Does Clickbait Actually Attract More Clicks? Three Clickbait Studies You Must Read

Maria D. Molina
Department of Advertising and Public Relations
Michigan State University
molinad2@msu.edu

S. Shyam Sundar
Media Effects Research Laboratory
The Pennsylvania State University
sss12@psu.edu

Md Main Uddin Rony
College of Information Studies
University of Maryland
mrony@umd.edu

Naeemul Hassan
Philip Merrill College of Journalism
College of Information Studies
University of Maryland
nhassan@umd.edu

Thai Le
College of Information Sciences and Technology
The Pennsylvania State University
tql3@psu.edu

Dongwon Lee
College of Information Sciences and Technology
The Pennsylvania State University
dongwon@psu.edu

ABSTRACT

Studies show that users do not reliably click more often on headlines classified as clickbait by automated classifiers. Is this because the linguistic criteria (e.g., use of lists or questions) emphasized by the classifiers are not psychologically relevant in attracting interest, or because their classifications are confounded by other unknown factors associated with assumptions of the classifiers? We address these possibilities with three studies—a quasi-experiment using headlines classified as clickbait by three machine-learning models (Study 1), a controlled experiment varying the headline of an identical news story to contain only one clickbait characteristic (Study 2), and a computational analysis of four classifiers using real-world sharing data (Study 3). Studies 1 and 2 revealed that clickbait did not generate more curiosity than non-clickbait. Study 3 revealed that while some headlines generate more engagement, the detectors agreed on a classification only 47% of the time, raising fundamental questions about their validity.

CCS CONCEPTS

• **Human-centered computing** → Interaction design; Empirical studies in interaction design.

KEYWORDS

clickbait, engagement, content perception, machine learning

ACM Reference Format:
Maria D. Molina, S. Shyam Sundar, Md Main Uddin Rony, Naeemul Hassan, Thai Le, and Dongwon Lee. 2021. Does Clickbait Actually Attract More Clicks? Three Clickbait Studies You Must Read. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 08–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3411764.3445753

CHI '21, May 08–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445753

1 INTRODUCTION

Click-through is the currency of the modern Internet, with content creators striving to garner clicks by baiting users. As a result, we are bombarded daily with a plethora of clickbait, or headlines designed to persuade users to click on them by evoking curiosity and intrigue [1, 2, 6, 36, 40]. Clickbait media used to be associated with low quality and often misleading articles from unreliable sources (e.g., conspiracy, junk science, and satire sites) [7, 28, 36], but increasingly, respectable news outlets of established mainstream media have resorted to these strategies as well in order to gain user attention in a crowded information space [36, 38]. Because clickbait is often associated with scams and misinformation, many argue that articles utilizing clickbait should be detected and demoted or downranked in news aggregators [1], with extensive efforts in the academic community devoted to building algorithms for automatic detection of clickbait. This is exemplified by the 2017 Clickbait Challenge where teams of scholars worked on clickbait detection solutions to automatically detect this potentially malicious content [34]. But, are clickbait headlines clickbaity? Do users actually click more on clickbait? Literature across different fields show mixed results. On the one hand, research reveals a clear user preference toward clickbait content [36, 44]. However, other research reveals that this is not always the case [29, 30, 37]. There are a few possible reasons for the conflicting findings. First, it is possible that the differences derive from the distinct operationalizations of clickbait, with each study utilizing different characteristics of clickbait to study its effects. For example, two studies [29, 37] tested two styles or characteristics of clickbait—forward-referencing (akin to the "wh" characteristic in the present study, (i.e., who, what, where, when, why) and questions (i.e., headlines that ask a question), and compared user engagement and perceptions between these two styles and non-clickbait headlines. In Scacco and Muddiman [37], users engaged more with non-clickbait compared to either of the two clickbait styles. In Molyneux and Coddington [29], the question headline was perceived as lower quality than the non-clickbait headline, although the effects were small. On the other hand, Chakraborty [6] and Rony and colleagues [36] employed several characteristics of clickbait, including questions like in Scacco and Muddiman [37] and Molyneux and Coddington [29], but also other characteristics such as listicles, demonstrative adjectives and hyperboles. Likewise, Venneti and Alam [44] utilized several linguistic characteristics in their study, such as the use of exclamation and question marks, length and structure of the headline, as well as entities of the headline such as important events and figures. Contrary to Scacco and Muddiman [37] and Molyneux and Coddington [29], Rony and colleagues [36] and Venneti and Alam [44] found that, on average, clickbait received more engagement than non-clickbait. The difference in the characteristics used to define clickbait among these studies suggests that the conflicting findings might be because some characteristics of clickbait might be more successful than others at generating engagement. Initial evidence of this possibility is revealed in Lockwood's study [24]. The author's analysis of headlines of academic articles revealed that while positive framing and arousing framing increased attention to an academic article, framing the title as a question made no difference, and utilizing wordplay negatively influenced attention.

A second possibility for the mixed results of the effects of clickbait found in the literature could be the different methodological approaches to study clickbait engagement. Notably, past studies conducted using experimental designs and content analysis [29, 37] find that non-clickbait is more engaging than clickbait. Studies using automatic detection [36, 44] reveal the opposite pattern. While it is possible that the null results from experimental studies is due to the low external validity of the method, which does not allow participants to interact with headlines as they normally would during their regular browsing, it is also possible that the engagement captured through computational methods do not represent engagement with clickbait *per se*, but third variables embedded in the assumptions of each clickbait detector. For example, while some computational models are trained with data labeled as clickbait (or non-clickbait) by human coders (see [6] for example), other models utilize source assumptions for labeling of data (see [26] for example). This represents a fundamental difference in the conceptual definition and understanding of clickbait. Likewise, classification systems to detect clickbait are built using different machine learning approaches and techniques. For example, while some detectors are built using traditional machine learning (Naïve Bayes) [13], others utilize more complex deep learning models [36]. The distinct assumptions of classification systems are reflected in the manner in which their training data are coded (manual coding compared to source assumption), the type of machine learning model employed, and so on, thereby resulting in systems that, although accurate, classify clickbait using their unique idiosyncratic criteria. Such diversity in assumptions and approaches used by automated clickbait detectors raise important questions about the validity of such systems.

In this paper, we explore whether clickbait is indeed "clickbaity" by analyzing whether some characteristics of clickbait are more engaging than others and whether the differences in engagement between clickbait and non-clickbait derived from computational models represent different assumptions of the detectors other than "clickbaitiness" of the headlines *per se*. For this purpose, we conducted two experimental studies involving human subjects and a computational study of scraped clickbait headlines that compared four clickbait classifiers using a testing sample of real-world share data. In the following section, we explain why certain characteristics of headlines might be more engaging than others and why differences in the basic assumptions of computational models might be problematic for its classification.

2 COGNITIVE PROCESSING OF HEADLINES

The headline serves an important role in a news story because it helps orient readers to the information, summarizing its key ideas, and serving as an attention grabber [2, 12]. As Dor [12] explains, a headline is "a communicative device whose function is to produce the optimal level of affinity between the content of the story and the reader's context of interpretation, in order to render the story optimally relevant for the reader" (p. 720). Rooted in traditional print journalism as essentially a tool for improving usability of newspapers, the headline has become a critically important feature of digital media because a click on it is directly related to revenue. While news organizations acknowledge the importance of a click, they work to accomplish that goal without compromising their journalistic standards [40]. This is not the case for other types of content that circulate online. Many unreliable websites (e.g., junk science and conspiracy theories) utilize strategies such as clickbait to lure audiences into reading their content, but end up not meeting users' expectations once they click on the headline [6].

But, how do clickbaits gain user attention? The way in which headlines are written or framed has powerful effects on how the story is perceived. By framing the headline in a particular manner, the author makes some aspects of the text more salient compared to others, in turn promoting a particular definition, interpretation, or proposition [16]. Making particular pieces of information more noticeable and meaningful enhances the likelihood that the reader will perceive them, process them, store them in memory, and engage with them. For example, in Tannenbaum [41], the author presented three different groups of participants with the same article varying only in how the headline of a trial story was framed—guilty, innocent, or neutral. Participants who were assigned to the guilty condition were more likely to identify the person of the news article as guilty, while those in the innocent condition were more likely to think the person was innocent. Similarly, in Ecker and colleagues [15], participants exposed to misleading headlines in commentary articles were less likely to engage in inferential reasoning, compared to participants who were presented with headlines congruent with the article. The different effects in user perception and processing of information occurs because the headline is the first item to be encoded into memory and it helps readers assign relevance. The headline provides signals to the reader such that "the observer allocates cognitive resources toward certain environmental features because the communicator deems that information as more relevant to the observer than other information" [27] (p. 131). As such, users will read the story with a particular schema derived from the headline [4, 11, 15]. This schema, in turn, facilitates retrieval of information related to the headline, thus biasing the manner in which the story is processed by readers.

2.1 Effects of Headlines in User Engagement with Content

Not only do headlines influence how readers interpret the text, but even preceding that, it helps users decide if they should click and read the story in the first place. Even though clicking behavior is often associated with interest, there are several additional predictors of this type of engagement, including cognitive, affective, and pragmatic reasons [22]. For example, Tenenboim and Cohen [43] reveal that while users click more often on sensationalist content, they comment more on public-affairs content. This is because sensationalist content arouses curiosity and clicking on it enables self-experience. On the other hand, commenting behavior helps construct group identity and allows for self-expression. This example reveals that even though users might be equally interested in sensationalist and public-affairs content, they express their interest through different engagement actions available on social media. Similarly, Kormelink and Meijer [22] found that users will sometimes find an article interesting, yet will not click on it. This occurs when the headline is informationally complete (the user feels there is nothing additional to learn from reading the story) or when there is an associative gap (the headline does not tell enough for the user to want to click on it). In other words, for users to be persuaded into clicking, there should be a perfect balance between providing enough information to raise curiosity and leaving the user wanting to know more about the topic.

While clicking is an important metric directly associated with advertisement revenue, there are other important social media metrics to consider. One of them is the number of shares an article receives. The number of shares is an important metric for content creators because it reflects a mechanism to increase future readership or clicks. Simply put, sharing contributes to the virality of content, in turn increasing the number of possible clicks. But, what motivates users to share one piece of online content over another? Reasons include information utility, opinion leadership, emotional impact, relevance, entertainment, and social cohesion [3, 14, 33, 42]. Importantly, a large percentage of articles in social media are shared without being clicked upon, which means the users are persuaded to share by the headline alone. A large-scale Twitter analysis [17] revealed that nearly 60% of the shared URLs are never clicked upon. This means that the headline of the article on its own is sufficient to trigger the information utility, emotional impact, and other motivations needed for a user to share an article.

Content creators are well aware of the importance of headlines in generating content engagement. As such, they rely on different strategies to persuade users into clicking. One strategy is the use of linguistic properties to persuade users by generating a "curiosity gap" [6, 25]. The next section will expound on the psychology behind this strategy—also known as clickbait.

2.2 The Psychology of Clickbait

As a tool of persuasion, clickbait employs linguistic strategies that take advantage of the curiosity gap [6, 25], where the headline generates "enough curiosity among the readers such that they become compelled to click on the link to fill the knowledge gap" [6] (p. 1). This cognitive phenomenon, known as "information gap" [25], pertains to the curiosity triggered when a user's informational reference point is elevated beyond the users' current understanding or knowledge of a particular topic. When the user is alerted about such a gap, s/he will do the needful to close that gap, which in the case of clickbait headlines means reading the article to satiate the need to know. Importantly, however, attaining curiosity is a difficult task. For example, research reveals that clickbait headlines might raise curiosity or annoyance [2]. However, they are more likely to raise curiosity when they are perceived as creative, and thus are often preferred over merely informative headlines [19]. Ecker and colleagues [15] found that mismatched headlines influence memory and comprehension, but only to the extent that users do not perceive that they are being deceived. Loewenstein's [25] "information gap" hypothesis, in fact, explains several precursors to achieve the curiosity gap, namely awareness of the information gap and previous knowledge about the topic. The author states that curiosity will be greater when information is perceived as likely to close that knowledge gap and when the piece is perceived as providing insight (or a quick solution) rather than incremental solutions. Given the several factors needed to arouse curiosity, it is not surprising that we see mixed findings regarding the effects of clickbait headlines on actual user engagement [5, 36, 37, 44]. This raises the question: What will it take for a headline to accomplish the ideal level of curiosity to persuade users to click on it or share it?

