IIMB-WP N0. 592/2019

WORKING PAPER NO: 592

The Antecedents and Rebroadcast Consequences of Clickbait

Prithwiraj Mukherjee
Assistant Professor
Marketing
Indian Institute of Management Bangalore
Bannerghatta Road, Bangalore – 5600 76
pmukherjee@iimb.ac.in

Souvik Dutta
Assistant Professor
Economics and Social Sciences
Indian Institute of Management Bangalore
Bannerghatta Road, Bangalore – 5600 76
souvik@iimb.ac.in

Dalhia Mani
Assistant Professor
Entrepreneurship
Indian Institute of Management Bangalore
Bannerghatta Road, Bangalore – 5600 76
dalhia.mani@iimb.ac.in

Year of Publication – July 2019

The Antecedents and Rebroadcast Consequences of Clickbait

Prithwiraj Mukherjee[*]
Souvik Dutta[†]
Dalhia Mani[‡]

July 16, 2019

Abstract

Clickbait is a method of framing articles' titles to induce readers to click on them, and is a common feature of online media today. We use a publicly available data set consisting of articles from 25 media organizations, each of which is rated as clickbait or not by human respondents, and augment it with Twitter retweet count, sentiment analysis and topic modeling. We demonstrate that human interest articles are positively associated with clickbait. We also show that the fraction of people with journalistic backgrounds in an organization's top management team is positively related to its clickbait usage. Finally we show that clickbait is rebroadcast less than non-clickbait on social media. Our results serve as a cautionary message to media organizations and digital marketers, who may be inadvertently harming the reach of their content by using clickbait.

Keywords: Clickbait, Sharing, Upper Echelons Theory, Sentiment Analysis, Latent Dirichlet Allocation, Propensity Score Matching

[*] Assistant Professor of Marketing, Indian Institute of Management Bangalore, pmukherjee@iimb.ac.in

[†] Assistant Professor of Economics and Social Sciences, Indian Institute of Management Bangalore, souvik@iimb.ac.in

[‡] Assistant Professor of Entrepreneurship, Indian Institute of Management Bangalore, dalhia.mani@iimb.ac.in

We are thankful to Joel Bothello and Sai Yayavaram for their helpful comments and suggestions. We also thank Kiran Jonnalagadda for helping us implement the Python code to capture retweet counts from Twitter

Statement of Intended Contribution

We present a study of clickbait that goes beyond existing machine learning approaches that identify it. We identify its antecedents from a content and organizational characteristics' perspective. Further, we show that clickbait articles are less rebroadcast on social media than non-clickbait. We quantify this deleterious effect of clickbait on online content's reach via sharing. Thus, we present one of the earliest rigorous investigations of both antecedents and consequences of clickbait, and add to the sharing and word of mouth literature.

We also contribute to extant research in upper echelons theory, by envisioning clickbait as an organizational outcome in media companies, which is related to top management team composition. We demonstrate that media organizations with a higher fraction of people from journalistic or editorial backgrounds in their top management teams are less likely to employ clickbait.

Our work is important not just to academic researchers, but very managerially relevant to digital media organizations, cautioning them against harming their content's reach by using clickbait. The antecedents we discover can help them avoid common elements that constitute it. Beyond these, we present multiple methods of text mining, propensity score matching etc., which we deem useful to analytics educators and practitioners. Finally, anyone with a general interest in digital news and social media should find our work interesting.

"Optimizing headlines for attention can be a good thing. But when that starts to become the only thing people think about, it's time for some new content metrics." — Berger (2014a)

Introduction

The news business is witnessing unprecedented transformation with the advent of digital media; between 2006 and 2011, US print media lost about 20% of their paid subscribers (Pattabhiramaiah et al., 2017b) and saw a 50% decline in ad revenue between 2007 and 2012 (Lambrecht and Misra, 2016). Reasons could be competition from free online portals as well as cannibalization of sales from the newspapers' own online portals. Digital portals are thus important revenue sources not just for new-age web-only companies like Buzzfeed and Mashable, but also for outlets over a century old, like The New York Times and The Guardian. Such digital news outlets additionally must now rely on online readership, a large fraction of which comes through social media sites like Facebook and Twitter or news aggregators like Google and Reddit. Internet surfers today click on very few of the hundreds of news links they are potentially exposed to. Many of these media companies rely on per-impression advertising revenues, where advertisers pay them every time a unique reader lands on a given article, viewing advertisements accompanying the article. To compete for readers' attention in such an environment, media outlets often use a tactic called "clickbait" - designing headlines to arouse readers' curiosity, thus inducing them to click on their links.

Despite the general negative perceptions of clickbait (see the "clickbait and its stakeholders" section), it is surprisingly persistent in online media. Our data set (see Table 1 and the "data and its augmentation" section for more details) reveals that even older, so-called "staid" journalistic outfits like The New York Times and The Guardian indulge in significant amounts of clickbait. Another revelation is that the publicly funded BBC, now the subject of a UK Office of Communications investigation for clickbait, has a surprisingly large fraction of its online headlines framed as clickbait, comparable to traditional tabloids like The Independent.

Thus, we answer the following question here - *what kind of organization indulges in clickbait and in what contexts?* Based on upper echelons theory (Hambrick and Mason, 1984; Hambrick, 2007) we hypothesize that organizations with a higher fraction of people with journalistic or editorial backgrounds in the top management team are less likely to employ clickbait than organizations with a higher fraction of people with non-editorial backgrounds (marketing, human resources, finance, venture capital, etc). Additionally, we hypothesize that organizations with paid content and offline presence are less likely to indulge in clickbait than those having free-only and web-only content. Invoking research on curiosity (Loewenstein, 1994), we also hypothesize that clickbait consists of more positive valence, and that lifestyle-related content is more likely to be clickbait. Using a probit model, we find evidence for some of these hypotheses, as outlined in the "results" section.

Specifically, we demonstrate that organizations with larger fractions of people from journalistic and editorial backgrounds in their top management teams are less likely to employ clickbait than organizations with smaller fractions of people from journalistic and editorial backgrounds. Furthermore, we present evidence for web-only organizations employing more clickbait than organizations with offline presences also. We find that clickbait is positively associated with moderately but not extremely positive title valence and that clickbait articles tend to have title valence dissimilar to the corresponding main article valence. We also find that human interest articles dealing with topics like health, social media, lifestyle, people and travel but not entertainment are positively associated with clickbait.

We answer another question in this paper - *is clickbait shared more or less than non-clickbait?* Based on popular perceptions surrounding clickbait, we hypothesize that by virtue of inducing irritation and annoyance, clickbait is disliked by readers and thus shared less, given that it conflicts with most motivations individuals may have for sharing (Berger, 2014b). Using propensity score matching in conjunction with linear regression, we find overwhelming evidence that this is indeed true. Even after controlling for selection bias, media organizations' Twitter follower counts, organizational and content variables, we find that clickbait articles receive 68.59 fewer retweets than non-clickbait articles. This is especially stark, as articles in our data set are retweeted 144.72 on average. Thus, our findings serve as a warning to media organizations employing clickbait, as they may be inadvertently sacrificing their content's reach. With our findings, we thus add to the growing literature on the determinants of sharing on social media (Berger and Schwartz, 2011; Berger and Milkman, 2012; Zhang et al., 2017; Tellis et al., 2019).

We use the Webis Clickbait Challenge 2017 data set, a corpus of over 19,000 articles sampled from several media sources' Twitter handles by Potthast et al. (2018b), with each headline rated as "clickbait" or not by human workers on Amazon Mechanical Turk. We add to this data set by scraping from Twitter, the number of times each such link was rebroadcast, i.e. retweeted. Additionally, we augment this dataset by performing sentiment analysis (Berger and Milkman, 2012) and topic modeling via Latent Dirichlet Analysis (Tirunillai and Tellis, 2014; Zhang et al., 2017) on each article in the corpus. We use probit regressions to test our first set of hypotheses about the antecedents of clickbait, and propensity score matching to compare the number of retweets between clickbait and non-clickbait articles.

Our primary intended audience for this paper are digital marketers, media managers and researchers in digital marketing, fitting well with the roadmaps laid out by Berger (2014b) on word of mouth research and Lamberton and Stephen (2016) and Kannan and Li (2017) on digital marketing research. Additionally, the multitude of methods makes our work of interest to marketing analytics educators. Finally, we expect anyone with an interest in social media and digital news to be interested in this work, given the popular interest in clickbait.

Clickbait and its Stakeholders

Coined by a blogger - Geiger (2006), clickbait has led to interesting ethical debates on journalistic practices as well as a lot of annoyance from readers. Considered misleading (and often conflated with fake news), almost every article on this topic, whether machine learning approaches to identify it (e.g. Rony et al., 2017; Potthast et al., 2018a) or commentaries in popular media (e.g. Berger, 2014a; Frampton, 2015; DeMers, 2017), describe clickbait as a nuisance. Comedian John Stewart opines (Smith, 2014), "It's like carnival barkers, and they all sit out there and go, *'Come on in here and see a three-legged man!'*. So you walk in and it's a guy with a crutch." Ben Smith, editor of Buzzfeed News uses this John Stewart quote to distance his organization from clickbait. He says (Smith, 2014), "But it suggests that Stewart, like many people in the media industry, confuses what we do with true clickbait. We have admittedly (and at times deliberately) not done a great job of explaining why we have always avoided clickbait at BuzzFeed[1]. In fact — and here is a trade secret I'd decided a few years ago we'd be better off not revealing — clickbait stopped working around 2009."

Frampton (2015) characterizes clickbait thus, "Put simply, [clickbait] is a headline which tempts the reader to click on the link to the story. But the name is used pejoratively to describe headlines which are sensationalized, turn out to be adverts or are simply misleading," echoing a general sentiment of many journalists. Ingram (2014) opines that the internet era has amplified, rather than invented clickbait. Indeed, salacious headlines like "Plastic surgeon builds himself a new wife" and "Handyman, 66, makes 5 neighbors pregnant" (Schaffer, 1995) have been used by tabloids and gossip magazines for over a century now. However, we recognize a fundamental difference between curiosity-inducing print headlines and online clickbait; the purpose of the former is to sell an entire newspaper or magazine, and thus such headlines are likely to be on the front page, visible to prospective buyers in news stands and magazine stores. Clickbait, on the other hand, drives traffic to one article, and one article alone. Thus, a few catchy front page headlines may sell an entire paper, while clickbait works on one article alone.

While Smith (2014) of Buzzfeed News couches his condemnation of clickbait in economic terms - "... clickbait stopped working around 2009," other stalwarts of journalism rue the clicks-driven commercial focus of new-age journalism, without directly referring to clickbait (which is possibly a symptom of this development). The primary targets of these critiques are the new online gatekeepers of news - Google and Facebook. A Pew Research Center report (Shearer and Matsa, 2018) finds that 68% of US adults get their news from social media, even

[1] Smith's sentiments may not resonate with readers - our data set shows that 62.79% of Buzzfeed's headlines were classified as "clickbait" by human survey respondents. See Table 1 for more details

though 57% do not trust the accuracy of all the content. 21% of readers surveyed in the study mentioned that they prefer ease of access over content. More strikingly, the report estimates that about 43% of America gets its news today from Facebook, followed by 21% from Youtube (a Google subsidiary) and 12% from Twitter. These numbers give these gatekeepers, especially Facebook, a lot of power; changes in their content propagation algorithms can significantly affect the fortunes of media companies relying on them for visibility. Former Washington Post editor Robert G Kaiser, who outlines how digital business models may be killing journalism laments (Kaiser, 2014), "Now, however, in the first years of the 21st century, accelerating technological transformation has undermined the business models that kept American news media afloat, raising the possibility that the great institutions on which we have depended for news of the world around us may not survive." Another stalwart of journalism, Jill Abramson, former Chief Editor of The New York Times names Google and Facebook as the "chief villains" (Graves, 2019) behind the 2018-19 layoffs at top media houses including Buzzfeed, Conde Nast, Huffington Post and Gannett. American politicians too have taken cognizance of this phenomenon, with Rhode Island Democratic Representative David Cicilline saying, "A free, diverse press cannot survive unless we confront the power of Facebook and Google." (Neidig, 2019). Financial survival and profitability thus seem to be at a conflict with traditional journalistic values at a lot of media organizations, which are increasingly under pressure to drive traffic to their portals via search engines and social media.

