Clickbait Detection

Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen

Bauhaus-Universität Weimar

<first name>.<last name>@uni-weimar.de

Abstract This paper proposes a new model for the detection of clickbait, i.e., short messages that lure readers to click a link. Clickbait is primarily used by online content publishers to increase their readership, whereas its automatic detection will give readers a way of filtering their news stream. We contribute by compiling the first clickbait corpus of 2992 Twitter tweets, 767 of which are clickbait, and, by developing a clickbait model based on 215 features that enables a random forest classifier to achieve 0.79 ROC-AUC at 0.76 precision and 0.76 recall.

1 Introduction

Clickbait refers to a certain kind of web content advertisement that is designed to entice its readers into clicking an accompanying link. Typically, it is spread on social media in the form of short teaser messages that may read like the following examples:

- A Man Falls Down And Cries For Help Twice. The Second Time, My Jaw Drops
- 9 Out Of 10 Americans Are Completely Wrong About This Mind-Blowing Fact
- Here's What Actually Reduces Gun Violence

When reading such and similar messages, many get the distinct impression that something is odd about them; something unnamed is referred to, some emotional reaction is promised, some lack of knowledge is ascribed, some authority is claimed. Content publishers of all kinds discovered clickbait as an effective tool to draw attention to their websites. The level of attention captured by a website determines the prize of displaying ads there, whereas attention is measured in terms of unique page impressions, usually caused by clicking on a link that points to a given page (often abbreviated as "clicks"). Therefore, a clickbait's *target link* alongside its *teaser message* usually redirects to the sender's website if the reader is afar, or else to another page on the same site. The content found at the linked page often encourages the reader to share it, suggesting clickbait for a default message and thus spreading it virally. Clickbait on social media has been on the rise in recent years, and even some news publishers have adopted this technique. These developments have caused general concern among many outspoken bloggers, since clickbait threatens to clog up social media channels, and since it violates journalistic codes of ethics.

In this paper, we present the first approach to automatic clickbait detection. Our contributions are twofold: (1) we collect and annotate the first publicly available clickbait corpus of 3000 Twitter tweets, sampled from the top Twitter publishers, and (2) we develop and evaluate the first clickbait detection model. After discussing related work in Section 2, Section 3 reports on corpus construction, Section 4 on our clickbait model, and Section 5 on its evaluation.

2 Related Work

The rationale why clickbait works is widely attributed to teaser messages opening a so-called "curiosity gap," increasing the likelihood of readers to click the target link to satisfy their curiosity. Loewenstein's information-gap theory of curiosity [19] is frequently cited to provide a psychological underpinning (p. 87): "the information-gap theory views curiosity as arising when attention becomes focused on a gap in one's knowledge. Such information gaps produce the feeling of deprivation labeled curiosity. The curious individual is motivated to obtain the missing information to reduce or eliminate the feeling of deprivation." Loewenstein identifies stimuli that may spark involuntary curiosity, such as riddles or puzzles, event sequences with unknown outcomes, expectation violations, information possessed by others, or forgotten information. The effectiveness by which clickbait exploits this cognitive bias results from data-driven optimization. Unlike with printed front page headlines, for example, where feedback about their potential contribution to newspaper sales is indirect, incomplete, and delayed, clickbait is optimized in real-time, recasting the teaser message to maximize click-through [16]. Some companies allegedly rely mostly on clickbait for their traffic. Their success on social networks recently caused Facebook to take action against clickbait as announced by El-Arini and Tang [8]. Yet, little is known about Facebook's clickbait filtering approach; no corresponding publications have surfaced. El-Arini and Tang's announcement mentions only that context features such as dwell time on the linked page and the ratio of clicks to likes are taken into account.

To the best of our knowledge, clickbait has been subject to research only twice to date, both times by linguists: first, Vijgen [26] studies articles that compile lists of things, so-called "listicles." Listicles are often under suspicion to be clickbait. The authors study 720 listicles published at BuzzFeed in two weeks of January 2014, which made up about 30% of the total articles published in this period. The titles of listicles, which are typically shared as teaser messages, exert a very homogeneous structure: all titles contain a cardinal number—the number of items listed—and 85% of the titles start with it. Moreover, these titles contain strong nouns and adjectives to convey authority and sensationalism. Moreover, the main articles consistently achieve easy readability according to the Gunning fog index [10]. Second, Blom and Hansen [3] study phoricity in headlines as a means to arouse curiosity. They analyze 2000 random headlines from a Danish news website and identify two common forms of forward-references: discourse deixis and cataphora. The former are references at discourse level ("*This* news will blow your mind."), and the latter at phrase level ("*This* name is hilarious."). Based on a dictionary of basic deictic and cataphoric expressions, the share of such phoric expressions at 10 major Danish news websites reveals that they occur mostly in commercial, ad-funded, and tabloid news websites. However, no detection approach is proposed.