Industry and academia have explored this question and agree on several stylistic features that induce greater engagement with content [5, 36]. Seven characteristics are recurrent: questions, lists, wh words (why, where, when, what), demonstrative adjectives (e.g., here, this, these), positive superlatives, (e.g., best, bigger), negative superlatives (e.g., worst, ever), and modals (e.g., should, would) (see Table 1 for definitions). Nonetheless, despite agreement that these stylistic features are common among clickbait headlines, studies have found conflicting results. One possibility for these mixed results is that some characteristics are more successful than others. In this paper, we explore which of these seven characteristics or stylistic features of clickbait are better at generating engagement, if indeed they are more effective than non-clickbait headline, and what are the psychological mechanisms that lead to higher engagement. More formally, we propose:

> **RQ1:** What is the relationship between the seven characteristics of clickbait and a) user clicking behavior (read more) and b) user sharing behavior?
>
> **RQ2:** What is the relationship between the seven characteristics of clickbait and a) perceived deception, b) perceived entertainment, c) perceived credibility, and c) curiosity arousal?

3 ASSUMPTIONS OF CLICKBAIT DETECTORS AND EFFECTS ON CLASSIFICATION

As explained in the previous section, to be considered clickbait, a headline should evoke interest while also leaving the user wanting to know more about the topic. Identifying which headlines meet this criterion is challenging, especially for a computational model because the concept must be defined with sufficient level of concreteness with features that represent "clickbaitiness" for machine distinction. As such, there are important decisions and assumptions that scholars must make when building computational models for the detection of clickbait. The first decision to make is what data to use for training the algorithm and how to label it. While some scholars opt for human labeling of data by asking annotators to identify if a headline is clickbait or not (e.g., [6]), others are based on weak supervision techniques such as assuming that certain sources

Table 1: Definition of Clickbait characteristics

Clickbait Characteristic	Definition
Questions	An interrogative or inquiry left open-ended and assumed to be answered within the associated article.
List	A statement in list format based on a particular theme. The expectation is that the reader will encounter a series of facts, examples, or tips about that theme upon reading the article.
Wh Words	Function words such as what, which, when, where, who, whom, whose, why, whether and how. Different than question headlines, this clickbait characteristic does not ask an actual question or inquiry when utilizing such function words.
Demonstrative Adjectives	Demonstrative adjectives (e.g. this or that) serve to indicate an entity being referred to and help to distinguish that entity from other entities.
Positive Superlatives	When several entities are compared, the positive superlative refers to the entity that is at the highest limit of the group in a particular characteristic (e.g. best, closest).
Negative Superlatives	When several entities are compared, the negative superlative refers to the entity that is at the lowest limit of the group in a particular characteristic (e.g. worst, least).
Modals	An auxiliary verb that expresses possibility, suggestion, or obligation (e.g. could, must).

are more likely to produce clickbait headlines (e.g., [13, 26]) and thereby labeling all headlines from a given source as clickbaits. Still others (e.g., [23]) use machine learning to generate synthetic clickbait headlines. These differences in approaches to assembling training data represent fundamental differences in the conceptual definition of clickbait. While human coding represents independent coders' understanding of what clickbait is (or what criteria were given by researchers for labeling data), source assumptions assume that reliable news organizations (e.g., New York Times) do not use clickbait strategies, deliberately overlooking the fact that this is not always the case [36].

Another important decision to make when building automated clickbait detectors is the type of machine learning model to use. For example, some detectors are based on traditional machine learning such as Naïve Bayes or Support Vector Machine (e.g. [13]), while others employ a more complex deep learning model (e.g. [36]). Each model has unique assumptions and characteristics that can affect the classification decision. Classical Naïve Bayes with bag-of-words features, for example, does not fully take into account the sequential dependency between words. This means that each feature or word is independent, and there is no intermediate representation between word and output. Thus, for Naïve Bayes models with bag-of-words (excluding punctuation marks), the headline "Want to Hear Biden and Harris Plans for Next Year?" is the same as "Biden and Harris Want to Hear Plans for Next Year." The model would not differentiate between the two, and classify both in the same way (i. e., both clickbait or both non-clickbait). On the other hand, more sophisticated deep learning models learn representations or features of the input texts and use these features for prediction, and unlike in Naïve Bayes (with bag-of-words), these features can be interpreted as interactions among the words in the sentence. Additionally, deep learning models can learn shortcuts that might generalize to some unseen text but are sometimes not very meaningful [18]. Thus, for deep learning models, the aforementioned headlines might not be equivalent. In other words, the model could classify the first as clickbait and the second as non-clickbait (or the other way around). While a Naïve Bayes classifier can learn the dependency among words located near each other by using n-gram features with n>1 or by converting the headline using word2vec instead of bag-of-words, this still does not consider the relative position of a word or phrase in a sentence. A deep learning architecture such as recurrent neural network (RNN) [8] or BERT [10] takes into account both word dependency and its relative position in a sentence.

The different characteristics of clickbait detectors (e.g., how data were labeled for training: human coding vs. source assumption vs. machine generation) or the type of machine learning model employed (Naïve Bayes, Support Vector Machine, Deep Learning), represent another possible explanation for the mixed results we see in the literature, and may explain why computationally based studies reveal greater engagement with clickbait (vs. non-clickbait headlines), while experimental studies do not show this pattern. It is possible that the different assumptions of each detector might have resulted in systems that, although accurate, classify clickbait using their unique conceptualization and operationalization of clickbait. If this is the case, when comparing the classification of different clickbait detectors on the same headlines, the agreement between the detectors would be low. On the other hand, if indeed the classification systems are all classifying the same concept, the agreement among the classifiers should be high. We test this question further in this paper. More formally:

> **RQ3:** What is the relationship between clickbait detectors varying in the labeling of data used for training (human annotated data vs. weak supervision with source assumptions) and type of model (traditional machine learning vs. deep learning) and their level of agreement with headline classification?

In summary, we aim to investigate if clickbait headlines are actually more engaging than non-clickbait headlines by testing possible reasons behind the mixed findings in the literature. Specifically, we test two possibilities: 1) that some clickbait characteristics generate more curiosity than others and thus differential engagement levels and 2) that differences in engagement between clickbait and non-clickbait derived from computational models represent other variables such as topic distinctions and assumptions of the system rather than "clickbaitiness." We explore these questions through two experiments and a computational analysis of real-world sharing of news headlines. In the first experiment, we scraped a series of headlines from reliable sources (the top circulated print media and most watched broadcast media according to Nielsen rating) and unreliable online sources (junk science, conspiracy, and satire sites) [36] and passed them through three automated clickbait detectors–two deep learning models and a traditional machine learning model (Naïve Bayes). Then, we selected headlines classified as clickbait by all three detectors and that possessed only one of the seven clickbait characteristics (or none). We presented these headlines to participants and asked their likelihood of sharing them with their network and/or clicking the headline to read the article further. In Study 2, we conducted a similar experiment, but to control for potential content effects, we randomly assigned participants to one of eight headlines for the same news story, varying only the clickbait (or not clickbait) characteristic utilized. Participants were then directed to read the article and asked their perceptions of both the headline and the associated article. Finally, in Study 3, we analyzed real-world sharing behavior of a series of headlines classified as clickbait by four classifiers (we used the same three classifiers as Study 1 and added one more classifier, resulting in a total of two deep learning models and two traditional machine learning models) and containing one of the seven clickbait characteristics. Methods and results of each study are explained in the following sections.

4 STUDY 1 METHOD

For study 1, we conducted an 8 (Characteristic: 7 Clickbait Characteristics + 1 Non-Clickbait) x 2 (Type of Content: Political vs. Non-Political) mixed method quasi-experiment. The 8 characteristics were a within-subject factor and the type of content was a between-subject factor. We decided on a within-subject factor for the clickbait characteristic because it automatically controls for individual differences among participants and represents a stronger test of user preference. Furthermore, we included political and non-political headlines as part of our study design because research suggests a difference in engagement between these two types of content, with non-political headlines receiving higher engagement than political headlines [22]. Including both in our manipulation allows us to account for the possibility that clickbait might be more successful for non-political content compared to political content. We gathered headlines for this study by scraping them from reliable and unreliable online sources, as defined by Rony and colleagues [36], and coded them computationally using their model to determine the clickbait characteristic they possess. Then, we passed the headlines through three different high-accuracy clickbait detectors that identified whether a headline is clickbait or not. We passed the headlines through three classifiers to increase the robustness of our study. Classifier 1 (93% accuracy) is a deep learning model trained on 32,000 headlines derived from news organizations and coded by three volunteer coders[1] [36]. Classifier 2 (93% accuracy) is a traditional machine learning algorithm (Naïve Bayes), with headlines labeled based on source assumptions[2] [13]. Classifier 3 (90% accuracy) is also a deep learning model, but it is based on a 12,000-headline dataset labeled based on credibility of the source[3] [26]. As a final step, we chose two headlines per clickbait characteristic for stimulus sampling purposes. To be selected, a headline had to 1) be identified as clickbait by the three detectors and 2) use only one of the clickbait characteristics (otherwise we manually modified) (See Table 2 for final list of headlines).

4.1 Participants

Participants for this study were recruited from Amazon Mechanical Turk (N=150). The final sample consisted of 149, after deleting one for incomplete data. An *a priori* power analysis revealed that in order to detect a medium-size effect (.25), with an error of .05 and a power of .80, a sample of 114 participants was needed. Participants' age ranged from 20 to 75 (M= 38.31, SD= 12.84), and 50.3% self-identified as female. Participants were predominantly Caucasian (77.2%) and highly educated, with 39.3% having a bachelor's degree and 15.4% a master's degree or higher.

4.2 Procedure

After acknowledging consent, participants were told that we are conducting a study with the purpose of identifying people's preference for different types of headlines. During the study, they were presented with a series of headlines through an interaction (See Figure 1). For each headline, participants could click on "read more" if they would be likely to click on the headline and "share" if they would be likely to share it with their network of friends. They could select one of the options, both, or none. After reading the instructions, participants were provided with a test run so they could have a feel for the interaction before starting the actual study. Participants were randomly assigned to either the political or the non-political condition. Then, they received 8 headlines (one for each clickbait characteristic and non-clickbait) randomized in order of presentation. Each headline was presented as in Figure 1, allowing participants to interact with the headline as they would during their normal course of browsing. Upon interacting with each headline, participants were asked questions about their perceptions of the headline. After participants went through all 8 assigned headlines, they were asked demographic questions and questions about their political orientation.

4.3 Measures

4.3.1 "Read More" and "Sharing" Intention. To measure participants' likelihood of reading the article associated with the headline and sharing the headline, we created an interaction (See Figure 1) such that participants were provided with the headline and two buttons. Participants were instructed to click on "read more" if they would be likely to click on the headline and read the article associated with it (the article was however not displayed to participants). Similarly, participants could click on "share" if they would be likely to share the headline with their network of friends on social media. This interaction resulted in two variables, one indicating if the participant clicked on "read more" or not, coded as 1=clicked or 0=did not click, and the other indicating if the participant clicked on "share" or not, coded as 1= shared or 0 = did not share.

