Crowdsourcing a Large Corpus of Clickbait on Twitter

Martin Potthast[1] **Tim Gollub**[2] **Kristof Komlossy**[2] **Sebastian Schuster**[2]
Matti Wiegmann[2,3] **Erika Patricia Garces Fernandez**[2] **Matthias Hagen**[4] **Benno Stein**[2]

[1]Leipzig University [2]Bauhaus-Universität Weimar
[3]German Aerospace Center (DLR) [4]Martin-Luther-Universität Halle-Wittenberg

Abstract

Clickbait has become a nuisance on social media. To address the urging task of clickbait detection, we constructed a new corpus of 38,517 annotated Twitter tweets, the Webis Clickbait Corpus 2017. To avoid biases in terms of publisher and topic, tweets were sampled from the top 27 most retweeted news publishers, covering a period of 150 days. Each tweet has been annotated on 4-point scale by five annotators recruited at Amazon's Mechanical Turk. The corpus has been employed to evaluate 12 clickbait detectors submitted to the Clickbait Challenge 2017.

Download: https://webis.de/data/webis-clickbait-17.html Challenge: https://clickbait-challenge.org

1 Introduction

For publishers of online content, the most popular revenue model is advertising. On freely accessible web pages ads are displayed along the content, and publishers earn fees for each ad seen by the visitors of these pages. Online content is often expected "to pay for itself:" it has to attract enough visitors such that the earned fees exceed the investment made in creating it. Assuming that each piece of content has its target audience, generating profit boils down to letting the respective target audience know where to find it. Ironically, this requires the content itself to be advertised.

Notwithstanding other means of advertisement, today, most publishers advertise their content on social media. By spreading "teaser messages" (be it with or without paying promotion fees), publishers engage in viral marketing. Teaser messages consist of two or three parts: (1) a short text, (2) an optional media attachment such as an image or a video, and (3) a link to the publisher's page where the advertised content is found. From a publisher's perspective, the ideal teaser message discloses its advertised content to the least possible extent, so that its target audience (better: everyone) is tempted to visit the publisher's pages. Conversely, from a reader's perspective, the ideal teaser message is a self-contained summary, so that visiting the publisher's page is subject only to cases where not everything important could be fitted into the summary, or where the reader wants to learn more about the backgrounds. The two ideals are naturally at odds, and the social network operators may need to reconcile between them.

Figure 1 shows examples of teaser messages in the above sense, advertising news content on Twitter. The left-to-right ordering of the messages in the figure is indicative of the degree to which an average reader might feel sufficiently informed without clicking the link (left), less informed but rather curious and urged to click (right), or something in-between (middle). The messages more to the right side of the figure are examples of what is called "clickbait" today. Clickbait is an umbrella term that encompasses all kinds of teaser messages in social media capable of instigating an (on average) increased click-through.

This paper introduces the Webis Clickbait Corpus 2017, a large-scale corpus of teaser messages, which has been built to address automatic clickbait detection. All messages in the corpus have been carefully sampled and annotated via crowdsourcing regarding their degree of clickbaitiness. After reviewing related work in Section 2, Section 3 defines clickbait and its operationalization, and Sections 4 and 5 detail the corpus construction process, including a first corpus analysis.

Figure 1: Examples of teaser messages from news publishers on Twitter. The tweets have been redacted, removing information identifying the publishers in order not to pillory them.

2 Related Work

To the best of our knowledge, Table 1 lists all clickbait-related resources available to date. The Webis Clickbait Corpus 2017 (Webis-Clickbait-17, for short) was preceded by the Webis Clickbait Corpus 2016 built by Potthast et al. (2016) to study clickbait detection for the first time. Given that clickbait itself is a fairly recent subject of inquiry, all of these resources have been published within the past two years. The resources can be distinguished in terms of genre, teaser type, approaches to acquisition and annotation, whether the linked articles are archived, and their size. A discussion of each aspect follows.

Genre. All resources to date, including ours, cover news. News are of particular importance, since the use of clickbait in this domain raises concerns: rather than maximizing the informative value of the teaser messages employed for news dissemination, news publishers also misuse them to maximize their ad revenue. Because of their important societal role, though, news publishers must be held to high standards, and the self-imposed codes of conduct of major news publishers clearly reject manipulative means to increase readership. Clickbait has been dismissed as "digital yellow journalism," i.e., as nothing new or noteworthy,[1] but this view is short-sighted: our previous work shows that nearly 25% of a sample of 2,992 tweets shared by the top-20 most retweeted news publishers are clickbait (Potthast et al., 2016)—a finding which is now confirmed by our 10 times larger corpus. Besides news, clickbait does occur in other genres, too. For example, on entertainment platforms such as YouTube (Qu et al., 2018). In fact, all kinds of social media where users are referred to another web page are susceptible to clickbait.

