | | |
| $ARS_{jt}$ | 0.668 | 0.550 | 1 | | | |
| $ATN_{jt}$ | 0.614 | 0.339 | 0.470 | 1 | | |
| $NRU_{jt}$ | 0.548 | -0.441 | 0.662 | 0.685 | 1 | |
| $PTU_{jt}$ | 0.440 | 0.741 | 0.220 | 0.643 | -0.518 | 1 |

*Table 6. Correlation Table (Publisher Level Study)*

To test our second research conjecture, i.e. the quadratic relationship and the moderation effect, we built up the following econometric model for estimating the coefficients:

$$\begin{aligned} &f(y|X)\\ &= \gamma_0 + \gamma_1 CFQ_{j,t-1} + \gamma_2 CFQ_{j,t-1}^2 + \gamma_3 AGE_{j,t} + \gamma_4 AGE_{j,t} * CFQ_{j,t-1} + \gamma_5 AGE_{j,t} * CFQ_{j,t-1}^2\\ &+ \gamma_6 ARS_{j,t} + \gamma_7 ATN_{j,t} + \gamma_8 NRU_{j,t} + \gamma_9 PTU_{j,t} + \gamma_{10} PDM_{j,t} + \xi_{it-1} \end{aligned}$$

In similar vein as the article-level investigation, we employed both Poisson regression and negative binomial regression respectively. In addition, we also applied the Hausman test to determine whether fixed-effect model outperformed the random-effect model. The null hypothesis was rejected. Thus, we chose the fixed effect model to fit our data. The estimated results were presented in Table 7 below.

| Variable | Coefficient | | | |
|---|---|---|---|---|
| | Poisson Regression | | Negative Binomial Regression | |
| | Model 1 | Model 2 | Model 3 | Model 4 |
| $CFQ_{jt-1}$ | 0.085*<br>(0.056) | 0.062**<br>(0.041) | 0.030**<br>(0.048) | 0.026**<br>(0.018) |
| $CFQ^2_{jt-1}$ | -0.015*<br>(0.009) | -0.014**<br>(0.019) | -0.005**<br>(0.003) | -0.007***<br>(0.004) |
| $AGE_{jt-1}$ | 0.096*<br>(0.022) | 0.103**<br>(0.153) | 0.072**<br>(0.120) | 0.066**<br>(0.118) |
| $AGE_{jt-1}*CFQ_{jt-1}$ | -0.023**<br>(0.014) | -0.017*<br>(0.036) | -0.027**<br>(0.067) | -0.011**<br>(0.024) |

| | | | | |
|---|---|---|---|---|
| $AGE_{jt-1}*CFQ^2_{jt-1}$ | -0.002*<br>(0.003) | -0.003***<br>(0.007) | -0.006**<br>(0.015) | -0.009***<br>(0.011) |
| $ARS_{jt}$ | -- | 0.016*<br>(0.033) | -- | 0.020*<br>(0.089) |
| $ATN_{jt}$ | -- | 0.008**<br>(0.016) | -- | 0.007***<br>(0.012) |
| $NRU_{jt}$ | -- | 0.001*<br>(0.002) | -- | 0.003<br>(0.002) |
| $PTU_{jt}$ | -- | 0.375**<br>(0.590) | -- | 0.179**<br>(0.336) |
| $PDM_j$ | Included but values not presented | | | |
| cons | 8.726**<br>(12.770) | 8.968***<br>(9.632) | 9.209***<br>(11.589) | 8.483***<br>(10.226) |
| R2: within | 0.226 | 0.211 | 0.303 | 0.328 |
| R2: Between | 0.563 | 0.577 | 0.656 | 0.667 |
| R2: Overall | 0.488 | 0.500 | 0.559 | 0.604 |
| *p-value <=0.1; **p-value <=0.05; ***p-value <=0.01 | | | | |

*Table 7. Results for Publisher Level Model Estimation (Fixed effect model)*

The estimated coefficients in the Table 7 provided evidences to support the second research conjecture. There was an inverted-U shaped relationship between the number of clickbait created by a publisher and its visit traffic. In addition, although the aged publisher could promptly achieve the increase in visit traffic by spreading the clickbait, however, they were more prone to suffer the disastrous drop in visit traffic by continuing increasing to create such baited headlines.