[1] https://github.com/bhargaviparanjape/clickbait/tree/master/dataset
[2] https://github.com/peterldowns/clickbait-classifier
[3] https://github.com/saurabhmathur96/clickbait-detector#data

Table 2: Stimulus sampling headlines for Study 1

	Political	Non-Political
Non-Clickbait	Congress Actually Passes Law To Protect U.S. Citizens, Cops Must Report Killings	Colorado Says No To Marijuana Smoking In Hospital Delivery Rooms
	Associated Press Reveals Secret Deal Allowing Iran To Expand Nuke Program!	Carnival's Princess Cruise Lines Agree to Fine, Plea Guilty to Pollution Charges
Question	Want to Hear Gov. Tom Wolf Plans For Next Year? Tune In Tonight	Afraid of the Disneyland Legionnaire disease outbreak? Don't be fooled by Mickey Mouse science
	"Yes I Respect Porn Stars. Don't You Respect Porn Stars?" – Giulliani Grills Melania Trump Spox	Did Zayn Malik Delete His Instagram Page to Avoid Rumors of Cheating?
List	10 Reasons National Review Doesn't Want Trump In The White House	6 Ways Emma Watson Is Adult Hermione Granger
	The 10 ways Obama actively sought to destroy America	Here Are 25 Commonly Accepted 'Facts' That Are Total Bullsh*t
Wh	What Happened While The Media Distracted Us With Bathroom Issues	Why Taylor Swift Only Wears Crop Tops and Skirts
	Video Meant to Make People Feel Sorry For The 'Refugee Crisis' Proves Why The Refugees Must Go Back	How Female Execs Are Striking Back At Pharma For Using Women as "Eye Candy"
Demonstrative Adj.	This Chart Captures Every Sexist Slur Trump Supporters Tweeted to Megyn Kelly	This Dark Side of The Internet Is Costing Young People Their Jobs And Social Lives
	Judge Jeanine Completely Skewers Obama and Gruber in this EPIC Rant	Drake Takes His Lint Roller Everywhere, And These GIFs Prove It
Positive Superlative	The Best Takedown Of Democrats Admin's Gun Control Arguments	Best BGT Performance Ever: Watch Boys Get Rousing Ovation for Their Anti-Bullying Song
	Donald Trump Being Attacked By A Bald Eagle Is the Best Thing Ever	This Daylight Savings Trailer Is The Best Thing You'll See All Day
Negative Superlative	Trump Says Obama's Presidency is 'Worst Thing To Ever Happen to Israel.'	This Clip of Miley, Taylor & Katy Freaking The F**k Out Over A Bee Never Gets Old
	Most Unbelievably Scandalous Video Hillary Wishes Never Saw Daylight	Awards Shows Bring Out the Worst in Us
Modal	Voices: We Won't Be OK While There Are Still Two Baltimores	Tove Lo's Fierce New Tattoo Will Make Cat Lovers Jealous
	We Must Change Park Ranger Uniforms to Make Illegal Aliens Feel Better	Khloe Kardashian Looks Super Skinny After Losing 40 Pounds! You Won't Believe the Before and After Pics

4.3.2 Headline Perception. To measure headline perception, after interacting with each headline, we asked participants to indicate how well each word from a list of 23 adjectives describes the headline they just read. Questions were asked on a 1-7 scale. Items were adapted from Sundar's [39] news content perception scale, Ormond et al.'s [32] and Dark and Ritchie's [9] deception scales, and Naylor's curiosity scale [31]. The authors additionally added items to reflect the possible entertainment value of headlines. An exploratory factor analysis using Oblimin rotation revealed four factors: credible (e.g., accurate, believable, well-written, persuasive; M= 3.59, SD= 1.51, α = .91), deceptive (e.g., dishonest, deceptive, fake, tricky; M= 3.39, SD= 1.41, α = .84), entertaining (e.g., humorous, enjoyable, entertaining; M= 3.24, SD= 1.70, α = .87), and curiosity arousing (e.g., want to know more, boring (r), intriguing; M= 3.91, SD=1.78 α = .85).

4.3.3 Political Orientation. We measured participants' political orientation through four questions proposed by Janoff-Bulman and colleagues [20]. Items included, "Where would you place yourself on a scale from 1 (Strong Democrat) to 7 (Strong Republican)", and "Where would you place yourself on a scale from 1 (Very Liberal) to 7 (Very Conservative)" (M= 3.63, SD= 1.59, α = .85).

5 STUDY 1 RESULTS

First, we calculated the distribution of clicks on the "read more" and "share" buttons by characteristic to get a general idea of the relative success of each characteristic at engaging users. Table 3 reveals that participants clicked on "read more" and "share" more often for non-clickbait headlines (compared to the seven clickbait characteristics). To test if these differences were statistically significant, we ran two logistic regressions, one using participants' response on "read more" button as the dependent variable and the other using participants' response on "share" button as the dependent variable. For both regressions, participants' political orientation was entered first as a control variable, followed by the type of content (political vs. non-political) and headline characteristic as independent variables. The final step included an interaction term between the type of content and headline characteristic.

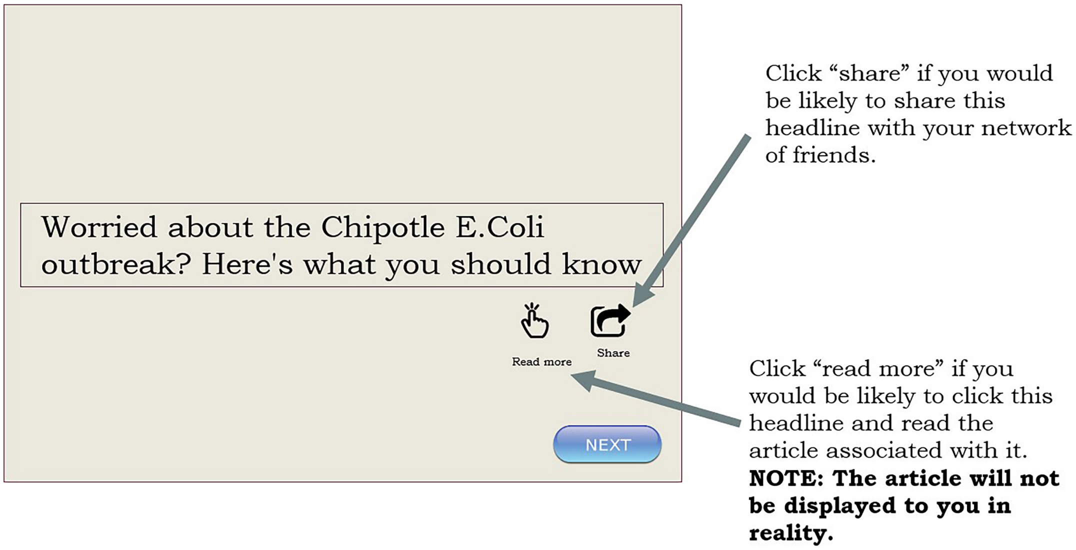

Figure 1: Training interaction provided to participants. As in the training, for the real study, participants could click "read more," "share," both options, or none of them, for each of the eight headlines presented.

Table 3: Number of clicks on "read more" and "share" by characteristic

	Read More	Share
Demonstrative Adjective	71	18
List	77	22
Modals	55	16
Negative Superlative	62	22
Positive Superlative	65	17
Question	59	13
Wh	74	18
Non-Clickbait	84	23
Total	547	149

When we entered "read more" as the dependent variable, data revealed that headline characteristic was a significant predictor of clicking behavior (Wald χ^2 = 18.06, p = .01). Post-hoc pairwise comparison based on odds ratio revealed that the odds of non-clickbait headlines receiving a click was 2.22 times greater than modal headlines (p <.001), 1.81 times greater than negative superlatives (p= .01),1.68 times greater than positive superlatives (p= .03), and 1.97 times greater than questions (p= .003). On the other hand, non-clickbait headlines were as successful as demonstrative adjectives, lists, and "wh" headlines. This means that non-clickbait headlines performed better than four of the seven clickbait characteristics, but as good as the other three. Furthermore, when comparing between clickbait characteristics, the odds of the "wh" headlines receiving a click was 1.70 times higher than modals (p= .03); the odds of listicles receiving a click was 1.83 times higher than modals (p= .01); the odds of listicles receiving a click was 1.63 times greater than questions (p= .04). Results revealed no significant main effect of type of content and no interaction effect.

When entering "share" as the dependent variable, results revealed no significant main effects and no significant interaction effect. This indicated that all headlines were as likely to be shared regardless of linguistic characteristic or type of content.

To test if headlines were perceived differently by participants as a function of the clickbait characteristics and type of content, we ran a series of 8 (7 Clickbait Characteristics + 1 Non-Clickbait) x 2 (Type of Content: Political vs. Non-Political) repeated-measures analyses of variance using a mixed model approach.

When entering curiosity as the dependent variable, results revealed a significant main effect of characteristic, F (7, 1029) = 5.84, p <.0001. Tukey HSD post-hoc comparisons (see Table 4) revealed that non-clickbait headlines elicited more curiosity than all clickbait headlines (the difference is statistically significant for all comparisons except for the comparison with demonstrative adjectives and lists). This finding runs counter to the belief that clickbait headlines induce more curiosity, and thus receive more user engagement.

The main effect should be interpreted in light of the interaction effect between characteristic and type of content, F (7, 1029) = 3.17, p =.003. Patterns of the interaction (See Figure 2) reveal that while non-political content written using "wh" words and modals were perceived as less curiosity-arousing compared to other clickbait characteristics and non-clickbait, when political content is written using these characteristics, the headline arouses a higher level of curiosity, reaching similar numbers as those of non-clickbait headlines.

With credibility as the dependent variable, data revealed a significant effect of characteristic, F (7, 1029) = 7.47, p <.0001. Post-hoc comparisons using Tukey HSD difference test are reported in Table 4 and reveal that overall non-clickbait headlines

Table 4: Perceptual differences as a function of characteristics.

	Curiosity	Credibility	Deception	Entertaining
Demonstrative Adjective	M = 4.12$_{ab}$ SE=0.14	M = 3.74 $_{ab}$ SE=0.12	M = 3.16$_{bc}$ SE=0.12	M = 3.55 $_{ab}$ SE=0.13
List	M = 4.01$_{abc}$ SE=0.14	M = 3.52 $_{bcd}$ SE=0.12	M = 3.62$_{a}$ SE=0.12	M = 3.62$_{a}$ SE=0.13
Modals	M = 3.51$_{c}$ SE=0.14	M = 3.25$_{d}$ SE=0.12	M = 3.05$_{c}$ SE=0.12	M = 3.38 $_{ab}$ SE=0.13
Negative Superlative	M = 3.72$_{bc}$ SE=0.14	M = 3.34$_{cd}$ SE=0.12	M = 3.50$_{ab}$ SE=0.12	M = 3.62 $_{a}$ SE=0.13
Positive Superlative	M = 3.86$_{bc}$ SE=0.14	M = 3.55 $_{bcd}$ SE=0.12	M = 3.11$_{c}$ SE=0.12	M = 3.55 $_{ab}$ SE=0.13
Question	M = 3.71$_{bc}$ SE=0.14	M = 3.66 $_{abc}$ SE=0.12	M = 3.15$_{bc}$ SE=0.12	M = 3.22$_{b}$ SE=0.13
Wh	M = 3.92$_{bc}$ SE=0.14	M = 3.60$_{bcd}$ SE=0.12	M = 3.32$_{abc}$ SE=0.12	M = 3.43 $_{ab}$ SE=0.13
Non-Clickbait	M = 4.48$_{a}$ SE=0.14	M = 4.02$_{a}$ SE=0.12	M = 3.03 SE=0.12 $_{c}$	M = 3.27 $_{ab}$ SE=0.13

Note: Vertical means with no lower-case subscript in common differ at p< .05 using Tukey HSD post-hoc comparisons.

