A Large-Scale Analysis of Health Journalism by Reliable and Unreliable Media

Sameer Dhoju[a], Md Main Uddin Rony[a], Muhammad Ashad Kabir[b], Naeemul Hassan[a]

a Department of Computer and Information Science, University of Mississippi, USA
b School of Computing and Mathematics, Charles Sturt University, Australia

Abstract

Media outlets play crucial roles in disseminating health information. Previous studies have examined how health journalism is practiced by reliable and unreliable media outlets. However, most of the existing works are conducted over a relatively small set of samples. In this study, we investigate a large collection (about 30 thousand) of health-related news articles which were published by 29 reliable and 20 unreliable media outlets and identify several differences in health journalism practice. Our analysis shows that there are significant structural, topical, and semantic disparities in the way reliable and unreliable media outlets conduct health journalism. We argue, in this age of 'fake news', these findings will be useful to combat online health disinformation.

Keywords:

Health Communication

Introduction

Of the 20 most-shared articles on Facebook in 2016 with the word "cancer" in the headline, more than half the reports were discredited by doctors and health authorities [1]. The spread of health-related hoaxes is not new. However, the advent of Internet, social networking sites (SNS), and click-through-rate (CTR)-based pay policies have made it possible to create hoaxes/"fake news", publish at a larger scale and reach to a broader audience with a higher speed than ever [2]. Misleading or erroneous health news can be dangerous as it can lead to a critical situation. Houston reported a measles outbreak in Europe due to lower immunization rate which experts believed was the result of anti-vaccination campaigns caused by a false news about MMR vaccine [3]. Moreover, misinformation can spoil the credibility of the health-care providers and create a lack of trust in taking medicine, food, and vaccines. Recently, researchers have started to address the fake news problem in general [4; 5]. However, health disinformation is a relatively unexplored area. According to a report from Pew Research Center [6], 72% of adult internet users search online for information about a range of health issues. So, it is important to ensure that the health information which is available online is accurate and of good quality. There are some authoritative and reliable entities such as National Institutes of Health (NIH)[1] or Health On the Net[2] which provide high-quality health information. Also, there are some fact-checking sites such as Snopes.com[3] and Quackwatch.org[4] that regularly debunk health and medical related misinformation. Nonetheless, these sites are incapable of busting the deluge of health disinformation continuously produced by unreliable health information outlets

(e.g., RealFarmacy.com, Health Nut News). Moreover, the bots in social networks significantly promote unsubstantiated health-related claims [7]. Researchers have tried developing automated health hoax detection techniques but had limited success due to several reasons such as small training data size and lack of consciousness of users [8-11].

The objective of this paper is to identify discriminating features that can potentially separate a reliable health news from an unreliable health news by investigating a large-scale dataset. We examine how reliable media and unreliable media outlets conduct health journalism. First, we prepare a large dataset of health-related news articles which were produced and published by a set of reliable media outlets and unreliable media outlets. Then, using a systematic content analysis, we identify the features which separate a reliable outlet sourced health article from an unreliable sourced one. These features incorporate the structural, topical, and semantic differences in health articles from these outlets. For instance, our structural analysis finds that the unreliable media outlets use clickbaity headlines in their health-related news significantly more than what reliable outlets do. Our topical analysis finds that while the reliable outlets discuss "cancer" along with research and studies, in the unreliable outlets "cancer" is associated with autism and vaccination. The semantic analysis shows that on average a health news from reliable media contains more reference quotes than an average unreliable sourced health news. We argue that these features can be critical in understanding health misinformation and designing systems to combat such disinformation. In future, our goal is to develop a machine learning model using these features to distinguish unreliable media sourced health news from reliable articles.

Related Work

There has been extensive research on how scientific medical research outcomes should be disseminated to general people by following health journalism protocols [12-16]. For instance, Lopes et al. suggest that it is necessary to integrate journalism studies, strategic communication concepts, and health professional knowledge to successfully disseminate professional findings. Some researchers particularly focused on the spread of health misinformation in social media [17]. For example, [8] analyzes Zika related misinformation on Twitter. In particular, it shows that tracking health misinformation in social media is not trivial, and requires some expert supervision. It exploited crowdsource to annotate a collection of Tweets and used the annotated data to build a rumor classification model. One limitation of this work is that the used dataset is too small (6 rumors) to make a general conclusion. Moreover, it didn't consider the features in the actual news articles unlike us. Ghenai and Mejova [9] examines the individuals on social media that are posting questionable health-related information,