Besides, some dedicated individuals have taken the initiative: Gianotto [9] implements a browser plugin that transcribes clickbait teaser messages based on a rule set so that they convey a more "truthful," or rather ironic meaning. We employ the rule set premises as features and as a baseline for evaluation. Beckman [2], Mizrahi [20], Stempeck [24], and Kempe [15] manually re-share clickbait teaser messages, adding spoilers. Eidnes [7] employs recurrent neural networks to generate nonsense clickbait for fun.

3 A Twitter Clickbait Corpus

To sample our corpus, we focus on Twitter as a social media platform used by many content publishers. To obtain an unbiased choice of publishers, we sample from the top 20 most prolific publishers on Twitter as determined by their influence in terms of re-tweets. Table 1 (left) overviews these publishers. Well-known English-speaking newspapers are

Table 1. Left: Top 20 publishers on Twitter according to NewsWhip [21] in 2014. The darker a cell, the more prolific the publisher; white cells indicate missing data. Right: Our clickbait corpus in terms of tweets with links posted in week 24, 2015, tweets sampled for manual annotation, and tweets labeled as clickbait (absolute and relative) by majority vote of three assessors.

Publisher	**Twitter re-tweets** ($\times 10^6$)												Σ	**Clickbait corpus**			
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	2014	Tweets	Sample	Clickbait	
BBC News	2.70	2.48	2.71	2.87	3.12	3.25	3.56	3.39	3.79	4.02	3.96	3.75	39.6	694	150	25	17%
New York Times		1.28	1.84	2.00	2.11	2.28	2.48	2.35	2.42	2.60	2.22	2.18	23.8	875	150	32	21%
Mashable	1.42	1.46	1.66	1.83	1.77	1.83	1.95	1.66	1.86	1.82	1.78	1.60	20.6	803	150	49	33%
ABC News	0.79	1.15	1.76	1.56	1.62	1.80	1.91	1.68	1.67	1.36	1.28	1.06	17.6	279	150	13	9%
CNN	1.18	1.16	1.21	1.25	1.25	1.17	1.39	1.31	1.35	1.53	1.27	0.97	15.0	345	150	25	17%
The Guardian		1.07	1.16	1.13	1.23	1.32	1.42	1.27	1.37	1.51	1.35	1.19	14.0	744	150	22	15%
Huffington Post	0.96	0.72	0.85	0.77	0.83	0.93	1.14	1.14	1.12	1.17	1.02	0.90	11.6	770	150	69	46%
Forbes	0.80	0.81	0.96	0.89	0.96	1.03	1.11	1.13	1.18	1.10	0.78	0.75	11.5	721	150	57	38%
Bleacher Report	0.58	0.57	0.68	0.72	0.74	0.84	0.84	0.83	1.09	1.12	1.18	1.04	10.2	196	150	13	9%
Fox News	0.59	0.54	0.68	0.82	0.83	0.92	0.95	0.96	0.97	1.04	0.99	0.87	10.2	378	150	12	8%
BuzzFeed	0.76	0.80	0.81	0.84	0.74	0.74	0.99	0.85	0.86	0.90	0.86	0.82	10.0	695	150	63	42%
NBC News	0.60	0.64	0.78	0.75	0.72	0.75	0.86	0.89	0.95	0.95	0.93	0.87	9.7	408	150	21	14%
Yahoo!		0.57	0.66	0.59	0.67	0.77	0.90	1.00	0.84	0.82	0.77	0.58	8.2	195	150	34	23%
Daily Mail	0.51	0.48	0.51	0.54		0.65	0.73	0.67	0.75	0.69	0.79	0.63	6.9	516	150	33	22%
ESPN	0.44	0.40	0.50	0.52	0.53	0.49	0.58	0.53	0.76	0.78	0.69	0.67	6.9	142	142	34	24%
Wall Street Journal		0.56	0.64	0.63	0.66	0.68	0.81	0.77	0.88	0.90			6.5	747	150	28	19%
Business Insider	0.46	0.50	0.54	0.49	0.47	0.50	0.55	0.52	0.59	0.71	0.56	0.63	6.5	779	150	76	51%
The Telegraph		0.50	0.59				0.89	0.89	0.90	0.94	0.91	0.82	6.4	699	150	32	21%
Washington Post		0.39	0.40	0.41	0.42	0.51	0.62	0.73	0.75	0.80	0.70	0.64	6.4	691	150	62	41%
The Independent	0.34			0.39	0.46	0.45	0.69	0.67	0.60	0.75	0.82	0.67	5.8	530	150	67	45%
													Σ	11207	2992	767	26%