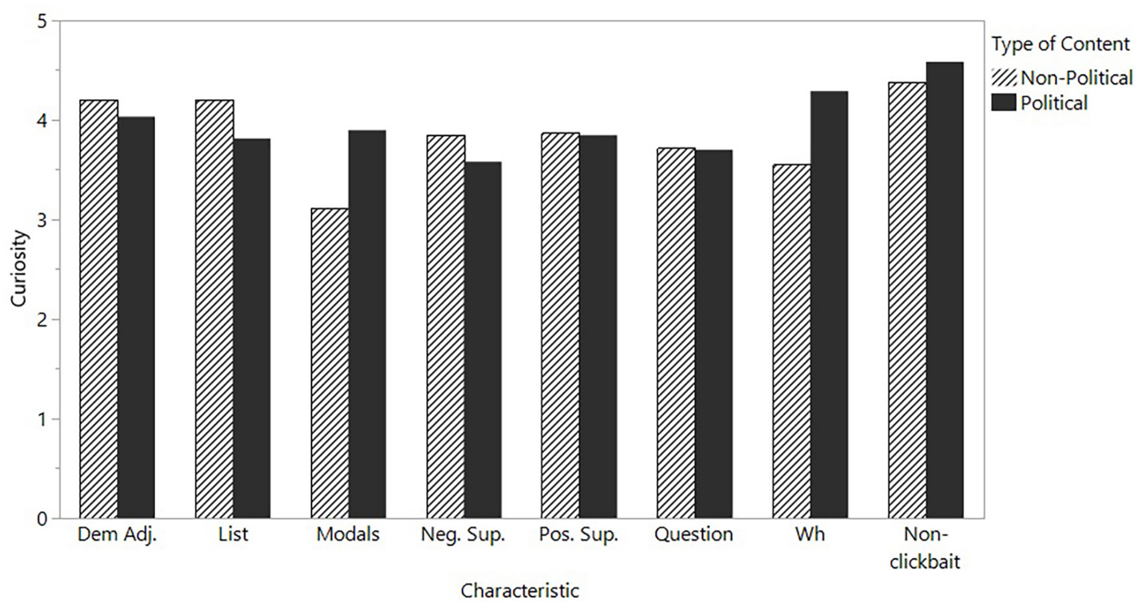

Figure 2: Aroused curiosity as a function of characteristic and type of content

were perceived as more credible than clickbait (the difference was statistically significant for all comparisons except for the comparison with demonstrative adjectives and questions). This effect should be interpreted in light of a significant interaction effect (See Figure 3) between characteristic and type of content, F (7, 1029) = 3.75, p <.001, revealing that for all characteristics except for modals, questions and wh, non-political content was perceived as more credible than political content. Importantly, looking at the interaction effect, non-clickbait content (both political and non-political) were perceived among the most credible headlines.

When entering perceived deception as the dependent variable, data revealed a significant main effect for type of content such that political headlines were perceived as more deceitful (M= 3.62, SD = 0.12) than non-political headlines (M= 2.87, SD = 0.11), F (1, 147) = 19.52, p <.0001. Results also revealed a significant effect of characteristic F (7, 1029) = 6.45, p <.0001. Post-hoc pairwise comparisons using Tukey HSD test (Table 4) indicate that lists, negative superlatives, and "wh" headlines were perceived as more deceitful compared to the other characteristics. Non-clickbait was perceived as the least deceitful, significantly lower than lists and negative

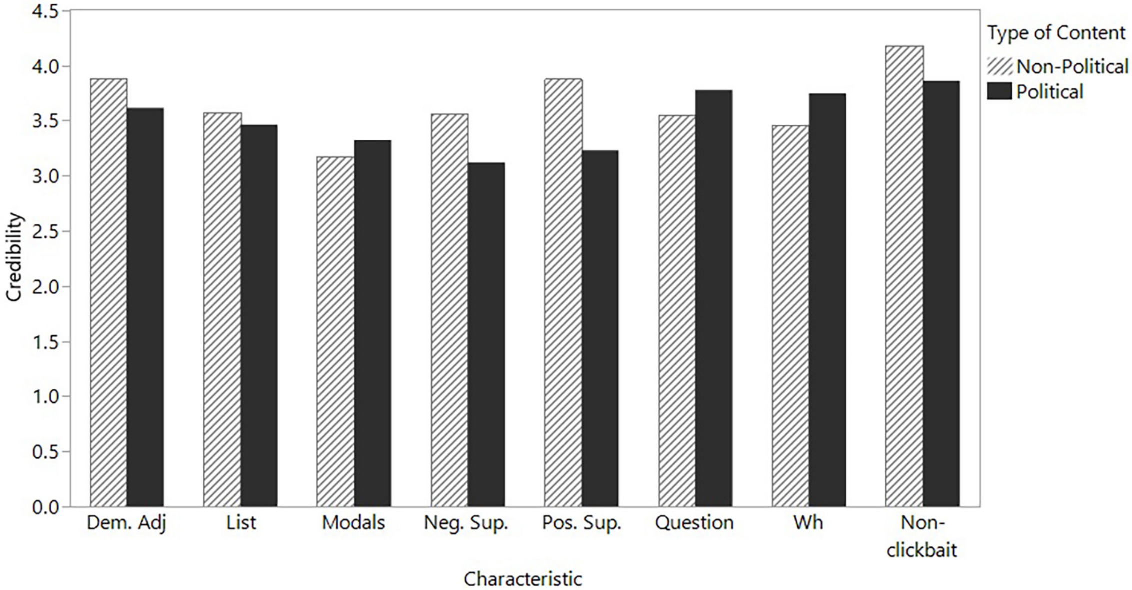

Figure 3: Perceived credibility as a function of characteristic and type of content

superlatives. Nonetheless, the main effects should be interpreted based on a significant interaction effect between headline characteristic and type of content, F (7, 1029) = 9.76, p <.0001. The trend of the interaction (See Figure 4) reveals that political content tends to be perceived as more deceitful compared to non-political content, except for headlines using questions. In this case, not only is there no statistically significant difference between political and non-political headlines, but political headlines with questions were perceived as less deceptive than political headlines using positive superlatives, lists, and negative superlatives.

Finally, when we entered entertaining as the dependent variable, data revealed a significant main effect of characteristic, F (7, 1029) = 2.99, p = .004. Pairwise post-hoc comparisons using Tukey HSD (See Table 4) revealed that all headlines were perceived as equally entertaining, except for question-based clickbait that ranked least entertaining compared to lists and negative superlatives. Data also showed an interaction effect between characteristic and type of content, F (7, 1029) = 3.76, p < .001 (see Figure 5), such that lists and negative superlatives were perceived as more entertaining in non-political headlines, compared to political headlines. However, political headlines using demonstrative adjectives were perceived as more entertaining than non-political headlines using demonstrative adjectives.

6 STUDY 1 DISCUSSION

In answering RQ1, results of Study 1 reveal that indeed some clickbait characteristics result in more clicks (to read more) than others. Specifically, headlines using "wh" word and using lists received more clicks than those using modals, and headlines using lists received more clicks than those using questions. Despite these differences among clickbait headlines, non-clickbait received more clicks overall than four of the seven clickbait characteristics and performed the same as the remaining three. Furthermore, there were no statistically significant differences in sharing behavior as a function of characteristics. These results indicate that despite the common understanding that clickbait headlines generate more clicks than non-clickbait, this is not the case. Users are as likely (and sometimes more likely) to engage with non-clickbait headlines. The overall preference for non-clickbait headlines can be explained by how users perceived them. For starters, headlines using demonstrative adjectives and lists aroused more curiosity than the other clickbait characteristics. However, non-clickbait was similarly arousing as these two characteristics, and was more arousing than the remining five clickbait characteristics. This runs counter to the proposition that clickbait headlines will be more arousing than non-clickbait headlines by inducing a curiosity gap [1, 6]. Likewise, non-clickbait headlines were perceived by participants as equally entertaining as clickbait headlines. As expected, non-clickbait was also perceived as less deceitful and more credible than clickbait headlines. Interestingly, demonstrative adjectives and questions were perceived as equally credible. It is also important to note the interaction effects between clickbait characteristics and type of content suggesting that some clickbait characteristics might be perceived differently depending on whether the content is political. Nonetheless, despite these slight variations, the interaction effects still revealed that non-clickbait induce about the same (if not more) curiosity than clickbait, are perceived as equally (if not more) credible and are perceived as equally (or less) deceitful.

Results of Study 1 provide initial evidence that clickbait headlines might not be as successful as we think. Nonetheless, there are important limitations of this study to consider. First, there are content characteristics that are not accounted for in this study. We utilized headlines scraped from online sources in the interest of

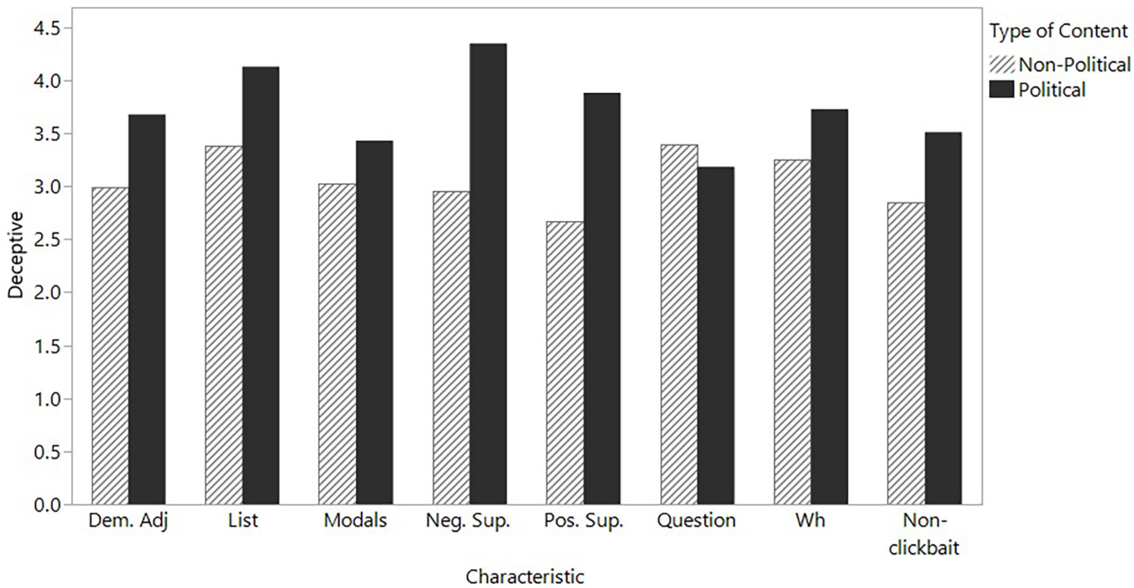

Figure 4: Perceived deceitfulness as a function of characteristic and type of content

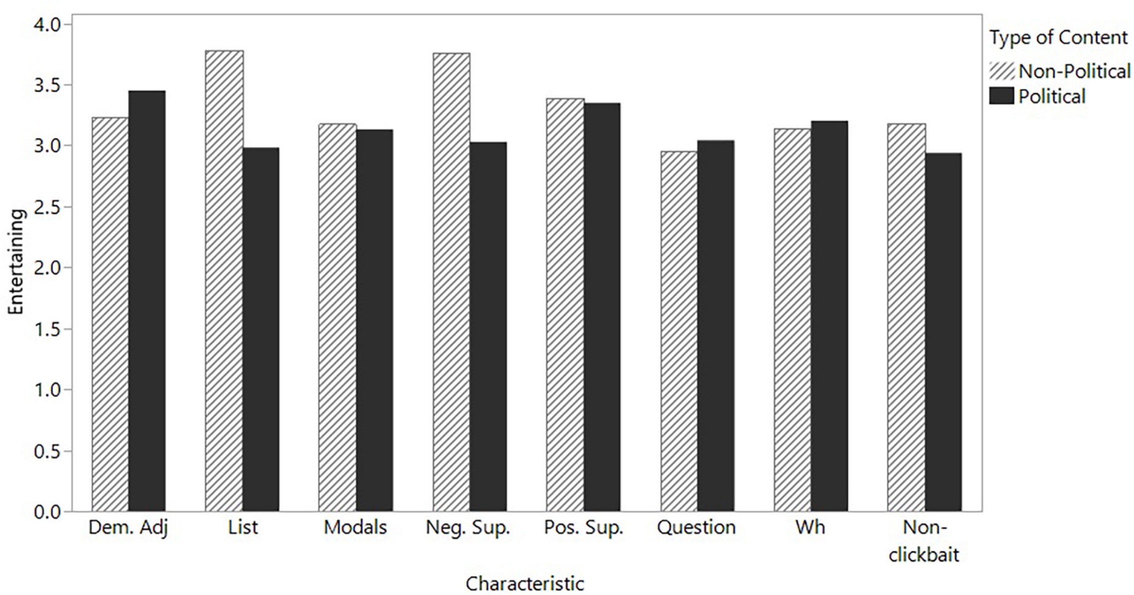

Figure 5: Perceived entertainment as a function of characteristic and type of content

ecological validity, but it resulted in headlines pertaining to different topics and associated with different articles (See Table 2), thus creating a content confound when comparing the various types of headlines. Secondly, we did not show the actual text after users clicked "read more," which might have reduced participants' desire to click. Furthermore, our sample consisted of M-Turkers, who might have higher levels of digital literacy or might be more motivated to systematically think about their engagement with headlines (at least in a study context) than the regular user. To account for such limitations, we conducted Study 2, which is described next.