---

[1] https://www.nih.gov/
[2] https://www.hon.ch/en/
[3] https://www.snopes.com/
[4] http://www.quackwatch.org/

and in particular promoting cancer treatments which have been shown to be ineffective. It develops a feature based supervised classification model to automatically identify users who are comparatively more susceptible to health misinformation. There are other works which focus on automatically identifying health misinformation. For example, Kinsora et al. [18] developed a classifier to detect misinformative posts in health forums. One of the limitations of this work is that the training data is only labeled by two individuals. Researchers have also worked on building tools that can help a user to easily consume health information. Kostkova et al. [10] developed the "VAC Medi+board", an interactive visualization platform integrating Twitter data and news coverage from a reliable source called MediSys[5]. It covers public debate related to vaccines and helps users to easily browse health information on a certain vaccine-related topic.

Our study significantly differs from these already existing researches. Instead of depending on a small sample of health hoaxes like some of the existing works, we take a different approach and focus on the source outlets. This gives us the benefit of investigating with a larger dataset. We investigate the journalistic practice of reliable and unreliable outlets, an area which has not been studied to the best of our knowledge.

Data Preparation

The results are presented here. Authors may choose a combination of text, tables, figures, and graphs to convey the results of their work to the reader. There are no set limitations on the number of tables, figures, and graphs that may be used in papers, posters, and proposals. Large figures and tables may span two columns. Please number tables and figures and reference them appropriately in the text.

Media Outlet Selection

The first challenge is to identify reliable and unreliable outlets. The matter of reliability is subjective. We decided to consider the outlets which have been cross-checked as reliable or unreliable by credible sources.

Reliable Media

We identified 29 reliable media outlets from three sources– i) 11 of them are certified by the Health On the Net [19], a non-profit organization that promotes transparent and reliable health information online. It is officially related with the World Health Organization (WHO) [20]. ii) 8 from U.S. government's health-related centers and institutions (e.g., CDC, NIH, NCBI), and iii) 10 from the most circulated broadcast [21] mainstream media outlets (e.g., CNN, NBC). Note, the mainstream outlets generally have a separate section for health information (e.g., https://www.cnn.com/health). As our goal is to collect health-related news, we restricted ourselves to their health portals only.

Unreliable Media

Dr. Melissa Zimdars, a communication and media expert, prepared a list of false, misleading, clickbaity, and satirical media outlets [22; 23]. Similar lists are also maintained by Wikipedia [24] and informationisbeautiful.net [25]. We identified 6 media outlets which primarily spread health-related misinformation and are present in these lists. Another source for identifying unreliable outlets is Snopes.com, a popular hoax-debunking website that fact-checks news of different domains including health. We followed the health or medical hoaxes debunked by Snopes.com and identified 14 media outlets which sourced those hoaxes. In total, we identified 20 unreliable

outlets. Table 1 lists the Facebook page ids of all the reliable and unreliable outlets that have been used in this study.

Table 1– List of Facebook page ids of the reliable and unreliable outlets. Some of them are unavailable now

Reliable	everydayhealth, WebMD, statnews, AmericanHeart, BBCLifestyleHealth, CBSHealth, FoxNewsHealth, WellNYT, latimesscience, tampabaytimeshealth, philly.comhealth, AmericanHeart, AmericanCancerSociety, HHS, CNNHealth, cancer.gov, FDA, mplus.gov, NHLBI, kidshealthparents, ahrq.gov, healthadvocateinc, HealthCentral, eMedicineHealth, C4YWH, BabyCenter, MayoClinic, MedicineNet, healthline
Unreliable	liveahealth, healthexpertgroup, healthysolo, organichealthcorner, justhealthylifestyle1, REALfarmacy.com, thetruthaboutcancer, BookforHealthyLife, viralstories.bm, justhealthyway, thereadersfile, pinoyhomeremedies, onlygenuinehealth, greatremediesgreathealth, HealthRanger, thefoodbabe, AgeofAutism, HealthNutNews, consciouslifenews, HealthImpactNews

Data Collection

The next challenge is to gather news articles published by the selected outlets. We identified the official Facebook pages of each of the 49 media outlets and collected all the link-posts[6] shared by the outlets within January 1, 2015 and April 2, 2018 using Facebook Graph API. For each post, we gathered the corresponding news article link, the status message, and the posting date.

