Crowdsourcing a Large Corpus of Clickbait on Twitter

Today, **teasers** in social media are the primary means to disseminate online news.

Since only clicks generate **revenue**, publishers started to **fish for clicks** by omitting or exaggerating vital information in the teasers. This practice is called **clickbaiting**.

Clickbaiting **harms the quality** of social media, much as spam does for e-mail. To advance automatic **clickbait detection**, we compiled a large Twitter clickbait dataset.

1 Aquisition

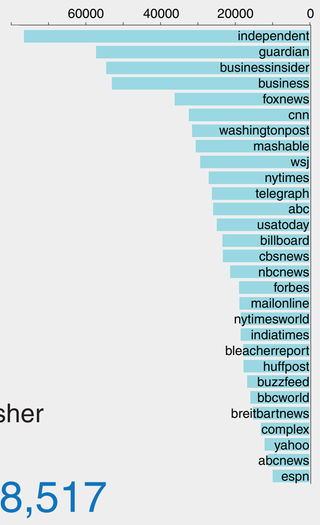
Observed: 27 publishers

Filters:
- No videos in tweets
- Exactly one hyperlink in tweet
- Article archiving succeeded
- Main content extraction succeeded

Crawling period: 5 month
Dec. 1, 2016 - Apr. 30, 2017

Sampling strategy
Maximum 10 tweets per day and publisher to avoid topical and publisher biases.

Totally sampled tweets: 38,517



2 Annotation

For each tweet, we crowdsourced **five clickbait annotations**.

Preliminary studies showed that a **4-point scale** is better to assess clickbait than a binary scale.

We use the annotations' mode as the **clickbait score** of a tweet.

A tweet is Clickbait if (1) the tweet withholds information required to undestand what the content of the article is, and if (2) the tweet exaggerates the article to create misleading expectations for the reader. —*Facebook*

Clickbait is saying "this town" or "this state" or "this celebrity" instead of saying Los Angeles or Colorado or Justin Timberlake. It's over-promising and under-delivering. It's leaving one crucial piece of information the reader may want to know. —*HuffPoSpoilers*

We provided **definitions** ↑ and **examples** ↓ to the annotators.

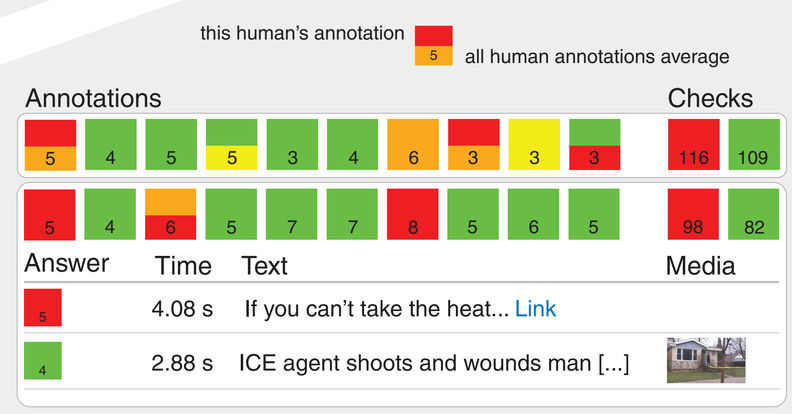
Not Click Baiting	Heavily Click Baiting
David Bowie, the British singer and famous actor, dies aged 69 Link	You'll never belive who tripped and fell on the red carpet.. Link
Biggest known example of 'Ginat Huntsman Spider' found in Queensland, Australia Link	These heartbreaking wishes of children will change your life Link

3 Review

To assure **annotation quality**, we manually inspected all annotations. If annotators frequently missed checks or deviated from others, their annotations were rejected.

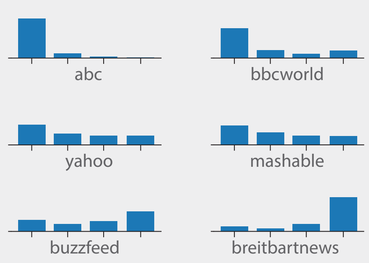
We had to **reject and repeat ~30%** of all annotations.

In total, we inspected more than **250.000 annotations**.



Interface for quick, visual review of annotations. Each line contains 12 annotations for different tweets, 2 of which are check instances.

4 Statistics



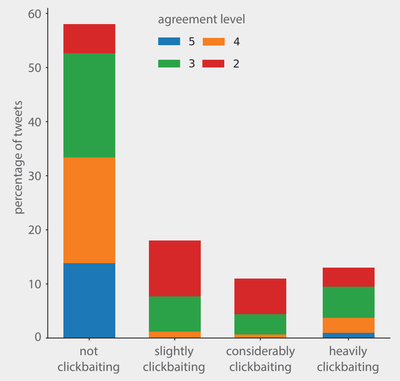
Selection of per-publisher distributions over the 4-point clickbait scale.

Most publishers follow the **general distribution** →

← **Buzzfeed** and **Breitbart** are notable exceptions

← **TV networks** are particularly non-clickbait

General distribution of clickbait classes with annotator agreement.



5 Release

The corpus has been used for the **clickbait challenge 2017**.

Team	MSE	F1	Prec	Rec	Acc
albacore	0.032	0.670	0.732	0.619	0.855
zingel	0.033	0.683	0.719	0.650	0.856
anchovy	0.034	0.679	0.717	0.645	0.855
emperor	0.036	0.641	0.714	0.581	0.845
...					
baseline	0.044	0.552	0.758	0.434	0.832

Current Clickbait-Challenge Leaderboard. Participation is still possible upon request.

www.clickbait-challenge.org

supported by

Google | DIGITAL NEWS INITIATIVE

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

(1) Leipzig University *firstname.lastname@uni-leipzig.de*
(2) Bauhaus-Universität Weimar *firstname.lastname@uni-weimar.de*
(3) German Aerospace Center (DLR) *firstname.lastname@dlr.de*
(4) Martin-Luther-Universität Halle-Wittenberg *firstname.lastname@uni-halle.de*