7 STUDY 2 METHOD

For Study 2, we chose one of the political non-clickbait headlines utilized in Study 1 (Secret Deal Allowing Iran To Expand Nuke

Table 5: List of headlines for Study 2

Characteristic	Headline
Demonstrative Adjective	This Secret Deal Allows Iran to Make Nukes in Half the Time
List	Four Things to Know about Secret Nuke Deal with Iran
Modals	Secret Deal Expanding Iran's Nuke Program that you Should Know
Negative Superlative	The Worst Secret Deal: Allows Iran to Expand Nuke Program
Positive Superlative	Secret deal is the Best Thing that Happened to Iran Nuke Program
Question	Want to Know the Secret Behind Iran's Nuclear Deal?
"Wh"	What you did not Know About Iran's Secret Deal to Build Nukes
Non-clickbait	Secret Deal Allows Iran to Expand Nuke Program

Program) and systematically changed it to contain one of the seven clickbait characteristics. We chose a political headline because the use of clickbait in this area is of particular concern. Many argue that the use of clickbait media results in the "dumbing down of news" [22]. The clickbait headlines were written by a former journalist and resulted in a total of eight headlines (7 clickbait + 1 non-clickbait) (See Table 5 for exact headlines).

7.1 Participants

Participants for Study 2 (N=249) consisted of 89 students recruited from communication and information science and technology courses at two US universities (one located in the Northeast and another in the South), as well as 160 participants recruited from Amazon Mechanical Turk. An *a priori* power analysis revealed that to detect a medium-size effect (.25), with an error of .05 and a power of .80, a sample of 240 participants would be needed. We recruited students and M-Turk participants to account for the possibility that the superior engagement of non-clickbait headlines (vs. clickbait) in Study 1 could be due to the higher digital literacy of M-Turkers. Of our total sample, 61.8% self-identified as male, 36.9% as female, 1.2% other, and .1% did not report. Our sample was predominantly white (65.5%) and their ages ranged between 18 and 68 years (M= 29.89, SD= 11.42).

7.2 Procedure

Upon consenting to participate in this study, participants were directed to the same instructions and training interaction as in Study 1 (see Figure 1). They were then randomly assigned to one of the 8 conditions (7 clickbait headlines + 1 non-clickbait headline). Once they received their assigned headline, they could click "read more," and/or "share" as in Study 1. Upon completing the interaction, participants were redirected to a questionnaire asking about their perceptions of the headline. Participants who clicked on "read more" received a prompt saying that before we take them to the story, we would like to ask their quick impressions of the headline they just read, while those who did not click on "read more" received a prompt simply stating that we will now ask them about their quick impression of the headline. The different prompts were included to assure those that clicked "read more" that they will indeed see the story afterwards. The fact that participants received the actual text when clicking "read more" addresses the possibility that in Study 1 users did not click "read more" for the clickbait headlines because they knew that they would not get to read the story anyway. After completing the questionnaire eliciting their perceptions of the headline, all participants were taken to the story associated with the headline. The story remained constant across conditions, but the headline varied depending on the condition (see Figure 6). After reading the story, participants were directed to a questionnaire asking about their perceptions of the news story, their likelihood of sharing the story with their network of friends, and their elaboration of the content of the story. At the end of the questionnaire, they were asked about their demographics and political orientation.

7.3 Measures

7.3.1 Headline Reading and Sharing Intention. To measure participants' likelihood of reading the article associated with the headline and sharing the headline, we followed the same procedure as in Study 1 and created two variables, one indicating if the participant clicked on "read more" or not, coded as 1=clicked or 0=did not click, and the other indicating if the participant clicked on "share" or not, coded as 1= shared or 0 = did not share.

7.3.2 Headline Perceptions. To measure headline perceptions, we utilized a similar scale as in Study 1 and asked participants to indicate how well each word in a list of 24 adjectives reflected the headline they just read. The final scales were constructed based on an exploratory factor analysis using Oblimin Rotation. We constrained the EFA to four factors to imitate the scales utilized in Study 1 for comparability purposes. The final scales pertained to credibility (e.g.: accurate, informative, objective, M= 3.77, SD= 1.29, α = .90), deception (e.g.: dishonest, biased, fake, M = 3.81, SD = 1.29, α = .84), curiosity (e.g.: interesting, want to know more, aroused my curiosity, M= 4.54, SD= 1.45, α = .76), and entertaining (e.g.: enjoyable, entertaining, humorous, M= 2.89, SD= 1.39, α = .75).

7.3.3 Story Perceptions. Story perceptions were measured via the four subscales proposed by Sundar [39] including credibility (e.g.: objective, fair, unbiased, M= 4.15, SD= 1.30, α = .86), liking (e.g.: enjoyable, pleasing, interesting, M= 3.77, SD= 1.14, α = .71), quality (e.g.: clear, coherent, comprehensive, M= 4.69, SD= 1.14, α = .79), and representativeness (e.g.: important, relevant, timely, M= 4.76, SD= 1.29, α = .78). Questions were administered on a 1-7 scale.

7.3.4 News Sharing Intention. Participants' intention to share the news was measured with one item asking participants to rate on a 1-7 scale: How likely they would be to share this story with their network of friends and family (M= 3.13, SD= 2.01).

This Secret Deal Allows Iran to Make Nukes in Half the Time

VIENNA (AP) — A document obtained by The Associated Press shows that key restrictions on Iran's nuclear program will ease in slightly more than a decade, halving the time Tehran would need to build a bomb.

The document is the only secret text linked to last year's agreement between Iran and six foreign powers. It says that after a period between 11 to 13 years, Iran can replace its 5,060 inefficient centrifuges with up to 3,500 advanced machines.

Since those are five times as efficient, the time Iran would need to make a weapon would drop from a year to six months.

Iran says its enrichment is peaceful, but the program could be used for nuclear warheads. Two diplomats providing the information Monday demanded anonymity because they weren't authorized to do so.

The Worst Secret Deal: Allows Iran to Expand Nuke Program

VIENNA (AP) — A document obtained by The Associated Press shows that key restrictions on Iran's nuclear program will ease in slightly more than a decade, halving the time Tehran would need to build a bomb.

The document is the only secret text linked to last year's agreement between Iran and six foreign powers. It says that after a period between 11 to 13 years, Iran can replace its 5,060 inefficient centrifuges with up to 3,500 advanced machines.

Since those are five times as efficient, the time Iran would need to make a weapon would drop from a year to six months.

Iran says its enrichment is peaceful, but the program could be used for nuclear warheads. Two diplomats providing the information Monday demanded anonymity because they weren't authorized to do so.

Figure 6: Sample story stimulus: demonstrative adjective condition (left), negative superlative condition (right). The story remained constant, but the headline differed based on the condition to which participants were assigned (eight total).

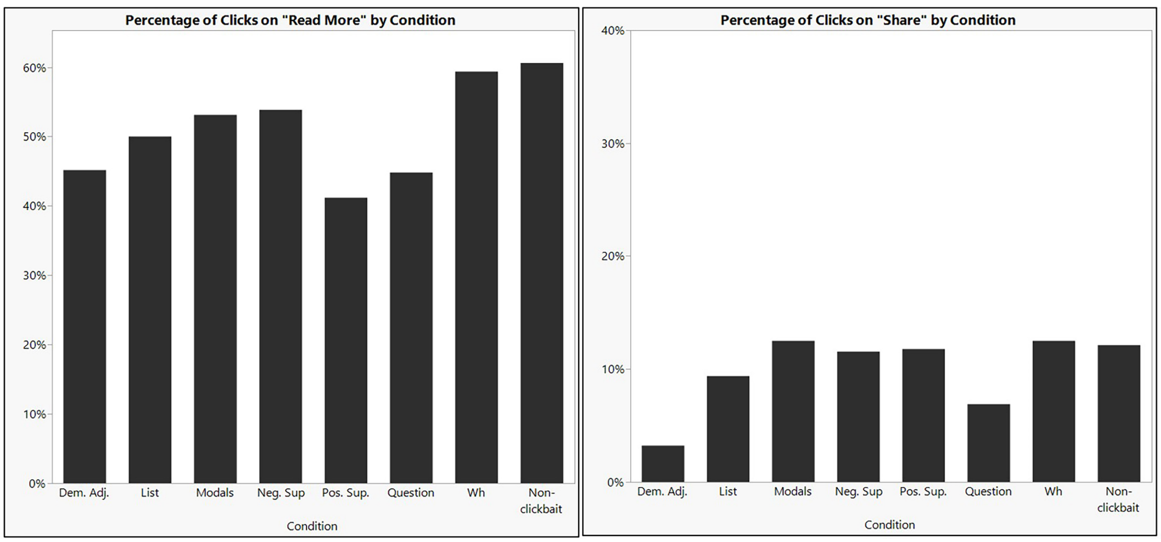

Figure 7: Percentage of participants who clicked on "Read more" (left) or "Share" (right) by condition.

7.3.5 News Story Elaboration. To assess the extent to which participants elaborated or systematically processed information from the article, we administered 12 items from a validated measure of message elaboration [35]. The scale asked participants the degree to which they engaged in a series of behaviors while reading the message, on a 1-7 scale, such as: "attempting to analyze the issues in the message, unconcerned with the ideas(r), and expending a great deal of cognitive effort" (M= 4.71, SD= 1.06, α = .86).

8 STUDY 2 RESULTS

To assess the effects of the headline characteristics on user engagement with the headline (clicking on "read more" and "share"), we first explored descriptive statistics. Figure 7 reveals that, overall, users clicked "read more" on non-clickbait at a higher rate, followed by "wh" words. When looking at sharing behavior, on the other hand, there is a relatively low number of shares overall. Questions and demonstrative-adjective headlines seem to be shared relatively less than the other characteristics.

To test if these differences in user engagement are statistically significant, we conducted two logistic regressions, one for users' click on "read more" and one for users' click on "share." We entered participant type (student or M-Turker), political orientation, gender, and age as control variables. Results revealed no significant effect of characteristic when entering "read more" as the dependent variable (Wald χ^2 = 4.42, p = .73), nor when entering "share" as the dependent variable (Wald χ^2 = 2.91, p = .89). This means that user engagement was not different across clickbait and non-clickbait headlines.

To assess if participants perceived the headlines any differently, we conducted a series of one-way ANOVAs for each dependent variable of interest. For all analyses, we entered participant type, political orientation, gender, and age as control variables. Results revealed no significant difference for any of the four dependent variables of interest: credibility (F (7, 236) = 1.37, p =.22), deceptive

(F (7, 236) = 1.14, p =.34), curiosity inducing (F (7, 236) = 1.22, p =.30), entertaining (F (7, 236) = 0.50, p =.84).

Following analyses of the headline, we examined if there were any differences in how participants perceived the news story, their intention to share the story, and their elaboration as a function of headline characteristic. For this purpose, we ran a series of one-way ANOVAs. Results revealed no statistically significant difference on perceived story credibility (F (7, 236) = 088, p =.52), representativeness (F (7, 236) = 0.83, p =.57), quality (F (7, 236) = 0.94, p =.48), liking (F (7, 236) = 1.16, p =.33), users' elaboration of the story (F (7, 235) = 1.66, p =.12), or likelihood of sharing the story (F (7, 236) = 0.24, p =.98).