News Article Scraping

We used a Python package named Newspaper3k[7] to gather the news article related data. Given a news article link, this package provides the headline, body, author name (if present), and publish date of the article. It also provides the visual elements (image, video) used in an article. In total, we collected data for 29, 047 articles from reliable outlets and 15, 017 from unreliable outlets.
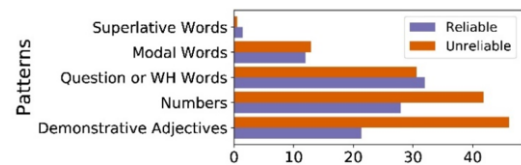

Figure 1– Distribution of clickbait patterns

Filtering non-Health News Articles

Even though we restricted ourselves to health-related outlets, we observed that the outlets also published or shared non-health (e.g., sports, entertainment, weather) news. We removed these non-health articles from our dataset and only kept *health, food & drink*, or *fitness & beauty* related articles. Specifically, for each news article, we used the document categorization service provided by Google Cloud Natural Language API to determine its topic. If an article doesn't belong to one of the three above

mentioned topics, it is filtered out. This step reduced the dataset size to 27,589; 18,436 from reliable outlets and 9,153 from unreliable outlets. We used this health-related dataset only in all the experiments of this paper. Figure 1 shows the health-related news percentage distribution for reliable outlets and unreliable outlets using box-plots. For each of the 29 reliable outlets, we measure the percentage of health news and then use these 29 percentage values to draw the box-plot for the reliable outlets; likewise, for unreliable. We observe that the reliable outlets (median 72%) publish news on health topics comparatively less than unreliable outlets (median 85%).
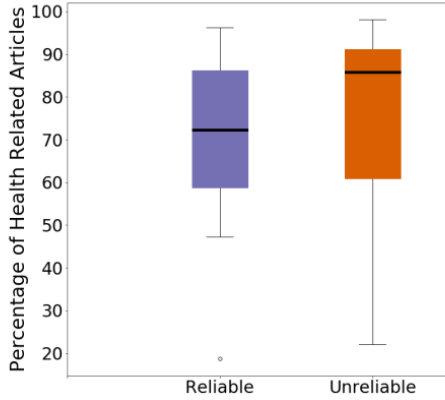

Figure 2– Comparison between reliable and unreliable outlets with respect to the presence of health-related news contents

Analysis

Using this dataset, we conduct content analysis to examine structural, topical, and semantic differences in health news from reliable and unreliable outlets.

Structural Difference

We particularly examine two structural elements- headlines and visual media of news articles.

Headline

The headline is a key element of a news article. According to a study done by American Press Institute and the Associated Press [26], only 4 out of 10 Americans read beyond the headline. So, it is important to understand how reliable and short headline does. We observe that the average headline unreliable outlets construct the headlines of their health-related news. According to Breaux [27], a longer headline results in significantly higher click-through-rate (CTR) than a length of an article from reliable outlets and an article from unreliable outlets is 8.56 words and 12.13 words, respectively. So, on average, an unreliable outlet's headline has a higher chance of receiving more clicks or attention than a reliable outlet's headline. To further investigate this, we examine the *clickbaitiness* of the headlines. The term clickbait refers to a form of web content (headline, image, thumbnail, etc.) that employs writing formulas, linguistic techniques, and suspense creating visual elements to trick readers into clicking links but does not deliver on its promises [28]. Chen et al. [29] reported that clickbait usage is a common pattern in false news articles. We investigate to what extent the reliable and unreliable outlets use clickbait headlines in their health articles. For each article headline, we test whether it is a clickbait or not using two supervised clickbait detection models– a sub-word embedding based deep learning model [30] and a feature engineering based Multinomial Naive Bayes model [31]. Agreement between these models was measured as 0.44 using Cohen's κ. We mark a headline as a clickbait if both models labeled it as clickbait. We observe, 27.29% (5,031 out of 18,436) of the headlines from reliable outlets are click bait. In unreliable outlets, the percentage is significantly higher, 40.03% (3,664 out of 9,153). So, it is evident that the unreliable outlets use more clickbaits than reliable outlets.