Facebook itself condemns clickbait thus[2], "Clickbait headlines intentionally omit crucial information or exaggerate the details of a story to make it seem like a bigger deal than it really is. This gets attention and lures visitors into clicking on a link, but they then quickly return to News Feed. We've heard from people that they prefer to see clearly written headlines that help them decide how they want to spend their time." Facebook expends considerable resources in tweaking its algorithms to prevent clickbait from reaching its users. In a detailed blog post, Facebook engineers Babu et al. (2017) explain steps taken by the social media giant to tackle clickbait. Another gatekeeper, Google, does not comment publicly on clickbait, though some commentators (e.g. DeMers, 2017) speculate that its algorithms are also tweaked to deal with clickbait, especially in promoted content.

Finally, clickbait is perceived as a nuisance by readers, sometimes even leading to vigilante action. A popular Twitter handle called `@savedyouaclick`, that summarizes alleged clickbait articles in a single line (see Figure 1), has over 270,000 followers, as well as a dedicated subreddit `/r/savedyouaclick` with over 770,000 subscribers. Several Twitter users too do the same, with the hashtag `#savedyouaclick` being very popular on the portal. A cursory glance at blogs and social media posts also indicates that readers generally perceive clickbait

[2] https://www.facebook.com/facebookmedia/blog/drive-reach-and-referrals-without-clickbait

negatively.

Insert Figure 1 about here

Hypothesis Development

In this section, we present theoretical arguments leading to our hypotheses in two broad domains: (a) the antecedents of clickbait, and (b) does clickbait get rebroadcast as much as non-clickbait on social media?

Antecedents of Clickbait

We divide this section into two broad domains: (a) the contents of clickbait and (b) the types organizations that may indulge in more (or less) clickbait. We provide brief literature reviews and theoretical arguments for our hypotheses here.

Content in clickbait

Digital media outlets employ clickbait to induce people to click on their links, by arousing their curiosity. Thus, we focus here on the kind of content that could arouse curiosity, noting that any stimulus that induces curiosity, if present in a title, should also lead to it being perceived as clickbait. Loewenstein (1994) notes that curiosity is a departure from the rational choice paradigm in economics, especially when it usually offers no other benefit than satisfying curiosity itself. Psychologists today concur on five dimensions of curiosity (Kashdan et al., 2018a,b): (a) *deprivation sensitivity*, an unpleasant state where an individual seeks relief by filling a gap in their knowledge (Loewenstein, 1994) (b) *joyous exploration*, a pleasurable state where an individual is filled with wonder about their world (Deci, 1992) (c) *social curiosity*, where individuals are curious about other people, leading to behaviors like gossip, voyeurism and eavesdropping (Renner, 2006) (d) *stress tolerance*, where novelty may induce anxiety but an individual is reluctant to act to satisfy this curiosity (Silvia, 2008) and (e) *thrill seeking*, where an individual is willing to take risks to acquire varied, complex and intense experiences (Zuckerman, 1979, 2014).

Of these five dimensions, deprivation sensitivity and social curiosity seem to be the most germane to the domain of clickbait[3]. Social curiosity can lead to gossip, and a need to know about others is a fundamental need for humans to function effectively in societies (Foster, 2004). We contend that stories with a human angle, in entertainment, society and lifestyle-related domains can trigger social curiosity and thus hypothesize,

[3] In certain online behaviors like illegal movie downloads, pornography consumption, etc., individuals are known to risk legal action, exposing their computers to malware, etc. Many providers of such content also use clickbait, but this is beyond the scope of our study

H 1. *The perception of clickbait is positively associated with articles dealing in human interest topics like entertainment, society and lifestyle*

In their information gap theory, Golman and Loewenstein (2018) propose that information's valence and clarity drive people's actions, with the latter being akin to uncertainty avoidance. When valence and clarity oppose each other, the latter tends to be a stronger driving force. The drive to seek clarity (or avoid uncertainty) to bridge information gaps thus could be contextualized in clickbait as well; a headline could be deliberately framed in such a way by an editor that it is dissonant with the body of text it is associated with. Therefore, we hypothesize,

H 2. *The perception of clickbait is positively associated with dissimilarity between an article's title and the main article's valence*

Finally, Marvin and Shohamy (2016) find that information with positive valence enhances curiosity more than information with negative sentiment valence[4]. Wiggin et al. (2018) establish that curiosity can lead to indulgent reward seeking, which in the context of clickbait, can be associated with joyous exploration. Therefore we hypothesize,

H 3. *The perception of clickbait is positively associated with an article's title's valence*

Organizations and clickbait

We now focus on the types of organization that are more likely to employ clickbait. Organizational responses to dwindling offline subscriptions and new-age online media competitors have been varied, ranging from The Independent going online-only, to The Guardian making all its online content free, to organizations like The Washington Post and The New York Times instituting paywalls. However, all of them operate online portals today. The Guardian operates its paid offline newspaper as before, but makes all its online content free, soliciting donations from readers in order to maintain editorial independence. In a study of The New York Times paywall, Pattabhiramaiah et al. (2017a) find that the paywall has a positive spillover on the newspaper's offline readership. Many newspapers have taken the seemingly counter-intuitive step of increasing their prices. Pattabhiramaiah et al. (2017b) explain this by showing that this may be because of the lowered ability of the newspaper to command advertising revenues. Clearly, subscription fees substitute for advertising revenues; Lambrecht and Misra (2016) model such a tradeoff for online goods, showing that firms should offer more free than paid content in periods of high demand. In our context, it is clear that the purpose of clickbait is to drive traffic to an online portal, for the purpose of ad revenues. We expect that media organizations without paid subscribers are more likely to do clickbait than those with paid subscribers, as a result of the non-contractual nature of their business. Thus we hypothesize,

[4] We recognize however, that extremely negative emotions can induce morbid curiosity (Oosterwijk, 2017)

H 4. *Organizations without paid subscribers are more likely to use clickbait than organizations with some paid subscribers*

Furthermore, organizations with an offline component (newspaper, magazine, TV, radio) are more likely to have been around longer, targeting both online and offline readers. We expect a residual inertial effect of their older offline models on their online practices. Thus we hypothesize,

H 5. *Web-only organizations are more likely to do clickbait than organizations with an offline outlet*

Finally we investigate how composition of a media organization's top management team is related to its propensity to employ clickbait. At the outset, we recognize that clickbait is not only undesirable to readers, but also to journalists and editors, irrespective of their organization's decision to use it. However, given the commercial pressures associated with survival in the digital age, we expect the upper managements in media organizations to make conscious decisions regarding the content in their online portals. We expect an inherent tension between people with editorial or journalistic backgrounds, and the others, like venture capitalists, marketing managers, etc. We thus envision clickbait to be an organizational decision, driven to a great deal by the top management team of a media organization, influenced by the backgrounds of its individual members.

Early evidence that organizational outcomes and orientations are related to top management team composition comes from Dearborn and Simon (1958), who find that executives look at business problems from their own departmental points of view. This intuition is then formalized in upper echelons theory (Hambrick and Mason, 1984; Hambrick, 2007) which posits that an organization's top management team composition shapes organizational decisions. For example, Chaganti and Sambharya (1987) find that tobacco companies' strategic orientations are related to top executives' backgrounds. Bantel and Jackson (1989) find that more innovative banks are seen to be managed by teams that are more educated and have diversity in functional backgrounds. Wiersema and Bantel (1992) find that organizations that are more likely to undergo change in corporate strategy have younger top managements with shorter tenures in the organizations, higher education levels, and more diversity in functional specializations. Jaworski and Kohli (1993) uncover the relationship between top managers' emphasis on market orientation and organization's actual market orientation. Barker III and Mueller (2002) find that research and development spending is positively associated with CEOs' background in marketing, engineering or research and development itself. Menz (2012) provides a comprehensive review of this stream of research. In line with upper echelons theory, we expect that media organizations with a higher fraction of people from editorial backgrounds are more likely to stick to old school journalistic principles (see the “clickbait and its stakeholders” section) and therefore hypothesize,

H 6. *Organizations with a higher fraction of people with editorial backgrounds in their top management team are less likely to employ clickbait than organizations with a lower fraction of people with editorial backgrounds in their top management team*

Rebroadcast Consequences of Clickbait

We now ask the following question: *Is clickbait shared more than non-clickbait?* We present here a brief review of extant research on the antecedents of sharing on social media. Berger (2011) and Berger and Milkman (2012), find that physiological arousal can induce content to be shared; high arousal content inducing awe, anger and anxiety tend to be shared more than content with low arousal like sadness. Akpinar and Berger (2017) study online ads to find that content with emotional appeal is shared more than content with informative appeal. Schulze et al. (2014) establish that viral campaign ads for hedonic but not utilitarian goods are successful in being rebroadcast on social media. Vousughi et al. (2018) find that false news diffuses faster and wider because it is usually more novel than true news, and that false news about politics is more viral than news about natural disasters, financial news or urban legends. Recently, Tellis et al. (2019) present several factors driving rebroadcasting of videos on social media. They find that positive emotions enhance sharing while prominent brand placement works adversely. They also find that emotional ads are shared more on general social media platforms like Facebook, Google+ and Twitter, while informational ads are shared more on a professional network like LinkedIn.

To investigate clickbait itself, we look at the different motivations people may have for sharing. Berger (2014b) provides a comprehensive review of this domain, noting that motivations could range from (a) *impression management* where individuals convey positive impressions of themselves, (b) *emotion regulation* where individuals manage their emotions, (c) *information acquisition* where individuals seek inputs from others, (d) *social bonding* where individuals seek to connect with others, and (e) *persuading others*. In our domain, clickbait seems undesirable to readers as the “clickbait and its stakeholders” section demonstrates, and seems incompatible with the sharing motivations outlined above. It induces irritation and often annoyance, and given this, it is unlikely to enhance sharers’ perceptions with their peers. Except in the limited case of emotion regulation, where venting and vengeance are possible, it is unlikely that an individual would want to be an intermediary source that rebroadcasts clickbait. As Berger (2014a) summarizes, “... clickbait is bad because it over-promises and under-delivers.” Thus, notwithstanding an individual’s own propensity to click on clickbait, we hypothesize,

H 7. *Clickbait is rebroadcast less than non-clickbait on social media*

Data Set and its Augmentation

Our principal data source is a corpus of online articles (with each article being rated by humans for its "clickbaitiness") from the Webis Clickbait Challenge 2017, a contest for machine learning enthusiasts to identify clickbait (Potthast et al., 2018b). While Potthast and colleagues present an extensive description of their data collection procedure, we provide a brief summary here. To create a corpus of media articles, Potthast and colleagues crawl Twitter and sample 38,517 tweets from popular media handles like Buzzfeed, ESPN, The New York Times and The Daily Mail between Dec 1, 2016 and April 30, 2017. Of these, 19,538 articles are available as a training set for their machine learning challenge. This data set is available in two archives, downloadable publicly[5]. The remaining data are not publicly available, as they are used to validate contestants' clickbait detection algorithms in a private server (Potthast et al., 2018a).