9 STUDY 2 DISCUSSION

In Study 2, we controlled for the topic of the headlines and included students and M-Turkers in our sample to account for content confounds of Study 1, yet we found no significant differences between clickbait and non-clickbait headlines in terms of likelihood of clicking and sharing. If anything, users clicked more on non-clickbait headlines, although not by a statistically significant margin.

Analyses also revealed that participants perceived the headline the same regardless of its characteristics. Again, these results run counter to past research suggesting that headline style influences user perception of the headline and its associated content [15, 41]. While data suggest that the headline selected for Study 2 elicited rather high curiosity (M= 4.54), meaning that participants found the article to be interesting, we should note that this study used only one headline. More testing is required for enhancing the external validity of our findings.

Nonetheless, the null findings of Study 2 suggests three possibilities, 1) clickbait is not as clickbaity as we think, 2) the "clickbaitiness" of clickbait headlines is not solely determined by the characteristics suggested by industry [5] and identified by the clickbait detector in Rony et al. [36], but by other factors that help create a psychological information gap, or 3) the higher engagement of clickbait headlines (compared to non-clickbait) found in computational analysis [36, 44] might not represent engagement with clickbait *per se*, but third variables (or common-cause variables) attributable to the assumptions of each clickbait detector (e.g., labeling procedure for training data, type of machine learning model). If this is the case, then each detector is operating with a unique and distinct conceptualization of clickbait, which calls into question the validity of clickbait detectors. If the same concept is being captured by the different models, then they should have high agreement in classifying the same set of clickbait headlines. If they do not have high agreement, it would represent a validity issue of content classification systems for clickbait detection, with results being confounded with other variables such as topic distinction and system assumptions. We explore these issues in Study 3 through computational analysis of real-world number of shares of headlines scraped from reliable and unreliable sources (political and non-political). This time, we passed the headlines through four different clickbait classifiers (the same three as in Study 1 plus a new traditional machine-learning classifier) varying in their basic assumptions of clickbait and in the type of machine-learning utilized. We added a fourth classifier in Study 3 in order to have a fully crossed factorial design that can assess differences in engagement based on the characteristics of the classifiers: 2 (Type of Model: Conventional Machine Learning vs. Deep Learning) x 2 (Data: Annotated Data vs. Source Assumptions).

10 STUDY 3 METHOD

For Study 3, we scraped headlines from reliable and unreliable sources (political and non-political) and computationally identified the clickbait characteristics utilized by them (see supplemental material for complete list of headlines). Then, we passed the headlines through four different clickbait detectors. The first three detectors were the same as in Study 1. We added a fourth detector – a support vector machine learning model trained on manually annotated data (96% accuracy). This addition resulted in a 2 (Type of Model: Traditional Machine Learning vs. Deep Learning) x 2 (Data: Annotated Data vs. Weak Supervision with Source Assumptions) comparison allowing us to assess differences in engagement based on the characteristics of the classifiers. As a final step, we linked the scraped headlines to actual share data retrieved from sharedcount.com. Data from sharedcount.com includes total number of Facebook shares, Facebook comments, Facebook reactions, and number of pins. In total, we ended up with data for 371 headlines after deleting errors and headlines that used other characteristics aside from the seven analyzed in this paper.

11 STUDY 3 RESULTS

First, we analyzed descriptive statistics to assess the classification agreement between the four clickbait detectors. Data reveal that the four classifiers agreed on the classification 47.17% of the times. Of the175 headlines that were classified similarly by the four classifiers, 139 were clickbait classifications and 36 were non-clickbait. Furthermore, as Figure 8 shows, the level of agreement with the classification also varied based on the characteristic used by the headline. For instance, while the four classifiers agreed on the clickbait classification more times for the negative superlative characteristic (compared to the other six characteristics), the four classifiers never agreed on a non-clickbait classification for the negative superlative or question characteristics, as illustrated by the absence of the "non-clickbait agreement" bar.

We then proceeded to compare classifiers pairwise; specifically, the percentage of times that two classifiers agreed on either a clickbait or a non-clickbait classification. Table 6 reveals that the highest agreement was between detector 1 and 4, both of which used manually annotated data for training.

To analyze whether one clickbait characteristic received more engagement than another, as proposed by RQ1, we ran a series of negative binomial regressions with maximum likelihood estimation. This was the most appropriate analysis given overdispersion of the count data. Given the rather low agreement between detectors explained above, we ran a separate analysis for each detector so that we could assess 1) if there is a feature that is more successful at engaging users across all classifiers, and 2) if the type of model (classical machine learning vs. deep learning) and the training dataset (annotated data set vs. weak supervision with source assumption)

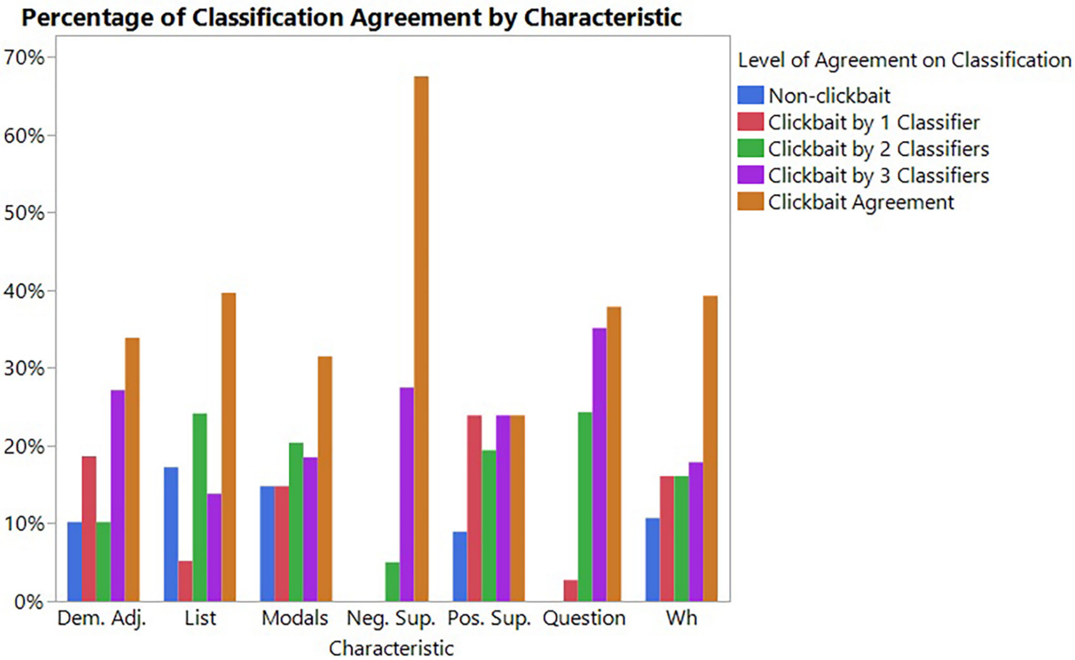

Figure 8: Percentage of times the clickbait detectors agreed on a classification by characteristic.

Table 6: Pairwise comparison agreement between detectors

Comparison	Percentage of Total Headlines Agreed as Clickbait	Percentage of Total Headlines Agreed as Non-Clickbait	Total Agreement
Detector 1 &. Detector 2	59.57%	14.02%	73.59%
Detector 1 & Detector 3	50.94%	18.87%	69.81%
Detector 1 & Detector 4	54.72%	23.18%	77.90%
Detector 2 & Detector 3	49.06%	15.36%	64.42%
Detector 2 & Detector 4	48.52%	15.36%	63.88%
Detector 3 & Detector 4	47.17%	27.49%	74.66%

Note: Percentages of total headlines agreed as non-clickbait (or clickbait) refer to the percentage of the total number of headlines where the two classifiers agreed on the non-clickbait or clickbait determination.

yield different results[4]. For all analyses, we entered the type of content (political vs. non-political) and the headline features as the independent variables, with total engagement (combined number of shares, comments, reactions and pins) as the dependent variable.

When analyzing Detector 1 (Deep Learning, Manually Annotated Data), we found a significant effect of types of content, such that non-political headlines received more engagement than political headlines, $b = 0.55$, Wald $\chi^2 = 4.64$, $p = .03$. There was also a significant effect of features, Wald χ^2 =42.59, $p < .0001$. Post-hoc comparisons using Tukey HSD revealed that demonstrative adjectives, lists, modals, and “wh” words received higher engagement than non-clickbait headlines.

When analyzing Detector 2 (Traditional Machine Learning: Naïve Bayes, Weak Supervision), we also found a significant effect of features, Wald χ^2 =31.72, $p < .0001$. Post-hoc comparisons revealed that demonstrative adjectives, lists, and “wh” words performed better than non-clickbait headlines. Analysis with Detector 2 additionally yielded significant pairwise differences between clickbait headlines, such that demonstrative adjectives, lists and “wh” words received more engagement than positive superlatives.

When analyzing Detector 3 (Deep Learning, Weak Supervision), there was also a significant effect of feature, Wald χ^2 =31.99, $p < .0001$. Again, post-hoc comparisons revealed that demonstrative adjectives, lists, and “wh” words performed better than non-clickbait.

[4]When running the analysis using headlines where at least three detectors agreed on the classification (a rather liberal approach given that for some headlines one classifier would have deemed it as non-clickbait) there was a near significant effect of features on total engagement, Wald $\chi 2$ =12.32, $p = .055$. However, post-hoc comparisons using Tukey HSD yielded no significant pairwise differences.

Data also revealed that demonstrative adjectives and "wh" words received more engagement than positive superlatives.

Finally, analysis of Detector 4 (Traditional Machine Learning: Support Vector Machine, Manually Annotated Data) yielded a significant effect of feature, Wald χ^2 =23.15, p = .002. Tukey HSD pairwise comparison revealed that the difference is only between demonstrative adjectives and positive superlatives, with demonstrative adjectives receiving higher engagement.

12 STUDY 3 DISCUSSION

Results of Study 3 are intriguing as it reveals rather low agreement among the four clickbait detectors, despite the detectors individually having high accuracy. This is likely due to the different assumptions and characteristics of each detector. For instance, two of the models were trained using annotated data while the other two used source assumptions as ground truth. This means that in the former, clickbait is defined in terms of the perceptions of the volunteers who classified those headlines and their understanding of what is and what is not clickbait. On the other hand, the other two models relied on source credibility to train the model, such that headlines from sources like the Wall Street Journal are understood as being non-clickbait, while those from outlets like Buzzfeed are assumed to be clickbait. Two of the models utilized a deep learning model, while the other two utilized more classical machine-learning models (Naïve Bayes and Support Vector Machine). Results of Study 3 reveal that the differences in assumptions made by each model results in a low agreement when comparing the four detectors together and generate different results in terms of user engagement. It is worth noting, however, that when examining the pairwise comparison among classifiers, the highest total agreement was between the two human-labeled classifiers—classifiers 1 and 4 (see Table 6).

Regardless of the low agreement between the clickbait detectors, three out of the four models consistently revealed that clickbait headlines using demonstrative adjectives, lists and "wh" words resulted in higher engagement compared to non-clickbait headlines. This means that at least in an uncontrolled environment, these characteristics are more successful in luring users into clicking. Similarly, two of the four models reveal that positive superlatives are not as good as other clickbait characteristics (e.g., demonstrative adjectives) at engaging users. These differences notwithstanding, it is important to note that some of the headlines classified as non-clickbait by classifiers still contained one of the seven characteristics of clickbait analyzed in this paper (see supplemental material for a list of headlines). This means that it is not the characteristics alone that distinguish a headline as clickbait or not clickbait. It is likely that there are other underlying linguistic properties that may elicit information gap more effectively when read in tandem with a particular characteristic.