We further investigate the linguistic patterns used in the clickbait headlines. In particular, we analyze the presence of some common patterns which are generally employed in clickbait according to [27; 32]. The patterns are- 1) Presence of demonstrative adjectives (e.g., this, these, that). 2) Presence of numbers (e.g., 10, ten). 3) Presence of modal words (e.g., must, should, could, can). 4) Presence of question or WH words (e.g., what, who, how). 5) Presence of superlative words (e.g., best, worst, never). Figure 2 shows the distribution of these patterns among the clickbait headlines of reliable and unreliable outlets. Note, one headline may contain more than one pattern. For example, this headline *"Are these the worst 9 diseases in the world?"* contains four of the above patterns. This is the reason why summation of the percentages isn't equal to one. We see that unreliable outlets use demonstrative adjective and numbers significantly more compared to the reliable outlets.

Use of visual media

We examined how often the outlets use images in the articles. Our analysis finds that on average an article from reliable outlets uses 13.83 images and an article from unreliable outlets uses 14.22 images. Figure 3a shows density plots of the average number of images per article for both outlet categories. We observe that a good portion of unreliable outlet sourced articles uses a high number of images (more than 20).
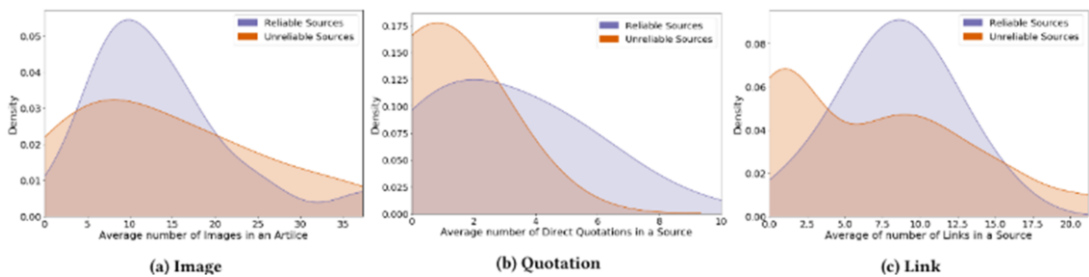

(a) Image   **(b) Quotation**   **(c) Link**

Figure 3–Distribution of average number of image/quotation/link per article from reliable and unreliable outlets

Figure 4– Topic modeling (k = 3) of articles from reliable outlets (top, denoted as RT) and from unreliable outlets (bottom, denoted as UT).

Topical Difference

The health domain is considerably broad and it covers many topics. We hypothesize that there are differences between the health topics in reliable and unreliable outlet articles. We test this hypothesis using an unsupervised & a supervised analysis.

Topic Modeling

We use *Latent Dirichlet Allocation(LDA)* algorithm to model the topics in the news articles. The number of topics, *k*, was set as 3. Figure 4 shows three topics for each of the outlet categories. Each topic is modeled by the top-10 important words in that topic. The font size of words is proportional to the importance. Figure 4a and 4d indicate that "cancer" is a common topic in reliable and unreliable outlets. Although, the words *study, said, percent, research,* and their font sizes in Figure 4a indicate that the topic "cancer" is associated with research studies, facts, and references in reliable outlets. On the contrary, unreliable outlets have the words *vaccine, autism,* and *risk* in Figure 4d which suggests the discussion regarding how vaccines put people under autism and cancer risk, an unsubstantiated claim, generally propagated by unreliable media[8,9]. Figure 4e and 4f suggest the discussions about weight loss, skin, and hair care products (e.g., essential oil, lemon). Topics in Figure 4b and 4c discuss mostly flu, virus, skin infection, exercise, diabetes and so on.
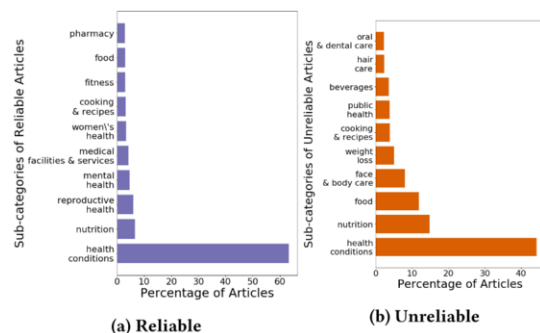

Figure 5 –Top-10 topics in reliable and unreliable outlets.

Topic Categorization

In addition to topic modeling, we categorically analyze the articles' topics using Google Cloud Natural Language API. Figure 5 shows the top-10 topics in the reliable and unreliable outlets. In the case of reliable, the distribution is significantly dominated by *health condition*. On the other hand, in the case of

unreliable outlets, percentages of *nutrition* and *food* are noticeable. Only 4 of the 10 categories are common in two outlet groups. Unreliable topics have *weight loss, hair care, face & body care*. This finding supports our claim from topic modeling analysis.