# 4 Discussion and Conclusion

The media consumption moving online was supposed to enable people to access information more effectively, instead, to some extent, people were yet overwhelmed. The prevalence of non-factual content in online media made people more difficult retrieve their desirable information. Differing from the fake news or malicious rumour in the internet, their deluge has been effectively reduced in terms of various initiatives, like legislation or relatively mature technology-based solution. The clickbait, residing in the grey area between fake news and factual report, has not been thoroughly studied so far. In prior literatures, though not many, the research topics could be classified into two streams, i.e. design and evaluation of clickbait detection methods (e.g. Biyani et al. 2016; Rony et al. 2017) and the various consequences of clickbait prevailing (e.g. Beleslin et al. 2017; Scacco and Muddiman 2016). However, these studies did not well answer the question, “Why the clickbait prevails in the Internet”. Disclosing the antecedents driving the prevalence of clickbait ought to help us enact appropriate regulations and policies to effectively reduce its quantity and popularity. The findings from this work contributed to bridge such gap in the extant literatures.

As an exploratory study, we have proposed two intriguing research conjectures regarding to the “demand side” of clickbait. In particular, we asserted 1) the rhetoric characterized the clickbait and enticed individuals to click; and 2) there was a quadratic (inverted U-shaped) relationship between number of clickbait posted by publisher and its visit traffic, and such relationship was moderated by publisher’s age. The first conjecture was to unpack the prevalence of clickbait in the individual level, and the latter was proposed to outline consequential impact of such prevalence on the publishers. To test these two conjectures, we collected the longitudinal data in collaboration with a leading digital media company in China, and applied sophisticated analytical frameworks, i.e. a series of self-designed and developed text mining methods and econometric analysis, for empirical validation. We obtained three important findings.

First, besides verifying that the rhetorical characteristics indeed enticed individual to click the baited headlines, we obtained an in-depth understanding by revealing the significant role of hyperbole, insinuation, and visual rhetoric in enticing individual to click the baited headlines. Interestingly, the puzzle, which was widely used for attracting attentions in online advertisement (Alwitt 2002; Bizzozero et al. 2016; Fazio et al. 1992; Jiang et al. 2012), was not significantly associated to individual decision on clicking. From the regulator's perspective, it is instrumental in proactively examining those potentially baited headlines in terms of targeting these rhetorical features. This finding could also benefit those editors who created the headlines reflecting the factual basis. They could leverage these three identified rhetorical features to promote the articles by attracting more attentions and readership.

Second, our findings revealed an interesting phenomenon. Although using the clickbait could be an incentive to temporally increasing the visit traffic, continuing such strategy yet facilitated the user churn over time. Understanding such consequence by publishers could discourage them to increasingly create clickbait because of the impending negative outcome. This could effectively restrain the growth of clickbait in the online media. Indeed, this could also help those publishers falling into the dilemmatic bottleneck in user growth.

Last but not least, we found those aged publishers were prone to achieve the outperformance in visit traffic growth. However, it was also important to point out the user churn might also impend more easily and promptly. Remarkably, comparing with those newly established publishers, the aged publishers had higher influence on spreading the clickbait. Besides, according to the paradigm of organizational ecology (Hannan and Freeman 1993), the aged organizations intended to employ less aggressive but moderate strategy. Thus, by knowing the risk of such double-edged sword by creating clickbait, the influential and aged publishers were not willing to create clickbait impulsively, which also indirectly restricted the popularity of clickbait in the Internet.

As an exploratory study, we should be aware of the caveats of this work, which served as suggestions for future research. First, we only extracted four types of rhetorical features in the headlines though the practitioners have concurred with the generalizability of these four most representative ones. More sophisticated methods for extracting more rhetorical characteristics are suggested. Second, to further understand individual psychological rationales, we encourage future works to conduct psychometric analysis by different methods, e.g. surveys, lab and field experiments, and even NeuroIS. Third, the causality issues are not perfectly addressed in this study though we included the lagged variables in the model. The experimentation can be considered to rule out the endogeneity. Last but not least, our text mining framework was designed and implement for processing Chinese language. Future studies can implement our works into different languages, which can further improve the precession and effectiveness.