13 SUMMARY OF FINDINGS

In sum, we conducted three studies to assess if clickbait headlines are actually "clickbaity" (see Table 7 for a comparison of findings). In Study 1, we found that among the seven clickbait characteristics, users were more likely to click on "read more" for "wh" words and lists compared to modals. They were more likely to click on "read more" for lists compared to questions. However, results revealed an overall preference for non-clickbait. In fact, non-clickbait received significantly more clicks than modals, negative superlatives, positive superlatives, and questions, and the same as demonstrative adjectives, lists, and "wh" words. Moreover, when assessing user perception of the headlines, non-clickbait headlines elicited more curiosity than all 7 types of clickbait headlines, were perceived as less deceitful and considered more credible. The interaction effects between clickbait characteristic and type of content on user perceptions suggest that the success of clickbait is also contingent on the type of content of the headline. For example, non-political headlines were perceived as more credible than political headlines, except for modals, questions and "wh" headlines. Similarly, political headlines were generally perceived as more deceitful than non-political headlines, except for headlines using questions, in which case non-political headlines were perceived as more deceitful.

In Study 2, we assessed user preference in a more controlled environment to account for the possibility that the general preference for non-clickbait found in Study 1 could be due to content differences across the headlines. However, even when we systematically varied the same headline to possess one of the seven characteristics or non-clickbait, we found no statistically significant differences in engagement between them. This means that users were as likely to click on "read more" and share for any of the clickbait characteristics or non-clickbait, and their perceptions of the headline (deception, curiosity, entertaining, credible) and news story (credibility, representativeness, quality, liking) were the same regardless of headline type.

The findings of Study 1 and 2 suggested two possibilities for the positive effects of clickbait on engagement found in previous computational studies [36, 44]. First, that in these studies clickbait headlines are not "clickbaity" because of the characteristics suggested by industry [5] and identified by the clickbait detector in [36] alone, but by other factors that help create a psychological knowledge gap. Or, that the preference for clickbait is a function of third variables derived from the assumptions of the clickbait detectors. We explored these possibilities in Study 3 by comparing the classification of 4 different clickbait detectors varying in the nature of training data and the type of machine learning employed. While three of the four classifiers suggest that users engage more with headlines using demonstrative adjectives, "wh" words, and lists, we found low overall agreement among the 4 classifiers, such that the four of them agreed on a clickbait vs. non-clickbait classification only 47% of the time. We discuss the implications of our three studies in the next section.

14 GENERAL DISCUSSION

Overall, our findings suggest that determining whether a headline is clickbait or not is quite complex. It does not simply depend on a few key linguistic characteristics like using a listicle or framing it as a question. It also depends on the nature of the automated classifier used to distinguish clickbaits from non-clickbaits. In fact, results of our study expose the unreliability of clickbait classifiers by showing low agreement between them. Like our Study 1 and 2, past studies [29, 37] conducted using experimental designs and content analysis find that non-clickbait is more engaging than clickbait. Studies using automatic detection [36] reveal the opposite

Table 7: Summary of significant main effects of clickbait characteristic by study

Study 1	
Outcome	**Comparison**
Read more	Non-clickbait > modal; Non-clickbait > negative superlatives; Non-clickbait > positive superlatives; Non-clickbait > questions; Non-clickbait = demonstrative adjectives; Non-clickbait = demonstrative adjectives; Non-clickbait = lists; Non-clickbait = "wh"; Wh > modals; Lists >modals; List > questions
Curiosity	Non-clickbait >modals; Non-clickbait >negative superlatives; Non-clickbait >positive superlatives; Non-clickbait >questions; Non-clickbait >"wh"
Credibility	Non-clickbait > list; Non-clickbait > modal; Non-clickbait > negative superlative; Non-clickbait > positive superlative; Non-clickbait > "wh"
Deception	List >demonstrative adjective; List>modals; List >positive superlatives; List> question; List > Non-clickbait; Negative superlative >modals; Negative superlative >positive superlatives; Negative superlative >no-clickbait
Entertaining	List > question; Negative superlative > question
Study 2	
Outcome	**Comparison**
Headline engagement (read more & share)	No significant differences
Headline perceptions (deception, curiosity, entertaining, credible)	No significant differences
Story perceptions (credibility, representativeness, quality, liking)	No significant differences
Elaboration of the story	No significant differences
Likelihood of sharing the story	No significant differences
Study 3 – Detector 1 (Deep Learning, trained on Manually Annotated Data)	
Outcome	**Comparison**
Shares	Demonstrative adjectives > non-clickbait; Lists > non-clickbait; "Wh" > non-clickbait; Modals > non-clickbait
Study 3 – Detector 2 (Traditional Machine Learning: Naïve Bayes, Weak Supervision)	
Outcome	**Comparison**
Shares	Demonstrative adjectives > non-clickbait; Lists > non-clickbait; "Wh" > non-clickbait; Demonstrative adjectives > positive superlative; Lists > positive superlatives; "Wh" > positive superlatives
Study 3 – Detector 3 (Deep Learning, Weak Supervision)	
Outcome	**Comparison**
Shares	Demonstrative adjectives > non-clickbait; Lists > non-clickbait; "Wh" > non-clickbait; Demonstrative adjectives > positive superlative; "Wh" words > positive superlatives
Study 3 – Detector 4 (Traditional Machine Learning: Support Vector Machine, trained on Manually Annotated Data)	
Outcome	**Comparison**
Shares	Demonstrative adjectives > positive superlative

pattern, as did our Study 3. But, in analyzing each detector in isolation, we find that the four models only agreed 47% of the time, which indicates the complexity of defining and classifying clickbait computationally. The noise associated with this low level of agreement raises fundamental questions about the validity of clickbait detectors and taxonomies touting objective characteristics that result from clickbait analysis. In reality, clickbait determination could be highly subjective. As the author of one of the clickbait detectors used in this study [13] states "I know it when I see it." However, classifying clickbait might be a difficult task precisely for that reason—we know it when we see it, but we cannot quite define it operationally. It is possible that clickbait, and more precisely the curiosity gap that it is supposed to generate, represents an abstract concept difficult to define at the granularity needed for computational detection. The higher engagement of clickbait over non-clickbait found in computational studies might be due to other variables such as topic distinctions and assumptions of the system rather than "clickbaitiness." It is also possible that the results of our study represent the current digital user, one who no longer falls for clickbait headlines due to either their considerable prior experience with this type of headline (including the disappointment and frustration they may have felt) or due to the increasing number of media literacy campaigns in recent years aimed to educating the public. Culturally, we may have reached an inflection point where we have come to recognize clickbaits for what they really are and therefore deliberately avoid clicking on them.

That said, three classifiers agreed that demonstrative adjectives, "wh" words and listicles performed better than non-clickbait when using real-world engagement data. These stylistic markers might be essential for clickbait headlines to be "clickbaity," and are worthy of further testing. However, it is worth noting that we did not find the same effects in Study 1 and 2. This is important because, in Study 1, we selected 1) clickbait headlines that possessed only one of the seven characteristics, and 2) non-clickbait headlines that did not contain any of the characteristics (or manually adjusted to ensure that this was the case), whereas in Study 2, we further controlled for any possible content differences by manipulating the same headline to possess one and only one or none of the characteristics. This was not the case for Study 3, where even headlines considered as non-clickbait by the four classifiers appeared to have at least traces of some of the linguistic characteristics associated with clickbait headlines. For example, the headlines, "Greatest Military Coup Ever Could Now Be Underway on U.S. Soil," and "Media Silent as Biggest Protests Since French Revolution Sweep France," both use positive superlative words, yet were classified as non-clickbait by the four classifiers. This raises the question whether the characteristics themselves determine the 'clickbaitiness' of a headline or are there other linguistic mechanisms that promote user engagement? Furthermore, in Study 1 and 2, we assessed users' perception of the headline and found a null effect (Study 2) and a slight superiority of non-clickbait headlines (Study 1). As data scientists are well aware [5], analysis of real-world engagement can be rife with confounding variables. For example, it is possible that one type of clickbait headline, more than other types, tends to be placed on the upper part of the news site, generating more clicks partly because of its placement rather than the headline alone. This type of confound could lead to a form of Simpson's Paradox [5], where the direction of an association may be reversed when analyzing the subgroups of that population [21]. This is what we may be witnessing in Study 3. The many third variables associated with real-world engagement data coupled with the low reliability of the detectors found in our study sheds light on the possibility that the greater engagement of clickbait reported in computational studies should be interpreted with caution because the engagement might not be due to actual "clickbaitiness" or the knowledge gap hypothesis, but rather due to contextual factors and interface elements unrelated to clickbait.

Our findings hold several practical implications. First, our study echoes recent works by other scholars [37] and suggests that news organizations should steer away from utilizing clickbait characteristics in their headline writing, especially for topics that rely on users' perceptions of integrity and objectivity. For starters, headlines written as non-clickbait receive the same or equal engagement compared to clickbait headlines. Moreover, users perceive non-clickbait as more credible and less deceptive. They are also able to induce user curiosity to a higher extent than clickbait, probably because they are processed more deeply by users. All this suggests that respectable news organizations should refrain from resorting to clickbaits to boost engagement with their stories, as it is not only ineffective in attracting users but also detrimental to their credibility. In other words, resorting to clickbait is not worth the credibility risk, especially in the news domain where trust in news organizations has already been on a decline. A recent survey of U.S. adults indicated that credibility of news organizations declined in 2020 compared to previous years, "indicating not only a lack of trust but suggesting that audiences have grown cautious when consuming news" [45] (para. 1). One way to boost credibility and differentiate from other types of content online is for news organizations to stick to non-clickbait headlines when presenting content to their audiences.

Our findings also raise fundamental doubts about the reliability of computational clickbait detection solutions. Particularly, the low agreement among clickbait detectors reveals that clickbait detection is a rather difficult task. Our results suggest that the success of future detectors should be analyzed beyond their individual performance (the four classifiers in this study had high accuracy scores), but also based on their concurrent validity. In other words, detectors should be assessed by comparing them to other clickbait measures to assess their ability to accurately and consistently detect clickbait. One idea can be to test how the detector performs in comparison to human classifiers. Insights into those results can provide ideas to improve detection accuracy by assessing potential areas of disagreement. Another approach proposed by scholars studying fake news [28] is for the computational sciences to collaborate with the social sciences when building machine learning models for the detection of abstract concepts such as clickbait. For example, the social sciences can help identify features essential for clickbait detection or other abstract concepts, which can then be used to build algorithms for the detection of such content.

This research is not without limitations. First, in our studies, we limited our classification of clickbait to seven characteristics. It is possible that other characteristics of clickbait not represented in our research are more successful than the ones we selected. Similarly, we utilized headlines that only possessed one of the seven characteristics. The goal of our research was to identify which of those characteristics is more engaging. However, there could be a potential additive effect that we did not account for. Secondly, in our user studies (Study 1 and 2), we used a sample of M-Turkers and students; though no difference was found across populations, it is not a representative sample of the U.S. population which might be less technologically savvy compared to Turkers and college students. However, while our sample is not representative, college students are as likely to fall for persuasive linguistic strategies such as clickbait, as the normal population. In fact, a study [46] found that college students performed worse than high school students in a reasoning task to evaluate online information. It is also worth noting that the scraped headlines utilized for Study 3 is considerably small, compared to other field studies conducted to analyze headline engagement. Nonetheless, our findings align with Scacco and Muddiman [37], whose field analysis of 5288 headlines revealed higher engagement of non-clickbait headline. Though our Study 3 does reveal a slight preference for a few clickbait characteristics (demonstrative adjectives, "wh" words, and listicles), these findings should also be analyzed with caution given potential confounds of real-world engagement analysis discussed earlier, and the rather low agreement rate among the detectors. Likewise, in Study 1 and 2, we presented headlines that were classified as clickbait by three classifiers, but we found in Study 3 that the classifiers have low agreement. The low agreement raises questions about the validity of our Study 1 and Study 2. Nonetheless it is worth noting that the null effect of clickbait headlines is consistent with findings in other experimental studies [29, 30, 37].