Semantic Difference

We analyze what efforts the outlets make for a logical and meaningful health news. Specifically, we consider to what extent the outlets use quotations and hyperlinks. Use of quotation and hyperlinks in a news article is associated with credibility [33; 34]. Presence of quotation and hyperlinks indicates that an article is logically constructed and supported with credible factual information.

Quotation

We use the Stanford QuoteAnnotator[10] to identify the quotations from a news article. Figure 3b shows density plots of the number of quotations per article for reliable and unreliable outlets. We observe that unreliable outlets use less number of quotations compared to reliable outlets. We find that the average number of quotations per article is 1 and 3 in unreliable and reliable outlets, respectively. This suggests that the reliable outlet sources articles are more credible and unreliable outlets are less credible.

Hyperlink

We examine the use of the hyperlink in the articles. On average, a reliable outlet sourced article contains 8.4 hyperlinks and an unreliable outlet sourced article contains 6.8 hyperlinks. Figure 3c shows density plots of the number of links per article for reliable and unreliable outlets. The peaks indicate that most of the articles from reliable outlets have close to 8 (median) hyperlinks. On the other hand, most of the unreliable outlet articles have less than 2 hyperlinks. This analysis again suggests that the reliable sourced articles are more credible than unreliable outlet articles.

Conclusion and Future Work

In this paper, we closely looked at structural, topical, and semantic differences between articles from reliable and unreliable outlets. Our findings reconfirm some of the existing claims such as unreliable outlets use clickbaity headlines to catch the attention of users. In addition, this study finds new patterns that can potentially help separate health disinformation. For example, we find that less quotation and hyperlinks are more associated with unreliable outlets. However, there are some limitations to this study. For instance, we didn't consider the videos, cited experts, comments of the users, and other information. In the future, we want to overcome these limitations and leverage the findings of this study to combat health disinformation.

References

[1] Katie Forster. (accessed October 30, 2018). Revealed: How dangerous fake health news conquered Facebook. https://www.independent.co.uk/life-style/healthhand-families/health-news/fake-news-health-facebook-cruel-damaging-socialmedia-mike-adams-natural-health-ranger-conspiracy-a7498201.html

[2] Mathew Ingram. (accessed October 30, 2018). The internet didn't invent viral content or clickbait journalism - there's just more of it now, and it happens faster. https://gigaom.com/2014/04/01/the-internet-didnt-invent-

---

[8]www.webmd.com/brain/autism/do-vaccines-cause-autism
[9]https://www.skepticalraptor.com/skepticalraptorblog.php/polio-vaccine-causescancer-myth/

[10] https://stanfordnlp.github.io/CoreNLP/quote.html

viral-contentor-clickbait-journalism-theres-just-more-of-it-now-and-it-happens-faster/

[3] Muiris Houston. (accessed October 31, 2018). Measles back with a vengeance due to fake health news. https://www.irishtimes.com/opinion/measles-back-with-avengeance-due-to-fake-health-news-1.3401960

[4] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. Science 359, 6380 (2018), 1094–1096.

[5] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter 19, 1 (2017), 22–36.

[6] Susannah Fox. (accessed October 30, 2018). The social life of health information. http://www.pewresearch.org/fact-tank/2014/01/15/the-social-life-ofhealth-information/

[7] Gaby Galvin. (accessed October 30, 2018). How Bots Could Hack Your Health. https://www.usnews.com/news/healthiest-communities/articles/2018-07-24/how-social-media-bots-could-compromise-public-health

[8] Amira Ghenai and Yelena Mejova. 2017. Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter. In Healthcare Informatics (ICHI), 2017 IEEE International Conference on. IEEE, 518–518

[9] Amira Ghenai and Yelena Mejova. 2018. Fake Cures: User-centric Modeling of Health Misinformation in Social Media. In 2018 ACM Conference on ComputerSupported Cooperative Work and Social Computing (CSCW). ACM.

[10] Patty Kostkova, Vino Mano, Heidi J Larson, and William S Schulz. 2016. Vac medi+ board: Analysing vaccine rumours in news and social media. In Proceedings of the 6th International Conference on Digital Health Conference. ACM.

[11] Emily K Vraga and Leticia Bode. 2017. Using Expert Sources to Correct Health Misinformation in Social Media. Science Communication 39, 5 (2017), 621–645.