Teaser type. Two teaser types have been covered for corpus construction so far, namely the headline of a news article, and tweets about news articles. Headlines must be considered an approximation of an actual teaser message on social media; they are often used as textual component of a teaser, but not necessarily verbatim, and typically in combination with media attachments. The corpora using headlines lack associated media attachments, however. Only our previous corpus (Potthast et al., 2016) and the one presented in this paper incorporate realistic teaser messages. The teasers found on platforms like Facebook, LinkedIn, YouTube, and the like, are all similar to tweets, but differ in terms of allowed text length, media arrangement, and available meta data. Other teaser types are found on portal pages of websites where users are referred to more specific content found elsewhere (e.g., consider the teaser messages found on a news publisher's front page). Yet others may be found in discussion forums, comment boards, and emails, where more free-form text is allowed. However, when used as clickbait, these kinds of messages more often resemble spam, which is distinctly different from clickbait, but shares the goal of luring readers to their sender's web pages. One may expect that the clickbait message style will eventually become part of the repertoire of spammers.

Acquisition approach. Key to the representativeness of a language resource is the corpus acquisition strategy applied. Here, the news domain presents difficulties for sampling, since an average researcher hardly has access to all news published in a given period of time, so that directly sampling from the population of news is infeasible. To circumvent this issue, three strategies have been devised for clickbait corpus acquisition, namely reputation-based, gatekeeper-based, and importance-based acquisition.

With reputation-based acquisition, teaser messages (or news articles) are selected from publishers with a good reputation to obtain a relatively clean sample of non-clickbait, and from publishers with correspondingly bad reputation to obtain as much clickbait as possible. For instance, according to the

[1] http://theconcourse.deadspin.com/shut-up-about-clickbait-1551902024

Publication	Genre	Teaser	Acquisition	Annotation (Scale, People, κ)			Articles	Size
Agrawal (2016)	News	Headline	Gatekeeper-based	Binary	3	0.84	No	2,388
Potthast et al. (2016)	News	Tweet	Importance-based	Binary	3	0.35	Yes	2,992
Biyani et al. (2016)	News	Headline	Reputation-based?	Binary	?	?	?	4,073
Chakraborty et al. (2016)	News	Headline	Reputation-based	Binary	3	0.79	No	15,000
Rony et al. (2017)	News	Headline	Reputation-based	Binary	3	0.79?	No	32,000
Webis-Clickbait-17	News	Tweet	Importance-based	Graded	5	0.36	Yes	38,517

Table 1: Overview of clickbait corpora. All are in English; question marks (?) indicate lack of clarity.

literature, the latter publishers include BuzzFeed, Huffington Post, Upworthy, ViralNova, Scoopwhoop, and ViralStories. This approach has been applied by Chakraborty et al. (2016) whose corpus formed the basis for the one of Rony et al. (2017). The description of Biyani et al. (2016) is very superficial, but they do mention certain publishers, and the fact that almost half of their teasers are clickbait suggests that they have actively searched for such messages. Gatekeeper-based acquisition makes use of, e.g., web forums where news are curated. Agrawal (2016) obtain their selection of news from Reddit forums, where one forum pre-selects news of interest for the benefit of users, and where another collects clickbait to "spoil" it, e.g., by giving away the information withheld from a teaser. Both, reputation-based and gatekeeper-based acquisition, directly induce a ground truth. However, especially the reputation-based approach must be taken with a grain of salt, since many publishers with a comparably low reputation, such as BuzzFeed and Huffington Post, still publish a significant amount of news without resorting to clickbait teaser messages. Rony et al. (2017), Chakraborty et al. (2016), and Agrawal (2016) report to have double-checked the induced labels; each instance has been reviewed by at least three student volunteers. But the substantial inter-annotator agreements must be taken with a grain of salt, too, since as Agrawal (2016) reports, "the reason for such a high value of inter-assessor agreement is due to the fact that these headlines are already in their well defined categories owing to the nature of data collection." This conscientious observation hints at a bias that has been introduced by reputation-based acquisition and *not* disguising the sources or at least randomizing the messages before review.

By comparison, under importance-based acquisition as applied in (Potthast et al., 2016) as well as for our new corpus, publishers are ranked with respect to a specific criterion, such as sales or shares, and teasers are sampled from the top-ranked publishers. This way, it is ensured that a corpus covers the news seen by the largest portion of the target audience. Moreover, no assumptions are made about the nature of a publisher, so that the proportion of clickbait observed is not artificially amplified as is the case with the aforementioned acquisition approaches. Importance-based acquisition does not induce a ground truth, so that all teasers collected must be reviewed and annotated.