The Webis Clickbait Challenge 2017 *archive* data set (size: 94.3 GB) contains urls of each article embedded in the tweets, as well as web archives of each article concerned. This is a master data set containing all the raw data about the 19,538 articles in the corpus. The *training* data set (size: 894 MB) contains titles, text contents as paragraphs, post time and number of media files in each media post, and clickbait ratings by Mechanical Turk workers, but not the urls or names of publishers. We extract the urls which are separately available in the larger *archive* data set, thus inferring the publisher of each post. While we directly infer most publishers from their urls (e.g. The New York Times uses urls of the form `nyti.ms` and Buzzfeed uses `bzfd.it`), the identities of a few publishers like The Guardian who uses third party link shortening services (e.g. `bit.ly`, `tinyurl.com` `trib.al`) are inferred by an R script that pings these shortened urls, and then decodes their longer versions. We also infer the number of times a url has been shared on Twitter using a Python script implementing the Selenium browser automation tool, that searches Twitter for each article's url and records the number of retweets of each tweet featuring the said link has received. This Twitter search can capture original links and all its link shortened variants as well, allowing us to capture all shares of a given url on the platform. We record the number of shares as $retweets$, used as a dependent variable to test Hypothesis 7.

After omitting links with incorrect ID matches between archive and training data sets, links that are no longer detected on Twitter by our script (some posts may be deleted) and incorrectly parsed headlines[6], we are left with a corpus of 19,384 usable articles for our analysis. We record each article's publisher as $publisher$ in our data set - this serves two purposes: to append media organization characteristics as we describe in the next paragraph, and to cluster standard errors

[5] url: `https://webis.de/data/webis-clickbait-17.html`

[6] Some list format articles direct the user to click through multiple pages where each page has its own headline. Potthast et al. (2018b)'s algorithm sometimes incorrectly appends all these sub-headlines into a single string, showing headlines with 200+ words, which we omit

by publisher as detailed in the "results" section.

We append the following media organization characteristics to our data set: (a) $paid = 1$ if the organization has at least some way to have paid subscribers (eg. The New York Times) and 0 otherwise (e.g. Buzzfeed). This is to test Hypothesis 4, (b) $webonly = 1$ if it is a web-only portal (e.g. Buzzfeed) and 0 otherwise (i.e. it has a paper, TV, radio etc. presence; e.g. The New York Times). This is to test Hypothesis 5, (c) $editorial$ = fraction of people in the top management team who have an editorial, journalistic or news production background. This is to test Hypothesis 6. To capture $editorial$, we characterize the backgrounds of each person in an organization's top management team, usually available as an executive bio on its "about us" page or equivalent. Human resource professionals, marketing managers, venture capitalists, lawyers, etc. are assigned a score of 0 while people with journalistic, editorial, news production backgrounds are assigned a score of 1. Seven organizations do not have such pages, but we trace their top management teams using other online sources[7] on the web. We also create a variable $general = 1$ if the media organization covers several kinds of news (e.g. The New York Times) or 0 if it is specialized (e.g. Bleacher Report specializes in sports) which is used as a control variable. Finally we record the number of followers of each media organization's Twitter handle, $twitterfollows$ as a control variable to test Hypothesis 7.

We are unsuccessful in obtaining reliable top management team information for two organizations - Breitbart News and Yahoo! While the former is famously secretive about its ownership patterns and organizational structure, the latter has been in a state of organizational turmoil for many years now. Thus, we omit articles from these two organizations in our data, leaving us with 17,943 observations. Table 1 provides a list of publishers and their proportions of clickbait. Finally, based on each article's post date, we create a dummy variable $weekend = 1$ if posted on a weekend and 0 if posted on a weekday, to be used as a control variable. We also record $titlelength$ as the number of words in an article's title, to be used as a control variable.

Insert Table 1 about here

Potthast et al. (2018b) have each headline rated by five Amazon Mechanical Turk workers (who are presented a definition of clcikbait, a title and a link to the associated url that they can then visit) on a Likert scale (0 = "Not clickbaiting"; .33 = "Slightly clickbaiting"; .66 = "Considerably clickbaiting"; 1 = "Heavily clickbaiting"), and the raw scores are available in their *training* data set along with mean, median and mode. They use the mode of the five ratings to come up with a binary 0/1 $clickbait$ variable, yielding 24.37% of all posts being classified as clickbait. We use the same conceptualization as our dependent variable $clickbait$ here. This is used as the dependent variable for Hypotheses 1 - 6, and as an endogenous treatment condition

[7] We use sources like Bloomberg, Wikipedia and LinkedIn, taking care to ensure convergence between multiple such sources in case the top management team is not explicitly mentioned on the organization's "about us" page

for Hypothesis 7. Some examples of clickbait and non-clickbait headlines are provided in Table 2.

Insert Table 2 about here

We now proceed to describe our further augmentation of the data set using sentiment analysis and topic modeling.

Sentiment Analysis

We now describe our methods for extracting sentiments from each article's title and body. In its essence, sentiment analysis involves assigning a sentiment score to each word in a body of text, based on predefined dictionaries. We perform sentiment analysis using R's *syuzhet* package, which consists of four training dictionaries - the syuzhet lexicon developed by the Nebraska Literary Lab, the opinion lexicon (Hu and Liu, 2004), the AfinnWord database (Hansen et al., 2011) and the NRC emotion lexicon (Mohammad and Turney, 2010). These widely accepted sentiment analysis lexicons contain several thousands of English words, with emotion scores between -1 and 1 associated with each word. To make the sentiment scores independent of title length, we take the mean of all the non-zero sentiment words as the sentiment score, with some exceptions where all words are neutral, being assigned a score of zero. The sentiment score of each title, $titlesenti$ (used to test Hypothesis 3) and body, $parasenti$ is thus a number between -1 and 1. We also calculate another variable $senti\_dissim = |titlesenti - parasenti|$, i.e., the absolute value of the difference between title sentiment and main body sentiment to capture dissimilarity between sentiments in the title and main body of each article. This variable is used to test Hypothesis 2. Figures 2 and 3 present distributions of title sentiment and paragraph sentiment in our data while Figure 4 shows the dissimilarities between title and paragraph sentiments in our data.

Insert figures 2, 3 and 4 consecutively about here

Topic Modeling

In addition to sentiment analysis, we perform topic modeling on the body text of each article, using Latent Dirichlet Allocation (Blei et al., 2002, 2003; Tirunillai and Tellis, 2014; Zhang et al., 2017). The idea of topic modeling and Latent Dirichlet Allocation is to let an unsupervised machine learning algorithm discover clusters of similar themed articles from a set of text articles, based on words appearing in them. It is an increasingly popular method of segmenting unstructured text data today, especially in social media and news analysis. In Latent Dirichlet Allocation, the number of clusters k is pre-specified to the algorithm, similar to $k-$means clustering used to segment structured numeric data. After removing filler words from the text corpus, we implement Latent Dirichlet Allocation with Gibbs sampling using R's *topicmodels* package (Hornik and Grün, 2011).

One issue in Latent Dirichlet Allocation is in determining the optimal number of clusters. While higher numbers of clusters have lower perplexity and higher likelihood, it gets increasingly difficult for human coders to unambiguously label each cluster as their number increases. We thus run our code several times from 2, 3 ... 25 clusters and also for 30, 50 and 70 clusters and both perplexity decreases and likelihood increases monotonically as the number of clusters goes up[8]. As Latent Dirichlet Allocation on our corpus is extremely computationally intensive, with each run taking anywhere between 6 and 15 hours, we do not attempt to find the optimum number of clusters beyond which perplexity increases and/or likelihood decreases, especially as it is extremely difficult for manual coders to unambiguously label anything more than 15-20 clusters, and that Latent Dirichlet Allocation is prone to over-fitting. Such a situation is already noted in Zhang et al. (2017). Similar to their approach, we resolve this issue by running the probit model (model 2) of the "results" section for 2, 3 ... 25 clusters and observing their Akaike (AIC) and Bayesian (BIC) information criteria. We observe a sharp drop in AIC and BIC up to 12 clusters after which both AIC and BIC seem to taper off (see Figure 5), oscillating up and down a little bit as in Figure 5. Based on this, we set our number of clusters as 12.

Insert Figure 5 about here

Table 3 presents the top 20 keywords present in each of the 12 clusters discovered by our Latent Dirichlet Allocation implementation, with labels being assigned manually after discussion among co-authors. Cluster probabilities output by our code are appended to the data, with each probability being named after the respective cluster title. The variables $entertainment$, $health$, $socialmedia$, $lifestyle$ and $travel$ are all related to Hypothesis 1, and used as theoretical variables, while the rest are used as control variables in our further analysis. Figure 6 provides distributions of these probabilities.

Insert Figure 6 about here

Finally, Table 4 provides descriptive statistics of dependent, theoretical and control variables in our data set.

Insert tables 3 and 4 consecutively about here

[8] Topic modeling is run including articles for Breitbart and Yahoo! in the corpus, and these are later dropped as reliable data on their top management teams are not found as described in the "data and its augmentation" section. This allows us a larger sample of articles to perform topic modeling on

Results

In this section we present results of two tests: (a) a test of Hypotheses 1 - 6 presented in the "antecedents of clickbait" subsection using a probit model and (b) a test of Hypothesis 7 presented in the "rebroadcast consequences of clickbait" subsection using propensity score matching coupled with ordinary least squares regression. In (b), we use propensity score matching as we recognize that $clickbait$ is an endogenous treatment, and thus use the probit model of (a) as the selection criterion for $clickbait$, and test Hypothesis 7 after that, controlling for number of followers of the media organizations' Twitter handles.

Antecedents of Clickbait

We test Hypotheses 1 - 6 here using the following probit model

$$P(clickbait = 1) = \text{probit}(X\beta) \tag{1}$$

where X represents the matrix of covariates presented in Table 4 and β is a vector of coefficients of theoretical and control variables. We estimate this model clustering standard errors by publisher. The estimated coefficients are presented in Table 5 (model 2 incorporates non-linear effects of $titlesenti$, $parasenti$ and $titlelength$, while model 1 does not). For both models, we choose $sports$ as the base cluster because its probability has the lowest correlation with $clickbait$.

Insert Table 5 about here

From the results we see that while the coefficient of $entertainment$ is not significant even at a 0.1 significance level, the coefficients of $health$, $socialmedia$, $lifestyle$ and $people$ are significant at a 0.01 significance level and $travel$ is significant at a 0.1 significance level. This gives mixed support to Hypothesis 1 which states that articles with a human angle, dealing with entertainment and lifestyle-related topics are more likely to be perceived as clickbait. Hypothesis 2 is supported at a 0.05 significance level as evidenced by the coefficient on $senti\_dissim$, indicating that titles whose valence is dissimilar from the main article are more likely to be perceived as clickbait. Model 1 supports Hypothesis 3 at a 0.05 significance level, though the negative (and significant at 0.05 significance level) coefficient on $titlesenti^2$ in Model 2 suggests an inverted U-shaped relationship, indicating that extremely positive titles are less likely to be perceived as clickbait as compared to moderately positive titles, possibly because the element of curiosity may be missing. However, we encourage further research to test this post hoc explanation.