among them, but also publishers which have been pointed out for making excessive use of clickbait, including Business Insider [11], the Huffington Post [20], and BuzzFeed [1]; BuzzFeed has publicly opposed the allegations [23]. BBC News has been the most prolific publisher throughout 2014, increasing their number of re-tweets steadily from 2.7 million in January to more than 3.7 million in December for a total of 39.6 million. The New York Times comes in second with a total of 23.8 million. On third rank, the online-only news publisher Mashable is listed, showing that these companies compete with traditional media.

For our corpus, we collected tweets sent by the publishers in week 24 of 2015 that included links, as shown in Table 1 (right). We randomly sampled 150 tweets per publisher for a total of 2992 tweets (one publisher sent only 142 tweets in that time). Each tweet was annotated independently by three assessors who rated them being clickbait or not. Judgments were made only based on the tweet's plain text and image (i.e., the teaser message), and not by clicking on links. We obtain a "fair" inter-annotator agreement with a Fleiss' κ of 0.35. Taking the majority vote as ground truth, a total of 767 tweets (26%) are considered clickbait. Table 1 (right, column "Clickbait") shows the distribution of clickbait across publishers. According to our annotation, Business Insider sends 51% clickbait, followed by Huffington Post, The Independent, BuzzFeed, and the Washington Post with more than 40% each. Most online-only news publishers (Business Insider, Huffington Post, BuzzFeed, Mashable) send at least 33% clickbait, Bleacher Report being the only exception with a little less than 10%. TV networks (CNN, NBC, ABC, Fox) are generally at the low end of the distribution. Altogether, these figures suggest that all of the top 20 news publishers employ clickbait on a regular basis, supporting the allegations raised by bloggers.

4 Clickbait Detection Model

Our clickbait detection model is based on 215 features; Table 2, column "Feature (type)," gives an overview. The features divide into three categories pertaining to (1) the teaser message, (2) the linked web page, and (3) meta information.

(1) Teaser message. Our primary feature engineering focus is on capturing the characteristics of a clickbait's teaser message, which is why most features are in this category. We subdivide the teaser message features into three subcategories: the first subcategory (1a) comprises basic text statistics. Features 1-9 are bag-of-words features, where Features 7 and 8 are Twitter-specific and Feature 9 consists of automatically generated image tags for images sent as part of a tweet, obtained from the Imagga tagging service [13]. Feature 10 computes the sentiment polarity of a tweet using the Stanford NLP library, and Features 11-13 measure a tweet's readability, where Features 12 and 13 are based on the Terrier stop word list [22] and the Dale-Chall list of easy words [5]. Features 14-16 quantify contractions and punctuation use, and Features 17-19 length statistics. The second and third subcategory (1b) and (1c) of teaser message features comprise dictionary features, where each feature encodes whether or not a tweet contains a word from a given dictionary of specific words or phrases. Features 20 and 21 are two dictionaries obtained from Gianotto [9], where the first contains common clickbait phrases and the second clickbait patterns in the form of regular expressions. Finally, Features 22-203 are all 182 General Inquirer dictionaries [25].

(2) Linked web page. Analyzing the web pages linked from a tweet, Features 204-209 are again bag-of-words features, whereas Features 210 and 211 measure readability and length of the main content when extracted with Boilerpipe [17].