Annotation approach. Previous clickbait corpora consider only a binary scale for annotation, where a teaser is either clickbait or not. In our own annotation experiments, we found the binary scale to be lacking, since, as illustrated by the examples depicted in Figure 1, the degree to which a certain teaser message tries to manipulate its reader varies depending on the person reading it. For example, it may be that one person is hardly affected by any of these examples, whereas another more strongly experiences an urge to click. This is also why a single opinion does not suffice to arrive at an objective judgment about a teaser message, but a number of opinions should be collected and averaged.

Article archival. Teaser messages do not occur in isolation but link to some content on another web page. While the task of clickbait detection has typically been cast as an assessment of teaser messages in isolation, the linked content may still play an important role for subsequent analyses. Only our two corpora Webis-Clickbait-16 and 17 incorporate the linked web pages.

Size. Finally, in terms of size, our corpus is on par with that of Rony et al. (2017), albeit, as discussed above, our sampling and annotation approaches are significantly more sophisticated.

3 Defining, Understanding, and Operationalizing Clickbait

Clickbait turns out to be an elusive concept. Characterizing clickbait as an enumeration of properties exhibited by a teaser message proves difficult. Having reviewed thousands of examples, we would say that clickbait withholds crucial information, and that it is emotional, sensational, condescending, inflammatory, know-it-all, or any combination thereof. However, this list of properties is likely not exhaustive, leaves lots of room for interpretation, and none appear to be a strict necessity. Furthermore, these properties are difficult to be operationalized. But despite the apparent difficulty of capturing in a clear-cut definition the essence of what makes clickbait clickbait, the term has been in use since 1999, according to Merriam Webster's. Apparently, the underlying phenomenon, namely that messages are spread designed to entice readers into clicking a link, has been around ever since the ad revenue business model took off on the web. And the phenomenon has been nagging enough so that web users took note of it and coined the term. This is evidence that people are able to tell apart clickbait from other forms of messages found online and gives us hope that machine learning technology, too, can be taught to detect clickbait.

A better way of understanding and explaining clickbait is by describing it in terms of what happens at publisher site and at reader site. The following four definitions contrast the emergence of clickbait and publisher intentions with the effects of clickbait and reader perception:

- *Emergence.* Clickbait is the result of optimizing teaser messages to maximize click-through.
- *Intention.* Clickbait are teaser messages whose authors intend to lure as many people as possible to a web page, disregarding the content's target audience.
- *Effects.* Clickbait are teaser messages that manipulate its readers into clicking a link.
- *Perception.* Clickbait are teaser messages perceived by (some) readers as bait to click a link.

These definitions capture important aspects of clickbait that a mere listing of properties cannot convey, and they reflect how clickbait is described, e.g., at Merriam Webster and on Wikipedia. The question remains, though, which of these definitions is best suited to *operationalize* clickbait.

For publishers, the availability of facilities to analyze user interactions with postings in real time on social networks naturally has given rise to their optimization to maximize click-through. Today's shape and form of clickbait can be attributed to Upworthy (Koechley, 2012), who "pioneered" and popularized clickbait by flooding Facebook at a scale that prompted Facebook to change its algorithms to reduce it (El-Arini and Tang, 2014). Other publishers quickly learned from Upworthy's example. To learn more, we asked professional journalists who use social network analytics on a daily basis: On the upside, seeing a recently finished article spread is gratifying, and they are thankful for the opportunity to adjust teasers, e.g., to fix mistakes or misconceptions. On the downside, this feedback about personal success or failure is also available to their department heads. When encountering clickbait, some form of malicious intent, be it voluntary or forced, can hence be presumed. From a reader's perspective, clickbait is mainly a form of manipulation. The manipulative power of clickbait has been attributed to the curiosity gap (Loewenstein, 1994), a gap of knowledge created in a reader's mind by withholding crucial information from a teaser message. Once experienced, the gap causes an urge to be filled, which may be explained as a kind of curiosity. This is at least what is proclaimed by Koechley (2012) as to why clickbait worked for Upworthy, and what has henceforth been cited. The connection between Loewenstein's theory and clickbait has not been shown scientifically, though, so that it remains an open question how the relatively short teaser messages can have such a strong effect. More generally, we say that clickbait appeals to human urges and exploits cognitive dissonances to fulfill its goal.

Our best bet at operationalizing clickbait is the fact that clickbait is perceptible to (some) readers, whereas proving malicious intent of individual journalists or publishers as a whole (i.e., reputation-based annotation) is virtually impossible. An obvious limitation of the former operationalization is that we cannot capture clickbait that eludes conscious reflection while still successfully manipulating (some) readers. This is why each teaser is judged by more than one annotator, and why our assessors have the freedom to judge the baitiness of a teaser message on a Likert scale rather enforcing a binary decision.