We do not find any empirical support for Hypothesis 4 even at a 0.1 significance level as evidenced by the coefficient of $paid$, but very strong evidence for Hypothesis 5 as evidenced by the positive (and significant at 0.05 significance level) coefficient of $webonly$. This suggests that while web-only organizations indeed do more clickbait than their offline-also counterparts, clickbait levels are similar whether or not these organizations have a paid service.

Finally, we find strong evidence for Hypothesis 6 as evidenced by the negative (and significant at 0.05 significance level) coefficient of $editorial$. This corroborates that organizations with higher fraction of people with editorial or journalistic backgrounds in their top management teams are less likely to employ clickbait.

Rebroadcast Consequences of Clickbait

We now evaluate whether clickbait is shared more or less than non-clickbait; as per Hypothesis 7 we expect that it is rebroadcast less on social media. We recognize here that “clickbait” is an endogenous treatment condition in an experiment (with “non-clickbait” as the control condition). Thus, we use propensity score matching, a well established method to deal with this selection bias to generate a sample of non-clickbait (control) articles that closely match clickbait (treatment) articles based on observable characteristics in our data (e.g. Rosenbaum and Rubin, 1983; Dehejia and Wahba, 2002; Rishika et al., 2013; Hofmann et al., 2017).

We briefly discuss our propensity score matching method here. As in the potential outcome approach (Rubin, 1974), we define average treatment effect on the treated (ATT) as,

$$ATT = E[Y(1)|D = 1] - E[Y(0)|D = 1] \tag{2}$$

In equation (2), $Y(1)$ indicates the potential outcome (number of retweets) with treatment (clickbait) and $Y(0)$ is the potential outcome (number of retweets) without treatment (non-clickbait). D is a treatment indicator: $D = 1$ for clickbait and 0 for non-clickbait. The second term on the right hand side of equation (2) is thus the average number of retweets the clickbait (treatment) articles would have got if they were non-clickbait (control) articles instead. This is unobservable because we cannot observe the number of retweets of the same clickbait articles if they were non-clickbait (a hypothetical scenario). However, we do observe $E[Y(0)|D = 0]$, the number of retweets for observation in the non-clickbait (control) group, and therefore specify a difference Δ as,

$$\begin{aligned}\Delta &= E[Y(1)|D = 1] - E[Y(0)|D = 0]\\ &= E[Y(1)|D = 1] - E[Y(0)|D = 1] + E[Y(0)|D = 1] - E[Y(0)|D = 0]\\ &= ATT + SB\end{aligned} \tag{3}$$

The second term SB in equation (3) is the *selection bias*, and ATT (average treatment effect on the treated) can be estimated by the difference (Δ) in mean observed retweets (outcome) between clickbait (treatment) and non-clickbait (control) conditions if $SB = 0$. The objective of propensity score matching is thus to select observations from the control group that most closely match observations in the treatment group on observable characteristics, allowing us then to infer causality of the treatment (clickbait) variable on the outcome (number of retweets), analogous to a randomized experiment (Rubin, 2006).

The first step in propensity score matching is to estimate the propensity scores for all observations in the data. We use model 2 of the probit model specified by equation (1) for this purpose using $P(clickbait = 1)$ as the propensity score. We then match the titles in the treatment (clickbait) and control (non-clickbait) groups using these calculated propensity scores using a one-to-one matching algorithm, to match 3,833 clickbait (treatment) titles with 3,833 corresponding non-clickbait (control) titles. The next step is to ensure that the common support condition holds (Rosenbaum and Rubin, 1983). This means that we check if there is a sufficient overlap between the characteristics of clickbait and non-clickbait observations. As per Lechner (2002)'s recommendation, we plot distributions of propensity scores before matching and after matching as box plots in Figure 7, and histograms in Figure 8, offering some visual confirmation of the common support condition. Six observations from the treatment (clickbait) group lack common support and hence are removed from the analysis.

Insert figures 7 and 8 consecutively about here

To further assess the quality of the matching procedure, we compare the mean values of the covariates before and after matching. The results are reported in Table 6. All mean differences that were significant before the matching became insignificant after the matching. Thus the variables that are used for propensity score matching are not significantly different across the control (non-clickbait) and the treatment (clickbait) group and hence satisfy balance of covariates.

Insert Table 6 about here

Table 7 presents the results of the propensity score matching with respect to the outcome (number of retweets) variable. These results indicate that there is a mean difference of 62.35 retweets between non-clickbait articles (137.10) and clickbait articles (74.75). In other words, articles which are clickbait have on average 62.35 fewer retweets as compared to non-clickbait articles, after controlling for selection effects. Given that an article in our data set gets on average 144.72 retweets, a difference of 62.35 retweets is stark. This difference is even higher at 88.96 retweets when we do not account for selection effects. Both these differences are significant at a 0.01 significance level.

Insert Table 7 about here

Though propensity score matching provides a robust estimation of the average number of retweets of clickbait and non-clickbait articles after accounting for selection bias, retweets can also depend on other characteristics like Twitter follower count. Since propensity score matching creates a control (non-clickbait) group that is similar to the treatment (clickbait) group except for whether the title is clickbait or not, we can find the impact of clickbait on retweets by running the following linear regression,

$$retweets = \alpha + \delta clickbait + \gamma X + \varepsilon \tag{4}$$

where $retweets$ (number of retweets) is the dependent variable and X is a set of covariates in Table 4. We can estimate the coefficient δ which captures the impact of clickbait on retweets. According to Hypothesis 7, the coefficient δ must be negative and this is confirmed by the regression result in Table 8.

Model 1 in Table 8 shows that after controlling for Twitter follower count, the average number of retweets of a clickbait article is 69.85 retweets less than the average number of retweets of a non-clickbait article. Model 2 in Table 8 shows that after controlling for Twitter follower count and other variables, the average number of retweets of a clickbait article is 68.71 retweets less than the average number of retweets of a non-clickbait article. The effects are larger in size when compared to ATT as in Table 7.

Insert Table 8 about here

Discussion

We now present a brief discussion about our study's theoretical and managerial implications, along with suggestions for future research.

Theoretical and Managerial Implications

Clickbait has been discussed a lot in popular media, but received scant academic attention, except for machine learning approaches to detect it. Given its prevalence today, and the changing landscape of the news business, we thus present, to the best of our knowledge, the first study of clickbait from a marketing research angle. We look at clickbait through the lenses of curiosity research and upper echelons theory to identify its antecedents. Hypotheses 1 - 3 deal with the contents of clickbait; we observe that human interest articles dealing with topics like health, social media, lifestyle, people and travel but not entertainment are more likely to be perceived as clickbait. Articles with moderately positive title valence and whose main content valence is dissimilar to the title valence, are more likely to be perceived as clickbait.

Hypotheses 4 - 6 shed light on the organizational antecedents of clickbait. We find very strong evidence for our upper echelons theory-based Hypothesis 6 that media organizations with higher proportions of people with editorial or journalistic backgrounds in their top management team are less likely to employ clickbait (than organizations with lower proportions of people with editorial or journalistic backgrounds), due to clickbait's incompatibility with traditional journalistic principles. Also as proposed in Hypothesis 5, we find that web-only portals tend to use more clickbait than those with offline presences also. However, we are unable to show support for Hypothesis 4, that organizations with a paid outlet are less likely to do clickbait than purely free outlets. This actually indicates the prevalence of clickbait across different types of revenue models, and its mainstreaming today.

We also shed light on whether clickbait is rebroadcast more or less than non-clickbait on social media. We find overwhelming evidence, even after controlling for a media organization's Twitter follower count, that clickbait articles are rebroadcast less than non-clickbait articles. This has important consequences for media managers and digital marketers alike. Though our data set does not have direct website traffic numbers, we quantify the negative effect of clickbait on social media rebroadcasting, thus reducing the reach of an organization's online content. This is therefore an unambiguous warning for both media organizations employing clickbait, as well as brands advertising along with the associated content. Our work thus contributes to the sharing literature (e.g. Berger and Schwartz, 2011; Berger and Milkman, 2012; Berger, 2014b; Zhang et al., 2017; Tellis et al., 2019), shedding light on an important phenomenon that has emerged in digital media over the last two decades.

Scope for Future Research

Our work contributes to the growing literature on sharing on social media (Berger, 2014b) and the general area of digital marketing (Lamberton and Stephen, 2016; Kannan and Li, 2017). Given the importance of online news and the emergence of clickbait today, we suggest that our study can be extended in several possible ways. First, a longitudinal investigation of clickbait than our data set allows can shed light on its evolution over time. Second, we suggest a systematic investigation into which discrete emotions like awe, anger and anxiety are related to clickbait, in the manner of Berger and Milkman (2012). We are unable to do this due to resource constraints. Third, experimental investigations of how attitudes to advertisements accompanying clickbait may be different from attitudes to advertisements accompanying non-clickbait in various contexts would be highly beneficial to digital marketers and media organizations alike.

Conclusion

We present a study of an important and recently emergent phenomenon in digital news media, that has thus far received great attention from popular media, but little attention from academic research. Our study is thus the first of its kind to systematically investigate clickbait's antecedents and rebroadcast consequences, using well established theories of curiosity and upper echelons theory to frame our hypotheses. Apart from shedding light on these, our study is also useful to digital marketing educators and practitioners, showcasing the use of multiple relevant tools and methodologies including web scraping, sentiment analysis, topic modeling using Latent Dirichlet Allocation, linear and probit regressions and propensity score matching. We also reiterate our words of caution to media organizations and digital marketers, that clickbait reduces rebroadcasting, and therefore the reach of online articles. Anyone should consider using it very carefully, if at all.

References

Akpinar, E. and Berger, J. (2017). Valuable virality. *Journal of Marketing Research*, 54(2):318–330.

Babu, A., Liu, A., and Zhang, J. (2017). New updates to reduce clickbait headlines. *Facebook Newsroom*.

Bantel, K. and Jackson, S. (1989). Top management and innovations in banking: does the composition of the top team make a difference? *Strategic Management Journal*, 10:107–124.

Barker III, V. L. and Mueller, G. C. (2002). CEO characteristics and firm R&D spending. *Management Science*, 48(6):782–801.

Berger, J. (2011). Arousal increases social transmission of information. *Psychological Science*, 22(7):891–893.

Berger, J. (2014a). Why clickbait fails. *Jonah's Blog*.

Berger, J. (2014b). Word of mouth and interpersonal communication: A review and directions for future research. *Journal of Consumer Psychology*, 24(4):586–607.

Berger, J. and Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, 49(2):192–205.

Berger, J. and Schwartz, E. M. (2011). What drives immediate and ongoing word of mouth? *Journal of Marketing Research*, 48(5):869–880.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2002). Latent dirichlet allocation. In *Advances in neural information processing systems*, pages 601–608.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Chaganti, R. and Sambharya, R. (1987). Strategic orientation and characteristics of upper management. *Strategic Management Journal*, 8(4):393–401.

Dearborn, D. C. and Simon, H. A. (1958). Selective perception: A note on the departmental identifications of executives. *Sociometry*, 21(2):140–144.

Deci, E. L. (1992). The relation of interest to the motivation of behavior: A self-determination theory perspective. *The Role of Interest in Learning and Development*, 44.

Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161.

DeMers, J. (2017). Is clickbait dying, or stronger than ever? *Forbes*.

Foster, E. K. (2004). Research on gossip: Taxonomy, methods, and future directions. *Review of General Psychology*, 8(2):78–99.

Frampton, B. (2015). Clickbait: The changing face of online journalism. *BBC*.

Geiger, J. (2006). Definition of clickbait.

Golman, R. and Loewenstein, G. (2018). Information gaps: A theory of preferences regarding the presence and absence of information. *Decision*, 5(3):143.

Graves, L. (2019). Jill Abramson on media layoffs: 'The villains are Facebook and Google'. *The Guardian*.

Hambrick, D. C. (2007). *Upper echelons theory: An update*. Academy of Management Briarcliff Manor, NY 10510.

Hambrick, D. C. and Mason, P. A. (1984). Upper echelons: The organization as a reflection of its top managers. *Academy of Management Review*, 9(2):193–206.

Hansen, L. K., Arvidsson, A., Nielsen, F. A., Colleoni, E., and Etter, M. (2011). Good friends, bad news-affect and virality in twitter. In *Future information technology*, pages 34–43. Springer.

Hofmann, J., Clement, M., Völckner, F., and Hennig-Thurau, T. (2017). Empirical generalizations on the impact of stars on the economic success of movies. *International Journal of Research in Marketing*, 34(2):442–461.

Hornik, K. and Grün, B. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Ingram, M. (2014). The internet didn't invent viral content or clickbait journalism — there's just more of it now, and it happens faster. *GigaOm*.

Jaworski, B. and Kohli, A. (1993). Market Orientation: Antecedents and Consequences. *Journal of Marketing*, 57(3):53–70.

Kaiser, R. G. (2014). The bad news about the news. *The Brookings Institution*.

Kannan, P. K. and Li, H. A. (2017). Digital marketing: A framework, review and research agenda. *International Journal of Research in Marketing*, 34(1):22–45.

Kashdan, T. B., Disabato, D. J., Goodman, F. R., and Naughton, C. (2018a). The five dimensions of curiosity. *Harvard Business Review*, 96(5):58–60.

Kashdan, T. B., Stiksma, M. C., Disabato, D. J., McKnight, P. E., Bekier, J., Kaji, J., and Lazarus, R. (2018b). The five-dimensional curiosity scale: Capturing the bandwidth of curiosity and identifying four unique subgroups of curious people. *Journal of Research in Personality*, 73:130–149.

Lamberton, C. and Stephen, A. T. (2016). A thematic exploration of digital, social media, and mobile marketing: Research evolution from 2000 to 2015 and an agenda for future inquiry. *Journal of Marketing*, 80(6):146–172.

Lambrecht, A. and Misra, K. (2016). Fee or free: when should firms charge for online content? *Management Science*, 63(4):1150–1165.

Lechner, M. (2002). Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165(1):59–82.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116(1):75.

Marvin, C. B. and Shohamy, D. (2016). Curiosity and reward: Valence predicts choice and information prediction errors enhance learning. *Journal of Experimental Psychology: General*, 145(3):266.

Menz, M. (2012). Functional top management team members: A review, synthesis, and research agenda. *Journal of Management*, 38(1):45–80.

Mohammad, S. M. and Turney, P. D. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

Neidig, H. (2019). Media layoffs bring heat on Facebook, Google. *The Hill*.

Oosterwijk, S. (2017). Choosing the negative: A behavioral demonstration of morbid curiosity. *PloS one*, 12(7):e0178399.

Pattabhiramaiah, A., Sriram, S., and Manchanda, P. (2017a). Paywalls: monetizing online content. *Journal of Marketing*, 83(4).

Pattabhiramaiah, A., Sriram, S., and Sridhar, S. (2017b). Rising prices under declining preferences: the case of the US print newspaper industry. *Marketing Science*, 37(1):97–122.

Potthast, M., Gollub, T., Hagen, M., and Stein, B. (2018a). The clickbait challenge 2017: towards a regression model for clickbait strength. *arXiv preprint arXiv:1812.10847*.

Potthast, M., Gollub, T., Komlossy, K., Schuster, S., Wiegmann, M., Fernandez, E. P. G., Hagen, M., and Stein, B. (2018b). Crowdsourcing a large corpus of clickbait on twitter. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1498–1507.

Renner, B. (2006). Curiosity about people: The development of a social curiosity measure in adults. *Journal of Personality Assessment*, 87(3):305–316.

Rishika, R., Kumar, A., Janakiraman, R., and Bezawada, R. (2013). The effect of customers' social media participation on customer visit frequency and profitability: an empirical investigation. *Information systems research*, 24(1):108–127.

Rony, M. M. U., Hassan, N., and Yousuf, M. (2017). Diving deep into clickbaits: Who use them to what extents in which topics with what effects? In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 232–239. ACM.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge University Press.

Schaffer, D. (1995). Shocking secrets revealed! the language of tabloid headlines. *ETC: A Review of General Semantics*, 52(1):27–46.

Schulze, C., Schöler, L., and Skiera, B. (2014). Not all fun and games: Viral marketing for utilitarian products. *Journal of Marketing*, 78(1):1–19.

Shearer, E. and Matsa, K. E. (2018). News use across social media platforms 2018. Technical report, Pew Research Center.

Silvia, P. J. (2008). Appraisal components and emotion traits: Examining the appraisal basis of trait curiosity. *Cognition and Emotion*, 22(1):94–113.

Smith, B. (2014). Why BuzzFeed doesn't do clickbait. *BuzzFeed*.

Tellis, G. J., MacInnis, D. J., Tirunillai, S., and Zhang, Y. (2019). What Drives Virality (Sharing) of Online Digital Content? The Critical Role of Information, Emotion, and Brand Prominence. *Journal of Marketing*, pages 1–20.

Tirunillai, S. and Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, 51(4):463–479.

Vousughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359:1146–1151.

Wiersema, M. and Bantel, K. (1992). Top management team demography and corporate strategic change. *Academy of Management Journal*, 35(1):91–121.

Wiggin, K. L., Reimann, M., and Jain, S. P. (2018). Curiosity tempts indulgence. *Journal of Consumer Research*, 45(6):1194–1212.

Zhang, Y., Moe, W. W., and Schweidel, D. A. (2017). Modeling the role of message content and influencers in social media rebroadcasting. *International Journal of Research in Marketing*, 34(1):100–119.

Zuckerman, M. (1979). Sensation-seeking: beyond optimal arousal. *Hillside, NJ: Earlbaum*.

Zuckerman, M. (2014). *Sensation seeking (psychology revivals): Beyond the optimal level of arousal*. Psychology Press.

Table 1: A list of publishers and their clickbait usage in our data

Publisher Name	Number of Observations	Percentage of Clickbait
ABC News (Australia)	744	11.83
ABC News (US)	673	5.05
BBC	739	24.36
Billboard	762	10.63
Bleacher Report	548	8.03
Bloomberg	739	20.84
Business Insider	686	31.34
BuzzFeed	735	62.72
CBS News	722	12.19
CNN	717	15.06
Complex	762	16.53
Daily Mail	732	22.4
ESPN	360	23.05
Forbes	737	38.53
Fox News	757	5.81
The Guardian	789	14.7
HuffPost	776	21.39
Independent	720	34.17
Indiatimes	745	37.58
Mashable	698	36.67
NBC News	790	11.52
The New York Times	775	20.13
The Telegraph	737	16.69
The Washington Post	766	20.50
The Wall Street Journal	734	12.81
Total	17,943	21.39

Table 2: Examples of clickbait and non-clickbait titles from our data

Headline	**Publisher**	**Clickbait?**
A Superior Chicken Soup	The New York Times	*Yes*
Leah Remini's Reddit AMA reveals juicy secrets of Scientology	Mashable	*Yes*
Can Work Life Balance Be a Reality? This Company Makes it Possible	NBC News	*Yes*
Visit Myanmar's Capital Now! There's Still a Lot Not to See	The Wall Street Journal	*Yes*
Panama Papers: Europol links 3,500 names to suspected criminals	The Guardian	*No*
100 Women 2016: On the frontline with the women policing the peace in Afghanistan	BBC	*No*
Older Viewers and Conservatives Are Watching Less NFL, Survey Finds	The Wall Street Journal	*No*
5 Dead, 7 Injured After Tornadoes in Alabama and Tennessee	ABC News (US)	*No*

Table 3: Top 20 keywords in each cluster

	ENTERTAINMENT	CRIME	SPORTS	HEALTH	SOCIALMEDIA	LIFESTYLE
1	show	polic	game	women	twitter	food
2	year	offic	team	peopl	facebook	just
3	star	year	season	school	com	wear
4	film	told	play	health	news	beauti
5	music	report	year	student	breitbart	dress
6	best	man	player	studi	report	photo
7	song	famili	win	research	comment	dog
8	perform	attack	three	univers	media	hair
9	movi	investig	back	year	pic	fashion
10	award	peopl	time	children	send	christma
11	week	kill	top	care	email	eat
12	album	old	run	work	video	design
13	love	death	defens	educ	follow	brand
14	celebr	charg	second	mani	share	product
15	play	accord	week	age	tip	buzzfe
16	fan	arrest	coach	medic	made	best
17	releas	case	just	risk	page	time
18	episod	citi	bowl	drug	tweet	old
19	night	home	leagu	chang	site	reaction
20	artist	court	best	caus	articl	style

	PEOPLE	INTLPOLITICS	TRAVEL	USPOLITICS	ECONOMY	GEOPOLITICS
1	just	countri	citi	trump	year	state
2	time	govern	year	presid	compani	unit
3	peopl	parti	water	hous	market	countri
4	know	minist	car	obama	million	attack
5	don	polit	flight	white	busi	north
6	work	nation	area	elect	tax	russia
7	realli	year	travel	state	percent	offici
8	feel	peopl	build	donald	job	forc
9	got	vote	world	american	billion	secur
10	life	european	air	republican	pay	china
11	didn	elect	imag	democrat	rate	russian
12	talk	support	around	administr	price	militari
13	back	prime	space	nation	money	war
14	live	leader	near	campaign	plan	syria
15	tri	british	power	washington	cost	govern
16	alway	london	park	senat	bank	nation
17	love	world	airlin	law	fund	group
18	happen	india	time	order	industri	korea
19	person	europ	hotel	immigr	invest	foreign
20	everi	union	mile	vote	month	report

Table 4: Summary statistics

Variable	**Variable type**	**Mean**	**Standard Deviation**
Dependent variables:			
$clickbait$	Dummy	0.21	0.41
$retweets$	Continuous	144.72	527.15
Theoretical variables:			
$titlesenti$	Continuous	-0.01	0.44
$editorial$	Continuous	0.37	0.24
$senti\_dissim$	Continuous	0.32	0.24
$entertainment$	Continuous	0.08	0.11
$health$	Continuous	0.08	0.10
$socialmedia$	Continuous	0.07	0.07
$lifestyle$	Continuous	0.07	0.08
$people$	Continuous	0.10	0.08
$travel$	Continuous	0.08	0.09
$paid$	Dummy	0.59	0.50
$webonly$	Dummy	0.32	0.46
Common control variables for both dependent variables:			
$economy$	Continuous	0.09	0.12
$crime$	Continuous	0.09	0.11
$sports$	Continuous	0.08	0.13
$intlpolitics$	Continuous	0.08	0.09
$uspolitics$	Continuous	0.10	0.12
$geopolitics$	Continuous	0.07	0.10
$titlelength$	Continuous	11.47	4.23
$parasenti$	Continuous	0.10	0.22
$titlesenti^2$	Continuous	0.19	0.23
$parasenti^2$	Continuous	0.06	0.08
$titlelength^2$	Continuous	149.44	136.20
$general$	Dummy	0.66	0.47
$weekend$	Dummy	0.28	0.45
Control variables for rebroadcast:			
$twitterfollows$	Continuous	11.66	11.64