[12] Nicole K Dalmer. 2017. Questioning reliability assessments of health information on social media. Journal of the Medical Library Association: JMLA 105,1.

[13] Irja Marije de Jong, Frank Kupper, Marlous Arentshorst, and Jacqueline Broerse. 2016. Responsible reporting: neuroimaging news in the age of responsible research and innovation. Science and engineering ethics 22, 4 (2016).

[14] Marjorie Kagawa-Singer and Shaheen Kassim-Lakha. 2003. A strategy to reduce cross-cultural miscommunication and increase the likelihood of improving health outcomes. Academic Medicine 78, 6 (2003), 577–587.

[15] Gary Schwitzer. 2008. How do US journalists cover treatments, tests, products, and procedures? An evaluation of 500 stories. PLoS medicine 5, 5 (2008), e95.

[16] Miriam Shuchman and Michael S Wilkes. 1997. Medical scientists and health news reporting: a case of miscommunication. Annals of Internal Medicine 126, 12 (1997), 976–982.

[17] Felisbela Lopes, Teresa Ruão, Zara Pinto Coelho, and Sandra Marinho. 2009. Journalists and health care professionals: what can we do about it?. In 2009 Annual Conference of the International Association for Media and Communication Research (IAMCR), "Human Rights and Communication". 1–15.

[18] Alexander Kinsora, Kate Barron, Qiaozhu Mei, and VG Vinod Vydiswaran. 2017. Creating a Labeled Dataset for Medical Misinformation in Health Forums. In Healthcare Informatics (ICHI), 2017 IEEE International Conference on. IEEE, 456–461.

[19] HEALTH ON THE NET. (accessed September 24, 2018). https://www.hon.ch/en/

[20] World Health Organization (WHO). (accessed September 24, 2018). http://www.who.int/

[21] Michael Schneider. (accessed September 24, 2018). Most-Watched Television Networks: Ranking 2016's Winners and Losers. https://www.indiewire.com/2016/12/cnn-fox-news-msnbc-nbratings-2016-winners-losers-1201762864/

[22] Wikipedia. (accessed September 24, 2018). Wikipedia:Zimdars' fake news list. https://bit.ly/2ziHafj

[23] Melissa Zimdars. 2016. My 'fake news list' went viral. But made-up stories are only part of the problem. https://www.washingtonpost.com/posteverything/wp/2016/11/18/my-fake-news-list-went-viral-but-made-up-stories-are-onlypart-of-the-problem. (2016).

[24] Wikipedia. (accessed September 24, 2018). List of fake news websites. https://bit.ly/2moBDvA

[25] informationisbeautiful.net. 2016. Unreliable/Fake News Sites & Sources. https://docs.google.com/spreadsheets/d/1xDDmbr54qzzG8wUrRdxQlC1dixJSIYqQUaXVZBqsJs

[26] American Press Institute and the Associated Press-NORC Center for Public Affairs Research. (accessed September 28, 2018). The Personal News Cycle: How Americans choose to get their news. https://www.americanpressinstitute.org/publications/reports/survey-research/how-americans-get-news/

[27] Chris Breaux. (accessed September 28, 2018). "You'll Never Guess How Chartbeat's Data Scientists Came Up With the Single Greatest Headline". http://blog.chartbeat.com/2015/11/20/youll-never-guess-how-chartbeatsdata-scientists-came-up-with-the-single-greatest-headline

[28] Bryan Gardiner. (accessed September 28, 2018). "You'll Be Outraged at How Easy It Was to Get You to Click on This Headline".

[29] 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. ACM, 15–19.

[30] 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? In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. ACM, 232–239

[31] Saurabh Mathur. (accessed September 24, 2018). Clickbait Detector. https://github.com/saurabhmathur96/clickbait-detector

[32] Matthew Opatrny. (accessed September 28, 2018). "9 Headline Tips to Help You Connect with Your Target Audience". https://www.outbrain.com/blog/9-headlinetips-to-help-marketers-and-publishers-connect-with-their-target-audiences/

[33] Juliette De Maeyer. 2012. The journalistic hyperlink: Prescriptive discourses about linking in online news. Journalism Practice 6, 5-6 (2012), 692–701. https://www.wired.com/2015/12/psychology-of-clickbait/

[34] S Shyam Sundar. 1998. Effect of source attribution on perception of online news stories. Journalism & Mass Communication Quarterly 75, 1 (1998), 55–68.