4 Corpus Acquisition

In order to make a valuable contribution to the research community, we set out to render the Webis Clickbait Corpus 2017 authentic, representative, rich in terms of potential features, unbiased, and large-scale. An overview of the main characteristics of our acquisition approach is provided in Table 2. We follow and extend the approach taken to construct our previous, small Webis-Clickbait-16 corpus (Potthast et al., 2016), rendering both corpora comparable. In what follows, we outline the design choices we were required to make regarding the selection of publishers, the crawling of their news, and the sampling of news items for annotation.

4.1 Platform and Publisher Selection

As teaser type for our corpus, we chose Twitter tweets. Twitter was chosen because (1) the platform has a large user base, and (2) virtually all major US news publishers disseminate their articles through Twitter. Facebook would have been equally suited, but we opted for Twitter for its more rigorous limitations of shape and form of teaser messages. This way, we hoped to ensure less variability in terms of how clickbait can be constructed, lest the more free-form teaser messages on Facebook may have caused more ambiguity. At any rate, all publishers under consideration disseminate their news on both platforms simultaneously. A cursory comparison of teasers shared on both platforms for the same news item shows that their contents are typically roughly equivalent, so that our platform choice may not significantly limit the generalizability of results obtained based on our corpus.

Our sample of news publishers is governed by publisher importance in terms of retweets. Restricting ourselves to English-language publishers, we obtain a ranking of the top-most retweeted news publishers from the NewsWhip social media analytics service.[2] Taking the top 27 publishers,[3] we ensured that the news covered reflect what a large population of English-speaking internet users is confronted with on a daily basis. Although many of the top-most publishers (say, the top 10) remain the same over long time spans, others make an appearance on lower ranks and are displaced again as time goes by and publishers compete for user attention. Our sample of publishers reflects the ranking of October 2016.

4.2 Crawling, Archiving, and Preprocessing News Teasers and Articles

Given the Twitter handles of the 27 news publishers selected, we used Twitter's API to record every tweet they published in the period from December 1, 2016, through April 30, 2017. To enable the research community to develop and experiment with a rich set of features, we included the tweet text, media attachements, and the meta data provided by Twitter. Due to technical limitations we had to omit tweets that contained video content. Further discarding tweets that did not contain exactly one hyperlink, we crawled the news article advertised. Besides technical difficulties arising from the HTTP forwarding associated with links on Twitter as well as publisher-specific forwarding, we found it insufficient to only download the HTML content of the linked news articles: many news publishers paginate articles, and some are at the forefront of dynamic web page design, so that we found cases where the article content was not part of the initial download, but resulted from JavaScript-based content loading. That, and the fact that the articles of some publishers tend to disappear or be reorganized rather quickly after publication, led us to adopt a more reproducible archiving approach. Using the tools underlying the Wayback Machine of the Internet Archive, we recorded the whole communication that takes place between a client (browser) requesting the web page of a news article and the publisher's web server hosting it, storing it within web archive (WARC) files (including, e.g., HTML, CSS, Javascript, and images). This way, every article page that forms part of our corpus can be reviewed as it was on the day we crawled it, allowing for corpus reviews even after years, hence maximizing its reproducibility.

Nevertheless, users of our corpus will not have to handle the raw WARC files. For convenience, we applied publisher-specific wrappers extracting a set of content-related fields (cf. fields prefixed with "target" in Table 2). In this regard, we discarded Boilerpipe and other generic content extraction tools, since their performance was surprisingly poor.

[2] https://www.newswhip.com

[3] We aimed for the top 30, but accidentally used the wrong Twitter handles of three, so that we excluded them altogether.

Corpus Acquisition	
Platform:	Twitter
Crawling period:	Dec 1 2016 – Apr 30 2017
Crawled tweets:	459,541
Publishers:	27. Names: abc, bbcworld, billboard, bleacherreport, breitbartnews, business, businessinsider, buzzfeed, cbsnews, cnn, complex, espn, forbes, foxnews, guardian, huffpost, independent, indiatimes, mailonline, mashable, nbcnews, nytimes, telegraph, usatoday, washingtonpost, wsj, yahoo
Filters:	- No videos in tweets. - Exactly one hyperlink in tweet. - Article archiving succeeded. - Main content extraction succeeded.
Recorded fields:	12. Names: postId, postTimestamp, postText, postMedia, postPublisher, targetUrl, targetTitle, targetDescription, targetKeywords, targetParagraphs, targetCaptions, targetWarcArchive
Sampling strategy:	Maximally 10 tweets per day and publisher
Sampled tweets:	38,517