(3) Meta information. Feature 212 encodes a tweet's sender, Feature 213 whether media (e.g., an image or a video) has been attached to a tweet, Feature 214 whether a tweet has been retweeted, and Feature 215 the part of day in which the tweet was sent (i.e., morning, afternoon, evening, night)

5 Evaluation

We randomly split our corpus into datasets for training and testing at a 2:1 training-test ratio. To avoid overfitting, we discard all features that have non-trivial weights in less than 1% of the training dataset only. The features listed in Table 2 remained, whereas many individual features from the bag-of-words feature types were discarded (see the feature IDs marked with a *). Before training our clickbait detection model, we balance the training data by oversampling clickbait. We compare the three well-known learning algorithms logistic regression [18], naive Bayes [14], and random forest [4] as implemented in Weka 3.7 [12] using default parameters. To assess detection performance, we measure precision and recall for the clickbait class, and the area under the curve (AUC) of the receiver operating characteristic (ROC). We evaluate the performance of all features combined, each feature category on its own, and each individual feature (type) in isolation. Table 2 shows the results.

All features combined achieve a ROC-AUC of 0.74 with random forest, 0.72 with logistic regression, and 0.69 with naive Bayes. The precision scores on all features do not differ much across classifiers, the recall ranges from 0.66 with naive Bayes to 0.73 with random forest. Interestingly, the teaser message features (1a) alone compete or even outperform all features combined in terms of precision, recall, and ROC-AUC, using naive Bayes and random forest. The character n-gram features and the word 1-gram feature (IDs 1-4) appear to contribute most to this performance. Character n-grams are known to capture writing style, which may partly explain their predictive power for clickbait. The other features from category (1a) barely improve over chance as measured by ROC-AUC, yet, some at least achieve high precision, recall, or both. We further employ feature selection based on the χ^2 test to study the dependency of performance on the number of high-performing features. Selecting the top 10, 100, and 1000 features, overall performance with random forest outperforms that of feature category (1a) with 0.79 ROC-AUC. Features from all categories are selected, but mostly n-gram features from the teaser message and the linked web page.

Table 2. Evaluation of our clickbait detection model. Some features are feature types that expand to many individual frequency-weighted features (i.e., IDs 1-9 and IDs 204-209). As classifiers, we evaluate logistic regression (LR), naive Bayes (NB), and random forest (RF).