$sports$ is used as baseline topic in future regressions

Table 5: Probit model results. Dependent variable: $clickbait$

Variable	**Model 1**	**Model 2**
$titlesenti$	0.132***	0.159***
	(0.0264)	(0.0289)
$editorial$	-0.490**	-0.491**
	(0.205)	(0.208)
$senti\_dissim$	0.183***	0.295***
	(0.0537)	(0.0674)
$entertainment$	0.439	0.437
	(0.535)	(0.545)
$health$	1.602***	1.610***
	(0.502)	(0.508)
$socialmedia$	1.953***	1.889***
	(0.532)	(0.536)
$lifestyle$	3.151***	3.100***
	(0.508)	(0.509)
$people$	3.011***	3.028***
	(0.527)	(0.537)
$travel$	1.000**	0.991*
	(0.507)	(0.513)
$paid$	0.0752	0.0738
	(0.115)	(0.112)
$webonly$	0.352***	0.383***
	(0.127)	(0.124)
$economy$	1.224**	1.227**
	(0.537)	(0.545)
$crime$	-0.170	-0.164
	(0.538)	(0.544)
$intlpolitics$	0.121	0.104
	(0.520)	(0.525)
$uspolitics$	-0.524	-0.507
	(0.525)	(0.531)
$geopolitics$	-0.0532	-0.0942
	(0.488)	(0.495)
$titlelength$	-0.0135	-0.0600***
	(0.0106)	(0.0227)
$parasenti$	-0.0485	-0.120
	(0.0857)	(0.0835)
$general$	0.133	0.134
	(0.109)	(0.109)
$titlesenti^2$		-0.154**
		(0.0638)
$parasenti^2$		0.233
		(0.229)
$titlelength^2$		0.00153**
		(0.000767)
$Constant$	-1.817***	-1.529***
	(0.502)	(0.509)
Observations	17,943	17,943
AIC	16339.3	16306.25
BIC	16495.2	16485.54

Robust standard errors clustered by publisher in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 6: Summary statistics and covariate comparison before and after matching

		Control Group					
	Treatment Group	Before Matching			After Matching		
Covariate	Mean	Mean	Mean diff	t-stat	Mean	Mean diff	t-stat
$titlesenti$	0.059	-0.029	**0.088**	**11.060**	0.060	-0.001	-0.110
$editorial$	0.319	0.383	**-0.065**	**-15.070**	0.321	-0.003	-0.620
$senti\_dissim$	0.338	0.314	**0.024**	**5.430**	0.337	0.001	0.090
$entertainment$	0.086	0.084	0.002	1.070	0.088	-0.002	-0.870
$health$	0.094	0.077	**0.017**	**9.650**	0.095	-0.001	-0.470
$socialmedia$	0.082	0.069	**0.013**	**10.380**	0.084	-0.002	-1.010
$lifestyle$	0.108	0.061	**0.047**	**31.900**	0.108	0.000	0.080
$people$	0.131	0.090	**0.041**	**26.000**	0.134	-0.003	-1.220
$travel$	0.081	0.080	0.001	0.610	0.077	0.004	1.620
$paid$	0.551	0.596	**-0.045**	**-5.030**	0.547	0.005	0.410
$webonly$	0.466	0.275	**0.192**	**22.960**	0.459	0.007	0.640
$economy$	0.099	0.095	0.003	1.540	0.095	0.004	1.420
$crime$	0.066	0.094	**-0.028**	**-14.210**	0.066	0.001	0.310
$intlpolitics$	0.066	0.080	**-0.014**	**-8.600**	0.066	0.001	0.320
$uspolitics$	0.066	0.107	**-0.041**	**-18.380**	0.069	-0.003	-1.300
$geopolitics$	0.052	0.081	**-0.029**	**-15.350**	0.051	0.001	0.600
$titlelength$	11.390	11.490	-0.100	-1.300	11.342	0.048	0.490
$parasenti$	0.155	0.093	**0.062**	**15.580**	0.151	0.004	0.810
$titlesenti^2$	0.204	0.191	**0.013**	**3.100**	0.205	-0.001	-0.180
$parasenti^2$	0.071	0.057	**0.013**	**8.790**	0.067	0.003	1.510
$titlelength^2$	148.610	149.700	-1.090	-0.440	146.140	2.470	0.820
$general$	0.651	0.668	-0.016	-1.900	0.658	-0.007	-0.650
$weekend$	0.313	0.278	**0.035**	**4.220**	0.300	0.013	1.240

Notes: Mean difference for each covariate is calculated by subtracting the mean of clickbait (control) group from the mean of the non-clickbait (treatment) group. The t-statistics for these differences in mean are also reported. Mean differences when significant at .05 are indicated in bold.

Table 7: Results of propensity score matching with respect to the outcome variable

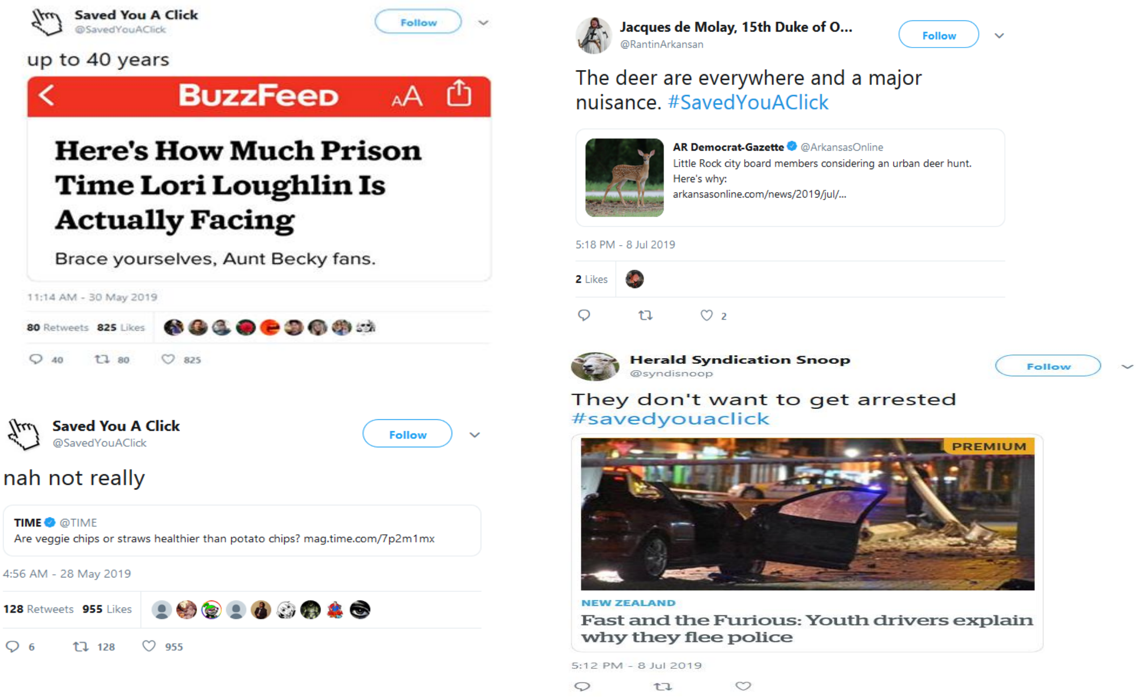
	Treated	Controls	Difference	Standard error	t-stat
Number of retweets UNMATCHED	74.79	163.75	-88.96	9.57	-9.29***
Number of retweets MATCHED (ATT)	74.75	137.10	**-62.35**	8.30	-7.51***

Notes: ***, $p < 0.01$

Table 8: Ordinary Least Squares Regression. Dependent variable: $retweets$.

Variable	Model 1	Model 2
$clickbait$	-69.85***	-68.59***
	(22.83)	(19.27)
$twitterfollows$	3.722***	3.656***
	(0.799)	(0.679)
$titlesenti$		15.24
		(12.00)
$editorial$		-139.2
		(88.31)
$senti\_dissim$		-15.44
		(17.25)
$entertainment$		-207.8
		(245.9)
$health$		-362.8*
		(210.3)
$socialmedia$		-409.5*
		(209.4)
$lifestyle$		-422.1*
		(205.4)
$people$		-367.2**
		(172.5)
$travel$		-422.7*
		(215.3)
$paid$		-2.288
		(38.54)
$webonly$		34.48
		(35.26)
$economy$		-463.1*
		(249.5)
$crime$		-312.7
		(187.5)
$intlpolitics$		-525.3**
		(239.3)
$uspolitics$		-88.19
		(220.7)
$geopolitics$		-244.5
		(187.3)
$titlelength$		-6.430
		(5.235)
$parasenti$		5.254
		(19.87)
$general$		15.37
		(29.41)
$titlesentisq$		19.76
		(15.41)
$parasentisq$		-16.73
		(35.42)
$titlelengthsq$		0.158
		(0.138)
$Constant$	104.7***	507.1**
	(35.94)	(209.6)
Observations	7,666	7,666
R-squared	0.035	0.081

Robust standard errors clustered by publisher in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$



Saved You A Click
@SavedYouAClick

up to 40 years

Here's How Much Prison Time Lori Loughlin Is Actually Facing

Brace yourselves, Aunt Becky fans.

11:14 AM - 30 May 2019

80 Retweets 825 Likes

40 80 825

Saved You A Click
@SavedYouAClick

nah not really

TIME @TIME
Are veggie chips or straws healthier than potato chips? mag.time.com/7p2m1mx

4:56 AM - 28 May 2019

128 Retweets 955 Likes

6 128 955

Jacques de Molay, 15th Duke of O...
@RantinArkansan

The deer are everywhere and a major nuisance. #SavedYouAClick

AR Democrat-Gazette @ArkansasOnline
Little Rock city board members considering an urban deer hunt. Here's why:
arkansasonline.com/news/2019/jul/...