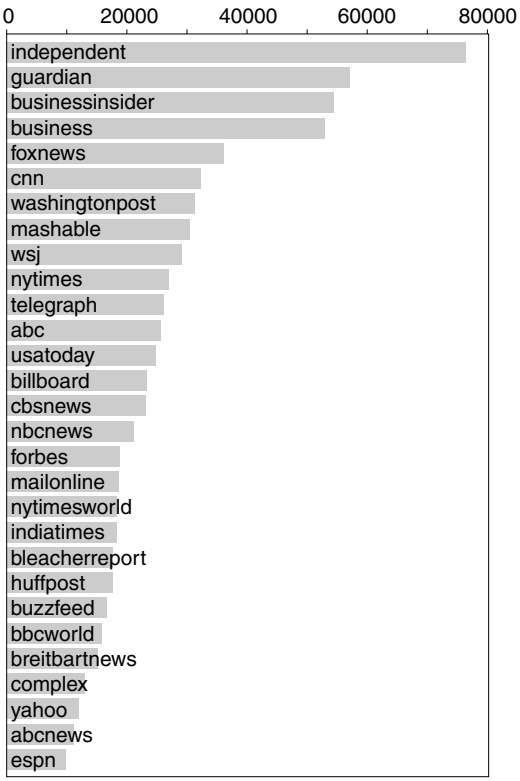

Table 2: Corpus acquisition overview (left), and number of tweets crawled from every publisher (right).

4.3 Sampling News

When sampling from a stream of news, special attention must be paid to avoid topic bias. This pertains particularly to the task of constructing a clickbait corpus, since clickbait is more a matter of teaser style rather than teaser topic. A clickbait detection model trained on too narrow a range of topics may not generalize well since it will likely pick up stylistic differences that result from differences between topics rather than differences between the style of clickbait and that of non-clickbait. Our previously published corpus Webis-Clickbait-16 suffers from this shortcoming, since it covers only one week worth of tweets. Although news coverage switches topics rather quickly, this is hardly enough to ensure topic diversity. To obtain a sample of tweets that has a high topic diversity, we crawled news as described above for five months in a row, yielding almost half a million tweets that fit our criteria and that were successfully archived.

From this population, we drew a random sample for annotation where, for budgetary reasons, the goal was to draw at least 30,000 tweets and at most 40,000. Since the distribution of tweets per publisher is highly skewed, we apply stratified sampling to avoid a corresponding publisher bias. Similarly, we ensure that tweets are sampled from each day of the five months worth of tweets to ensure coverage of the whole time period. Selecting a maximum of ten tweets per day and publisher yielded a set of 38,517 tweets and archived articles, which were then subjected to manual annotation.

5 Corpus Annotation

The annotation of the acquired tweets regarding their clickbaitiness was implemented with the crowdsourcing platform Amazon Mechanical Turk (AMT). To obtain high-quality annotations we invested substantial effort into preliminary evaluations of different task designs, alternative annotation scales, and, in particular, into the review of the annotations made. The following subsections detail our considerations, experiences, and insights.