Feature (type)		Precision			Recall			ROC-AUC		
ID	Description	LR	NB	RF	LR	NB	RF	LR	NB	RF
	all features	0.70	0.71	0.70	0.70	0.66	0.73	0.72	0.69	0.74
	top 10 as per χ^2 ranking	0.70	0.70	0.68	0.67	0.72	0.65	0.71	0.70	0.66
	top 100 as per χ^2 ranking	0.71	0.72	0.72	0.65	0.65	0.71	0.73	0.72	0.76
	top 1000 as per χ^2 ranking	0.64	0.70	0.76	0.58	0.65	0.76	0.60	0.69	0.79
(1a) *Teaser message*		0.60	0.74	0.71	0.55	0.72	0.73	0.54	0.74	0.73
1*	character 1-grams	0.71	0.68	0.71	0.65	0.56	0.71	0.72	0.68	0.71
2*	character 2-grams	0.64	0.73	0.71	0.60	0.70	0.72	0.60	0.75	0.74
3*	character 3-grams	0.63	0.74	0.74	0.58	0.74	0.75	0.61	0.76	0.77
4*	word 1-grams	0.70	0.74	0.72	0.66	0.66	0.71	0.70	0.76	0.72
5*	word 2-grams	0.64	0.63	0.61	0.68	0.45	0.46	0.58	0.58	0.55
6*	word 3-grams	0.55	0.55	0.55	0.69	0.69	0.69	0.50	0.50	0.50
7	hashtags	0.64	0.65	0.65	0.32	0.32	0.32	0.50	0.49	0.50
8	@ mentions	0.71	0.72	0.71	0.37	0.37	0.37	0.53	0.53	0.53
9	image tags as per Imagga [13]	0.55	0.59	0.57	0.41	0.50	0.51	0.48	0.52	0.51
10	sentiment polarity (Stanford NLP)	0.64	0.64	0.64	0.57	0.57	0.57	0.58	0.58	0.58
11	readability (Flesch-Kincaid)	0.63	0.63	0.61	0.54	0.55	0.54	0.59	0.59	0.56
12	stop words-to-words ratio	0.67	0.67	0.60	0.59	0.62	0.48	0.65	0.65	0.57
13	easy words-to-words ratio	0.09	0.09	0.49	0.30	0.30	0.70	0.50	0.50	0.50
14	has abbreviations	0.57	0.57	0.55	0.48	0.59	0.47	0.50	0.48	0.47
15	number of dots	0.63	0.64	0.63	0.42	0.37	0.42	0.54	0.54	0.54
16	starts with number	0.72	0.72	0.72	0.72	0.72	0.72	0.55	0.55	0.55
17	length of longest word	0.62	0.61	0.60	0.49	0.55	0.44	0.57	0.57	0.55
18	mean word length	0.58	0.57	0.61	0.51	0.56	0.56	0.50	0.48	0.54
19	length in characters	0.67	0.64	0.64	0.59	0.61	0.56	0.62	0.62	0.58
(1b) *Teaser message: Downworthy*		0.64	0.64	0.64	0.69	0.69	0.69	0.54	0.54	0.54
20	common clickbait phrases	0.65	0.65	0.65	0.70	0.70	0.70	0.54	0.54	0.54
21	common clickbait patterns	0.58	0.58	0.58	0.69	0.69	0.69	0.50	0.50	0.50
(1c) *Teaser message: General Inquirer (GI)*		0.66	0.70	0.67	0.60	0.64	0.67	0.65	0.68	0.70
22	GI dict. You	0.71	0.71	0.71	0.73	0.73	0.73	0.58	0.58	0.58
23	GI dict. POLIT	0.63	0.70	0.63	0.52	0.44	0.52	0.58	0.58	0.58
24	GI dict. Intrj	0.67	0.67	0.67	0.71	0.71	0.71	0.57	0.57	0.57
25	GI dict. HU	0.63	0.66	0.63	0.52	0.40	0.52	0.57	0.57	0.57
26	GI dict. Space	0.63	0.62	0.62	0.56	0.55	0.52	0.57	0.56	0.56
27	GI dict. Understated	0.64	0.58	0.65	0.67	0.32	0.69	0.56	0.55	0.56
28	GI dict. PowTot	0.65	0.65	0.65	0.49	0.49	0.49	0.59	0.59	0.56
...	+ 175 further GI dictionaries									
(2) *Linked web page*		0.64	0.64	0.64	0.67	0.67	0.67	0.56	0.56	0.56
204*	character 1-grams	0.63	0.57	0.62	0.55	0.62	0.61	0.60	0.54	0.61
205*	character 2-grams	0.58	0.62	0.61	0.49	0.62	0.62	0.50	0.59	0.61
206*	character 3-grams	0.58	0.67	0.62	0.49	0.59	0.64	0.52	0.63	0.61
207*	word 1-grams	0.60	0.72	0.65	0.50	0.64	0.65	0.54	0.71	0.64
208*	word 2-grams	0.58	0.70	0.67	0.48	0.60	0.65	0.51	0.70	0.66
209*	word 3-grams	0.56	0.65	0.63	0.44	0.55	0.58	0.46	0.61	0.63
210	main content readability (Flesch-Kincaid)	0.57	0.59	0.60	0.59	0.61	0.54	0.45	0.54	0.55
211	main content word length	0.61	0.63	0.58	0.56	0.68	0.50	0.54	0.56	0.51
(3) *Meta information*		0.62	0.74	0.74	0.55	0.72	0.75	0.54	0.74	0.77
212	sender name	0.65	0.65	0.65	0.60	0.60	0.60	0.67	0.67	0.67
213	has media attachment	0.55	0.55	0.55	0.53	0.53	0.53	0.47	0.47	0.47
214	is retweet	0.60	0.60	0.60	0.39	0.39	0.39	0.51	0.51	0.51
215	part of day as per server time	0.60	0.60	0.60	0.53	0.53	0.53	0.51	0.51	0.51

Finally, as a baseline for comparison, the Downworthy rule sets [9] achieve about 0.69 recall at about 0.64 precision, whereas their ROC-AUC is only 0.54. This baseline is not only outperformed by combinations of other features, but also individual features, such as the General Inquirer dictionary "You" (9 pronouns indicating another person is being addressed directly) as well as several others. Furthermore, sentiment analysis alone appears to be insufficient to detect clickbait (Feature 10), whereas in feature combinations it does possess some predictive power.