These limitations notwithstanding, our research adds to our current understanding of users' engagement with clickbait headlines by exploring possible reasons for the mixed findings regarding their relative superiority over non-clickbait headlines. Our studies reveal that indeed clickbait is not as "clickbaity" as we tend to think, and that results might vary depending on the conceptual definition of clickbait. For example, in our study, the four clickbait detectors yielded somewhat different engagement results, and agreed on a classification only about 47% of the time. Moreover, three of the four detectors suggest user preference for demonstrative adjectives, lists, and "wh" words. However, these findings were not consistent when we tested this premise in a controlled and semi-controlled setting, suggesting that other factors pertaining to content of the story may indeed play a significant role in attracting clicks from users. Furthermore, other content genres, such as gossip, advertisements, and warnings, may fare better with clickbaits compared to straight news headlines, which many online users may see as inherently unworthy of clicks.

In conclusion, this paper exposes the unreliability of clickbait detecting classifiers by demonstrating low agreement between them. While results of this research reveal that some characteristics (demonstrative adjectives, "wh" words, and listicles) attract more clicks when analyzing real-world engagement data, the noise associated with the low level of agreement among classifiers raises fundamental questions about the validity of these systems. It is possible that the higher engagement of clickbait in computational studies (such as our Study 3) is driven by other third variables that go beyond the linguistic features of the headline, e.g., topic differences and article placement within the website that may have nothing to do with "clickbaitiness" or curiosity effects. It is also possible that the modest effects of clickbaits are due perhaps to greater media literacy and user wariness arising from their prior experience with clickbait headlines.

ACKNOWLEDGMENTS

This research is in part supported by the U. S. National Science Foundation (NSF) awards #1742702, #1820609, #1909702, #1915801, and #1934782.

REFERENCES

[1] Prakhar Biyani, Kostas Tsioutsiouliklis, and John Blackmer. 2016. "8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (AAAI'16), 94–100.

[2] Jonas Nygaard Blom and Kenneth Reinecke Hansen. 2015. Click bait: Forward-reference as lure in online news headlines. *J. Pragmat.* 76, , 87–100. DOI: https://doi.org/10.1016/j.pragma.2014.11.010

[3] Piotr S. Bobkowski. 2015. Sharing the news: Effects of informational utility and opinion leadership on online news sharing. *Journal. Mass Commun. Q.* 92, 2, 320–345. DOI: https://doi.org/10.1177/1077699015573194

[4] John D. Bransford and Marcia K. Johnson. 1972. Contextual prerequisites for understanding: Some investigations of comprehension and recall. *J. Verbal Learn. Verbal Behav.* 11, 6 (Dec. 1972), 717–726. DOI: https://doi.org/10.1016/s0022-5371(72)80006-9

[5] Chris Breaux. 2015. You'll never guess how chartbeat came up with the greatest headline. Retrieved August 3, 2020 from https://blog.chartbeat.com/2015/11/20/youll-never-guess-how-chartbeats-data-scientists-came-up-with-the-single-greatest-headline/

[6] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: detecting and preventing clickbaits in online news media. In *International Conference on Advances in Social Network Analysis and Mining (ASONAM)* (Aug. 2016).

[7] Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. 2015. Misleading online content: Recognizing clickbait as "false news." In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection* (WMDD '15), Nov. 2015, New York, NY, 15–19. DOI: https://doi.org/10.1145/2823465.2823467

[8] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv14123555.* Retrieved from http://arxiv.org/abs/1412.3555

[9] Peter R. Darke and Robin J. B. Ritchie. 2007. The defensive consumer: Advertising deception, defensive processing, and distrust. *J. Mark. Res.* 44, 1, 114–127. DOI: https://doi.org/10.1509/jmkr.44.1.114

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv181004805.*Retrieved from http://arxiv.org/abs/1810.04805

[11] D. James Dooling and Roy Lachman. 1971. Effects of comprehension on retention of prose. *J. Exp. Psychol.* 88, 2 (May 1971), 216–222. DOI: https://doi.org/10.1037/h0030904

[12] Daniel Dor. 2003. On newspaper headlines as relevance optimizers. J. Pragmat. 35, 5 (Jan. 2003), 695–721. DOI: https://doi.org/10.1016/s0378-2166(02)00134-0

[13] Peter Downs. 2020. Clickbait classifier. Retrieved August 10, 2020 from https://github.com/peterldowns/clickbait-classifier

[14] Andrew Duffy, Edson Tandoc, and Rich Ling. 2019. Too good to be true, too good not to share: the social utility of fake news. *Inf. Commun. Soc.* (June 2019), 1–15. DOI: https://doi.org/10.1080/1369118X.2019.1623904

[15] Ullrich K. H. Ecker, Stephan Lewandowsky, Ee Pin Chang, and Rekha Pillai. 2014. The effects of subtle misinformation in news headlines. *J. Exp. Psychol. Appl.* 20, 4 (DEc. 2014), 323–335. DOI: https://doi.org/10.1037/xap0000028

[16] Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *J. Commun.* 43, 4, 51–58. DOI: https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

[17] Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. 2016. Social clicks: What and who gets read on Twitter? *ACM SIGMETRICS Perform. Eval. Rev.* 44, 1 (June 2016), 179–192. DOI: https://doi.org/10.1145/2964791.2901462

[18] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *arXiv200407780.* Retrieved from http://arxiv.org/abs/2004.07780

[19] Elly Ifantidou. 2009. Newspaper headlines and relevance: Ad hoc concepts in ad hoc contexts. *J. Pragmat.* 41, 4 (April 2009), 699–720. DOI: https://doi.org/10.1016/j.pragma.2008.10.016

[20] Ronnie Janoff-Bulman, Sana Sheikh, and Kate G. Baldacci. 2008. Mapping moral motives: Approach, avoidance, and political orientation. *J. Exp. Soc. Psychol.* 44, 4 (July 2008), 1091–1099. DOI: https://doi.org/10.1016/j.jesp.2007.11.003

[21] Rogier Kievit, Willem Eduard Frankenhuis, Lourens Waldorp, and Denny Borsboom. 2013. Simpson's paradox in psychological science: a practical guide. *Front. Psychol.* 4, (Aug. 2013). DOI: https://doi.org/10.3389/fpsyg.2013.00513

[22] Tim Groot Kormelink and Irene Costera Meijer. 2017. What clicks actually mean: Exploring digital news user practices. *Journal. Lond. Engl.* 19, 5 (Jan. 2017), 668–683. DOI: https://doi.org/10.1177/1464884916688290

[23] Thai Le, Kai Shu, Maria D. Molina, Dongwon Lee, S. Shyam Sundar, and Huan Liu. 2019. 5 sources of clickbaits you should know! using synthetic clickbaits to improve prediction and distinguish between bot-generated and human-written headlines. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '19),* ACM, New York, NY, USA, 33–40. https://doi.org/10.1145/3341161.3342875

[24] Gwilym Lockwood. 2016. Academic clickbait: articles with positively-framed titles, interesting phrasing, and no wordplay get more attention online. *The Winnower* (June 2016). DOI: https://doi.org/10.15200/winn.146723.36330

[25] George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychol. Bull.* 116, 1 (Jul. 1994), 75–98.

[26] Saurabh Mathur. 2019. Clickbait-detector. (August 2019). DOI: https://doi.org/10.17605/OSF.IO/T3UJ9

[27] Matthew T. McCrudden and Gregory Schraw. 2006. Relevance and goal-focusing in text processing. *Educ. Psychol. Rev.* 19, 2 (2Oct. 2006), 113–139. DOI: https://doi.org/10.1007/s10648-006-9010-7

[28] Maria D. Molina, S. S. Sundar, T Le, and D Lee. 2019. "Fake news" is not simply false information: A concept explication and taxonomy of online content. *Am. Behav. Sci.* (Oct. 2019), 1–33. DOI: https://doi.org/10.1177/0002764219878224

[29] Logan Molyneux and Mark Coddington. 2020. Aggregation, Clickbait and their effect on perceptions of journalistic credibility and quality. *Journal. Pract.* 14, 4 (April 2020), 429–446. DOI: https://doi.org/10.1080/17512786.2019.1628658

[30] Kevin Munger, Mario Luca, Jonathan Nagler, and Joshua Tucker. 2020. The (null) effects of clickbait headlines on polarization, trust, and learning. *Public Opin. Q.* 84, 1, 49–73. DOI: https://doi.org/10.1093/poq/nfaa008

[31] F. D. Naylor. 1981. A state-trait curiosity inventory. *Aust. Psychol.* 16, 2 (March 1981), 172–183. DOI: https://doi.org/10.1080/00050068108255893

[32] Dustin Ormond, Merrill Warkentin, Allen C. Johnston, and Samuel C. Thompson. 2016. Perceived deception: Evaluating source credibility and self-efficacy. *J. Inf. Priv. Secur.* 12, 4 (Dec. 2016), 197–217. DOI: https://doi.org/10.1080/15536548.2016.1243857

[33] Cherniece J. Plume and Emma L. Slade. 2018. Sharing of sponsored advertisements on social media: A uses and gratifications perspective. *Inf. Syst. Front.* 20, 3 (June 2018), 471–483. DOI: https://doi.org/10.1007/s10796-017-9821-8

[34] Martin Potthast, Tim Gollub, Matthias Hagen, and Benno Stein. 2018. The clickbait challenge 2017: Towards a regression model for clickbait strength. *arXiv181210847* Retrieved from http://arxiv.org/abs/1812.10847

[35] Rodney A. Reynolds. 1997. A validation test of a message elaboration measure. *Commun. Res. Rep.* 14, 3 (May 1997), 269–278. DOI: https://doi.org/10.1080/08824099709388670

[36] Md Main Uddin Rony, Naeemul Hassan, and Mohammad Yousuf. 2017. Diving deep into clickbaits: Who use them to what extents in which topics with what effects? arXiv:1703.09400v1. Retrieved from https://arxiv.org/abs/1703.09400

[37] Joshua M. Scacco and Ashley Muddiman. 2020. The curiosity effect: Information seeking in the contemporary news environment. *New Media Soc.* 22, 3,429–448. DOI: https://doi.org/10.1177/1461444819863408

[38] Craig Silverman. 2015. Lies, damn lies and viral content. DOI: https://doi.org/10.7916/D8Q81RHH

[39] S. Shyam Sundar. 1999. Exploring receivers' criteria for perception of print and online news. *Journal. Mass Commun. Q.* 76, 2 (June 1999), 373–386. DOI: https://doi.org/10.1177/107769909907600213

[40] Edson C. Tandoc. 2014. Journalism is twerking? How web analytics is changing the process of gatekeeping. *New Media Soc.* 16, 4 (May 2014), 559–575. DOI: https://doi.org/10.1177/1461444814530541

[41] Percy H. Tannenbaum. 1953. The effect of headlines on the interpretation of news stories. *Journal. Q.* 30, 2 (Spring 1953), 189.

[42] Gerard J. Tellis, Deborah J. MacInnis, Seshadri Tirunillai, and Yanwei Zhang. 2019. What drives virality (sharing) of online digital content? The critical role of information, emotion, and brand prominence. *J. Mark.* 83, 4 (April 2019), 1–20. DOI: https://doi.org/10.1177/0022242919841034

[43] Ori Tenenboim and Akiba A. Cohen. 2015. What prompts users to click and comment: A longitudinal study of online news. *Journalism* 16, 198–217. DOI: https://doi.org/10.1177/1464884913513996

[44] Lasya Venneti and Aniket Alam. Clickbaits: Curious hypertexts for news narratives in the digital medium. In *Proceedings of Workshop on Narrative and Hypertext '17 (NHT17),* July 2017, Prague, Czech Republic.

[45] Amy Watson. Credibility of major news organizations in the United States from 2017 to 2020. Retrieved from https://www.statista.com/statistics/239784/credibility-of-major-news-organizations-in-the-us/

[46] Sam Wineburg, Sarah McGrew, Joel Breakstone, and Teresa Ortega. Evaluating information: The cornerstone of civic online reasoning. Retrieved from https://purl.stanford.edu/fv751yt5934