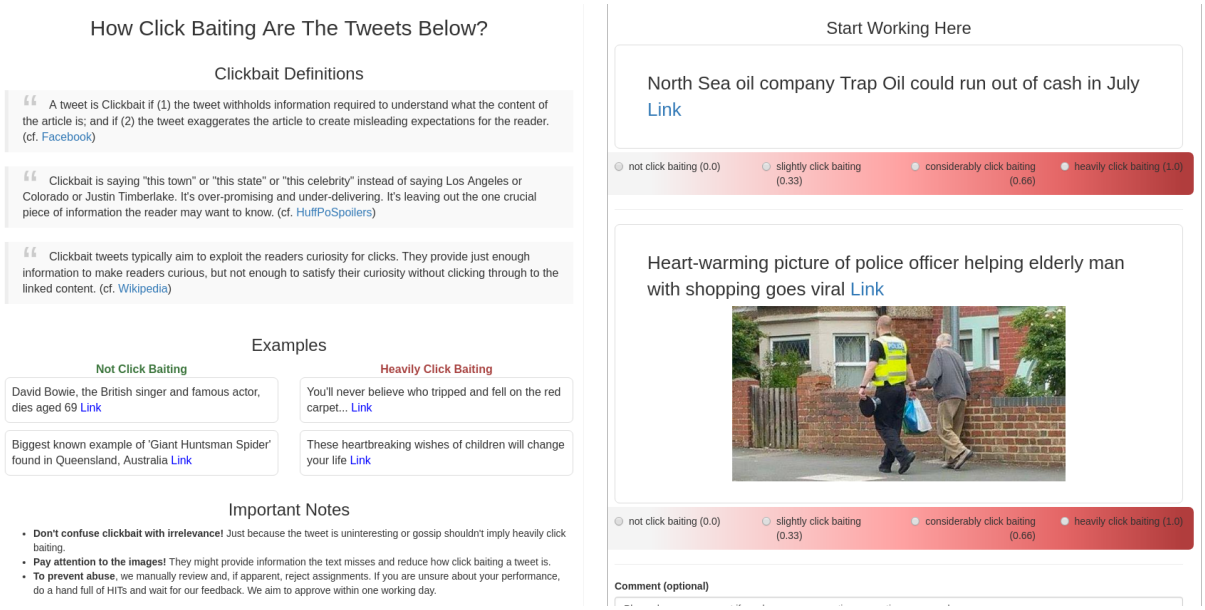

Figure 2: Screenshot of the final AMT task design.

5.1 Task Design

Designing a successful crowdsourcing task entails a number of design decisions, which can hardly be made without pilot studies. We conducted a series of AMT pilot studies to optimize the task design. In this regard, we employed the 2992 tweets of our previous clickbait corpus Webis-Clickbait-16. Figure 2 shows the final task design we used.

Clickbait is a concept with which crowd workers may not be familiar. To develop their "mental clickbait model," we quoted colloquial clickbait definitions from Facebook, HuffPoSpoilers, and Wikipedia as part of the task instructions (i.e., the aforementioned definitions are not suited for laymen). In addition, two typical clickbait and two non-clickbait examples were shown, drawn from our previous corpus. Since the annotations of the first pilot studies' tweets still showed an overly high variance, we extended the instructions again: first, since gossip was often misclassified as clickbait, we added a note that workers should not mix up irrelevance with clickbait. Second, since information in images attached to tweets was often overlooked, we added a note that special attention has to be paid to images. To avoid that workers get a high number of rejections because of misunderstanding the task, we suggested to complete at most two task instances and to await their approval before continuing.

The task instructions were followed by a list of tweets, each of which to be annotated regarding its clickbaitiness. We compared different annotation scales, the lengths of the tweet lists (five versus ten), and the number of *check instances* (one or two). As check instances we employed those tweets where all workers agreed; in order to bootstrap a pool of check instances, we started with tweets from our previous corpus. Check instances are of great value for the assessment of crowd workers. In our case it turned out that having two check instances per task instance—one that is clickbait and one that is not—allowed us to pinpoint underperforming workers of two kinds: (1) those with a tendency towards one of the extreme values, and (2) those who do not use the entire grading scale. Because of the (high) number of required check instances, we decided to have ten instead of five tweets per task in the final task design.

Within all studies, we payed one cent per tweet, an amount commensurate with the average time it took workers to complete a task so that a reasonable, albeit not generous, hourly wage would result. An exception was a particular pilot study where, in addition to the assessment of clickbait, workers had to mark words in the tweets which they considered indicative for their judgments. Having such a subtask was recently recommended for a search result relevance annotation task (McDonnell et al., 2016). In our case, this additional task increased the overall task completion time considerably while not improving worker agreement. We hence omitted this additional step in the final task design.

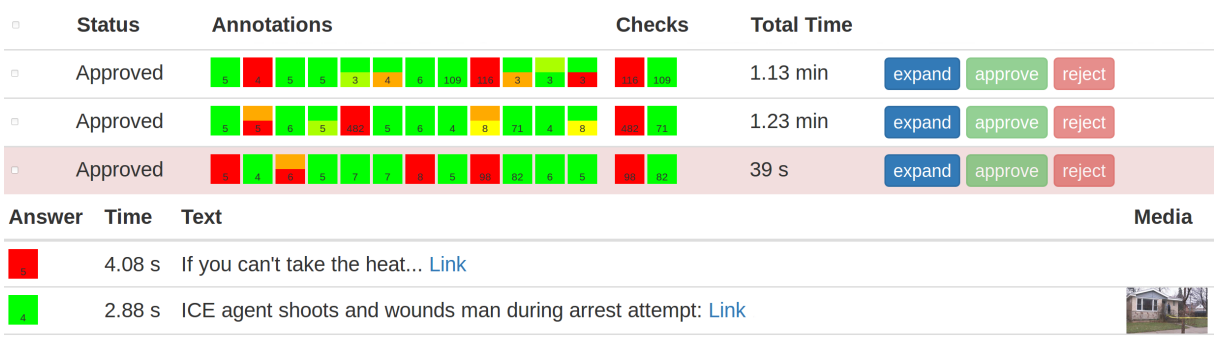

Figure 3: Screenshot of the system used for reviewing the crowdsourced clickbait assessments. Each row in the table refers to a HIT that has been worked on by a specific crowd worker.