5:18 PM - 8 Jul 2019

2 Likes

2

Herald Syndication Snoop
@syndisnoop

They don't want to get arrested #savedyouaclick

NEW ZEALAND
Fast and the Furious: Youth drivers explain why they flee police

5:12 PM - 8 Jul 2019

Figure 1: Examples of tweets by popular online vigilante Twitter handle `@savedyouaclick` and ordinary users using the hashtag `#savedyouaclick` to put out spoilers on clickbait headlines

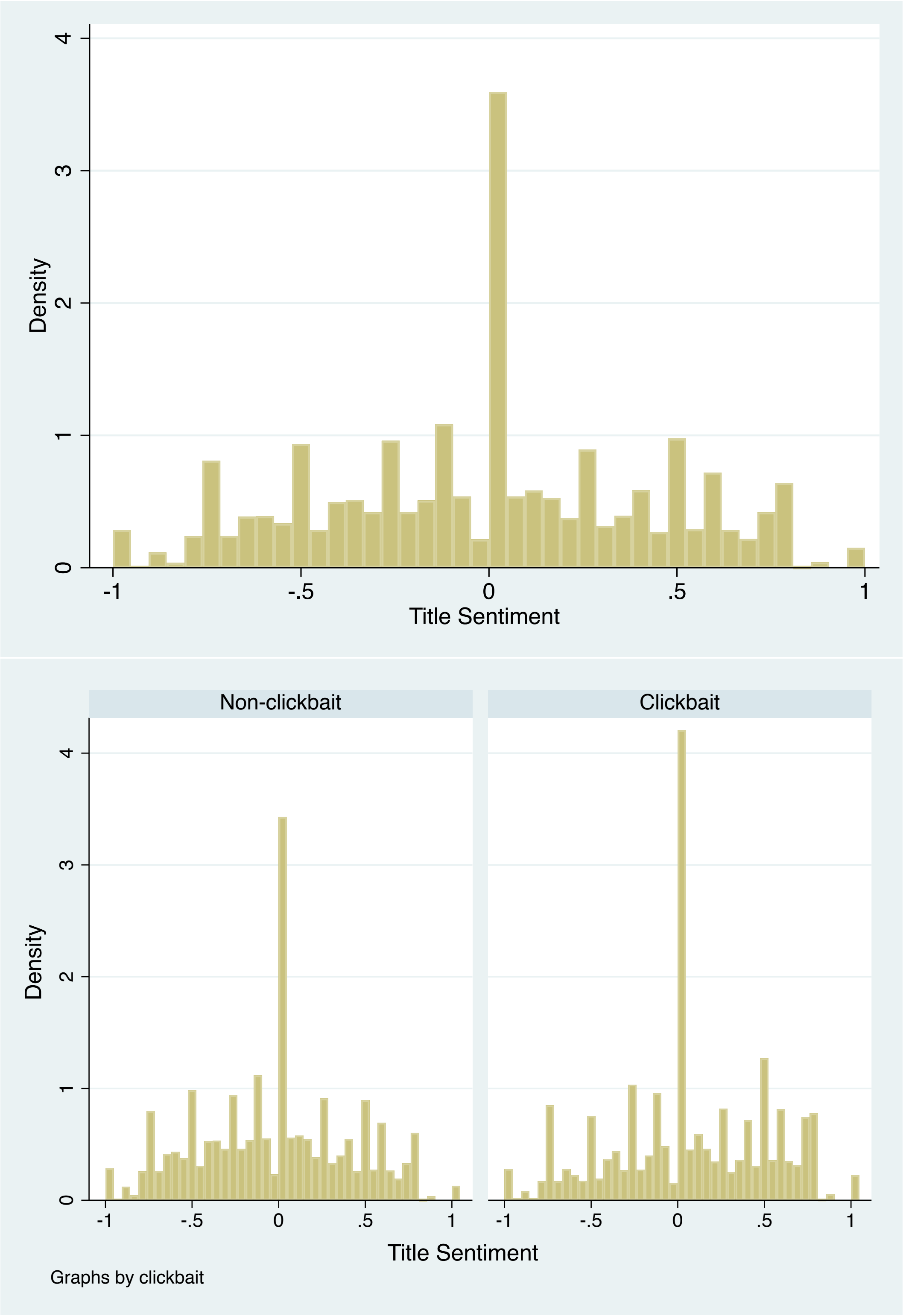


Figure 2: Distribution of sentiments in titles across all posts (upper figure) and separately for non-clickbait and clickbait (lower figure) in our data

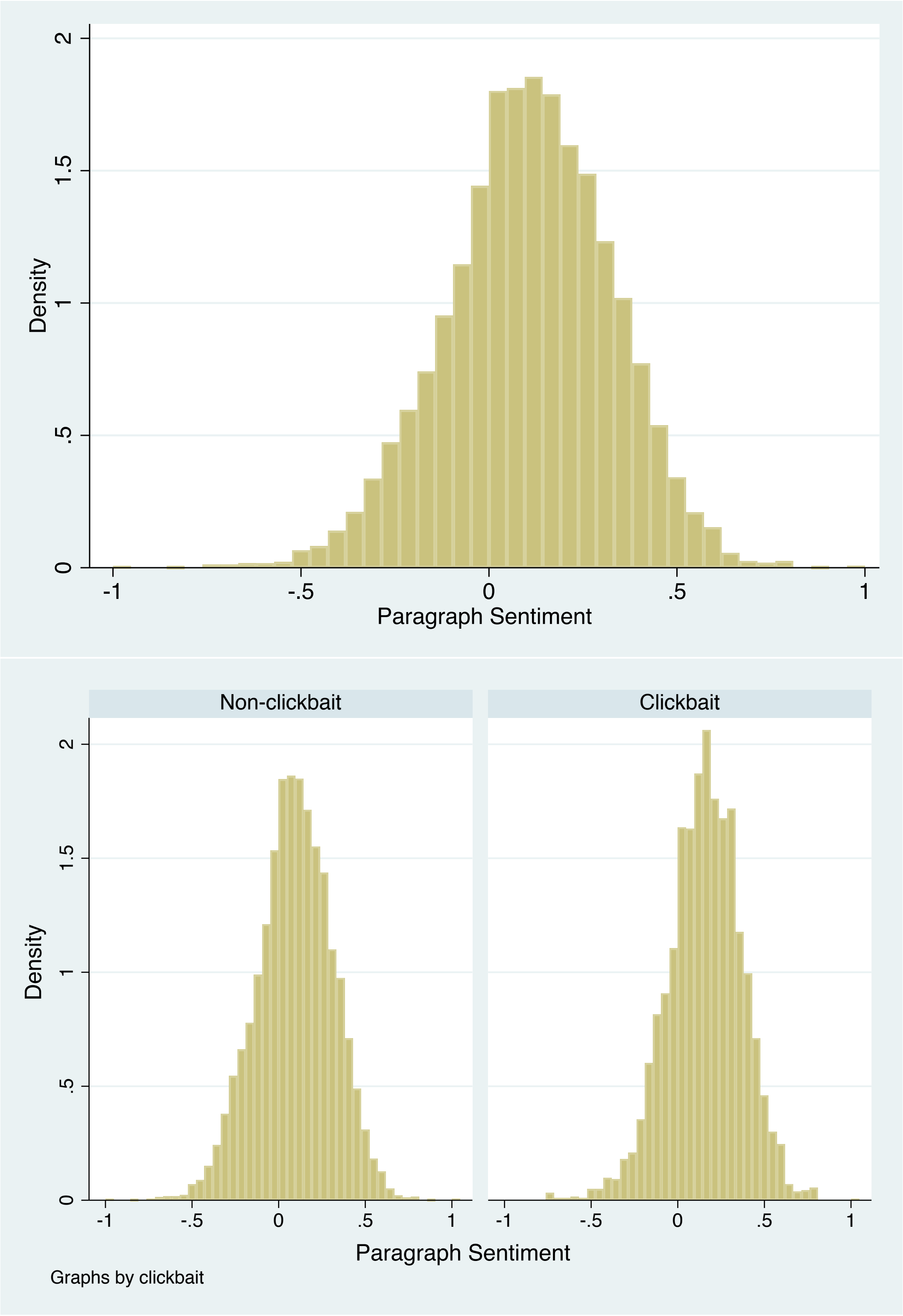


Figure 3: Distribution of sentiments in paragraphs across all posts (upper figure) and separately for non-clickbait and clickbait (lower figure) in our data

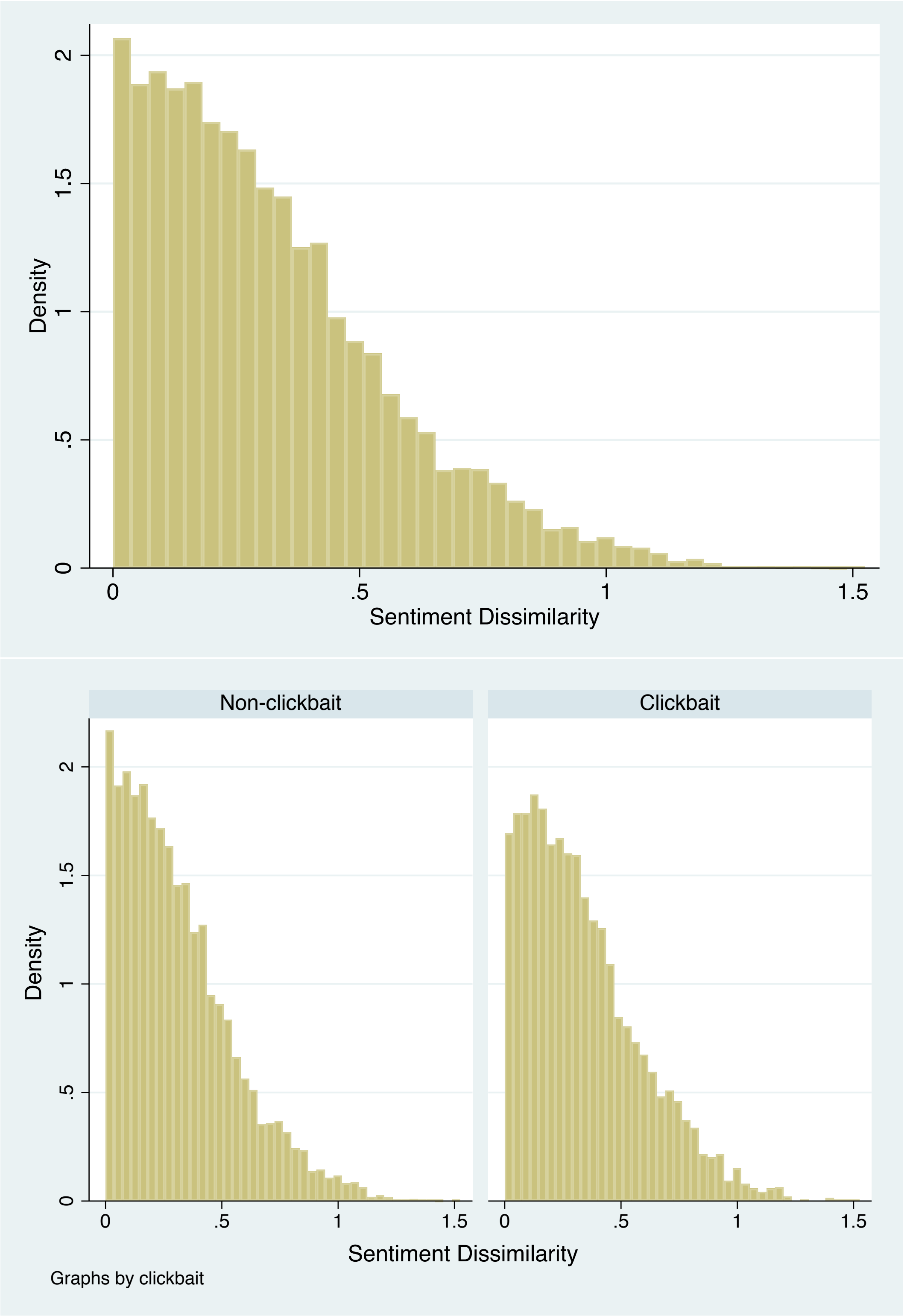


Figure 4: Distribution of sentiment dissimilarity in paragraphs across all posts (upper figure) and separately for non-clickbait and clickbait (lower figure) in our data