## References

Ajzen, I. 1985. "From Intentions to Actions: A Theory of Planned Behavior," In *Action control.* Springer Berlin Heidelberg. pp. 11-39.

Alwitt, L. F. 2002. "Suspense and Advertising Responses," *Journal of Consumer Psychology* (12:1), pp.35-49.

Anand, A., Chakraborty, T., and Park, N. 2017. "We Used Neural Networks to Detect Clickbaits: You won't Believe What Happened Next!" In *European Conference on Information Retrieval.* Springer, Cham. pp. 541-547.

Beleslin, I., Njegovan, B. R., & Vukadinović, M. S. 2017. "Clickbait Titles: Risky Formula for Attracting Readers and Advertisers," *XVII International Scientific Conference on Industrial Systems* (IS'17) Novi Sad, Serbia.

Benoit, W. L., and Smythe, M. J. 2003. "Rhetorical Theory as Message Reception: A Cognitive Response Approach to Rhetorical Theory and Criticism," *Communication Studies* (54:1), pp.96-114.

Biyani, P., Tsioutsiouliklis, K., and Blackmer, J. 2016. "8 Amazing Secrets for Getting More Clicks": Detecting Clickbaits in News Streams Using Article Informality. In *AAAI*, pp. 94-100.

Bizzozero, P., Flepp, R., and Franck, E. 2016. "The Importance of Suspense and Surprise in Entertainment Demand: Evidence from Wimbledon," *Journal of Economic Behavior & Organization* (130), pp.47-63.

Cao, X., and Le, T. 2017. "Machine Learning Based Detection of Clickbait Posts in Social Media," arXiv preprint arXiv:1710.01977.

Cao, Z., and Ye, J. 2009. "Attention Savings and Emoticons Usage in BBS," In Fourth International Conference on Computer Sciences and Convergence Information Technology, ICCIT'09. IEEE. pp. 416-419.

Chakraborty, A., Paranjape, B., Kakarla, S., and Ganguly, N. 2016. "Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media," In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on* IEEE. pp. 9-16.

Chen, Yimin, Niall J. Conroy, and Victoria L. Rubin. 2015. "Misleading Online Content: Recognizing Clickbait as False News." Proceedings of *the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM.

Cohen, P., Cohen, J., West, S. G., and Aiken, L. S. 2003. Applied Multiple regression/Correlation Analysis for the Behavioral Sciences, Routledge.

Cook, C. 2016. "Money Under Fire: The Ethics of Revenue Generation for Oppositional News Outlets," *Beyond clickbait and commerce*, 66.

Cui, A., Zhang, M., Liu, Y., Ma, S., and Zhang, K. 2012. "Discover Breaking Events with Popular Hashtags in Twitter," In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 1794-1798). ACM.

Deighton, J. 1985. *Rhetorical strategies in advertising*. ACR North American Advances.

Farhi, P. 2007. "Salvation? The Embattled Newspaper Business is Betting Heavily on Web Advertising Revenue to Secure Its Survival. But That Wager is Hardly a Sure Thing," *American Journalism Review* (29:6), pp.18-24.

Fazio, R. H., Herr, P. M., and Powell, M. C. 1992. "On the Development and Strength of Category–Brand Associations in Memory: The Case of Mystery Ads," Journal of Consumer Psychology (1:1), pp.1-13.

Flower, L., and Hayes, J. R. 1980. The Cognition of Discovery: Defining a Rhetorical Problem. *College composition and communication* 31(1), pp. 21-32.

Fulgoni, G. M., and Lipsman, A. 2017. "The Downside of Digital Word of Mouth And the Pursuit of Media Quality: How Social Sharing Is Disrupting Digital Advertising Models and Metrics," *Journal of Advertising Research* (57:2), pp.127-131.

Gairola, S., Lal, Y. K., Kumar, V., and Khattar, D. 2017. "A Neural Clickbait Detection Engine," arXiv preprint arXiv:1710.01507.

Hannan, M. T., and Freeman, J. 1993. *Organizational Ecology*. Harvard university press.

Hart, R. P., and Daughton, S. M. 2015. *Modern Rhetorical Criticism*. Routledge.

Hilbe, J. M. 2011. *Negative Binomial Regression*. Cambridge University Press.