5.2 Annotation Scale

Existing clickbait corpora use a binary scale for assessing clickbait. Although there are teaser messages that are obviously clickbait ("You won't believe what happened next!"), we noticed that making a binary decision about a teaser message is often not useful: even a teaser that can be rephrased in a more informative way may not be considered as clickbait if it conveys the idea of what the linked article is about. Hence, to account for different degrees of clickbait, we compared the use of a graded scale to the use of a binary scale. The graded scale has four values (Likert scale with forced choice) and ranges from "not" over "slightly" and "considerably" to "heavily" clickbaiting. For the ten annotations that we collected for each of the 2992 tweets in our previous corpus, the binary scale achieved an agreement of 85% (on average 8.5 annotators voted for the majority class). For the four point graded scale, the agreement dropped to 63%. However, when binarizing these annotations by combining the first two and the second two classes, an agreement of 84% is obtained. This shows that the graded scale is the more flexible choice: it can be abstracted into a binary scale without loss of agreement, and it allows for casting clickbait detection as a regression problem.

5.3 Reviewing Process

To guarantee a high quality dataset, all crowdsourced assessments were reviewed and, if necessary, discarded, resubmitting the respective assignment to AMT. Figure 3 shows a screenshot of the system that we have developed for efficient review of massive amounts of annotations: each row in the table shows a set of annotations (a single assignment) that has been submitted by a crowd worker. The Annotations column in turn shows a colored square for each tweet of the assignment. The upper half of a square color-codes the annotations made by the specific crowd worker, the lower half color-codes how (and how many) other workers annotated the tweet. By default, the color encodes the statistical mode of the annotations; other statistical location parameters such as mean and median are available from a menu. As mentioned above, each assignment included two check instances, which were displayed in the Checks column and which serve as an effective heads-up indicator: only if a worker frequently misclassified check instances, a more in-depth reviewing of the annotations was conducted. To this end, each of the rows in the table is expandable to display the tweets underlying the squares. As a further indicator, task completion time was measured and displayed as well (see the Total Time column). Obviously, a low agreement paired with a short completion time gives a strong bias that a worker did not work conscientiously. Altogether, more than 600 000 individual annotations were reviewed with the system for the clickbait corpus project; our rejection rate was 35.9%.

5.4 Corpus Analysis

The histogram in Figure 4 (left) shows the distribution of tweets across the four classes of our graded scale in the form of stacked bars. To classify a tweet into one of the four classes, the mode of its annotations is used, where, in case of multiple modes, the fifth annotation is used to determine the class. The different colors in the bars encode different levels of agreement. With a value of 0.21 in terms of Fleiss' κ, the annotator agreement is between slight and fair. However, when binarizing the classes

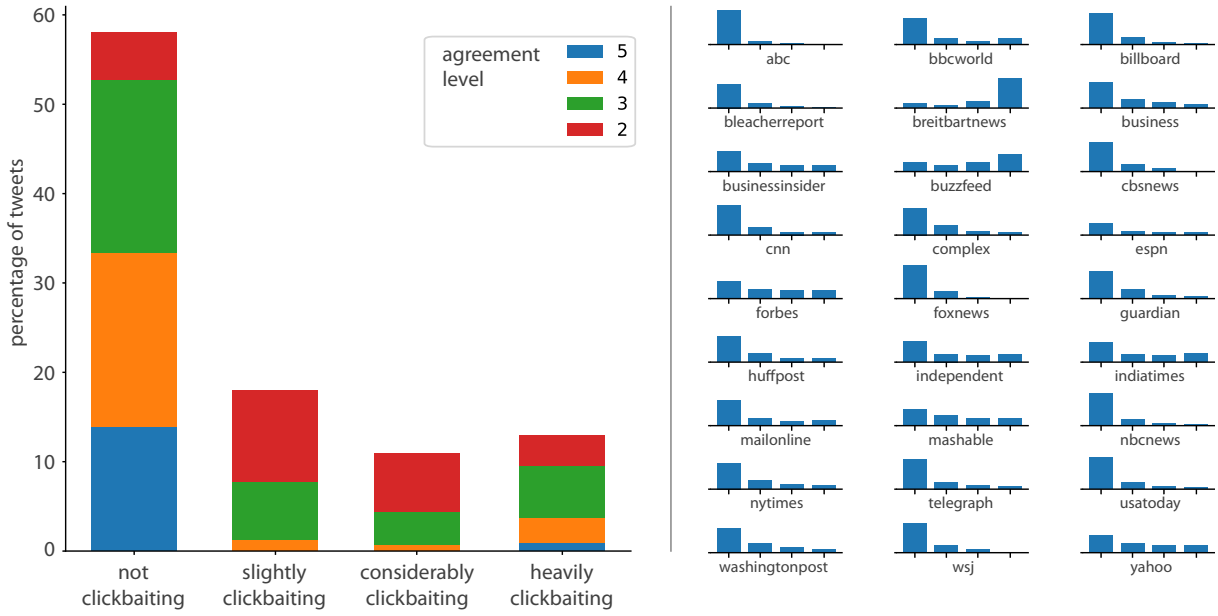

Figure 4: Left: Distribution of tweets over the four clickbait classes; the level of agreement about the class is color-coded. Right: Class distribution per publisher; no color coding is applied.