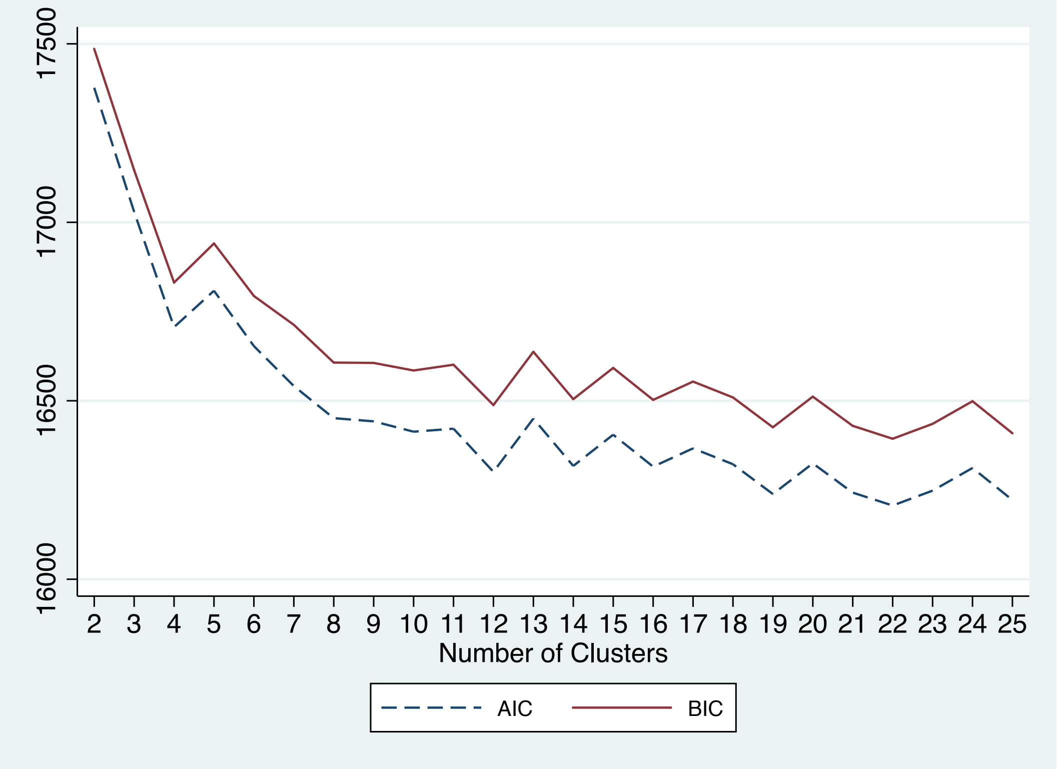


Figure 5: AIC and BIC for probit model versus number of clusters - presented for model 2 in the "results" section

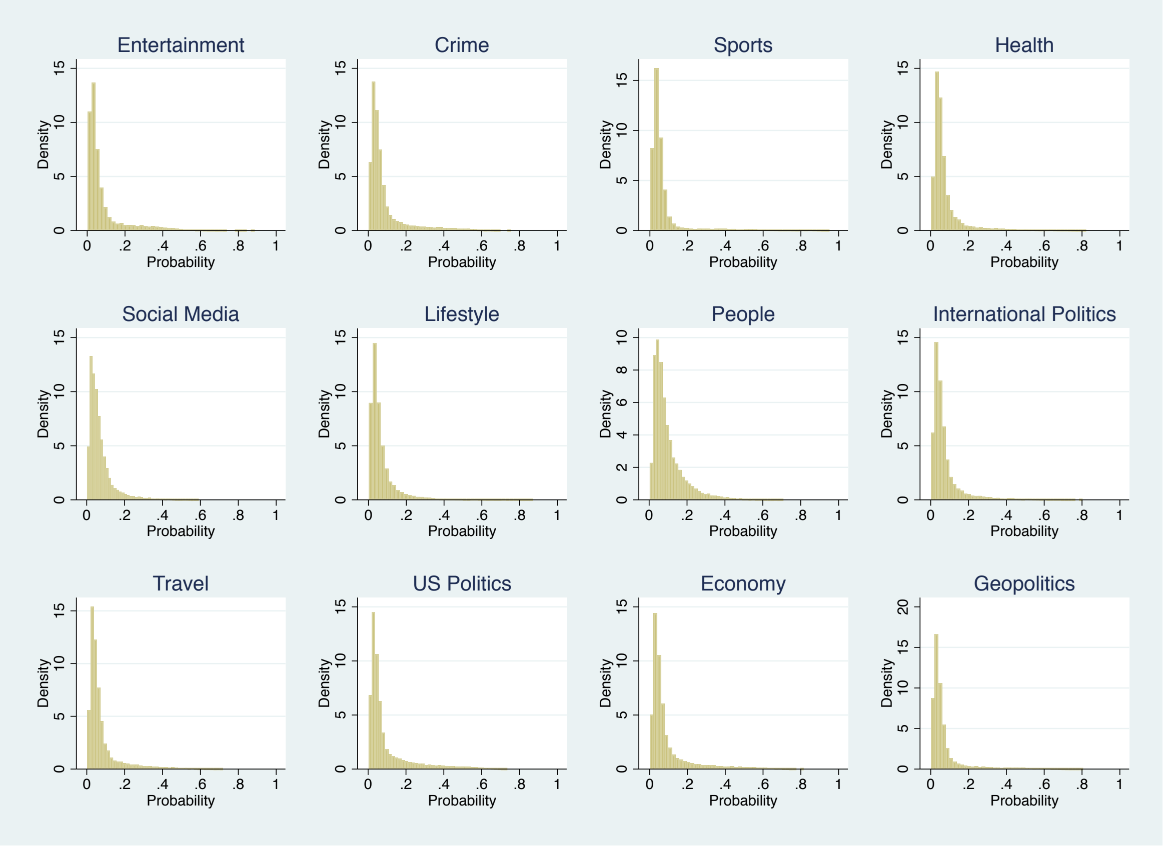


Figure 6: Distribution of probabilities of various topics in our data

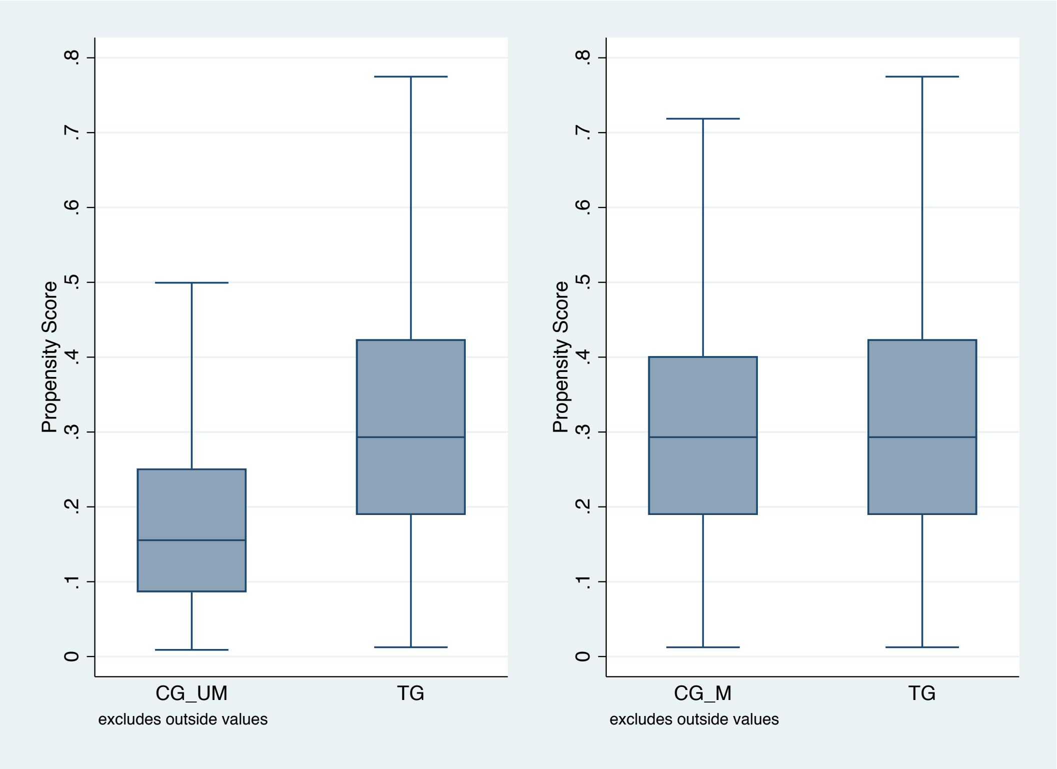


Figure 7: Distribution of Propensity Score Before and After Matching

Notes: TG-Treatment Group, CG_UM-Control Group *before* matching, CG_M-Control Group *after* matching

.

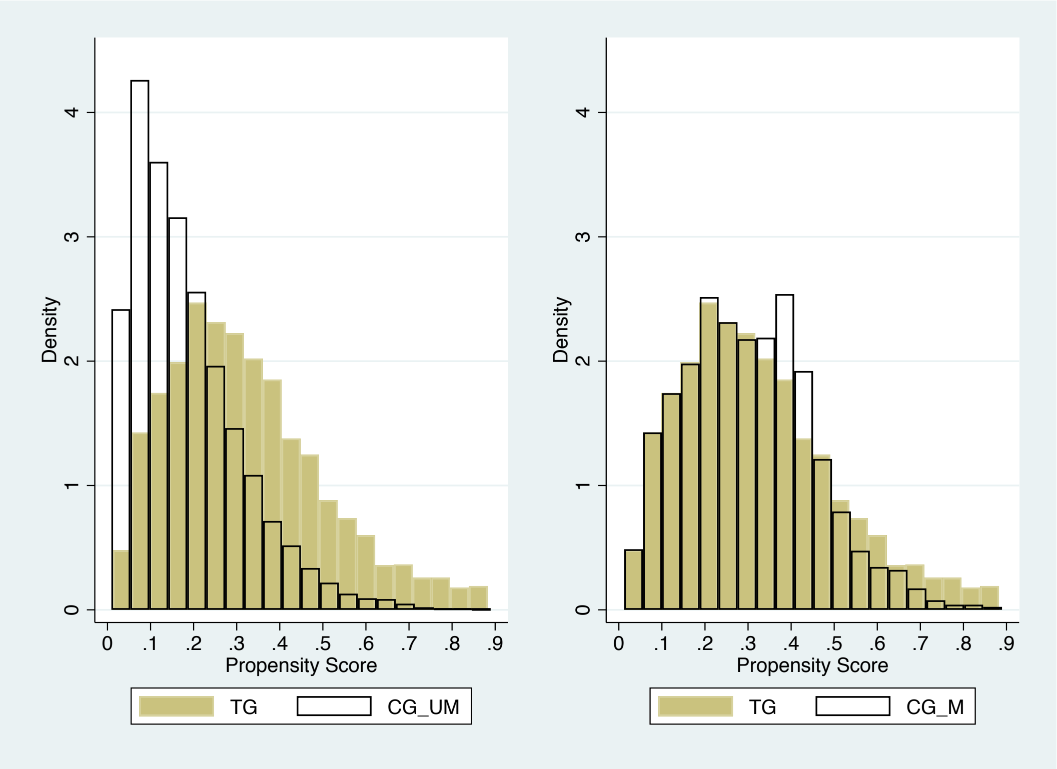


Figure 8: Distribution of Propensity Score Before and After Matching

Notes: TG-Treatment Group, CG_UM-Control Group *before* matching, CG_M-Control Group *after* matching

.