Jiang, Q., Tan, C.H., and Wei, K.K. 2012. "Cross-Website Navigation Behavior And Purchase Commitment: A Pluralistic Field Research". In *Pacific Asia Conference on Information Systems (*PACIS) Proceedings. Paper 193.

Loewenstein, G. 1994. "The Psychology of Curiosity: A Review and Reinterpretation," *Psychological bulletin* (116:1), pp.75.

Lombardi, C. 2017. "Competition and the Public Interest in the Digital Market for Information" Discussion Paper, Europa-Kolleg Hamburg, Institute for European Integration. No. 1/17.

López-Sánchez, D., Herrero, J. R., Arrieta, A. G., and Corchado, J. M. 2017. "Hybridizing Metric Learning and Case-based Reasoning for Adaptable Clickbait Detection," *Applied Intelligence*, pp.1-16.

McQuarrie, Edward F. and David G. Mick. 1996, "Figures of Rhetoric in Advertising Language," *Journal of Consumer Research* (22), pp.424-438.

Orosa, B. G. 2017. "Use of Clickbait in the Online News Media of the 28 EU Member Countries," http://www.revistalatinacs.org/072paper/1218/68en.html

Phillips, B. J., and McQuarrie, E. F. 2002. "The Development, Change, and Transformation of Rhetorical Style in Magazine Advertisements 1954–1999," *Journal of Advertising* (31:4), pp.1-13.

Rochlin, N. 2017. "Fake News: Belief in Post-truth," *Library Hi Tech* (35:3), pp.386-392.

Potthast, M., Köpsel, S., Stein, B., and Hagen, M. 2016. "Clickbait Detection," In *European Conference on Information Retrieval,* Springer, Cham. pp. 810-817.

Rony, M. M. U., Hassan, N., and Yousuf, M. 2017. “BaitBuster: A Clickbait Identification Framework,” *AAAI*.

Rubin, Victoria L. 2017. "Deception Detection and Rumor Debunking for Social Media." *The SAGE Handbook of Social Media Research Methods*, 342.

Scacco, J. M., and Muddiman, A. 2016. “Investigating the Influence of “Clickbait” News Headlines,” https://engagingnewsproject.org/wp-content/uploads/2016/08/ENP-Investigating-the-Influence-of-Clickbait-News-Headlines.pdf

Scaraboto, D., Rossi, C. A. V., and Costa, D. 2012. “How Consumers Persuade Each Other: Rhetorical Strategies of Interpersonal Influence in Online Communities,” *BAR-Brazilian Administration Review* (9:3), pp.246-267.

Schilling, M. A., and Phelps, C. C. 2007. “Interfirm Collaboration Networks: The Impact of Large-scale Network Structure on Firm Innovation,” *Management Science* (53:7), pp. 1113-1126.

Schlosser, A. E., White, T. B., and Lloyd, S. M. 2006. “Converting Web Site Visitors into Buyers: How Web Site Investment Increases Consumer Trusting Beliefs and Online Purchase Intentions,” *Journal of Marketing* (70:2), pp.133-148.

Vatz, R. E. 1973. “The Myth of the Rhetorical Situation,” *Philosophy & rhetoric*, pp.154-161.

Vellingiri, J., Kaliraj, S., Satheeshkumar, S., and Parthiban, T. 2015. “A Novel Approach for User Navigation Pattern Discovery and Analysis for Web Usage Mining,” *Journal of Computer Science* (11:2), pp.372-382.

Wang, C. L., Zhang, Y., Ye, L. R., and Nguyen, D. D. 2005. “Subscription to Fee-based Online Services: What Makes Consumer Pay for Online Content?” *Journal of Electronic Commerce Research* (6:4), pp.304.

Zhang, W., Kang, L., Jiang, Q., and Pei, L. 2018. ”From buzz to bucks: The impact of social media opinions on the locus of innovation,” *Electronic Commerce Research and Applications* (30), pp. 125-137.

Zheng, H. T., Yao, X., Jiang, Y., Xia, S. T., and Xiao, X. 2017. “Boost Clickbait Detection based on User Behavior Analysis,” In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data,* Springer, Cham. pp. 73-80.

Zhou, Y. 2017. “Clickbait Detection in Tweets Using Self-attentive Network,” arXiv preprint arXiv:1710.05364.