6 Conclusion

This paper presents the first machine learning approach to clickbait detection: the goal is to identify messages in a social stream that are designed to exploit cognitive biases to increase the likelihood of readers clicking an accompanying link. Clickbait's practical success, and the resulting flood of clickbait in social media, may cause it to become another form of spam, clogging up social networks and being a nuisance to its users. The adoption of clickbait by news publishers is particularly worrisome. Automatic clickbait detection would provide for a solution by helping individuals and social networks to filter respective messages, and by discouraging content publishers from making use of clickbait. To this end, we contribute the first evaluation corpus as well as a strong baseline detection model. However, the task is far from being solved, and our future work will be on contrasting clickbait between different social media, and improving detection performance.

Bibliography

[1] Saif Ajani. A Full 63% of BuzzFeed's Posts are Clickbait. http://keyhole.co/blog/buzzfeed-clickbait/, 2015.

[2] Jake Beckman. Saved You A Click—Don't click on that. I already did. https://twitter.com/savedyouaclick, 2015.

[3] Jonas Nygaard Blom and Kenneth Reinecke Hansen. Click Bait: Forward-reference as Lure in Online News Headlines. *Journal of Pragmatics*, 76:87–100, 2015.

[4] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[5] Joseph Rocca. Dale-Chall Easy Word List. http://countwordsworth.com/download/DaleChallEasyWordList.txt, 2013.

[6] Jesse Davis and Mark Goadrich. The Relationship Between Precision-Recall and ROC curves. In *Proc. of ICML'06*, pages 233–240.

[7] Lars Eidnes. Auto-Generating Clickbait With Recurrent Neural Networks. http://larseidnes.com/2015/10/13/auto-generating-clickbait-with-recurrent-neural-networks/, 2015.

[8] Khalid El-Arini and Joyce Tang. News Feed FYI: Click-baiting. http://newsroom.fb.com/news/2014/08/news-feed-fyi-click-baiting/, 2014.

[9] Alison Gianotto. Downworthy—A Browser Plugin to Turn Hyperbolic Viral Headlines Into What They Really Mean. http://downworthy.snipe.net, 2014.

[10] Robert Gunning. The Fog Index After Twenty Years. *J. Bus. Commun.*, 6(2):3–13, 1969.

[11] Keach Hagey. Henry Blodget's Second Act. http://www.wsj.com/articles/SB10000872396390444840104577555180608254796, 2011.

[12] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18, July 2009.

[13] Imagga Image Tagging Technology. http://imagga.com, 2015.

[14] George H. John and Pat Langley. Estimating Continuous Distributions in Bayesian Classifiers. In *Proc. of UAI'95*, pages 338–345.

[15] Rebecca Kempe. Clickbait Spoilers—Channeling Traffic from Clickbaiting Sites Back to Reputable Providers of Original Content. http://www.clickbaitspoilers.org, 2015.

[16] Peter Koechley. Why The Title Matters More Than The Talk. http://blog.upworthy.com/post/26345634089/why-the-title-matters-more-than-the-talk, 2012.

[17] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate Detection Using Shallow Text Features. In *Proc. of WSDM'10*, pages 441–450.

[18] Saskia le Cessie and Johannes C. van Houwelingen. Ridge Estimators in Logistic Regression. *Applied Statistics*, 41(1):191–201, 1992.

[19] George Loewenstein. The Psychology of Curiosity: A Review and Reinterpretation. *Psychological Bulletin*, 116(1):75, 1994.

[20] Alex Mizrahi. HuffPo Spoilers—I Give In To Click-bait So You Don't Have To. https://twitter.com/huffpospoilers, 2015.

[21] NewsWhip Media Tracker. http://www.newswhip.com, 2015.

[22] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *OSIR @ SIGIR 2006*.

[23] Ben Smith. Why BuzzFeed Doesn't Do Clickbait. http://www.buzzfeed.com/bensmith/why-buzzfeed-doesnt-do-clickbait, 2015.

[24] Matt Stempeck. Upworthy Spoiler—Words That Describe the Links That Follow. https://twitter.com/upworthyspoiler, 2015.

[25] Philip J. Stone, Dexter C. Dunphy, and Marshall S. Smith. *The General Inquirer: A Computer Approach to Content Analysis*. MIT press, 1966.

[26] Bram Vijgen. The Listicle: An Exploring Research on an Interesting Shareable New Media Phenomenon. *Studia Universitatis Babes-Bolyai-Ephemerides*, (1):103–122, 2014.