as described before, κ becomes 0.36, which corresponds to the respective value of 0.35 reported for our previous clickbait corpus (Potthast et al., 2016). Also note that the distribution of tweets across the binarized classes matches that of our previous corpus. Recalling that our previous corpus has been assessed by few trained experts, we conclude that our crowdsourcing strategy lives up to the state of the art and that it can be considered as successful: the two independently designed and operationalized annotation studies still achieve the same result, and hence our annotation experiment can be understood as a reproduction of our previous efforts, only at a larger scale. We summarize this results as follows: to scale the annotation of clickbait corpora, resorting to trained experts is not a necessary requirement.

Figure 4 (right) shows the distribution of tweets across the classes per publisher. Two of the 27 publishers, breitbartnews and buzzfeed, do obviously not follow the overall distribution; both send significantly more clickbait than the others. However, their contribution to the total amount of clickbaiting tweets (and hence to the corpus' publisher bias) is moderate only. Interestingly, TV networks in particular are among the publisher with least clickbait, avoiding even weakly clickbaiting messages.

6 Conclusion

Clickbait has quickly become one of the pests of social media, not unlike spam for email. While its working mechanisms are still barely understood, nearly every major publisher on social media employs it already to a greater or lesser extent to increase their readership. Although clickbait is apparently an effective marketing instrument, the ends do not justify the means: clickbait is also an effective instrument of manipulation. For example, the propagators of fake news make use of clickbait, too, to spread their disinformation. For news dissemination in general, clickbait must be rejected as well, since readers expect to be comprehensively informed, and journalistic codes of ethics typically prohibit unethical means of marketing.

To lay the groundwork for the emerging and ongoing investigation of clickbait by computer linguists and natural language processing alike, we have constructed the Webis Clickbait Corpus 2017, the first large-scale corpus of clickbait which has been carefully designed to be as representative as possible of teaser messages sent by the major publishers on Twitter. This way, our corpus will serve qualitative as well as quantitative research alike, allowing for deeper insights into the nature of clickbait, and the construction of technology to handle it, respectively. To foster the emerging research community on this subject, we share the resource itself and all the technologies that helped to compile it free of charge under permissible open source licenses.

To kickstart the use of our corpus, and to foster the development of new detection technology for clickbait messages, we have organized the Clickbait Challenge 2017, a shared task which used our corpus for evaluation. The corpus was divided into a training and test set, each comprising roughly half of the total number of tweets. Twelve teams of researchers participated in the challenge, submitting as many approaches for evaluation. A detailed review of the submitted approaches is out of the scope of this paper, however, it can be found in (Potthast et al., 2018).

References

A. Agrawal. 2016. Clickbait detection using deep learning. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pages 268–272, Oct.

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, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 94–100.

Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, San Francisco, CA, USA, August 18-21, 2016*, pages 9–16.

Khalid El-Arini and Joyce Tang. 2014. News Feed FYI: Click-baiting. http://web.archive.org/web/20150529104738/http://newsroom.fb.com/news/2014/08/news-feed-fyi-click-baiting/.

Peter Koechley. 2012. Why The Title Matters More Than The Talk. http://web.archive.org/web/20150611110506/http://blog.upworthy.com/post/26345634089/why-the-title-matters-more-than-the-talk.

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

Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why is that relevant? collecting annotator rationales for relevance judgments. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait Detection. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, *Advances in Information Retrieval. 38th European Conference on IR Research (ECIR 16)*, volume 9626 of *Lecture Notes in Computer Science*, pages 810–817, Berlin Heidelberg New York, March. Springer.

Martin Potthast, Tim Gollub, Matthias Hagen, and Benno Stein. 2018. The Clickbait Challenge 2017: Towards a Regression Model for Clickbait Strength. In *Proceddings of the Clickbait Chhallenge (to appear)*.

Jiani Qu, Anny Marleen Hißbach, Tim Gollub, and Martin Potthast. 2018. Towards Crowdsourcing Clickbait Labels for YouTube Videos. In *Proceedings of the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 18)*, July.

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? *CoRR*, abs/1703.09400.